

MIXED FINITE ELEMENT FORMULATION AND ERROR ESTIMATES BASED ON PROPER ORTHOGONAL DECOMPOSITION FOR THE NONSTATIONARY NAVIER–STOKES EQUATIONS*

ZHENDONG LUO[†], JING CHEN[‡], I. M. NAVON[§], AND XIAOZHONG YANG[†]

Abstract. In this paper, proper orthogonal decomposition (POD) is used for model reduction of mixed finite element (MFE) for the nonstationary Navier–Stokes equations and error estimates between a reference solution and the POD solution of reduced MFE formulation are derived. The basic idea of this reduction technique is that ensembles of data are first compiled from transient solutions computed equation system derived with the usual MFE method for the nonstationary Navier–Stokes equations or from physics system trajectories by drawing samples of experiments and interpolation (or data assimilation), and then the basis functions of the usual MFE method are substituted with the POD basis functions reconstructed by the elements of the ensemble to derive the POD-reduced MFE formulation for the nonstationary Navier–Stokes equations. It is shown by considering numerical simulation results obtained for the illustrating example of cavity flows that the error between POD solution of reduced MFE formulation and the reference solution is consistent with theoretical results. Moreover, it is also shown that this result validates the feasibility and efficiency of the POD method.

Key words. mixed finite element method, proper orthogonal decomposition, the nonstationary Navier–Stokes equations, error estimate

AMS subject classifications. 65N30, 35Q10

DOI. 10.1137/070689498

1. Introduction. The mixed finite element (MFE) method is one of the important approaches for solving systems of partial differential equations, for example, the nonstationary Navier–Stokes equations (see [1], [2], or [3]). However, the computational model for the fully discrete system of MFE solutions of the nonstationary Navier–Stokes equations yields very large systems that are computationally intensive. Thus, an important problem is how to simplify the computational load and save time-consuming calculations and resource demands in the actual computational process in a way that guarantees a sufficiently accurate and efficient numerical solution. Proper orthogonal decomposition (POD), also known as Karhunen–Loève expansions in signal analysis and pattern recognition (see [4]), or principal component analysis in statistics (see [5]), or the method of empirical orthogonal functions in geophysical fluid dynamics (see [6], [7]) or meteorology (see [8]), is a technique offering adequate approximation for representing fluid flow with reduced number of degrees of freedom, i.e., with lower dimensional models (see [9]), so as to alleviate the computational load

*Received by the editors April 25, 2007; accepted for publication (in revised form) May 19, 2008; published electronically October 24, 2008.

<http://www.siam.org/journals/sinum/47-1/68949.html>

[†]School of Mathematics and Physics, North China Electric Power University, Beijing 102206, China (zhdluo@163.com). This work was supported in part by the National Science Foundation of China (NSF10871022 and 10771065).

[‡]Corresponding author. College of Science, China Agricultural University, Beijing 100083, China (jing_quchen@163.com). This author was supported in part by the National Science Foundation of China (NSF10871022 and 10771213).

[§]Department of Scientific Computing, Florida State University, Dirac Sci. Lib. Bldg., #483, Tallahassee, FL 32306-4120 (navon@scs.fsu.edu). This author was supported in part by NASA MAP grant Modeling, Analysis and Prediction Program (NNG06GC67G).

and provide CPU and memory requirements savings, and has found widespread applications in problems related to the approximation of large-scale models. Although the basic properties of the POD method are well established and studies have been conducted to evaluate the suitability of this technique for various fluid flows (see [10]–[12]), its applicability and limitations for reduced MFE formulation for the non-stationary Navier–Stokes equations are not well documented.

The POD method mainly provides a useful tool for efficiently approximating a large amount of data. The method essentially provides an orthogonal basis for representing the given data in a certain least squares optimal sense; that is, it provides a way to find optimal lower dimensional approximations of the given data. In addition to being optimal in a least squares sense, POD has the property that it uses a modal decomposition that is completely data dependent and does not assume any prior knowledge of the process used to generate the data. This property is advantageous in situations where a priori knowledge of the underlying process is insufficient to warrant a certain choice of basis. Combined with the Galerkin projection procedure, POD provides a powerful method for generating lower dimensional models of dynamical systems that have a very large or even infinite dimensional phase space. In many cases, the behavior of a dynamic system is governed by characteristics or related structures, even though the ensemble is formed by a large number of different instantaneous solutions. POD method can capture these temporal and spatial structures by applying a statistical analysis to the ensemble of data. In fluid dynamics, Lumley first employed the POD technique to capture the large eddy coherent structures in a turbulent boundary layer (see [13]); this technique was further extended in [14], where a link between the turbulent structure and dynamics of a chaotic system was investigated. In Holmes, Lumley, and Berkooz [9], the overall properties of POD are reviewed and extended to widen the applicability of the method. The method of snapshots was introduced by Sirovich [15], and is widely used in applications to reduce the order of POD eigenvalue problem. Examples of these are optimal flow control problems [16]–[18] and turbulence [9, 13, 14, 19, 20]. In many applications of POD, the method is used to generate basis functions for a reduced order model, which can simplify and provide quicker assessment of the major features of the fluid dynamics for the purpose of flow control as demonstrated by Ko et al. [18] or design optimization as shown by Ly and Tran [17]. This application is used in a variety of other physical applications, such as in [17], which demonstrates an effective use of POD for a chemical vapor deposition reactor. Some reduced order finite difference models and MFE formulations and error estimates based on POD for the upper tropical Pacific Ocean model (see [21]–[25]), as well as a finite difference scheme based on POD for the nonstationary Navier–Stokes equations (see [26]), have been derived. However, to the best of our knowledge, there are no published results addressing the use of POD to reduce the MFE formulation of the nonlinear nonstationary Navier–Stokes equations and provide estimates of the error between reference solution and the POD-reduced MFE solution.

In this paper, POD is used to reduce the MFE formulation for the nonstationary Navier–Stokes equations and to derive error estimates between reference solution and the POD-reduced MFE solution. It is shown by considering the results obtained for numerical simulations of cavity flows that the error between POD solution of reduced MFE formulation and reference solution is consistent with theoretically derived results. Moreover, it is also shown that this validates the feasibility and efficiency of the POD method. Though Kunisch and Volkwein have presented some Galerkin POD methods for parabolic problems and a general equation in fluid dynamics in [27], [28],

our method is different from their approaches, whose methods consist of Galerkin projection where the original variables are substituted for a linear combination of POD basis and the error estimates of the velocity field therein are only derived, their POD basis being generated with the solution of the physical system at all time instances. In particular, the velocity field is only approximated in [28], while both velocity and pressure fields are simultaneously approximated in our present method. While the singular value decomposition approach combined with POD methodology is used to treat the Burgers equation in [29] and the cavity flow problem in [12], the error estimates have not completely been derived, in particular, a reduced formulation of MFE for the nonstationary Navier–Stokes has not yet been derived up to now. Therefore, our method improves upon existing methods since our POD basis is generated with the solution of the physical system only at time instances which are both useful and of interest for us.

2. MFE approximation for the nonstationary Navier–Stokes equations and snapshots generate. Let $\Omega \subset R^2$ be a bounded, connected, and polygonal domain. Consider the following nonstationary Navier–Stokes equations.

Problem (I) Find $\mathbf{u} = (u_1, u_2)$, p such that, for $T > 0$,

$$(2.1) \quad \begin{cases} \mathbf{u}_t - \nu \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla) \mathbf{u} + \nabla p = \mathbf{f} & \text{in } \Omega \times (0, T), \\ \operatorname{div} \mathbf{u} = 0 & \text{in } \Omega \times (0, T), \\ \mathbf{u}(x, y, t) = \boldsymbol{\varphi}(x, y, t) & \text{on } \partial\Omega \times (0, T), \\ \mathbf{u}(x, y, 0) = \boldsymbol{\varphi}(x, y, 0) & \text{in } \Omega, \end{cases}$$

where \mathbf{u} represents the velocity vector, p the pressure, ν the constant inverse Reynolds number, $\mathbf{f} = (f_1, f_2)$ the given body force, and $\boldsymbol{\varphi}(x, y, t)$ the given vector function. For the sake of convenience, without lost generality, we may as well suppose that $\boldsymbol{\varphi}(x, y, t)$ is a zero vector in the following theoretical analysis.

The Sobolev spaces used in this context are standard (see [30]). For example, for a bounded domain Ω , we denote by $H^m(\Omega)$ ($m \geq 0$) and $L^2(\Omega) = H^0(\Omega)$ the usual Sobolev spaces equipped with the seminorm and the norm, respectively,

$$|v|_{m,\Omega} = \left\{ \sum_{|\boldsymbol{\alpha}|=m} \int_{\Omega} |D^{\boldsymbol{\alpha}} v|^2 dx dy \right\}^{1/2} \quad \text{and} \quad \|v\|_{m,\Omega} = \left\{ \sum_{i=0}^m |v|_{i,\Omega}^2 \right\}^{1/2} \quad \forall v \in H^m(\Omega),$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$, α_1 and α_2 are two nonnegative integers, and $|\boldsymbol{\alpha}| = \alpha_1 + \alpha_2$. Especially, the subspace $H_0^1(\Omega)$ of $H^1(\Omega)$ is denoted by

$$H_0^1(\Omega) = \{v \in H^1(\Omega); u|_{\partial\Omega} = 0\}.$$

Note that $\|\cdot\|_1$ is equivalent to $|\cdot|_1$ in $H_0^1(\Omega)$. Let $L_0^2(\Omega) = \{q \in L^2(\Omega); \int_{\Omega} q dx dy = 0\}$, which is a subspace of $L^2(\Omega)$. It is necessary to introduce the Sobolev spaces dependent on time t in order to discuss the generalized solution for Problem (I). Let Φ be a Hilbert space. For all $T > 0$ and integer $n \geq 0$, for $t \in [0, T]$, define

$$H^n(0, T; \Phi) = \left\{ v(t) \in \Phi; \int_0^T \sum_{i=0}^n \left\| \frac{d^i}{dt^i} v(t) \right\|_{\Phi}^2 dt < \infty \right\},$$

which is endowed with the norm

$$\|v\|_{H^n(\Phi)} = \left[\sum_{i=0}^n \int_0^T \left\| \frac{d^i}{dt^i} v(t) \right\|_{\Phi}^2 dt \right]^{\frac{1}{2}} \quad \text{for } v \in H^n(\Phi),$$

where $\|\cdot\|_{\Phi}$ is the norm of space Φ . Especially, if $n = 0$,

$$\|v\|_{L^2(\Phi)} = \left(\int_0^T \|v(t)\|_{\Phi}^2 dt \right)^{\frac{1}{2}}.$$

And define

$$L^\infty(0, T; \Phi) = \left\{ v(t) \in \Phi; \operatorname{ess\,sup}_{0 \leq t \leq T} \|v(t)\|_{\Phi} < \infty \right\},$$

which is endowed with the norm

$$\|v\|_{L^\infty(\Phi)} = \operatorname{ess\,sup}_{0 \leq t \leq T} \|v(t)\|_{\Phi}.$$

The variational formulation for Problem (I) is written as:

Problem (II) Find $(\mathbf{u}, p) \in H^1(0, T; X) \times L^2(0, T; M)$ such that, for all $t \in (0, T)$,

$$(2.2) \quad \begin{cases} (\mathbf{u}_t, \mathbf{v}) + a(\mathbf{u}, \mathbf{v}) + a_1(\mathbf{u}, \mathbf{u}, \mathbf{v}) - b(p, \mathbf{v}) = (\mathbf{f}, \mathbf{v}) \quad \forall \mathbf{v} \in X, \\ b(q, \mathbf{u}) = 0 \quad \forall q \in M, \\ \mathbf{u}(x, 0) = \mathbf{0} \quad \text{in } \Omega, \end{cases}$$

where $X = H_0^1(\Omega)^2$, $M = L_0^2(\Omega)$, $a(\mathbf{u}, \mathbf{v}) = \nu \int_{\Omega} \nabla \mathbf{u} \cdot \nabla \mathbf{v} dx dy$, $a_1(\mathbf{u}, \mathbf{v}, \mathbf{w}) = \frac{1}{2} \int_{\Omega} \sum_{i,j=1}^2 [u_i \frac{\partial v_j}{\partial x_i} w_j - u_i \frac{\partial w_j}{\partial x_i} v_j] dx dy$ ($\mathbf{u}, \mathbf{v}, \mathbf{w} \in X$), and $b(q, \mathbf{v}) = \int_{\Omega} q \operatorname{div} \mathbf{v} dx dy$.

Throughout the paper, C indicates a positive constant which is possibly different at different occurrences, being independent of the spatial and temporal mesh sizes, but may depend on Ω , the Reynolds number, and other parameters introduced in this paper.

The following property for trilinear form $a_1(\cdot, \cdot, \cdot)$ is often used (see [1], [2], or [3]).

$$(2.3) \quad a_1(\mathbf{u}, \mathbf{v}, \mathbf{w}) = -a_1(\mathbf{u}, \mathbf{w}, \mathbf{v}), \quad a_1(\mathbf{u}, \mathbf{v}, \mathbf{v}) = 0 \quad \forall \mathbf{u}, \mathbf{v}, \mathbf{w} \in X.$$

The bilinear forms $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ have the following properties:

$$(2.4) \quad a(\mathbf{v}, \mathbf{v}) \geq \nu |\mathbf{v}|_1^2 \quad \forall \mathbf{v} \in H_0^1(\Omega)^2,$$

$$(2.5) \quad |a(\mathbf{u}, \mathbf{v})| \leq \nu |u|_1 |\mathbf{v}|_1 \quad \forall \mathbf{u}, \mathbf{v} \in H_0^1(\Omega)^2,$$

and

$$(2.6) \quad \sup_{\mathbf{v} \in H_0^1(\Omega)^2} \frac{b(q, \mathbf{v})}{|\mathbf{v}|_1} \geq \beta \|q\|_0 \quad \forall q \in L_0^2(\Omega),$$

where β is a positive constant. Define

$$(2.7) \quad N = \sup_{\mathbf{u}, \mathbf{v}, \mathbf{w} \in X} \frac{a_1(\mathbf{u}, \mathbf{v}, \mathbf{w})}{|\mathbf{u}|_1 |\mathbf{v}|_1 |\mathbf{w}|_1}; \quad \|\mathbf{f}\|_{-1} = \sup_{\mathbf{v} \in X} \frac{(\mathbf{f}, \mathbf{v})}{|\mathbf{v}|_1}.$$

The following result is classical (see [1], [2], or [3]).

THEOREM 2.1. *If $\mathbf{f} \in L^2(0, T; H^{-1}(\Omega)^2)$, then Problem (II) has at least a solution which, in addition, is unique provided that $\nu^{-2} N \|\mathbf{f}\|_{L^2(H^{-1})} < 1$, and there is the following prior estimate:*

$$\|\nabla \mathbf{u}\|_{L^2(L^2)} \leq \nu^{-1} \|\mathbf{f}\|_{L^2(H^{-1})} \equiv R, \quad \|\mathbf{u}\|_0 \leq \nu^{-1/2} \|\mathbf{f}\|_{L^2(H^{-1})} = R\nu^{-1/2}.$$

Let $\{\mathfrak{S}_h\}$ be a uniformly regular family of triangulation of $\bar{\Omega}$ (see [31], [32], or [33]), indexed by a parameter $h = \max_{K \in \mathfrak{S}_h} \{h_K; h_K = \text{diam}(K)\}$; i.e., there exists a constant C , independent of h , such that $h \leq Ch_K \forall K \in \mathfrak{S}_h$.

We introduce the following finite element spaces X_h and M_h of X and M , respectively. Let $X_h \subset X$ (which is at least the piecewise polynomial vector space of m th degree, where $m > 0$ is integer) and $M_h \subset M$ (which is the piecewise polynomial space of $(m-1)$ th degree). Write $\hat{X}_h = X_h \times M_h$.

We assume that (X_h, M_h) satisfies the following approximate properties: $\forall v \in H^{m+1}(\Omega)^2 \cap X$ and $\forall q \in M \cap H^m(\Omega)$,

$$(2.8) \quad \inf_{\mathbf{v}_h \in X_h} \|\nabla(\mathbf{v} - \mathbf{v}_h)\|_0 \leq Ch^m |\mathbf{v}|_{m+1}, \quad \inf_{q_h \in M_h} \|q - q_h\|_0 \leq Ch^m |q|_m,$$

together the so-called discrete LBB condition, i.e.,

$$(2.9) \quad \sup_{\mathbf{v}_h \in X_h} \frac{b(q_h, \mathbf{v}_h)}{\|\nabla \mathbf{v}_h\|_0} \geq \beta \|q_h\|_0 \quad \forall q_h \in M_h,$$

where β is a positive constant independent of h .

There are many spaces X_h and M_h satisfying the discrete LBB conditions (see [33]). Here, we provide some examples as follows.

Example 2.1. The first-order finite element space $X_h \times M_h$ can be taken as Bernardi–Fortin–Raugel’s element (see [33]), i.e.,

$$(2.10) \quad \begin{aligned} X_h &= \{\mathbf{v}_h \in X \cap C^0(\bar{\Omega})^2; \mathbf{v}_h|_K \in P_K \quad \forall K \in \mathfrak{S}_h\}, \\ M_h &= \{\varphi_h \in M; \varphi_h|_K \in P_0(K) \quad \forall K \in \mathfrak{S}_h\}, \end{aligned}$$

where $P_K = P_1(K)^2 \oplus \text{span}\{\mathbf{n}_i \prod_{j=1, j \neq i}^3 \lambda_{Kj}, i = 1, 2, 3\}$, \mathbf{n}_i are the unit normal vector to side F_i opposite the vertex A_i of triangle K , λ_{Ki} ’s are the barycenter coordinates corresponding to the vertex A_i ($i = 1, 2, 3$) on K (see [31], [32]), and $P_m(K)$ is the space of piecewise polynomials of degree m on K .

Example 2.2. The first-order finite element space $X_h \times M_h$ can also be taken as Mini’s element, i.e.,

$$(2.11) \quad \begin{aligned} X_h &= \{\mathbf{v}_h \in X \cap C^0(\Omega)^2; \mathbf{v}_h|_K \in P_K \quad \forall K \in \mathfrak{S}_h\}, \\ M_h &= \{q_h \in M \cap C^0(\Omega); q_h|_K \in P_1(K) \quad \forall K \in \mathfrak{S}_h\}, \end{aligned}$$

where $P_K = P_1(K)^2 \oplus \text{span}\{\lambda_{K1}\lambda_{K2}\lambda_{K3}\}^2$.

Example 2.3. The second-order finite element space $X_h \times M_h$ can be taken as

$$(2.12) \quad \begin{aligned} X_h &= \{\mathbf{v}_h \in X \cap C^0(\Omega)^2; \mathbf{v}_h|_K \in P_K \quad \forall K \in \mathfrak{S}_h\}, \\ M_h &= \{q_h \in M \cap C^0(\Omega); q_h|_K \in P_1(K) \quad \forall K \in \mathfrak{S}_h\}, \end{aligned}$$

where $P_K = P_2(K)^2 \oplus \text{span}\{\lambda_{K1}\lambda_{K2}\lambda_{K3}\}^2$.

Example 2.4. The third-order finite element space $X_h \times M_h$ can be taken as

$$(2.13) \quad \begin{aligned} X_h &= \{\mathbf{v}_h \in X \cap C^0(\Omega)^2; \mathbf{v}_h|_K \in P_K \quad \forall K \in \mathfrak{S}_h\}, \\ M_h &= \{q_h \in M \cap C^0(\Omega); q_h|_K \in P_2(K) \quad \forall K \in \mathfrak{S}_h\}, \end{aligned}$$

where $P_K = P_3(K)^2 \oplus \text{span}\{\lambda_{K1}\lambda_{K2}\lambda_{K3}\lambda_{Ki}, i = 1, 2, 3\}^2$.

It has been proved (see [33]) that, for the finite element space $X_h \times M_h$ in Examples 2.1–2.4, there exists a restriction operator $r_h: X \rightarrow X_h$ such that, for any $\mathbf{v} \in X$,

$$(2.14) \quad \begin{aligned} b(q_h, \mathbf{v} - r_h \mathbf{v}) &= 0 \quad \forall q_h \in M_h, \quad \|\nabla r_h \mathbf{v}\|_0 \leq C \|\nabla \mathbf{v}\|_0, \\ \|\nabla(\mathbf{v} - r_h \mathbf{v})\|_0 &\leq Ch^k |\mathbf{v}|_{k+1} \quad \text{if } \mathbf{v} \in H^{k+1}(\Omega)^2, \quad k = 1, 2, 3. \end{aligned}$$

The spaces $X_h \times M_h$ used throughout the next part in this paper mean those in Examples 2.1–2.4, which satisfy the discrete LBB condition (2.9) (see [33] for a more detailed proof).

In order to find a numerical solution for Problem (II), it is necessary to discretize Problem (II). We introduce a MFE approximation for the spatial variable and FDS (finite difference scheme) for the time derivative. Let L be the positive integer, denote the time step increment by $k = T/L$ (T being the total time), $t^{(n)} = nk$, $0 \leq n \leq L$; $(\mathbf{u}_h^n, p_h^n) \in X_h \times M_h$ the MFE approximation corresponding to $(u(t^{(n)}), p(t^{(n)})) \equiv (\mathbf{u}^n, p^n)$. Then, applying a semi-implicit Euler scheme for the time integration, the fully discrete MFE solution for Problem (I) may be written as:

Problem (III) Find $(\mathbf{u}_h^n, p_h^n) \in X_h \times M_h$ such that

$$(2.15) \quad \begin{cases} (\mathbf{u}_h^n, \mathbf{v}_h) + ka(\mathbf{u}_h^n, \mathbf{v}_h) + ka_1(\mathbf{u}_h^{n-1}, \mathbf{u}_h^n, \mathbf{v}_h) - kb(p_h^n, \mathbf{v}_h) \\ \quad = k(f^n, \mathbf{v}_h) + (\mathbf{u}_h^{n-1}, \mathbf{v}_h) \quad \forall \mathbf{v}_h \in X_h, \\ b(q_h, \mathbf{u}_h^n) = 0 \quad \forall q_h \in M_h, \\ \mathbf{u}_h^0 = \mathbf{0} \quad \text{in } \Omega, \end{cases}$$

where $1 \leq n \leq L$.

Put $A(\mathbf{u}_h^n, \mathbf{v}_h) = (\mathbf{u}_h^n, \mathbf{v}_h) + ka(\mathbf{u}_h^n, \mathbf{v}_h) + ka_1(\mathbf{u}_h^{n-1}, \mathbf{u}_h^n, \mathbf{v}_h)$. Since $A(\mathbf{u}_h^n, \mathbf{u}_h^n) = (\mathbf{u}_h^n, \mathbf{u}_h^n) + ka(\mathbf{u}_h^n, \mathbf{u}_h^n) + ka_1(\mathbf{u}_h^{n-1}, \mathbf{u}_h^n, \mathbf{u}_h^n) = \|\mathbf{u}_h^n\|_0 + k\nu \|\nabla \mathbf{u}_h^n\|_0$, $A(\cdot, \cdot)$ is coercive in $X_h \times X_h$. And $kb(\cdot, \cdot)$ also satisfies the discrete LBB condition in $X_h \times M_h$; therefore, by MFE theory (see [1], [32], or [33]), we obtain the following result.

THEOREM 2.2. *Under the assumptions (2.8), (2.9), if $\mathbf{f} \in H^{-1}(\Omega)^2$ satisfies $N \sum_{i=1}^n \|\mathbf{f}^i\|_{-1} < \nu^2$, then Problem (III) has a unique solution $(\mathbf{u}_h^n, p_h^n) \in X_h \times M_h$ and satisfies*

$$(2.16) \quad \|\mathbf{u}_h^n\|_0^2 + k\nu \sum_{i=1}^n \|\nabla \mathbf{u}_h^i\|_0^2 \leq k\nu^{-1} \sum_{i=1}^n \|\mathbf{f}^i\|_{-1}^2,$$

if $k = O(h^2)$,

$$(2.17) \quad \|\mathbf{u}^n - \mathbf{u}_h^n\|_0 + k^{1/2} \sum_{i=1}^n \|\nabla(\mathbf{u}^i - \mathbf{u}_h^i)\|_0 + k^{1/2} \sum_{i=1}^n \|p^i - p_h^i\|_0 \leq C(h^m + k),$$

where $(\mathbf{u}, p) \in [H_0^1(\Omega) \cap H^{m+1}(\Omega)]^2 \times [H^m(\Omega) \cap M]$ is the exact solution for the problem (I), C is a constant dependent on $|\mathbf{u}^n|_{m+1}$ and $|p^n|_m$, and $1 \leq n \leq L$.

If Reynolds number $Re = \nu^{-1}$, triangulation parameter h , finite element space $X_h \times M_h$, the time step increment k , and \mathbf{f} are given, by solving Problem (III), we can obtain a solution ensemble $\{u_{1h}^n, u_{2h}^n, p_h^n\}_{n=1}^L$ for Problem (III). Then we choose ℓ (for example, $\ell = 20$, or 30 , in general, $\ell \ll L$) instantaneous solutions $\mathbf{U}_i(x, y) = (u_{1h}^{n_i}, u_{2h}^{n_i}, p_h^{n_i})^T$ ($1 \leq n_1 < n_2 < \dots < n_\ell \leq L$) (which are useful and of interest for us) from the L group of solutions $(u_{1h}^n, u_{2h}^n, p_h^n)^T$ ($1 \leq n \leq L$) for Problem (III), which are referred to as snapshots.

3. A reduced MFE formulation based POD technique for the nonstationary Navier–Stokes equations. In this section, we use the POD technique to deal with the snapshots in section 2 and produce an optimal representation in an average sense.

Recall $\hat{X}_h = X_h \times M_h$. For $\mathbf{U}_i(x, y) = (u_{1h}^{n_i}, u_{2h}^{n_i}, p_h^{n_i})^T$ ($i = 1, 2, \dots, \ell$) in section 2, we set

$$(3.1) \quad \mathcal{V} = \text{span}\{\mathbf{U}_1, \mathbf{U}_2, \dots, \mathbf{U}_\ell\},$$

and refer to \mathcal{V} as the ensemble consisting of the snapshots $\{\mathbf{U}_i\}_{i=1}^\ell$, at least one of which is assumed to be nonzero. Let $\{\boldsymbol{\psi}_j\}_{j=1}^l$ denote an orthonormal basis of \mathcal{V} with $l = \dim \mathcal{V}$. Then each member of the ensemble can be expressed as

$$(3.2) \quad \mathbf{U}_i = \sum_{j=1}^l (\mathbf{U}_i, \boldsymbol{\psi}_j)_{\hat{X}} \boldsymbol{\psi}_j \quad \text{for } i = 1, 2, \dots, \ell,$$

where $(\mathbf{U}_i, \boldsymbol{\psi}_j)_{\hat{X}} \boldsymbol{\psi}_j = ((\nabla \mathbf{u}_h^{n_i}, \nabla \boldsymbol{\psi}_{u_j})_0 \boldsymbol{\psi}_{u_j}, (p_h^{n_i}, \boldsymbol{\psi}_{p_j})_0 \boldsymbol{\psi}_{p_j})$, $(\cdot, \cdot)_0$ is L^2 -inner product, and $\boldsymbol{\psi}_{u_j}$ and $\boldsymbol{\psi}_{p_j}$ are orthonormal bases corresponding to \mathbf{u} and p , respectively.

Since $\mathcal{V} = \text{span}\{\mathbf{U}_1, \mathbf{U}_2, \dots, \mathbf{U}_\ell\} = \text{span}\{\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \dots, \boldsymbol{\psi}_l\}$, $b(p_h^{n_i}, \mathbf{u}_h^{n_j}) = 0$ ($1 \leq i, j \leq \ell$) implies $b(\boldsymbol{\psi}_{p_i}, \boldsymbol{\psi}_{u_j}) = 0$ ($1 \leq i, j \leq l$).

DEFINITION 3.1. *The method of POD consists in finding the orthonormal basis such that for every d ($1 \leq d \leq l$) the mean square error between the elements \mathbf{U}_i ($1 \leq i \leq \ell$) and corresponding d th partial sum of (3.2) is minimized on average:*

$$(3.3) \quad \min_{\{\boldsymbol{\psi}_j\}_{j=1}^d} \frac{1}{\ell} \sum_{i=1}^{\ell} \left\| \mathbf{U}_i - \sum_{j=1}^d (\mathbf{U}_i, \boldsymbol{\psi}_j)_{\hat{X}} \boldsymbol{\psi}_j \right\|_{\hat{X}}^2$$

such that

$$(3.4) \quad (\boldsymbol{\psi}_i, \boldsymbol{\psi}_j)_{\hat{X}} = \delta_{ij} \quad \text{for } 1 \leq i \leq d, 1 \leq j \leq i,$$

where $\|\mathbf{U}_i\|_{\hat{X}} = [\|\nabla \mathbf{u}_h^{n_i}\|_0^2 + \|\nabla \mathbf{u}_{2h}^{n_i}\|_0^2 + \|p_h^{n_i}\|_0^2]^{\frac{1}{2}}$. A solution $\{\boldsymbol{\psi}_j\}_{j=1}^d$ of (3.3) and (3.4) is known as a POD basis of rank d .

We introduce the correlation matrix $\mathbf{K} = (K_{ij})_{\ell \times \ell} \in R^{\ell \times \ell}$ corresponding to the snapshots $\{\mathbf{U}_i\}_{i=1}^\ell$ by

$$(3.5) \quad K_{ij} = \frac{1}{\ell} (\mathbf{U}_i, \mathbf{U}_j)_{\hat{X}}.$$

The matrix \mathbf{K} is positive semidefinite and has rank l . The solutions of (3.3) and (3.4) can be found in [10], [15], or [28], for example.

PROPOSITION 3.2. *Let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_l > 0$ denote the positive eigenvalues of \mathbf{K} and $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_l$ the associated orthonormal eigenvectors. Then a POD basis of rank $d \leq l$ is given by*

$$(3.6) \quad \boldsymbol{\psi}_i = \frac{1}{\sqrt{\lambda_i}} \sum_{j=1}^{\ell} (\mathbf{v}_i)_j \mathbf{U}_j,$$

where $(\mathbf{v}_i)_j$ denotes the j th component of the eigenvector \mathbf{v}_i . Furthermore, the following error formula holds:

$$(3.7) \quad \frac{1}{\ell} \sum_{i=1}^{\ell} \left\| \mathbf{U}_i - \sum_{j=1}^d (\mathbf{U}_i, \boldsymbol{\psi}_j)_{\hat{X}} \boldsymbol{\psi}_j \right\|_{\hat{X}}^2 = \sum_{j=d+1}^l \lambda_j.$$

Let $\mathcal{V}^d = \text{span}\{\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \dots, \boldsymbol{\psi}_d\}$ and $X^d \times M^d = \mathcal{V}^d$ with $X^d \subset X_h \subset X$ and $M^d \subset M_h \subset M$. Set the Ritz-projection $P^h: X \rightarrow X_h$ (if P^h is restricted to Ritz-projection from X_h to X^d , it is written as P^d) such that $P^h|_{X_h} = P^d: X_h \rightarrow X^d$ and $P^h: X \setminus X_h \rightarrow X_h \setminus X^d$ and L^2 -projection $\rho^d: M \rightarrow M^d$ denoted by, respectively,

$$(3.8) \quad a(P^h \mathbf{u}, \mathbf{v}_h) = a(\mathbf{u}, \mathbf{v}_h) \quad \forall \mathbf{v}_h \in X_h$$

and

$$(3.9) \quad (\rho^d p, q_d)_0 = (p, q_d)_0 \quad \forall q_d \in M^d,$$

where $\mathbf{u} \in X$ and $p \in M$. Due to (3.8) and (3.9) the linear operators P^h and ρ^d are well-defined and bounded:

$$(3.10) \quad \|\nabla(P^d \mathbf{u})\|_0 \leq \|\nabla \mathbf{u}\|_0, \quad \|\rho^d p\|_0 \leq \|p\|_0 \quad \forall \mathbf{u} \in X \text{ and } p \in M.$$

LEMMA 3.3. *For every d ($1 \leq d \leq l$) the projection operators P^d and ρ^d satisfy, respectively,*

$$(3.11) \quad \frac{1}{\ell} \sum_{i=1}^{\ell} \|\nabla(\mathbf{u}_h^{n_i} - P^d \mathbf{u}_h^{n_i})\|_0^2 \leq \sum_{j=d+1}^l \lambda_j,$$

$$(3.12) \quad \frac{1}{\ell} \sum_{i=1}^{\ell} \|\mathbf{u}_h^{n_i} - P^d \mathbf{u}_h^{n_i}\|_0^2 \leq Ch^2 \sum_{j=d+1}^l \lambda_j,$$

and

$$(3.13) \quad \frac{1}{\ell} \sum_{i=1}^{\ell} \|p_h^{n_i} - \rho^d p_h^{n_i}\|_0^2 \leq \sum_{j=d+1}^l \lambda_j,$$

where $\mathbf{u}_h^{n_i} = (u_{1h}^{n_i}, u_{2h}^{n_i})$ and $(u_{1h}^{n_i}, u_{2h}^{n_i}, p_h^{n_i})^T \in \mathcal{V}$.

Proof. For any $\mathbf{u} \in X$ we deduce from (3.8) that

$$\begin{aligned} \nu \|\nabla(\mathbf{u} - P^h \mathbf{u})\|_0^2 &= a(\mathbf{u} - P^h \mathbf{u}, \mathbf{u} - P^h \mathbf{u}) \\ &= a(\mathbf{u} - P^h \mathbf{u}, \mathbf{u} - \mathbf{v}_h) \\ &\leq \nu \|\nabla(\mathbf{u} - P^h \mathbf{u})\|_0 \|\nabla(\mathbf{u} - \mathbf{v}_h)\|_0 \quad \forall \mathbf{v}_h \in X_h. \end{aligned}$$

Therefore, we obtain that

$$(3.14) \quad \|\nabla(\mathbf{u} - P^h \mathbf{u})\|_0 \leq \|\nabla(\mathbf{u} - \mathbf{v}_h)\|_0 \quad \forall \mathbf{v}_h \in X_h.$$

If $\mathbf{u} = \mathbf{u}_h^{n_i}$, and P^h is restricted to Ritz-projection from X_h to X^d , i.e., $P^h \mathbf{u}_h^{n_i} = P^d \mathbf{u}_h^{n_i} \in X^d$, taking $\mathbf{v}_h = \sum_{j=1}^d (\mathbf{u}_h^{n_i}, \boldsymbol{\psi}_{u_j})_X \boldsymbol{\psi}_{u_j} \in X^d \subset X_h$ (where $\boldsymbol{\psi}_{u_j}$ is the component of $\boldsymbol{\psi}_j$ corresponding to \mathbf{u}) in (3.14), we can obtain (3.11) from (3.7).

In order to prove (3.12), we consider the following variational problem:

$$(3.15) \quad (\nabla \mathbf{w}, \nabla \mathbf{v}) = (\mathbf{u} - P^h \mathbf{u}, \mathbf{v}) \quad \forall \mathbf{v} \in X.$$

Thus, $\mathbf{w} \in [H_0^1(\Omega) \cap H^2(\Omega)]^2$ and satisfies $\|\mathbf{w}\|_2 \leq C \|\mathbf{u} - P^h \mathbf{u}\|_0$. Taking $\mathbf{v} = \mathbf{u} - P^h \mathbf{u}$ in (3.15), from (3.14) we obtain that

$$\begin{aligned} \|\mathbf{u} - P^h \mathbf{u}\|_0^2 &= (\nabla \mathbf{w}, \nabla(\mathbf{u} - P^h \mathbf{u})) \\ (3.16) \quad &= (\nabla(\mathbf{w} - \mathbf{w}_h), \nabla(\mathbf{u} - P^h \mathbf{u})) \\ &\leq \|\nabla(\mathbf{w} - \mathbf{w}_h)\|_0 \|\nabla(\mathbf{u} - P^d \mathbf{u})\|_0 \quad \forall \mathbf{w}_h \in X_h. \end{aligned}$$

Taking $\mathbf{w}_h = r_h \mathbf{w}$, from (2.14) and (3.16) we have

$$\begin{aligned} \|\mathbf{u} - P^h \mathbf{u}\|_0^2 &\leq Ch \|\mathbf{w}\|_2 \|\nabla(\mathbf{u} - P^h \mathbf{u})\|_0 \\ &\leq Ch \|\mathbf{u} - P^h \mathbf{u}\|_0 \|\nabla(\mathbf{u} - P^h \mathbf{u})\|_0. \end{aligned}$$

Thus, we obtain that

$$(3.17) \quad \|\mathbf{u} - P^h \mathbf{u}\|_0 \leq Ch \|\nabla(\mathbf{u} - P^h \mathbf{u})\|_0.$$

Therefore, if $\mathbf{u} = \mathbf{u}_h^{n_i}$ and P^h is restricted to Ritz-projection from X_h to X^d , i.e., $P^h \mathbf{u}_h^{n_i} = P^d \mathbf{u}_h^{n_i} \in X^d$, by (3.17) and (3.11) we obtain (3.12).

Using Hölder inequality and (3.9) can yield

$$\begin{aligned} \|p_h^{n_i} - \rho^d p_h^{n_i}\|_0^2 &= (p_h^{n_i} - \rho^d p_h^{n_i}, p_h^{n_i} - \rho^d p_h^{n_i}) \\ &= (p_h^{n_i} - \rho^d p_h^{n_i}, p_h^{n_i} - q_d) \\ &\leq \|p_h^{n_i} - \rho^d p_h^{n_i}\|_0 \|p_h^{n_i} - q_d\|_0 \quad \forall q_d \in M^d, \end{aligned}$$

and consequently,

$$(3.18) \quad \|p_h^{n_i} - \rho^d p_h^{n_i}\|_0 \leq \|p_h^{n_i} - q_d\|_0 \quad \forall q_d \in M^d.$$

Taking $q_d = \sum_{j=1}^d (p_h^{n_i}, \psi_{pj})_0 \psi_{pj}$ (where ψ_{pj} is the component of $\boldsymbol{\psi}_j$ corresponding to p) in (3.18), from (3.7) we can obtain (3.13), which completes the proof of Lemma 3.3. \square

Thus, using $\mathcal{V}^d = X^d \times M^d$, we can obtain the reduced formulation for Problem (III) as follows.

Problem (IV) Find $(\mathbf{u}_d^n, p_d^n) \in \mathcal{V}^d$ such that

$$(3.19) \quad \begin{cases} (\mathbf{u}_d^n, \mathbf{v}_d) + ka(\mathbf{u}_d^n, \mathbf{v}_d) + ka_1(\mathbf{u}_d^{n-1}, \mathbf{u}_d^n, \mathbf{v}_d) - kb(p_d^n, \mathbf{v}_d) \\ \quad = k(f^n, \mathbf{v}_d) + (\mathbf{u}_d^{n-1}, \mathbf{v}_d) \quad \forall \mathbf{v}_d \in X^d, \\ b(q_d, \mathbf{u}_d^n) = 0 \quad \forall q_d \in M^d, \\ \mathbf{u}_d^0 = \mathbf{0}, \end{cases}$$

where $1 \leq n \leq L$.

Remark 3.4. Problem (IV) is a reduced MFE formulation based on the POD technique for Problem (III), since it includes only $3d$ ($d \ll l \leq \ell \ll L$) degrees of freedom and is independent of the spatial grid scale h , while Problem (III) includes $3N_p + N_K \approx 5N_p$ for Mini's element of Example 2.2 (where N_p is the number of vertices in \mathfrak{S}_h and N_K the number of elements in \mathfrak{S}_h) and $3d \ll 5N_p$ (for example, in section 5, $d \leq 7$, while $N_p = 32 \times 32 = 1024$). The number of degrees of freedom of Example 2.1 is also approximately $5N_p$, but Example 2.3 and Example 2.4 are more. When one computes actual problems, one may obtain the ensemble of snapshots from physical system trajectories by drawing samples from experiments and interpolation (or data assimilation). For example, for weather forecast, one can use the previous weather prediction results to construct the ensemble of snapshots, and then restructure the POD basis for the ensemble of snapshots by above (3.3)–(3.6), and finally combine it with a Galerkin projection to derive a reduced order dynamical system; i.e., one needs only to solve the above Problem (IV), which has only $3d$ degrees of freedom,

but it is unnecessary to solve Problem (III). Thus, the forecast of future weather change can be quickly simulated, which is a result of major importance for real-life applications. Since the development and change of a large number of future nature phenomena are closely related to previous results (for example, weather change, biology anagenesis, and so on), using existing results as snapshots in order to structure POD basis, by solving corresponding PDEs, one may truly capture the laws of change of natural phenomena. Therefore, these POD methods provide useful and important applications.

4. Existence and error analysis of the solution of the reduced MFE formulation based on POD technique for the nonstationary Navier–Stokes equations. This section is devoted to discussing the existence and error estimates for Problem (IV).

We see from (3.6) that $\mathcal{V}^d = X^d \times M^d \subset \mathcal{V} \subset X_h \times M_h \subset X \times M$, where $X_h \times M_h$ is one of those spaces in Examples 2.1–2.4. Therefore, we have in the following result.

LEMMA 4.1. *There exists also an operator $r_d: X_h \rightarrow X^d$ such that, for all $\mathbf{u}_h \in X_h$,*

$$(4.1) \quad b(q_d, \mathbf{u}_h - r_d \mathbf{u}_h) = 0 \quad \forall q_d \in M^d, \quad \|\nabla r_d \mathbf{u}_h\|_0 \leq c \|\nabla \mathbf{u}_h\|_0,$$

and, for every d ($1 \leq d \leq l$),

$$(4.2) \quad \frac{1}{\ell} \sum_{i=1}^{\ell} \|\nabla(\mathbf{u}_h^{n_i} - r_d \mathbf{u}_h^{n_i})\|_0^2 \leq C \sum_{j=d+1}^l \lambda_j.$$

Proof. We use the Mini's and the second finite element as examples. Noting that for any $q_d \in M^d$ and $K \in \mathfrak{S}_h$, $\nabla q_d|_K \in P_0(K)$, using Green formula, we have

$$\begin{aligned} b(q_d, \mathbf{u}_h - r_d \mathbf{u}_h) &= - \int_{\Omega} \nabla q_d (\mathbf{u}_h - r_d \mathbf{u}_h) dx dy \\ &= - \sum_{K \in \mathfrak{S}_h} \nabla q_d|_K \int_K (\mathbf{u}_h - r_d \mathbf{u}_h) dx dy. \end{aligned}$$

Define r_d as follows:

$$(4.3) \quad r_d \mathbf{u}_h|_K = P^d \mathbf{u}_h|_K + \gamma_K \lambda_{K1} \lambda_{K2} \lambda_{K3} \quad \forall \mathbf{v}_h \in X_h \text{ and } K \in \mathfrak{S}_h,$$

where $\gamma_K = \int_K (\mathbf{u}_h - P^d \mathbf{u}_h) dx / \int_K \lambda_{K1} \lambda_{K2} \lambda_{K3} dx$. Thus, the first equality of (4.1) holds. Using (3.10)–(3.12) yields the inequality of (4.1). Then, if $\mathbf{u}_h = \mathbf{u}_h^{n_i}$, using (3.11)–(3.12), by simply computing we deduce (4.2). \square

Set

$$\begin{aligned} V &= \{\mathbf{v} \in X; b(q, \mathbf{v}) = 0 \quad \forall q \in M\}, \\ V_h &= \{\mathbf{v}_h \in X_h; b(q_h, \mathbf{v}_h) = 0 \quad \forall q_h \in M_h\}, \\ V^d &= \{\mathbf{v}_d \in X^d; b(q_d, \mathbf{v}_d) = 0 \quad \forall q_d \in M^d\}. \end{aligned}$$

Using dual principle and inequalities (3.11) and (3.12), we deduce the following result (see [1], [31]–[33]).

LEMMA 4.2. *There exists an operator $R_d: V \cup V_h \rightarrow V^d$ such that, for all $\mathbf{v} \in V \cup V_h$,*

$$(\mathbf{v} - R_d \mathbf{v}, \mathbf{v}_d) = 0 \quad \forall \mathbf{v}_d \in V^d, \quad \|\nabla R_d \mathbf{v}\|_0 \leq C \|\nabla \mathbf{v}\|_0,$$

and, for every d ($1 \leq d \leq l$),

$$(4.4) \quad \frac{1}{\ell} \sum_{i=1}^{\ell} \|\mathbf{u}_h^{n_i} - R_d \mathbf{u}_h^{n_i}\|_{-1}^2 \leq \frac{Ch^2}{\ell} \sum_{i=1}^{\ell} \|\nabla(\mathbf{u}_h^{n_i} - R_d \mathbf{u}_h^{n_i})\|_0^2 \leq Ch^2 \sum_{j=d+1}^l \lambda_j,$$

where $\|\cdot\|_{-1}$ denotes the normal of space $H^{-1}(\Omega)^2$ (see (2.7)).

We have the following result for the solution of Problem (IV).

THEOREM 4.3. *Under the hypotheses of Theorem 2.2, Problem (IV) has a unique solution $(\mathbf{u}_d^n, p_d^n) \in X^d \times M^d$ and satisfies*

$$(4.5) \quad \|\mathbf{u}_d^n\|_0^2 + k\nu \sum_{i=1}^n \|\nabla \mathbf{u}_d^i\|_0^2 \leq k\nu^{-1} \sum_{i=1}^n \|f^i\|_{-1}^2.$$

Proof. Using the same technique as the proof of Theorem 2.2, we could prove that Problem (IV) has a unique solution $(\mathbf{u}_d^n, p_d^n) \in X^d \times M^d$ and satisfies (4.5). \square

In the following theorem, error estimates of the solution for Problem (IV) are derived.

THEOREM 4.4. *Under the hypotheses of Theorem 2.2, if $h^2 = O(k)$, $k = O(\ell^{-2})$, snapshots are equably taken, and $\mathbf{f} \in H^{-1}(\Omega)^2$ satisfies $2\nu^{-2}N \sum_{i=1}^n \|\mathbf{f}^i\|_{-1} < 1$, then the error between the solution (\mathbf{u}_d^n, p_d^n) for Problem (IV) and the solution (\mathbf{u}_h^n, p_h^n) for Problem (III) has the following error estimates, for $n = 1, 2, \dots, L$,*

$$(4.6) \quad \begin{aligned} & \|\mathbf{u}_h^{n_i} - \mathbf{u}_d^{n_i}\|_0 + k^{1/2} \|p_h^{n_i} - p_d^{n_i}\|_0 + k^{1/2} \|\nabla(\mathbf{u}_h^{n_i} - \mathbf{u}_d^{n_i})\|_0 \\ & \leq C \left(k^{1/2} \sum_{j=d+1}^l \lambda_j \right)^{1/2}, \quad i = 1, 2, \dots, \ell; \\ & \|\mathbf{u}_h^n - \mathbf{u}_d^n\|_0 + k^{1/2} \|p_h^n - p_d^n\|_0 + k^{1/2} \|\nabla(\mathbf{u}_h^n - \mathbf{u}_d^n)\|_0 \\ & \leq Ck + C \left(k^{1/2} \sum_{j=d+1}^l \lambda_j \right)^{1/2}, \quad n \notin \{n_1, n_2, \dots, n_\ell\}. \end{aligned}$$

Proof. Subtracting Problem (IV) from Problem (III), taking $\mathbf{v}_h = \mathbf{v}_d \in X^d$ and $q_h = q_d \in M^d$, can yield

$$(4.7) \quad \begin{aligned} & (\mathbf{u}_h^n - \mathbf{u}_d^n, \mathbf{v}_d) + ka(\mathbf{u}_h^n - \mathbf{u}_d^n, \mathbf{v}_d) - kb(p_h^n - p_d^n, \mathbf{v}_d) + ka_1(\mathbf{u}_h^{n-1}, \mathbf{u}_h^n, \mathbf{v}_d) \\ & - ka_1(\mathbf{u}_d^{n-1}, \mathbf{u}_d^n, \mathbf{v}_d) = (\mathbf{u}_h^{n-1} - \mathbf{u}_d^{n-1}, \mathbf{v}_d) \quad \forall \mathbf{v}_d \in X^d, \end{aligned}$$

$$(4.8) \quad b(q_d, \mathbf{u}_h^n - \mathbf{u}_d^n) = 0 \quad \forall q_d \in M^d,$$

$$(4.9) \quad \mathbf{u}_h^0 - \mathbf{u}_d^0 = \mathbf{0}.$$

We obtain, from (2.3), (2.7), Theorem 2.2, and Theorem 4.3, by Hölder inequality, that

$$(4.10) \quad \begin{aligned} & |a_1(\mathbf{u}_h^{n-1}, \mathbf{u}_h^n, \mathbf{v}_d) - a_1(\mathbf{u}_d^{n-1}, \mathbf{u}_d^n, \mathbf{v}_d)| \\ & = |a_1(\mathbf{u}_h^{n-1} - \mathbf{u}_d^{n-1}, \mathbf{u}_h^n, \mathbf{v}_d) + a_1(\mathbf{u}_d^{n-1}, \mathbf{u}_h^n - \mathbf{u}_d^n, \mathbf{v}_d)| \\ & \leq C[\|\nabla(\mathbf{u}_h^{n-1} - \mathbf{u}_d^{n-1})\|_0 + \|\nabla(\mathbf{u}_h^n - \mathbf{u}_d^n)\|_0] \|\nabla \mathbf{v}_d\|_0, \end{aligned}$$

especially, if $\mathbf{v}_d = P^d \mathbf{u}_h^n - \mathbf{u}_d^n$, then

$$\begin{aligned}
& |a_1(\mathbf{u}_h^{n-1}, \mathbf{u}_h^n, P^d \mathbf{u}_h^n - \mathbf{u}_d^n) - a_1(\mathbf{u}_d^{n-1}, \mathbf{u}_d^n, P^d \mathbf{u}_h^n - \mathbf{u}_d^n)| \\
&= |a_1(\mathbf{u}_h^{n-1} - \mathbf{u}_d^{n-1}, \mathbf{u}_h^n, P^d \mathbf{u}_h^n - \mathbf{u}_d^n) + a_1(\mathbf{u}_d^{n-1}, \mathbf{u}_h^n - \mathbf{u}_d^n, P^d \mathbf{u}_h^n - \mathbf{u}_d^n)| \\
(4.11) \quad &= |a_1(\mathbf{u}_h^{n-1} - \mathbf{u}_d^{n-1}, \mathbf{u}_h^n, P^d \mathbf{u}_h^n - \mathbf{u}_h^n) + a_1(\mathbf{u}_h^{n-1} - \mathbf{u}_d^{n-1}, \mathbf{u}_h^n, \mathbf{u}_h^n - \mathbf{u}_d^n)| \\
&\quad + |a_1(\mathbf{u}_d^{n-1}, \mathbf{u}_h^n - \mathbf{u}_d^n, P^d \mathbf{u}_h^n - \mathbf{u}_h^n)| \\
&\leq C \|\nabla(\mathbf{u}_h^n - P^d \mathbf{u}_h^n)\|_0^2 + \varepsilon [\|\nabla(\mathbf{u}_h^{n-1} - \mathbf{u}_d^{n-1})\|_0^2 + \|\nabla(\mathbf{u}_h^n - \mathbf{u}_d^n)\|_0^2] \\
&\quad + N \|\nabla \mathbf{u}_h^n\|_0 \|\nabla(\mathbf{u}_h^{n-1} - \mathbf{u}_d^{n-1})\|_0 \|\nabla(\mathbf{u}_h^n - \mathbf{u}_d^n)\|_0,
\end{aligned}$$

where ε is a small positive constant which can be chosen arbitrarily.

Write $\bar{\partial}_t \mathbf{u}_h^n = [\mathbf{u}_h^n - \mathbf{u}_h^{n-1}]/k$ and note that $\bar{\partial}_t \mathbf{u}_d^n \in V^d$ and $\bar{\partial}_t R_d \mathbf{u}_h^n \in V^d$. From Lemma 4.2, (4.7), and (4.10), we have that

$$\begin{aligned}
& \|\bar{\partial}_t \mathbf{u}_h^n - \bar{\partial}_t \mathbf{u}_d^n\|_{-1} \leq \|\bar{\partial}_t \mathbf{u}_h^n - \bar{\partial}_t R_d \mathbf{u}_h^n\|_{-1} + \|\bar{\partial}_t R_d \mathbf{u}_h^n - \bar{\partial}_t \mathbf{u}_d^n\|_{-1} \\
&\leq \|\bar{\partial}_t \mathbf{u}_h^n - \bar{\partial}_t R_d \mathbf{u}_h^n\|_{-1} + \sup_{\mathbf{v} \in V} \frac{(\bar{\partial}_t R_d \mathbf{u}_h^n - \bar{\partial}_t \mathbf{u}_d^n, \mathbf{v})}{\|\nabla \mathbf{v}\|_0} \\
&= \|\bar{\partial}_t \mathbf{u}_h^n - \bar{\partial}_t R_d \mathbf{u}_h^n\|_{-1} + \sup_{\mathbf{v} \in V} \frac{(\bar{\partial}_t \mathbf{u}_h^n - \bar{\partial}_t \mathbf{u}_d^n, R_d \mathbf{v})}{\|\nabla \mathbf{v}\|_0} \\
(4.12) \quad &= \|\bar{\partial}_t \mathbf{u}_h^n - \bar{\partial}_t R_d \mathbf{u}_h^n\|_{-1} + \sup_{\mathbf{v} \in V} \frac{1}{\|\nabla \mathbf{v}\|_0} [b(p_h^n - p_d^n, R_d \mathbf{v}) \\
&\quad - a(\mathbf{u}_h^n - \mathbf{u}_d^n, R_d \mathbf{v}) - a_1(\mathbf{u}_h^{n-1}, \mathbf{u}_h^n, R_d \mathbf{v}) + a_1(\mathbf{u}_d^{n-1}, \mathbf{u}_d^n, R_d \mathbf{v})] \\
&= \|\bar{\partial}_t \mathbf{u}_h^n - \bar{\partial}_t R_d \mathbf{u}_h^n\|_{-1} + \sup_{\mathbf{v} \in V} \frac{1}{\|\nabla \mathbf{v}\|_0} [b(p_h^n - \rho^d p_h^n, R_d \mathbf{v}) \\
&\quad - a(\mathbf{u}_h^n - \mathbf{u}_d^n, R_d \mathbf{v}) - a_1(\mathbf{u}_h^{n-1}, \mathbf{u}_h^n, R_d \mathbf{v}) + a_1(\mathbf{u}_d^{n-1}, \mathbf{u}_d^n, R_d \mathbf{v})] \\
&\leq \|\bar{\partial}_t \mathbf{u}_h^n - \bar{\partial}_t R_d \mathbf{u}_h^n\|_{-1} + C [\|p_h^n - \rho^d p_h^n\|_0 \\
&\quad + \|\nabla(\mathbf{u}_h^{n-1} - \mathbf{u}_d^{n-1})\|_0 + \|\nabla(\mathbf{u}_h^n - \mathbf{u}_d^n)\|_0].
\end{aligned}$$

By using (2.9), (4.7), (4.10), (4.12), and Lemma 4.1, we have that

$$\begin{aligned}
& \beta \|\rho^d p_h^n - p_d^n\|_0 \leq \sup_{\mathbf{v}_h \in X_h} \frac{b(\rho^d p_h^n - p_d^n, \mathbf{v}_h)}{\|\nabla \mathbf{v}_h\|_0} = \sup_{\mathbf{v}_h \in X_h} \frac{b(p_h^n - p_d^n, r_d \mathbf{v}_h)}{\|\nabla \mathbf{v}_h\|_0} \\
&= \sup_{\mathbf{v}_h \in X_h} \frac{1}{\|\nabla \mathbf{v}_h\|_0} [(\bar{\partial}_t \mathbf{u}_h^n - \bar{\partial}_t \mathbf{u}_d^n, r_d \mathbf{v}_h) + a(\mathbf{u}_h^n - \mathbf{u}_d^n, r_d \mathbf{v}_h) \\
(4.13) \quad &\quad + a_1(\mathbf{u}_h^{n-1}, \mathbf{u}_h^n, r_d \mathbf{v}_h) - a_1(\mathbf{u}_d^{n-1}, \mathbf{u}_d^n, r_d \mathbf{v}_h)] \\
&\leq C [\|\bar{\partial}_t \mathbf{u}_h^n - \bar{\partial}_t \mathbf{u}_d^n\|_{-1} + \|\nabla(\mathbf{u}_h^{n-1} - \mathbf{u}_d^{n-1})\|_0 + \|\nabla(\mathbf{u}_h^n - \mathbf{u}_d^n)\|_0] \\
&\leq C [\|\bar{\partial}_t \mathbf{u}_h^n - \bar{\partial}_t R_d \mathbf{u}_h^n\|_{-1} + \|p_h^n - \rho^d p_h^n\|_0 \\
&\quad + \|\nabla(\mathbf{u}_h^{n-1} - \mathbf{u}_d^{n-1})\|_0 + \|\nabla(\mathbf{u}_h^n - \mathbf{u}_d^n)\|_0].
\end{aligned}$$

Thus, we obtain that

$$\begin{aligned}
(4.14) \quad & \|p_h^n - p_d^n\|_0 \leq \|p_h^n - \rho^d p_h^n\|_0 + \|\rho^d p_h^n - p_d^n\|_0 \leq C [\|\nabla(\mathbf{u}_h^{n-1} - \mathbf{u}_d^{n-1})\|_0 \\
&\quad + \|\nabla(\mathbf{u}_h^n - \mathbf{u}_d^n)\|_0 + \|\bar{\partial}_t \mathbf{u}_h^n - \bar{\partial}_t R_d \mathbf{u}_h^n\|_{-1} + \|p_h^n - \rho^d p_h^n\|_0].
\end{aligned}$$

Taking $\mathbf{v}_d = P^d \mathbf{u}_h^n - \mathbf{u}_d^n$ in (4.7), it follows from (4.8) that

$$\begin{aligned}
(4.15) \quad & (\mathbf{u}_h^n - \mathbf{u}_d^n, \mathbf{u}_h^n - \mathbf{u}_d^n) - (\mathbf{u}_h^{n-1} - \mathbf{u}_d^{n-1}, \mathbf{u}_h^n - \mathbf{u}_d^n) + ka(\mathbf{u}_h^n - \mathbf{u}_d^n, \mathbf{u}_h^n - \mathbf{u}_d^n) \\
& = (\mathbf{u}_h^n - \mathbf{u}_d^n - (\mathbf{u}_h^{n-1} - \mathbf{u}_d^{n-1}), \mathbf{u}_h^n - P^d \mathbf{u}_h^n) + ka(\mathbf{u}_h^n - P^d \mathbf{u}_h^n, \mathbf{u}_h^n - P^d \mathbf{u}_h^n) \\
& \quad + kb(p_h^n - \rho^d p_h^n, \mathbf{u}_h^n - \mathbf{u}_d^n) + kb(p_h^n - p_d^n, \mathbf{u}_h^n - P^d \mathbf{u}_h^n) \\
& \quad - ka_1(\mathbf{u}_h^{n-1}, \mathbf{u}_h^n, P^d \mathbf{u}_h^n - \mathbf{u}_d^n) + ka_1(\mathbf{u}_d^{n-1}, \mathbf{u}_d^n, P^d \mathbf{u}_h^n - \mathbf{u}_d^n).
\end{aligned}$$

Thus, noting that $a(a-b) = [a^2 - b^2 + (a-b)^2]/2$ (for $a \geq 0$ and $b \geq 0$), by (4.11), (4.14), Hölder inequality, Cauchy inequality, and Proposition 3.2, we obtain that

$$\begin{aligned}
(4.16) \quad & \frac{1}{2} [\|\mathbf{u}_h^n - \mathbf{u}_d^n\|_0^2 - \|\mathbf{u}_h^{n-1} - \mathbf{u}_d^{n-1}\|_0^2 + \|\mathbf{u}_h^n - \mathbf{u}_d^n - (\mathbf{u}_h^{n-1} - \mathbf{u}_d^{n-1})\|_0^2] \\
& + \nu k \|\nabla(\mathbf{u}_h^n - \mathbf{u}_d^n)\|_0^2 \leq \frac{1}{2} \|\mathbf{u}_h^n - \mathbf{u}_d^n - (\mathbf{u}_h^{n-1} - \mathbf{u}_d^{n-1})\|_0^2 + \frac{1}{2} \|\mathbf{u}_h^n - P^d \mathbf{u}_h^n\|_0^2 \\
& + Ck \|\nabla(\mathbf{u}_h^n - P^d \mathbf{u}_h^n)\|_0^2 + Ck \|p_h^n - \rho^d p_h^n\|_0^2 + C \|\bar{\partial}_t \mathbf{u}_h^n - \bar{\partial}_t R_d \mathbf{u}_h^n\|_{-1}^2 \\
& + (\varepsilon_1 + C\varepsilon_2 + \varepsilon)k [\|\nabla(\mathbf{u}_h^n - \mathbf{u}_d^n)\|_0^2 + \|\nabla(\mathbf{u}_h^{n-1} - \mathbf{u}_d^{n-1})\|_0^2] \\
& + \frac{1}{2} k [N^2 \gamma^{-1} \|\nabla \mathbf{u}_h^n\|_0^2 \|\nabla(\mathbf{u}_h^{n-1} - \mathbf{u}_d^{n-1})\|_0^2 + \gamma \|\nabla(\mathbf{u}_h^n - \mathbf{u}_d^n)\|_0^2],
\end{aligned}$$

where ε_1 and ε_2 are two small positive constants which can be chosen arbitrarily. Taking $\varepsilon + \varepsilon_1 + C\varepsilon_2 = \nu/4$, it follows from (4.16) that

$$\begin{aligned}
(4.17) \quad & [\|\mathbf{u}_h^n - \mathbf{u}_d^n\|_0^2 - \|\mathbf{u}_h^{n-1} - \mathbf{u}_d^{n-1}\|_0^2] + \nu k \|\nabla(\mathbf{u}_h^n - \mathbf{u}_d^n)\|_0^2 \\
& \leq \|\mathbf{u}_h^n - P^d \mathbf{u}_h^n\|_0^2 + Ck \|\nabla(\mathbf{u}_h^n - P^d \mathbf{u}_h^n)\|_0^2 + Ck \|p_h^n - \rho^d p_h^n\|_0^2 \\
& \quad + C \|\bar{\partial}_t \mathbf{u}_h^n - \bar{\partial}_t R_d \mathbf{u}_h^n\|_{-1}^2 + \frac{1}{2} k \gamma \|\nabla(\mathbf{u}_h^{n-1} - \mathbf{u}_d^{n-1})\|_0^2 \\
& \quad + k N^2 \gamma^{-1} \|\nabla \mathbf{u}_h^n\|_0^2 \|\nabla(\mathbf{u}_h^{n-1} - \mathbf{u}_d^{n-1})\|_0^2, \quad 1 \leq n \leq L.
\end{aligned}$$

If $h^2 = O(k)$, $2\nu^{-2} N \sum_{j=1}^n \|\mathbf{f}^j\|_{-1} < 1$, $n = n_i$ ($i = 1, 2, \dots, \ell$), summing (4.17) from $n = n_1, n_2, \dots, n_i$ ($i = 1, 2, \dots, \ell$), let $n_0 = 0$, and noting that $\mathbf{u}_h^0 - \mathbf{u}_d^0 = \mathbf{0}$ and $\ell \leq L$, from Lemmas 3.3, 4.1, and 4.2, we obtain that

$$\begin{aligned}
(4.18) \quad & \|\mathbf{u}_h^{n_i} - \mathbf{u}_d^{n_i}\|_0^2 + \nu k \|\nabla(\mathbf{u}_h^{n_i} - \mathbf{u}_d^{n_i})\|_0^2 \leq C \sum_{j=1}^{n_i} \|\mathbf{u}_h^{n_j} - P^d \mathbf{u}_h^{n_j}\|_0^2 \\
& + Ck \sum_{j=1}^{n_i} [\|\nabla(\mathbf{u}_h^{n_j} - P^d \mathbf{u}_h^{n_j})\|_0^2 + \|p_h^{n_j} - \rho^d p_h^{n_j}\|_0^2] \\
& + C \sum_{j=1}^{n_i} [\|\mathbf{u}_h^{n_j} - R_d \mathbf{u}_h^{n_j}\|_{-1}^2 + \|\mathbf{u}_h^{n_{j-1}} - R_d \mathbf{u}_h^{n_{j-1}}\|_{-1}^2] \\
& \leq Ck \sum_{j=d+1}^l \lambda_j, \quad i = 1, 2, \dots, \ell.
\end{aligned}$$

Thus, we obtain that

$$(4.19) \quad \begin{aligned} & \|\mathbf{u}_h^{n_i} - \mathbf{u}_d^{n_i}\|_0 + (\nu k)^{1/2} \|\nabla(\mathbf{u}_h^{n_i} - \mathbf{u}_d^{n_i})\|_0 \\ & \leq C \left(k^{1/2} \sum_{j=d+1}^l \lambda_j \right)^{1/2}, \quad i = 1, 2, \dots, \ell. \end{aligned}$$

Combining (4.19) and (4.14), by Lemmas 3.3, 4.1, and 4.2, we obtain the first inequality of (4.6).

If $n \neq n_i$ ($i = 1, 2, \dots, \ell$), we may as well let $t^{(n)} \in (t^{(n_{i-1})}, t^{(n_i)})$ and $t^{(n)}$ be the nearest point to $t^{(n_i)}$. Expanding \mathbf{u}^n and p^n into Taylor series with respect to $t^{(n_i)}$ yields that

$$(4.20) \quad \begin{aligned} \mathbf{u}^n &= \mathbf{u}^{n_i} - \eta_i k \frac{\partial \mathbf{u}(\xi_1)}{\partial t}, \quad t^{(n)} \leq \xi_1 \leq t^{(n_i)}, \\ p^n &= p^{n_i} - \eta_i k \frac{\partial p(\xi_2)}{\partial t}, \quad t^{(n)} \leq \xi_2 \leq t^{(n_i)}, \end{aligned}$$

where η_i is the step number from $t^{(n)}$ to $t^{(n_i)}$. If $h^2 = O(k)$, $2\nu^{-2}N \sum_{j=1}^n \|\mathbf{f}^j\|_{-1} < 1$, $k = O(\ell^{-2})$, summing (4.17) for n_1, \dots, n_{i-1}, n , let $n_0 = 0$, and noting that $\mathbf{u}_h^0 - \mathbf{u}_d^0 = \mathbf{0}$, from Lemmas 4.1 and 4.2 and Lemma 3.3, we obtain that

$$(4.21) \quad \|\mathbf{u}_h^n - \mathbf{u}_d^n\|_0^2 + k\gamma \|\nabla(\mathbf{u}_h^n - \mathbf{u}_d^n)\|_0^2 \leq C\eta^2 k^3 + Ck^{1/2} \sum_{j=d+1}^l \lambda_j.$$

Since snapshots are equably taken, $\eta_i \leq L/(2\ell)$. If $k = O(\ell^{-2})$, we obtain that

$$(4.22) \quad \|\mathbf{u}_h^n - \mathbf{u}_d^n\|_0 + k^{1/2} \|\nabla(\mathbf{u}_h^n - \mathbf{u}_d^n)\|_0 \leq Ck + C \left(k^{1/2} \sum_{j=d+1}^l \lambda_j \right)^{1/2}.$$

Combining (4.22) and (4.14), by Lemmas 3.3, 4.1, and 4.2, we obtain the second inequality of (4.6). \square

Combining Theorem 2.2 and Theorem 4.4 yields the following result.

THEOREM 4.5. *Under Theorem 2.2 and Theorem 4.4 hypotheses, the error estimate between the solutions for Problem (II) and the solutions for the reduced order basic Problem (IV) is, for $n = 1, 2, \dots, L$, $m = 1, 2, 3$,*

$$(4.23) \quad \begin{aligned} & \|\mathbf{u}^n - \mathbf{u}_d^n\|_0 + k^{1/2} \|p^n - p_d^n\|_0 + k^{1/2} \|\nabla(\mathbf{u}^n - \mathbf{u}_d^n)\|_0 \\ & \leq Ck + Ch^m + C \left(k^{1/2} \sum_{j=d+1}^l \lambda_j \right)^{1/2}. \end{aligned}$$

Remark 4.6. Though the constants C in Theorems 4.4 and 4.5 are directly independent on k , they are indirectly dependent on L . Therefore, if $k \rightarrow 0$, that implies $L \rightarrow \infty$. The condition $k = O(\ell^{-2})$, which implies $L = O(\ell^2)$, in Theorem 4.4 shows the relation between the number ℓ of snapshots and the number L at all time instances. Therefore, it is unnecessary to take total transient solutions at all time instances $t^{(n)}$ as snapshots (see, for instance, in [27]–[29]). Theorems 4.4 and 4.5 have

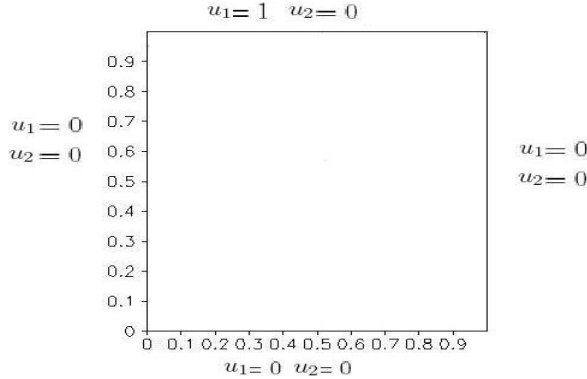


FIG. 1. Physical model of the cavity flows: $t = 0$; i.e., $n = 0$ initial values on boundary.

presented the error estimates between the solution of the reduced MFE formulation Problem (IV) and the solution of usual MFE formulation Problem (III) and Problem (II), respectively. Since our methods employ some MFE solutions (\mathbf{u}_h^n, p_h^n) ($n = 1, 2, \dots, L$) for Problem (III) as assistant analysis, the error estimates in Theorem 4.5 are correlated to the spatial grid scale h and time step size k . However, when one computes actual problems, one may obtain the ensemble of snapshots from physical system trajectories by drawing samples from experiments and interpolation (or data assimilation). Therefore, the assistant (\mathbf{u}_h^n, p_h^n) ($n = 1, 2, \dots, L$) could be replaced with the interpolation functions of experimental and previous results, thus rendering it unnecessary to solve Problem (III), and requiring only to directly solve Problem (IV) such that Theorem 4.4 is satisfied.

5. Some numerical experiments. In this section, we present some numerical examples of the physical model of cavity flows for Mini's element and different Reynolds numbers by the reduced formulation Problem (IV), thus validating the feasibility and efficiency of the POD method.

Let the side length of the cavity be 1 (see Figure 1). We first divide the cavity into $32 \times 32 = 1024$ small squares with side length $\Delta x = \Delta y = \frac{1}{32}$, and then link the diagonal of the square to divide each square into two triangles in the same direction, which consists of triangularization \mathfrak{S}_h . Take time step increment as $k = 0.001$. Except that u_1 is equal to 1 on upper boundary, all other initial value, boundary values, and (f_1, f_2) are all taken as 0 (see Figure 1).

We obtain 20 values (i.e., snapshots) at time $t = 10, 20, 30, \dots, 200$ by solving the usual MFE formulation, i.e., Problem (III). It is shown by computing that eigenvalues satisfy $[k^{1/2} \sum_{i=7}^{20} \lambda_i]^{1/2} \leq 10^{-3}$. When $t = 200$, we obtain the solutions of the reduced formulation Problem (IV) based on the POD method of MFE depicted graphically in Figures 2 to 5 on the right-hand side employed six POD bases for $Re = 750$ and required six POD bases for $Re = 1500$, while the solutions obtained with usual MFE formulation Problem (III) are depicted graphically in Figures 2 to 5 on the left-hand side. (Since these figures are equal to solutions obtained with 20 bases, they are also referred to as the figures of the solution with full bases.)

Figure 6 shows the errors between solutions obtained with a different number of POD bases and solutions obtained with full bases. Comparing the usual MFE formulation Problem (III) with the reduced MFE formulation Problem (IV) containing six POD bases implementing 3000 times the numerical simulation computations, we

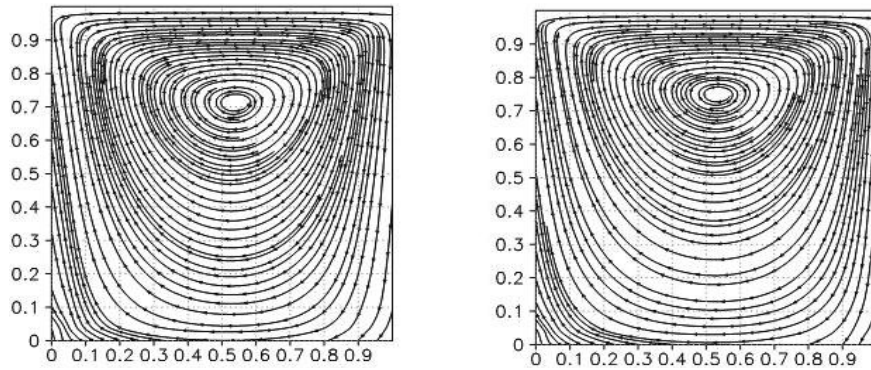


FIG. 2. When $Re = 750$, velocity stream line figure for usual MFE solutions (on left-hand side figure) and $d = 6$, the solution of the reduced MFE formulation (on right-hand side figure).

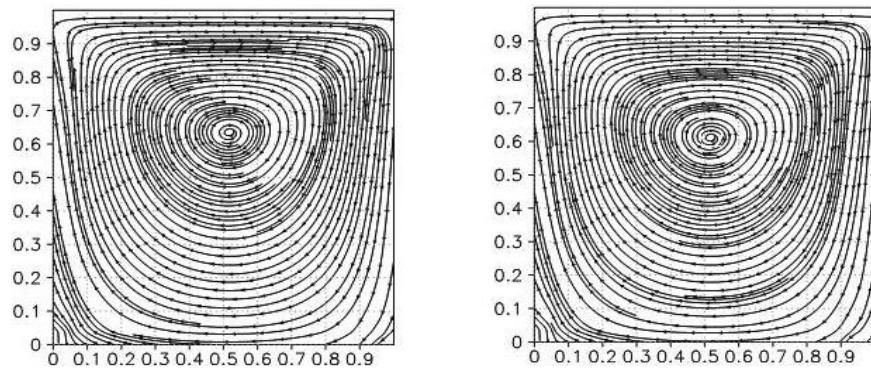


FIG. 3. When $Re = 1500$, velocity stream line figure for usual MFE solutions (on left-hand side figure) and $d = 6$, the solution of the reduced MFE formulation (on right-hand side figure).

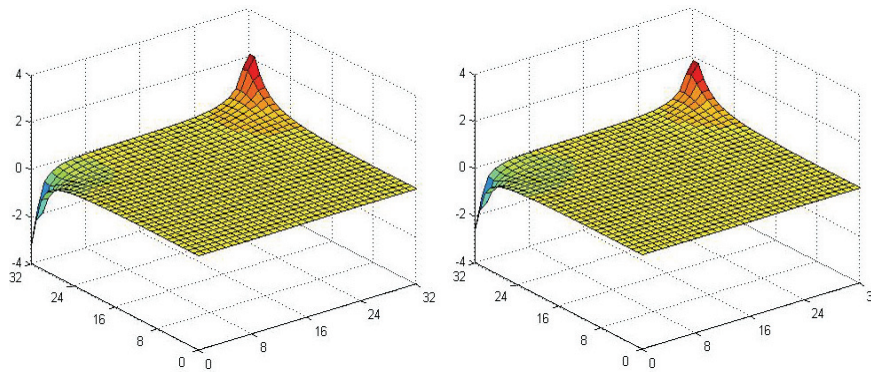


FIG. 4. When $Re = 750$, pressure figure for usual MFE solution (on left-hand side figure) and $d = 6$ solution of reduced MFE formulation (on right-hand side figure).

find that for usual MFE formulation Problem (III) the required CPU time is 6 minutes, while for the reduced MFE formulation Problem (IV) with 6 POD bases the corresponding time is only three seconds; i.e., the usual MFE formulation Problem (III) required a CPU time which is by a factor of 120 larger than that required by

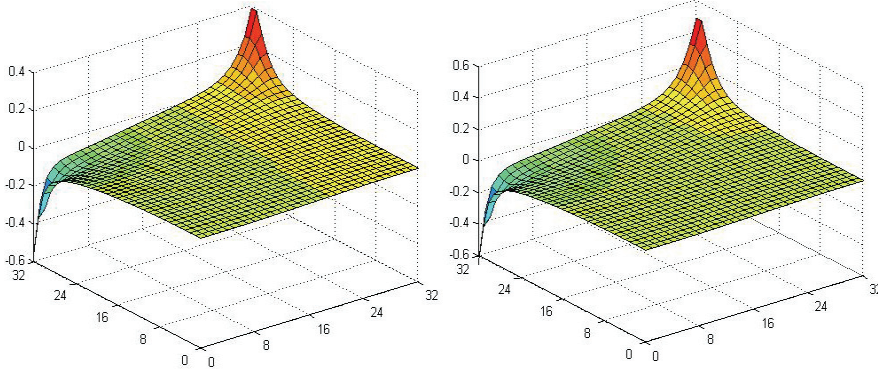


FIG. 5. When $Re = 1500$, the pressure figure for usual MFE solution (on left-hand side figure) and $d = 6$ solution of reduced MFE formulation (on right-hand side figure).

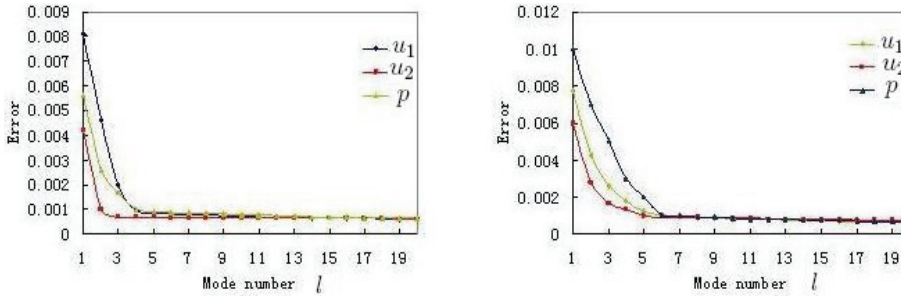


FIG. 6. Error for $Re = 750$ on left-hand side; error for $Re = 1500$ on right-hand side.

the reduced MFE formulation Problem (IV) with 6 POD bases, while the error between their respective solutions does not exceed 10^{-3} . It is also shown that finding the approximate solutions for the nonstationary Navier–Stokes equations with the reduced MFE formulation Problem (IV) is computationally very effective. The results for numerical examples are consistent with those obtained for the theoretical case.

6. Conclusions. In this paper, we have employed the POD technique to derive a reduced formulation for the nonstationary Navier–Stokes equations. We first reconstruct optimal orthogonal bases of ensembles of data which are compiled from transient solutions derived by using the usual MFE equation system, while in actual applications, one may obtain the ensemble of snapshots from physical system trajectories by drawing samples from experiments and interpolation (or data assimilation). For example, for weather forecast, one may use previous weather prediction results to construct the ensemble of snapshots to restructure the POD basis for the ensemble of snapshots by methods of the above section 3. We have also combined the optimal orthogonal bases with a Galerkin projection procedure, thus yielding a new reduced MFE formulation of lower dimensional order and of high accuracy for the nonstationary Navier–Stokes equations. We have then proceeded to derive error estimates between our reduced MFE approximate solutions and the usual MFE approximate solutions, and have shown, using numerical examples, that the error between the reduced MFE approximate solution and the usual MFE solution is consistent with the theoretical error results, thus validating both feasibility and efficiency

of our reduced MFE formulation. Future research work in this area will aim to extend the reduced MFE formulation, applying it to a realistic operational atmospheric numerical weather forecast system and to more complicated PDEs. We have shown both by theoretical analysis as well as by numerical examples that the reduced MFE formulation presented herein has extensive potential applications.

Though Kunisch and Volkwein have presented some Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics, i.e., for the nonstationary Navier–Stokes equations in [28], our method is different from their approaches, whose methods consist of Galerkin projection approaches where the original variables are substituted for linear combination of POD basis and the error estimates of the velocity field therein are only derived, their POD basis being generated with the solutions of the physical system at all time instances, while our POD basis is generated with only few solutions of the physical system which are useful and of interest for us. Especially, only the velocity field is approximated in [28], while both the velocity field and the pressure are all synchronously approximated in our present method, and error estimates of velocity field and pressure approximate solutions are also synchronously derived. Thus, our method appears to be more optimal than that in [28].

Acknowledgments. The authors thank all referees and Professor Karl Kunisch for their valued suggestions to this paper very much.

REFERENCES

- [1] V. GIRAULT AND P. A. RAVIART, *Finite Element Methods for Navier–Stokes Equations, Theorem and Algorithms*, Springer-Verlag, Berlin, 1986.
- [2] J. G. HEYWOOD AND R. RANNACHER, *Finite element approximation of the nonstationary Navier–Stokes problem, I. Regularity of solutions and second order estimates for spatial discretization*, SIAM J. Numer. Anal., 19 (1982), pp. 275–311.
- [3] Z. D. LUO, *The third order estimate of mixed finite element for the Navier–Stokes problems*, Chinese Quart. J. Math., 10 (1995), pp. 9–12.
- [4] K. FUKUNAGA, *Introduction to Statistical Recognition*, Academic Press, New York, 1990.
- [5] I. T. JOLLIFFE, *Principal Component Analysis*, Springer-Verlag, Berlin, 2002.
- [6] D. T. CROMMELIN AND A. J. MAJDA, *Strategies for model reduction: Comparing different optimal bases*, J. Atmospheric Sci., 61 (2004), pp. 2206–2217.
- [7] A. J. MAJDA, I. TIMOFEYEV, AND E. VANDEN-EIJNDEN, *Systematic strategies for stochastic mode reduction in climate*, J. Atmospheric Sci., 60 (2003), pp. 1705–1722.
- [8] F. SELTEN, *Baroclinic empirical orthogonal functions as basis functions in an atmospheric model*, J. Atmospheric Sci., 54 (1997), pp. 2100–2114.
- [9] P. HOLMES, J. L. LUMLEY, AND G. BERKOOZ, *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*, Cambridge University Press, Cambridge, UK, 1996.
- [10] G. BERKOOZ, P. HOLMES, AND J. L. LUMLEY, *The proper orthogonal decomposition in analysis of turbulent flows*, Ann. Rev. Fluid Mech., 25 (1993), pp. 539–575.
- [11] W. CAZEMIER, R. W. C. P. VERSTAPPEN, AND A. E. P. VELDMAN, *Proper orthogonal decomposition and low-dimensional models for driven cavity flows*, Phys. Fluids, 10 (1998), pp. 1685–1699.
- [12] D. AHLMAN, F. SÖDELUND, J. JACKSON, A. KURDILA, AND W. SHYY, *Proper orthogonal decomposition for time-dependent lid-driven cavity flows*, Numerical Heat Transfer Part B–Fundamentals, 42 (2002), pp. 285–306.
- [13] J. L. LUMLEY, *Coherent Structures in Turbulence*, in Transition and Turbulence, R. E. Meyer, ed., Academic Press, New York, 1981, pp. 215–242.
- [14] Y. N. AUBRY, P. HOLMES, J. L. LUMLEY, AND E. STONE, *The dynamics of coherent structures in the wall region of a turbulent boundary layer*, J. Fluid Dyn., 192 (1988), pp. 115–173.
- [15] L. SIROVICH, *Turbulence and the dynamics of coherent structures: Parts I–III*, Quart. Appl. Math., 45 (1987), pp. 561–590.
- [16] R. D. JOSLIN, M. D. GUNZBURGER, R. A. NICOLAIDES, G. ERLEBACHER, AND M. Y. HUSSAINI, *A self-contained automated methodology for optimal flow control validated for transition delay*, AIAA J., 35 (1997), pp. 816–824.

- [17] H. V. LY AND H. T. TRAN, *Proper orthogonal decomposition for flow calculations and optimal control in a horizontal CVD reactor*, Quart. Appl. Math., 60 (2002), pp. 631–656.
- [18] O. K. REDIONITIS, J. KO, X. YUE, AND A. J. KURDILA, *Synthetic Jets, Their Reduced Order Modeling and Applications to Flow Control*, AIAA Paper number 99-1000, 37 Aerospace Sciences Meeting & Exhibit, Reno, NV, 1999.
- [19] P. MOIN AND R. D. MOSER, *Characteristic-eddy decomposition of turbulence in channel*, J. Fluid Mech., 200 (1989), pp. 417–509.
- [20] M. RAJAEI, S. K. F. KARLSSON, AND L. SIROVICH, *Low dimensional description of free shear flow coherent structures and their dynamical behavior*, J. Fluid Mech., 258 (1994), pp. 1401–1402.
- [21] Y. H. CAO, J. ZHU, Z. D. LUO, AND I. M. NAVON, *Reduced order modeling of the upper tropical pacific ocean model using proper orthogonal decomposition*, Comput. Math. Appl., 52 (2006), pp. 1373–1386.
- [22] Y. H. CAO, J. ZHU, I. M. NAVON, AND Z. D. LUO, *A reduced order approach to four-dimensional variational data assimilation using proper orthogonal decomposition*, Internat. J. Numer. Methods Fluids, 53 (2007), pp. 1571–1583.
- [23] Z. D. LUO, J. ZHU, R. W. WANG, AND I. M. NAVON, *Proper orthogonal decomposition approach and error estimation of mixed finite element methods for the tropical Pacific Ocean reduced gravity model*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 4184–4195.
- [24] Z. D. LUO, J. CHEN, J. ZHU, R. W. WANG, AND I. M. NAVON, *An optimizing reduced order FDS for the tropical Pacific Ocean reduced gravity model*, Internat. J. Numer. Methods Fluids, 55 (2007), pp. 143–161.
- [25] R. W. WANG, J. ZHU, Z. D. LUO, AND I. M. NAVON, *An equation-free reduced order modeling approach to tropic pacific simulation*, accepted for publication in the Advances in Geosciences book series of World Scientific Publishing, 2007.
- [26] Z. D. LUO, R. W. WANG, J. CHEN, AND J. ZHU, *Finite difference scheme based on proper orthogonal decomposition for the nonstationary Navier–Stokes equations*, Sci. China Ser. A: Math., 50 (2007), pp. 1186–1196.
- [27] K. KUNISCH AND S. VOLKWEIN, *Galerkin proper orthogonal decomposition methods for parabolic problems*, Numer. Math., 90 (2001), pp. 177–148.
- [28] K. KUNISCH AND S. VOLKWEIN, *Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics*, SIAM J. Numer. Anal., 40 (2002), pp. 492–515.
- [29] K. KUNISCH AND S. VOLKWEIN, *Control of Burgers’ equation by a reduced order approach using proper orthogonal decomposition*, J. Optim. Theory Appl., 102 (1999), pp. 345–371.
- [30] R. A. ADAMS, *Sobolev Space*, Academic Press, New York, 1975.
- [31] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
- [32] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
- [33] Z. D. LUO, *Mixed Finite Element Methods and Applications*, Chinese Science Press, Beijing, 2006.

DATA APPROXIMATION USING SHAPE-PRESERVING PARAMETRIC SURFACES*

PAOLO COSTANTINI[†] AND FRANCESCA PELOSI[†]

Abstract. In this paper we present a new, composite, method for the construction of shape-preserving surfaces approximating a set of spatial data, obtained by combining together: a scheme for detecting the shape of the data, a new class of tensor-product splines, and a suitable linear least-squares strategy. This method is mainly conceived for the reconstruction of objects in reverse engineering.

Key words. approximation, parametric surfaces, tensor-product splines, shape preservation

AMS subject classifications. 65D07, 65D10, 65D17

DOI. 10.1137/070694843

1. Introduction. The construction of a mathematical model reproducing given objects plays an important role in many practical fields (medicine, geology, engineering, fine arts) essentially because it allows the possibility both of storing them in archives, databases, or electronic museum catalogs and of analyzing their physical, mechanical, aesthetic, etc., characteristics. Typically, such a construction is obtained computing a surface approximating a set of data points taken from the object itself; moreover, the measurements often involve a huge amount of data and are (very) accurate, so that the points furnish a good representation of the shape of the object.

It is clear that the surface must reproduce the main visual features (corners, ridges, patches with “visually uniform” curvature, etc.), and this requirement pushes our construction within the frame of the so-called *shape-preserving approximation*. It is worthwhile to recall that, despite its importance in practical applications and in contrast to the contiguous field of *shape-preserving interpolation*, shape-preserving approximation has not received a considerable attention. Indeed, the few available methods concern the construction of parametric curves and, to the best of our knowledge, only the paper [15] has been published on data approximation using parametric surfaces with shape constraints. The motivation is twofold: first, multivariate approximation is not an easy problem and becomes very hard when shape constraints are added,¹ and, second, it is intrinsically difficult to define the shape of the data and thus to properly set the shape constraints.

Suppose we are given a set of points $\{(t_\mu, \mathbf{P}_\mu), \mu = 0, \dots, M\}$ where $\mathbf{P}_\mu \in \mathbb{R}^d$, $d = 2, 3$, and $t_\mu \in \mathbb{R}$ is the associated value of the parameter. In order to better understand our strategy, consider for the moment the simpler and better known problem of constructing interpolating curves. A typical scheme is usually composed of the following steps. First we compute the piecewise linear curve interpolating the data points, and, second, we use it for extracting their geometric characteristics (e.g., discrete curvature and torsion). Then we construct an interpolating curve using some

*Received by the editors June 19, 2007; accepted for publication (in revised form) June 4, 2008; published electronically October 24, 2008. This work was supported by MIUR under project FIRB, contract RBAU0128CL.

<http://www.siam.org/journals/sinum/47-1/69484.html>

[†]Dipartimento di Scienze Matematiche ed Informatiche, Università di Siena, 53100 Siena, Italy (costantini@unisi.it, pelosi@unisi.it).

¹Note that also for the much easier problem of *functional bivariate interpolation* there are few effective methods.

kind of *splines in tension*, that is, splines which depend on a set of *tension parameters* and which can be modified (stretched) so that their shape tends to the shape of the piecewise linear curve. In other words (a) we construct a reference curve, (b) define the shape of the data *as* the shape of the reference curve, and (c) construct an interpolating (smoother) curve which can reproduce as far as we want the shape of the reference curve.

The method we have recently proposed for the construction of shape-preserving approximating curves ([3],[4],[5]) follows essentially the above steps.

If we want to proceed as in (a) we need at first (a.1) to define a knot sequence in the parameter interval $[t_0, t_M]$, extracting from the parameter values $\{t_0, \dots, t_M\}$ a subsequence $\{u_0, \dots, u_m\}$ (typically $m \ll M$), and then (a.2) to construct a piecewise reference curve which gives a good reproduction of the shape of the data. For (a.1) we have used the so-called *zero moment approach* ([1] and references quoted therein, [18]; see also section 2), which provides an efficient tool for selecting the *significant* knots, where the data exhibit a change in the shape. Starting with our experiences on spatial shape-preserving interpolating curves, in [3] and [4] we used piecewise linear splines as reference curves, but we soon realized that they were not completely satisfactory for these purposes. Therefore, on the basis of experimental evidence (see, e.g., [5], Figure 3) we have used the space S_2^0 of C^0 piecewise quadratic splines for constructing the reference curve required in (a.2). We have then (b) defined the shape of the data as the shape of the selected reference curve and for (c) we have used a new spline space, isomorphic to the space S_4^2 of C^2 , four degree splines, which “tends” to S_2^0 for limit values of the tension parameters.

We remark that the advantages of using a curve from this new space instead of the (C^0) reference curve itself are in the possibility of locally modifying the shape and thus in reproducing either smooth sections or sharp corners of the underlying object.

The aim of the present paper is to describe a similar scheme for the construction of parametric surfaces, approximating a set of spatial data. Such surfaces are given by tensor product, variable degree polynomial splines, whose parameters are defined on rectangular grids of the form $\{u_0, \dots, u_m\} \otimes \{v_0, \dots, v_n\}$. The sections in which this paper is subdivided follow essentially the constructive steps given above.

More specifically, in the next one we present an algorithm for the selection of the knot lines, describe the construction of the reference surface, and define the shape of the data. We anticipate that the shape will be defined in terms of the *zero moment* of the reference surface. This choice (which does not involve the Gaussian curvature and only indirectly the mean curvature) is motivated by the good performance in detecting the shape of the data,² by the simplicity in the theoretical definition of the shape constraints, and in the acceptable complexity of their practical implementation. In order to avoid possible misunderstandings, we remark that the reference surface is given by a C^0 quadratic tensor-product spline, and we cannot expect the *classical* shape preserving properties of bilinear splines; therefore, in this paper (as well as in [5], which deals with approximating curves) there are no *standard* tension type theorems, which typically provide conditions for the elimination of extraneous oscillations. Indeed, our choice is motivated by the good flexibility of the biquadratic patches which gives a good reproduction of the changes in the shape of the data.

In section 3 we briefly recall some preliminary results, define the space of tensor-product variable degree splines together with their convergence properties, and then

²A graphical comparison of the *zero moments* with respect to the mean and Gaussian curvature for the approximating surface is given in Figures 9 and 10.

we present a global and a local algorithm for the construction of the final shape preserving surface.

In section 4 we report the graphical examples, and we close the paper with final comments, remarks, and anticipations of related works in section 5.

2. The shape of the data. In this section we want to define the *shape of the data*. Starting with the experimental observation that several data sets derived from profiles or sections of real objects are well represented by C^0 piecewise quadratic curves (see [5]), we develop the natural extension of this idea for surface data.

2.1. The zero moment analysis. We assume we are given a set of spatial data, together with the corresponding parameter values.³ In order to cover the main practical applications, we will consider data—taken from a function $\mathcal{P} = \mathcal{P}(t, r)$ —both with a tensor product topology

$$(2.1) \quad \mathcal{P} = \{(t_\mu, r_\nu, \mathbf{P}_{\mu,\nu}) : \mathbf{P}_{\mu,\nu} \in \mathbb{R}^3; \mu = 0, \dots, M; \nu = 0, \dots, N\}$$

or scattered

$$(2.2) \quad \mathcal{P} = \{(t_\mu, r_\mu, \mathbf{P}_\mu) : \mathbf{P}_\mu \in \mathbb{R}^3; \mu = 0, \dots, L\}.$$

In any case we assume that the parameter values are contained in a rectangular domain \mathcal{D} , where, setting $T := \{t_0, t_1, \dots\}$, $R := \{r_0, r_1, \dots\}$

$$\mathcal{D} = [\min(T), \max(T)] \times [\min(R), \max(R)].$$

As already said in the introduction, the first step of our construction is the detection of proper sequences of grid lines, taken in the parameter domain. In other words, we extract from the data parameters two subsequences of knots, $\mathcal{U} = \{u_0, \dots, u_m\}$, with $u_i = t_{\mu_i}$ and $\mathcal{V} = \{v_0, \dots, v_n\}$ with $v_j = r_{\nu_j}$ and subdivide \mathcal{D} using the rectangles

$$\mathcal{D}_{i,j} := [u_i, u_{i+1}] \times [r_j, r_{j+1}]$$

such that the corresponding subsets of points $\mathcal{P}_{i,j}$ define subregions of *uniform shape*.

In the following subsection we will also use \mathcal{U} and \mathcal{V} for defining the knot sequences of the tensor product splines.

The method we propose for such selection is based on some results on zero moments, described in [1]. First we briefly summarize the basic results of [1], and then we recall how to use them in order to select the knot sequences (for details see [17]).

Let $\mathbf{x} : \mathcal{T} \rightarrow \mathbb{R}^3$ be a parametric surface from the parameter manifold $\mathcal{T} \subset \mathbb{R}^2$. In accordance with [1], [18], we state the following.

DEFINITION 1. For any $(u, v) \in \mathcal{T}$ the zero moment of \mathbf{x} at (u, v) is given by the barycenter $\mathbf{M}_\epsilon^0(\mathbf{x}(u, v))$ of $\mathbf{x}(\mathcal{T}) \cap \Theta_\epsilon(\mathbf{x}(u, v))$:

$$(2.3) \quad \mathbf{M}_\epsilon^0(\mathbf{x}(u, v)) := \frac{1}{|\mathbf{x}(\mathcal{T}) \cap \Theta_\epsilon|} \int_{\mathcal{I}_\epsilon} \mathbf{x} \, dA,$$

where dA is the area element; $\Theta_\epsilon := \Theta_\epsilon(\mathbf{x}(u, v))$ is the Euclidean ball in \mathbb{R}^3 of radius ϵ , centered at $\mathbf{x}(u, v)$, and the domain of the integral is defined as $\mathcal{I}_\epsilon := \{(t, r) \in \mathcal{T} : \mathbf{x}(t, r) \in \Theta_\epsilon(\mathbf{x}(u, v))\}$.

³This heavy assumption will be discussed in section 5.

Using the zero moment we can define difference vector

$$(2.4) \quad \mathbf{n}_\epsilon(u, v; \mathbf{x}) := \mathbf{M}_\epsilon^0(\mathbf{x}(u, v)) - \mathbf{x}(u, v),$$

which will be called the ϵ -normal; the main result of [1] relates such a vector with the mean curvature of $\mathbf{x}(u, v)$ and the normal vector of the surface at (u, v) .

THEOREM 1. *Let $\mathbf{x} : \mathcal{T} \rightarrow \mathbb{R}^3$ be a regular parametric surface. For $(u, v) \in \mathcal{T}$ consider a ball of radius ϵ with center $\mathbf{x}(u, v)$. Then*

$$(2.5) \quad \mathbf{n}_\epsilon(u, v; \mathbf{x}) = -\epsilon^2 \frac{1}{6} H(u, v) \mathbf{n}(u, v) + o(\epsilon^2),$$

where $H(u, v)$ and $\mathbf{n}(u, v)$ are, respectively, the mean curvature and the normal vector of the surface \mathbf{x} at (u, v) .

From the above theorem we have that $\|\mathbf{n}_\epsilon(u, v; \mathbf{x})\|/\epsilon^2$ is proportional to the mean curvature $H(u, v)$, and thus it can be used to select corners or sharp changes in the shape of $\mathbf{x}(u, v)$. Moreover, since the vector $\mathbf{n}_\epsilon(u, v; \mathbf{x})$ lies parallel to the normal $\mathbf{n}(u, v)$, it gives information about convex or concave behaviors of $\mathbf{x}(u, v)$, in a neighborhood of (u, v) . In virtue of these considerations, $\mathbf{n}_\epsilon(u, v; \mathbf{x})$ will be called *local ϵ shape of \mathbf{x}* .

In view of the contents of the following sections, we remark that the ϵ -normal can be defined also at irregular points, still providing geometric information.

In order to select the sequences \mathcal{U} and \mathcal{V} of knots from the parameter values, we have to look for both sharp corners and changes in convexity of the data. We need two algorithms, respectively, Algorithm 1 for data of the forms (2.1) and Algorithm 2 for data of the form (2.2). Such algorithms are extremely important for the success of our method but also rather complicated and, in order to avoid interruptions in the reading of the paper, we have preferred to postpone their description until the final appendix. Here we limit ourselves to say that they are based on the discrete (and easier) counterpart of (2.3) studied in [17], namely

$$\mathbf{M}_\epsilon^0(\mathbf{P}_{\mu, \nu}) := \frac{1}{\text{card}(\Theta_\epsilon(\mathbf{P}_{\mu, \nu}))} \sum_{\mathbf{P}_{q, l} \in \text{wn}\Theta_\epsilon(\mathbf{P}_{\mu, \nu})} \mathbf{P}_{q, l}$$

in the case (2.1) or

$$\mathbf{M}_\epsilon^0(\mathbf{P}_\mu) := \frac{1}{\text{card}(\Theta_\epsilon(\mathbf{P}_\mu))} \sum_{\mathbf{P}_q \in \Theta_\epsilon(\mathbf{P}_\mu)} \mathbf{P}_q$$

in the case (2.2), and the changes of the corresponding *discrete ϵ -normals*

$$\mathbf{n}_\epsilon(\mathbf{P}_{\mu, \nu}) := \mathbf{M}_\epsilon^0(\mathbf{P}_{\mu, \nu}) - \mathbf{P}_{\mu, \nu} \quad \text{or} \quad \mathbf{n}_\epsilon(\mathbf{P}_\mu) := \mathbf{M}_\epsilon^0(\mathbf{P}_\mu) - \mathbf{P}_\mu$$

are compared with tolerances tol_{peak} and tol_{cc} for detecting, respectively, the peaks and the lines of convexity changes of the data points. Clearly, a good choice of the tolerances ϵ , tol_{peak} , and tol_{cc} is crucial for a suitable detection of the data subsets with *uniform shape*, as distinctly appears, for instance, in Figure 1.

2.2. The reference surface. For an arbitrary sequence of ordered knots $\{z_0, z_1, \dots, z_q\}$, let

$$\mathbf{S}_2^0[z_0, z_q] := \{\mathbf{s} : [z_0, z_l] \rightarrow \mathbb{R}^3 : \mathbf{s} \in C^0[z_0, z_l] \text{ s.t. } \mathbf{s}|_{[z_i, z_{i+1}]}\text{ has components in } \mathbb{P}_2 \text{ for } i = 0, \dots, q-1\},$$

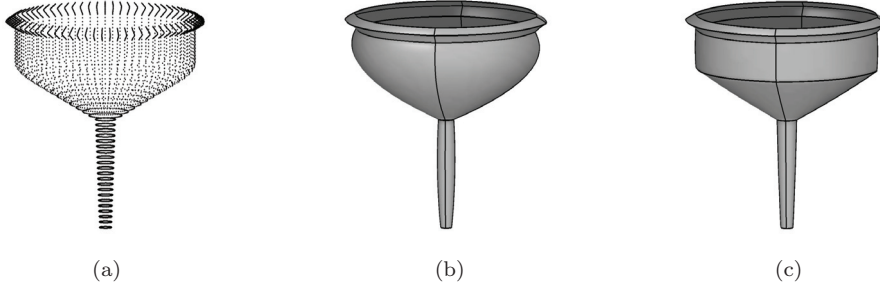


FIG. 1. Influence of the input parameters on Algorithm 1: (a) Data points. (b) Gridlines and reference surface, large tol_{peak} . (c) Gridlines and reference surface, small tol_{peak} .

be the space of parametric C^0 , quadratic splines. We take the parameter sequences $\mathcal{U} = \{u_0, \dots, u_m\}$ and $\mathcal{V} = \{v_0, \dots, v_n\}$ given by the algorithms described in the previous subsection and form the tensor product space of C^0 biquadratic parametric splines

$$(2.6) \quad \mathbf{S}_{2,\mathcal{U}}^0 \otimes \mathbf{S}_{2,\mathcal{V}}^0 := \mathbf{S}_2^0[u_0, u_m] \otimes \mathbf{S}_2^0[v_0, v_n].$$

Let $\mathbf{f} : \mathcal{D} \mapsto \mathbb{R}^3$. We introduce the following semi-norm

$$(2.7) \quad |\mathbf{f}| := \sqrt{\sum_{\mu=0}^M \sum_{\nu=0}^N \|\mathbf{f}(t_\mu, r_\nu)\|_2^2} \quad \text{or} \quad |\mathbf{f}| := \sqrt{\sum_{\mu=0}^L \|\mathbf{f}(t_\mu, r_\mu)\|_2^2},$$

respectively, for the cases (2.1) or (2.2). Justified by practical applications (where we have many and well spread data), we assume the following property.

ASSUMPTION 1. *The semi-norm (2.7) is a norm for all the (finite dimensional) tensor-product spline spaces introduced in the present and in the following sections.*

Following the scheme proposed in [5] for spatial curves, we construct $\boldsymbol{\sigma}^* \in \mathbf{S}_{2,\mathcal{U}}^0 \otimes \mathbf{S}_{2,\mathcal{V}}^0$, the best approximation to data.

$$|\boldsymbol{\sigma}^* - \mathcal{P}| \leq |\boldsymbol{\sigma} - \mathcal{P}|, \quad \forall \boldsymbol{\sigma} \in \mathbf{S}_{2,\mathcal{U}}^0 \otimes \mathbf{S}_{2,\mathcal{V}}^0.$$

Remark 1. Since, obviously, the minimization of $|\boldsymbol{\sigma} - \mathcal{P}|$ is equivalent to that of $|\boldsymbol{\sigma} - \mathcal{P}|^2$, the computation of $\boldsymbol{\sigma}^*$ requires the solution of three independent discrete least squares problems for the three components of the surface.

We use $\boldsymbol{\sigma}^*$ for a comprehensive visual description of the data set; that is, we set the following definition.

DEFINITION 2. *The (local) shape of the data \mathcal{P} is given by $\mathbf{n}_\epsilon(u, v; \boldsymbol{\sigma}^*)$.*

We conclude this section by observing that the use of the *local ϵ shape* of $\boldsymbol{\sigma}^*$ has the advantage of being both simple and effective in the description of the data shape.

3. The tensor product spline spaces. In this section we want to define the tensor product space of *quartic-like* variable degree polynomial splines (VDPS for short) and analyze its main properties. We start recalling some basic facts on one-dimensional splines, referring for details to [2] and [5].

3.1. C^2 VDPS curves. Let $\mathcal{Z} = \{z_0, z_1, \dots, z_q\}$ be an ordered knot sequence, let $h_i := z_{i+1} - z_i$, $i = 0, \dots, q-1$, and let $\mathcal{K} = \{k_0, \dots, k_q\}$, $k_i \geq 4$ be a given sequence

of integers. In the following \mathcal{B}_j^ℓ will denote the j th ℓ -degree Bernstein polynomial and λ_j^ℓ the corresponding Bézier control net. In other words $\mathcal{B}_j^\ell = B^{(\ell)}\lambda_j^\ell$ where $B^{(\ell)}$ is the ℓ -degree Bernstein operator. For each interval $[z_i, z_{i+1}]$ we consider the five-dimensional polynomial space:

$$(3.1) \quad VP_{4,k_i,k_{i+1}} = \text{span} \{ (1-w)^{k_i}, \mathbb{P}_2, w^{k_{i+1}} \},$$

with $w = (z - z_i)/h_i$. Any $\phi_i \in VP_{4,k_i,k_{i+1}}$ can be expressed as

$$(3.2) \quad \phi_i(w) = e_i^-(1-w)^{k_i} + b_{i,0}^2 \mathcal{B}_0^2(w) + b_{i,1}^2 \mathcal{B}_1^2(w) + b_{i,2}^2 \mathcal{B}_2^2(w) + e_i^+ w^{k_{i+1}},$$

with $e_i^-, e_i^+, b_{i,0}^2 \in \mathbb{R}$. Note that the break points $\xi_{i,0}, \xi_{i,1}, \dots, \xi_{i,4}$ of the corresponding piecewise linear function

$$\lambda_i(w) = e_i^- \lambda_0^{k_i}(w) + b_{i,0}^2 \lambda_0^2(w) + b_{i,1}^2 \lambda_1^2(w) + b_{i,2}^2 \lambda_2^2(w) + e_i^+ \lambda_{k_{i+1}}^{k_{i+1}}(w)$$

are located at

$$\xi_{i,0} = z_i, \quad \xi_{i,1} = z_i + \frac{h_i}{k_i}, \quad \xi_{i,2} = z_i + \frac{h_i}{2}, \quad \xi_{i,3} = z_{i+1} - \frac{h_i}{k_{i+1}}, \quad \xi_{i,4} = z_{i+1}$$

and have values

$$\begin{aligned} \beta_{i,0} &= b_{i,0}^2 + e_i^-, \quad \beta_{i,1} = \left(1 - \frac{2}{k_i}\right) b_{i,0}^2 + \frac{2}{k_i} b_{i,1}^2, \quad \beta_{i,2} = b_{i,1}^2, \\ \beta_{i,3} &= \frac{2}{k_{i+1}} b_{i,1}^2 + \left(1 - \frac{2}{k_{i+1}}\right) b_{i,2}^2, \quad \beta_{i,4} = b_{i,2}^2 + e_i^+. \end{aligned}$$

For further references, we explicitly write the break points of λ_i :

$$(3.3) \quad \{(\xi_{i,0}, \beta_{i,0}), (\xi_{i,1}, \beta_{i,1}), (\xi_{i,2}, \beta_{i,2}), (\xi_{i,3}, \beta_{i,3}), (\xi_{i,4}, \beta_{i,4})\}.$$

Obviously $\phi_i \in \mathbb{P}_{\bar{k}}$ where $\bar{k} = \max\{k_i, k_{i+1}\}$; in [2], Theorem 2.1, the following result is given.

THEOREM 2. *The sequence $b_{i,0}^{\bar{k}}, \dots, b_{i,\bar{k}}^{\bar{k}}$ of control points of ϕ_i are obtained via repeated convex combinations of $\{\beta_{i,0}, \dots, \beta_{i,4}\}$.*

The convex combinations quoted in the theorem can be described by *degree elevation operators* (see [12]) and can be represented by a matrix (constructed as the product of the bidiagonal matrices describing each convex combination/degree elevation step), namely

$$(3.4) \quad \begin{bmatrix} b_{i,0}^{\bar{k}} \\ b_{i,1}^{\bar{k}} \\ \vdots \\ b_{i,\bar{k}}^{\bar{k}} \end{bmatrix} = A^{(\bar{k})} \begin{bmatrix} \beta_{i,0} \\ \beta_{i,1} \\ \vdots \\ \beta_{i,4} \end{bmatrix}; \quad A^{(\bar{k})} \in \mathbb{R}^{(\bar{k}+1) \times 5}.$$

Given the well-known properties of Bernstein polynomials, Theorem 2 says that ϕ_i lies in the convex hull of $\{\beta_{i,0}, \dots, \beta_{i,4}\}$ and satisfies the usual end tangent conditions.

Using (3.1) we easily define the space of C^2 *quartic-like* VDPS

$$VS_{4,\mathcal{K}}^2 = \{s \in C^2[z_0, z_q] \text{ s.t. } s(z) = \phi_i(w); \phi_i \in VP_{4,k_i,k_{i+1}}, z \in [z_i, z_{i+1}], w = (z - z_i)/h_i, i = 0, \dots, q-1\}.$$

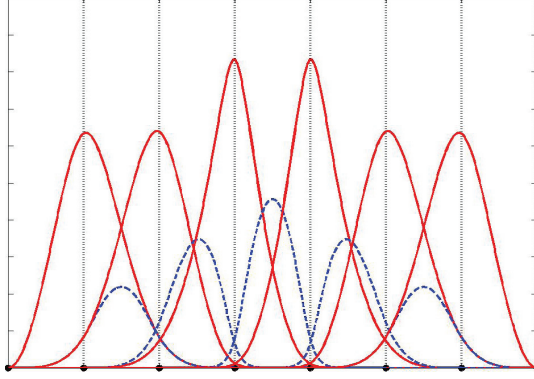


FIG. 2. Some B-spline basis functions: N_i^S (solid), N_i^D (dashed) with $\mathcal{K} = \{4, 4, 4, 10, 10, 4, 4, 4\}$.

Let us take an extended knot sequence $z_{-2} < z_{-1} < z_0, z_q < z_{q+1} < z_{q+2}$, and an extended degree sequence $k_{-2} = k_{-1} = k_0, k_q = k_{q+1} = k_{q+2}$. In [5] it is shown that $VS_{4,\mathcal{K}}^2$ admits a stochastic and compactly supported B-spline-like basis, denoted by $\{N_{-1}^D, N_0^S, N_0^D, \dots, N_q^S, N_q^D\}$ (we have adopted this notation for relating this paper with the geometric construction described in [2], [5]; here it suffices to qualitatively say that the N_i^S and the N_i^D are related, respectively, to break and to internal points; see Figure 2). Note that, for $H \in \{S, D\}$, $N_i^H = N_i^H(\cdot; \mathcal{Z}, \mathcal{K})$. Figure 2 shows some B-spline basis functions. A spline curve in \mathbb{R}^3 can be constructed as

$$(3.5) \quad \mathbf{VS}_{4,\mathcal{K}}^2 = \{\mathbf{s} : [z_0, z_q] \rightarrow \mathbb{R}^3 \text{ s.t. has components in } VS_{4,\mathcal{K}}^2\},$$

and any spline curve $\mathbf{s} \in \mathbf{VS}_{4,\mathcal{K}}^2$ can be expressed as

$$(3.6) \quad \mathbf{s} = \sum_{i=0}^q \mathcal{L}_i^S N_i^S + \sum_{i=-1}^q \mathcal{L}_i^D N_i^D.$$

The points $\mathcal{L}_{-1}^D, \mathcal{L}_0^S, \mathcal{L}_0^D, \dots, \mathcal{L}_{q-1}^D, \mathcal{L}_q^S, \mathcal{L}_q^D$ are called *pseudo-de Boor control points* and play the same role as the classical control points for quartic splines. In particular, the control points $\{\beta_{i,0}, \dots, \beta_{i,4}\}$ of (3.3), corresponding to $\mathbf{s}|_{[z_i, z_{i+1}]}$, can be computed with some steps of a corner-cutting procedure (for details see [2]).

Remark 2. We easily see from (3.2) that the computational cost for the evaluation of ϕ_i is approximately the same as for a polynomial of \mathbb{P}_4 . This property is clearly inherited by the splines of \mathbf{VS} (and later by tensor product spline surfaces) which have a computational cost equivalent to that of quadratic splines in \mathbf{S}_4^2 .

3.2. C^2 VDPS surfaces. Given two extended sequences of knots

$$\mathcal{U} = \{u_{-2}, \dots, u_{m+2}\}, \quad \mathcal{V} = \{v_{-2}, \dots, v_{n+2}\}$$

and two extended sequences of degrees

$$\mathcal{K}^u = \{k_{-2}^u, \dots, k_{m+2}^u\}, \quad \mathcal{K}^v = \{k_{-2}^v, \dots, k_{n+2}^v\},$$

we consider the univariate spaces VS_{4,\mathcal{K}^u}^2 and VS_{4,\mathcal{K}^v}^2 as specified in the above subsection. The tensor-product space is given by

$$(3.7) \quad VS_{4,\mathcal{K}^u}^2 \otimes VS_{4,\mathcal{K}^v}^2 = \text{span} \left\{ \begin{aligned} &N_{i,j}^{DD}, \quad i = -1, \dots, m, \quad j = -1, \dots, n; \\ &N_{i,j}^{DS}, \quad i = -1, \dots, m, \quad j = 0, \dots, n; \\ &N_{i,j}^{SD}, \quad i = 0, \dots, m, \quad j = -1, \dots, n; \\ &N_{i,j}^{SS}, \quad i = 0, \dots, m, \quad j = 0, \dots, n \end{aligned} \right\},$$

where the basis functions are given by

$$\begin{aligned} N_{i,j}^{DD} &:= N_i^D(\cdot; \mathcal{U}, \mathcal{K}^u) N_j^D(\cdot; \mathcal{V}, \mathcal{K}^v), \quad N_{i,j}^{DS} := N_i^D(\cdot; \mathcal{U}, \mathcal{K}^u) N_j^S(\cdot; \mathcal{V}, \mathcal{K}^v), \\ N_{i,j}^{SD} &:= N_i^S(\cdot; \mathcal{U}, \mathcal{K}^u) N_j^D(\cdot; \mathcal{V}, \mathcal{K}^v), \quad N_{i,j}^{SS} := N_i^S(\cdot; \mathcal{U}, \mathcal{K}^u) N_j^S(\cdot; \mathcal{V}, \mathcal{K}^v). \end{aligned}$$

The plot of some tensor product basis functions is shown in Figure 3 for different degree sequences.

From Theorem 2 of [5], the following properties immediately hold:

- (a) $N_{i,j}^{DD}(u, v) > 0$ for $u \in (u_{i-1}, u_{i+2}) \times (v_{j-1}, v_{j+2})$ and $N_{i,j}^{DD}(u, v) = 0$ otherwise;
- $N_{i,j}^{DS}(u, v) > 0$ for $u \in (u_{i-1}, u_{i+2}) \times (v_{j-2}, v_{j+2})$ and $N_{i,j}^{DS}(u, v) = 0$ otherwise;
- $N_{i,j}^{SD}(u, v) > 0$ for $u \in (u_{i-2}, u_{i+2}) \times (v_{j-1}, v_{j+2})$ and $N_{i,j}^{SD}(u, v) = 0$ otherwise;
- $N_{i,j}^{SS}(u, v) > 0$ for $u \in (u_{i-2}, u_{i+2}) \times (v_{j-2}, v_{j+2})$ and $N_{i,j}^{SS}(u, v) = 0$ otherwise;
- (b) $\sum_{i=-1}^m \sum_{j=-1}^n N_{i,j}^{DD}(u, v) + \sum_{i=-1}^m \sum_{j=0}^n N_{i,j}^{DS}(u, v) + \sum_{i=0}^m \sum_{j=-1}^n N_{i,j}^{SD}(u, v) + \sum_{i=0}^m \sum_{j=0}^n N_{i,j}^{SS}(u, v) \equiv 1$
- (c) $N_{i,j}^{HH}(u, v)$ for $HH \in \{DD, DS, SD, SS\}$ is an element of the functional tensor product space (3.7).

Let $s \in VS_{4,\mathcal{K}^u}^2 \otimes VS_{4,\mathcal{K}^v}^2$. Obviously, $\phi_{i,j} := s|_{\mathcal{D}_{i,j}}$ is a bivariate polynomial in the tensor-product space obtained from (3.1), that is, $VP_{4,k_i^u, k_{i+1}^u} \otimes VP_{4,k_j^v, k_{j+1}^v}$. Following the notations of (3.2) and setting $\tilde{u} = (u - u_i)/(u_{i+1} - u_i)$, $\tilde{v} = (v - v_j)/(v_{j+1} - v_j)$ we have

$$\begin{aligned} VP_{4,k_i^u, k_{i+1}^u} \otimes VP_{4,k_j^v, k_{j+1}^v} &= \text{span} \left(\left\{ (1 - \tilde{u})^{k_i^u}, \mathcal{B}_0^2(\tilde{u}), \mathcal{B}_1^2(\tilde{u}), \mathcal{B}_2^2(\tilde{u}), \tilde{u}^{k_{i+1}^u} \right\} \right. \\ &\quad \left. \otimes \left\{ (1 - \tilde{v})^{k_j^v}, \mathcal{B}_0^2(\tilde{v}), \mathcal{B}_1^2(\tilde{v}), \mathcal{B}_2^2(\tilde{v}), \tilde{v}^{k_{j+1}^v} \right\} \right), \end{aligned}$$

and, recalling (3.3), $\phi_{i,j}$ admits a bivariate control polygon $\lambda_{i,j}$ defined by the control points

$$(3.8) \quad (\xi_{\mu,\nu}^{uv}, \beta_{\mu,\nu}^{uv}); \quad \xi_{\mu,\nu}^{uv} = \xi_{\mu}^u \xi_{\nu}^v, \quad \beta_{\mu,\nu}^{uv} \in \mathbb{R}, \quad \mu, \nu = 0, \dots, 4$$

The space of variable degree parametric surface is given by

$$(3.9) \quad \mathbf{VS}_{4,\mathcal{K}^u}^2 \otimes \mathbf{VS}_{4,\mathcal{K}^v}^2 := \{ \mathbf{s} : [u_0, u_m] \times [v_0, v_n] \mapsto \mathbb{R}^3 \text{ has component in (3.7)} \}$$

and any $\mathbf{s} \in \mathbf{VS}_{4,\mathcal{K}^u}^2 \otimes \mathbf{VS}_{4,\mathcal{K}^v}^2$ can be represented in the form

$$(3.10) \quad \begin{aligned} \mathbf{s} &= \sum_{i=-1}^m \sum_{j=-1}^n \mathcal{L}_{i,j}^{DD} N_{i,j}^{DD} + \sum_{i=-1}^m \sum_{j=0}^n \mathcal{L}_{i,j}^{DS} N_{i,j}^{DS} \\ &\quad + \sum_{i=0}^m \sum_{j=-1}^n \mathcal{L}_{i,j}^{SD} N_{i,j}^{SD} + \sum_{i=0}^m \sum_{j=0}^n \mathcal{L}_{i,j}^{SS} N_{i,j}^{SS}. \end{aligned}$$

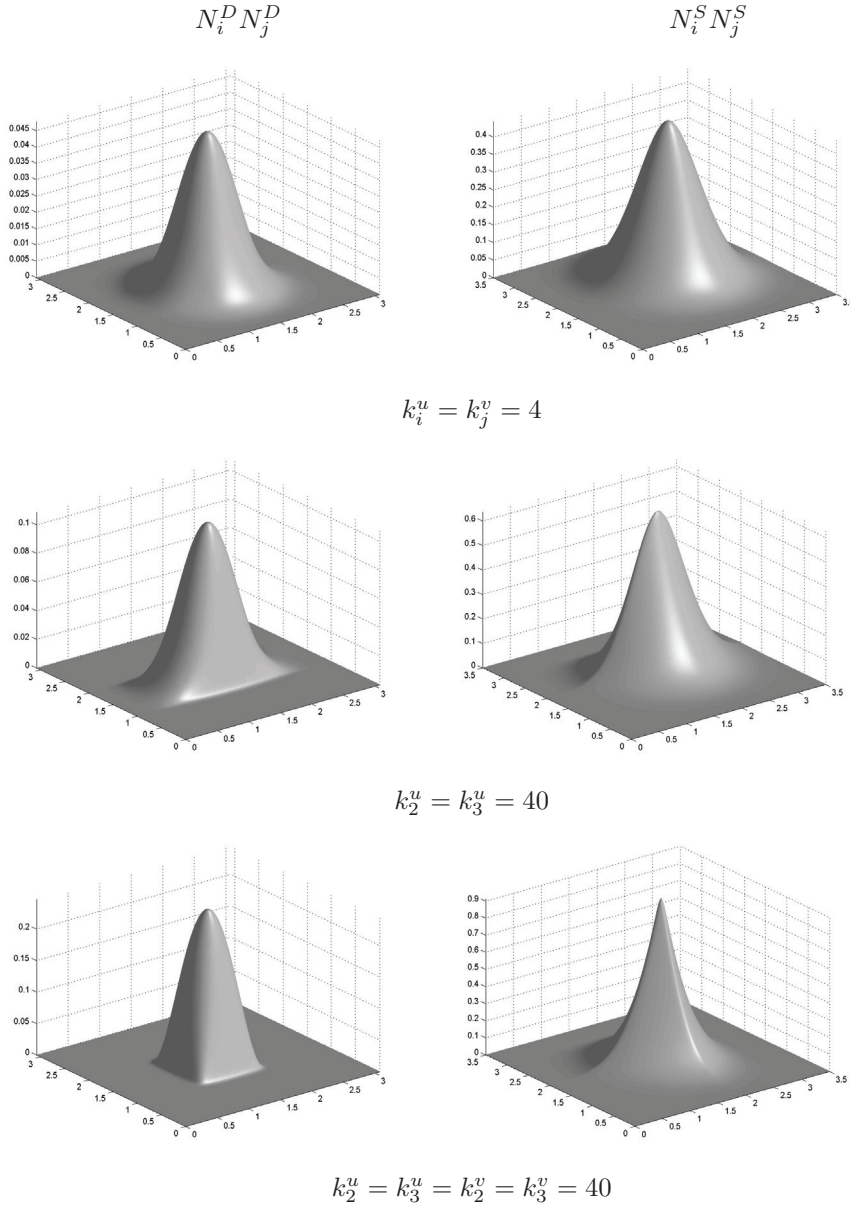


FIG. 3. Some tensor product B-spline basis functions with different degrees.

As a consequence of the properties of the basis functions, we have that the *pseudo-de Boor control net*, defined by the piecewise bilinear function interpolating the points $\mathcal{L}_{i,j}^{DD}$, $\mathcal{L}_{i,j}^{DS}$, $\mathcal{L}_{i,j}^{SD}$, and $\mathcal{L}_{i,j}^{SS}$ in (3.10), plays the role of the usual de Boor control net for C^2 quartic spline surface.

Remark 3. The basis functions $N_{i,j}^{HH}(u,v)$, $HH \in \{DD, DS, SD, SS\}$, tend to the corresponding basis elements of the space $\mathcal{S}_{2,u}^0 \otimes \mathcal{S}_{2,v}^0$ for limit values of the degrees. Despite the technical details, the convergence results of subsection 3.4 are essentially based on this property.

3.3. Boundary conditions. Given the rectangular shape of the parameter domain \mathcal{D} , the sections of the surface \mathbf{s} along the isoparametric lines $u = u_i$ and $v = v_j$ are VDPS spline curves from (3.5). This means that for each parametric lines of the form

$$\mathbf{s}(u, v_j), \quad j = 0, 1, \dots, n; \quad \mathbf{s}(u_i, v), \quad i = 0, 1, \dots, m,$$

we need to add suitable boundary conditions.

For open parametric lines, we consider restricted polynomial spaces at the first and last subintervals of the knot sequences of the form

$$VP_{4,k_1} = \text{span} \{ \mathbb{P}_2, w^{k_1} \}, \quad VP_{4,k_{q-1}} = \text{span} \{ (1-w)^{k_{q-1}}, \mathbb{P}_2 \},$$

and we construct the univariate space $\mathbf{VS}_{4,\mathcal{K},\text{open}}^2$ as

$$(3.11) \quad \mathbf{VS}_{4,\mathcal{K},\text{open}}^2 = \left\{ \mathbf{s} \in \mathbf{VS}_{4,\mathcal{K}}^2 : \mathbf{s}|_{[z_0, z_1]} \text{ and } \mathbf{s}|_{[z_{q-1}, z_q]} \right. \\ \left. \text{have components in } VP_{4,k_1} \text{ and } VP_{4,k_{q-1}} \right\},$$

so that $\dim(\mathbf{VS}_{4,\mathcal{K},\text{open}}^2) = \dim(\mathbf{S}_2^0)$. Also this space admits a B-spline basis, $\mathcal{N}_0^S, \mathcal{N}_0^D, \dots, \mathcal{N}_{q-1}^D, \mathcal{N}_q^S$, obtained with a slight modification of the scheme used for (3.5).

In the case of closed grid lines, we consider the univariate space $\mathbf{VS}_{4,\mathcal{K},\text{closed}}^2$ of the form

$$(3.12) \quad \mathbf{VS}_{4,\mathcal{K},\text{closed}}^2 = \left\{ \mathbf{s} \in \mathbf{VS}_{4,\mathcal{K}}^2 : \mathbf{s} = \sum_{i=0}^{q-1} \left(\mathcal{L}_i^S \mathcal{N}_i^S + \mathcal{L}_i^D \mathcal{N}_i^D \right) \right\},$$

where the basis functions \mathcal{N}_i^S and \mathcal{N}_i^D are obtained by imposing periodic extension of the boundary control points:

$$\begin{aligned} \mathcal{N}_0^S &:= N_0^S + N_q^S, & \mathcal{N}_i^S &:= N_i^S, & i &= 1, \dots, q-1, \\ \mathcal{N}_0^D &:= N_0^D + N_q^D, & \mathcal{N}_i^D &:= N_i^D, & i &= 1, \dots, q-2, \\ \mathcal{N}_{q-1}^D &:= N_{q-1}^D + N_{-1}^D. \end{aligned}$$

We limit ourselves to say that the properties of B-splines do not hold for the first and last group of the above functions and refer for details to [5].

Now, due to the tensor product structure, we can construct four different subspaces of $\mathbf{VS}_{4,\mathcal{K}^u}^2 \otimes \mathbf{VS}_{4,\mathcal{K}^v}^2$ in (3.9):

(I) The grid line curves $\mathbf{s}(u, v_j)$ and $\mathbf{s}(u_i, v)$ are open in both extreme knots, and then we form the tensor product space:

$$\mathbf{TVS}_{oo} := \mathbf{VS}_{4,\mathcal{K}^u,\text{open}}^2 \otimes \mathbf{VS}_{4,\mathcal{K}^v,\text{open}}^2,$$

where $\dim(\mathbf{TVS}_{oo}) = (2m+1)(2n+1)$. We can set

$$\begin{aligned} \mathbf{TVS}_{oo} = \text{span} \{ & \mathcal{N}_{i,j}^{DD}, \quad i = 0, \dots, m-1, \quad j = 0, \dots, n-1; \\ & \mathcal{N}_{i,j}^{DS}, \quad i = 0, \dots, m-1, \quad j = 0, \dots, n; \quad \mathcal{N}_{i,j}^{SD}, \quad i = 0, \dots, m, \quad j = 0, \dots, n-1; \\ & \mathcal{N}_{i,j}^{SS}, \quad i = 0, \dots, m, \quad j = 0, \dots, n \}. \end{aligned}$$

(II) The grid line curves $\mathbf{s}(u, v_j)$ are closed curves and $\mathbf{s}(u_i, v)$ are open, and then we form the tensor product space:

$$\mathbf{TVS}_{co} := \mathbf{VS}_{4,\mathcal{K}^u,\text{closed}}^2 \otimes \mathbf{VS}_{4,\mathcal{K}^v,\text{open}}^2,$$

where $\dim(\mathbf{TVS}_{co}) = 2m(2n + 1)$ and

$$\mathbf{TVS}_{co} = \text{span} \{ \mathcal{N}_{i,j}^{DD}, \mathcal{N}_{i,j}^{SD}, i = 0, \dots, m-1, j = 0, \dots, n-1; \\ \mathcal{N}_{i,j}^{DS}, \mathcal{N}_{i,j}^{SS}, i = 0, \dots, m-1, j = 0, \dots, n \}.$$

(III) The grid line curves $\mathbf{s}(u, v_j)$ are open and $\mathbf{s}(u_i, v)$ are closed, and then we form the tensor product space:

$$\mathbf{TVS}_{oc} := \mathbf{VS}_{4, \mathcal{K}^u, \text{open}}^2 \otimes \mathbf{VS}_{4, \mathcal{K}^v, \text{closed}}^2,$$

where $\dim(\mathbf{TVS}_{oc}) = 2n(2m + 1)$ and

$$\mathbf{TVS}_{oc} = \text{span} \{ \mathcal{N}_{i,j}^{DD}, \mathcal{N}_{i,j}^{DS}, i = 0, \dots, m-1, j = 0, \dots, n-1; \\ \mathcal{N}_{i,j}^{SD}, \mathcal{N}_{i,j}^{SS}, i = 0, \dots, m, j = 0, \dots, n-1 \}.$$

(IV) The grid line curves $\mathbf{s}(u, v_j)$ and $\mathbf{s}(u_i, v)$ are closed in both extreme knots, and then we form the tensor product space:

$$\mathbf{TVS}_{cc} := \mathbf{VS}_{4, \mathcal{K}^u, \text{closed}}^2 \otimes \mathbf{VS}_{4, \mathcal{K}^v, \text{closed}}^2,$$

where $\dim(\mathbf{TVS}_{cc}) = 4mn$ and we set

$$\mathbf{TVS}_{cc} = \text{span} \{ \mathcal{N}_{i,j}^{DD}, \mathcal{N}_{i,j}^{DS}, \mathcal{N}_{i,j}^{SD}, \mathcal{N}_{i,j}^{SS}, i = 0, \dots, m-1, j = 0, \dots, n-1 \}.$$

Note that type (II) is equivalent to impose that

$$\frac{\partial^{(r)}}{\partial u^{(r)}} \mathbf{s}(u, v)|_{(u_0, v)} = \frac{\partial^{(r)}}{\partial u^{(r)}} \mathbf{s}(u, v)|_{(u_m, v)},$$

and type (III) is equivalent to impose

$$\frac{\partial^{(r)}}{\partial v^{(r)}} \mathbf{s}(u, v)|_{(u, v_0)} = \frac{\partial^{(r)}}{\partial v^{(r)}} \mathbf{s}(u, v)|_{(u, v_n)}$$

for $r = 0, 1, 2$.

3.4. Convergence properties. It is clear from Definition (3.1) that, for limit values of the degrees k_i and k_{i+1} , the space $VP_{4, k_i, k_{i+1}}$ “tends” to the space \mathbb{P}_2 , that is, for any $\phi \in VP_{4, k_i, k_{i+1}}$, $\phi(w) = e_i^- (1-w)^{k_i} + b_{i,0}^2 \mathcal{B}_0^2(w) + b_{i,1}^2 \mathcal{B}_1^2(w) + b_{i,2}^2 \mathcal{B}_2^2(w) + e_i^+ w^{k_{i+1}}$, we have

$$\lim_{k_i, k_{i+1} \rightarrow \infty} \phi(w) = b_{i,0}^2 \mathcal{B}_0^2(w) + b_{i,1}^2 \mathcal{B}_1^2(w) + b_{i,2}^2 \mathcal{B}_2^2(w); \quad \forall w \in [a, b] \subset (z_i, z_{i+1}).$$

This property is clearly inherited by the spline space $VS_{4, \mathcal{K}}$ which, roughly speaking, “tends” to S_2^0 , the space of continuous quadratic splines. A first result is obtained as a straightforward consequence of Theorem 3 of [2].

THEOREM 3. *Let $\mathbf{s} \in \mathbf{TVS}_{xx}$, $xx \in \{oo, oc, co, cc\}$*

$$\mathbf{s}(u, v) = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \mathcal{L}_{i,j}^{DD} \mathcal{N}_{i,j}^{DD}(u, v) + \sum_{i=0}^{m-1} \sum_{j=0}^n \mathcal{L}_{i,j}^{DS} \mathcal{N}_{i,j}^{DS}(u, v) \\ + \sum_{i=0}^m \sum_{j=0}^{n-1} \mathcal{L}_{i,j}^{SD} \mathcal{N}_{i,j}^{SD}(u, v) + \sum_{i=0}^m \sum_{j=0}^n \mathcal{L}_{i,j}^{SS} \mathcal{N}_{i,j}^{SS}(u, v),$$

and let $\sigma_s \in \mathbf{S}_{2,\mathcal{U}}^0 \otimes \mathbf{S}_{2,\mathcal{V}}^0$ be defined by the control net $\mathcal{L}_{i,j}^{DD}, \mathcal{L}_{i,j}^{DS}, \mathcal{L}_{i,j}^{SD}$, and $\mathcal{L}_{i,j}^{SS}$ of s . Then, for all admissible indices, we have

$$\begin{aligned} \lim_{k_r^u, k_l^v, k_{l+1}^v \rightarrow \infty} s(u_r, v) &= \sigma_s(u_r, v), & v \in [v_l, v_{l+1}]; \\ \lim_{k_r^u, k_{r+1}^u, k_l^v \rightarrow \infty} s(u, v_l) &= \sigma_s(u, v_l), & u \in [u_r, u_{r+1}]; \end{aligned}$$

and

$$\lim_{k_r^u, k_{r+1}^u, k_l^v, k_{l+1}^v \rightarrow \infty} s(u, v) = \sigma_s(u, v), \quad (u, v) \in [u_r, u_{r+1}] \times [v_l, v_{l+1}].$$

The above theorem is not sufficient for our shape-preserving purposes. Let σ^* be our reference surface as defined in subsection 2.2 and let s^* be the best approximation in \mathbf{TVS}_{xx} ; that is,

$$|s^* - \mathcal{P}| \leq |s - \mathcal{P}|, \quad \forall s \in \mathbf{TVS}_{xx}.$$

Recalling Remark 1, we can easily state the following result.

THEOREM 4. *If $k_i^u, k_j^v \rightarrow \infty$; $i = 0, \dots, m$, $j = 0, \dots, n$, then*

$$|s^* - \sigma^*| \rightarrow 0.$$

Proof. For any $s \in \mathbf{TVS}_{xx}$ let us denote with σ_s the C^0 biquadratic spline surface in $\mathbf{S}_{2,\mathcal{U}}^0 \otimes \mathbf{S}_{2,\mathcal{V}}^0$ defined on the control net of s and, conversely, for any $\sigma \in \mathbf{S}_{2,\mathcal{U}}^0 \otimes \mathbf{S}_{2,\mathcal{V}}^0$ let us denote with s_σ the variable degree tensor product surface having the same control net as σ . Recalling the notations of subsection 3.2, let $\mathcal{K}_\ell^u = (k_{0,\ell}^u, \dots, k_{m,\ell}^u)$, $\mathcal{K}_\ell^v = (k_{0,\ell}^v, \dots, k_{n,\ell}^v)$ with $\lim_{\ell \rightarrow \infty} k_{i,\ell}^u = \lim_{\ell \rightarrow \infty} k_{j,\ell}^v = \infty$, all i, j , and let us denote $s = s(\mathcal{K}_\ell^u, \mathcal{K}_\ell^v)$. From Theorem 3 we have

$$\lim_{\ell \rightarrow \infty} |s^*(\mathcal{K}_\ell^u, \mathcal{K}_\ell^v) - \sigma_{s^*(\mathcal{K}_\ell^u, \mathcal{K}_\ell^v)}| = 0$$

and so $|s^*(\mathcal{K}_\ell^u, \mathcal{K}_\ell^v) - \sigma^*| \rightarrow 0$ if, and only if, $|\sigma_{s^*(\mathcal{K}_\ell^u, \mathcal{K}_\ell^v)} - \sigma^*| \rightarrow 0$.

Suppose that $|\sigma_{s^*(\mathcal{K}_\ell^u, \mathcal{K}_\ell^v)} - \sigma^*| \not\rightarrow 0$; then, for some subsequence $\{\mathcal{K}_{\ell(r)}^u, \mathcal{K}_{\ell(r)}^v\}$ and for a positive constant C , for sufficiently large r we have

$$|\sigma_{s^*(\mathcal{K}_{\ell(r)}^u, \mathcal{K}_{\ell(r)}^v)} - \sigma^*| \geq C.$$

We recall that the best approximations σ^* and $s^*(\mathcal{K}_{\ell(r)}^u, \mathcal{K}_{\ell(r)}^v)$ are unique; therefore we have $|\sigma_{s^*(\mathcal{K}_{\ell(r)}^u, \mathcal{K}_{\ell(r)}^v)} - \mathcal{P}| > |\sigma^* - \mathcal{P}|$ and for sufficiently large r , again by Theorem 3,

$$|s^*(\mathcal{K}_{\ell(r)}^u, \mathcal{K}_{\ell(r)}^v) - \mathcal{P}| > |\sigma_{s^*(\mathcal{K}_{\ell(r)}^u, \mathcal{K}_{\ell(r)}^v)} - \mathcal{P}|,$$

which leads to a contradiction. \square

We intend now to relate the convergence with shape properties. Using (2.3), (2.4), and Theorem 4 we have immediately the following result.

THEOREM 5. *If $k_i^u, k_j^v \rightarrow \infty$; $i = 0, \dots, m$, $j = 0, \dots, n$, then*

$$\mathbf{n}_\epsilon(u, v; s^*) \rightarrow \mathbf{n}_\epsilon(u, v; \sigma^*).$$

Recalling Definition 2 we can summarize the above property by saying that \mathbf{s}^* is *asymptotically shape-preserving*. The consequent shape-preserving criterium is formalized below (the notation is in accordance with Corollary 2).

COROLLARY 1. *Let $\delta_{i,j} > 0$, $i = 0, \dots, m$, $j = 0, \dots, n$. Then there exist threshold sequences $\tilde{\mathcal{K}}^u$, $\tilde{\mathcal{K}}^v$ such that for any $\mathcal{K}^u \geq \tilde{\mathcal{K}}^u$, $\mathcal{K}^v \geq \tilde{\mathcal{K}}^v$ (the inequalities hold componentwise), and for $(u, v) \in \mathcal{D}_{i,j}$*

$$(3.13) \quad \|\mathbf{n}_\epsilon(u, v; \mathbf{s}^*) - \mathbf{n}_\epsilon(u, v; \boldsymbol{\sigma}^*)\| \leq 2\delta_{i,j}.$$

However we note that its practical use requires extensive checks of the ϵ -normals associated to the progressive increases of the degree sequences. A weaker but cheaper criterium is given below.

Let $\mathbf{f}_{i,j}^* = \mathbf{s}_{|\mathcal{D}_{i,j}}^*$ and $\phi_{i,j}^* = \boldsymbol{\sigma}_{|\mathcal{D}_{i,j}}^*$ and let the corresponding control points (the 3D analogous of (3.8)) be denoted, respectively, by

$$(\boldsymbol{\xi}_{\mu,\nu}^{uv}, \boldsymbol{\beta}_{\mu,\nu}^{uv}(\mathbf{f}_{i,j}^*)), (\boldsymbol{\xi}_{\mu,\nu}^{uv}, \boldsymbol{\beta}_{\mu,\nu}^{uv}(\phi_{i,j}^*)); \mu, \nu = 0, \dots, 4.$$

LEMMA 1. *Let $\delta_{i,j} > 0$ and let $\|\boldsymbol{\beta}_{\mu,\nu}^{uv}(\mathbf{f}_{i,j}^*) - \boldsymbol{\beta}_{\mu,\nu}^{uv}(\phi_{i,j}^*)\| \leq \delta_{i,j}$; $\mu, \nu = 0, \dots, 4$. Then, for any $(u, v) \in \mathcal{D}_{i,j}$, $\|\mathbf{f}_{i,j}^*(u, v) - \phi_{i,j}^*(u, v)\| \leq \delta_{i,j}$.*

Proof. Let $\bar{k}^u = \max\{k_i^u, k_{i+1}^u\}$, $\bar{k}^v = \max\{k_j^v, k_{j+1}^v\}$, and let $B^{(\bar{k}^u)}$, $B^{(\bar{k}^v)}$ the Bernstein operators and $A^{(\bar{k}^u)}$, $A^{(\bar{k}^v)}$ the degree elevation convex combination operators (3.4). Obviously the components of $\phi_{i,j}^*$ belong to $VP_{4,k_i^u, k_{i+1}^u} \otimes VP_{4,k_j^v, k_{j+1}^v}$, and therefore we have (the matrix products apply componentwise)

$$\begin{aligned} \mathbf{f}_{i,j}^* - \phi_{i,j}^* &= B^{(\bar{k}^u)} A^{(\bar{k}^u)} \boldsymbol{\beta}^{uv}(\mathbf{f}_{i,j}^*) A^{(\bar{k}^v)T} B^{(\bar{k}^v)T} - B^{(\bar{k}^u)} A^{(\bar{k}^u)} \boldsymbol{\beta}^{uv}(\phi_{i,j}^*) A^{(\bar{k}^v)T} B^{(\bar{k}^v)T} \\ &= B^{(\bar{k}^u)} A^{(\bar{k}^u)} (\boldsymbol{\beta}^{uv}(\mathbf{f}_{i,j}^*) - \boldsymbol{\beta}^{uv}(\phi_{i,j}^*)) A^{(\bar{k}^v)T} B^{(\bar{k}^v)T}, \end{aligned}$$

where $\boldsymbol{\beta}^{uv}(\mathbf{f}_{i,j}^*)$ and $\boldsymbol{\beta}^{uv}(\phi_{i,j}^*)$ denote the matrices of control points. By the property of convex combination and of Bernstein operators we have, for any (u, v) ,

$$\|\mathbf{f}_{i,j}^*(u, v) - \phi_{i,j}^*(u, v)\| \leq \max_{\mu, \nu=0, \dots, 4} \|\boldsymbol{\beta}_{\mu,\nu}^{uv}(\mathbf{f}_{i,j}^*) - \boldsymbol{\beta}_{\mu,\nu}^{uv}(\phi_{i,j}^*)\| \leq \delta_{i,j}$$

which proves the claim. \square

For any $(u, v) \in \mathcal{D}_{i,j}$, we take the ϵ -balls $\Theta_\epsilon(\mathbf{f}_{i,j}^*(u, v))$ and $\Theta_\epsilon(\phi_{i,j}^*(u, v))$ and define the set

$$\begin{aligned} \overline{\mathcal{D}}_{i,j}(\epsilon; u, v) &:= \{(t, r) \in \mathcal{D}_{i,j} \text{ s.t. } \mathbf{f}_{i,j}^*(t, r) \in \Theta_\epsilon(\mathbf{f}_{i,j}^*(u, v)) \cap \Theta_\epsilon(\phi_{i,j}^*(u, v))\} \\ &\cap \{(t, r) \in \mathcal{D}_{i,j} \text{ s.t. } \phi_{i,j}^*(t, r) \in \Theta_\epsilon(\mathbf{f}_{i,j}^*(u, v)) \cap \Theta_\epsilon(\phi_{i,j}^*(u, v))\} \end{aligned}$$

and the *modified barycenters*

$$\begin{aligned} \overline{\mathbf{M}}_\epsilon^0(\mathbf{f}_{i,j}^*(u, v)) &:= \frac{1}{|\mathbf{f}_{i,j}^*(\overline{\mathcal{D}}_{i,j}(\epsilon; u, v))|} \int_{\overline{\mathcal{D}}_{i,j}(\epsilon; u, v)} \mathbf{f}_{i,j}^* dA, \\ \overline{\mathbf{M}}_\epsilon^0(\phi_{i,j}^*(u, v)) &:= \frac{1}{|\phi_{i,j}^*(\overline{\mathcal{D}}_{i,j}(\epsilon; u, v))|} \int_{\overline{\mathcal{D}}_{i,j}(\epsilon; u, v)} \phi_{i,j}^* dA. \end{aligned}$$

We make the assumption that the subsets $\overline{\mathcal{D}}_{i,j}(\epsilon; u, v)$ are nonempty and defer to Remark 4 its justification.

We have the following result.

LEMMA 2. Let $\delta_{i,j} > 0$ and let $\|\beta_{\mu,\nu}^{uv}(\mathbf{f}_{i,j}^*) - \beta_{\mu,\nu}^{uv}(\phi_{i,j}^*)\| \leq \delta_{i,j}$; $\mu, \nu = 0, \dots, 4$. Then, for any $(u, v) \in \mathcal{D}_{i,j}$

$$\left\| \overline{M}_\epsilon^0(\mathbf{f}_{i,j}^*(u, v)) - \overline{M}_\epsilon^0(\phi_{i,j}^*(u, v)) \right\| \leq \delta_{i,j}.$$

Proof. Let $q \in \mathbb{N}$, $q \geq k_i^u, k_{i+1}^u, k_j^v$, and k_{j+1}^v , and let $A^{(q)}$ be the degree elevation convex combination operator (3.4). Let $\lambda_{q,q}^{uv}(\mathbf{f}_{i,j}^*)$ and $\lambda_{q,q}^{uv}(\phi_{i,j}^*)$ be, respectively, the q -degree control nets of $\mathbf{f}_{i,j}^*$ and $\phi_{i,j}^*$. We recall that $\lambda_{q,q}^{uv}(\mathbf{f}_{i,j}^*)$ and $\lambda_{q,q}^{uv}(\phi_{i,j}^*)$ are continuous piecewise bilinear functions with knots at $\bar{u}_{i,\alpha} := u_i + \alpha(u_{i+1} - u_i)/q$, $\alpha = 0, \dots, q$, and $\bar{v}_{j,\gamma} := v_j + \gamma(v_{j+1} - v_j)/q$, $\gamma = 0, \dots, q$, and that

$$\lambda_{q,q}^{uv}(\mathbf{f}_{i,j}^*; \bar{u}_{i,\alpha}, \bar{v}_{j,\gamma}) = A^{(q)} \beta^{uv}(\mathbf{f}_{i,j}^*) A^{(q)T},$$

$$\lambda_{q,q}^{uv}(\phi_{i,j}^*; \bar{u}_{i,\alpha}, \bar{v}_{j,\gamma}) = A^{(q)} \beta^{uv}(\phi_{i,j}^*) A^{(q)T},$$

where $\beta^{uv}(\mathbf{f}_{i,j}^*)$ and $\beta^{uv}(\phi_{i,j}^*)$ denote the matrices of control points (the matrix products apply componentwise). Clearly we have

$$(3.14) \quad \left\| \lambda_{q,q}^{uv}(\mathbf{f}_{i,j}^*; \bar{u}_{i,\alpha}, \bar{v}_{j,\gamma}) - \lambda_{q,q}^{uv}(\phi_{i,j}^*; \bar{u}_{i,\alpha}, \bar{v}_{j,\gamma}) \right\| \leq \delta_{i,j}.$$

We also recall that

$$(3.15) \quad \lim_{q \rightarrow \infty} \lambda_{q,q}^{uv}(\mathbf{f}_{i,j}^*; u, v) = \mathbf{f}_{i,j}^*(u, v), \quad \lim_{q \rightarrow \infty} \lambda_{q,q}^{uv}(\phi_{i,j}^*; u, v) = \phi_{i,j}^*(u, v).$$

Now we set $R_{\alpha,\gamma} = [\bar{u}_{i,\alpha}, \bar{u}_{i,\alpha+1}] \times [\bar{v}_{j,\gamma}, \bar{v}_{j,\gamma+1}]$, which we use to define

$$\overline{\Delta}_{i,j}^{(q)}(\epsilon; u, v) := \left\{ \bigcup_{(\alpha,\gamma)} R_{\alpha,\gamma} \quad \text{s.t.} \quad \alpha, \gamma = 0, \dots, q-1; R_{\alpha,\gamma} \subset \overline{\mathcal{D}}_{i,j}(\epsilon; u, v) \right\}.$$

Note that $\overline{\Delta}_{i,j}^{(q)}(\epsilon; u, v) \subset \overline{\mathcal{D}}_{i,j}(\epsilon; u, v)$ and

$$(3.16) \quad \lim_{q \rightarrow \infty} \text{dist} \left(\overline{\Delta}_{i,j}^{(q)}(\epsilon; u, v), \overline{\mathcal{D}}_{i,j}(\epsilon; u, v) \right) = 0.$$

We observe that the values

$$\begin{aligned} \overline{M}_\epsilon^0(\lambda_{q,q}^{uv}(\mathbf{f}_{i,j}^*; u, v)) &:= \frac{1}{|\lambda_{q,q}^{uv}(\mathbf{f}_{i,j}^*; \overline{\Delta}_{i,j}^{(q)}(\epsilon; u, v))|} \int_{\overline{\Delta}_{i,j}^{(q)}(\epsilon; u, v)} \lambda_{q,q}^{uv}(\mathbf{f}_{i,j}^*) dA, \\ \overline{M}_\epsilon^0(\lambda_{q,q}^{uv}(\phi_{i,j}^*; u, v)) &:= \frac{1}{|\lambda_{q,q}^{uv}(\phi_{i,j}^*; \overline{\Delta}_{i,j}^{(q)}(\epsilon; u, v))|} \int_{\overline{\Delta}_{i,j}^{(q)}(\epsilon; u, v)} \lambda_{q,q}^{uv}(\phi_{i,j}^*) dA, \end{aligned}$$

can be computed, respectively, as an average of the points

$$\lambda_{q,q}^{uv}(\mathbf{f}_{i,j}^*; \bar{u}_{i,\alpha}, \bar{v}_{j,\gamma}), \lambda_{q,q}^{uv}(\phi_{i,j}^*; \bar{u}_{i,\alpha}, \bar{v}_{j,\gamma}); (\bar{u}_{i,\alpha}, \bar{v}_{j,\gamma}) \in \overline{\Delta}_{i,j}^{(q)}(\epsilon; u, v);$$

therefore, from (3.14) we have

$$\left\| \overline{M}_\epsilon^0(\lambda_{q,q}^{uv}(\mathbf{f}_{i,j}^*; u, v)) - \overline{M}_\epsilon^0(\lambda_{q,q}^{uv}(\phi_{i,j}^*; u, v)) \right\| \leq \delta_{i,j},$$

and the claim follows immediately from (3.15) and (3.16). \square

Let us define the *modified ϵ -normals*

$$(3.17) \quad \begin{aligned} \bar{\mathbf{n}}_\epsilon(u, v; \mathbf{f}_{i,j}^*) &:= \overline{\mathbf{M}}_\epsilon^0(\mathbf{f}_{i,j}^*(u, v)) - \mathbf{f}_{i,j}^*(u, v), \\ \bar{\mathbf{n}}_\epsilon(u, v; \phi_{i,j}^*) &:= \overline{\mathbf{M}}_\epsilon^0(\phi_{i,j}^*(u, v)) - \phi_{i,j}^*(u, v). \end{aligned}$$

From Lemmas 1 and 2 we have immediately the following results.

THEOREM 6. *Let $\delta_{i,j} > 0$ and let*

$$\|\beta_{\mu,\nu}^{uv}(\mathbf{f}_{i,j}^*) - \beta_{\mu,\nu}^{uv}(\phi_{i,j}^*)\|_\infty \leq \delta_{i,j}; \quad \mu, \nu = 0, \dots, 4.$$

Then, for any $(u, v) \in \mathcal{D}_{i,j}$

$$\|\bar{\mathbf{n}}_\epsilon(u, v; \mathbf{f}_{i,j}^*) - \bar{\mathbf{n}}_\epsilon(u, v; \phi_{i,j}^*)\|_\infty \leq 2\delta_{i,j}.$$

COROLLARY 2. *Let the values $\delta_{i,j} > 0$, $i = 0, \dots, m$, $j = 0, \dots, n$ be given. Then there exist threshold sequences $\tilde{\mathcal{K}}^u$ and $\tilde{\mathcal{K}}^v$ such that for any $\mathcal{K}^u \geq \tilde{\mathcal{K}}^u$, $\mathcal{K}^v \geq \tilde{\mathcal{K}}^v$ (the inequalities hold componentwise)*

$$(3.18) \quad \|\beta_{\mu,\nu}^{uv}(\mathbf{f}_{i,j}^*) - \beta_{\mu,\nu}^{uv}(\phi_{i,j}^*)\|_\infty \leq \delta_{i,j}; \quad \mu, \nu = 0, \dots, 4,$$

so that, for $(u, v) \in \mathcal{D}_{i,j}$,

$$\|\bar{\mathbf{n}}_\epsilon(u, v; \mathbf{f}_{i,j}^*) - \bar{\mathbf{n}}_\epsilon(u, v; \phi_{i,j}^*)\| \leq 2\delta_{i,j}.$$

We anticipate that in the schemes presented in the next section ϵ , which defines the radius of the ϵ -balls and the ϵ -normals, is a given input parameter and it is not adaptively changed during the execution of the algorithms. On the contrary, the algorithms modify the degree sequences, in order to fulfill the shape-preserving criteria.

Remark 4. From Theorem 4 and Lemma 1 we have that, for suitable sequences of the tolerances $\delta_{i,j}$ or, equivalently, for suitable sequences of the degrees, the center of the ϵ -balls $\Theta_\epsilon(\mathbf{f}_{i,j}^*(u, v))$ and $\Theta_\epsilon(\phi_{i,j}^*(u, v))$ are close enough to guarantee that $\overline{\mathcal{D}}_{i,j}(\epsilon; u, v) \neq \emptyset$.

Remark 5. Let

$$\mathcal{I}_{i,j}(\epsilon, \mathbf{f}_{i,j}^*; u, v) = \{(t, r) \in \mathcal{D}_{i,j} \text{ s.t. } \mathbf{f}_{i,j}^*(t, r) \in \Theta_\epsilon(\mathbf{f}_{i,j}^*(u, v))\},$$

$$\mathcal{I}_{i,j}(\epsilon, \phi_{i,j}^*; u, v) = \{(t, r) \in \mathcal{D}_{i,j} \text{ s.t. } \phi_{i,j}^*(t, r) \in \Theta_\epsilon(\phi_{i,j}^*(u, v))\}.$$

From Theorem 4 and Lemma 1 we have that, if $k_i^u, k_j^v \rightarrow \infty$, all i, j , then

$$\max \{ \text{dist}(\mathcal{I}_{i,j}(\epsilon, \mathbf{f}_{i,j}^*; u, v), \overline{\mathcal{D}}_{i,j}(\epsilon; u, v)), \text{dist}(\mathcal{I}_{i,j}(\epsilon, \phi_{i,j}^*; u, v), \overline{\mathcal{D}}_{i,j}(\epsilon; u, v)) \} \rightarrow 0.$$

Therefore, for the typical piecewise regular surfaces used in data approximation, (3.17) furnishes a good approximation of the corresponding ϵ -normals $\bar{\mathbf{n}}_\epsilon(u, v; \mathbf{f}_{i,j}^*)$ and $\bar{\mathbf{n}}_\epsilon(u, v; \phi_{i,j}^*)$ given in (2.4).

Remark 6. The implementation of (3.18) is much cheaper than (3.13) since it requires one only to check the distances of the Bézier control points of \mathbf{s}^* and σ^* .

Remark 7. In view of (2.5) we require that $\delta_{i,j} = o(\epsilon^2)$.

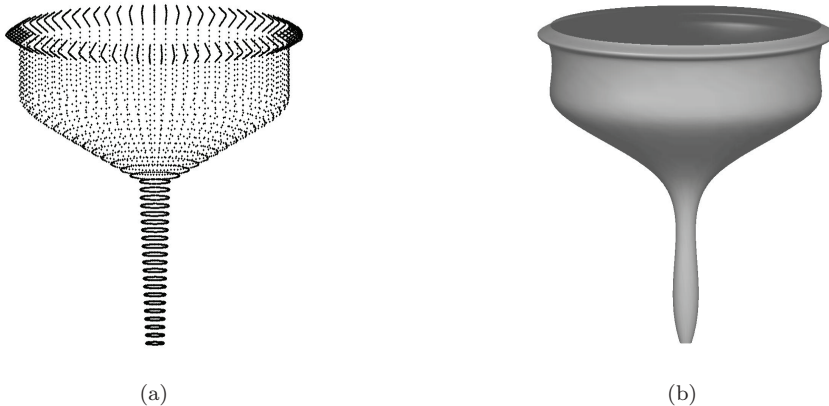


FIG. 4. *Example 1: (a) Data points. (b) Unconstrained approximation.*

3.5. The algorithms. The aim of this subsection is to propose two schemes for the effective construction of an approximating surface. We remark that simpler, unconstrained approximation methods are often inadequate for a good reproduction of the underlying object. Consider, for instance, the funnel data of Figure 4(a); the surface depicted in Figure 4(b) obtained using an unconstrained least square approximation in the space $\mathcal{S}_4^2 \otimes \mathcal{S}_4^2$ does not represent the main characteristics of the object.

The reader should compare this plot with Figure 5(d).

Global Algorithm

Let the data (2.1) or (2.2) and the real, positive numbers $\epsilon, \delta_{i,j}$ be given.

1. Compute the knot sequences \mathcal{U} and \mathcal{V} using Algorithm 1 or Algorithm 2 of section 6.
2. Compute σ^* as specified in subsection 2.1.
3. Set $k_i^u = 4, i = 0, \dots, m; k_j^v = 4, j = 0, \dots, n$.
4. Compute s^* as described in subsection 3.4.
5. While (3.13) or (3.18) is not satisfied
 - 5.1. Set $k_i^u = k_i^u + 1, i = 0, \dots, m; k_j^v = k_j^v + 1, j = 0, \dots, n$.
 - 5.2. Compute s^* as described in subsection 3.4.

This global algorithm produces an approximating surface s^* which is also shape preserving in the sense that its ϵ -normals or modified ϵ -normals agree in each $\mathcal{D}_{i,j}$ with those of the reference surface σ^* up to $2\delta_{i,j}$.

The main feature of such a global approach is in the convergence properties. On the other hand, its main drawback relies intrinsically in its global nature; a single, cumbersome, patch can force all the degrees to reach large values, and these in turn force the surface to have everywhere a nonsmooth and unpleasant appearance.

Therefore a local scheme, in which the only degrees related to the cumbersome patches of the data are increased, would be in some cases preferable. Unfortunately, in this case we cannot prove results similar to those of Corollaries 1 and 2. The very reason of this nonconvergence is that, even if $f_{i,j}^* = s_{|\mathcal{D}_{i,j}}^*$ tends to locally have a piecewise biquadratic shape when $k_i^u, k_{i+1}^u, k_j^v, k_{j+1}^v \rightarrow \infty$, it does not tend to $\phi_{i,j}^* = \sigma_{|\mathcal{D}_{i,j}}^*$. In other words, s^* belongs to a TVS space which does no more tend to the space of $\mathcal{S}_{2,\mathcal{U}}^0 \otimes \mathcal{S}_{2,\mathcal{V}}^0$. Therefore the best approximation taken from such a TVS can be, in general, far from σ^* .

We intend to show how it is possible to overcome this difficulty and obtain a local scheme, changing the linear least squares problem into a *weighted* linear least

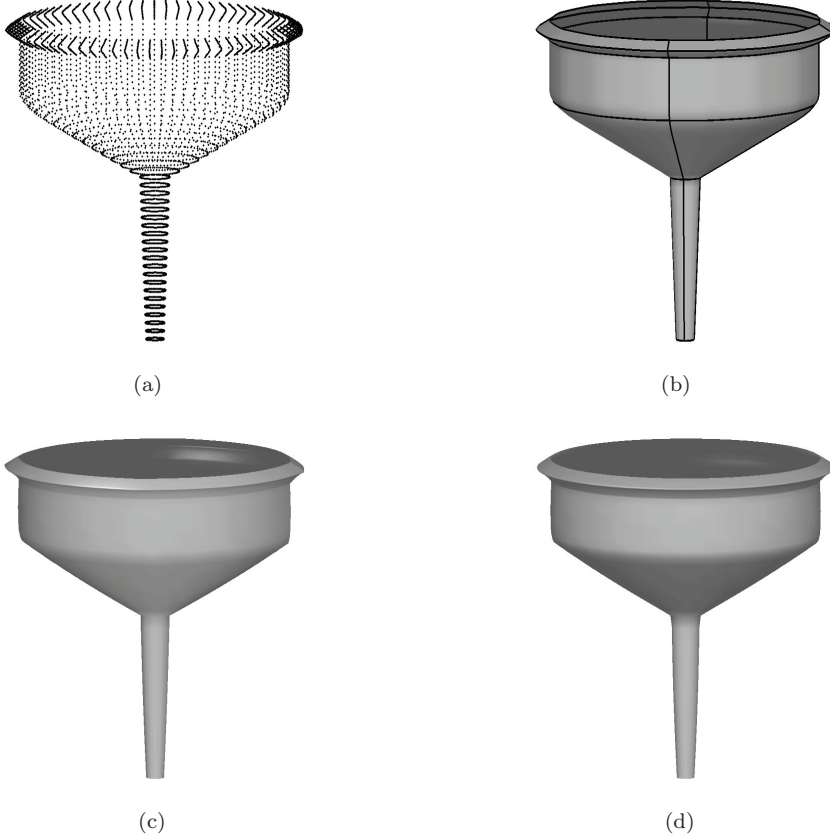


FIG. 5. *Example 1: (a) Data points. (b) Selected gridlines. (c) Reference surface σ^* . (d) Shape-preserving approximation in TVS_{co} .*

squares problem (using basically the same idea proposed in [4]). The key fact is that we accept a compromise, obtaining the convergence at the price of a reduction in the approximation power. In order to force the spline to locally mimic σ^* when a local increase is applied, we work with an extension of the approximation problem. Let

$$\{w_{\bar{\mu}, \bar{\nu}}; \bar{\mu} = 0, \dots, 2m, \bar{\nu} = 0, \dots, 2n\}$$

be a sequence of positive weights. The basic idea is to push s^* towards σ^* , by inserting

$$\sigma^*(\gamma_{\bar{\mu}}, \zeta_{\bar{\nu}}); \bar{\mu} = 0, \dots, 2m, \bar{\nu} = 0, \dots, 2n$$

as weighted points in the approximation problem, with

$$\gamma_{2i} = u_i, \quad i = 0, \dots, m; \quad \gamma_{2i+1} = (u_{i+1} + u_i)/2, \quad i = 0, \dots, m-1$$

and

$$\zeta_{2j} = v_j, \quad j = 0, \dots, n; \quad \zeta_{2j+1} = (v_{j+1} + v_j)/2, \quad j = 0, \dots, n-1.$$

The points we are going to approximate are

$$\mathcal{Q} = \{(t_\mu r_\nu, \mathbf{P}_{\mu, \nu})\} \cup \{(\gamma_{\bar{\mu}}, \zeta_{\bar{\nu}}, \sigma^*(\gamma_{\bar{\mu}}, \zeta_{\bar{\nu}}))\}$$

in case (2.1) or

$$\mathcal{Q} = \{(t_\mu, r_\mu, \mathbf{P}_\mu)\} \cup \{(\gamma_{\bar{\mu}}, \zeta_{\bar{\nu}}, \boldsymbol{\sigma}^*(\gamma_{\bar{\mu}}, \zeta_{\bar{\nu}}))\}$$

in case (2.2). We take a modified semi-norm, given by

$$(3.19) \quad \langle \mathbf{f} \rangle := \sqrt{\sum_{\mu=0}^M \sum_{\nu=0}^N \|\mathbf{f}(t_\mu, r_\nu)\|_2^2 + \sum_{\bar{\mu}=0}^{2m} \sum_{\bar{\nu}=0}^{2n} w_{\bar{\mu}, \bar{\nu}} \|\mathbf{f}(\gamma_{\bar{\mu}}, \zeta_{\bar{\nu}})\|_2^2}$$

or by

$$(3.20) \quad \langle \mathbf{f} \rangle := \sqrt{\sum_{\mu=0}^L \|\mathbf{f}(t_\mu, r_\mu)\|_2^2 + \sum_{\bar{\mu}=0}^{2m} \sum_{\bar{\nu}=0}^{2n} w_{\bar{\mu}, \bar{\nu}} \|\mathbf{f}(\gamma_{\bar{\mu}}, \zeta_{\bar{\nu}})\|_2^2}.$$

We assume that the following property holds.

ASSUMPTION 2. *The semi-norm (3.19) or (3.20) is a norm for all the (finite dimensional) tensor-product spline spaces.*

We find $\mathbf{s}^* \in \mathbf{TVS}$, the best weighted least squares approximation:

$$\langle \mathbf{s}^* - \mathcal{Q} \rangle \leq \langle \mathbf{s} - \mathcal{Q} \rangle, \quad \forall \mathbf{s} \in \mathbf{TVS}.$$

Note that if $w_{\bar{\mu}, \bar{\nu}} = 0$, all $\bar{\mu}, \bar{\nu}$ we have the old approximation problem, and, for given $\bar{\mu}, \bar{\nu}$,

$$(3.21) \quad \lim_{w_{\bar{\mu}, \bar{\nu}} \rightarrow \infty} \mathbf{s}^*(\gamma_{\bar{\mu}}, \zeta_{\bar{\nu}}) = \boldsymbol{\sigma}^*(\gamma_{\bar{\mu}}, \zeta_{\bar{\nu}}).$$

We can use both the weights and the degrees to control the behavior of the surface in $\mathcal{D}_{i,j}$. Let

$$MN_{(i,j)} = \{(\bar{\mu}, \bar{\nu}) : \bar{\mu} = 2i, 2i+1, 2i+2, \bar{\nu} = 2j, 2j+1, 2j+2\}.$$

We have the following result, whose proof is a trivial consequence of Theorem 3 and (3.21).

THEOREM 7. *Let*

$$w_{\bar{\mu}, \bar{\nu}}, k_p^u, k_q^v \rightarrow \infty \quad \text{for } (\bar{\mu}, \bar{\nu}) \in MN_{(i,j)}, \quad p = i, i+1, \quad q = j, j+1,$$

and then

$$\mathbf{s}^*(u, v) \rightarrow \boldsymbol{\sigma}^*(u, v), \quad \forall (u, v) \in \mathcal{D}_{i,j}.$$

Now, we also have the following local versions of Corollaries 1 and 2.

COROLLARY 3. *Let $\delta_{i,j} > 0$. There exist threshold values $\tilde{w}_{\bar{\mu}, \bar{\nu}}$, \tilde{k}_p^u , and \tilde{k}_q^v with $(\bar{\mu}, \bar{\nu}) \in MN_{(i,j)}$, $p = i, i+1$, $q = j$, and $j+1$ such that for any $w_{\bar{\mu}, \bar{\nu}} \geq \tilde{w}_{\bar{\mu}, \bar{\nu}}$, $k_p^u \geq \tilde{k}_p^u$,*

and $k_q^v \geq \tilde{k}_q^v$, and for any $(u, v) \in \mathcal{D}_{i,j}$

$$(3.22) \quad \|\mathbf{n}_\epsilon(u, v; \mathbf{s}^*) - \mathbf{n}_\epsilon(u, v; \boldsymbol{\sigma}^*)\| \leq 2\delta_{i,j}.$$

COROLLARY 4. Let $\delta_{i,j} > 0$ be given. There exist threshold values $\tilde{w}_{\bar{\mu}, \bar{\nu}}$, \tilde{k}_p^u , and \tilde{k}_q^v with $(\bar{\mu}, \bar{\nu}) \in MN_{(i,j)}$, $p = i, i+1$, and $q = j, j+1$ such that for any $w_{\bar{\mu}, \bar{\nu}} \geq \tilde{w}_{\bar{\mu}, \bar{\nu}}$, $k_p^u \geq \tilde{k}_p^u$, and $k_q^v \geq \tilde{k}_q^v$

$$(3.23) \quad \|\beta_{\mu,\nu}^{uv}(\mathbf{f}_{i,j}^*) - \beta_{\mu,\nu}^{uv}(\boldsymbol{\phi}_{i,j}^*)\|_\infty \leq \delta_{i,j}; \quad \mu, \nu = 0, \dots, 4,$$

so that, for $(u, v) \in \mathcal{D}_{i,j}$,

$$\|\bar{\mathbf{n}}_\epsilon(u, v; \mathbf{f}_{i,j}^*) - \bar{\mathbf{n}}_\epsilon(u, v; \boldsymbol{\phi}_{i,j}^*)\| \leq 2\delta_{i,j}.$$

Clearly, the larger the weights, the more the approximation to the *true data* \mathcal{P} deteriorates, and we would like to keep the weights as small as possible while maintaining the “pleasantness” of the surface. Note that if we increase two consecutive degrees k_i^u, k_{i+1}^u (k_j^v, k_{j+1}^v) the effect spreads over the strip corresponding to $u_i \leq u \leq u_{i+1}$ ($v_j \leq v \leq v_{j+1}$), and if we increase the weights $w_{\mu,\nu}$ with $(\mu, \nu) \in MN^{(i,j)}$, the effect is local to $\mathcal{D}_{i,j}$.

A possible, simple strategy is described in the following algorithm.

Local Algorithm

Let the data (2.1) or (2.2) and the real, positive numbers ϵ , $\delta_{i,j}$, Δ_w be given.

1. Compute the knot sequences \mathcal{U} and \mathcal{V} using Algorithm 1 or Algorithm 2 of section 6.
2. Compute $\boldsymbol{\sigma}^*$ as specified in subsection 2.1.
3. Set $k_i^u = 4$, $i = 0, \dots, m$; $k_j^v = 4$, $j = 0, \dots, n$;
 $w_{\bar{\mu}, \bar{\nu}} = 0$, $\bar{\mu} = 0, \dots, 2m$, $\bar{\nu} = 0, \dots, 2n$.
4. Compute \mathbf{s}^* as described in subsection 3.4.
5. While (3.22) or (3.23) is not satisfied
 - 5.1. For all i, j compute
$$ERR_{i,j} = \max \left\{ \|\beta_{\alpha,\gamma}^{uv}(\mathbf{f}_{i,j}^*) - \beta_{\alpha,\gamma}^{uv}(\boldsymbol{\phi}_{i,j}^*)\|_\infty; \alpha, \gamma = 0, \dots, 4 \right\}.$$
 - 5.2. Find I, J such that $ERR_{I,J} \geq ERR_{i,j}$, all i, j .
 - 5.3. For $p = I, I+1$, $q = J, J+1$, $(\bar{\mu}, \bar{\nu}) \in MN_{(i,j)}$ set
 $k_p^u = k_p^u + 1$; $k_q^v = k_q^v + 1$; $w_{\bar{\mu}, \bar{\nu}} = w_{\bar{\mu}, \bar{\nu}} + \Delta_w$
 - 5.4. Compute \mathbf{s}^* as described in subsection 3.4.

Obviously, steps 5.2 and 5.3 could be modified to treat simultaneously more sub-rectangles (for instance, all $\mathcal{D}_{i,j}$ whose errors exceed a threshold value), thus reducing the computational cost required by instruction 5.4.

4. Graphical examples. In this section we want to discuss the practical applications of the theory developed in the previous sections with the aid of some graphical tests. As previously said in the introduction, our method is mainly conceived for applications in reverse engineering, where the surface is typically represented by well-identifiable patches with a clear geometric shape. Figures 5 and 6 are representative samples from this field. We have also added other additional examples using geological (Figure 7) and medical (Figure 8) data, which, being taken beyond the main field of applications,⁴ provide severe tests for our method. Indeed, if compared

⁴Geological data are often used as typical examples of fractal phenomena, in which a smooth representation could lead to erroneous interpretations.

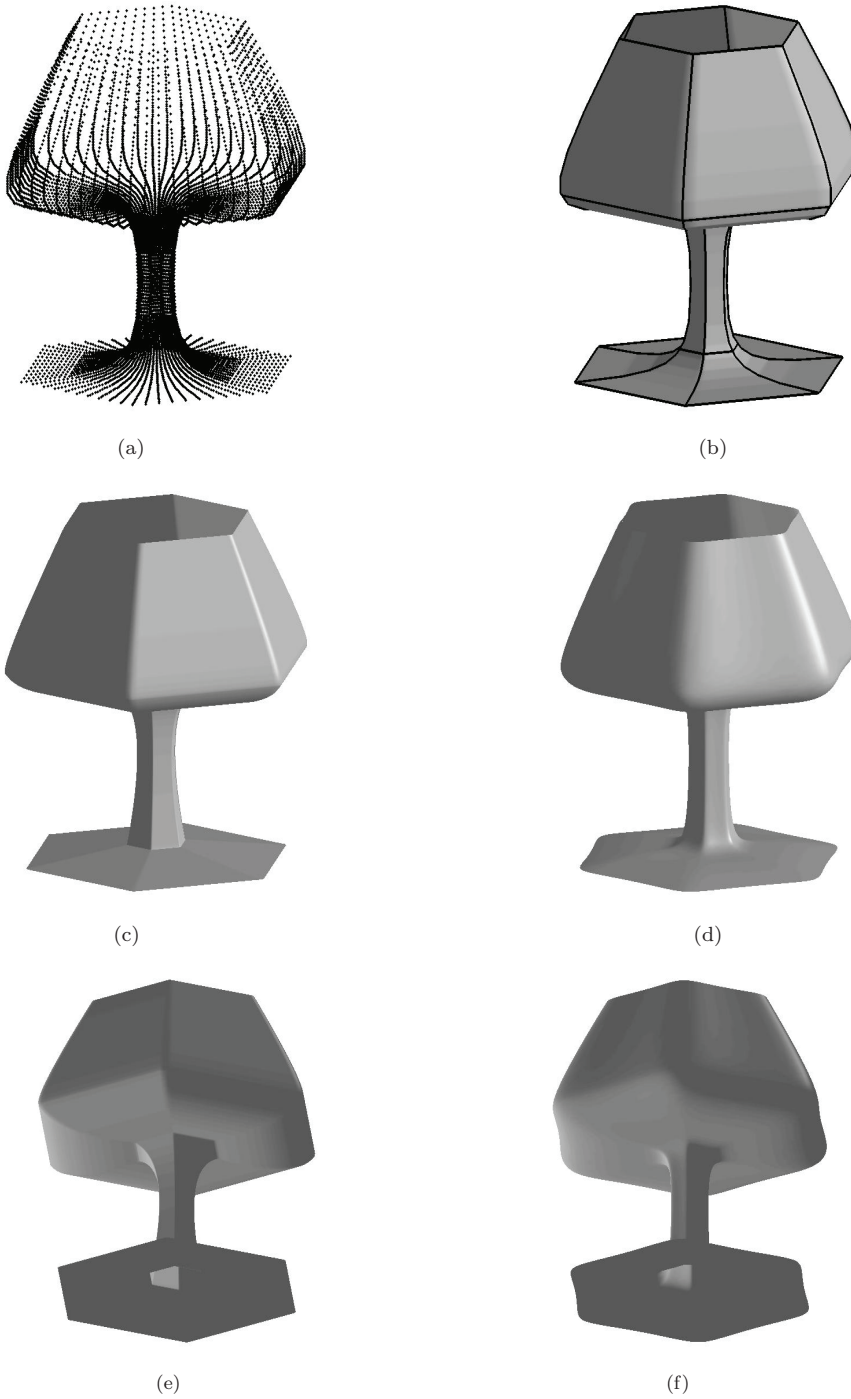


FIG. 6. Example 2: (a) Data points. (b) Selected gridlines. (c) Reference surface σ^* . (d) Shape-preserving approximation in TVS_{co} . (e) Rotated biquadratic surface σ^* . (f) Rotated shape-preserving approximation s^* .

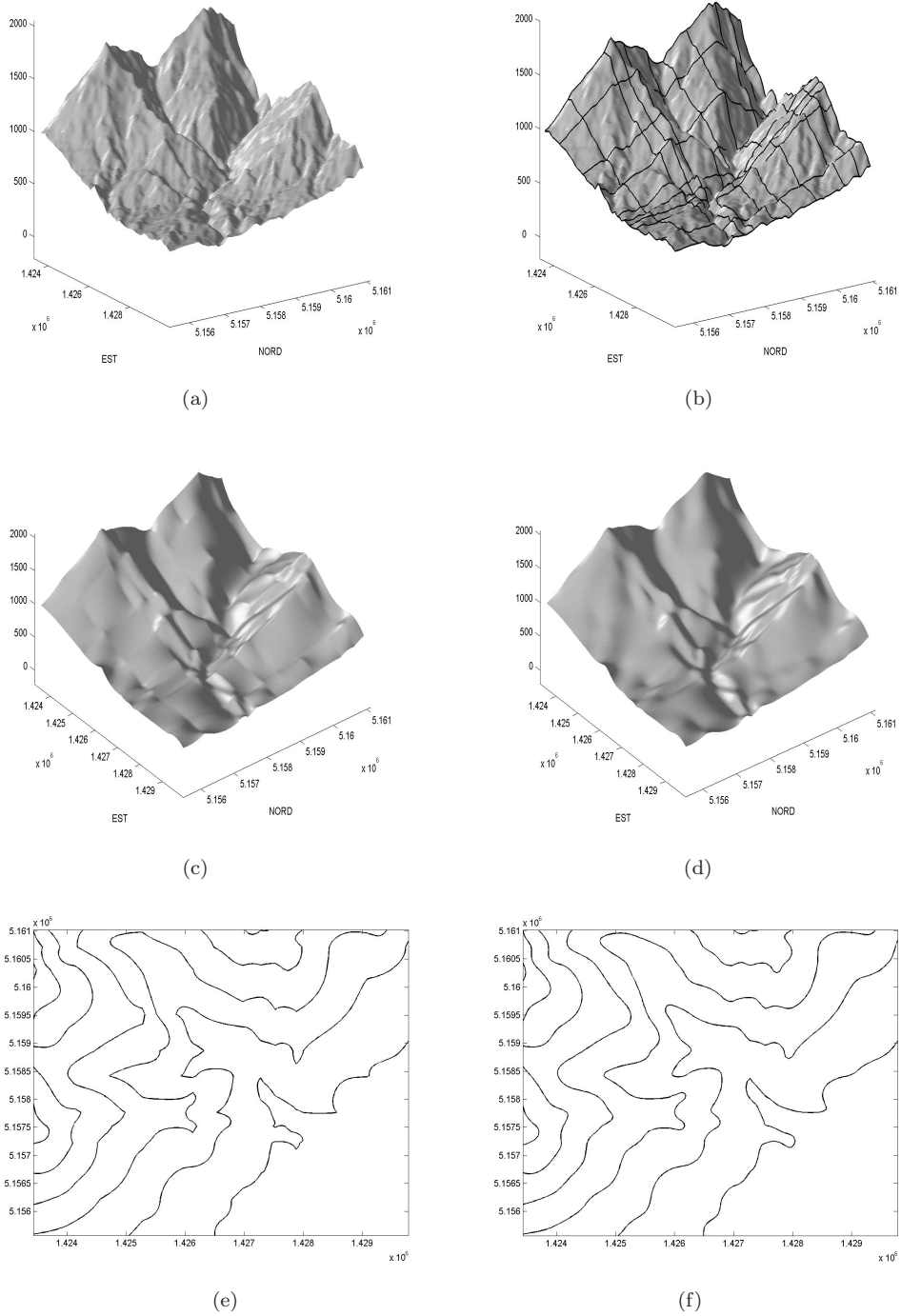


FIG. 7. Example 3: (a) Data surface. (b) Selected gridlines. (c) Reference surface σ^* . (d) Shape-preserving approximation $s^* \in TVSo$. (e) Contour plot of σ^* . (f) Contour plot of s^* .

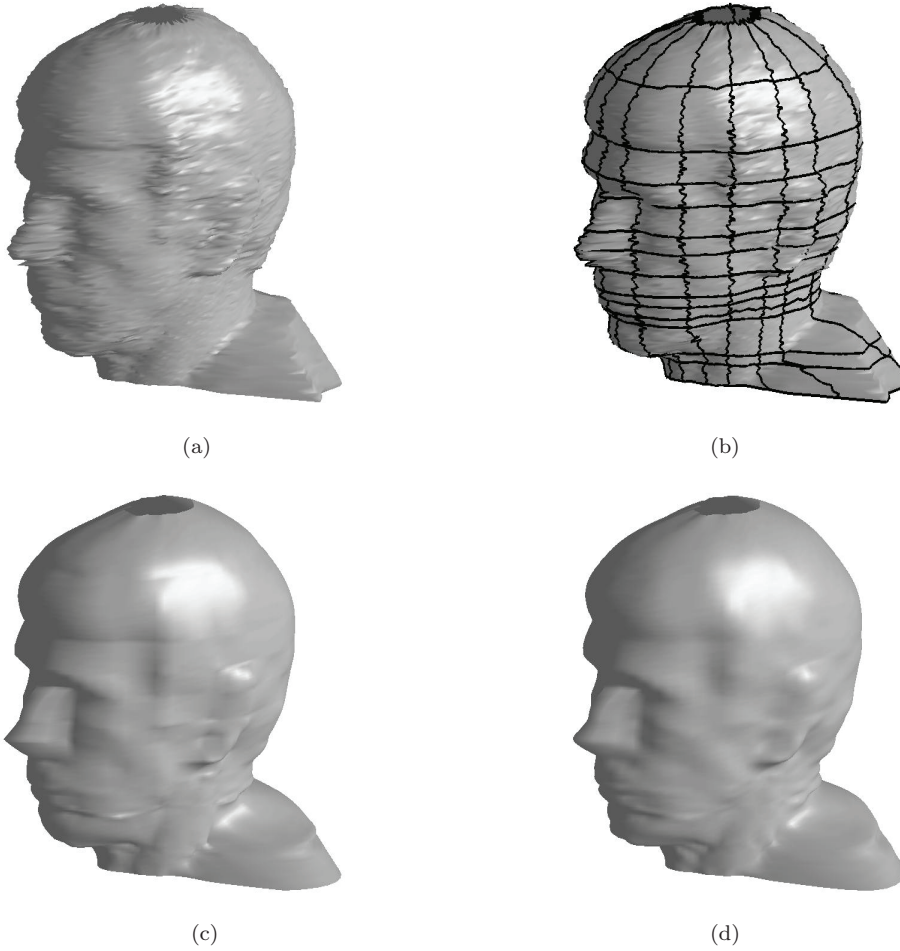


FIG. 8. *Example 4: (a) Data surface. (b) The selected gridlines. (c) Reference surface σ^* . (d) Shape-preserving approximation in $TV S_{co}$.*

with other pictures, Figures 7 and 8 are aesthetically poor. However, such kind of data are typically represented using a huge amount of (triangular) surface patches; here we intend to show that the shape of cumbersome data can be efficiently analyzed with the zero moment approach and that appreciably results in the representation of complex objects can be obtained with a single mathematical model composed of a (relatively) limited number of patches.

We conclude this section with a comparison of the zero-moments with the mean and Gaussian curvature. In Figure 9 the surface approximating the goblet data of Figure 6 has been enriched with some color maps. The first two rows show the normalized mean and Gaussian curvature. The last row shows the length of the normalized ϵ -normals, where the sign depends on the position with respect to the tangent plane. Figure 10 shows the normalized mean curvature, Gaussian curvature, and ϵ -normals as bivariate functions in the parameter domain. From this and similar examples we infer that, despite the inherent differences of the plots, the locations of the significant changes substantially agree. Therefore, the ϵ -normals seem to provide an effective tool both to analyze the data and to set up the shape constraints.

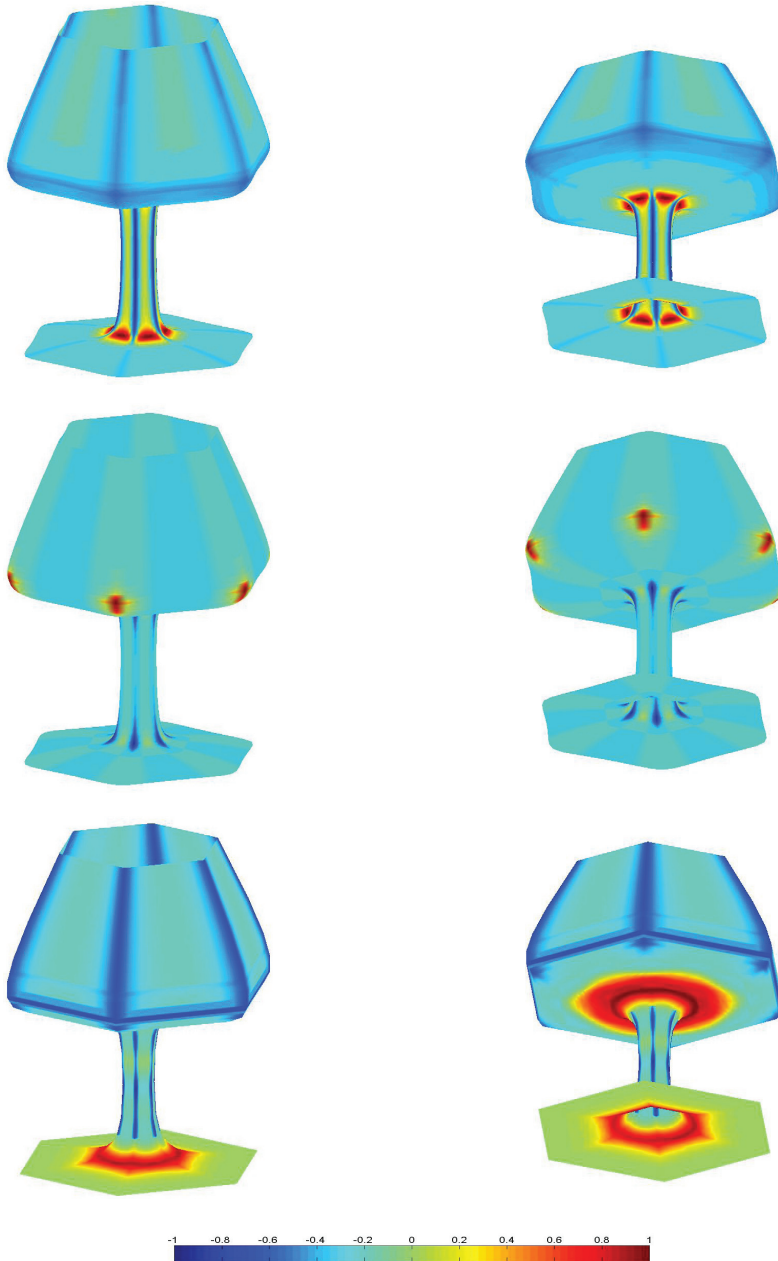


FIG. 9. Shape analysis of the “goblet” surface. Top: normalized mean curvature. Middle: normalized Gaussian curvature. Bottom: normalized length of ϵ -normals.

5. Conclusions. We have presented a new method for the construction of tensor-product parametric surfaces. Such a method, which, as far as we know, is the first one dealing with shape-preserving approximation of spatial data, is composed of three principal steps: a scheme for detecting the shape of the data, the application of a new class of tensor-product spline, and a suitable linear least-squares strategy, inserted in a global or in a local algorithm. Its main advantage relies in the possibility of

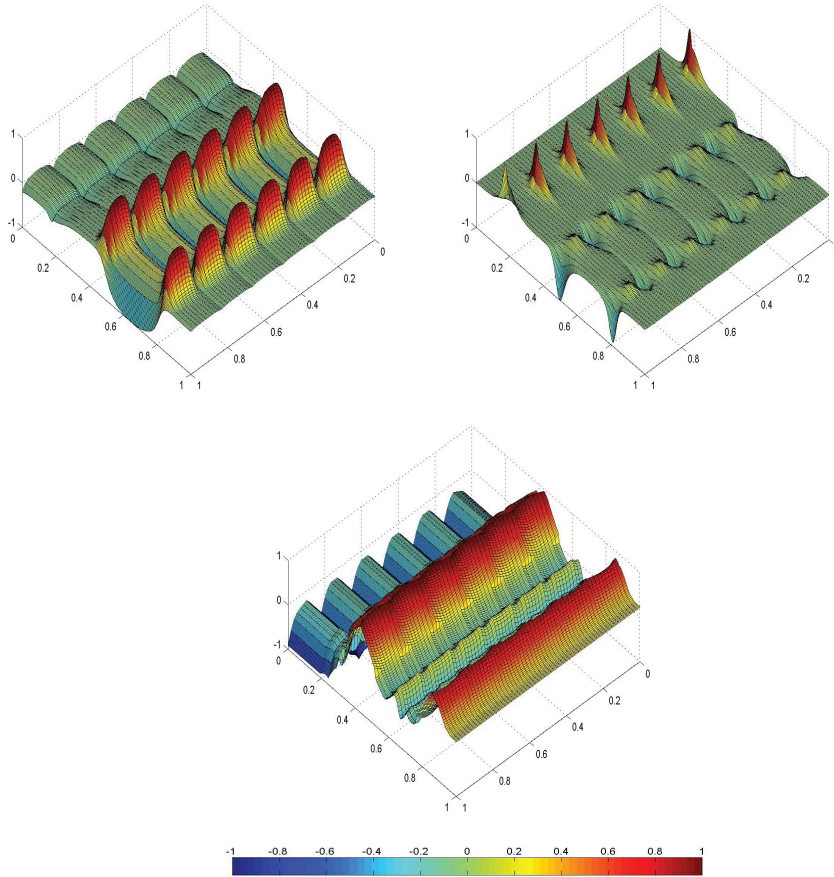


FIG. 10. Shape analysis of the “goblet surface”. In clockwise sense: normalized mean curvature, Gaussian curvature, and length of ϵ -normals represented as functions in the parameter domain.

using the same mathematical model (our variable degree polynomial spline spaces) for representing objects composed of patches with different shapes, and by sharp or smooth edges. The practical results, reported in the previous section, indicate that appreciable results are obtained at the price of limited computational expenses.

It is important to point out that in the tensor-product approximation all the points are uniformly treated. However, in practical applications it is common that some feature curves both summarize the shape of real objects and are crucial for understanding their physical properties. In other words, there are privileged subsets of the data points that form a network of fundamental grid lines and must be reproduced with much more accuracy than the remaining ones. In [6] is proposed a specific method based on surfaces defined via the Boolean sum of variable degree spline operators.

Several practical experiments, similar to those reported in Figures 7 and 8, suggest the opportunity of using spline surfaces on triangulations. This approach requires both the definition of new polynomials over triangular domains, approaching in the limit quadratic triangular elements, and the definition of the corresponding spline spaces. The zero-moment analysis must also be adapted to triangulations. These new results will be collected in a forthcoming paper.

It is worthwhile to point out that both in (2.1) or in (2.2) we have assumed to know the parameter values corresponding to the data points. It is well known that

the computation of the parameters, especially for arbitrary spatial data set, is a very difficult task. For more general applications, our algorithm should be integrated with some efficient tool, for instance, those proposed in [9], [10], and references therein.

We conclude this paper with some open questions.

The first one concerns a possible *adaptive strategy* for the zero-moment analysis. A crucial aspect is obviously in the size of ϵ : small values can detect useless details, and large values can ignore significant changes of the shape. Following the idea of adaptive quadrature rules, we could stop the process when we find no or few differences between the results of two zero-moment analysis obtained with, say, ϵ and $\epsilon/2$. Obviously, when necessary, this scheme could be separately applied and stopped to different subdivision levels, that is, to different subrectangles with decreasing area.

The second concerns the possible relations between ϵ -normals and the Bézier control points. It is well known that important consequences derive from the variation diminishing property, both for planar (see, e.g., [12]) and for spatial (see [11]) curves, and, in particular, that the geometric characteristics of the curve can be deduced from the discrete curvature and torsion of the control points. From a different point of view, we can say that there are some geometric properties which are invariant with respect to the refinement of the net⁵ and are thus inherited by the curve. Nothing similar occurs for parametric surfaces; the few results concerning the relations between the discrete geometry of the control points and the mean or Gaussian curvature of the surface are either very restrictive or very cumbersome (see, e.g., [8], [14], [16]). An investigation on the possible relations between the discrete ϵ -normals of the control points and the ϵ -normals of the final surface would therefore be very useful.

6. Appendix. Knot Selection: Algorithm 1.

Let the data (2.1) and the real positive numbers ϵ , tol_{peak} , and $\text{tol}_{\text{cc}} \in [0, 1]$ be given. The algorithm computes the knot sequences \mathcal{U} and \mathcal{V} .

1. Zero-moment computation. For each $(t_\mu, r_\nu, \mathbf{P}_{\mu,\nu}) \in \mathcal{P}$:

1.1 compute the barycenter $M_\epsilon^0(\mathbf{P}_{\mu,\nu})$ as

$$M_\epsilon^0(\mathbf{P}_{\mu,\nu}) := \frac{1}{\text{card}(\Theta_\epsilon(\mathbf{P}_{\mu,\nu}))} \sum_{\mathbf{P}_{q,l} \in \Theta_\epsilon(\mathbf{P}_{\mu,\nu})} \mathbf{P}_{q,l},$$

where $\Theta_\epsilon(\mathbf{P}_{\mu,\nu}) := \Theta_\epsilon(t_\mu, r_\nu, \mathbf{P}_{\mu,\nu})$ is the ball of radius ϵ centered at $(t_\mu, r_\nu, \mathbf{P}_{\mu,\nu})$;

1.2 compute the ϵ -normal, $\mathbf{n}_\epsilon(\mathbf{P}_{\mu,\nu}) := \mathbf{n}_\epsilon(t_\mu, r_\nu; \mathbf{P}_{\mu,\nu})$ as

$$\mathbf{n}_\epsilon(\mathbf{P}_{\mu,\nu}) := M_\epsilon^0(\mathbf{P}_{\mu,\nu}) - \mathbf{P}_{\mu,\nu}.$$

2. Peak selection

2.1 Find the points $(t_\mu, r_\nu, \mathbf{P}_{\mu,\nu}) \in \mathcal{P}$ such that

$$\|\mathbf{n}_\epsilon(\mathbf{P}_{\mu,\nu})\| = \max_{\mathbf{P}_{q,l} \in \Theta_{\epsilon/2}(\mathbf{P}_{\mu,\nu})} \{\|\mathbf{n}_\epsilon(\mathbf{P}_{q,l})\| : \|\mathbf{n}_\epsilon(\mathbf{P}_{q,l})\| > \epsilon^2 \text{tol}_{\text{peak}}\}.$$

2.2 $\mathcal{P}_{\text{peak}} \leftarrow \mathcal{P}_{\text{peak}} \cup \{(t_\mu, r_\nu, \mathbf{P}_{\mu,\nu})\}$.

⁵It is well known that the *degree elevation* or the (repeated) *de Casteljau* algorithms [12] produce finer and finer sets of control points converging to the Bézier curve.

3. Convexity change selection

3.1 For each $(t_\mu, r_\nu, \mathbf{P}_{\mu,\nu}) \in \mathcal{P}$ compute

$$\alpha_{\mu,\nu}^u := \frac{\mathbf{n}_\epsilon(\mathbf{P}_{\mu,\nu}) \cdot \mathbf{n}_\epsilon(\mathbf{P}_{\mu+1,\nu})}{\|\mathbf{n}_\epsilon(\mathbf{P}_{\mu,\nu})\| \|\mathbf{n}_\epsilon(\mathbf{P}_{\mu+1,\nu})\|}, \quad \alpha_{\mu,\nu}^v := \frac{\mathbf{n}_\epsilon(\mathbf{P}_{\mu,\nu}) \cdot \mathbf{n}_\epsilon(\mathbf{P}_{\mu,\nu+1})}{\|\mathbf{n}_\epsilon(\mathbf{P}_{\mu,\nu})\| \|\mathbf{n}_\epsilon(\mathbf{P}_{\mu,\nu+1})\|}.$$

3.2 Find the points $(t_\mu, r_\nu, \mathbf{P}_{\mu,\nu}) \in \mathcal{P}$ such that

$$\alpha_{\mu,\nu}^u = \min_{\mathbf{P}_{q,l} \in \Theta_{\epsilon/2}(\mathbf{P}_{\mu,\nu})} \{\alpha_{q,l}^u : \alpha_{q,l}^u < -\text{tol}_{\text{cc}}\}.$$

3.3 $\mathcal{P}_{\text{cc}}^u \leftarrow \mathcal{P}_{\text{cc}}^u \cup \{\mathbf{P}_{\mu,\nu}\}$.

3.4 Find the points $(t_\mu, r_\nu, \mathbf{P}_{\mu,\nu}) \in \mathcal{P}$ such that

$$\alpha_{\mu,\nu}^v = \min_{\mathbf{P}_{q,l} \in \Theta_{\epsilon/2}(\mathbf{P}_{\mu,\nu})} \{\alpha_{q,l}^v : \alpha_{q,l}^v < -\text{tol}_{\text{cc}}\}.$$

3.5 $\mathcal{P}_{\text{cc}}^v \leftarrow \mathcal{P}_{\text{cc}}^v \cup \{(t_\mu, r_\nu, \mathbf{P}_{\mu,\nu})\}$.

4. Grid-lines extraction

4.1 For each point $(t_\mu, r_\nu, \mathbf{P}_{\mu,\nu}) \in \mathcal{P}_{\text{peak}}$,

- $\mathcal{U} \leftarrow \mathcal{U} \cup \{t_\mu\}$,
- $\mathcal{V} \leftarrow \mathcal{V} \cup \{r_\nu\}$.

4.2 For each t_μ such that $(t_\mu, r_l, \mathbf{P}_{\mu,\nu}) \in \mathcal{P}_{\text{cc}}^u$, for some $l \in \{0, 1, \dots, M\}$

- compute the indicator $\mathcal{I}_\mu^{\text{cc}}$ as

$$\mathcal{I}_\mu^{\text{cc}} := \frac{\text{card}(\mathcal{C}_\mu^u)}{2(N+1)} + \frac{1}{2(N+1)} \sum_{l \in \mathcal{C}_\mu^u} |\alpha_{\mu,l}^u|,$$

where $\mathcal{C}_\mu^u = \{l : (t_\mu, r_l, \mathbf{P}_{\mu,l}) \in \mathcal{P}_{\text{cc}}^u\}$;

- if $\mathcal{I}_\mu^{\text{cc}} > \text{tol}_{\text{cc}}$, then $\mathcal{U} \leftarrow \mathcal{U} \cup \{t_\mu\}$.

4.3 For each r_ν such that $(t_q, r_\nu, \mathbf{P}_{\mu,\nu}) \in \mathcal{P}_{\text{cc}}^v$, for some $q \in \{0, 1, \dots, N\}$

- compute the indicator $\mathcal{J}_\nu^{\text{cc}}$ as

$$\mathcal{J}_\nu^{\text{cc}} := \frac{\text{card}(\mathcal{C}_\nu^v)}{2(M+1)} + \frac{1}{2(M+1)} \sum_{q \in \mathcal{C}_\nu^v} |\alpha_{q,\nu}^v|,$$

where $\mathcal{C}_\nu^v = \{q : (t_q, r_\nu, \mathbf{P}_{q,\nu}) \in \mathcal{P}_{\text{cc}}^v\}$;

- if $\mathcal{J}_\nu^{\text{cc}} > \text{tol}_{\text{cc}}$, then $\mathcal{V} \leftarrow \mathcal{V} \cup \{r_\nu\}$.

Knot Selection: Algorithm 2.

Let the data (2.2) and the real positive numbers ϵ , tol_{peak} , and $\text{tol}_{\text{cc}} \in [0, 1]$ be given.

The algorithm computes the knot sequences \mathcal{U} and \mathcal{V} .

1. **Zero-moment computation.** For each $(t_\mu, r_\mu, \mathbf{P}_\mu) \in \mathcal{P}$

1.1 Compute the barycenter $M_\epsilon^0(\mathbf{P}_\mu)$ as

$$M_\epsilon^0(\mathbf{P}_\mu) := \frac{1}{\text{card}(\Theta_\epsilon(\mathbf{P}_\mu))} \sum_{\mathbf{P}_q \in \Theta_\epsilon(\mathbf{P}_\mu)} \mathbf{P}_q,$$

where $\Theta_\epsilon(\mathbf{P}_\mu) := \Theta_\epsilon(t_\mu, r_\mu, \mathbf{P}_\mu)$ in the ball of radius ϵ centered at $(t_\mu, r_\mu, \mathbf{P}_\mu)$.

1.2 Compute the ϵ -normal, $\mathbf{n}_\epsilon(\mathbf{P}_\mu) := \mathbf{n}_\epsilon(t_\mu, r_\mu; \mathbf{P}_\mu)$ as

$$\mathbf{n}_\epsilon(\mathbf{P}_\mu) := M_\epsilon^0(\mathbf{P}_\mu) - \mathbf{P}_\mu.$$

2. Peak selection

2.1 Find the points $(t_\mu, r_\mu, \mathbf{P}_\mu) \in \mathcal{P}$ such that

$$\|\mathbf{n}_\epsilon(\mathbf{P}_\mu)\| = \max_{\mathbf{P}_q \in \Theta_{\epsilon/2}(\mathbf{P}_\mu)} \{ \|\mathbf{n}_\epsilon(\mathbf{P}_q)\| : \|\mathbf{n}_\epsilon(\mathbf{P}_q)\| > \epsilon^2 \text{tol}_{\text{peak}} \}.$$

2.2 $\mathcal{P}_{\text{peak}} \leftarrow \mathcal{P}_{\text{peak}} \cup \{(t_\mu, r_\mu, \mathbf{P}_\mu)\}$.

3. Convexity change selection

3.1 For each $(t_\mu, r_\mu, \mathbf{P}_\mu) \in \mathcal{P}$:

– find the set

$$\mathcal{C}_\mu := \left\{ q : \mathbf{P}_q \in \Theta_\epsilon(\mathbf{P}_\mu) \text{ and } \frac{\mathbf{n}_\epsilon(\mathbf{P}_\mu) \cdot \mathbf{n}_\epsilon(\mathbf{P}_q)}{\|\mathbf{n}_\epsilon(\mathbf{P}_\mu)\| \|\mathbf{n}_\epsilon(\mathbf{P}_q)\|} < -\text{tol}_{\text{cc}} \right\};$$

– compute the indicator of convexity change $\mathcal{F}_\mu^{\text{cc}}$ as

$$\mathcal{F}_\mu^{\text{cc}} = \frac{1}{2}(N_\mu + A_\mu),$$

where

$$N_\mu = \frac{\text{card}(\mathcal{C}_\mu)}{\text{card}(\Theta_\epsilon(\mathbf{P}_\mu))}, \quad A_\mu = \frac{\sum_{q: \mathbf{P}_q \in \mathcal{C}_\mu} (1 - d_q) \beta_q}{\sum_{q: \mathbf{P}_q \in \mathcal{C}_\mu} (1 - d_q)},$$

$$\beta_q = \frac{|\mathbf{n}_\epsilon(\mathbf{P}_\mu) \cdot \mathbf{n}_\epsilon(\mathbf{P}_q)|}{\|\mathbf{n}_\epsilon(\mathbf{P}_\mu)\| \|\mathbf{n}_\epsilon(\mathbf{P}_q)\|}, \quad d_q = \frac{\|\mathbf{P}_\mu - \mathbf{P}_q\|}{\epsilon}.$$

3.2 Find the points $(t_\mu, r_\mu, \mathbf{P}_\mu) \in \mathcal{P}$ such that

$$\mathcal{F}_\mu^{\text{cc}} = \max_{\mathbf{P}_q \in \Theta_{\epsilon/2}(\mathbf{P}_\mu)} \{ \mathcal{F}_q^{\text{cc}} : \mathcal{F}_q^{\text{cc}} > \text{tol}_{\text{cc}} \}.$$

3.3 $\mathcal{P}_{\text{cc}} \leftarrow \mathcal{P}_{\text{cc}} \cup \{(t_\mu, r_\mu, \mathbf{P}_\mu)\}$.

4. **Grid-lines extraction.** For each $\mathbf{P}_\mu \in \mathcal{P}_{\text{peak}} \cup \mathcal{P}_{\text{cc}}$,

– $\mathcal{U} \leftarrow \mathcal{U} \cup \{t_\mu\}$,

– $\mathcal{V} \leftarrow \mathcal{V} \cup \{r_\mu\}$.

In the above algorithms, after the computation of the zero-moments for discrete data (described at step 1), we look for sharp corners or *peaks*, in accordance with the magnitude of $\|\mathbf{n}_\epsilon(\mathbf{P}_{\mu,\nu})\|$ or $\|\mathbf{n}_\epsilon(\mathbf{P}_\mu)\|$ (step 2). Step 3 in Algorithms 1 and 2 describes the selection of convexity changes exploiting the different topology of data. For (2.1), with a tensor product topology of the data, we select *convexity changes* by considering the angles $\alpha_{\mu,\nu}^u, \alpha_{\mu,\nu}^v$ between consecutive ϵ -normals in u and v directions. Then we assign to each parametric line $u = t_\mu, \mu = 0, 1, \dots, M$ ($v = r_\nu, \nu = 0, 1, \dots, N$) an indicator $\mathcal{I}_\mu^{\text{cc}} \in [0, 1]$ ($\mathcal{I}_\nu^{\text{cc}} \in [0, 1]$) depending on the angles $\alpha_{\mu,l}^u, l = 0, 1, \dots, N$ ($\alpha_{q,\nu}^v, q = 0, 1, \dots, M$) and on the amount of convexity changes on such line. In order to select as knots the lines associated to the most significant changes in convexity, we choose the parameters with indicator $\mathcal{I}_\mu^{\text{cc}}$ or $\mathcal{I}_\nu^{\text{cc}}$ greater than a threshold tolerance (step 4 of Algorithm 1). For the sparse data (2.2), the indicator $\mathcal{F}_\mu^{\text{cc}}, \mu = 0, 1, \dots, M$, is a function of the angles between $\mathbf{n}_\epsilon(\mathbf{P}_\mu)$ and the ϵ -normals of all points contained in a ball centered at \mathbf{P}_μ and of the amount of convexity changes inside the ball. Again, great values of $\mathcal{F}_\mu^{\text{cc}} \in [0, 1]$ correspond to significant convexity changes. Both *peak* points and *convexity changes* are used in step 4 for the selection of the knot sequences \mathcal{U} and \mathcal{V} .

Acknowledgment. The authors wish to thank the referees for their helpful comments and remarks.

REFERENCES

- [1] U. CLARENZ, M. RUMPF, AND A. TELEA, *Robust feature detection and local classification for surfaces based on moment analysis*, IEEE Trans. Visualization and Computer Graphics, 10 (2004), pp. 516–524.
- [2] P. COSTANTINI, *Properties and applications of new polynomial spaces*, Internat. J. Wavel. Mult. Inform. Process., 4 (2006), pp. 489–507.
- [3] P. COSTANTINI AND F. PELOSI, *Shape-preserving approximation by space curves*, Numer. Algebra, 27 (2001), pp. 219–316.
- [4] P. COSTANTINI AND F. PELOSI, *Shape preserving approximation of spatial data*, Adv. Comput. Math., 20 (2004), pp. 25–51.
- [5] P. COSTANTINI AND F. PELOSI, *Shape preserving data approximation using new spline spaces*, in Mathematical Methods for Curves and Surfaces: Tromsø 2004, M. Dæhlen, K. Mørken, and L. L. Schumaker, eds., Nashboro Press, 2005, pp. 81–92.
- [6] P. COSTANTINI, F. PELOSI, AND M. L. SAMPOLI, *Boolean surfaces with shape constraints*, Comput. Aided Design (Special issue: Constrained Design of Curves and Surfaces), 40 (2008), pp. 62–75.
- [7] I. CRAVERO AND C. MANNI, *Detecting the shape of spatial data via zero moments*, in Mathematical Methods for Curves and Surfaces: Tromsø 2004, M. Dæhlen, K. Mørken, and L. L. Schumaker, eds., Nashboro Press, (2005), pp. 93–102.
- [8] M. S. FLOATER, *A weak condition for the convexity of tensor-product Bézier and B-spline surfaces*, Adv. Comput. Math., 2 (1994), pp. 67–80.
- [9] M. S. FLOATER, *Parametrization and smooth approximation of surface triangulations*, Comput. Aided Geom. Design, 14 (1997), pp. 231–250.
- [10] M. S. FLOATER AND K. HORMANN, *Parameterization of triangulations and unorganized points*, in Tutorials on Multiresolution in Geometric Modelling, A. Iske, E. Quak, and M. S. Floater, eds., Springer-Verlag, Heidelberg, Germany, 2002, pp. 287–315.
- [11] T. N. T. GOODMAN, *Total positivity and the shape of curves*, in Total Positivity and its Applications, M. Gasca and C. A. Micchelli, eds., Kluwer, Dordrecht, 1996, pp. 157–186.
- [12] J. HOSCHEK AND D. LASSER, *Fundamentals of Computer Aided Geometric Design*, A. K. Peters, Ltd, Wellesley, MA, 1993.
- [13] B. JÜTTLER, *Shape preserving least-squares approximation by polynomial parametric spline curves*, Comput. Aided Geom. Design, 14 (1997), pp. 731–747.
- [14] B. JÜTTLER, *Linear convexity conditions for parametric tensor-product Bézier surface patches*, The Mathematics of Surfaces VII, T. N. T. Goodman and R. Martin, eds., IMA Publications, 1998.
- [15] B. JÜTTLER, *Convex surface fitting with parametric Bézier surfaces*, M. Daehlen, T. Lyche, and L. L. Schumaker, eds., Mathematical Methods for Curves and Surfaces II, Vanderbilt University Press, Nashville, TN, 1998, pp. 263–270.
- [16] G. D. KORAS AND P. D. KAKLIS, *Convexity conditions for parametric tensor-product B-spline surfaces*, Adv. Comput. Math., 10 (1999), pp. 291–309.
- [17] F. PELOSI, *Shape detection for bivariate data*, Dipartimento di Scienze Matematiche ed Informatiche, Università di Siena, Rapporto n.492 (2008), Comput. Geom. Design, submitted.
- [18] H. POTTMANN, Q. X. HUANG, Y. L. YANG, AND S. KOLPL, *Integral Invariants for robust geometry Processing*, Technical report 146, Geometry Preprint Series, Vienna Univ. Technology, November 2005.

GEGENBAUER TAU METHODS WITH AND WITHOUT SPURIOUS EIGENVALUES*

MARIOS CHARALAMBIDES[†] AND FABIAN WALEFFE[‡]

Abstract. It is proven that a class of Gegenbauer tau approximations to a fourth order differential eigenvalue problem of a hydrodynamic type provides real, negative, and distinct eigenvalues, as is the case for the exact solutions. This class of Gegenbauer tau methods includes Chebyshev and Legendre Galerkin and “inviscid” Galerkin but does not include Chebyshev and Legendre tau. Rigorous and numerical results show that the results are sharp: positive or complex eigenvalues arise outside of this class. The widely used modified tau approach is proved to be equivalent to the Galerkin method.

Key words. spurious eigenvalues, Gegenbauer, spectrum, stable polynomials, positive pairs

AMS subject classifications. 65D30, 65L10, 65L15, 65M70, 65N35, 26C10

DOI. 10.1137/070704228

1. Introduction. The Chebyshev tau method used by Orszag [17] to obtain exponentially accurate solutions of the Orr–Sommerfeld equation yields two eigenvalues with large positive real parts. Such eigenvalues also occur for the Stokes modes in a channel given by the fourth order differential equation

$$(1.1) \quad (D^2 - \alpha^2)^2 u = \lambda (D^2 - \alpha^2) u,$$

with the boundary conditions $u = Du = 0$ at $x = \pm 1$, where λ is the eigenvalue, $u = u(x)$ is the eigenfunction, $D = d/dx$, and α is a real wavenumber. The Stokes eigenvalues λ are real and negative as can be checked by multiplying (1.1) by u^* , the complex conjugate of $u(x)$, and by integrating by parts twice using the no-slip boundary conditions. In fact the Stokes spectrum has been known analytically since Rayleigh [7, section 26.1]. Yet, the Chebyshev tau method applied to (1.1) yields two eigenvalues with large positive real parts, for any order of approximation and for any numerical accuracy. Such eigenvalues are obviously *spurious* for (1.1). Gottlieb and Orszag [11, Chapter 13] introduced the eigenvalue problem

$$(1.2) \quad \begin{aligned} D^4 u &= \lambda D^2 u & \text{in } -1 \leq x \leq 1, \\ u &= Du = 0 & \text{at } x = \pm 1 \end{aligned}$$

as an even simpler one-dimensional (1D) model of incompressible fluid flow. This is the $\alpha \rightarrow 0$ limit of the eigenvalue problem (1.1) and of the Orr–Sommerfeld equation [17]. For any fixed α , problem (1.2) is also the asymptotic equation for large λ solutions of the Stokes and the Orr–Sommerfeld equations. The eigensolutions of (1.2) are known analytically. They consist of even modes $u(x) = 1 - \cos(n\pi x)/\cos(n\pi)$, with $\lambda = -n^2\pi^2$, and odd modes $u(x) = x - \sin(q_n x)/\sin(q_n)$, with $\lambda = -q_n^2$, where $q_n = \tan q_n$ so that $n\pi < q_n < (2n + 1)\pi/2 \forall$ integers $n > 0$. The key properties of

*Received by the editors October 1, 2007; accepted for publication (in revised form) June 5, 2008; published electronically October 24, 2008.

<http://www.siam.org/journals/sinum/47-1/70422.html>

[†]Department of Business Administration, Frederick University Cyprus, 7 Yianni Frederickou Street, Pallouriotissa, PO Box 24729, 1303 Nicosia, Cyprus (bus.chm@fit.ac.cy).

[‡]Department of Mathematics, University of Wisconsin, Madison, WI 53706 (waleffe@math.wisc.edu). This author’s work was supported in part by NSF grant DMS-0204636.

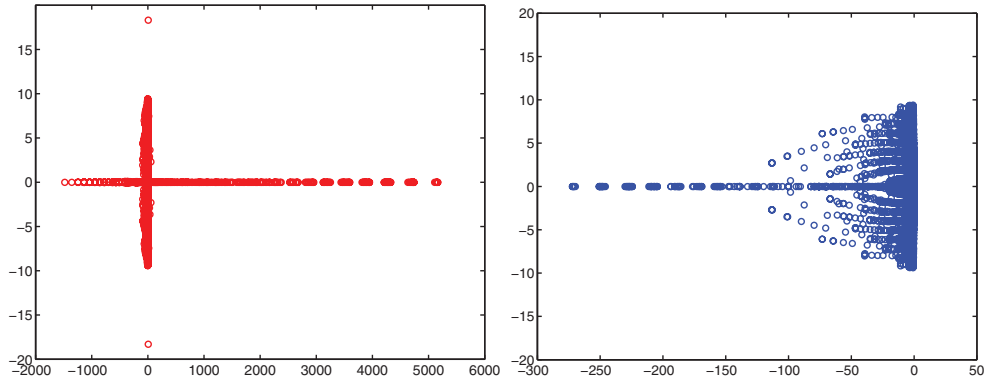


FIG. 1.1. Eigenvalues of a 3D steady state solution of the Navier–Stokes equations for plane Couette flow [21] computed with Chebyshev tau (left) and Chebyshev Galerkin (right) methods for identical resolutions (8773 modes after symmetry reductions. The solutions themselves are indistinguishable). Note the difference in horizontal scales. The Chebyshev tau method produces 274 eigenvalues with positive real parts, 273 of which are spurious. The Chebyshev Galerkin method returns only one positive eigenvalue, the physical one, equal to 0.03681 at Reynolds number 1000 [21, Figure 4].

these solutions are that the eigenvalues are *real*, *negative*, and *distinct*, and the even and odd mode eigenvalues interlace. These properties also hold for the Stokes modes, the solutions of (1.1) but not for the Orr–Sommerfeld modes.

The Chebyshev tau method provides spectrally accurate approximations to the lower magnitude eigenvalues, but it also yields two large positive eigenvalues for problems (1.1) and (1.2) [11, Table 13.1]. Those positive eigenvalues are clearly *spurious* since it is known that (1.1) and (1.2) should only have negative eigenvalues. The Chebyshev tau method yields two spurious eigenvalues for no-slip (i.e., clamped) boundary conditions $u = Du = 0$ at $x = \pm 1$ but none for the free-slip boundary conditions $u = D^2u = 0$ at $x = \pm 1$. The latter problem reduces to the second order problem $D^2v = \lambda v$, with $v(\pm 1) = 0$, for which a class of Jacobi and Gegenbauer tau methods has been proven to yield real, negative, and distinct eigenvalues [4, 5]. For mixed boundary conditions, e.g., $u(\pm 1) = Du(-1) = D^2u(1) = 0$, there is one spurious eigenvalue (this is a numerical observation).

For a 1D problem such as (1.1) or (1.2), the spurious eigenvalues are easy to recognize, and they appear as minor nuisances. Boyd [2, section 7.6] even questions the value of distinguishing between “spurious” and numerically inaccurate eigenvalues. However, in many applications, large *negative* eigenvalues are inconsequential, while “spurious” *positive* eigenvalues are very significant, and in higher dimensions, spurious eigenvalues are not as easy to pick out and set aside. In a recent application, three-dimensional (3D) *unstable* traveling wave solutions of the Navier–Stokes equations were calculated with both free-slip and no-slip boundary conditions, and anything in between, by Newton’s method [19, 20]. In that application, the Chebyshev tau method provides hundreds of spurious unstable eigenvalues, depending on the resolution and the exact type of boundary conditions, not all of which have very large magnitudes (Figure 1.1, left). A simple change in the *test* functions from Chebyshev polynomials $T_n(x)$ to $(1 - x^2)T_n(x)$ or $(1 - x^2)^2T_n(x)$ eliminates all of those spurious eigenvalues (Figure 1.1, right). This, and more, is proven below for the test problem (1.2) in the broader context of Gegenbauer tau methods, which include Chebyshev and Legendre tau and Chebyshev and Legendre Galerkin methods.

Another practical consequence of the spurious eigenvalues is that the Chebyshev tau method is unconditionally unstable when applied to the time-dependent version of (1.1) or (1.2), with $\partial/\partial t$ in place of λ . Such time-dependent problems appear as building blocks in the Navier–Stokes simulations of channel-type flows. Gottlieb and Orszag [11, p. 145] proposed a *modified tau* method for the time-dependent problems and proved that the modified method was stable for even solutions. The modified tau method (section 6) is a key idea behind several successful time integration schemes for the Navier–Stokes equations [3, section 7.3], [14]. The modified tau method amounts to using two more expansion polynomials for the fourth order differential operator on the left-hand sides of (1.2) and (1.1) than for the second order operator on the right hand sides. That modified tau method was adapted to eigenvalue problems by Gardner, Trogdon, and Douglass [8] and McFadden, Murray, and Boisvert [16]. McFadden, Murray, and Boisvert showed the equivalence between the modified Chebyshev tau method and the Chebyshev Galerkin method by direct calculation. Zebib [22] had given numerical evidence that the Galerkin method removed spurious eigenvalues. The modified tau method idea was adapted to the collocation formulation by Huang and Sloan [13].

A heuristic “explanation” for spurious eigenvalues is that there is a “mismatch” between the number of boundary conditions applied to the fourth order operator on the left-hand side of (1.2) and those applied to the second order operator on the right-hand side. That interpretation fits with the modified tau method, which uses two more polynomials for the fourth order operator than for the second order operator. However, it is incorrect, since, while the tau method for Chebyshev polynomials of the first kind $T_n(x)$ gives spurious eigenvalues, for instance, the tau method for Chebyshev polynomials of the second kind $U_n(x)$ does not.

All of these various methods are best seen in the context of the Gegenbauer class with residuals weighted by $W^{(\gamma)}(x) = (1 - x^2)^{\gamma-1/2}$ (section 2), where $\gamma = 0$ corresponds to Chebyshev and $\gamma = 1/2$ to Legendre polynomials. Dawkins, Dunbar, and Douglass [6] proved the existence of spurious positive eigenvalues for (1.2) when $\gamma < 1/2$. The proof is straightforward. For (1.2), the polynomial equation for $\mu = 1/\lambda$ can be derived explicitly (section 4). All coefficients of that polynomial are real and positive, except the constant term which is negative when $\gamma < 1/2$. Hence there is one real positive μ and a “spurious” positive eigenvalue when $\gamma < 1/2$ (details are given in section 5.2). For $\gamma = 1/2$, the Legendre tau case, the constant term is zero, and hence there is one $\mu = 0$ eigenvalue or a $\lambda = 1/\mu = \infty$ eigenvalue. Perturbation analysis shows that the $\lambda = \infty$ eigenvalues become very large positive eigenvalues for $\gamma < 1/2$ and very large negative eigenvalues for $\gamma > 1/2$. We provide a quicker derivation of those results in section 3. Dawkins, Dunbar, and Douglass’s results do not prove that there are no spurious eigenvalues for $\gamma > 1/2$ since there could be complex eigenvalues with positive real parts. In section 5, we prove that the Gegenbauer tau method applied to (1.2) provides eigenvalues that are *real*, *negative*, and *distinct* when $1/2 < \gamma \leq 7/2$. This provides a complete characterization of the Gegenbauer tau spectrum for problem (1.2). Numerical calculations confirm that the range $1/2 < \gamma \leq 7/2$ is sharp. Spurious positive eigenvalues exist for $\gamma < 1/2$ [6], and complex eigenvalues arise for $\gamma > 7/2$, for sufficiently high polynomial order. In section 6, we prove that the modified tau method is mathematically equivalent to the Galerkin approach.

Obviously, $\gamma = 1/2$ is a critical value for the weight function $(1 - x^2)^{\gamma-1/2}$. The boundaries $x = \pm 1$ have infinite weight for $\gamma < 1/2$ and zero weight for $\gamma > 1/2$, but we do not know a valid heuristic explanation for “spurious” eigenvalues beyond that observation, if one exists. Section 3 provides further insights into the nature of

the spurious eigenvalues and gives *some* support for the view that “spurious” and numerically inaccurate eigenvalues are related. In Figure 1.1, for *fixed resolution*, the 273 spurious eigenvalues for $\gamma = 0$ (Chebyshev tau) escape to $+\infty$ as $\gamma \nearrow 1/2$ (Legendre tau). They come back from $-\infty$ as γ increases beyond $1/2$. Thus there is indeed a connection between large positive eigenvalues and large negative eigenvalues, but whether we have “spurious” positive eigenvalues or inconsequential very negative eigenvalues is sharply controlled by γ , irrespective of the order of approximation n (see also (3.6)).

2. Tau and Galerkin methods.

DEFINITION 2.1. *A Gegenbauer tau method approximates the solution $u(x)$ of a differential equation in $-1 \leq x \leq 1$ by a polynomial of degree n , $u_n(x)$, that satisfies the m boundary conditions exactly. The remaining $n + 1 - m$ polynomial coefficients are determined by requiring that the residual be orthogonal to all polynomials of degree $n - m$ (or less) with respect to the Gegenbauer weight $W^{(\gamma)}(x) = (1 - x^2)^{\gamma-1/2}$, with $\gamma > -1/2$.*

For problem (1.2), the residual

$$(2.1) \quad R_{n-2}(x) \equiv \lambda D^2 u_n(x) - D^4 u_n(x)$$

is a polynomial of degree $n - 2$ in x . The polynomial approximation $u_n(x)$ is determined from the four boundary conditions $u_n(\pm 1) = Du_n(\pm 1) = 0$ and the requirement that $R_{n-2}(x)$ is orthogonal to *all* polynomials $q_{n-4}(x)$ of degree $n - 4$ or less with respect to the weight function $W^{(\gamma)}(x) = (1 - x^2)^{\gamma-1/2} \geq 0$ in the interval $(-1, 1)$

$$(2.2) \quad \int_{-1}^1 R_{n-2}(x) q_{n-4}(x) W^{(\gamma)}(x) dx = 0 \quad \forall q_{n-4}(x).$$

This provides $n - 3$ equations which, together with the four boundary conditions, yield $n + 1$ equations for the $n + 1$ undetermined coefficients in the polynomial approximation $u_n(x)$. For the Gegenbauer weight function $W^{(\gamma)}(x) = (1 - x^2)^{\gamma-1/2}$, the residual can be written explicitly as

$$(2.3) \quad R_{n-2}(x) = \tau_0 \lambda G_{n-2}^{(\gamma)}(x) + \tau_1 \lambda G_{n-3}^{(\gamma)}(x)$$

for some x -independent coefficients τ_0 and τ_1 , where $G_n^{(\gamma)}(x)$ is the Gegenbauer polynomial of degree n . This follows from orthogonality of the Gegenbauer polynomials in $-1 < x < 1$ with respect to the weight $(1 - x^2)^{\gamma-1/2}$, which implies the orthogonality of the Gegenbauer polynomial of degree k to *any* polynomial of degree $k - 1$ or less with respect to that weight function.

Gegenbauer (a.k.a. ultraspherical) polynomials are a special subclass of the Jacobi polynomials [1]. The latter are the most general class of polynomial solutions of a Sturm–Liouville eigenproblem that is singular at ± 1 , as required for faster-than-algebraic convergence [3]. Gegenbauer polynomials are the most general class of polynomials with the odd-even symmetry $G_n^{(\gamma)}(x) = (-1)^n G_n^{(\gamma)}(-x)$. This is a one-parameter family of polynomials, with the parameter $\gamma > -1/2$. Chebyshev polynomials correspond to $\gamma = 0$ and Legendre polynomials to $\gamma = 1/2$. We use a (slightly) nonstandard normalization of Gegenbauer polynomials since the standard normalization [1] is singular in the Chebyshev case. Some key properties of Gegenbauer polynomials used in this paper are given in Appendix A. Note that if $\lambda = 0$, then, from (2.1), the residual $R_{n-2}(x)$ must be a polynomial of degree $n - 4$ implying

that $\tau_0 = \tau_1 = 0$ in (2.3) and $D^4 u_n(x) = 0$ for all x in $(-1, 1)$. The boundary conditions $u_n(\pm 1) = Du_n(\pm 1) = 0$ then imply that $u_n(x) = 0$ for all x in $[-1, 1]$, the trivial solution. Hence we can assume that $\lambda \neq 0$ in the Gegenbauer tau method applied to (1.2). Following common usage [11, 3, 22, 16], we have the following definition.

DEFINITION 2.2. *The Gegenbauer Galerkin method approximates the solution $u(x)$ of a differential equation in $-1 \leq x \leq 1$ by a polynomial of degree n , $u_n(x)$, that satisfies the m boundary conditions exactly. The remaining $n+1-m$ polynomial coefficients are determined by imposing the fact that the residual be orthogonal to all polynomials of degree n (or less) that satisfy the homogeneous boundary conditions with respect to the Gegenbauer weight $W^{(\gamma)}(x) = (1-x^2)^{\gamma-1/2}$, with $\gamma > -1/2$.*

Strictly speaking, this a *Petrov-Galerkin* method since the test functions are not identical to the trial functions because of the Gegenbauer weight $(1-x^2)^{\gamma-1/2}$ [3].

For problem (1.2), $u_n(x)$ is determined from the boundary conditions $u_n(\pm 1) = Du_n(\pm 1) = 0$ and the orthogonality with respect to the weight $W^{(\gamma)}(x) = (1-x^2)^{\gamma-1/2}$ of the residual (2.1) to all polynomials of degree n that vanish together with their derivative at $x = \pm 1$. Such polynomials can be written as $(1-x^2)^2 q_{n-4}(x)$, where $q_{n-4}(x)$ is an arbitrary polynomial of degree $n-4$, and the weighted residual equations read

$$(2.4) \quad \int_{-1}^1 R_{n-2}(x) (1-x^2)^2 q_{n-4}(x) W^{(\gamma)}(x) dx = 0 \quad \forall q_{n-4}(x).$$

The Gegenbauer Galerkin method is therefore equivalent to the tau method for the weight $W^{(\gamma+2)}(x) = (1-x^2)^2 W^{(\gamma)}(x)$, and its residual has the explicit form

$$(2.5) \quad R_{n-2}(x) = \tau_0 \lambda G_{n-2}^{(\gamma+2)}(x) + \tau_1 \lambda G_{n-3}^{(\gamma+2)}(x).$$

So the *Chebyshev (or Legendre) Galerkin* method for clamped boundary conditions, $u_n(\pm 1) = Du_n(\pm 1) = 0$, is in fact a tau method for Chebyshev (or Legendre) polynomials of the *third kind* (proportional to the second derivative of Chebyshev (or Legendre) polynomials (A.5)). Since we consider a range of the Gegenbauer parameter γ , the Gegenbauer tau method also includes some Gegenbauer Galerkin methods.

This suggests an intermediate method where the test functions are polynomials that vanish at $x = \pm 1$ (inviscid boundary conditions only).

DEFINITION 2.3. *The Gegenbauer “inviscid Galerkin” method determines $u_n(x)$ from the four boundary conditions $u_n(\pm 1) = Du_n(\pm 1) = 0$ and the orthogonality of the residual to all polynomials of degree $n-2$ that vanish at $x = \pm 1$ with respect to the weight function $W^{(\gamma)}(x) = (1-x^2)^{\gamma-1/2}$.*

Such test polynomials can be written in the form $(1-x^2)q_{n-4}(x)$, where $q_{n-4}(x)$ is an arbitrary polynomial of degree $n-4$, and so the weighted residual equations read

$$(2.6) \quad \int_{-1}^1 R_{n-2}(x) (1-x^2) q_{n-4}(x) W^{(\gamma)}(x) dx = 0 \quad \forall q_{n-4}(x).$$

The *Gegenbauer inviscid Galerkin* method is therefore equivalent to the Gegenbauer tau method with weight $W^{(\gamma+1)}(x)$, and its residual for (1.2) is

$$(2.7) \quad R_{n-2}(x) = \tau_0 \lambda G_{n-2}^{(\gamma+1)}(x) + \tau_1 \lambda G_{n-3}^{(\gamma+1)}(x).$$

Thus the Chebyshev (or Legendre) inviscid Galerkin method is a tau method for Chebyshev (or Legendre) polynomials of the *second kind*. Since we consider a range of

the Gegenbauer parameter γ , the Gegenbauer tau method also includes some Gegenbauer inviscid Galerkin methods.

For completeness, we list the *collocation* approach, where $u_n(x)$ is determined from the boundary conditions $u_n(\pm 1) = Du_n(\pm 1) = 0$ and from enforcing $R_{n-2}(x_j) = 0$ at the $n - 3$ interior Gauss-Lobatto points x_j such that $DG_{n-2}(x_j) = 0$, $j = 1, \dots, n - 3$, [3, section 2.2]. The residual (2.1) has the form [10, equation (4.5)]

$$(2.8) \quad R_{n-2}(x) = (A + Bx) DG_{n-2}(x),$$

for some A and B independent of x . That residual can be written in several equivalent forms by using the properties of Gegenbauer polynomials (Appendix A). We do not have rigorous results for the collocation method.

3. Legendre and near-Legendre tau cases. Here we provide a quicker and more complete derivation of earlier results [6] about spurious eigenvalues for the Legendre and near-Legendre tau cases. This section provides a useful technical introduction to the problem, but it is not necessary to derive the main results of this paper. Dawkins, Dunbar, and Douglass [6], focusing only on even modes, use the monomial basis x^{2k} to derive an explicit form for the generalized eigenvalue problem $Aa = \lambda Ba$ for the Legendre tau method. In the monomial basis, the matrix A is upper triangular and nonsingular, and the matrix B is upper Hessenberg, but its first row is identically zero; hence there exists one infinite eigenvalue. A perturbation analysis is used to show that the infinite eigenvalue of the Legendre tau method becomes a large positive eigenvalue for the Gegenbauer tau methods with $\gamma < 1/2$ and a large negative eigenvalue for $\gamma > 1/2$.

In the Legendre tau method, the polynomial approximation $u_n(x)$ of degree n to problem (1.2) satisfies the four boundary conditions $u_n = Du_n = 0$ at $x = \pm 1$. Thus $u_n(x) = (1 - x^2)^2 p_{n-4}(x)$, and the polynomial $p_{n-4}(x)$ is determined from the weighted residual equation (2.2), with $\gamma = 1/2$ and $W^{(1/2)}(x) = 1$,

$$(3.1) \quad \int_{-1}^1 (\mu D^4 u_n - D^2 u_n) q_{n-4}(x) dx = 0 \quad \forall q_{n-4}(x).$$

The mathematical problem is fully specified, except for an arbitrary multiplicative constant for $u_n(x)$. Choosing various polynomial bases for $p_{n-4}(x)$ and $q_{n-4}(x)$ will lead to distinct matrix problems, but those problems are all similar to each other and provide exactly the same eigenvalues, in exact arithmetic.

We use the bases $G_l^{(5/2)}(x)$ for $p_{n-4}(x)$ and $G_k^{(1/2)}(x)$ for $q_{n-4}(x)$, with $k, l = 0, \dots, n - 4$, where $G_n^{(\gamma)}(x)$ is the Gegenbauer polynomial of degree n for index γ (see Appendix A, and recall that $G_k^{(1/2)}(x) = P_k(x)$ are Legendre polynomials). Thus we write $u_n(x) = \sum_{l=0}^{n-4} a_l (1 - x^2)^2 G_l^{(5/2)}(x)$ for some $n - 3$ coefficients a_l to be determined. The tau equation (3.1) provides the matrix eigenproblem $\mu Aa = Ba$ or $\mu \sum_{l=0}^{n-4} A(k, l) a_l = \sum_{l=0}^{n-4} B(k, l) a_l$ with

$$(3.2) \quad A(k, l) = \int_{-1}^1 D^4 \left[(1 - x^2)^2 G_l^{(5/2)}(x) \right] G_k^{(1/2)}(x) dx,$$

$$(3.3) \quad B(k, l) = \int_{-1}^1 D^2 \left[(1 - x^2)^2 G_l^{(5/2)}(x) \right] G_k^{(1/2)}(x) dx,$$

for $k, l = 0, \dots, n-4$. Using (B.1), these expressions simplify to

$$(3.4) \quad A(k, l) = \mathcal{C}_l \int_{-1}^1 \left[D^2 G_{l+2}^{(1/2)}(x) \right] G_k^{(1/2)}(x) dx,$$

$$(3.5) \quad B(k, l) = \mathcal{C}_l \int_{-1}^1 G_{l+2}^{(1/2)}(x) G_k^{(1/2)}(x) dx,$$

where $\mathcal{C}_l = \frac{1}{15}(l+1)(l+2)(l+3)(l+4)$.

Since the Legendre polynomials $G_n^{(1/2)}(x) = P_n(x)$ are orthogonal with respect to the unit weight, (3.5) yields that $B(k, l) \propto \delta_{k, l+2}$, where $\delta_{k, l+2}$ is the Kronecker delta so that $B(0, l) = B(1, l) = 0$, for all l , and B has nonzero elements only on the second subdiagonal. For $A(k, l)$, use (A.10) to express $D^2 G_{l+2}^{(1/2)}(x)$ as a linear combination of $G_l^{(1/2)}(x)$, $G_{l-2}^{(1/2)}(x)$, etc. Orthogonality of the Legendre polynomials $G_n^{(1/2)}(x)$ then implies that $A(k, l)$ is upper triangular with nonzero diagonal elements. Hence A is nonsingular, while the nullspace of B is two-dimensional. The eigenvalue problem $\mu Aa = Ba$ therefore has two $\mu = 0$ eigenvalues. Since the only nonzero elements of B consists of the subdiagonal $B(l+2, l)$, the two *right* eigenvectors corresponding to $\mu = 0$ are $a = [0, \dots, 0, 1]^T$ and $[0, \dots, 0, 1, 0]^T$, respectively. In other words,

$$(3.6) \quad u_n(x) = (1-x^2)^2 G_{n-4}^{(5/2)}(x) \quad \text{and} \quad u_n(x) = (1-x^2)^2 G_{n-5}^{(5/2)}(x)$$

satisfy the boundary conditions and the tau equation (3.1) with $\mu = 0 = 1/\lambda$, for all $n \geq 5$. One mode is even, and the other one is odd. Likewise, the *left* eigenvectors $b^T = [1, 0, \dots, 0]$ and $[0, 1, 0, \dots, 0]$ satisfy $\mu b^T A = b^T B$ with $\mu = 0$. These results are for the Legendre case, and $\mu = 0$ corresponds to $\lambda = 1/\mu = \infty$.

Now consider the Gegenbauer tau equations for $\gamma - 1/2 = \epsilon$ with $|\epsilon| \ll 1$, the near-Legendre case. The equation is (3.1) but with the extra weight factor $W^{(\gamma)}(x) = (1-x^2)^\epsilon$ inside the integral. We can figure out what happens to the $\mu = 0$ eigenvalues of the $\epsilon = 0$ Legendre case by perturbation. The matrices A and B and the left and right eigenvectors, denoted a and b , respectively, as well as the eigenvalue μ , now depend on ϵ . Let

$$(3.7) \quad A = A_0 + \epsilon A_1 + O(\epsilon^2), \quad B = B_0 + \epsilon B_1 + O(\epsilon^2),$$

$$(3.8) \quad a = a_0 + \epsilon a_1 + O(\epsilon^2), \quad b = b_0 + \epsilon b_1 + O(\epsilon^2),$$

$$(3.9) \quad \mu = \mu_0 + \epsilon \mu_1 + O(\epsilon^2),$$

where A_0 and B_0 are the matrices obtained above in (3.4) and (3.5) for $\epsilon = 0$, while a_0 and b_0^H are the corresponding right and left eigenvectors so that $\mu_0 A_0 a_0 = B_0 a_0$ and $\mu_0 b_0^H A_0 = b_0^H B_0$. Substituting these ϵ -expansions in the eigenvalue equation $\mu Aa = Ba$ and canceling out the zeroth order term, we obtain

$$(3.10) \quad \mu_1 A_0 a_0 + \mu_0 A_1 a_0 + \mu_0 A_0 a_1 = B_1 a_0 + B_0 a_1 + O(\epsilon).$$

Multiplying by b_0^H cancels out the $\mu_0 b_0^H A_0 a_1 = b_0^H B_0 a_1$ terms, so we obtain

$$(3.11) \quad \mu_1 = \frac{b_0^H B_1 a_0 - \mu_0 b_0^H A_1 a_0}{b_0^H A_0 a_0}.$$

This expression is general but simplifies further since we are interested in the perturbation of the zero eigenvalues $\mu_0 = 0$. This expression for μ_1 is quite simple since b_0 ,

a_0 , and A_0 are the zeroth order objects. All we need to compute when $\mu_0 = 0$ is the first order correction B_1 to the matrix B . But since a_0 and b_0 have only one nonzero component as given at the end of the previous paragraph, we need only to calculate two components of the B matrix. For n even, all that is needed are the first order corrections to $B(0, n-4)$ for the even mode and to $B(1, n-5)$ for the odd mode. For n odd, we need $B(0, n-5)$ for the even mode and $B(1, n-4)$ for the odd mode; however, since the even and odd modes decouple in this problem, it suffices to compute both the even and odd modes in only one case of n even or odd. The matrix elements in the $\epsilon \neq 0$ cases are still given by (3.4) and (3.5) but with the extra $(1-x^2)^\epsilon$ weight factor inside the integrals. Since $G_1^{(1/2)}(x) = x$ and $G_n^{(1/2)}(x) = P_n(x)$, the Legendre polynomial of degree n , we obtain

$$(3.12) \quad B(0, n-4) = C_{n-4} \int_{-1}^1 P_{n-2}(x) (1-x^2)^\epsilon dx = \epsilon B_1(0, n-4) + O(\epsilon^2),$$

$$(3.13) \quad B(1, n-5) = C_{n-5} \int_{-1}^1 x P_{n-3}(x) (1-x^2)^\epsilon dx = \epsilon B_1(1, n-5) + O(\epsilon^2),$$

with C_l as defined in (3.5). The integrals are readily evaluated, and details are provided in Appendix B. Using (B.2), (B.5), and (B.8), we obtain for the even mode (for n even) that

$$(3.14) \quad \mu_1 = \frac{B_1(0, n-4)}{A_0(0, n-4)} = \frac{-4}{(n-2)^2(n-1)^2}.$$

This matches the formula in Dawkins, Dunbar, and Douglass [6, p. 456] since their $2N = n-4$ and $2\nu - 1 = 2\epsilon$. Likewise using (B.7) and (B.9) for the odd mode (with n even) yields

$$(3.15) \quad \mu_1 = \frac{B_1(1, n-5)}{A_0(1, n-5)} = \frac{-4}{(n-4)^2(n-1)^2}.$$

Again if n is odd, then μ_1 for the even mode is given by (3.14) but with $n-1$ in lieu of n . Likewise for n odd, the odd mode is given by (3.15) with $n+1$ in lieu of n . Finally since $\lambda = 1/\mu$, the $\lambda = \infty$ eigenvalues in the Legendre tau case become $\lambda = 1/(\epsilon\mu_1 + O(\epsilon^2)) \sim 1/(\epsilon\mu_1)$ in the near-Legendre cases. From (3.14) and (3.15), these eigenvalues will be $O(n^4/\epsilon)$. Furthermore they will be positive when $\epsilon < 0$ (i.e., *spurious* when $\gamma < 1/2$) but negative when $\epsilon > 0$.

4. Characteristic polynomials. For the model problem (1.2), we can bypass the matrix eigenproblem of section 3 to directly derive the characteristic polynomial for the eigenvalues $\mu = 1/\lambda$. To do so, invert (2.1) to express the polynomial approximation $D^2 u_n(x)$ in terms of the residual $R_{n-2}(x)$

$$(4.1) \quad D^2 u_n(x) = \mu \sum_{k=0}^{\infty} \mu^k D^{2k} R_{n-2}(x),$$

where $\mu = 1/\lambda$. The inversion (4.1) follows from the application of the geometric (Neumann) series for $(1 - \mu D^2)^{-1} = \sum_{k=0}^{\infty} \mu^k D^{2k}$ which terminates since $R_{n-2}(x)$ is a polynomial. Thus $u_n(x)$ can be computed in terms of the unknown tau coefficients by double integration of (4.1) and the application of the boundary conditions. We can assume that $\lambda \neq 0$ because $\lambda = 0$ with $u_n(\pm 1) = Du_n(\pm 1) = 0$ necessarily

corresponds to the trivial solution $u_n(x) = 0 \forall x$ in $[-1, 1]$, as noted in the previous section.

The Gegenbauer polynomials are even in x for n even and odd for n odd (A.8). The symmetry of the differential equation (1.2) and of the Gegenbauer polynomials allows decoupling of the discrete problem into even and odd solutions. This parity reduction leads to simpler residuals and simpler forms for the corresponding characteristic polynomials. The residual in the parity-separated Gegenbauer case contains only one term

$$(4.2) \quad R_{n-2}(x) = \tau_0 \lambda G_{n-2}^{(\gamma)}(x),$$

instead of (2.3), where $G_n^{(\gamma)}(x)$ is the Gegenbauer polynomial of degree n with n even for even solutions and odd for odd solutions. Substituting (4.2) in (4.1) and renormalizing $u_n(x)$ by τ_0 gives

$$(4.3) \quad D^2 u_n(x) = \sum_{k=0}^{\infty} \mu^k D^{2k} G_{n-2}^{(\gamma)}(x).$$

For $\gamma > 1/2$, the identity (A.5) in the form $2\gamma G_{n-2}^{(\gamma)}(x) = DG_{n-1}^{(\gamma-1)}(x)$ can be used to write (4.3) in the form

$$(4.4) \quad D^2 u_n(x) = \frac{1}{2\gamma} \sum_{k=0}^{\infty} \mu^k D^{2k+1} G_{n-1}^{(\gamma-1)}(x),$$

which integrates to

$$(4.5) \quad Du_n(x) = \frac{1}{2\gamma} \sum_{k=0}^{\infty} \mu^k D^{2k} G_{n-1}^{(\gamma-1)}(x) + C,$$

where C is an arbitrary constant.

4.1. Even solutions. For even solutions $u_n(x) = u_n(-x)$, n is even and $Du_n(x)$ is odd so $C = 0$ in (4.5). The boundary condition $Du_n(1) = 0$ gives the characteristic equation for μ (for n even and $\gamma > 1/2$):

$$(4.6) \quad \sum_{k=0}^{\infty} \mu^k D^{2k} G_{n-1}^{(\gamma-1)}(1) = 0.$$

4.2. Odd solutions. For odd solutions, $u_n(x) = -u_n(-x)$, n is odd and the boundary condition $Du_n(1) = 0$ requires that

$$(4.7) \quad C = -\frac{1}{2\gamma} \sum_{k=0}^{\infty} \mu^k D^{2k} G_{n-1}^{(\gamma-1)}(1).$$

Substituting this C value in (4.5) and integrating gives

$$(4.8) \quad 2\gamma u_n(x) = \sum_{k=0}^{\infty} \mu^k D^{2k-1} G_{n-1}^{(\gamma-1)}(x) - x \sum_{k=0}^{\infty} \mu^k D^{2k} G_{n-1}^{(\gamma-1)}(1),$$

where we must define

$$(4.9) \quad D^{-1} G_{n-1}^{(\gamma-1)}(x) = \int_0^x G_{n-1}^{(\gamma-1)}(s) ds = \frac{G_n^{(\gamma-1)}(x) - G_{n-2}^{(\gamma-1)}(x)}{2(n + \gamma - 2)}$$

since $u_n(x)$ and n are odd, where we have used (A.10) to evaluate the integral and the symmetry (A.8) so that $G_n(0) = G_{n-2}(0) = 0$ for n odd. The boundary condition $u_n(1) = 0$ yields the characteristic polynomial equation (for n odd and $\gamma > 1/2$):

$$(4.10) \quad \sum_{k=0}^{\infty} \mu^k D^{2k-1} G_{n-1}^{(\gamma-1)}(1) - \sum_{k=0}^{\infty} \mu^k D^{2k} G_{n-1}^{(\gamma-1)}(1) = 0.$$

For $\gamma > 3/2$, we can use identity (A.5) in the form $2(\gamma-1)G_{n-1}^{(\gamma-1)}(x) = DG_n^{(\gamma-2)}(x)$ to write the characteristic equation (4.10) as

$$(4.11) \quad \sum_{k=0}^{\infty} \mu^k D^{2k} G_n^{(\gamma-2)}(1) - \sum_{k=0}^{\infty} \mu^k D^{2k+1} G_n^{(\gamma-2)}(1) = 0.$$

For $1/2 < \gamma \leq 3/2$, this cannot be used since $\gamma - 2 < -1/2$, but using (4.9) for the D^{-1} term in the first sum, the characteristic equation (4.10) can be written as

$$(4.12) \quad \mu \sum_{k=0}^{\infty} \mu^k D^{2k+1} G_{n-1}^{(\gamma-1)}(1) - \frac{G_{n-2}^{(\gamma-1)}(1) - G_n^{(\gamma-1)}(1)}{2(n+\gamma-2)} - \sum_{k=0}^{\infty} \mu^k D^{2k} G_{n-1}^{(\gamma-1)}(1) = 0.$$

5. Zeros of characteristic polynomials. Here we prove that the zeros of the characteristic polynomial equations (4.6) and (4.10) are real, negative, and distinct for $1/2 < \gamma \leq 7/2$. Some background material is needed.

5.1. Stable polynomials and the Hermite Biehler theorem. A polynomial $p(z)$ is *stable* if and only if all of its zeros have negative real parts. Stable polynomials can arise as characteristic polynomials of a numerical method applied to a differential equation as in [5] for $Du = \lambda u$, with $u(1) = 0$, and in other dynamical systems applications. The characterization of stable polynomials that is most useful here is given by [12], [18, p. 197].

THEOREM 5.1 (the Hermite Biehler theorem). *The real polynomial $p(z) = \Omega(z^2) + z\Theta(z^2)$ is stable if and only if $\Omega(\mu)$ and $\Theta(\mu)$ form a positive pair.*

DEFINITION 5.2. *Two real polynomials $\Omega(\mu)$ and $\Theta(\mu)$ of degree n and $n-1$ (or n), respectively, form a positive pair if*

(a) *the roots $\mu_1, \mu_2, \dots, \mu_n$ of $\Omega(\mu)$ and $\mu'_1, \mu'_2, \dots, \mu'_{n-1}$ (or $\mu'_1, \mu'_2, \dots, \mu'_n$) of $\Theta(\mu)$ are all distinct, real, and negative;*

(b) *the roots interlace as follows: $\mu_1 < \mu'_1 < \mu_2 < \dots < \mu'_{n-1} < \mu_n < 0$ (or $\mu'_1 < \mu_1 < \dots < \mu'_n < \mu_n < 0$);*

(c) *the highest coefficients of $\Omega(\mu)$ and $\Theta(\mu)$ are of like sign.*

We will use the following theorem about positive pairs [15, p. 198].

THEOREM 5.3. *Any nontrivial real linear combination of two polynomials of degree n (or n and $n-1$) with interlacing roots has real roots.*

(Since such a linear combination changes sign $n-1$ times along the real axis, it has $n-1$ real roots. Since it is a real polynomial of degree n , the remaining root is real also.)

5.2. Eigenvalues for even modes. In [5] and [4] we study the Gegenbauer tau method for $D^2u = \lambda u$, with $u(\pm 1) = 0$, which leads to the characteristic polynomial $\sum_{k=0}^{\infty} \mu^k D^{2k} G_n^{(\gamma)}(1)$. The derivation of that result is entirely similar to that in sections 2 and 4. The strategy to prove that the Gegenbauer tau method for that second order

problem has real, negative, and distinct roots is to show the stability of the polynomial

$$(5.1) \quad p(z) = \sum_{k=0}^n z^k D^k G_n^{(\gamma)}(1)$$

for $-1/2 < \gamma \leq 3/2$, then to use the Hermite Biehler theorem to deduce the following theorem.

THEOREM 5.4. *For $-1/2 < \gamma \leq 3/2$, the polynomials*

$$(5.2) \quad \Omega_n^{(\gamma)}(\mu) = \sum_{k=0}^{\infty} \mu^k D^{2k} G_n^{(\gamma)}(1) \quad \text{and} \quad \Theta_n^{(\gamma)}(\mu) = \sum_{k=0}^{\infty} \mu^k D^{2k+1} G_n^{(\gamma)}(1)$$

form a positive pair. From (A.5) this is equivalent to stating that the polynomials

$$(5.3) \quad \Omega_n^{(\gamma)}(\mu) = \sum_{k=0}^{\infty} \mu^k D^{2k} G_n^{(\gamma)}(1) \quad \text{and} \quad \Omega_{n-1}^{(\gamma+1)}(\mu) = \sum_{k=0}^{\infty} \mu^k D^{2k} G_{n-1}^{(\gamma+1)}(1)$$

also form a positive pair. Combining the γ and $\gamma+1$ ranges in (5.3) yields that $\Omega_n^{(\gamma)}(\mu)$ has real, negative, and distinct roots for $-1/2 < \gamma \leq 5/2$.

The stability of (5.1) is proven in [5, Theorem 1] for the broader class of Jacobi polynomials. The basic ideas of the proof are along the lines of Gottlieb [9] and Gottlieb and Lustman's work [10] on the stability of the Chebyshev collocation method for the first and second order operators.

For the fourth order problem (1.2), $D^4 u = \lambda D^2 u$, with $u(\pm 1) = Du(\pm 1) = 0$, the Gegenbauer tau method gives the characteristic polynomial (4.6) for even solutions. This is the polynomial $\Omega_{n-1}^{(\gamma-1)}(\mu)$ of (5.3) that appears for the second order problem [4, 5] and is known to have real, negative, and distinct eigenvalues for $-1/2 < \gamma - 1 \leq 5/2$, that is, for $1/2 < \gamma \leq 7/2$. Hence it follows directly from (4.6) and Theorem 5.4 that the Gegenbauer tau approximation for even solutions of problem (1.2), $D^4 u = \lambda D^2 u$, with $u(\pm 1) = Du(\pm 1) = 0$, has real, negative, and distinct eigenvalues for $1/2 < \gamma \leq 7/2$.

This result is sharp. For $\gamma > 7/2$ and sufficiently large n , our numerical computations show that the polynomial has a pair of complex eigenvalues. For $\gamma < 1/2$ the polynomial (4.6) has a real positive eigenvalue as first proven in [6]. The proof goes as follows. For $\gamma < 1/2$ we cannot use (4.5) since $\gamma - 1$ is below the range of definition of Gegenbauer polynomials (Appendix A). Instead, integrate (4.3) and use identity (A.10) to obtain $\int_0^1 G_{n-2}^{(\gamma)} dx = (G_{n-1}^{(\gamma)}(1) - G_{n-3}^{(\gamma)}(1))/(2(\gamma + n - 2))$ since $G_{n-1}(0) = G_{n-3}(0) = 0$ for n even (A.8). Using formula (A.11), one shows that $G_n^{(\gamma)}(1)$ increases with n if $\gamma > 1/2$ but decreases with n if $\gamma < 1/2$. Hence the constant term is negative for $\gamma < 1/2$, while all the other coefficients of the characteristic polynomial can be shown to be positive using (A.5) and (A.11). Therefore there is one real positive eigenvalue as proved in [6]. For $\gamma = 1/2$, the Legendre case, the constant term vanishes, and there is a $\mu = 0$ ($\lambda = \infty$) eigenvalue as established in section 3.

Remark 1. Exact even solutions of $D^4 u = \lambda D^2 u$, with $u(\pm 1) = Du(\pm 1) = 0$, obey $D^3 u = \lambda Du + C$, with $C = 0$, since $D^3 u$ and Du are odd. Thus even solution eigenvalues of (1.2) are equal to the eigenvalues for odd solutions of the second order problem $D^2 w = \lambda w$ with $w = 0$ at $x = \pm 1$, with $w = Du$. The Gegenbauer tau version of this property is that eigenvalues for even Gegenbauer tau solutions of (1.2) of order n (even) and index γ are equal to the eigenvalues for odd Gegenbauer tau solutions of $D^2 w = \lambda w$, $w(\pm 1) = 0$, of order $n - 1$ (odd) and index $\gamma - 1$. This follows directly from (4.6) and [4, 5].

5.3. Eigenvalues for odd modes. The reduction of the fourth order problem (1.2) to the second order problem does not hold for odd modes which have the characteristic equation (4.10). In fact, all previous theoretical work focused only on the even modes [6, 11]. The same general strategy as for the even case led us to prove the stability of a shifted version of polynomial (5.1).

THEOREM 5.5. *Let $G_n^{(\gamma)}(x)$ denote the nonstandard Gegenbauer polynomial of degree n as defined in Appendix A; then the polynomial*

$$(5.4) \quad p_n^{(\gamma)}(z) = \frac{G_{n-1}^{(\gamma)}(1) - G_{n+1}^{(\gamma)}(1)}{2(n + \gamma)} + \sum_{k=0}^n z^k D^k G_n^{(\gamma)}(1)$$

is stable for $-1/2 < \gamma \leq 1/2$.

The proof is elementary but technical; it is given in Appendix C. We also need the following simple lemma. This lemma will help us determine the sign of the coefficients of the characteristic polynomials.

LEMMA 5.6. *With $G_n^{(\gamma)}(x)$ as defined in Appendix A, the expression*

$$(5.5) \quad D^{k+1}G_n^{(\gamma)}(1) - D^k G_n^{(\gamma)}(1) \geq 0$$

for $k = 0, \dots, n-1$ and $\gamma > -1/2$.

Proof. From (A.13)

$$(5.6) \quad D^k G_n^{(\gamma)}(1) = \frac{2^{k-1} \Gamma(\gamma + k) \Gamma(n + 2\gamma + k)}{(n - k)! \Gamma(\gamma + 1) \Gamma(2\gamma + 2k)},$$

where $\Gamma(z)$ is the standard gamma function. For $k > 0$ and given that $\gamma > -1/2$, the sign of the above expression is positive since all individual terms are positive. In the $k = 0$ case, the sign of the expression is determined by the term $\frac{\Gamma(\gamma)}{\Gamma(2\gamma)}$ since all other terms are positive. If $-1/2 < \gamma < 0$, both numerator and denominator are negative, and thus their ratio is positive. If $\gamma > 0$, the two terms are positive, and thus again their ratio is positive. For $\gamma = 0$ a simple limiting argument shows positiveness again; in fact from (A.14), $G_n^{(0)}(1) = T_n(1)/n = 1/n$.

Taking the ratio $D^{k+1}G_n^{(\gamma)}(1)/D^k G_n^{(\gamma)}(1)$ and making some simplifications gives

$$(5.7) \quad \frac{D^{k+1}G_n^{(\gamma)}(1)}{D^k G_n^{(\gamma)}(1)} = \frac{(2\gamma + n + k)(n - k)}{(2\gamma + 2k + 1)}.$$

Since $k \leq n - 1$, then $2\gamma + 2k + 1 \leq 2\gamma + (n - 1) + k + 1 = 2\gamma + n + k$. Thus

$$(5.8) \quad \frac{D^{k+1}G_n^{(\gamma)}(1)}{D^k G_n^{(\gamma)}(1)} \geq 1, \quad k = 0, \dots, n - 1,$$

and since both derivatives are positive, the lemma follows. \square

We now have all of the tools to prove the following theorem.

THEOREM 5.7. *The Gegenbauer tau approximation to problem (1.2) has real, negative, and distinct eigenvalues for $1/2 < \gamma \leq 7/2$. This γ range is sharp, spurious positive eigenvalues exist for $\gamma < 1/2$, and complex eigenvalues arise for $7/2 < \gamma$.*

Proof. This has already been proven in section 5.2 for even solutions. For odd solutions, we need to consider two separate cases.

Case 1 ($3/2 < \gamma \leq 7/2$). The characteristic polynomial (4.11)

$$(5.9) \quad \sum_{k=0}^{\infty} \mu^k D^{2k} G_n^{(\gamma-2)}(1) - \sum_{k=0}^{\infty} \mu^k D^{2k+1} G_n^{(\gamma-2)}(1) = \Omega_n^{(\gamma-2)}(\mu) - \Theta_n^{(\gamma-2)}(\mu)$$

is a linear combination of the polynomials $\Omega_n^{(\gamma-2)}(\mu)$ and $\Theta_n^{(\gamma-2)}(\mu)$, which form a positive pair for $3/2 < \gamma \leq 7/2$, by Theorem 5.4. Therefore, by Theorem 5.3, this characteristic polynomial has real roots. Then, by Lemma 5.6, we deduce that all of its coefficients are of the same sign, and hence all of its roots must be negative.

Case 2 ($1/2 < \gamma \leq 3/2$). The polynomial

$$(5.10) \quad p_{n-1}^{(\gamma-1)}(z) = \Lambda(z^2) + z\Phi(z^2),$$

with $p_n^{(\gamma)}(z)$ as in Theorem 5.5, is stable for the desired range of parameters by Theorem 5.5, and so the Hermite Biehler theorem (Theorem 5.1) implies that the polynomials

$$(5.11) \quad \Lambda(\mu) = \frac{G_{n-2}^{(\gamma-1)}(1) - G_n^{(\gamma-1)}(1)}{2(n+\gamma-2)} + \sum_{k=0}^{\infty} \mu^k D^{2k} G_{n-1}^{(\gamma-1)}(1) \\ = \frac{G_{n-2}^{(\gamma-1)}(1) - G_n^{(\gamma-1)}(1)}{2(n+\gamma-2)} + \Omega_{n-1}^{(\gamma-1)}(\mu)$$

and

$$(5.12) \quad \Phi(\mu) = \sum_{k=0}^{\infty} \mu^k D^{2k+1} G_{n-1}^{(\gamma-1)}(1) = \Theta_{n-1}^{(\gamma-1)}(\mu)$$

form a positive pair, with $\Omega_n^{(\gamma)}(\mu)$ and $\Theta_n^{(\gamma)}(\mu)$ as defined in Theorem 5.4. Thus $\mu\Phi(\mu)$ and $\Lambda(\mu)$ have interlacing roots, and any real linear combination of the two must have real roots (Theorem 5.3). Now the characteristic polynomial (4.12) is in fact the linear combination $\mu\Phi(\mu) - \Lambda(\mu)$, so it has real roots. Its constant term is equal to

$$(5.13) \quad \frac{G_n^{(\gamma-1)}(1) - G_{n-2}^{(\gamma-1)}(1)}{2(n+\gamma-2)} - G_{n-1}^{(\gamma-1)}(1),$$

which is negative if $-1/2 < \gamma-1 \leq 1/2$, that is, $1/2 < \gamma \leq 3/2$, from (A.11). All other coefficients of the characteristic polynomial $\mu\Phi(\mu) - \Lambda(\mu)$ are negative by Lemma 5.6. Since all coefficients have the same sign and all roots are real, all of the roots must be negative. \square

6. Galerkin and modified tau methods. Our main Theorem 5.7 can be expressed in terms of the inviscid Galerkin and Galerkin methods since these methods are equivalent to Gegenbauer tau methods with index $\gamma+1$ and $\gamma+2$, respectively, as shown in section 2.

COROLLARY 6.1. *The Gegenbauer inviscid Galerkin approximation to problem (1.2) has real, negative, and distinct eigenvalues for $-1/2 < \gamma \leq 5/2$.*

COROLLARY 6.2. *The Gegenbauer Galerkin approximation to problem (1.2) has real negative eigenvalues for $-1/2 < \gamma \leq 3/2$.*

COROLLARY 6.3. *Since Chebyshev corresponds to $\gamma = 0$ and Legendre to $\gamma = 1/2$, the Chebyshev and Legendre tau approximations to problem (1.2) have spurious eigenvalues, but the Chebyshev and Legendre inviscid Galerkin ($\gamma = 1$ and $3/2$, respectively) and Galerkin ($\gamma = 2$ and $5/2$, respectively) approximations provide real, negative, and distinct eigenvalues.*

Finally, we prove that the modified tau method introduced by Gottlieb and Orszag [11] and developed by various authors [8, 16] is equivalent to the Galerkin method. McFadden, Murray, and Boisvert [16] have already shown the equivalence between the modified Chebyshev tau and the Chebyshev Galerkin methods. Our simpler proof generalizes their results to the Gegenbauer class of approximations.

The idea for the modified tau method, widely used for time marching, starts with the substitution $v(x) = D^2u(x)$. Problem (1.2) reads

$$(6.1) \quad D^2u = v, \quad D^2v = \lambda v, \quad \text{with } u = Du = 0 \text{ at } x = \pm 1.$$

If we approximate $u(x)$ by a polynomial of degree n , then $v = D^2u$ suggests that v should be a polynomial of degree $n - 2$; however, the modified tau method approximates both $u(x)$ and $v(x)$ by polynomials of degree n ,

$$(6.2) \quad \begin{aligned} u_n(x) &= \sum_{k=0}^n \hat{u}_k G_k^{(\gamma)}(x), & D^2u_n(x) &= \sum_{k=0}^{n-2} \hat{u}_k^{(2)} G_k^{(\gamma)}(x), \\ v_n(x) &= \sum_{k=0}^n \hat{v}_k G_k^{(\gamma)}(x), & D^2v_n(x) &= \sum_{k=0}^{n-2} \hat{v}_k^{(2)} G_k^{(\gamma)}(x), \end{aligned}$$

where the superscripts indicate the Gegenbauer coefficients of the corresponding derivatives. These can be expressed in terms of the Gegenbauer coefficients of the original function using (A.10) twice, as in the Chebyshev tau method [3, 17]. Hence there are $2n + 2$ coefficients to be determined, $\hat{u}_0, \dots, \hat{u}_n$ and $\hat{v}_0, \dots, \hat{v}_n$. In the modified tau method, these are determined by the four boundary conditions and the $2n - 2$ tau equations obtained from orthogonalizing the residuals of both equations $D^2u = v$ and $D^2v = \lambda v$ to the first $n - 1$ Gegenbauer polynomials $G_0^{(\gamma)}(x), \dots, G_{n-2}^{(\gamma)}(x)$ with respect to the Gegenbauer weight $(1 - x^2)^{\gamma-1/2}$. In terms of the expansions (6.2), these weighted residual equations have the simple form

$$(6.3) \quad \begin{aligned} \hat{u}_k^{(2)} &= \hat{v}_k, & 0 \leq k \leq n - 2, \\ \hat{v}_k^{(2)} &= \lambda \hat{v}_k, & 0 \leq k \leq n - 2. \end{aligned}$$

McFadden, Murray, and Boisvert [16] showed that the modified Chebyshev tau is equivalent to the Chebyshev Galerkin method for this particular problem. We provide a simpler proof for the general setting of Gegenbauer polynomials.

THEOREM 6.4. *The modified Gegenbauer tau method proposed in [11] is equivalent to the Gegenbauer Galerkin method for problem (1.2).*

Proof. Let the polynomial approximations and their derivatives be as in (6.2). Then the tau equations (6.3) are equivalent to the residual equations

$$(6.4) \quad \begin{aligned} v_n(x) - D^2u_n(x) &= \hat{v}_{n-1} G_{n-1}^{(\gamma)}(x) + \hat{v}_n G_n^{(\gamma)}(x), \\ (\lambda - D^2)v_n(x) &= \lambda \hat{v}_{n-1} G_{n-1}^{(\gamma)}(x) + \lambda \hat{v}_n G_n^{(\gamma)}(x). \end{aligned}$$

Combining the two yields

$$(6.5) \quad (\lambda - D^2) D^2 u_n(x) = \hat{v}_{n-1} D^2 G_{n-1}^{(\gamma)}(x) + \hat{v}_n D^2 G_n^{(\gamma)}(x), \quad u_n(\pm 1) = Du_n(\pm 1) = 0,$$

which using (A.5) is equivalent to

$$(6.6) \quad (\lambda - D^2) D^2 u_n(x) = \tau_0 \lambda G_{n-2}^{(\gamma+2)}(x) + \tau_1 \lambda G_{n-3}^{(\gamma+2)}(x), \quad u_n(\pm 1) = Du_n(\pm 1) = 0,$$

with $\hat{v}_n = 4\lambda(\gamma+1)(\gamma+2)\tau_0$ and $\hat{v}_{n-1} = 4\lambda(\gamma+1)(\gamma+2)\tau_1$. This is exactly the Gegenbauer Galerkin method as given in (2.5). \square

Since the modified Gegenbauer tau method is equivalent to the Galerkin method, the results in section 5 imply that the modified tau method for problem 1.2 has real and negative eigenvalues for $-1/2 < \gamma \leq 3/2$. This includes Chebyshev for $\gamma = 0$ and Legendre for $\gamma = 1/2$.

Appendix A. Gegenbauer (ultraspherical) polynomials. The Gegenbauer (a.k.a. ultraspherical) polynomials $C_n^{(\gamma)}(x)$, $\gamma > -1/2$, of degree n , are the Jacobi polynomials, with $\alpha = \beta = \gamma - 1/2$, up to normalization [1, 22.5.20]. They are symmetric (even for n even and odd for n odd) orthogonal polynomials with weight function $W^{(\gamma)}(x) = (1-x^2)^{\gamma-1/2}$. Since the standard normalization [1, 22.3.4] is singular for the Chebyshev case $\gamma = 0$, we use a nonstandard normalization that includes the Chebyshev case but preserves the simplicity of the Gegenbauer recurrences. Set

$$(A.1) \quad G_0^{(\gamma)}(x) := 1, \quad G_n^{(\gamma)}(x) := \frac{C_n^{(\gamma)}(x)}{2^\gamma}, \quad n \geq 1.$$

These Gegenbauer polynomials satisfy the orthogonality relationship

$$(A.2) \quad \int_{-1}^1 (1-x^2)^{\gamma-1/2} G_m^{(\gamma)} G_n^{(\gamma)} dx = \begin{cases} 0, & m \neq n, \\ h_n^\gamma, & m = n, \end{cases}$$

where [1, 22.2.3]

$$(A.3) \quad h_0^\gamma = \frac{\pi 2^{-2\gamma} \Gamma(2\gamma+1)}{\Gamma^2(\gamma+1)} \quad \text{and} \quad h_n^\gamma = \frac{\pi 2^{-1-2\gamma} \Gamma(n+2\gamma)}{(n+\gamma)n! \Gamma^2(\gamma+1)}, \quad n \geq 1.$$

The Sturm–Liouville form of the Gegenbauer equation for $G_n^{(\gamma)}(x)$ is

$$(A.4) \quad D \left[(1-x^2)^{\gamma+1/2} D G_n^{(\gamma)}(x) \right] = -n(n+2\gamma) (1-x^2)^{\gamma-1/2} G_n^{(\gamma)}(x).$$

They satisfy the derivative recurrence formula

$$(A.5) \quad \frac{d}{dx} G_{n+1}^{(\gamma)} = 2(\gamma+1) G_n^{(\gamma+1)}$$

(for $C_n^{(\gamma)}$ this is formula [2, A.57]), and their three-term recurrence takes the simple form

$$(A.6) \quad (n+1) G_{n+1}^{(\gamma)} = 2(n+\gamma)x G_n^{(\gamma)} - (n-1+2\gamma) G_{n-1}^{(\gamma)}, \quad n \geq 2,$$

with

$$(A.7) \quad G_0^{(\gamma)}(x) = 1, \quad G_1^{(\gamma)}(x) = x, \quad G_2^{(\gamma)} = (\gamma+1)x^2 - \frac{1}{2}.$$

These recurrences can be used to verify the odd-even symmetry of Gegenbauer polynomials [1, 22.4.2]

$$(A.8) \quad G_n^{(\gamma)}(x) = (-1)^n G_n^{(\gamma)}(-x).$$

Differentiating the recurrence (A.6) with respect to x and subtracting from the corresponding recurrence for $\gamma + 1$ using (A.5) yields [1, 22.7.23]

$$(A.9) \quad (n + \gamma)G_n^{(\gamma)} = (\gamma + 1) \left[G_n^{(\gamma+1)} - G_{n-2}^{(\gamma+1)} \right], \quad n \geq 3.$$

Combined with (A.5), this leads to the important derivative recurrence between Gegenbauer polynomials of same index γ

$$(A.10) \quad \begin{aligned} G_0^{(\gamma)}(x) &= DG_1^{(\gamma)}(x), \quad 2(1 + \gamma)G_1^{(\gamma)}(x) = DG_2^{(\gamma)}(x), \\ 2(n + \gamma)G_n^{(\gamma)} &= \frac{d}{dx} \left[G_{n+1}^{(\gamma)} - G_{n-1}^{(\gamma)} \right]. \end{aligned}$$

Evaluating the Gegenbauer polynomial at $x = 1$, we find [1, 22.4.2]

$$(A.11) \quad \begin{aligned} G_n^{(\gamma)}(1) &= \frac{1}{2\gamma} C_n^{(\gamma)}(1) = \frac{1}{2\gamma} \binom{2\gamma + n - 1}{n} = \frac{(2\gamma + n - 1)(2\gamma + n - 2) \cdots (2\gamma + 1)}{n!} \\ &= \frac{\Gamma(2\gamma + n)}{n! \Gamma(2\gamma + 1)} \end{aligned}$$

for $n \geq 2$, with $G_1^{(\gamma)}(1) = G_0^{(\gamma)}(1) = 1$, where $\Gamma(z)$ is the standard gamma function [1]. Note that $G_n^{(\gamma)}(1) > 0$ for $\gamma > -1/2$ and that it decreases with increasing n if $-1/2 < \gamma < 1/2$ but increases with n if $1/2 < \gamma$.

Now (A.5) gives

$$(A.12) \quad \frac{d^k G_n^{(\gamma)}}{dx^k}(x) = \frac{2^k \Gamma(\gamma + k + 1)}{\Gamma(\gamma + 1)} G_{n-k}^{(\gamma+k)}(x),$$

which coupled with (A.11) gives

$$(A.13) \quad \frac{d^k G_n^{(\gamma)}}{dx^k}(1) = \frac{2^{k-1} \Gamma(\gamma + k) \Gamma(n + 2\gamma + k)}{(n - k)! \Gamma(\gamma + 1) \Gamma(2\gamma + 2k)}, \quad n \geq 1.$$

Gegenbauer polynomials correspond to Chebyshev polynomials of the first kind $T_n(x)$ when $\gamma = 0$, to Legendre polynomials $P_n(x)$ for $\gamma = 1/2$, and to Chebyshev polynomials of the second kind $U_n(x)$ for $\gamma = 1$. For the nonstandard normalization,

$$(A.14) \quad G_n^{(0)}(x) = \frac{T_n(x)}{n}, \quad G_n^{(1/2)}(x) = P_n(x), \quad G_n^{(1)}(x) = \frac{U_n(x)}{2}.$$

Appendix B. Integrals and asymptotics. As shown in section 3, the tau equation (3.1) provides a matrix eigenproblem of the form $\mu Aa = Ba$. To reduce the coefficients $A(k, l)$ and $B(k, l)$ defined in (3.2) and (3.3) to the expressions (3.4) and

(3.5), respectively, use (A.4) and (A.5) repeatedly:

$$\begin{aligned}
\text{(B.1)} \quad D^2 \left[(1-x^2)^2 G_l^{(5/2)}(x) \right] &= \frac{1}{5} D^2 \left[(1-x^2)^2 D G_{l+1}^{(3/2)}(x) \right] \\
&= -\frac{1}{5} (l+1)(l+4) D \left[(1-x^2) G_{l+1}^{(3/2)}(x) \right] \\
&= -\frac{1}{15} (l+1)(l+4) D \left[(1-x^2) D G_{l+2}^{(1/2)}(x) \right] \\
&= \frac{1}{15} (l+1)(l+2)(l+3)(l+4) G_{l+2}^{(1/2)}(x) \\
&\equiv \mathcal{C}_l G_{l+2}^{(1/2)}(x).
\end{aligned}$$

For the perturbation analysis described in section 3, we need the first order corrections to $B(0, n-4)$ and to $B(1, n-5)$. From (3.12) and (3.13)

$$\text{(B.2)} \quad B_1(0, n-4) = \lim_{\epsilon \rightarrow 0} \frac{\mathcal{C}_{n-4}}{\epsilon} \int_{-1}^1 P_{n-2}(x) (1-x^2)^\epsilon dx,$$

$$\text{(B.3)} \quad B_1(1, n-5) = \lim_{\epsilon \rightarrow 0} \frac{\mathcal{C}_{n-5}}{\epsilon} \int_{-1}^1 x P_{n-3}(x) (1-x^2)^\epsilon dx.$$

To evaluate $\int_{-1}^1 P_n(x) (1-x^2)^\epsilon dx$ to $O(\epsilon)$ for $|\epsilon| \ll 1$ and n even, use (A.4) for $\gamma = 1/2$ and integration by parts to derive

$$\begin{aligned}
\text{(B.4)} \quad \int_{-1}^1 P_n(x) (1-x^2)^\epsilon dx &= \frac{-1}{n(n+1)} \int_{-1}^1 D \left[(1-x^2) D P_n \right] (1-x^2)^\epsilon dx \\
&= \frac{\epsilon}{n(n+1)} \int_{-1}^1 D P_n (1-x^2)^\epsilon (-2x) dx.
\end{aligned}$$

This integral is 0 if n is odd since $P_n(x) = (-1)^n P_n(-x)$. Since we have an ϵ prefactor, we can now set $\epsilon = 0$ in the integral and do the remaining integral by parts to obtain

$$\begin{aligned}
\text{(B.5)} \quad \int_{-1}^1 P_n(x) (1-x^2)^\epsilon dx &\sim \frac{-2\epsilon}{n(n+1)} \int_{-1}^1 x D P_n dx \\
&= \frac{-2\epsilon}{n(n+1)} \int_{-1}^1 (D(x P_n) - P_n) dx \\
&= \frac{-2\epsilon}{n(n+1)} (P_n(1) + P_n(-1)) = \frac{-4\epsilon}{n(n+1)}
\end{aligned}$$

for n even (0 for n odd as it should be).

For the integral in (B.3), use the recurrence (A.6) for $\gamma = 1/2$ to write $(2n-5)xP_{n-3}(x) = (n-2)P_{n-2}(x) + (n-3)P_{n-4}(x)$ and evaluate the resulting two integrals from (B.5). Hence, for n even,

$$\text{(B.6)} \quad B_1(0, n-4) \sim -\frac{4\mathcal{C}_{n-4}}{(n-2)(n-1)},$$

$$\text{(B.7)} \quad B_1(1, n-5) \sim -\frac{4\mathcal{C}_{n-5}}{(n-4)(n-1)}.$$

Now for $A_0(0, n-4)$, $A_0(1, n-5)$, and n even, we have

$$(B.8) \quad A_0(0, n-4) = \mathcal{C}_{n-4} \int_{-1}^1 D^2 P_{n-2}(x) dx = (n-2)(n-1)\mathcal{C}_{n-4},$$

$$(B.9) \quad A_0(1, n-5) = \mathcal{C}_{n-5} \int_{-1}^1 x D^2 P_{n-3}(x) dx = (n-4)(n-1)\mathcal{C}_{n-5}.$$

Appendix C. Proof of Theorem 5.5. Consider

$$(C.1) \quad f_n(x; z) = \sum_{k=0}^{\infty} z^k D^k G_n^{(\gamma)}(x) + K \left(G_n^{(\gamma)}(x) - G_{n+2}^{(\gamma)}(x) \right),$$

where z is a solution of $f_n(1; z) = 0$ and K is

$$(C.2) \quad K = \frac{\left(G_{n-1}^{(\gamma)}(1) - G_{n+1}^{(\gamma)}(1) \right)}{2(n+\gamma) \left(G_n^{(\gamma)}(1) - G_{n+2}^{(\gamma)}(1) \right)} = \dots = \frac{n+2}{2(n+\gamma+1)(n+2\gamma-1)},$$

where we have used (A.11). Note that $f_n(1; z) = p_n^{(\gamma)}(z)$ defined in Theorem 5.5. Taking the x -derivative of $f_n(x; z)$ and using (A.10), we find

$$(C.3) \quad \frac{df_n(x; z)}{dx} = \sum_{k=0}^{\infty} z^k D^{k+1} G_n^{(\gamma)}(x) - 2K(n+\gamma+1)G_{n+1}^{(\gamma)}(x).$$

Thus $f_n(x; z)$ satisfies the following differential equation,

$$(C.4) \quad f_n(x; z) - (1+K)G_n^{(\gamma)}(x) + KG_{n+2}^{(\gamma)}(x) = z \frac{df_n(x; z)}{dx} + z2K(n+\gamma+1)G_{n+1}^{(\gamma)}(x).$$

Multiplying by $(1+x) \frac{df_n^*(x; z)}{dx}$, integrating in the Gegenbauer norm, and adding the complex conjugate, we get

$$(C.5) \quad \int_{-1}^1 \frac{d|f_n|^2}{dx} (1+x)w(x)dx - (1+K) \left(\int_{-1}^1 (1+x) \frac{df_n^*}{dx} G_n^{(\gamma)}(x)w(x)dx + C.C. \right) \\ + K \left(\int_{-1}^1 (1+x) \frac{df_n^*}{dx} G_{n+2}^{(\gamma)}(x)w(x)dx + C.C. \right) = (z+z^*) \int_{-1}^1 \left| \frac{df_n}{dx} \right|^2 (1+x)w(x)dx \\ + \left(z2K(n+\gamma+1) \int_{-1}^1 (1+x) \frac{df_n^*}{dx} G_{n+1}^{(\gamma)}(x)w(x)dx + C.C. \right),$$

where $C.C.$ denotes the complex conjugate. To simplify (C.5) we need to compute four simple integrals:

$$(C.6) \quad \begin{aligned} I_1 &= \int_{-1}^1 \frac{d|f_n|^2}{dx} (1+x)w(x)dx, \\ J_0 &= \int_{-1}^1 (1+x) \frac{df_n^*}{dx} G_n^{(\gamma)}(x)w(x)dx, \\ J_1 &= \int_{-1}^1 (1+x) \frac{df_n^*}{dx} G_{n+1}^{(\gamma)}(x)w(x)dx, \\ J_2 &= \int_{-1}^1 (1+x) \frac{df_n^*}{dx} G_{n+2}^{(\gamma)}(x)w(x)dx. \end{aligned}$$

Using integration by parts, the first integral becomes

$$(C.7) \quad I_1 = \int_{-1}^1 \frac{d|f_n|^2}{dx} (1+x)w(x)dx = - \int_{-1}^1 |f_n|^2 (1-2\gamma x) \frac{w(x)}{1-x} dx.$$

Therefore, the integral is negative for $-1/2 < \gamma \leq 1/2$ since for this range of parameters $1-2\gamma x$ is positive.

For the calculation of the three other integrals, we are going to need the expressions

$$(C.8) \quad \frac{df_n}{dx} = \sum_{k=0}^{\infty} z^k D^{k+1} G_n^{(\gamma)}(x) - 2K(\gamma+n+1)G_{n+1}^{(\gamma)}(x),$$

$$(C.9) \quad = \mathcal{P}_{n-2}(x; z) + 2(\gamma+n-1)G_{n-1}^{(\gamma)}(x) - 2K(\gamma+n+1)G_{n+1}^{(\gamma)}(x),$$

$$(C.10) \quad = \mathcal{P}_{n-1}(x; z) - 2K(\gamma+n+1)G_{n+1}^{(\gamma)}(x),$$

where $\mathcal{P}_{n-2}(x; z)$ and $\mathcal{P}_{n-1}(x; z)$ are polynomials of degree $n-2$ and $n-1$, respectively.

With the use of (C.9) and the orthogonality of the Gegenbauer polynomials, we find

$$(C.11) \quad \begin{aligned} J_0 &= \int_{-1}^1 (1+x) \frac{df_n^*}{dx} G_n^{(\gamma)}(x) w(x) dx \\ &= 2(n-1+\gamma) \int_{-1}^1 x G_{n-1}^{(\gamma)}(x) G_n^{(\gamma)}(x) w(x) dx - 2K(n+1+\gamma) \\ &\quad \int_{-1}^1 x G_{n+1}^{(\gamma)}(x) G_n^{(\gamma)}(x) w(x) dx \\ &= n \int_{-1}^1 \left(G_n^{(\gamma)}(x) \right)^2 w(x) dx - K(n+\gamma+1) \frac{n+1}{n+\gamma} \int_{-1}^1 \left(G_{n+1}^{(\gamma)}(x) \right)^2 w(x) dx \\ &= nh_n^{(\gamma)} - K \frac{(n+\gamma+1)(n+1)}{n+\gamma} h_{n+1}^{(\gamma)}, \end{aligned}$$

where we have also used (A.5) and (A.6). In the same way, we compute

$$(C.12) \quad \begin{aligned} J_2 &= \int_{-1}^1 (1+x) \frac{df_n^*}{dx} G_{n+2}^{(\gamma)}(x) w(x) dx \\ &= -K \int_{-1}^1 2(n+\gamma+1)x G_{n+1}^{(\gamma)}(x) G_{n+2}^{(\gamma)}(x) w(x) dx \\ &= -K(n+2)h_{n+2}^{(\gamma)} \end{aligned}$$

and

$$(C.13) \quad \begin{aligned} J_1 &= \int_{-1}^1 (1+x) \frac{df_n^*}{dx} G_{n+1}^{(\gamma)}(x) w(x) dx \\ &= -2K(n+\gamma+1) \int_{-1}^1 (1+x) \left(G_{n+1}^{(\gamma)}(x) \right)^2 w(x) dx \\ &\quad - 2K(n+\gamma+1) \int_{-1}^1 \left(G_{n+1}^{(\gamma)}(x) \right)^2 w(x) dx \\ &= -2K(n+\gamma+1)h_{n+1}^{(\gamma)}. \end{aligned}$$

Thus (C.5) transforms to

$$(C.14) \quad - \int_{-1}^1 |f_n|^2 \frac{(1-2\gamma x)w(x)}{(1-x)} dx - 2(1+K) \left(nh_n^{(\gamma)} - K \frac{(n+\gamma+1)(n+1)}{n+\gamma} h_{n+1}^{(\gamma)} \right) - 2K^2(n+2)h_{n+2}^{(\gamma)} = (z+z^*) \left(\int_{-1}^1 \left| \frac{df_n}{dx} \right|^2 (1+x)w(x) dx - 4K^2(n+1+\gamma)^2 h_{n+1}^{(\gamma)} \right).$$

Our task is to show that the left-hand side of the above expression is negative, whereas the coefficient of the term $z+z^*$ on the right-hand side is positive. A simple calculation shows that

$$(C.15) \quad nh_n^{(\gamma)} - K \frac{(n+\gamma+1)(n+1)}{(n+\gamma)} h_{n+1}^{(\gamma)} = nh_n^{(\gamma)} - \frac{(n+1)(n+2)}{(n+\gamma)(n+2\gamma-1)} h_{n+1}^{(\gamma)} \\ = \frac{\pi 2^{-1-2\gamma} \Gamma(n+2\gamma)}{\gamma^2 \Gamma^2(\gamma) n! (n+\gamma)} \left(n - \frac{(n+2)(n+2\gamma)}{2(n-1+2\gamma)(n+1+\gamma)} \right) \\ \geq \frac{\pi 2^{-1-2\gamma} \Gamma(n+2\gamma)}{\gamma^2 \Gamma^2(\gamma) n! (n+\gamma)} \left(n - \frac{(n+2)(n+1)}{(n-2)(2n+1)} \right).$$

The last parenthesis is positive for $n \geq 3$.

For the right-hand side, we use the notation in (C.10) to get

$$(C.16) \quad \int_{-1}^1 \left| \frac{df_n}{dx} \right|^2 (1+x)w(x) dx - 4K^2(n+1+\gamma)^2 h_{n+1}^{(\gamma)} \\ = \int_{-1}^1 (1+x) \left(|\mathcal{P}_{n-1}(x; z)|^2 - 2K(n+1+\gamma)G_{n+1}^{(\gamma)}(x) (\mathcal{P}_{n-1}(x; z) + \mathcal{P}_{n-1}^*(x; z)) \right. \\ \left. + 4K^2(n+1+\gamma)^2 (G_{n+1}^{(\gamma)}(x))^2 \right) w(x) dx - 4K^2(n+1+\gamma)^2 h_{n+1}^{(\gamma)} \\ = \int_{-1}^1 (1+x) |\mathcal{P}_{n-1}(x; z)|^2 w(x) dx + 0 + 0 + 4K^2(n+1+\gamma)^2 h_{n+1}^{(\gamma)} - 4K^2(n+1+\gamma)^2 h_{n+1}^{(\gamma)} \\ = \int_{-1}^1 (1+x) |\mathcal{P}_{n-1}(x; z)|^2 w(x) dx > 0.$$

Thus (C.14) becomes

$$(C.17) \quad - \int_{-1}^1 |f_n|^2 \frac{(1-2\gamma x)w(x)}{(1-x)} dx - 2(1+K) \left(nh_n^{(\gamma)} - K \frac{(n+\gamma+1)(n+1)}{n+\gamma} h_{n+1}^{(\gamma)} \right) \\ - 2K^2(n+2)h_{n+2}^{(\gamma)} = (z+z^*) \int_{-1}^1 (1+x) |\mathcal{P}_{n-1}(x; z)|^2 w(x) dx,$$

and $\Re(z) < 0$.

For $n = 2$ we get the following characteristic polynomial:

$$(C.18) \quad f_2^{(\gamma)}(1; z) = \frac{2}{3}(\gamma+1)(3z^2 + 3z + 1),$$

whose zeros

$$(C.19) \quad z_1 = -\frac{1}{2} - \frac{\sqrt{3}i}{6}, \quad z_2 = -\frac{1}{2} + \frac{\sqrt{3}i}{6}$$

have negative real parts for any γ .

Acknowledgment. We thank Jue Wang for helpful calculations in the early stages of this work.

REFERENCES

- [1] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of Mathematical Functions*, Dover, New York, 1965.
- [2] J. P. BOYD, *Chebyshev and Fourier Spectral Methods*, Dover, New York, 2001.
- [3] C. CANUTO, M. Y. HUSSAINI, A. QUARTERONI, AND T. A. ZANG, *Spectral Methods in Fluid Dynamics*, Springer, New York, 1988.
- [4] M. CHARALAMBIDES AND F. WALEFFE, *Spectrum of the Jacobi tau approximation for the second derivative operator*, SIAM J. Numer. Anal., 46 (2008), pp. 280–294.
- [5] G. CSORDAS, M. CHARALAMBIDES, AND F. WALEFFE, *A new property of a class of Jacobi polynomials*, Proc. Amer. Math. Soc., 133 (2005), pp. 3551–3560.
- [6] P. T. DAWKINS, S. R. DUNBAR, AND R. W. DOUGLASS, *The origin and nature of spurious eigenvalues in the spectral tau method*, J. Comput. Phys., 147 (1998), pp. 441–462.
- [7] P. DRAZIN AND W. H. REID, *Hydrodynamic Stability*, Cambridge University Press, Cambridge, UK, 1981.
- [8] D. R. GADNER, S. A. TROGDON, AND R. W. DOUGLASS, *A modified tau spectral method that eliminates spurious eigenvalues*, J. Comput. Phys., 80 (1989), pp. 137–167.
- [9] D. GOTTLIEB, *The stability of pseudospectral-Chebyshev methods*, Math. Comp., 36 (1981), pp. 107–118.
- [10] D. GOTTLIEB AND L. LUSTMAN, *The spectrum of the Chebyshev collocation operator for the heat equation*, SIAM J. Numer. Anal., 20 (1983), pp. 909–921.
- [11] D. GOTTLIEB AND S. A. ORSZAG, *Numerical Analysis of Spectral Methods: Theory and Applications*, SIAM, Philadelphia, 1977.
- [12] O. HOLTZ, *Hermite-Biehler, Routh-Hurwitz, and total positivity*, Linear Algebra Appl., 372 (2003), pp. 105–110.
- [13] W. HUANG AND D. M. SLOAN, *The pseudospectral method for solving differential eigenvalue problems*, J. Comput. Phys., 111 (1994), pp. 399–409.
- [14] J. KIM, P. MOIN, AND R. MOSER, *Turbulence statistics in fully developed channel flow at low Reynolds number*, J. Fluid Mech., 177 (1987), pp. 133–166.
- [15] B. J. LEVIN, *Distribution of Zeros of Entire Functions*, AMS, Providence, 1980.
- [16] G. B. MCFADDEN, B. T. MURRAY, AND R. F. BOISVERT, *Elimination of spurious eigenvalues in the Chebyshev tau spectral method*, J. Comput. Phys., 91 (1990), pp. 228–239.
- [17] S. A. ORSZAG, *Accurate solution of the Orr-Sommerfeld stability equation*, J. Fluid Mech., 50 (1971), pp. 689–703.
- [18] Q. I. RAHMAN AND G. SCHMEISSER, *Analytic Theory of Polynomials*, Oxford University Press, New York, 2002.
- [19] F. WALEFFE, *Exact coherent structures in channel flow*, J. Fluid Mech., 435 (2001), pp. 93–102.
- [20] F. WALEFFE, *Homotopy of exact coherent structures in plane shear flows*, Phys. Fluids, 15 (2003), pp. 1517–1534.
- [21] J. WANG, J. GIBSON, AND F. WALEFFE, *Lower branch coherent states in shear flows: Transition and control*, Phys. Rev. Lett., 98 (2007), p. 204501.
- [22] A. ZEBIB, *Removal of spurious modes encountered in solving stability problems by spectral methods*, J. Comput. Phys., 70 (1987), pp. 521–525.

NUMERICAL ALGORITHM FOR CALCULATING THE GENERALIZED MITTAG-LEFFLER FUNCTION*

HANSJÖRG SEYBOLD[†] AND RUDOLF HILFER[‡]

Abstract. A numerical algorithm for calculating the generalized Mittag-Leffler $E_{\alpha,\beta}(z)$ function for arbitrary complex argument z and real parameters $\alpha > 0$ and $\beta \in \mathbb{R}$ is presented. The algorithm uses the Taylor series, the exponentially improved asymptotic series, and integral representations to obtain optimal stability and accuracy of the algorithm. Special care is applied to the limits of validity of the different schemes to avoid instabilities in the algorithm.

Key words. special functions of mathematical physics, fractional calculus, generalized Mittag-Leffler functions, numerical algorithms

AMS subject classifications. 65D15, 65D20, 33E12, 30E10

DOI. 10.1137/070700280

1. Introduction. The (generalized) Mittag-Leffler function $E_{\alpha,\beta}(z)$ is an entire function with two parameters α and β . The function $E_{\alpha,1}(z)$ is named after Mittag-Leffler, who introduced it in 1903 in a publication on Laplace–Abel integrals [10, 11, 12]. Shortly after its introduction it was generalized by Wiman [21].

The generalized Mittag-Leffler function with a nonnegative argument is completely monotone if and only if $0 < \alpha \leq 1$, $\operatorname{Re}(\beta) \geq \alpha$ [18].

In the special case $\beta = 1$ and $0 < \alpha \leq 1$, complete monotonicity was already conjectured by Feller and later proved by Pollard in 1948 [2, 16]. Analytical investigations on the distribution of the zeros in the complex plane were published by Wiman [21, 22], and numerical results are given in [7, 19].

In recent years fractional calculus has become a popular topic in physics and engineering [8]. Fractional equations of motion are widely accepted for describing viscoelasticity and anomalous diffusion [5, 6, 8, 9, 24]. The generalized Mittag-Leffler function plays a central role in fractional calculus and its applications because it is closely related to the eigenfunction of the fractional derivative operator [8, 17]. Progress in this field requires the calculation of the exact numerical values of the generalized Mittag-Leffler function for arbitrary complex arguments and the study of its properties.

The Mittag-Leffler function shows very different behaviors in the complex plane with varying parameters α and β , so different approaches have to be used for different regions in the complex plane. We present a stable and robust numerical method based on recursion relations, exponentially improved asymptotics, and integral representations for calculating the generalized Mittag-Leffler function for real parameters $\alpha > 0$ and β and arbitrary complex argument z .

The article is organized as follows: First, we give an overview of the formulas used for the calculation in section 2 and introduce the partitioning of the complex plane in section 3. Then, we present the error estimates starting with the series expansions

*Received by the editors August 17, 2007; accept for publication (in revised form) June 12, 2008; published electronically October 24, 2008.

<http://www.siam.org/journals/sinum/47-1/70028.html>

[†]Computational Physics for Engineering Materials, IfB, ETH Zurich, 8093 Zurich, Switzerland (hseybold@ethz.ch).

[‡]ICP, Universität Stuttgart, Pfaffenwaldring 27, 70569 Stuttgart, Germany, and Institut für Physik, Universität Mainz, 55099 Mainz, Germany (hilfer@icp.uni-stuttgart.de).

(section 4). Finally, we discuss the numerical details of the algorithm in section 5 and present the results of extensive numerical calculations in the last section (section 6), comparing the speed and stability of the algorithm.

2. Overview. This section gives an overview of the numerical algorithm. The generalized Mittag-Leffler function is an entire function defined by the following power series:

$$(2.1) \quad E_{\alpha,\beta}(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k + \beta)}$$

for $z \in \mathbb{C}$, $\alpha > 0$, and $\beta \in \mathbb{R}$. If $|z| < 1$, finitely many terms of the power series are sufficient to approximate $E_{\alpha,\beta}(z)$ with arbitrary precision. The error estimates are given in section 4.

The case $\alpha \geq 1$ will be reduced to the case $\alpha < 1$ using the following recursion relation [2, 15]:

$$(2.2) \quad E_{\alpha,\beta}(z) = \frac{1}{2m+1} \sum_{h=-m}^m E_{\alpha/(2m+1),\beta} \left(z^{1/(2m+1)} e^{2\pi i h/(2m+1)} \right),$$

which is valid for all $\alpha, \beta \in \mathbb{R}$, $z \in \mathbb{C}$, and $m = [(\alpha - 1)/2] + 1$. Here $[x]$ denotes the largest integer less than or equal to x .

For large values $|z| \gg 1$ the exponentially improved asymptotic series with the Berry-type smoothing transition gives a fast approximation of the Mittag-Leffler function. If $|\arg(z)| > \pi\alpha$, the asymptotic series is given by [2]

$$(2.3) \quad E_{\alpha,\beta}(z) \sim - \sum_{k=1}^{\infty} \frac{z^{-k}}{\Gamma(\beta - \alpha k)},$$

while for $|\arg(z)| < \pi\alpha$ we find the following series:

$$(2.4) \quad E_{\alpha,\beta}(z) \sim \frac{1}{\alpha} z^{(1-\beta)/\alpha} e^{z^{1/\alpha}} - \sum_{k=1}^{\infty} \frac{z^{-k}}{\Gamma(\beta - \alpha k)}.$$

In the transition area around the Stokes lines $|\arg(z)| < \pi\alpha \pm \delta$, with $\delta < \pi\alpha/2$, we have the Berry-type smoothing given by

$$(2.5) \quad E_{\alpha,\beta}(z) \sim \frac{1}{2\alpha} z^{(1-\beta)/\alpha} e^{z^{1/\alpha}} \operatorname{erfc} \left(-c(\theta) \sqrt{|z|^{1/\alpha}/2} \right) - \sum_{k=1}^{\infty} \frac{z^{-k}}{\Gamma(\beta - \alpha k)}$$

around the lower Stokes line for $-3\pi\alpha/2 < \arg(z) < \pi\alpha/2$, where the parameter c is given by the relation $\frac{1}{2}c^2 = 1 + i\theta - e^{i\theta}$, with $\theta = \arg(z^{1/\alpha}) + \pi$ and the principle branch of c is chosen such that $c \approx \theta + \frac{i}{6}\theta^2 - \frac{1}{36}\theta^3$ for small θ [23]. Around the upper Stokes line $\pi\alpha/2 < \arg(z) < 3\pi\alpha/2$, one finds that

$$(2.6) \quad E_{\alpha,\beta}(z) \sim \frac{1}{2\alpha} z^{(1-\beta)/\alpha} e^{z^{1/\alpha}} \operatorname{erfc} \left(c(\theta) \sqrt{|z|^{1/\alpha}/2} \right) - \sum_{k=1}^{\infty} \frac{z^{-k}}{\Gamma(\beta - \alpha k)},$$

with $\frac{1}{2}c^2 = 1 + i\theta - e^{i\theta}$, $\theta = \arg(z^{1/\alpha}) - \pi$, and the same condition as before for small θ . This exponentially improved asymptotic series converges very rapidly for most values of $z \in \mathbb{C}$, with $|z| > 1$. The Stokes phenomenon and the Berry-type

smoothing for the Mittag-Leffler function have been investigated recently in [13, 23].

In an intermediate range $r_0 \leq |z| \leq r_1$ between the Taylor series and the asymptotic expansion, we use Wiman's integral representation [2, p. 210]

$$(2.7) \quad E_{\alpha,\beta}(z) = \frac{1}{2\pi i} \int_{\mathcal{C}} \frac{y^{\alpha-\beta} e^y}{y^\alpha - z} dy$$

for calculating $E_{\alpha,\beta}(z)$. The values r_0 where the Taylor series ends and r_1 where the asymptotic series starts will be specified later. The path of integration \mathcal{C} in the complex plane starts and ends at $-\infty$ and encircles the circular disc $|y| \leq |z|^{1/\alpha}$ in the positive sense. Equation (2.7) can be obtained by inserting the Hankel representation of the inverse Γ -function into the Taylor series and bending the contour in the complex plane.

3. Partitioning of the complex plane. In different regions of the complex plane the Mittag-Leffler function shows different kinds of behavior. Therefore different calculation schemes are needed for different regions. The following definitions are used to describe these regions. First,

$$(3.1) \quad \overline{\mathbb{D}(r)} = \{z \in \mathbb{C} : |z| \leq r\}$$

is the closure of the open disk $\mathbb{D}(r) = \{z \in \mathbb{C} : |z| < r\}$ of radius r centered at the origin. Next, we define the wedges

$$(3.2) \quad \mathbb{W}(\phi_1, \phi_2) = \{z \in \mathbb{C} : \phi_1 < |\arg(z)| < \phi_2\},$$

$$(3.3) \quad \overline{\mathbb{W}}(\phi_1, \phi_2) = \{z \in \mathbb{C} : \phi_1 \leq |\arg(z)| \leq \phi_2\},$$

where $\phi_2 - \phi_1$ is the opening angle measured in a positive sense and $\phi_1, \phi_2 \in (-\pi, \pi)$. On a disc

$$(3.4) \quad \mathbb{G}_0 = \overline{\mathbb{D}(r_0)}$$

of radius $r_0 < 1$, the Taylor series (2.1) gives a good approximation of the generalized Mittag-Leffler function. For the algorithm we choose $r_0 = 0.95$.

For large values of $|z| \in \mathbb{C}$ exponentially improved asymptotics can be used to calculate $E_{\alpha,\beta}(z)$. Equation (2.4) is used for z in

$$(3.5) \quad \mathbb{G}_1 = [\mathbb{C} \setminus \mathbb{D}(r_1)] \cap \mathbb{W}(-\pi\alpha + \delta, \pi\alpha - \delta)$$

and (2.3) for z in

$$(3.6) \quad \mathbb{G}_2 = [\mathbb{C} \setminus \mathbb{D}(r_1)] \cap \mathbb{W}(\pi\alpha + \tilde{\delta}, -\pi\alpha - \tilde{\delta}),$$

where $r_1 > r_0$ will be defined in (4.21) and δ and $\tilde{\delta}$ are numbers smaller than $\pi\alpha/2$. In the algorithm δ and $\tilde{\delta}$ are chosen to be

$$(3.7) \quad \delta = \pi\alpha/8 \quad \text{and} \quad \tilde{\delta} = \min\{\pi\alpha/8, (\pi + \pi\alpha)/2\}.$$

Close to the Stokes lines $|\arg(z)| = \pi\alpha$, the approximation scheme with the series (2.3) and (2.4) becomes unstable, so the Berry-type smoothed asymptotic series (2.6) is used in the area around the upper Stokes line

$$(3.8) \quad \mathbb{G}_3 = [\mathbb{C} \setminus \mathbb{D}(r_1)] \cap \overline{\mathbb{W}}(\pi\alpha - \delta, \pi\alpha + \tilde{\delta})$$

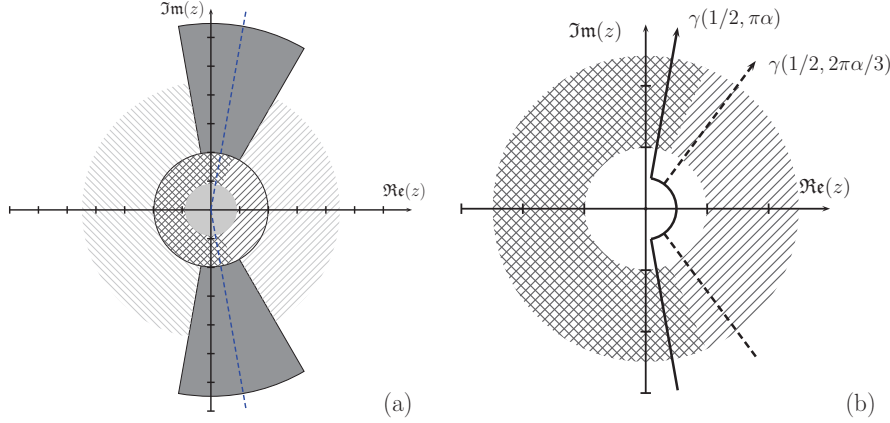


FIG. 1. Figure (a) shows the partitioning of the complex plane for calculating the generalized Mittag-Leffler function $E_{\alpha,\beta}(z)$. The thick dashed lines mark the Stokes line at $\arg(z) = \pm\pi\alpha$. In the central light gray region \mathbb{G}_0 , finitely many terms of the Taylor series (2.1) give a good approximation for the generalized Mittag-Leffler function. The black crosshatched region marks \mathbb{G}_6 (left) and the black hatched one \mathbb{G}_5 (right), where the integral representations (4.31), (4.32) and (4.25), (4.26), respectively, are used for the calculations. The asymptotic series (2.4) and (2.3) are used in the gray hatched regions \mathbb{G}_1 (left) and \mathbb{G}_2 (right). Berry-type smoothing is applied in the gray solid areas: (2.6) in the upper region \mathbb{G}_3 and (2.5) in the lower one \mathbb{G}_4 . In (b) the integration contours for the two cases $\vartheta = \pi\alpha$ (solid line) and $\vartheta = 2\pi\alpha/3$ (dashed line) are shown. While (4.25) and (4.26) are used in the crosshatched region corresponding to (\mathbb{G}_6) , (4.31) and (4.32) are used for the part where the hatching runs at 45° (\mathbb{G}_5).

and (2.5) close to the lower one

$$(3.9) \quad \mathbb{G}_4 = [\mathbb{C} \setminus \mathbb{D}(r_1)] \cap \overline{\mathbb{W}}(-\pi\alpha - \tilde{\delta}, -\pi\alpha + \delta).$$

In the transition area between the Taylor series and the asymptotic expansion, we make use of Hankel's integral representation for the generalized Mittag-Leffler function (2.7). We employ two different contour paths to avoid numerical problems which arise in other algorithms [4] when z is too close to the contour path. The two regions \mathbb{G}_5 and \mathbb{G}_6 are defined by

$$(3.10) \quad \mathbb{G}_5 = \mathbb{D}(r_1) \cap \overline{\mathbb{W}}(-5\pi\alpha/6, 5\pi\alpha/6) \setminus \mathbb{G}_0,$$

$$(3.11) \quad \mathbb{G}_6 = \mathbb{D}(r_1) \cap \mathbb{W}(5\pi\alpha/6, -5\pi\alpha/6) \setminus \mathbb{G}_0,$$

where $\overline{\mathbb{W}}(-5\pi\alpha/6, 5\pi\alpha/6)$ and $\mathbb{W}(5\pi\alpha/6, -5\pi\alpha/6)$ are defined as in (3.3) and (3.2), respectively. The different regions are shown in Figure 1.

4. Error estimates.

4.1. Taylor series. For $z \in \mathbb{G}_0$ finitely many terms of the Taylor series are sufficient to approximate the generalized Mittag-Leffler function for arbitrary $\alpha > 0$ and $\beta \in \mathbb{R}$. The maximum number of terms N taken into account is chosen such that the error

$$(4.1) \quad R_N(z) = \left| E_{\alpha,\beta}(z) - \sum_{k=0}^N \frac{z^k}{\Gamma(\alpha k + \beta)} \right| \leq \varepsilon$$

is smaller than a given accuracy ε .

THEOREM 4.1. Let $\varepsilon > 0$. If $|z| < 1$ and

$$(4.2) \quad N \geq \max \left\{ \left[(2 - \beta)/\alpha \right] + 1, \left[\frac{\ln(\varepsilon(1 - |z|))}{\ln(|z|)} \right] + 1 \right\},$$

then the error term $R_N(z) = \left| \sum_{k=N}^{\infty} \frac{z^k}{\Gamma(\alpha k + \beta)} \right|$ is smaller than the given accuracy ε .

We have $1/\Gamma(x) > 1$ for all $x > 2$. Setting $x = \alpha k + \beta$ one finds that $\frac{1}{\Gamma(\alpha N + \beta)} \leq 1$ for all $N \geq \left[\frac{(2 - \beta)}{\alpha} \right] + 1$, where $[a]$ is the smallest integer larger than a ; thus

$$(4.3) \quad R_N(z) = \left| \sum_{k=N}^{\infty} \frac{z^k}{\Gamma(\alpha k + \beta)} \right| \leq \left| \sum_{k=N}^{\infty} z^k \right| \leq \sum_{k=N}^{\infty} |z^k|.$$

Under the condition $|z| < 1$ the geometric series can be summed up, and one obtains together with (4.1) that

$$(4.4) \quad R_N(z) \leq \sum_{k=N}^{\infty} |z^k| = \frac{|z|^N}{1 - |z|} \stackrel{!}{\leq} \varepsilon \quad \text{and therefore} \quad N \geq \frac{\ln(\varepsilon(1 - |z|))}{\ln(|z|)}. \quad \square$$

For the algorithm we have chosen

$$(4.5) \quad M = \max \left\{ \left[\frac{(2 - \beta)}{\alpha} \right] + 1, \left[\frac{\ln(\varepsilon(1 - |z|))}{\ln(|z|)} \right] + 1 \right\}$$

for the maximum numbers of coefficients that have to be taken into account to reach an accuracy of ε .

4.2. Exponentially improved asymptotics. In this section we give an estimate of the error term of the asymptotic series. To apply the asymptotic series expansions for a given accuracy ε , two parameters have to be determined, which are the truncation point N of the series and a lower limit r_1 of $|z|$ such that the error is smaller than ε for $|z| > r_1$.

THEOREM 4.2. Let $\alpha \in (0, 1)$. For $N \approx \frac{1}{\alpha} |z|^{1/\alpha}$ and $|z| \geq (-2 \log \frac{\varepsilon}{C})^\alpha$ the estimate of the error term of the asymptotic series (2.3) and (2.4) fulfills the condition

$$(4.6) \quad R_N(z) = \left| E_{\alpha,\beta}(z) - \left(- \sum_{k=1}^{N-1} \frac{z^{-k}}{\Gamma(\beta - \alpha k)} \right) \right| \leq \varepsilon,$$

where C is a constant dependent only on α, β , and

$$(4.7) \quad R_N(z) = \frac{z^{N-1}}{2\pi i} \int_C \frac{t^{N\alpha - \beta} e^t}{t^\alpha - z} dt.$$

C is the classical Hankel contour, which runs along the positive real axis in the upper half-plane starting from $+\infty + i0$ encircling the origin and returning to $+\infty - i0$ along the positive real axis in the lower half-plane.

Without loss of generality we can assume that $\alpha N - \beta > -1$ by choosing N large enough. Then the circle can be shrunk to zero, and there remain two rays, above and below the cut of the complex plane along the negative real axis. Inserting $t = v e^{-\pi v}$ for the upper part of the contour and $t = v e^{\pi v}$ for the lower one, we obtain

$$(4.8) \quad R_N(z) = L_N(z) + U_N(z),$$

where

$$(4.9) \quad L_N(z) = e^{i\pi\beta} (z e^{i\pi\alpha})^{-N+1} \frac{1}{2\pi i} \int_0^\infty \frac{v^{\alpha N - \beta} e^{-v}}{v^\alpha - z e^{i\pi\alpha}} dv$$

and

$$(4.10) \quad U_N(z) = -e^{-i\pi\beta} (z e^{-i\pi\alpha})^{-N+1} \frac{1}{2\pi i} \int_0^\infty \frac{v^{\alpha N - \beta} e^{-v}}{v^\alpha - z e^{-i\pi\alpha}} dv$$

are the integrals along the upper and lower parts, respectively, of the remaining contour path. Following the arguments in [23] we obtain

$$(4.11) \quad U_N(z) \leq \frac{1}{2\pi \sin \alpha\pi} |z|^{-N} \Gamma(\alpha N - \beta + 1)$$

and similarly

$$(4.12) \quad L_N(z) \leq \frac{1}{2\pi \min\{\sin \alpha\pi, \sin \xi\}} |z|^{-N} \Gamma(\alpha N - \beta + 1),$$

where $\arg(z) \in (-\pi, -\pi\alpha - \xi)$ for $\xi > 0$. The angle ξ describes the distance of z to the Stokes line. Applying the argument of Boyd [1] yields

$$(4.13) \quad L_N(z) \leq C_1(\alpha, \beta) |z|^{-N} \Gamma(\alpha N - \beta + 1) \sqrt{N},$$

where C_1 now depends only on α and β . Combining (4.11) and (4.13) yields

$$(4.14) \quad R_N(z) \leq C_2(\alpha, \beta) |z|^{-N} \Gamma(\alpha N - \beta + 1) \sqrt{N},$$

where we used the fact that $N > 1$ and hence $\sqrt{N} > 1$.

To obtain an estimate for N and a minimal $|z|$, further approximations have to be applied. Using Stirling's formula for $R_N(z)$, (4.14) yields

$$(4.15) \quad \begin{aligned} R_N(z) &\leq C_2(\alpha, \beta) |z|^{-N} \Gamma(\alpha N - \beta + 1) \sqrt{N} \\ &= C_2(\alpha, \beta) |z|^{-N} (\alpha N - \beta) \Gamma(\alpha N - \beta) \sqrt{N} \\ &\leq C_3(\alpha, \beta) |z|^{-N} \sqrt{N} (\alpha N - \beta)^{(\alpha N - \beta + 3/2)} e^{-(\alpha N - \beta)}, \end{aligned}$$

with constant C_3 depending only on α and β . For $\beta \geq 0$ one has $(\alpha N - \beta) \leq \alpha N$. For $\beta < 0$ one obtains $(\alpha N - \beta)^\gamma \leq C'(\alpha N)^\gamma$ by applying the binomial series. Here γ was used as an abbreviation for $(\alpha N - \beta + 3/2)$. Combining these estimates with (4.15) we arrive at

$$(4.16) \quad R_N(z) \leq C_4(\alpha, \beta) (\alpha N)^{1-\beta} e^{N(-\alpha + \alpha \log(\alpha N) - \log |z|)},$$

where we applied the trivial identity $\sqrt{N} = (\alpha N)^{1/2} / \sqrt{\alpha}$ and C_4 absorbs all constants from the approximations. An optimal truncation should be obtained when

$$(4.17) \quad N \approx \frac{1}{\alpha} |z|^{1/\alpha},$$

which yields for real α and β

$$(4.18) \quad R_N(z) \leq C_4(\alpha, \beta)(\alpha N)^{1-\beta} e^{-N\alpha}.$$

Assuming that $|z| > 1$ yields $(|z|^{1/\alpha})^{1-\beta} \leq (|z|^{1/\alpha})^{|1-\beta|}$. Now we apply the inequality $x^y \leq (qy)^y e^{x/q}$, with $x, y, q > 0$ (see Theorem 4.4 below) for $q = 1/2$, $x = |z|^{1/\alpha}$, and $y = |1 - \beta|$. Thus the estimate

$$(4.19) \quad R_N(z) \leq C e^{-\frac{1}{2}|z|^{1/\alpha}}$$

holds for all α, β , and $|z| > 1$, where C is given by $C = C_4(\alpha, \beta)(1/2 \cdot |1 - \beta|)^{|1-\beta|}$. This can now be easily solved for $|z|$ to determine r_1 if we assume that the error $R_N(z)$ is smaller than a given error ε for $|z| > r_1$. Thus we obtain

$$(4.20) \quad |z| \geq r_1 = \left(-2 \log \frac{\varepsilon}{C}\right)^\alpha,$$

which yields the following conditions for the exponential asymptotics:

$$(4.21) \quad M = \left\lceil \frac{1}{\alpha} |z|^{1/\alpha} \right\rceil + 1, \quad r_1 = \left(-2 \log \frac{\varepsilon}{C}\right)^\alpha,$$

where $\lceil x \rceil$ is the smallest integer larger than x . The parameter M denotes the number of terms in the series that are taken into account to achieve the given accuracy. \square

Finally it is important to mention that the use of the Berry-type smoothing in the region close to the Stokes lines avoids numerical problems of the asymptotic series in the transition area between \mathbb{G}_2 (with (2.3)) and \mathbb{G}_3 (with (2.4)) and gives a much faster convergence due to the exponential corrections of the series.

4.3. Integral representation.

4.3.1. Basic formulas. For the area between the Taylor series and the exponentially improved asymptotics, Hankel's integral representation can be used for calculating the values of $E_{\alpha,\beta}(z)$. Another approach for calculating the Mittag-Leffler function was suggested in [4], but that algorithm neglects the numerical difficulties that arise for values of z close to the contour path. Also, it does not make use of the exponentially improved asymptotics and hence is much slower. We avoid the problems of [4] near the contour path by using two different integration formulas with different contours. In the following paragraph the integration formula will be derived from Hankel's integral representation by inserting the contour path into the definition, and the behavior of the integrands will be discussed. Finally we present the approximations and the error estimates of the integrals used in the numerical scheme.

We start from the classical Hankel representation of the generalized Mittag-Leffler function which is obtained by inserting the Hankel representation of the inverse Gamma function in the series expansion (2.1). The integral representation is given by [2, p. 210]

$$(4.22) \quad E_{\alpha,\beta}(z) = \frac{1}{2\pi i} \int_{\mathcal{C}} \frac{y^{\alpha-\beta} e^y}{y^\alpha - z} dy,$$

where the path of integration \mathcal{C} in the complex plane starts and ends at $-\infty$ and encircles the circular disc $|y| \leq |z|^{1/\alpha}$ in the positive sense.

Bending the contour in the complex plane and applying the Cauchy theorem, one obtains the following integral representation [21] for $0 < \alpha < 2$:

$$(4.23) \quad E_{\alpha,\beta}(z) = \frac{1}{2\pi i \alpha} \int_{\gamma(\varrho,\vartheta)} \frac{e^{\xi^{1/\alpha}} \xi^{\frac{1-\beta}{\alpha}}}{\xi - z} d\xi \quad \text{for } z \in G^{(-)}(\gamma),$$

$$(4.24) \quad E_{\alpha,\beta}(z) = \frac{1}{\alpha} z^{\frac{1-\beta}{\alpha}} \exp(z^{1/\alpha}) + \frac{1}{2\pi i \alpha} \int_{\gamma(\varrho,\vartheta)} \frac{e^{\xi^{1/\alpha}} \xi^{\frac{1-\beta}{\alpha}}}{\xi - z} d\xi \quad \text{for } z \in G^{(+)}(\gamma),$$

where the contour $\gamma(\varrho, \vartheta)$ starts at infinity in the lower half-plane, goes along a ray of $\arg(z) = -\vartheta$ towards the origin, encircles the origin with a circular arc of radius ϱ , and goes back to infinity in the upper half-plane along the ray $\arg(z) = +\vartheta$. Thus the complex plane is divided by the contour path in two parts, where $G^{(-)}(\gamma)$ and $G^{(+)}(\gamma)$ are the areas left and right of the contour path, respectively. Close to the contour path the numerical evaluation of the formulas (4.23) and (4.24) becomes inaccurate because of the singularity of the integral (4.23) at $r = |z|e^{\pm i\vartheta}$.

Combining formula (4.24) with $\vartheta = \pi\alpha$ and (4.23) with $\vartheta = 2\pi\alpha/3$, one can cover the whole complex plane. The regions $G^{(+)}(\pi\alpha)$ and $G^{(-)}(2\pi\alpha/3)$ have nonvanishing overlap. The overlap allows us to avoid the use of formulas (4.23) and (4.24) close to the contour path by choosing the partitioning described in section 3. The value of ϱ was set to 0.5 and thus lies in \mathbb{G}_1 , where the Taylor series is used for the calculation. Equation (4.24) is used with $\vartheta = \pi\alpha$ for $z \in \mathbb{G}_5$ and (4.23) with $\vartheta = 2\pi\alpha/3$ for $z \in \mathbb{G}_6$. When these integrals are evaluated, several cases arise. We distinguish the cases $\beta \leq 1$ and $\beta > 1$.

Let $z \in \mathbb{G}_5$. Summing up the different terms after inserting the parameterization of the contour path for $\vartheta = \pi\alpha$ and $\varepsilon = 1/2$ yields for $z \in G^{(+)}(\gamma)$ that

$$(4.25) \quad E_{\alpha,\beta}(z) = A(z; \alpha, \beta, 0) + \int_0^\infty B(r; \alpha, \beta, z, \pi\alpha) dr$$

for $\beta \leq 1$ and

$$(4.26) \quad E_{\alpha,\beta}(z) = A(z; \alpha, \beta, 0) + \int_{1/2}^\infty B(r; \alpha, \beta, z, \pi\alpha) dr + \int_{-\pi\alpha}^{\pi\alpha} C(\varphi; \alpha, \beta, z, 1/2) d\varphi$$

for $\beta > 1$, where the following abbreviations are used:

$$(4.27) \quad A(z; \alpha, \beta, x) = \frac{1}{\alpha} z^{(1-\beta)/\alpha} \exp\left[z^{1/\alpha} \cos(x/\alpha)\right],$$

$$(4.28) \quad B(r; \alpha, \beta, z, \phi) = \frac{1}{\pi} A(r; \alpha, \beta, \phi) \frac{r \sin[\omega(r, \phi, \alpha, \beta) - \phi] - z \sin[\omega(r, \phi, \alpha, \beta)]}{r^2 - 2rz \cos \phi + z^2},$$

$$(4.29) \quad C(\varphi; \alpha, \beta, z, \varrho) = \frac{\varrho}{2\pi} A(\varrho; \alpha, \beta, \varphi) \frac{\cos[\omega(\varrho, \varphi, \alpha, \beta)] + i \sin[\omega(\varrho, \varphi, \alpha, \beta)]}{\varrho(\cos \varphi + i \sin \varphi) - z},$$

$$(4.30) \quad \omega(x, y, \alpha, \beta) = x^{1/\alpha} \sin(y/\alpha) + y(1 + (1 - \beta)/\alpha).$$

In the case $\beta \leq 1$ we applied the limit $\varrho \rightarrow 0$. These equations are used to calculate $E_{\alpha,\beta}$ for $z \in \mathbb{G}_5$.

For $z \in \mathbb{G}_6$ we use a different contour with $\vartheta = 2\pi\alpha/3$. In this case the integral representations read

$$(4.31) \quad E_{\alpha,\beta}(z) = \int_0^\infty B(r; \alpha, \beta, z, 2\pi\alpha/3) dr, \quad \beta \leq 1,$$

$$(4.32) \quad E_{\alpha,\beta}(z) = \int_{1/2}^\infty B(r; \alpha, \beta, z, 2\pi\alpha/3) dr + \int_{-2\pi\alpha/3}^{2\pi\alpha/3} C(\varphi; \alpha, \beta, z, 1/2) d\varphi, \quad \beta > 1,$$

where the integrands have been defined in (4.27)–(4.30) above.

The integrand $C(\varphi; \alpha, \beta, z, \varrho)$ is oscillatory but bounded over the integration interval. Thus the integrals over C can be evaluated numerically using any appropriate quadrature formula. We use a robust adaptive 21-point Gauss–Kronrod scheme from the gnu scientific library (GSL) [3], which is based on the QUADPACK QAGS algorithm [14]. Other robust integration schemes such as the standard MATLAB Gauss–Lobatto scheme have also been used successfully. The integrals over $B(r; \alpha, \beta, z, \phi)$ involve unbounded intervals and have to be treated more carefully.

4.3.2. Integrands. As the integral representation is used only for $|z| > 0$, we assume z to be nonzero. It can be easily shown that the integrand $B(r; \alpha, \beta, z, \phi)$ in (4.28) behaves like $\mathcal{O}(r^{\frac{1-\beta}{\alpha}})$ for $r \rightarrow 0$ which yields to the following cases: For $\beta < 1 + \alpha$ the integrand $\int_a^\infty B(r; \alpha, \beta, z, \phi) dr$ is convergent in the limit $a \rightarrow 0$, but the integrand remains finite at $r = 0$ only for $\beta \leq 1$. Thus, for numerical integration the limit $a \rightarrow 0$ can be applied only in the case $\beta \leq 1$ in (4.31). For β exactly $1 + \alpha$ and $\phi = \pi\alpha$, the integrand can be further simplified and approaches a finite value in the limit $r \rightarrow 0$. The integrands for different cases of α and β are shown in Figure 2.

A special case in the integrand in (4.28) occurs when z lies on the contour line. Then the denominator becomes zero which causes problems in the numerical scheme. Although there are numerical algorithms for treating such integrands with singularities, it is much more accurate to avoid such a case if possible. In our case this has been done by choosing two different integral representations with different contour paths, so one can always be used in the case of z close to the contour of the other.

This also speeds up the algorithm because the integrand is smoother and the integration routine does not need to treat the singularities. The behavior of the integrand close to the contour path is shown in Figure 3.

4.3.3. Error estimates. Now we present the error estimates for the different integral formulas. The truncation points for the integrals over $B(r; \alpha, \beta, z, \phi)$ are determined such that the error $R(R_{\max}; \alpha, \beta, z, \phi)$ is smaller than a given accuracy ε .

For given $z \in \mathbb{G}_5$ (resp., $z \in \mathbb{G}_6$) and accuracy ε , we approximate the integrals by truncation. The error

$$(4.33) \quad R(R_{\max}; \alpha, \beta, z, \phi) \leq \left| \int_{R_{\max}}^\infty B(r; \alpha, \beta, z, \phi) dr \right| < \varepsilon$$

depends on the truncation point R_{\max} .

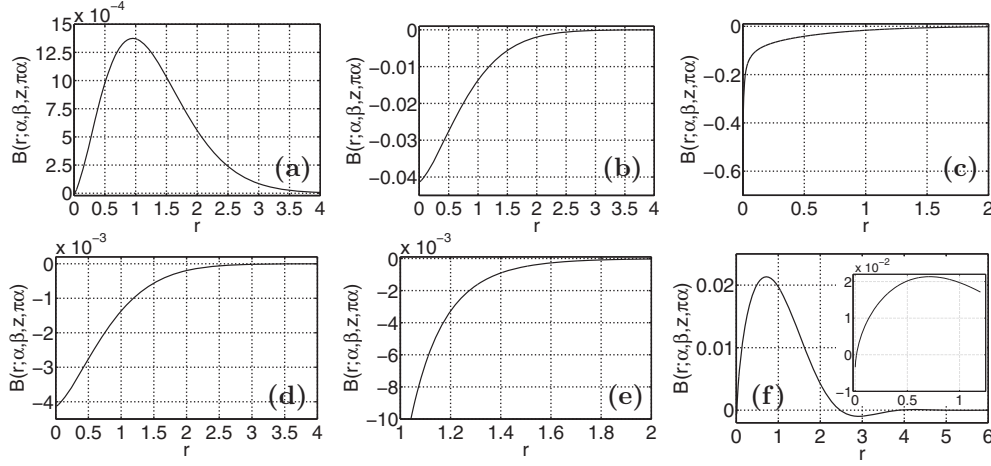


FIG. 2. (a)–(e) summarize the behavior of the integrand $B(r; \alpha, \beta, z, \phi)$ for different combinations of α and β , where $\phi = \pi\alpha$ and $z = 10$ are kept constant. (a) $\beta < 1$, $\beta < 1 + \alpha$: In this case the integral is convergent and the integrand approaches 0 for $r \rightarrow 0$. For the plot $\alpha = \beta = 2/3$ was chosen. (b) $\beta = 1$, $\beta < 1 + \alpha$: For $\beta = 1$ the integrand approaches a finite value in the limit $r \rightarrow 0$. The parameters for the figure are $\alpha = 2/3$ and $\beta = 1$. (c) $\beta > 1$, $\beta < 1 + \alpha$: If $\beta > 1$ but still smaller than $1 + \alpha$, the integral is still convergent but the integrand diverges in the limit $r \rightarrow 0$. The values for the plot have been chosen as $\alpha = 2/3$ and $\beta = 5/4$. (d) $\beta = 1 + \alpha$: This is a special case where the integrand approaches again a finite value in the limit $r \rightarrow 0$. The plot shows the integrand for $\alpha = 2/3$ and $\beta = 5/3$. (e) $\beta > 1 + \alpha$: For $\beta > 1 + \alpha$ the integrand diverges in the limit $r \rightarrow 0$ and the integral over $B(r; \alpha, \beta, z, \phi)$ does not converge anymore. (f) This figure shows the integrand $B(r; \alpha, \beta, z, \phi)$ for another value of $z = -10$ and $\phi = 2\pi\alpha/3$. The parameters α and β were chosen to be $\alpha = 2/3$ and $\beta = 1 + \alpha$. In this case the integrand does not approach a finite value in the limit $r \rightarrow 0$ as in the case $\phi = \pi\alpha$.

Inserting $\phi = \pi\alpha$ (resp., $\phi = 2\pi\alpha/3$) in $B(r; \alpha, \beta, z, \phi)$ one obtains, after simplifying the resulting terms,

$$(4.34) \quad B(r; \alpha, \beta, z, \pi\alpha) = \frac{1}{\pi\alpha} r^{(1-\beta)/\alpha} e^{-r^{1/\alpha}} \times \frac{r \sin[\pi(1-\beta)] - z \sin[\pi(1-\beta+\alpha)]}{r^2 - 2rz \cos \phi + z^2},$$

$$(4.35) \quad B(r; \alpha, \beta, z, 2\pi\alpha/3) = \frac{1}{\pi\alpha} r^{(1-\beta)/\alpha} e^{-(1/2)r^{1/\alpha}} \times \frac{r \sin[\omega(r, \frac{2\pi\alpha}{3}, \alpha, \beta) - \frac{2\pi\alpha}{3}] - z \sin[\omega(r, \frac{2\pi\alpha}{3}, \alpha, \beta)]}{r^2 - 2rz \cos \phi + z^2},$$

where $\omega(r, \frac{2\pi\alpha}{3}, \alpha, \beta) - \frac{2\pi\alpha}{3} = r^{1/\alpha} \frac{\sqrt{3}}{2} + \frac{2\pi}{3}(1-\beta)$ and $\omega(r, \frac{2\pi\alpha}{3}, \alpha, \beta) = r^{1/\alpha} \frac{\sqrt{3}}{2} + \frac{2\pi}{3}(\alpha + (1-\beta))$.

If we assume $r \geq 2|z|$, one finds for the denominator of (4.34) and (4.35) that

$$(4.36) \quad \frac{1}{|r^2 - 2rz \cos \phi + z^2|} = \frac{1}{r^2 \left| \frac{z}{r} - z_0 \right| \left| \frac{z}{r} - \bar{z}_0 \right|} \leq \frac{1}{r^2 \left(\left| \frac{z}{r} \right| - z_0 \right) \left(\left| \frac{z}{r} \right| - \bar{z}_0 \right)} \leq \frac{4}{r^2},$$

where ϕ is either $\pi\alpha$ or $2\pi\alpha/3$ and $z_0 = e^{i\phi}$.

As α, β, r , and ϕ are real numbers, one easily obtains

$$(4.37) \quad \left| \sin[\omega(r, \phi, \alpha, \beta) - \phi] - z \sin[\omega(r, \phi, \alpha, \beta)] \right| \leq \frac{3r}{2}$$

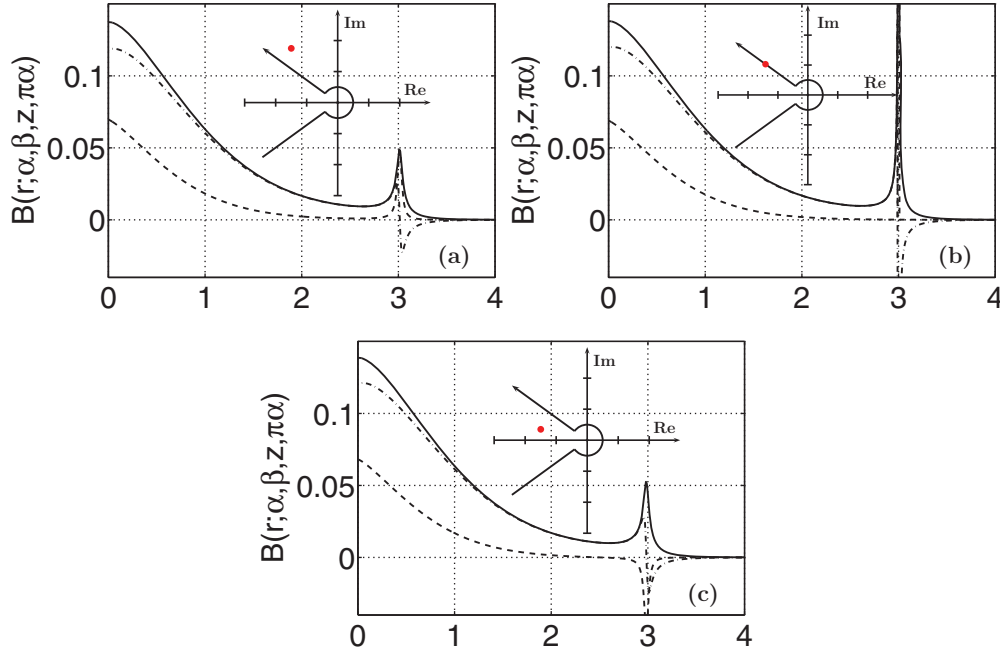


FIG. 3. The three plots show the divergence of the integrand $B(r; \alpha, \beta, z, \phi)$ close to the contour path. The real part is drawn with a dashed line and the imaginary part with a dashed-dotted line. The absolute value is marked with a black solid line. The parameters in the figures are $\alpha = 2/3$, $\beta = 1$, and $\phi = \pi\alpha$. (a) shows the integrand for z slightly above the contour path ($z = \exp(i\pi\alpha) + 0.03$), while in (c) the integrand is shown for slightly below ($z = \exp(i\pi\alpha) - 0.03$). The singularity is shown in (b) when z lies on the contour path. The insets show the contour path for the integration and the location of z relative to the contour path as a black dot.

for the numerator because in (4.36) r was assumed to be larger than $2|z|$. The angle ϕ is either $\pi\alpha$ or $2\pi\alpha/3$. Thus we obtain for the integrals

$$(4.38) \quad \left| \int_{R_{\max}}^{\infty} B(r; \alpha, \beta, z, \pi\alpha) dr \right| \leq \int_{R_{\max}}^{\infty} \frac{6}{\pi\alpha} r^{\frac{(1-\beta)}{\alpha}-1} e^{-r^{1/\alpha}} dr$$

$$\leq \frac{6}{\pi} \Gamma(1 - \beta, R_{\max}^{1/\alpha}),$$

$$(4.39) \quad \left| \int_{R_{\max}}^{\infty} B(r; \alpha, \beta, z, 2\pi\alpha/3) dr \right| \leq \int_{R_{\max}}^{\infty} \frac{6}{\pi\alpha} r^{\frac{(1-\beta)}{\alpha}-1} e^{-\frac{1}{2}r^{1/\alpha}} dr$$

$$\leq \frac{12}{2^{\beta}\pi} \Gamma\left(1 - \beta, \frac{1}{2}R_{\max}^{1/\alpha}\right)$$

after comparing with the definition of the incomplete gamma function. Now we apply the following two lemmas.

THEOREM 4.3 (lemma). *For the incomplete gamma function the following bounds hold:*

$$(4.40) \quad |\Gamma(1 - \beta, x)| \leq e^{-x} \quad \text{if } x \geq 1, \beta \geq 0,$$

$$(4.41) \quad |\Gamma(1 - \beta, x)| \leq (|\beta| + 2)x^{-\beta} e^{-x} \quad \text{if } x \geq |\beta| + 1, \beta < 0.$$

Proof. For $t \geq x \geq 1$ and $\beta \geq 0$ we have the inequality

$$(4.42) \quad t^{-\beta} e^{-t} \leq e^{-t},$$

and therefore

$$(4.43) \quad |\Gamma(1 - \beta, x)| \leq \int_x^\infty e^{-t} dt = e^{-x},$$

which proves the first inequality (4.40). For $\beta < 0$ we can find an $n \in \mathbb{N}$ such that

$$(4.44) \quad -(n + 1) \leq \beta < -n.$$

For $x > 0$ and $\beta < 0$ it holds that

$$(4.45) \quad \Gamma(1 - \beta, x) = x^{-\beta} e^{-x} \int_0^\infty e^{-u} (1 + (u/x))^{-\beta} du < x^{-\beta} e^{-x} \int_0^\infty e^{-u} (1 + (u/x))^N du,$$

where $N = |\beta| + 1$. This can be shown by using the binomial theorem for expanding $(1 + (u/x))^N$ and integrating (4.45) term by term:

$$(4.46) \quad \Gamma(1 - \beta, x) < x^{-\beta} e^{-x} \left(1 + \frac{N}{x} + N^2 x^2 + \dots + N^N x^N \right) < x^{-\beta} e^{-x} (N + 1),$$

where $x \geq N$ and $N + 1 \leq 2 - \beta$. This yields the second inequality of (4.41):

$$(4.47) \quad |\Gamma(1 - \beta, x)| \leq (|\beta| + 2)x^{-\beta} e^{-x}. \quad \square$$

THEOREM 4.4 (lemma). *The inequality*

$$(4.48) \quad x^y \leq (qy)^y e^{x/q}$$

holds for arbitrary $x, y, q > 0$.

Proof. Let $x = a \cdot y$ with $a > 0$. Then one can write (4.48) as $(ay)^y \leq (qy)^y e^{ay/q}$, which is equivalent to $a^y \leq q^y e^{ay/q}$. Dividing this by q^y and applying the logarithm on both sides yields $y \log(a/q) \leq y(a/q)$, which is the same as $\log(a/q) \leq (a/q)$ because $y > 0$. The last inequality is true because $\log(x) < x \forall x > 0$ and $a, q > 0$. \square

Using Theorem 4.3 with (4.38) (resp., (4.39)) and then applying Theorem 4.4 in the case $\beta < 0$, the following estimates for $\phi = \pi\alpha$ and $\phi = 2\pi\alpha/3$ are obtained:

$$(4.49) \quad R(R_{\max}; \alpha, \beta, z, \pi\alpha) \leq \begin{cases} \frac{6}{\pi} e^{-R_{\max}^{1/\alpha}}, & \beta \geq 0, R_{\max} \geq 1, \\ \frac{6}{\pi} (|\beta| + 2)(2|\beta|)^{|\beta|} e^{-\frac{1}{2}R_{\max}^{1/\alpha}}, & \beta < 0, R_{\max}^{1/\alpha} \geq |\beta| + 1, \end{cases}$$

$$(4.50) \quad R(R_{\max}; \alpha, \beta, z, 2\pi\alpha/3) \leq \begin{cases} \frac{12}{2^{\beta}\pi} e^{-\frac{1}{2}R_{\max}^{1/\alpha}}, & \beta \geq 0, R_{\max} \geq 1, \\ \frac{12}{2^{\beta}\pi} (|\beta| + 2)(4|\beta|)^{|\beta|} e^{-\frac{1}{4}R_{\max}^{1/\alpha}}, & \beta < 0, R_{\max}^{1/\alpha} \geq |\beta| + 1. \end{cases}$$

In the case $\phi = \pi\alpha$ we applied Theorem 4.4 with $x = R_{\max}^{1/\alpha}$, $y = -\beta$, and $q = 2$, whereas in the case $\phi = 2\pi\alpha/3$ Theorem 4.4 was used with the parameters $x = R_{\max}^{1/\alpha}$, $y = -\beta$, and $q = 4$.

Recalling that (4.36) requires $R_{\max} > 2|z|$ and solving (4.49) and (4.50) for the truncation point R_{\max} , one finds for $\phi = \pi\alpha$ that

$$(4.51) \quad R_{\max} \geq \begin{cases} \max \left\{ 1, 2|z|, \left(-\ln \frac{\pi\varepsilon}{6} \right)^\alpha \right\}, & \beta \geq 0, \\ \max \left\{ (|\beta| + 1)^\alpha, 2|z|, \left(-2 \ln \left(\frac{\pi\varepsilon}{6(|\beta| + 2)(2|\beta|)^{|\beta|}} \right) \right)^\alpha \right\} & \beta < 0 \end{cases}$$

while for $\phi = 2\pi\alpha/3$ we have

$$(4.52) \quad R_{\max} \geq \begin{cases} \max \left\{ 2^\alpha, 2|z|, \left(-2 \ln \frac{\pi 2^\beta \varepsilon}{12} \right)^\alpha \right\}, & \beta \geq 0, \\ \max \left\{ [2(|\beta| + 1)]^\alpha, 2|z|, \left[-4 \ln \frac{\pi 2^\beta \varepsilon}{12(|\beta| + 2)(4|\beta|)^{|\beta|}} \right]^\alpha \right\}, & \beta < 0. \end{cases}$$

5. Discussion of the algorithm. In this section we briefly summarize the algorithm and the different cases that have to be distinguished to calculate $E_{\alpha,\beta}$ for a given error ε . Also the numerical treatment of the constants and error terms is described and details of the algorithm are discussed.

For $|z| \leq 0.95$ the Taylor series (2.1) is used for all $\alpha > 0$ and $\beta \in \mathbb{R}$. We apply the recursion formula (2.2) in the case $\alpha > 1$. The maximum number of terms M in the Taylor series which are needed to achieve an accuracy of ε is given by (4.5). This number increases rapidly for $z \rightarrow 1$; thus the Taylor series becomes inefficient for the calculation of $E_{\alpha,\beta}$ if $z > 0.95$. Furthermore if M becomes very large, the calculation of the gamma function fails due to numerical overflows. This can be avoided by using the identity

$$(5.1) \quad z^k / \Gamma(\alpha k + \beta) = \exp(k \ln(z) - \ln \Gamma(\alpha k + \beta)),$$

which behaves in a more stable manner for large k and $|z|$ not too close to the origin, where the logarithm is singular. Thus for very small $|z| < 0.5$ we switch back to the definition (2.1). In this case only a few terms of the Taylor series are necessary to obtain good accuracy, so one does not have to worry about M . Also a real Lanczos approximation [3] of the reciprocal gamma function is used to avoid problems at the poles of $\Gamma(\alpha k + \beta)$ when β is smaller than 0.

The integral formulas are limited only by the floating arithmetic of the computer, where errors can accumulate due to the rounding errors of the numerical integration scheme.

Now if $|z| \geq r_1$, where r_1 is given by (4.21), the asymptotic formulas are used, where we distinguish between the two asymptotic formulas (2.3) in \mathbb{G}_1 (resp., (2.4) in \mathbb{G}_2) and the Berry-type smoothed regions \mathbb{G}_3 and \mathbb{G}_4 in an angle of $\pm\delta$ around the Stokes lines ((2.5) and (2.6)). In the algorithm we choose $\delta = \pi\alpha/8$ to avoid coming too close to the Stokes lines with (2.3) (resp., (2.4)). For $|\delta| \leq \pi\alpha/8$ the argument $\theta \in [-\pi/8, \pi/8]$ is sufficiently small that one can apply the approximation $c \approx \theta + \frac{1}{6}\theta^2 - \frac{1}{36}\theta^3$. Figure 4 shows a plot of $c(\theta)$. The error of the approximation is much smaller than 1×10^{-4} in the real part and 1×10^{-3} in the imaginary one so that this approximation in the exponent is negligible.

For large α ($\alpha \geq 8/9$) the angle $\pi\alpha + \delta$ would cross the negative real axis and produce a jump in θ ; thus we always make use of the normal asymptotic series in a

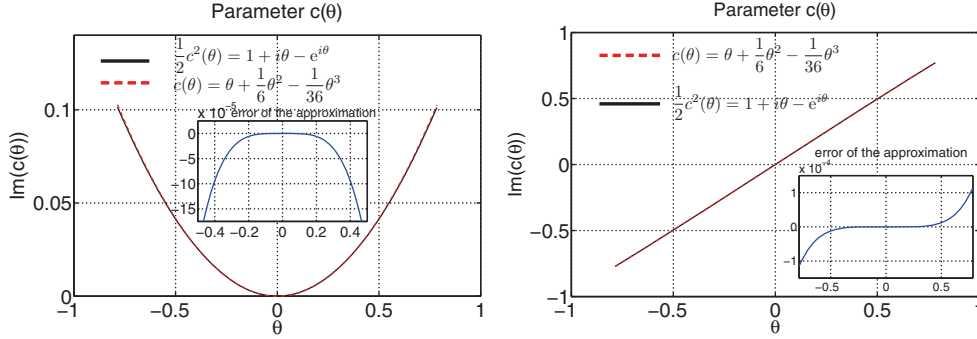


FIG. 4. The figures show the real (left) and imaginary part (right) of the parameter $c(\theta)$ of the Berry-type smoothed asymptotic series (2.5) (resp., (2.6)). The exact solution of $c(\theta)^2/2 = 1 + i\theta - e^{i\theta}$ is plotted as a black solid line, and the approximation $c \approx \theta + \frac{1}{6}\theta^2 - \frac{1}{36}\theta^3$ is marked with a dashed line. In the considered region of $\theta \in [-\pi/4, \pi/4]$ the difference between the two curves is negligible. The absolute error of the approximation is given in the inset plot.

wedge around the negative real axis, where the opening angle $\tilde{\delta}$ is between $\pi\alpha$ and π . More precisely $\tilde{\delta}$ is defined by $\tilde{\delta} = \min\{(\pi + \pi\alpha)/2, \delta\}$, where δ is the angle of the wedge around the Stokes line (cf. (3.7)). In the algorithm δ was chosen to be $\pi\alpha/8$.

Optimal truncation for the series should be achieved for M according to (4.21), which would mean that the number of necessary coefficients will increase with $\sim |z|^{1/\alpha}$. But for $z \rightarrow \infty$ the asymptotic series becomes better and better, so we apply an upper limit of $M < 100$ in the algorithm. This is also necessary to avoid overflows in the reciprocal gamma function for small negative values which reaches the floating point limit for about $\beta - \alpha k \approx 130$. As the term $z^{-k}/\Gamma(\beta - \alpha k)$ involves the multiplication of a very small number ($z^{-k} \ll 1$) with a very large number $1/\Gamma(\beta - \alpha k) \gg 1$, it is better to apply the $\exp - \log$ identity again for large k . In this case the calculation of $\log \Gamma(\beta - \alpha k)$ can fail due to the poles of the gamma function for negative integer values; thus we assume all terms to be zero for which $\beta - \alpha k$ is closer than 10^{-9} to a negative integer value.

In the algorithm the constant C in (4.19) is replaced with

$$(5.2) \quad C_0 = \frac{1}{2\pi} \left(\frac{1}{\sin \alpha\pi} + \frac{1}{\min\{\sin \alpha\pi, \sin \xi\}} \right) \approx \frac{1}{\pi \sin \pi\alpha},$$

obtained by combining (4.11) and (4.12). The choice to limit the constant C_0 on the Stokes line is supported analytically by the argument of Boyd [1, 23] that demands a smooth transition when crossing the Stokes line from one side to the other. Therefore the error cannot diverge when ξ goes to zero. Extensive numerical tests (see Table 2) and comparisons with the integral representations show that with (5.2) the relative errors between the asymptotic series and the integral representation are extremely small for $|z| \leq r_1 = (-2 \log(\varepsilon/C_0))^\alpha$. In practice the influence of the exact numerical value of C_0 is small as the constant appears inside a logarithm. In the Berry-type smoothed regions, the series (2.5) and (2.6) shows even better, namely, exponential convergence, than the normal series. Thus we can use the same estimate for r_1 as under optimal truncation.

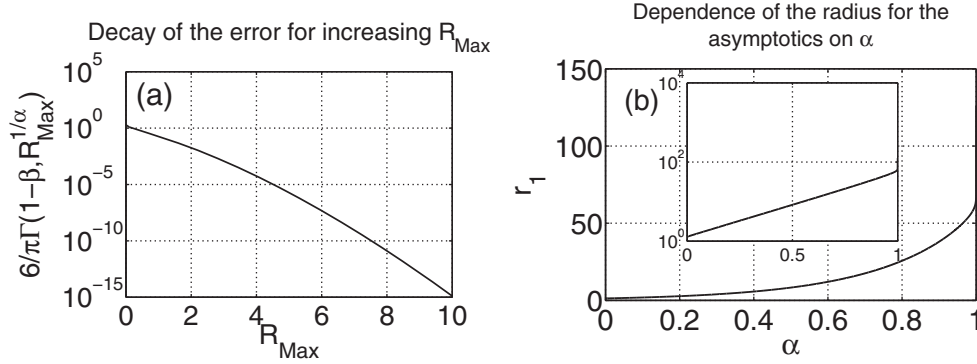


FIG. 5. The two plots summarize the results of the error estimates of the integral representation. (a) shows the decay of the error term (4.38) of the integral representation for $\alpha = \beta = 2/3$ on a semilog scale. It decays faster than exponentially with R_{max} . (b) shows the dependence of the radius r_1 on the parameter α for a given tolerance $\varepsilon = 1e - 12$ using the formula $r_1 = (-2\log(\varepsilon/C_0))^\alpha = 14.16$, with C_0 given by (5.2). The radius r_1 determines from where on the asymptotic formula that is used to calculate $E_{\alpha,\beta}(z)$. The inset shows a semilogarithmic plot of the same region.

If $0.95 < |z| < r_1$, the integral formulas are used in \mathbb{G}_5 and \mathbb{G}_6 . For the error estimates we have to distinguish here $\beta > 1$ and $\beta \leq 1$, where the case $\beta \leq 1$ yields two further subcases $\beta \geq 0$ and $\beta < 0$. The contours for the integral formulas for the areas \mathbb{G}_5 and \mathbb{G}_6 have been chosen such that z will never come close to the contour of integration. A plot of the areas \mathbb{G}_5 and \mathbb{G}_6 with the two contour paths $\gamma(1/2, \pi\alpha)$ and $\gamma(1/2, 2\pi\alpha/3)$ is shown in Figure 1(b). Figure 5(a) shows a plot of the right-hand side of (4.38) versus R_{max} . One can see that the accuracy is extremely good even for rather small R_{max} . The estimates for R_{max} are given in (4.51) and (4.52).

The integrand in (4.29) is oscillatory. The amplitude of these oscillations is mainly determined by the prefactor in (4.27). For real argument $z \in \mathbb{R}$, the imaginary part of the integrand is antisymmetric in (4.27) with respect to the origin. This means that the integral over the imaginary part of (4.29) vanishes and does not have to be calculated. For the real part one can also make use of the symmetry and perform the integration over only half of the contour arc.

A detailed description of the behavior of the generalized Mittag-Leffler function in the complex plane and the investigation of the distribution of the zeros can be found in [7]. Two contour plots of $E_{\alpha,\beta}(z)$ are shown in Figure 6.

The algorithm has been implemented for complex z with real and imaginary parts in double precision. The parameters α and β are represented as real double variables. A separate function for real double z has also been implemented. In this case the integration routines make use of the symmetry of the integrand and the fact that $\text{Im}(E_{\alpha,\beta}(x)) = 0$. This speeds up the integration routine and improves the stability of the algorithm.

6. Numerical tests and speed analysis of the algorithm. Extensive numerical calculations of the algorithm were performed to test the stability and the validity range for the parameters $\alpha, \beta \in \mathbb{R}$ and the argument $z \in \mathbb{C}$. The results are summarized in this section.

By incorporating the asymptotic formula the algorithm has been much improved. Especially for large values of $|z|$ the asymptotic series are much faster and more stable than the integral formulas. For some representative values the differences are reported

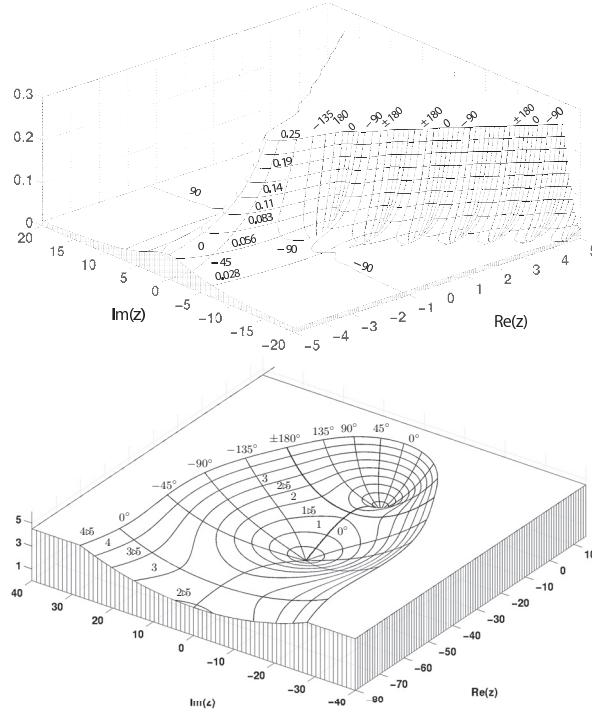


FIG. 6. The plots show the behavior of the Mittag-Leffler function $E_{\alpha,1}(z)$ in the complex plane for $\alpha = 0.8$ (top) and $\alpha = 2.25$ (bottom). In the three-dimensional plot the absolute value of the function is plotted with the contour lines on top. Furthermore, the contour lines of the argument are also marked.

in Table 1, where the speed of the integral routine is compared to the asymptotic formula in a range where both formulas can be used for the calculation. The speedup factor ranges typically between 20 and 40. Close to the Stokes lines the Berry-type smoothed formulas give a better approximation than the normal asymptotic series.

The convergence of the series is compared with the values obtained by the integral representation along rays with constant argument for $|z| \rightarrow \infty$. Good agreement and fast convergence is obtained in all sectors of the complex plane.

For very large values of $|z|$ the integral formula cannot be used due to rounding errors in the floating point arithmetic. Also the dependence on the parameters α and β is very important; for example, the amplitude equation (4.27) of the integrand $B(r; \alpha, \beta, z, \delta)$ in (4.28) behaves like $r^{(1-\beta)/\alpha}$. If this value becomes too large, the integration routines cannot achieve the required precision. An example of $B(r; \alpha, \beta, z, \delta)$ for different $\beta = 5, 10, 20$ is shown in Figure 7. Similar limitations hold for the integrand in (4.29), where the integrand oscillates with a higher and higher frequency while the value of the integral is close to zero (see Figure 8).

A comparison of the integral formula with the asymptotic series in a range around r_1 is shown in Figure 9. The range for which we still obtained reasonable results is summarized in Table 2. The tests were performed by calculating the values of $E_{\alpha,\beta}(z)$ on a rectangular grid with $\text{Re}_{\min} \leq \text{Re}(z) < \text{Re}_{\max}$ and $\text{Im}_{\min} \leq \text{Im}(z) < \text{Im}_{\max}$ with different lattice spacings. Furthermore the values of $E_{\alpha,\beta}$ are calculated along different rays from the origin covering the different areas of the partitioning (section 3) of the complex plane. For various combinations of α and β the range of validity with respect

TABLE 1

Comparison of the speedup for the calculation for $E_{\alpha,\beta}$ using the asymptotic series expansions. The tolerance ε is set to 10^{-16} . The last column indicates the equation used for the calculation of the asymptotic series. Each speedup factor is averaged over 100 calculations. All calculations were done using the complex algorithm.

α	β	z	$E_{\alpha,\beta}^{\text{int}}(z)$	$E_{\alpha,\beta}^{\text{int}}(z) - E_{\alpha,\beta}^{\text{asym}}(z)$	Speedup
0.6	0.8	7.0	4.24680224e + 11	0.0	40.8
0.6	0.8	20.0	4.50513132e + 64	0.0	21.1
0.6	0.8	-7.0	0.036402965145	6.2000e - 13	55.5
0.6	0.8	-50.0	0.004463867842	0.0	35.3
0.6	0.8	$7e^{0.6\pi i}$	0.00509750 + 0.03299810i	$1.06e - 12 - 2.28e - 12i$	57.1
0.6	0.8	$20e^{0.6\pi i}$	0.00282134 + 0.01075547i	0.0	39.7
0.6	1.25	7.0	9.86821285e + 10	0	22.0
0.6	1.25	20.0	4.76359640e + 63	0	14.0
0.6	1.25	-7.0	0.101261033685	$4.3e - 13$	31.2
0.6	1.25	-50.0	0.014419766303	0.0	16.01
0.6	1.25	$7e^{0.6\pi i}$	0.03339025 + 0.0980431i	$-8.0e - 14 - 2.70e - 13i$	32.7
0.6	1.25	$20e^{0.6\pi i}$	0.01128945 + 0.0342852i	0.0	16.1
0.6	-0.8	7.0	4.24680224e + 11	0.0	40.8
0.6	-0.8	20.0	4.50513132e + 64	0.0	21.1
0.6	-0.8	-7.0	0.036402965145	$6.2e - 13$	55.5
0.6	-0.8	-50.0	0.004463867842	0.0	35.3
0.6	-0.8	$7e^{0.6\pi i}$	0.01931826 + 0.0537209i	$-4.61e - 10 - 8.36e - 10i$	69.0
0.6	-0.8	$20e^{0.6\pi i}$	0.00592228 + 0.0179734i	0.0	35.2

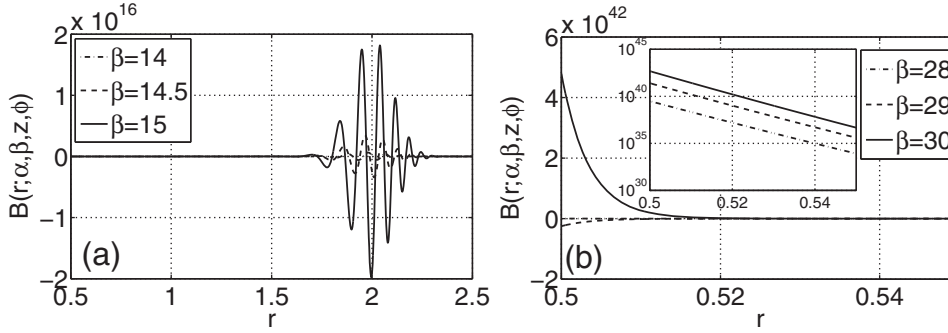


FIG. 7. The figure shows the integrand $B(r; \alpha, \beta, z, 2\pi\alpha/3)$ for $\alpha = 0.2$ and different values of large $|\beta|$. The argument is set to $z = -5 \in \mathbb{G}_5$ (left). One can see how strongly the amplitude increases with increasing β . The oscillations are especially difficult to handle as the large positive parts of the integral cancel out with the large negative ones. The right figure shows the integrand $B(r; \alpha, \beta, z, 2\pi\alpha)$ in the case $z = 5 \in \mathbb{G}_6$ for the parameters $\beta = 28$ to $\beta = 30$. Note the values on the y-axis. In the inset the semilog plot of the absolute value of the integrand in the same range is shown.

to the argument z was determined for the integral representation and the asymptotic series. The upper limit $|z|_{\max}^{\text{int}}$ for which the integral representation is still valid is calculated as follows: Let $E_{\alpha,\beta}^{\text{int}}$ be the values of the Mittag-Leffler function calculated with the integral representation and $E_{\alpha,\beta}^{\text{asym}}$ those obtained from the asymptotic series. As for $|z| \rightarrow \infty$ the asymptotic series becomes more and more accurate; thus the difference $|E_{\alpha,\beta}^{\text{int}} - E_{\alpha,\beta}^{\text{asym}}|$ decreases until it increases again for $|z| > |z|_{\max}^{\text{int}}$ due to the fact that the integral representation becomes inaccurate. For given α and β the value reported in Table 2 is the minimum of all $|z|_{\max}^{\text{int}}$ obtained on the different rays covering all sectors of the partitioning described in section 3.

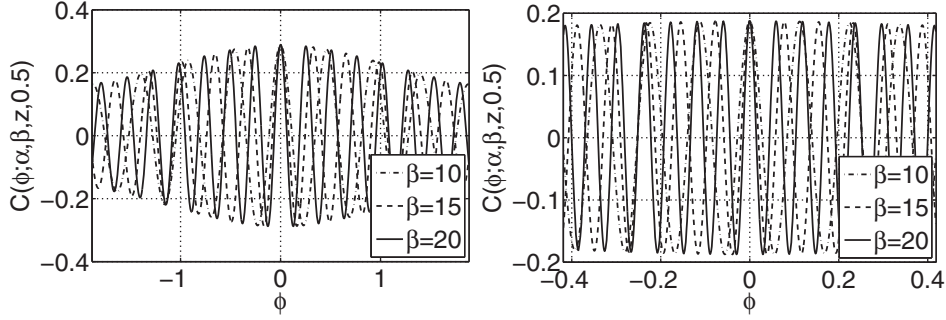


FIG. 8. The figure shows the integrand $C(\varphi; \alpha, \beta, z, \rho)$ in (4.29) for $\alpha = 0.9$ and different values of β (left). The same integrand but for $\alpha = 0.2$ is shown in the right figure. All other parameters are the same in both plots, where z was set to -5 and $\rho = 0.5$.

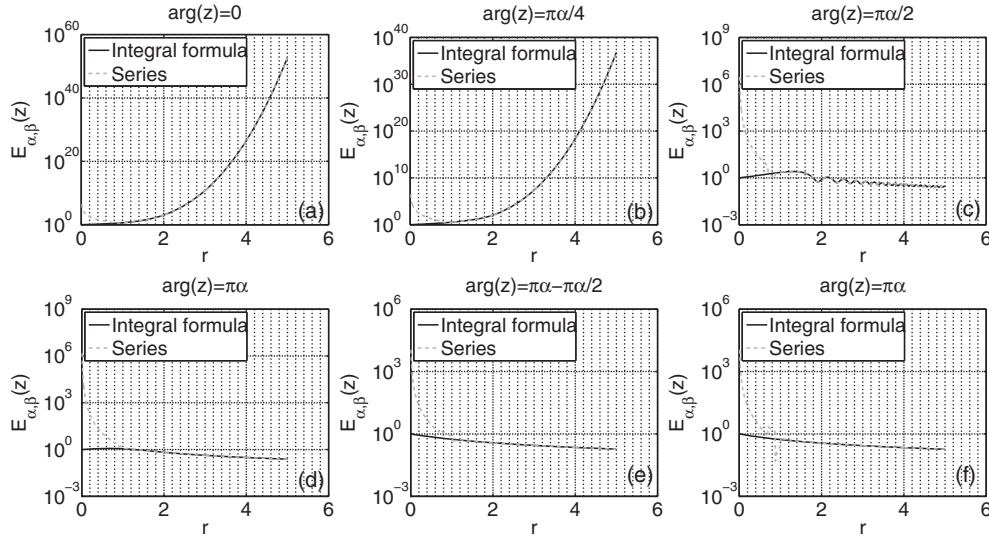


FIG. 9. The figures show the behavior of the Mittag-Leffler function in a range of $|z|$ around r_1 calculated with the asymptotic series (gray dashed line) and the integral formula (black solid line) for different arguments of z . Starting from (a) to (f) the arguments are (a) $\arg(z) = 0$, (b) $\arg(z) = \pi\alpha/4$, (c) $\arg(z) = \pi\alpha/2$, (d) $\arg(z) = \pi\alpha$, (e) $\arg(z) = \pi - \pi\alpha/4$, and (f) $\arg(z) = \pi$. The example is calculated for $\alpha = 1/3$, $\beta = 2$, and $\varepsilon = 10^{-10}$. The value of r_1 in this case is $r_1 = (-2 \log(\varepsilon/C))^\alpha = 3.53$.

In the case $|z| \rightarrow 0$ the asymptotic series becomes unstable. The lower bound $|z|_{\min}^{\text{asym}}$ of the asymptotic series is defined by the value of $|z|$ for which the relative error $\text{Err}_{\text{rel}} = |(E_{\alpha, \beta}^{\text{int}} - E_{\alpha, \beta}^{\text{asym}})/E_{\alpha, \beta}^{\text{int}}|$ exceeds 10^{-2} . The largest value of $|z|_{\min}^{\text{asym}}$ along the different rays is shown in Table 2.

For the parameters α and β we find that, in the range $0.05 < \alpha < 0.999$ and $-2 < \beta < 2$, the algorithm covers the whole complex plane up to machine precision in the asymptotic formula which is reached for positive values at $\text{Re}(z) \leq 1.5$ ($\alpha = 0.05$) and $|z| \leq 600$ ($\alpha = 0.999$) as the function grows like $1/\alpha \exp(x^{1/\alpha})$. For $\text{Re}(z) \rightarrow -\infty$ the limits are much larger. For a negative argument the function can be calculated at least for $\text{Re}(z) > -10^{50}$.

TABLE 2

Comparison of the asymptotic series and the integral formula for different combinations of α and β . The value $|z|_{\max}^{\text{int}}$ indicates when the integral formulas start to become inaccurate (see text). For $z < |z|_{\min}^{\text{asym}}$ the asymptotic formula begins to break down. This value is defined by the smallest number of $|z|$ for which the relative error is still smaller than 10^{-2} . The relative error is defined by $\text{Err}_{\text{rel}} = |(E_{\alpha,\beta}^{\text{int}} - E_{\alpha,\beta}^{\text{asym}})/E_{\alpha,\beta}^{\text{int}}|$, where $E_{\alpha,\beta}^{\text{int}}$ are the values of $E_{\alpha,\beta}$ obtained with the integral representation and $E_{\alpha,\beta}^{\text{asym}}$ are the values calculated with the series expansion. The numerical values of the relative error at the points $|z| = r_1$ are reported in the sixth column, where $r_1 = (-2 \log(\varepsilon/C_0))^\alpha$ is the lower radius for the asymptotic series. The value of ε is set to 10^{-11} , and C_0 is given by (5.2).

α	β	$ z _{\max}^{\text{int}}$	$ z _{\min}^{\text{asym}}$	r_1	Err_{rel} at r_1
0.1	-5	$ z _{\max} = 8.5 \pm 0.5$	1.3566	1.4809	1.0835e - 05
0.1	-1	$ z _{\max} = 8.5 \pm 0.5$	1.2718	1.4809	8.7296e - 08
0.1	-2/3	$ z _{\max} = 8.5 \pm 0.5$	1.2380	1.4809	3.2593e - 08
0.1	2/3	> 200	1.1391	1.4809	3.8697e - 10
0.1	1	> 200	1.1052	1.4809	8.2624e - 11
0.1	5/3	$ z _{\max} = 5.5 \pm 0.5$	1.1173	1.4809	3.0580e - 11
0.2	-5	$ z _{\max} = 13.5 \pm 0.5$	1.7785	2.1817	1.8535e - 09
0.2	-1	$ z _{\max} = 12.5 \pm 0.5$	1.6014	2.1817	6.0096e - 12
0.2	-2/3	$ z _{\max} = 12.5 \pm 0.5$	1.5863	2.1817	3.8440e - 12
0.2	2/3	> 200	1.3517	2.1817	7.3541e - 12
0.2	1	> 200	1.2778	2.1817	1.0214e - 12
0.2	5/3	$ z _{\max} = 136.5 \pm 0.5$	1.2304	2.1817	2.3144e - 12
0.5	-5	$ z _{\max} = 53.5 \pm 0.5$	4.4068	6.9547	4.4764e - 13
0.5	-1	$ z _{\max} = 40.5 \pm 0.5$	3.3007	6.9547	7.1942e - 14
0.5	-2/3	$ z _{\max} = 40.5 \pm 0.5$	3.2379	6.9547	5.5955e - 14
0.5	2/3	> 200	2.5417	6.9547	1.4747e - 11
0.5	1	> 200	2.0966	6.9547	1.7497e - 12
0.5	5/3	> 200	1.5969	6.9547	1.5321e - 14
0.5	10	> 200	2.7404	6.9547	7.4421e - 07
0.9	-5	> 200	15.2764	34.2477	3.5560e - 13
0.9	-1	> 200	9.7702	34.2477	1.3282e - 10
0.9	-2/3	$ z _{\max} = 190.5 \pm 0.5$	8.5021	34.2477	4.6096e - 13
0.9	2/3	> 200	6.5593	34.2477	1.0505e - 11
0.9	1	> 200	6.3687	34.2477	6.1782e - 12
0.9	5/3	> 200	3.6252	34.2477	8.7930e - 14
0.9	10	> 200	5.9216	34.2477	6.6471e - 08

The limitations of large $|\beta|$ can be overcome using the following relation:

$$(6.1) \quad z^n E_{\alpha,\beta+n\alpha}(z) = \sum_{k=n}^{\infty} \frac{z^{n+k}}{\Gamma(\alpha(n+k) + \beta)} = E_{\alpha,\beta}(z) - \sum_{k=0}^n \frac{z^k}{\Gamma(\alpha k + \beta)}.$$

This extends the range of β to arbitrary large and small values.

7. Conclusion. A numerical algorithm for calculating the generalized Mittag-Leffler function for arbitrary real parameters $\alpha > 0$ and β was presented, and the error estimates have been calculated. Different representations have been used for different values of $z \in \mathbb{C}$ to obtain optimal stability and accuracy. The algorithm is not only fast, but it also eliminates numerical instabilities in other codes [4]. Furthermore the algorithm was extended using exponentially improved asymptotics. The Berry-type smoothing was used to avoid numerical instabilities close to the Stokes lines, where the asymptotic formulas fail. A great improvement in the speed and stability of the algorithm especially for large values of z has been achieved using the asymptotic series as presented in section 6. Furthermore the algorithm has been analyzed in detail, and

several numerical techniques have been discussed to improve the calculation. Finally, a detailed analysis of the numerical stability and validity of the algorithm is given. The algorithm is available as in C and as a MATLAB script for download [20].

Acknowledgments. The authors thank Julian Engel for useful discussions and performing extensive test runs with the algorithm. Furthermore, we want to thank the referees for their constructive criticism and useful hints.

REFERENCES

- [1] W. C. BOYD, *Stieltjes transforms and the Stokes phenomenon*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 429 (1990), pp. 227–246.
- [2] A. ERDELYI, *Higher Transcendental Functions*, Vols. I–III, Krieger, Malabar, FL, 1981.
- [3] M. GALASSI, *The GNU Scientific Library*, Network Theory Ltd., Bristol, UK, 2006.
- [4] R. GORENFLO, Y. LUCHKO, AND I. LOUTCHKO, *Computation of the Mittag-Leffler function $E_{\alpha,\beta}(z)$ and its derivatives*, Fract. Calc. Appl. Anal., 5 (2002), pp. 491–518. Erratum: Fract. Calc. Appl. Anal., 6 (2003).
- [5] R. HILFER AND L. ANTON, *Fractional master equation and fractal time random walks*, Phys. Rev. E, 51 (1995), pp. 848–851.
- [6] R. HILFER, R. METZLER, A. BLUMEN, AND J. KLAFTER, *Strange kinetics*, Chem. Phys., Special Issue, 284 (2002), pp. 1–2.
- [7] R. HILFER AND H. SEYBOLD, *Computation of the generalized Mittag-Leffler function and its inverse in the complex plane*, Integral Transforms Spec. Funct., 17 (2006), pp. 637–652.
- [8] R. HILFER, *Applications of Fractional Calculus in Physics*, World Scientific, Singapore, 2000.
- [9] R. HILFER, *Experimental evidence for fractional time evolution in glass forming materials*, Chem. Phys., 284 (2002), pp. 399–408.
- [10] G. MITTAG-LEFFLER, *Sur la nouvelle fonction $E_\alpha(x)$* , C. R. Math. Acad. Sci. Paris, 137 (1903), p. 554.
- [11] G. MITTAG-LEFFLER, *Une généralisation de l'intégrale de Laplace-Abel*, C. R. Math. Acad. Sci. Paris, 136 (1903), p. 537.
- [12] G. MITTAG-LEFFLER, *Sur la représentation d'une branche uniforme d'une fonction monogène*, Acta Math., 29 (1905), p. 101.
- [13] R. PARIS, *Exponential asymptotics of the Mittag-Leffler function*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 458 (2002), pp. 3041–3052.
- [14] E. PIESSENS, E. DE DONCKER-KAPENGER, C. UEBERHUBER, AND D. KAHANER, *Quadpack, a subroutine package for automatic integration*, Springer, New York, 1983.
- [15] I. POLDUBNY, *Fractional Differential Equations*, 1st ed., Academic Press, New York, 1999.
- [16] H. POLLARD, *The completely monotonic character of the Mittag-Leffler function*, Bull. Amer. Math. Soc., 54 (1948), pp. 1115–1116.
- [17] S. SAMKO, K. KILBAS, AND O. MARICHEV, *Fractional Integrals and Derivatives*, Gordon and Breach, Berlin, 1993.
- [18] W. SCHNEIDER, *Completely monotone generalized Mittag-Leffler functions*, Expo. Math., 14 (1996), p. 3.
- [19] H. SEYBOLD AND R. HILFER, *Numerical results for the generalized Mittag-Leffler function*, Fract. Calc. Appl. Anal., 8 (2005), pp. 127–139.
- [20] H. SEYBOLD AND R. HILFER, <http://www.icp.uni-stuttgart.de/~hilfer/forschung>.
- [21] A. WIMAN, *Über den Fundamentalsatz in der Theorie der Funktionen $E_\alpha(x)$* , Acta Math., 29 (1905), pp. 191–201.
- [22] A. WIMAN, *Über die Nullstellen der Funktionen $E_\alpha(x)$* , Acta Math., 29 (1905), pp. 217–234.
- [23] R. WONG AND Y. ZHAO, *Exponential asymptotics of the Mittag-Leffler function*, Constr. Approx., 18 (2002), pp. 355–385.
- [24] W. WYSS, *The fractional diffusion equation*, J. Math. Phys., 27 (1986), pp. 2782–2785.

DISCONTINUOUS DISCRETIZATION FOR LEAST-SQUARES FORMULATION OF SINGULARLY PERTURBED REACTION-DIFFUSION PROBLEMS IN ONE AND TWO DIMENSIONS*

RUNCHANG LIN[†]

Abstract. In this paper, we consider the singularly perturbed reaction-diffusion problem in one and two dimensions. The boundary value problem is decomposed into a first-order system to which a suitable weighted least-squares formulation is proposed. A robust, stable, and efficient approach is developed based on local discontinuous Galerkin (LDG) discretization for the weak form. Uniform error estimates are derived. Numerical examples are presented to illustrate the method and the theoretical results. Comparison studies are made between the proposed method and other methods.

Key words. least-squares methods, local discontinuous Galerkin methods, singular perturbation problems, reaction-diffusion problems

AMS subject classification. 65N30

DOI. 10.1137/070700267

1. Introduction. In this paper, we are concerned with the singularly perturbed reaction-diffusion problem

$$(1.1) \quad \begin{cases} -\epsilon^2 \Delta u + cu = f & \text{in } \Omega = (0, 1)^d, \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

where $0 < \epsilon \ll 1$ is the perturbation parameter, c and f are continuous functions in $\bar{\Omega}$, $0 < c_0 \leq c \leq c_1$ in $\bar{\Omega}$ with two positive constants c_0 and c_1 , and $d = 1$ or 2 .

It is well known that exact solutions to singular perturbation problems (1.1) typically contain *layers*, which cause nonmonotonic numerical oscillations in the solutions from standard Galerkin finite element methods (FEMs). Stabilization techniques such as upwinding, Petrov–Galerkin, streamline diffusion, discontinuous Galerkin, and adaptive approximations have been developed to improve standard Galerkin methods. For an overview of these methods, we refer to the books by Miller, O’Riordan, and Shishkin [26], Morton [27], Roos, Stynes, and Tobiska [33], and the references therein. Nevertheless, singularly perturbed problems remain difficult to solve numerically.

The least-squares finite element method (LSFEM) is a general methodology, which is based on the minimization of the residuals in a least-squares sense. The method, for linear differential equations, leads to symmetric positive-definite algebraic systems which can be efficiently solved by iterative methods. Continuous LSFEMs have been applied to solve convection-reaction-diffusion problems; see, e.g., [4, 6, 8, 10, 11, 14, 16, 21, 22, 30, 31]. For more details on the theory of LSFEMs, we refer to the review paper of Bochev and Gunzburger [5] and the book by Jiang [20].

Although the least-squares method shows many attractive features, its use is restricted by several disadvantages [7, 10, 11]. In particular, comparing with the Galerkin method, the least-squares based weak formulation requires higher regularity

*Received by the editors August 16, 2007; accepted for publication (in revised form) June 23, 2008; published electronically October 24, 2008.

<http://www.siam.org/journals/sinum/47-1/70026.html>

[†]Department of Mathematical and Physical Sciences, Texas A & M International University, Laredo, TX 78041-1900 (rlin@tamui.edu).

of finite element spaces. In addition, as we will illustrate in section 5, the standard least-squares method is inefficient for solving singularly perturbed problems, especially within and near the layers.

Discontinuous approximation spaces have been used to discretize least-squares formulations for solving a variety of problems, which can remove additional regularity requirement of the LSFEM. Cao and Gunzburger [12] used, for the first time in the literature, least-squares methods with discontinuous elements to treat interface problems. Gerritsma and Proot [17] derived a discontinuous least-squares spectral element method for a sample first order ordinary differential equation. Bensow and Larson applied discontinuous LSFEMs to elliptic problems [2] and div-curl problems [3] with boundary singularities. In all of these works, special least-squares functionals are proposed in discontinuous finite element spaces. On the other hand, the least-squares technique has also been used as a stabilizer of DG methods. For instance, Houston, Jensen, and Süli [19] investigated a general family of hp -discontinuous Galerkin FEMs with least-squares stabilization for symmetric systems of first-order partial differential equations.

Recently, the author proposed a discontinuously discretized LSFEM for 1D singularly perturbed reaction-diffusion problems with constant coefficients [24]. We hereby extend the method and develop a robust and stable numerical approach for more general singularly perturbed reaction-diffusion problems in 1D and 2D spaces. We will demonstrate the efficiency of our methods both theoretically and numerically. Numerical comparison studies between the proposed method and the local discontinuous Galerkin (LDG) method, the continuous LSFEM, and the discontinuous LSFEM indicate that the method is a promising alternative to existing schemes.

This paper is organized as follows. Section 2 introduces definitions and notations used in this paper. In section 3, we present the singularly perturbed problem and its least-squares variational formulation. The LDG method is utilized for the associated discrete problems. We prove coercivity of the bilinear forms in an associated energy norm. An adaptive method is also provided in this section to reduce computational cost. In section 4, a priori error estimate results are presented in one and two spatial dimensions. In section 5, numerical examples are given, which verify the theoretical results. Some comparisons of our method to other methods are included. Conclusions are drawn in section 6.

2. Notations. Throughout this paper, we shall use C to denote a generic positive constant which is independent of ϵ and the mesh used. Vectors and scalars are denoted by bold and plain letters, respectively.

We will denote the inner products in $L^2(\Omega)$ and product spaces of $L^2(\Omega)$ by (\cdot, \cdot) . For $1 \leq p \leq \infty$ and $s \geq 0$, we use the standard notation for the Sobolev space $W_p^s(\Omega)$ with the norm $\|\cdot\|_{W_p^s(\Omega)}$ and the seminorm $|\cdot|_{W_p^s(\Omega)}$. $H^s(\Omega)$ is used to stand for the space $W_2^s(\Omega)$, whose norm and seminorm are denoted by $\|\cdot\|_{s,\Omega}$ and $|\cdot|_{s,\Omega}$, respectively. When no confusion may arise, the measure Ω will be omitted from the above norm designations. We recall the space $H_0^1(\Omega)$ consisting of all functions in $H^1(\Omega)$ that vanish on the boundary $\partial\Omega$, and the space

$$H(\text{div}; \Omega) = \{\mathbf{q} \in [L^2(\Omega)]^d : \nabla \cdot \mathbf{q} \in L^2(\Omega)\}$$

with corresponding norm

$$\|\mathbf{q}\|_{H(\text{div}; \Omega)}^2 = \|\nabla \cdot \mathbf{q}\|_0^2 + \|\mathbf{q}\|_0^2.$$

Here, we denote also the norms on product spaces of $H^s(\Omega)$ by $\|\cdot\|_{s,\Omega}$, or simply $\|\cdot\|_s$, where there is no chance for ambiguity. We define further the vector function space

$$\mathbf{H}(\Omega) = H(\operatorname{div}; \Omega) \times H_0^1(\Omega)$$

with norms

$$\|\mathbf{v}\|_0^2 = \|\mathbf{q}\|_0^2 + \|v\|_0^2 \quad \text{and} \quad \|\mathbf{v}\|_{\mathbf{H}(\Omega)}^2 = \|\mathbf{q}\|_{H(\operatorname{div}; \Omega)}^2 + \|v\|_1^2,$$

where, and in the remainder of this paper, the vector valued functions \mathbf{u} , \mathbf{v} , and \mathbf{w} have components

$$\mathbf{u} = \begin{bmatrix} \mathbf{p} \\ u \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} \mathbf{q} \\ v \end{bmatrix}, \quad \text{and} \quad \mathbf{w} = \begin{bmatrix} \mathbf{r} \\ w \end{bmatrix},$$

respectively, unless otherwise specified. As we shall see in section 3, \mathbf{p} , \mathbf{q} , and \mathbf{r} are vector functions of dimension d , which are corresponding to the gradients of u , v , and w , respectively.

Let u be a solution of (1.1). The following estimates hold (see, e.g., [34]):

$$\|u\|_s \leq \frac{C}{\epsilon^s} \|f\|_0$$

for $0 < \epsilon \leq 1$ and $s = 0, 1$, or 2 . These estimates are, in fact, sharp [35]. Thus the standard norm $\|u\|_1$ or $\|u\|_2$ does not provide an informative gauge when ϵ is small. It is natural to introduce the following ϵ -dependent norm in $H^1(\Omega)$ [33]:

$$\|v\|_{1,\epsilon}^2 = \epsilon^2 |v|_1^2 + \|v\|_0^2.$$

In addition, we present ϵ -dependent norms in $\mathbf{H}(\Omega)$ as

$$\begin{aligned} \|\mathbf{v}\|_{0,\epsilon}^2 &= \epsilon^2 \|\mathbf{q}\|_0^2 + \|v\|_0^2, \\ \|\mathbf{v}\|_{1,\epsilon}^2 &= \epsilon^4 \|\nabla \cdot \mathbf{q}\|_0^2 + \epsilon^2 |v|_1^2 + \|\mathbf{v}\|_{0,\epsilon}^2. \end{aligned}$$

Let $\mathcal{T}_h = \{\Omega_k\}_{k=1}^M$ be a shape regular triangulation on Ω with mesh size h . Let \mathcal{E} be the union of the boundaries of all elements Ω_k associated with the partition \mathcal{T}_h , and let $\mathcal{E}_{int} \subset \mathcal{E}$ be the set of all interior edges contained in Ω . Note that an anisotropic mesh (e.g., Shishkin mesh) will certainly improve numerical results, which, nevertheless, is not necessary.

We use the following broken Sobolev spaces:

$$H^s(\mathcal{T}_h) = \{v \in L^2(\Omega) : v|_{\Omega_k} \in H^s(\Omega_k), k = 1, \dots, M\},$$

where $H^s(\Omega_k)$ is the Sobolev space of order s on Ω_k , $s \geq 0$. The inner products and norms defined above can be taken over the elements Ω_k , which are denoted by $(\cdot, \cdot)_{\Omega_k}$ and $\|\cdot\|_{s,\Omega_k}$, respectively. For $v \in H^s(\mathcal{T}_h)$, we define its norms and seminorms as

$$(2.1) \quad \|v\|_s^2 = \sum_{k=1}^M \|v\|_{s,\Omega_k}^2, \quad |v|_s^2 = \sum_{k=1}^M |v|_{s,\Omega_k}^2,$$

$$(2.2) \quad \|v\|_{1,\epsilon}^2 = \sum_{k=1}^M \|v\|_{1,\epsilon,\Omega_k}^2,$$

where $s \geq 0$ and

$$\|v\|_{1,\epsilon,\Omega_k}^2 = \epsilon^2 \|v\|_{1,\Omega_k}^2 + \|v\|_{0,\Omega_k}^2.$$

We define also the space

$$\mathbf{H}(\mathcal{T}_h) = \{[\mathbf{q}, v]^T \in H(\operatorname{div}; \mathcal{T}_h) \times H^1(\mathcal{T}_h) : v|_{\partial\Omega} = 0\},$$

where

$$H(\operatorname{div}; \mathcal{T}_h) = \{\mathbf{q} \in [H^1(\mathcal{T}_h)]^d : \mathbf{q}|_{\Omega_k} \in [L^2(\Omega_k)]^d, \nabla \cdot \mathbf{q}|_{\Omega_k} \in L^2(\Omega_k)\}.$$

The inner products and norms in $\mathbf{H}(\mathcal{T}_h)$ can be defined analogously. In particular, for $\mathbf{v} \in \mathbf{H}(\mathcal{T}_h)$, we define

$$(2.3) \quad \|\mathbf{v}\|_{0,\epsilon}^2 = \sum_{k=1}^M \|\mathbf{v}\|_{0,\epsilon,\Omega_k}^2, \quad \|\mathbf{v}\|_{1,\epsilon}^2 = \sum_{k=1}^M \|\mathbf{v}\|_{1,\epsilon,\Omega_k}^2,$$

where

$$\begin{aligned} \|\mathbf{v}\|_{0,\epsilon,\Omega_k}^2 &= \epsilon^2 \|\mathbf{q}\|_{0,\Omega_k}^2 + \|v\|_{0,\Omega_k}^2, \\ \|\mathbf{v}\|_{1,\epsilon,\Omega_k}^2 &= \epsilon^4 \|\nabla \cdot \mathbf{q}\|_{0,\Omega_k}^2 + \epsilon^2 \|v\|_{1,\Omega_k}^2 + \|\mathbf{v}\|_{0,\epsilon,\Omega_k}^2. \end{aligned}$$

In (2.1)–(2.3), we use the same norm notations as in the continuous Sobolev spaces, which will cause no ambiguity.

For any element $\Omega_k \in \mathcal{T}_h$ and an edge $e \in \partial\Omega_k \cap \mathcal{E}_{int}$, let $\Omega_{k',e}$ be the unique element sharing e with Ω_k . Let $\mathbf{n}_{k,e}^+$ and $\mathbf{n}_{k,e}^-$ be the outward normal unit vectors of Ω_k and $\Omega_{k',e}$ to e , respectively. For $v \in H^1(\mathcal{T}_h)$, we denote $v_{k,e}^+$ and $v_{k,e}^-$, the interior and outer traces of v on e , with respect to Ω_k , respectively. Note that $v_{k,e}^-$ and $v_{k',e}^+$ are selfsame. Similarly, we can define traces $\mathbf{q}_{k,e}^+$ and $\mathbf{q}_{k,e}^-$ for $\mathbf{q} \in [H^1(\mathcal{T}_h)]^d$. As in [1], we define the *average* and *jump* of v and \mathbf{q} to element Ω_k across e by

$$\begin{aligned} \{v\}_{k,e} &= \frac{1}{2}(v_{k,e}^+ + v_{k,e}^-), & [[v]]_{k,e} &= v_{k,e}^+ \mathbf{n}_{k,e}^+ + v_{k,e}^- \mathbf{n}_{k,e}^-; \\ \{\mathbf{q}\}_{k,e} &= \frac{1}{2}(\mathbf{q}_{k,e}^+ + \mathbf{q}_{k,e}^-), & [[\mathbf{q}]]_{k,e} &= \mathbf{q}_{k,e}^+ \cdot \mathbf{n}_{k,e}^+ + \mathbf{q}_{k,e}^- \cdot \mathbf{n}_{k,e}^-. \end{aligned}$$

We denote v_k^+ , v_k^- , \mathbf{q}_k^+ , and \mathbf{q}_k^- , the interior and outer traces of v and \mathbf{q} , along $\partial\Omega_k$ with respect to Ω_k , respectively.

Finally, we define the finite element space associated with \mathcal{T}_h as

$$\mathbf{V}_h = [V^h]^d \times V_0^h \subset \mathbf{H}(\mathcal{T}_h),$$

where $V^h \subset H^1(\mathcal{T}_h)$ is the space of piecewise linear (1D and 2D) or bilinear (2D) polynomials allowing discontinuity along interelement edges, and V_0^h is the subspace of V^h , which consists of functions vanishing on the boundary $\partial\Omega$.

3. Least-squares finite element approximations. We rewrite (1.1) as the following system of first-order equations:

$$(3.1) \quad \begin{cases} \mathbf{p} - \nabla u = \mathbf{0} & \text{in } \Omega, \\ -\epsilon^2 \nabla \cdot \mathbf{p} + cu = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases}$$

For $\mathbf{u} \in \mathbf{H}(\Omega)$, define

$$A\mathbf{u} = \begin{bmatrix} \epsilon\sqrt{c_0}(\mathbf{p} - \nabla u) \\ -\epsilon^2\nabla \cdot \mathbf{p} + cu \end{bmatrix} \quad \text{and} \quad \mathbf{f} = \begin{bmatrix} \mathbf{0} \\ f \end{bmatrix}.$$

Notice that a weight $\epsilon\sqrt{c_0}$ is employed in the first component of $A\mathbf{u}$, where the same concerns as in the definitions of ϵ -dependent norms are reflected. Equation (3.1) can be written as

$$(3.2) \quad A\mathbf{u} = \mathbf{f} \quad \text{in } \Omega.$$

The homogenous boundary condition in (3.1) is satisfied since $\mathbf{u} \in \mathbf{H}(\Omega)$ implies $u \in H_0^1(\Omega)$. Note that problem (1.1) has a unique solution u in $H_0^1(\Omega) \cap H_{loc}^2(\Omega)$, which is, moreover, in $H_0^2(\Omega)$ when $\partial\Omega$ is sufficiently smooth [18]. We will assume from now on that problem (3.2) has a unique solution $\mathbf{u} \in \mathbf{H}(\Omega)$.

3.1. The least-squares formulation. Consider the least-squares functional \mathcal{J} in $\mathbf{H}(\Omega)$ defined by

$$\mathcal{J}(\mathbf{v}; f) = \|A\mathbf{v} - \mathbf{f}\|_0^2 = (A\mathbf{v} - \mathbf{f}, A\mathbf{v} - \mathbf{f}).$$

The least-squares method reads: find $\mathbf{u} \in \mathbf{H}(\Omega)$ such that

$$\mathcal{J}(\mathbf{u}; f) = \inf_{\mathbf{v} \in \mathbf{H}(\Omega)} \mathcal{J}(\mathbf{v}; f).$$

A necessary condition for \mathbf{u} to be a minimizer of the functional \mathcal{J} is that its first variation vanishes at \mathbf{u} , i.e.,

$$\lim_{t \rightarrow 0} \frac{d}{dt} \mathcal{J}(\mathbf{u} + t\mathbf{v}; f) = 2(A\mathbf{u} - \mathbf{f}, A\mathbf{v}) = 0 \quad \forall \mathbf{v} \in \mathbf{H}(\Omega).$$

The corresponding least-squares variational formulation for problem (1.1) thus follows: find $\mathbf{u} \in \mathbf{H}(\Omega)$ such that

$$(3.3) \quad B(\mathbf{u}, \mathbf{v}) = L(\mathbf{v}) \quad \forall \mathbf{v} \in \mathbf{H}(\Omega),$$

where the bilinear form $B : \mathbf{H}(\Omega) \times \mathbf{H}(\Omega) \rightarrow \mathbb{R}$ and linear functional $L : \mathbf{H}(\Omega) \rightarrow \mathbb{R}$ are defined as

$$(3.4) \quad B(\mathbf{w}, \mathbf{v}) = (A\mathbf{w}, A\mathbf{v}) = c_0\epsilon^2(\mathbf{r} - \nabla w, \mathbf{q} - \nabla v) + (-\epsilon^2\nabla \cdot \mathbf{r} + cw, -\epsilon^2\nabla \cdot \mathbf{q} + cv),$$

$$(3.5) \quad L(\mathbf{v}) = (\mathbf{f}, A\mathbf{v}) = (f, -\epsilon^2\nabla \cdot \mathbf{q} + cv)$$

for all $\mathbf{v}, \mathbf{w} \in \mathbf{H}(\Omega)$.

It is clear that the bilinear form $B(\cdot, \cdot)$ defined in (3.4) is symmetric. In addition, we have the following boundedness and coercivity results.

THEOREM 3.1. (i) *There exists a constant $C > 0$ independent of ϵ such that*

$$(3.6) \quad |B(\mathbf{w}, \mathbf{v})| \leq C\|\mathbf{w}\|_{1,\epsilon}\|\mathbf{v}\|_{1,\epsilon} \quad \forall \mathbf{v}, \mathbf{w} \in \mathbf{H}(\Omega).$$

(ii) *There exists a constant $\alpha > 0$ independent of ϵ such that*

$$(3.7) \quad B(\mathbf{v}, \mathbf{v}) \geq \alpha\|\mathbf{v}\|_{1,\epsilon}^2 \quad \forall \mathbf{v} \in \mathbf{H}(\Omega).$$

Proof. The boundedness property (3.6) is a direct consequence of the Cauchy-Schwartz and triangle inequalities.

We next prove the coercivity (3.7). From (3.4), one has

$$B(\mathbf{v}, \mathbf{v}) = c_0 \epsilon^2 \|\mathbf{q} - \nabla v\|_0^2 + \|\epsilon^2 \nabla \cdot \mathbf{q} + cv\|_0^2.$$

Using integration by parts and homogenous boundary conditions of v , we have

$$(3.8) \quad 2B(\mathbf{v}, \mathbf{v}) \geq \epsilon^4 \|\nabla \cdot \mathbf{q}\|_0^2 + \|cv\|_0^2 + c_0 \epsilon^2 \|\mathbf{q}\|_0^2 + c_0 \epsilon^2 \|\nabla v\|_0^2 + 2(\epsilon^2 \nabla \cdot \mathbf{q}, (c_0 - c)v).$$

Choose a constant $\delta = \frac{c_1}{c_0 + c_1}$. Then $0 < \delta < 1$. It follows that

$$(3.9) \quad \delta \epsilon^4 \|\nabla \cdot \mathbf{q}\|_0^2 + 2(\epsilon^2 \nabla \cdot \mathbf{q}, (c_0 - c)v) + \frac{1}{\delta} \|(c_0 - c)v\|_0^2 = \left\| \epsilon^2 \sqrt{\delta} \nabla \cdot \mathbf{q} + \frac{1}{\sqrt{\delta}} (c_0 - c)v \right\|_0^2 \geq 0.$$

By (3.8) and (3.9), we get

$$(3.10) \quad \begin{aligned} 2B(\mathbf{v}, \mathbf{v}) &\geq (1 - \delta) \epsilon^4 \|\nabla \cdot \mathbf{q}\|_0^2 + \|cv\|_0^2 - \frac{1}{\delta} \|(c_0 - c)v\|_0^2 + c_0 \epsilon^2 \|\mathbf{q}\|_0^2 + c_0 \epsilon^2 \|\nabla v\|_0^2 \\ &= (1 - \delta) \epsilon^4 \|\nabla \cdot \mathbf{q}\|_0^2 + \frac{1}{\delta} ((\delta c^2 - (c_0 - c)^2) v, v) + c_0 \epsilon^2 \|\mathbf{q}\|_0^2 + c_0 \epsilon^2 \|\nabla v\|_0^2. \end{aligned}$$

Note that

$$(\delta - 1)c + c_0 \geq -\frac{c_0}{c_0 + c_1} c_1 + c_0 = \frac{c_0^2}{c_0 + c_1}.$$

Hence

$$(3.11) \quad \delta c^2 - (c_0 - c)^2 = ((\delta - 1)c + c_0)c + c_0(c - c_0) \geq \frac{c_0^2}{c_0 + c_1} c \geq \frac{c_0^3}{c_0 + c_1}.$$

Finally, by (3.10) and (3.11), we arrive at

$$B(\mathbf{v}, \mathbf{v}) \geq \frac{1 - \delta}{2} \epsilon^4 \|\nabla \cdot \mathbf{q}\|_0^2 + \frac{c_0^3}{2c_1} \|v\|_0^2 + \frac{c_0}{2} \epsilon^2 \|\mathbf{q}\|_0^2 + \frac{c_0}{2} \epsilon^2 \|\nabla v\|_0^2,$$

which implies (3.7) by taking $\alpha = \min\{\frac{c_0}{2(c_0 + c_1)}, \frac{c_0^3}{2c_1}, \frac{c_0}{2}\}$. \square

The following result is a straightforward consequence of Theorem 3.1.

PROPOSITION 3.2. *The least-squares variational problem (3.3) is well posed.*

3.2. LDG discretization. We next discretize the least-squares formulation (3.3) with the LDG method [15]. Using integration by parts in each element, we get the LDG approximation for the least-squares variational formulation (3.3) as follows, by employing a process which will be referred to as the LDG-LS method in this paper. Find $\mathbf{u}_h = [\mathbf{p}_h, u_h]^T \in \mathbf{V}_h$ such that

$$(3.12) \quad B_h(\mathbf{u}_h, \mathbf{v}) = L_h(\mathbf{v}) \quad \forall \mathbf{v} \in \mathbf{V}_h,$$

where the bilinear form (or *primal form*) $B_h : \mathbf{H}(\mathcal{T}_h) \times \mathbf{H}(\mathcal{T}_h) \rightarrow \mathbb{R}$ and the linear form $L_h : \mathbf{H}(\mathcal{T}_h) \rightarrow \mathbb{R}$ are defined by

$$(3.13) \quad \begin{aligned} B_h(\mathbf{w}, \mathbf{v}) &= \sum_{k=1}^M \int_{\Omega_k} (c_0 \epsilon^2 \mathbf{r} \cdot \mathbf{q} + c_0 \epsilon^2 \nabla w \cdot \nabla v + \epsilon^4 \nabla \cdot \mathbf{r} \nabla \cdot \mathbf{q} + c^2 wv) \, dx \\ &+ \epsilon^2 \sum_{k=1}^M \int_{\Omega_k} (\mathbf{r} \cdot \nabla ((c - c_0)v) + (c_0 - c)w \nabla \cdot \mathbf{q}) \, dx \\ &- \epsilon^2 \sum_{k=1}^M \int_{\partial \Omega_k} (c_0 \widehat{w}_k \mathbf{q}_k^+ \cdot \mathbf{n}_k^+ + cv_k^+ \widehat{\mathbf{r}}_k \cdot \mathbf{n}_k^+) \, ds \end{aligned}$$

and

$$(3.14) \quad L_h(\mathbf{v}) = \sum_{k=1}^M (\mathbf{f}, A\mathbf{v})_{\Omega_k} = \sum_{k=1}^M (f, -\epsilon^2 \nabla \cdot \mathbf{q} + cv)_{\Omega_k}$$

for all $\mathbf{v}, \mathbf{w} \in \mathbf{H}(\mathcal{T}_h)$, respectively. In (3.13), \mathbf{n}_k^+ is the outward normal unit vector to $\partial\Omega_k$, \widehat{w} , and $\widehat{\mathbf{r}}$ are numerical fluxes defined by

$$(3.15) \quad \widehat{w}_{k,e} = \{w\}_{k,e} + [[w]]_{k,e} \cdot \boldsymbol{\lambda}_{k,e},$$

$$(3.16) \quad \widehat{\mathbf{r}}_{k,e} = \{\mathbf{r}\}_{k,e} + [[\mathbf{r}]]_{k,e} \boldsymbol{\mu}_{k,e}$$

on $e \in \mathcal{E}_{int}$ and by

$$(3.17) \quad \widehat{w}_{k,e} = w|_e,$$

$$(3.18) \quad \widehat{\mathbf{r}}_{k,e} = \mathbf{r}|_e$$

on $e \in \mathcal{E} \setminus \mathcal{E}_{int}$, where parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are vector-valued functions. Note that \mathbf{V}_h does not need to be a subspace of $\mathbf{H}(\Omega)$. Our method is therefore nonconforming in this sense.

Remark 3.1. For continuous functions, the numerical fluxes defined in (3.15)–(3.18) are the restrictions of the corresponding functions on associated interelement edges. A straightforward computation shows that $B_h(\cdot, \cdot)$ coincides with $B(\cdot, \cdot)$ in $\mathbf{H}(\Omega) \times \mathbf{H}(\Omega)$.

Remark 3.2. To see the difference between the *discontinuous LSFEMs* in [2, 3, 17] and the LDG-LS method developed in this paper, we note that, for discontinuous LSFEMs in the papers cited, special least-squares functionals are defined on discontinuous spaces, which lead to symmetric weak forms from minimization of the corresponding functionals. The LDG-LS method, on the other hand, discretizes a standard least-squares functional with the LDG method, whose formulation is in general nonsymmetric.

Remark 3.3. If $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are properly selected, then $B_h(\cdot, \cdot)$ preserves many good properties of $B(\cdot, \cdot)$. For instance, letting $\boldsymbol{\lambda}_k = \mathbf{n}_k^+ / 2$ and $\boldsymbol{\mu}_k = \mathbf{n}_k^+ / 2$, it is easy to verify that $\widehat{w}_k = w_k^+$, $\widehat{\mathbf{r}}_k \cdot \mathbf{n}_k^+ = \mathbf{r}_k^+ \cdot \mathbf{n}_k^+$, and hence

$$B_h(\mathbf{w}, \mathbf{v}) = \sum_{k=1}^M (A\mathbf{w}, A\mathbf{v})_{\Omega_k},$$

which is symmetric.

Before considering the coercivity of $B_h(\cdot, \cdot)$, we recall the trace inequalities when $d = 2$. For $\mathbf{v} \in \mathbf{V}_h$, we have (see, e.g., [9])

$$(3.19) \quad \|v_k^+\|_{0, \partial\Omega_k} \leq C_1 h^{1/2} |v|_{1, \Omega_k},$$

$$(3.20) \quad \|\mathbf{q}_k^+\|_{0, \partial\Omega_k} \leq C_2 h^{-1/2} \|\mathbf{q}\|_{0, \Omega_k},$$

where $\|\cdot\|_{0, \partial\Omega_k}$ is the $H^0(\partial\Omega_k)$ norm, and C_1 and C_2 are fixed constants satisfying (3.19) and (3.20) for all k , respectively. Then we have the following results.

THEOREM 3.3. *Let $B_h(\cdot, \cdot)$ be the bilinear form defined in (3.13) with $\boldsymbol{\lambda}_k = \mathbf{0}$ and $\boldsymbol{\mu}_k = \mathbf{0}$.*

(i) *There exists a constant $C > 0$ independent of ϵ such that*

$$(3.21) \quad |B_h(\mathbf{w}, \mathbf{v})| \leq C \|\mathbf{w}\|_{1, \epsilon} \|\mathbf{v}\|_{1, \epsilon} \quad \forall \mathbf{v}, \mathbf{w} \in \mathbf{V}_h.$$

(ii) Assume that $\min\{\frac{c_0}{c_0+c_1}, \frac{c_0^3}{c_1^3}, c_0\} > (c_1 - c_0)C_1C_2$ when $d = 2$. There exists a constant $\alpha^* > 0$ independent of ϵ such that

$$(3.22) \quad B_h(\mathbf{v}, \mathbf{v}) \geq \alpha^* \|\mathbf{v}\|_{1,\epsilon}^2 \quad \forall \mathbf{v} \in \mathbf{V}_h.$$

Proof. We first prove the coercivity of $B_h(\cdot, \cdot)$. Comparing (3.4) and (3.13), we have

$$B_h(\mathbf{w}, \mathbf{v}) = \sum_{k=1}^M \left(B(\mathbf{w}, \mathbf{v})_k + \epsilon^2 \int_{\partial\Omega_k} (c_0(w_k^+ - \widehat{w}_k) \mathbf{q}_k^+ \cdot \mathbf{n}_k^+ + cv_k^+(\mathbf{r}_k^+ - \widehat{\mathbf{r}}_k) \cdot \mathbf{n}_k^+) ds \right)$$

for all $\mathbf{v}, \mathbf{w} \in \mathbf{V}_h$, where

$$B(\mathbf{w}, \mathbf{v})_k = (A\mathbf{w}, A\mathbf{v})_{\Omega_k}$$

is the restriction of $B(\mathbf{w}, \mathbf{v})$ to the element Ω_k . Since $\boldsymbol{\lambda}_k = \mathbf{0}$ and $\boldsymbol{\mu}_k = \mathbf{0}$, we get

$$B_h(\mathbf{w}, \mathbf{v}) = \sum_{k=1}^M \left(B(\mathbf{w}, \mathbf{v})_k + \epsilon^2 \int_{\partial\Omega_k} c_0(w_k^+ - \{w\}_k) \mathbf{q}_k^+ \cdot \mathbf{n}_k^+ ds + \epsilon^2 \int_{\partial\Omega_k} cv_k^+(\mathbf{r}_k^+ - \{\mathbf{r}\}_k) \cdot \mathbf{n}_k^+ ds \right).$$

It follows that

$$(3.23) \quad B_h(\mathbf{v}, \mathbf{v}) = \sum_{k=1}^M B(\mathbf{v}, \mathbf{v})_k + c_0\epsilon^2 \sum_{k=1}^M \sum_{e \in \mathcal{C}\partial\Omega_k} \int_e I_{\Omega_{k,e}} ds + \epsilon^2 \sum_{k=1}^M \sum_{e \in \mathcal{C}\partial\Omega_k} \int_e J_{\Omega_{k,e}} ds,$$

where $I_{\Omega_{k,e}}$ and $J_{\Omega_{k,e}}$ are the interior traces of $(v - \{v\})\mathbf{q} \cdot \mathbf{n}_k^+ + v(\mathbf{q} - \{\mathbf{q}\}) \cdot \mathbf{n}_k^+$ and $(c - c_0)v(\mathbf{q} - \{\mathbf{q}\}) \cdot \mathbf{n}_k^+$ on edge e with respect to Ω_k , respectively.

For any element Ω_k , if an edge $e \in \mathcal{E}_{int}$, then $v|_e = 0$ and $\{v\}|_e = 0$, and hence $I_{\Omega_{k,e}} = 0$ and $J_{\Omega_{k,e}} = 0$, which contribute 0 in the last two summations in (3.23). On the other hand, if the edge $e \in \mathcal{E}_{int}$, then there is another element $\Omega_{k',e}$ sharing e with Ω_k , and there are two terms associated with the edge in each double summation in (3.23). In particular,

$$(3.24) \quad \int_e I_{\Omega_{k,e}} ds = \int_e \left(\frac{1}{2}(v_{k,e}^+ - v_{k,e}^-) \mathbf{q}_{k,e}^+ \cdot \mathbf{n}_{k,e}^+ + \frac{1}{2}v_{k,e}^+ (\mathbf{q}_{k,e}^+ - \mathbf{q}_{k,e}^-) \cdot \mathbf{n}_{k,e}^+ \right) ds,$$

$$(3.25) \quad \int_e I_{\Omega_{k',e}} ds = \int_e \left(\frac{1}{2}(v_{k,e}^- - v_{k,e}^+) \mathbf{q}_{k,e}^- \cdot \mathbf{n}_{k,e}^- + \frac{1}{2}v_{k,e}^- (\mathbf{q}_{k,e}^- - \mathbf{q}_{k,e}^+) \cdot \mathbf{n}_{k,e}^- \right) ds.$$

Adding (3.24) and (3.25), we get

$$\int_e (I_{\Omega_{k,e}} + I_{\Omega_{k',e}}) ds = \int_e (v_{k,e}^+ \mathbf{q}_{k,e}^+ \cdot \mathbf{n}_{k,e}^+ + v_{k,e}^- \mathbf{q}_{k,e}^- \cdot \mathbf{n}_{k,e}^-) ds,$$

since $\mathbf{n}_{k,e}^+ + \mathbf{n}_{k,e}^- = \mathbf{0}$. We thus conclude that

$$\sum_{k=1}^M \sum_{e \in \mathcal{C}\partial\Omega_k} \int_e I_{\Omega_{k,e}} ds = \sum_{k=1}^M \int_{\partial\Omega_k} v_k^+ \mathbf{q}_k^+ \cdot \mathbf{n}_k^+ ds.$$

Similarly,

$$(3.26) \quad \int_e J_{\Omega_{k,e}} ds = \int_e \frac{1}{2} (c - c_0) v_{k,e}^+ (\mathbf{q}_{k,e}^+ - \mathbf{q}_{k,e}^-) \cdot \mathbf{n}_{k,e}^+ ds,$$

$$(3.27) \quad \int_e J_{\Omega_{k',e}} ds = \int_e \frac{1}{2} (c - c_0) v_{k,e}^- (\mathbf{q}_{k,e}^- - \mathbf{q}_{k,e}^+) \cdot \mathbf{n}_{k,e}^- ds.$$

By adding (3.26) and (3.27), we then have

$$\int_e (J_{\Omega_{k,e}} + J_{\Omega_{k',e}}) ds = \int_e (c - c_0) \{v\}_{k,e} [[\mathbf{q}]]_{k,e} ds,$$

and hence

$$\sum_{k=1}^M \sum_{e \subset \partial \Omega_k} \int_e J_{\Omega_{k,e}} ds = \sum_{k=1}^M \frac{1}{2} \int_{\partial \Omega_k} (c - c_0) \{v\}_k [[\mathbf{q}]]_k ds.$$

Therefore, (3.23) reads

$$(3.28) \quad B_h(\mathbf{v}, \mathbf{v}) = \sum_{k=1}^M R_k + \sum_{k=1}^M \frac{\epsilon^2}{2} \int_{\partial \Omega_k} (c - c_0) \{v\}_k [[\mathbf{q}]]_k ds,$$

where

$$R_k = B(\mathbf{v}, \mathbf{v})_k + c_0 \epsilon^2 \int_{\partial \Omega_k} v_k^+ \mathbf{q}_k^+ \cdot \mathbf{n}_k^+ ds.$$

Now, by Theorem 3.1, we have

$$(3.29) \quad R_k \geq \alpha \|\mathbf{v}\|_{1,\epsilon,\Omega_k}^2$$

for each element Ω_k , where $\alpha = \min\{\frac{c_0}{2(c_0+c_1)}, \frac{c_0^3}{2c_1}, \frac{c_0}{2}\}$. Consider the second summation in (3.28). When $d = 1$, a straightforward calculation shows that (see [24])

$$(3.30) \quad \sum_{k=1}^M \int_{\partial \Omega_k} (c - c_0) \{v\}_k [[\mathbf{q}]]_k ds = 0.$$

When $d = 2$, since the interior and outer traces of $\{v\}[[\mathbf{q}]]$ on each edge are equal, we may rearrange terms in the sum. Applying Cauchy–Schwartz inequality, it follows that

$$\begin{aligned} & \left| \sum_{k=1}^M \int_{\partial \Omega_k} (c - c_0) \{v\}_k [[\mathbf{q}]]_k ds \right| = 2 \left| \sum_{k=1}^M \int_{\partial \Omega_k} (c - c_0) v_k^+ \mathbf{q}_k^+ \cdot \mathbf{n}_k^+ ds \right| \\ & \leq 2(c_1 - c_0) \sum_{k=1}^M \|v_k^+\|_{0,\partial \Omega_k} \|\mathbf{q}_k^+ \cdot \mathbf{n}_k^+\|_{0,\partial \Omega_k} \leq 2(c_1 - c_0) \sum_{k=1}^M \|v_k^+\|_{0,\partial \Omega_k} \|\mathbf{q}_k^+\|_{0,\partial \Omega_k}. \end{aligned}$$

Using the trace inequalities (3.19) and (3.20), we obtain

$$(3.31) \quad \sum_{k=1}^M \frac{\epsilon^2}{2} \int_{\partial \Omega_k} (c - c_0) \{v\}_k [[\mathbf{q}]]_k ds \geq - \sum_{k=1}^M C_3 \epsilon^2 |v|_{1,\Omega_k} \|\mathbf{q}\|_{0,\Omega_k},$$

where $C_3 = (c_1 - c_0)C_1C_2$. From (3.29)–(3.31) we conclude that, when $d = 1$,

$$B_h(\mathbf{v}, \mathbf{v}) \geq \sum_{k=1}^M (\alpha \|\mathbf{v}\|_{1,\epsilon,\Omega_k}^2) \geq \alpha \|\mathbf{v}\|_{1,\epsilon}^2,$$

and when $d = 2$,

$$B_h(\mathbf{v}, \mathbf{v}) \geq \sum_{k=1}^M (\alpha \|\mathbf{v}\|_{1,\epsilon,\Omega_k}^2 - C_3 \epsilon^2 |v|_{1,\Omega_k} \|\mathbf{q}\|_{0,\Omega_k}) \geq (\alpha - C_3/2) \|\mathbf{v}\|_{1,\epsilon}^2.$$

The coercivity result follows with $\alpha^* = \alpha$ when $d = 1$ and $\alpha^* = \alpha - C_3/2$ when $d = 2$.

The boundedness (3.21) is a direct consequence of the Cauchy–Schwartz, triangle, and trace inequalities. \square

In addition, we have the following analogue of Proposition 3.2.

PROPOSITION 3.4. *Assume that the conditions of Theorem 3.3 are fulfilled. The LDG-LS approximation problem (3.12) is well posed.*

Remark 3.4. Note that the boundedness result (3.21), as well as (3.6), is in the more desirable norm $\|\cdot\|_{1,\epsilon}$, comparing with the results in only standard H^1 norm from some other classical FEMs [27, 33]. This inequality leads naturally to a Céa type error estimate and a uniform error estimate in the $\|\cdot\|_{1,\epsilon}$ norm, which will be presented in section 4.

Remark 3.5. $\boldsymbol{\lambda} = \mathbf{0}$ and $\boldsymbol{\mu} = \mathbf{0}$ are sufficient but not necessary conditions for the results of Theorem 3.3. In section 5, we provide numerical results with nonzero numerical flux parameters. Moreover, a proper selection of $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ will produce numerical solutions of better quality. Our examples show that better computational results are obtained if $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are pointing toward the layers.

Remark 3.6. The LDG-LS method (3.12) is stable and robust with zero numerical flux parameters. In addition, if c is a constant function, then the assumption $\min\{\frac{c_0}{c_0+c_1}, \frac{c_0^3}{c_1}, c_0\} > (c_1 - c_0)C_1C_2$ for coercivity (3.22) is naturally true.

3.3. The hybrid adaptive method. We divide the solution domain Ω into two regions: the regular solution region Ω_C and the layer region Ω_D . In Ω_C the exact solution is smooth and the derivatives of the exact solution can be bounded by a constant that is independent of ϵ , where we may use continuous elements. In Ω_D the exact solution has large derivatives thus motivating the use of the discontinuous method in the region. Conforming with the triangulation \mathcal{T}_h , we may define $Z_D = \bigcup_{\Omega_k \cap \bar{\Omega}_D \neq \emptyset} \bar{\Omega}_k$, the region consisting of elements covering Ω_D , and $Z_C = \Omega \setminus Z_D$.

Let \mathbf{V}_h^* be a subspace of \mathbf{V}_h , such that the basis functions of \mathbf{V}_h^* are continuous in Z_C . The hybrid LDG/continuous least-squares (LDG/C-LS) FEM is: find $\mathbf{u}_h \in \mathbf{V}_h^*$ such that

$$(3.32) \quad B_h^*(\mathbf{u}_h, \mathbf{v}) = L_h(\mathbf{v}) \quad \forall \mathbf{v} \in \mathbf{V}_h^*,$$

where the bilinear form $B_h^* : \mathbf{V}_h^* \times \mathbf{V}_h^* \rightarrow \mathbb{R}$ is defined by

$$\begin{aligned} B_h^*(\mathbf{w}, \mathbf{v}) &= \sum_{k=1}^M \int_{\Omega_k} (c_0 \epsilon^2 \mathbf{r} \cdot \mathbf{q} + c_0 \epsilon^2 \nabla w \cdot \nabla v + \epsilon^4 \nabla \cdot \mathbf{r} \nabla \cdot \mathbf{q} + c^2 w v) \, d\mathbf{x} \\ &+ \epsilon^2 \sum_{k=1}^M \int_{\Omega_k} (\mathbf{r} \cdot \nabla((c - c_0)v) + (c_0 - c)w \nabla \cdot \mathbf{q}) \, d\mathbf{x} \\ &- \epsilon^2 \sum_{k=1}^M \int_{\partial\Omega_k} (c_0 \widehat{w}_k \mathbf{q}_k^+ \cdot \mathbf{n}_k^+ + c \widehat{\mathbf{x}}_k \cdot \mathbf{n}_k^+ v_k^+) \, ds \end{aligned}$$

for all $\mathbf{v}, \mathbf{w} \in \mathbf{V}_h^*$, the linear form L_h is given in (3.14), and the numerical fluxes are given in (3.15)–(3.18).

The continuous and discontinuous discretizations for the least-squares formulation (3.3) are naturally combined in the LDG/C-LS approximation (3.32), which saves the extra degrees of freedom required by the LDG-LS method. Moreover, the coercivity of B_h^* in the $\|\cdot\|_{1,\epsilon}$ norm and the well posedness of problem (3.32) can be verified. A computational comparison between the two methods is made in section 5.

4. Error estimates. In this section, we present some a priori error estimate results.

PROPOSITION 4.1. *The bilinear form $B_h(\cdot, \cdot)$ defined by (3.13) is consistent.*

Proof. Let $\mathbf{u} \in \mathbf{H}(\Omega)$ solve the problem (3.3). By the definitions (3.15)–(3.18), we have $\{u\}_{k,e} = u|_e$, $[[u]]_{k,e} = 0$, and $\{\mathbf{p}\}_{k,e} = \mathbf{p}|_e$ on all $e \in \mathcal{E}$; and $[[\mathbf{p}]]_{k,e} = \mathbf{0}$ on all $e \in \mathcal{E}_{int}$. Therefore, it follows that

$$B_h(\mathbf{u}, \mathbf{v}) = \sum_{k=1}^M B(\mathbf{u}, \mathbf{v})_k = \sum_{k=1}^M (\mathbf{f}, A\mathbf{v})_{\Omega_k} = L_h(\mathbf{v})$$

for all $\mathbf{v} \in \mathbf{V}_h$. The desired result thus follows. \square

Let \mathbf{u} and \mathbf{u}_h solve problems (3.3) and (3.12), respectively. It follows from Proposition 4.1 that $B_h(\cdot, \cdot)$ satisfies the Galerkin orthogonality

$$(4.1) \quad B_h(\mathbf{u} - \mathbf{u}_h, \mathbf{v}) = 0 \quad \forall \mathbf{v} \in \mathbf{V}_h.$$

We assume the numerical flux parameters $\boldsymbol{\lambda} = \mathbf{0}$ and $\boldsymbol{\mu} = \mathbf{0}$. Then, by Theorem 3.3 and (4.1), we have

$$\begin{aligned} \alpha^* \|\mathbf{u} - \mathbf{u}_h\|_{1,\epsilon}^2 &\leq B_h(\mathbf{u} - \mathbf{u}_h, \mathbf{u} - \mathbf{u}_h) \\ &= B_h(\mathbf{u} - \mathbf{u}_h, \mathbf{u} - \mathbf{v}) \leq C \|\mathbf{u} - \mathbf{u}_h\|_{1,\epsilon} \|\mathbf{u} - \mathbf{v}\|_{1,\epsilon} \end{aligned}$$

for all $\mathbf{v} \in \mathbf{V}_h$, which implies the following Céa type error estimate.

THEOREM 4.2. *Let \mathbf{u} and \mathbf{u}_h be solutions to (3.3) and (3.12), respectively. Assume that the conditions of Theorem 3.3 are fulfilled. Then*

$$(4.2) \quad \|\mathbf{u} - \mathbf{u}_h\|_{1,\epsilon} \leq C \inf_{\mathbf{v} \in \mathbf{V}_h} \|\mathbf{u} - \mathbf{v}\|_{1,\epsilon}.$$

Thus, the LDG-LS FEM is optimal in the ϵ -dependent norm. Consequently, we get the following a priori error estimate.

THEOREM 4.3. *Let \mathbf{u} and \mathbf{u}_h be solutions to (3.3) and (3.12), respectively. Assume that $u \in H^3(\Omega)$ and the conditions of Theorem 3.3 are fulfilled. Then*

$$(4.3) \quad \|\mathbf{u} - \mathbf{u}_h\|_{1,\epsilon} \leq Ch^{1/2}.$$

Proof. Let $I_h u$ and $I_h \mathbf{p}$ be the standard linear or bilinear finite element interpolation of u and \mathbf{p} (i.e., ∇u), respectively. Then, from approximation theory (e.g., [9]), we have

$$(4.4) \quad \begin{aligned} \|u - I_h u\|_{W_\infty^0(\Omega)} &\leq Ch|u|_{W_\infty^1(\Omega)}, \\ \|\mathbf{p} - I_h \mathbf{p}\|_{W_\infty^1(\Omega)} &\leq Ch|\mathbf{p}|_{W_\infty^2(\Omega)}, \end{aligned}$$

where C denotes a constant independent of the mesh size h .

By the coercivity of $B_h(\cdot, \cdot)$, we have

$$(4.5) \quad \begin{aligned} \alpha^* \|\mathbf{u} - I_h \mathbf{u}\|_{1,\epsilon}^2 &\leq B_h(\mathbf{u} - I_h \mathbf{u}, \mathbf{u} - I_h \mathbf{u}) \\ &= B(\mathbf{u} - I_h \mathbf{u}, \mathbf{u} - I_h \mathbf{u}) \\ &= (A(\mathbf{u} - I_h \mathbf{u}), A(\mathbf{u} - I_h \mathbf{u})), \end{aligned}$$

where the first identity is due to the fact that $\mathbf{u} - I_h \mathbf{u}$ is continuous in Ω (see Remark 3.1). Noting that the interpolation operator is linear, it follows that

$$A(\mathbf{u} - I_h \mathbf{u}) = A\mathbf{u} - AI_h \mathbf{u} = A\mathbf{u} - I_h A\mathbf{u} = \mathbf{f} - I_h \mathbf{f}.$$

By approximation theory, we have

$$(4.6) \quad \|f - I_h f\|_{W_\infty^0(\Omega)} \leq C|f|_{W_\infty^0(\Omega)}.$$

Using the estimates (4.4) and (4.6), we obtain

$$(4.7) \quad \begin{aligned} (A(\mathbf{u} - I_h \mathbf{u}), A(\mathbf{u} - I_h \mathbf{u})) &= (A(\mathbf{u} - I_h \mathbf{u}), \mathbf{f} - I_h \mathbf{f}) \\ &\leq Ch(\epsilon^2 |\mathbf{p}|_{W_\infty^2(\Omega)} + |u|_{W_\infty^1(\Omega)}) |f|_{W_\infty^0(\Omega)}. \end{aligned}$$

By (4.5) and (4.7), we get

$$(4.8) \quad \|\mathbf{u} - I_h \mathbf{u}\|_{1,\epsilon} \leq Ch^{1/2},$$

since u , and hence \mathbf{p} , is sufficiently smooth. The desired result follows immediately from Theorem 4.2 and (4.8) by selecting \mathbf{v} as $I_h \mathbf{u}$ in (4.2). \square

Remark 4.1. As shown in [10, 11], for second-order elliptic problems, the first-order system LSFEM has an optimal error estimate of $\mathcal{O}(h)$ in the H^1 norm. See also [5]. For singularly perturbed problems described in this paper, however, this optimal convergence rate cannot be achieved (cf. [33]).

Remark 4.2. In [36] and [28], uniform error estimates of $\mathcal{O}(h^{1/2})$ are obtained in an ϵ -dependent norm for singularly perturbed elliptic problems in 1D and 2D, respectively, by using exponentially fitted spline elements; cf. [33]. We obtain an estimate of the same order in Theorem 4.3. Our numerical examples show that this estimate is optimal.

Finally, by a standard procedure (see, e.g., [9]), we have the following maximum-norm error estimate, whose proof will not be included here.

THEOREM 4.4. *Let \mathbf{u} and \mathbf{u}_h be solutions to (3.3) and (3.12), respectively. Assume that conditions of Theorem 3.3 are fulfilled. Then*

$$(4.9) \quad \|u - u_h\|_{W_\infty^0(\Omega)} \leq Ch\|u\|_{W_\infty^1(\Omega)}.$$

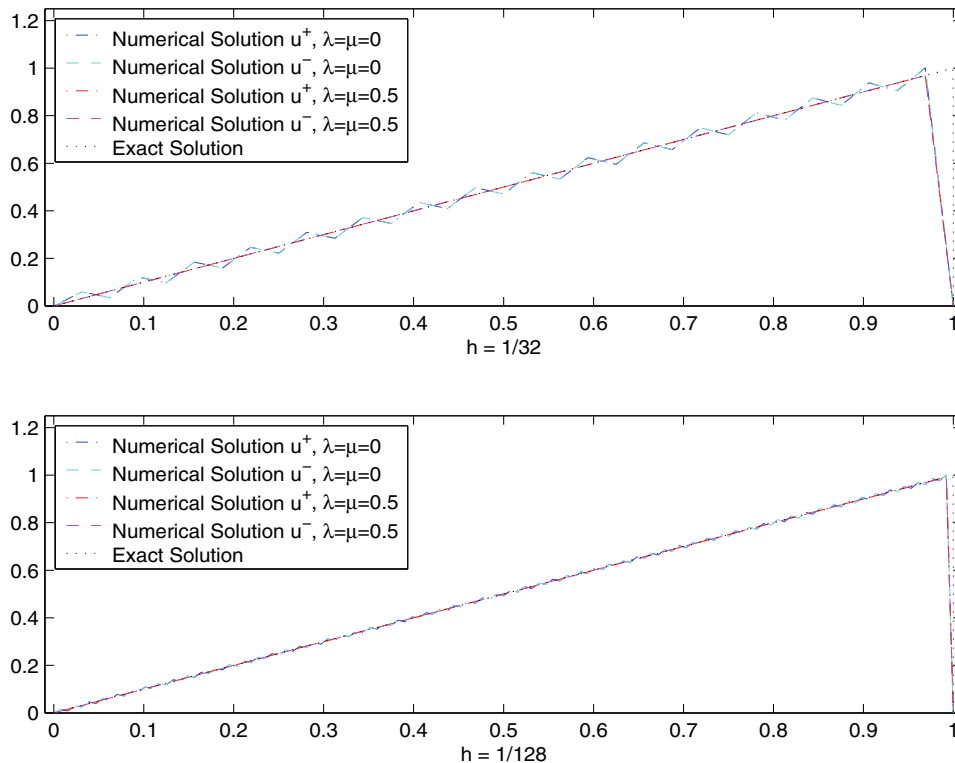


FIG. 1. *Example 5.1: Numerical solutions by the LDG-LS method with different numerical fluxes.*

5. Numerical experiments. In this section we present three numerical examples to illustrate the theoretical results of the methods developed in section 3. The stiffness matrices and load vectors are analytically calculated. High order Gaussian quadrature rules are used to calculate the norms of numerical errors over the computational regions (including the layers), which hereby causes no competitive extra errors in numerical integration. In the following examples, if not otherwise specified, we set $\epsilon^2 = 10^{-8}$. Uniform meshes are used for all examples.

Example 5.1. Consider the reaction-diffusion equation

$$(5.1) \quad \begin{cases} -\epsilon^2 u''(x) + u(x) = x & \text{in } (0, 1), \\ u(0) = u(1) = 0. \end{cases}$$

The analytical solution to (5.1) is

$$u(x) = x - \frac{e^{(x-1)/\epsilon} - e^{-(x+1)/\epsilon}}{1 - e^{-2/\epsilon}},$$

which has a typical exponential boundary layer at $x = 1$ when $\epsilon \ll 1$.

We first inspect the impact of the numerical fluxes. In Figure 1, we present the computational results of the LDG-LS method with numerical flux parameters (i) $\lambda = \mu = 0$ and (ii) $\lambda = \mu = 1/2$ (toward the boundary layer), and mesh size $h = 1/32$ and $1/128$, respectively. The numerical solutions have two traces (one-sided limits) at each interior mesh point, which are denoted in Figure 1 by u^+ and u^- , respectively.

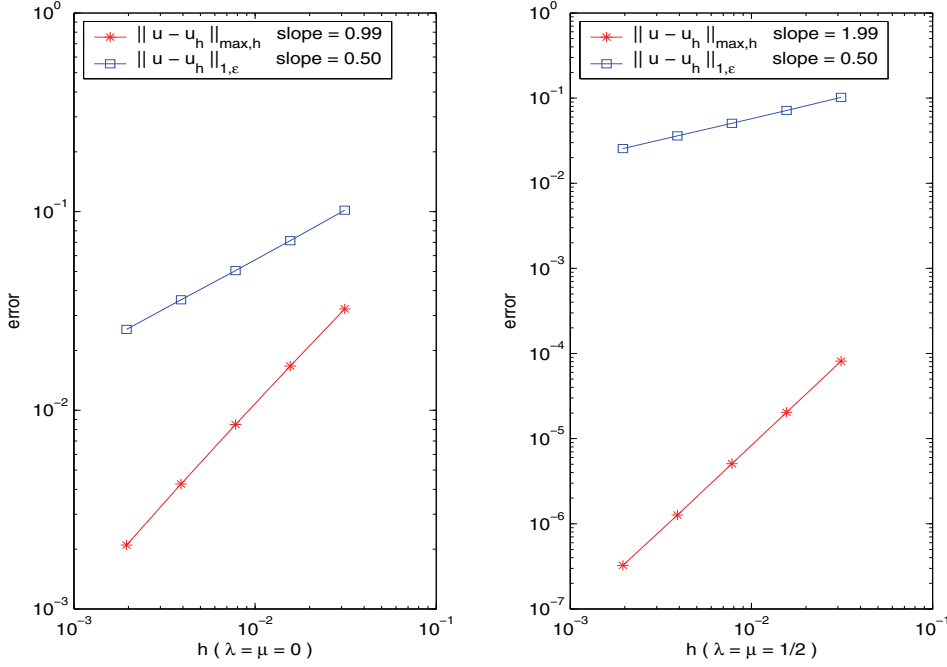


FIG. 2. Example 5.1: Convergence of the LDG-LS method with different numerical fluxes.

Figure 1 shows that the numerical results with (i) have oscillations. But those with (ii) have no oscillations even in quite a big mesh size. In addition, the jumps between u^+ and u^- are not visible. A more detailed investigation confirms that the jumps at the interior mesh points can be ignored when comparing them with other numerical errors.

Figure 2 shows log-log plots of numerical errors for the LDG-LS method with different numerical flux parameters, which are measured in the ϵ -dependent norm $\|\mathbf{u} - \mathbf{u}_h\|_{1,\epsilon}$ and the discrete maximum norm $\|u - u_h\|_{\max,h}$ at mesh points. Figure 2 confirms the estimates in Theorem 4.3. On the other hand, when selection (i) is used, the numerical results have first-order accuracy in the discrete maximum norm, as indicated in Theorem 4.4. When selection (ii) is used, the superconvergence phenomenon occurs. Superconvergence of singularly perturbed problems has been studied for continuous and discontinuous Galerkin methods; see, e.g., [23, 25, 32, 37, 38, 39]. In [13, 29], superconvergence results of LSFEM have been developed for second-order self-adjoint equations in 1D. Superconvergence analysis for least-squares methods in multidimensional nontensor product meshes is an ongoing research project.

Next, we compare the LDG-LS method with continuous LSFEMs, discontinuous LSFEMs, and DG methods. For comparison, we define a discontinuous least-squares functional in $\mathbf{H}(\mathcal{T}_h)$ as

$$\widehat{\mathcal{J}}(\mathbf{v}; f) = \sum_{k=1}^M \|A\mathbf{v} - \mathbf{f}\|_{0,\Omega_k}^2 + \sum_{k=1}^M \sum_{e \in \partial\Omega_K} \int_e ([v]_{k,e}^2 + \epsilon^2 [\mathbf{q}]_{k,e}^2) ds.$$

The corresponding FEM is referred to as *discontinuous LSFEM 1*. We also implemented the method proposed in [2], which is referred to as *discontinuous LSFEM*

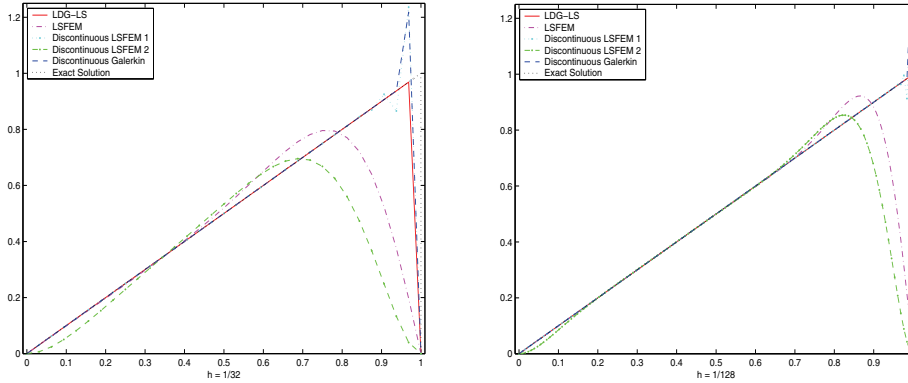


FIG. 3. Example 5.1: Numerical solutions by different numerical methods.

TABLE 1

Example 5.1: Numerical errors $\|\mathbf{u} - \mathbf{u}_h\|_{1,\epsilon}$ by the LDG-LS method for different ϵ values.

$\epsilon^2 \setminus h$	1/16	1/32	1/64	1/128	Order of convergence
10^0	2.106547e-2	7.758647e-3	2.799855e-3	1.000096e-3	1.476201
10^{-2}	1.459560e-1	6.232572e-2	2.470814e-2	9.363776e-3	1.362641
10^{-4}	3.112819e-1	2.045051e-1	1.206056e-1	6.187578e-2	0.867373
10^{-6}	4.314102e-1	2.819504e-1	1.732858e-1	1.120847e-1	0.647382
10^{-8}	4.368976e-1	3.089476e-1	2.183285e-1	1.536840e-1	0.504225
10^{-10}	4.368963e-1	3.089774e-1	2.184881e-1	1.544961e-1	0.499952
10^{-12}	4.368963e-1	3.089773e-1	2.184879e-1	1.544957e-1	0.499954

2. All numerical computations are conducted in uniform meshes with $h = 1/32$ and $h = 1/128$, respectively. In Figure 3, we present the numerical results of these methods. Here $\lambda = \mu = 1/2$ are used for the LDG-LS method. For methods with discontinuous elements, the average of the numerical approximations is plotted. It is observed that the solutions by the DG method and the discontinuous LSFEM 1 illustrate the typical “over-shooting” phenomenon near the boundary layer, where the magnitude of the numerical heap does not decrease as the mesh size decreases. On the other hand, the standard LSFEM and the discontinuous LSFEM 2 smear out the boundary layer, whose numerical errors in the discrete maximum norm do not converge to 0, though the methods converge with order 1/2 in the ϵ -dependent norm. The LDG-LS method provides the best numerical solutions here.

Finally, Table 1 is used to test for numerical independence of the LDG-LS method on ϵ . Here zero numerical fluxes are used. It is observed that when a singular perturbation occurs, the error $\|\mathbf{u} - \mathbf{u}_h\|_{1,\epsilon}$ converges with rate 1/2, which is independent of ϵ , as indicated in Theorem 4.3. On the other hand, when the problem is not singularly perturbed, superconvergence is observed (e.g., as $\epsilon^2 = 1$ or 10^{-2}); cf. [10]. In fact, when $\epsilon = 1$, norms $\|\cdot\|_{1,\epsilon}$ and $\|\cdot\|_1$ are the same.

Example 5.2. Consider the reaction-diffusion equation

$$(5.2) \quad \begin{cases} -\epsilon^2 u''(x) + (2-x)u(x) = f & \text{in } (0,1), \\ u(0) = u(1) = 0, \end{cases}$$

where f is chosen properly such that the solution u to (5.2) is

$$u(x) = \left(1 - e^{-x/\epsilon}\right) \left(1 - e^{(x-1)/\epsilon}\right).$$

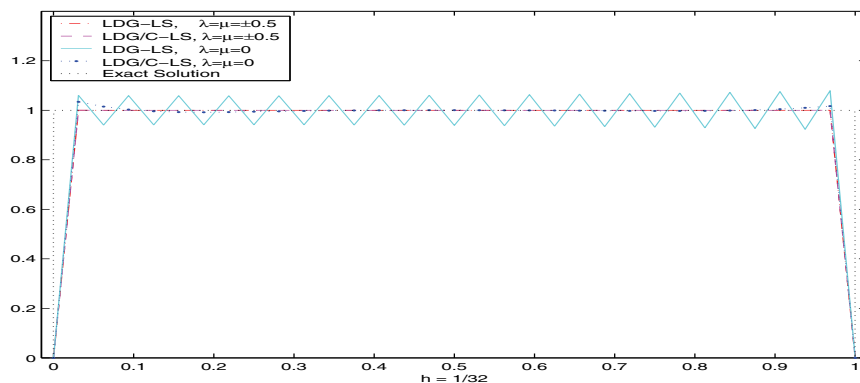


FIG. 4. Example 5.2: Numerical solutions by LDG-LS and LDG/C-LS methods ($h = 1/32$).

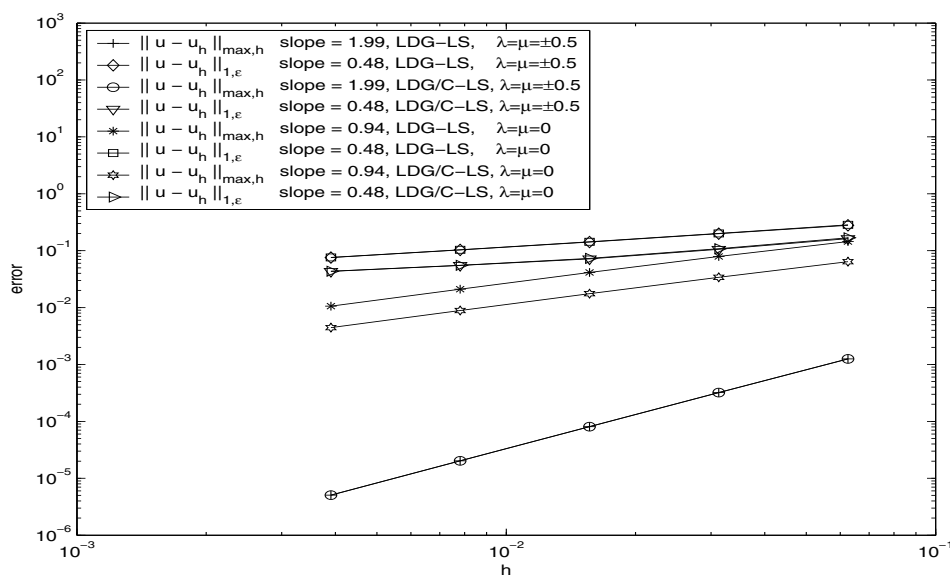


FIG. 5. Example 5.2: Convergence of LDG-LS and LDG/C-LS methods.

The exact solution has two boundary layers at $x = 0$ and 1 .

We compare the numerical solutions of the LDG-LS method and the LDG/C-LS method in this example. The numerical fluxes are first chosen toward the layers. In particular, for the LDG-LS scheme, we choose $\lambda = \mu = -1/2$ for elements located in $[0, 1/2]$ and $\lambda = \mu = 1/2$ for elements in $[1/2, 1]$. For the LDG/C-LS scheme, the numerical flux parameters are analogously chosen only for the first and last two elements, respectively, which contain the boundary layers. Recall that, in the case of this example, the LDG/C-LS method uses discontinuous basis functions only in two elements at each end, and uses continuous basis functions for the other elements. Analogous LDG-LS and LDG/C-LS schemes are developed with $\lambda = \mu = 0$ for comparison.

Figure 4 shows the numerical solutions by the two methods with different numerical fluxes in a uniform mesh ($h = 1/32$). When $\lambda = \mu = 0$, the LDG-LS solutions have oscillations in the entire region, while the LDG/C-LS solutions have only small

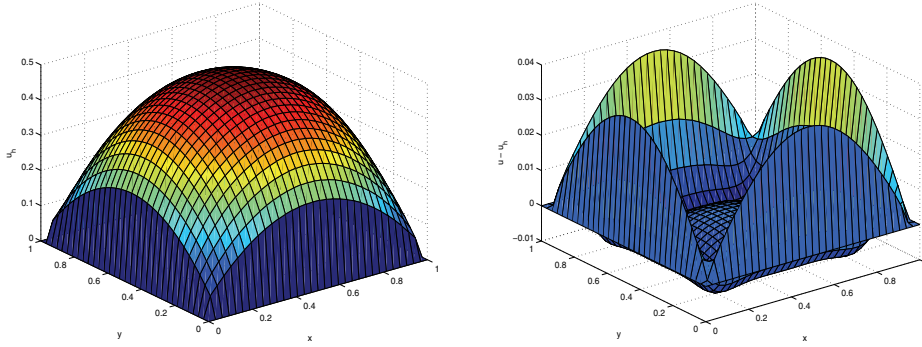


FIG. 6. Example 5.3: Numerical solutions and errors by the LDG/C-LS method ($h = 1/32$).

bumps near the boundary layers. On the other hand, solutions from both methods with $\lambda = \mu = \pm 1/2$ match the exact solutions very well. Figure 5 indicates that the convergence rates of the LDG-LS and LDG/C-LS methods are the same when identical numerical fluxes are applied, since the dominant error occurs near the boundary layers. Moreover, superconvergence is observed in the discrete maximum norm when the numerical flux parameters are $\pm 1/2$.

Note that the LDG/C-LS method produces an algebraic problem of about half the degrees of freedom as the LDG-LS method, which thus significantly reduces computational cost. This makes the method competitive with the standard DG or least-squares methods. Moreover, our numerical tests show that the resulting discrete problems by both methods have the same order condition numbers; in particular, $\mathcal{O}(h^{-2})$ condition numbers for the problem with nonsingular perturbation, and $\mathcal{O}(h^{-3})$ condition numbers for singularly perturbed cases.

Example 5.3. Consider the reaction-diffusion equation

$$(5.3) \quad \begin{cases} -\epsilon^2 \Delta u + 2u = f & \text{in } \Omega = (0, 1) \times (0, 1), \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

where f is chosen properly such that the solution u to (5.3) is

$$u(x, y) = \left(1 - e^{-x/\epsilon}\right) \left(1 - e^{-(x-1)/\epsilon}\right) y(1-y) + x(1-x) \left(1 - e^{-y/\epsilon}\right) \left(1 - e^{-(y-1)/\epsilon}\right).$$

The exact solution has exponential layers on the boundary $\partial\Omega$.

We use bilinear LDG/C-LS elements on uniform meshes in this example. Discontinuous basis functions are used only in two layers of elements along the boundary, and continuous basis functions are used for the other elements. Here, we use $\lambda = \mu = \mathbf{0}$ for all discontinuous elements.

Figure 6 shows the numerical solutions u_h and the errors $u - u_h$ by the LDG/C-LS method when $h = 1/32$. Figure 7 is the log-log plot of the numerical errors of the LDG/C-LS method. The convergence rates are calculated based on the last three data points. The numerical results agree with the theoretical predictions in Theorems 4.3 and 4.4.

6. Conclusion. A singularly perturbed reaction-diffusion problem with homogeneous Dirichlet boundary conditions is considered in this paper, for which we developed a stable numerical approach based on LDG discretization of least-squares

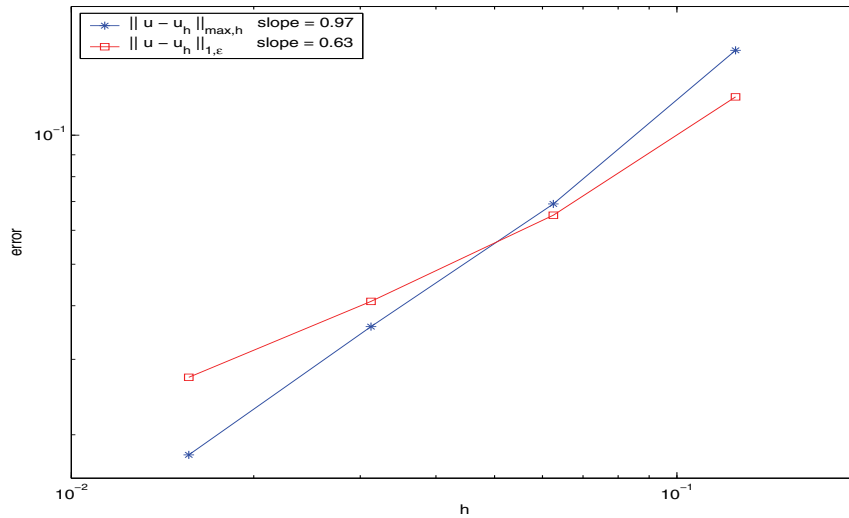


FIG. 7. Example 5.3: Convergence of the LDG/C-LS method.

formulation, which needs no special treatments. The coercivity and boundedness of the bilinear forms have been proven, which leads to the well posedness of the methods. We prove a uniform optimal energy norm error estimate. A maximum-norm error estimate is also provided. A hybrid adaptive method is derived for efficiency reasons. Numerical examples are presented, which are in agreement with the theoretical results. Comparisons have been conducted numerically. In addition, some superconvergence results have been observed.

The proposed approach is innovative, requiring neither special treatments nor manually adjusted parameters. Comparing with the other methods, such as the LDG method and the continuous and discontinuous LSFEMs, our method is more robust and accurate in solving problems with singular perturbation. The hybrid method is competitive with the standard LDG method and LSFEM in the sense of computational cost. This paper provides an efficient alternative to numerical approaches for solving singularly perturbed reaction-diffusion problems.

It is our belief that an analogous approach can be developed for general singularly perturbed convection-diffusion problems, which is an ongoing research project.

Acknowledgments. The author thanks the anonymous referees and the editor, Dr. Pavel Bochev, for their constructive comments and suggestions, which greatly improved this article.

REFERENCES

- [1] D. N. ARNOLD, F. BREZZI, B. COCKBURN, AND L. D. MARINI, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.
- [2] R. E. BENSOW AND M. G. LARSON, *Discontinuous/continuous least-squares finite element methods for elliptic problems*, Math. Models Methods Appl. Sci., 15 (2005), pp. 825–842.
- [3] R. E. BENSOW AND M. G. LARSON, *Discontinuous least-squares finite element method for the div-curl problem*, Numer. Math., 101 (2005), pp. 601–617.
- [4] P. B. BOCHEV, *Least-squares finite element methods for first-order elliptic systems*, Int. J. Numer. Anal. Model., 1 (2004), pp. 49–64.

- [5] P. B. BOCHEV AND M. D. GUNZBURGER, *Finite element methods of least-squares type*, SIAM Rev., 40 (1998), pp. 789–837.
- [6] P. BOCHEV AND M. GUNZBURGER, *On least-squares finite element methods for the Poisson equation and their connection to the Dirichlet and Kelvin principles*, SIAM J. Numer. Anal., 43 (2005), pp. 340–362.
- [7] J. H. BRAMBLE, R. D. LAZAROV, AND J. E. PASCIAK, *A least-squares approach based on a discrete minus one inner product for first order systems*, Math. Comp., 66 (1997), pp. 935–955.
- [8] J. H. BRAMBLE, R. D. LAZAROV, AND J. E. PASCIAK, *Least-squares for second-order elliptic problems*, Comput. Methods Appl. Mech. Engrg., 152 (1998), pp. 195–210.
- [9] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.
- [10] Z. CAI, R. D. LAZAROV, T. A. MANTEUFFEL, AND S. F. MCCORMICK, *First-order system least squares for second-order partial differential equations. I*, SIAM J. Numer. Anal., 31 (1994), pp. 1785–1799.
- [11] Z. CAI, T. A. MANTEUFFEL, AND S. F. MCCORMICK, *First-order system least squares for second-order partial differential equations. II*, SIAM J. Numer. Anal., 34 (1997), pp. 425–454.
- [12] Y. CAO AND M. D. GUNZBURGER, *Least-squares finite element approximations to solutions of interface problems*, SIAM J. Numer. Anal., 35 (1998), pp. 393–405.
- [13] G. F. CAREY AND Y. SHEN, *Convergence studies of least-squares finite elements for first-order systems*, Comm. Appl. Numer. Meth., 5 (1989), pp. 427–434.
- [14] G. F. CAREY AND Y. SHEN, *Least-squares finite element approximation of Fisher’s reaction-diffusion equation*, Numer. Meth. Partial Differential Equations, 11 (1995), pp. 175–186.
- [15] B. COCKBURN AND C.-W. SHU, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.
- [16] H.-Y. DUAN AND D.-L. ZHANG, *A finite element method for singularly perturbed reaction-diffusion problems*, Acta Math. Appl. Sin. Engl. Ser., 19 (2003), pp. 25–30.
- [17] M. I. GERRITSMAN AND M. M. J. PROOT, *Analysis of a discontinuous least squares spectral element method*, J. Sci. Comput., 17 (2002), pp. 297–306.
- [18] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 224, Springer-Verlag, Berlin, 1983.
- [19] P. HOUSTON, M. JENSEN, AND E. SÜLI, *hp-discontinuous Galerkin finite element methods with least-squares stabilization*, J. Sci. Comput., 17 (2002), pp. 3–25.
- [20] B.-N. JIANG, *The Least-Squares Finite Element Method. Theory and Applications in Computational Fluid Dynamics and Electromagnetics*, Springer-Verlag, Berlin, 1998.
- [21] B. N. JIANG AND L. A. POVINELLI, *Optimal least-squares finite element method for elliptic problems*, Comput. Methods Appl. Mech. Engrg., 102 (1993), pp. 199–212.
- [22] R. D. LAZAROV, L. TOBISKA, AND P. S. VASSILEVSKI, *Streamline diffusion least-squares mixed finite element methods for convection-diffusion problems*, East-West J. Numer. Math., 5 (1997), pp. 249–264.
- [23] J. LI, *Convergence and superconvergence analysis of finite element methods on highly nonuniform anisotropic meshes for singularly perturbed reaction-diffusion problems*, Appl. Numer. Math., 36 (2001), pp. 129–154.
- [24] R. LIN, *A discontinuous least-squares finite element method for singularly perturbed reaction-diffusion problems*, Dyn. Contin. Discrete Impuls. Syst. Ser. A Math. Anal., 14 (2007), Advances in Dynamical Systems, suppl. S2, pp. 243–246.
- [25] T. LINSS, *Uniform superconvergence of a Galerkin finite element method on Shishkin-type meshes*, Numer. Methods Partial Differential Equations, 16 (2000), pp. 426–440.
- [26] J. J. H. MILLER, E. O’RIORDAN, AND G. I. SHISHKIN, *Fitted Numerical Methods for Singular Perturbation Problems. Error Estimates in the Maximum Norm for Linear Problems in One and Two Dimensions*, World Scientific, River Edge, NJ, 1996.
- [27] K. W. MORTON, *Numerical Solution of Convection-Diffusion Problems*, Applied Mathematics and Mathematical Computation 12, Chapman & Hall, London, 1996.
- [28] E. O’RIORDAN AND M. STYNES, *A globally uniformly convergent finite element method for a singularly perturbed elliptic problem in two dimensions*, Math. Comp., 57 (1991), pp. 47–62.
- [29] A. I. PEHLIVANOV, G. F. CAREY, R. D. LAZAROV, AND Y. SHEN, *Convergence analysis of least-squares mixed finite elements*, Computing, 51 (1993), pp. 111–123.
- [30] A. I. PEHLIVANOV, G. F. CAREY, AND R. D. LAZAROV, *Least-squares mixed finite elements for second-order elliptic problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1368–1377.
- [31] A. I. PEHLIVANOV, G. F. CAREY, AND P. S. VASSILEVSKI, *Least-squares mixed finite element*

- methods for non-selfadjoint elliptic problems. I. Error estimates*, Numer. Math., 72 (1996), pp. 501–522.
- [32] H.-G. ROOS, *Superconvergence on a hybrid mesh for singularly perturbed problems with exponential layers*, ZAMM Z. Angew. Math. Mech., 86 (2006), pp. 649–655.
 - [33] H.-G. ROOS, M. STYNES, AND L. TOBISKA, *Numerical Methods for Singularly Perturbed Differential Equations. Convection-Diffusion and Flow Problems*, Springer-Verlag, Berlin, 1996.
 - [34] A. H. SCHATZ AND L. B. WAHLBIN, *On the finite element method for singularly perturbed reaction-diffusion problems in two and one dimensions*, Math. Comp., 40 (1983), pp. 47–89.
 - [35] D. R. SMITH, *Singular-Perturbation Theory. An Introduction with Applications*, Cambridge University Press, Cambridge, UK, 1985.
 - [36] M. STYNES AND E. O’RIORDAN, *An analysis of a singularly perturbed two-point boundary value problem using only finite element techniques*, Math. Comp., 56 (1991), pp. 663–675.
 - [37] Z. XIE AND Z. ZHANG, *Superconvergence of DG method for one-dimensional singularly perturbed problems*, J. Comput. Math., 25 (2007), pp. 185–200.
 - [38] Z. ZHANG, *Finite element superconvergence approximation for one-dimensional singularly perturbed problems*, Numer. Methods Partial Differential Equations, 18 (2002), pp. 374–395.
 - [39] Z. ZHANG, *Finite element superconvergence on Shishkin mesh for 2-D convection-diffusion problems*, Math. Comp., 72 (2003), pp. 1147–1177.

A POSTERIORI ERROR ESTIMATE AND ADAPTIVE MESH REFINEMENT FOR THE CELL-CENTERED FINITE VOLUME METHOD FOR ELLIPTIC BOUNDARY VALUE PROBLEMS*

CHRISTOPH ERATH[†] AND DIRK PRAETORIUS[‡]

Abstract. We extend a result of Nicaise [*SIAM J. Numer. Anal.*, 43 (2005), pp. 1481–1503] for the a posteriori error estimation of the cell-centered finite volume method for the numerical solution of elliptic problems. Having computed the piecewise constant finite volume solution u_h , we compute a Morley-type interpolant $\mathcal{I}u_h$. For the exact solution u , the energy error $\|\nabla_{\mathcal{T}}(u - \mathcal{I}u_h)\|_{L^2}$ can be controlled efficiently and reliably by a residual-based a posteriori error estimator η . The local contributions of η are used to steer an adaptive mesh-refining algorithm. A model example serves the Laplace equation in two dimensions with mixed Dirichlet–Neumann boundary conditions.

Key words. finite volume method, cell-centered method, diamond path, a posteriori error estimate, adaptive algorithm

AMS subject classifications. 65N30, 65N15

DOI. 10.1137/070702126

1. Introduction. Throughout, $\Omega \subset \mathbb{R}^2$ is a bounded and connected domain with Lipschitz boundary $\Gamma := \partial\Omega$. We assume that Γ is divided into a closed Dirichlet boundary $\Gamma_D \subseteq \Gamma$ with positive surface measure and a Neumann boundary $\Gamma_N := \Gamma \setminus \Gamma_D$. We consider the elliptic model problem

$$(1.1) \quad -\Delta u = f \quad \text{in } \Omega$$

with mixed boundary conditions

$$(1.2) \quad u = u_D \quad \text{on } \Gamma_D \quad \text{and} \quad \partial u / \partial \mathbf{n} = g \quad \text{on } \Gamma_N.$$

Here $f \in L^2(\Omega)$, $u_D \in H^1(\Gamma_D)$, and $g \in L^2(\Gamma_N)$ are given data, and $L^2(\cdot)$ and $H^1(\cdot)$ denote the standard Lebesgue and Sobolev spaces equipped with the usual norms $\|\cdot\|_{L^2(\cdot)}$ and $\|\cdot\|_{H^1(\cdot)}$. The weak form of (1.1) reads as follows: Find $u \in H^1(\Omega)$ with $u|_{\Gamma_D} = u_D$ and

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\Omega} f v \, dx + \int_{\Gamma_N} g v \, dx \quad \text{for all } v \in H_D^1(\Omega) := \{v \in H^1(\Omega) \mid v|_{\Gamma_D} = 0\}.$$

Recall that there is a unique solution u which we aim to approximate by a postprocessed finite volume scheme. For technical reasons, we assume that Γ_D as well as Γ_N are connected; see Theorem 5.1 below.

Let \mathcal{T} be a triangulation of Ω and \mathcal{E} the set of all edges of \mathcal{T} . Replacing the continuous diffusion flux $\int_E \partial u / \partial \mathbf{n}_E \, ds$ by a discrete diffusion flux $F_E^D(u_h)$, the cell-centered finite volume method provides a \mathcal{T} -elementwise constant approximation $u_h \in$

*Received by the editors September 6, 2007; accepted for publication (in revised form) June 9, 2008; published electronically October 24, 2008.

<http://www.siam.org/journals/sinum/47-1/70212.html>

[†]Corresponding author. University of Ulm, Institute for Numerical Mathematics, Helmholtzstraße 18, D-89069 Ulm, Germany (christoph.erath@uni-ulm.de).

[‡]Vienna University of Technology, Institute for Analysis and Scientific Computing, Wiedner Hauptstraße 8-10, A-1040 Vienna, Austria (dirk.praetorius@tuwien.ac.at).

$\mathcal{P}_0(\mathcal{T})$ of u . The classical choice of $F_E^D(u_h)$ is based on the admissibility of the triangulation \mathcal{T} in the sense of [10]. However, locally refined meshes are usually not admissible. Another choice of $F_E^D(u_h)$ is the diamond path method, which has been mathematically analyzed in [6, 7] for rectangular meshes with a maximum of one hanging node per edge. Optimal order of convergence $\|u - u_h\|_{1,h} = \mathcal{O}(h)$ of the error with respect to a discrete H^1 -norm $\|\cdot\|_{1,h}$ holds under the regularity assumption $u \in H^2(\Omega)$, which is usually not met in practice.

We aim to provide a mathematical criterion for steering an adaptive mesh-refining algorithm to recover the optimal order of convergence $\mathcal{O}(N^{-1/2})$ with respect to the number $N = \#\mathcal{T}$ of elements. Following Nicaise [11], we introduce a Morley-type interpolant $\mathcal{I}u_h$ which belongs to a certain $H^1(\Omega)$ -nonconforming finite element space, the definition of which is a generalization of the definition in [11, section 5] to the case of hanging nodes and mixed boundary conditions. Roughly speaking, the analytical idea is to ensure that $\mathcal{I}u_h$ has enough orthogonality properties which can be used to adapt the well-known a posteriori error analysis from the context of the finite element method; see, e.g., [12, 1]. For each element $T \in \mathcal{T}$ with corresponding edges \mathcal{E}_T , we define the refinement indicators

$$\begin{aligned} \eta_T^2 := & h_T^2 \|f - f_T\|_{L^2(T)}^2 + \sum_{E \in \{E \in \mathcal{E}_E \mid E \subset \partial T\}} h_E \|\llbracket \nabla_{\mathcal{T}}(\mathcal{I}u_h) \rrbracket\|_{L^2(E)}^2 \\ & + \sum_{E \in \mathcal{E}_T \cap \mathcal{E}_N} h_E \left\| \frac{\partial(u - \mathcal{I}u_h)}{\partial \mathbf{n}_E} \right\|_{L^2(E)}^2 + \sum_{E \in \mathcal{E}_T \cap \mathcal{E}_D} h_E \left\| \frac{\partial(u - \mathcal{I}u_h)}{\partial \mathbf{t}_E} \right\|_{L^2(E)}^2. \end{aligned}$$

Here $\llbracket \cdot \rrbracket$ denotes the jump, \mathbf{n}_E and \mathbf{t}_E denote the normal and tangential vector on E , respectively, f_T denotes the piecewise integral mean of the volume term, and h_E is the length of the edge E . We prove that the corresponding error estimator

$$\eta := \left(\sum_{T \in \mathcal{T}} \eta_T^2 \right)^{1/2}$$

is reliable and efficient in the sense that

$$C_{\text{rel}}^{-1} \|\nabla_{\mathcal{T}}(u - \mathcal{I}u_h)\|_{L^2(\Omega)} \leq \eta \leq C_{\text{eff}} \left[\|\nabla_{\mathcal{T}}(u - \mathcal{I}u_h)\|_{L^2(\Omega)} + \|h(f - f_T)\|_{L^2(\Omega)} \right].$$

Here $\nabla_{\mathcal{T}}$ denotes the \mathcal{T} -piecewise gradient, and the constants $C_{\text{eff}}, C_{\text{rel}} > 0$ depend only on the shape of the elements in \mathcal{T} but not on f , the local mesh width h , or the number of elements. Moreover, the efficiency estimate holds even locally:

$$\eta_T \leq C_{\text{eff}} \left[\|\nabla_{\mathcal{T}}(u - \mathcal{I}u_h)\|_{L^2(\omega_T)} + \|h(f - f_T)\|_{L^2(\omega_T)} \right],$$

where ω_T denotes the patch of the element $T \in \mathcal{T}$.

The proof of the reliability makes use of the Helmholtz decomposition to deal with mixed boundary conditions. For the proof of the efficiency estimate, the non-avoidance of hanging nodes needs the extended definition of edge patches. We stress that [11] treats only the Dirichlet problem $\Gamma_D = \Gamma$ and that the a posteriori error analysis is restricted to the case of regular meshes. Therefore the definition of $\mathcal{I}u_h$ had to be substantially modified.

The content of this paper is organized as follows. In section 2, we introduce the notation that is used below. In particular, we define the concept of an *almost regular triangulation*, which allows the analytical error analysis in the case of certain hanging

nodes. Section 3 gives a short summary on the classical cell-centered finite volume method for our model problem. We recall the ideas of the diamond path, where emphasis is laid on the treatment of nodes $a \in \Gamma_N$ that lie on the Neumann boundary Γ_N . In section 4, we define the Morley interpolant and collect the orthogonality properties used for the error analysis. Reliability and efficiency of the error estimator η are then proven in section 5. Numerical experiments, found in section 6, confirm the theoretical results and conclude the work. In particular, we observe that the proposed strategy even recovers the optimal order of convergence with respect to the energy norm $\|u - u_h\|_{1,h}$.

2. Preliminaries and notation. In this section, we introduce the notation for the triangulations that are considered below. In particular, we define the so-called *almost regular* triangulation which allows certain hanging nodes.

2.1. Almost regular triangulation. Throughout, \mathcal{T} denotes a triangulation of Ω , where \mathcal{N} and \mathcal{E} are the corresponding set of nodes and edges, respectively. We assume that the elements $T \in \mathcal{T}$ are triangles or rectangles, either of which are nondegenerate. For $T \in \mathcal{T}$, $h_T := \text{diam}(T)$ denotes the Euclidean diameter and ϱ_T is the corresponding height; i.e., the volume of T is $|T| = h_T \varrho_T$ in the case of T being a rectangle and $|T| = h_T \varrho_T / 2$ in the case of T being a triangle. Moreover, for an edge $E \in \mathcal{E}$, we denote by h_E its length.

Nodes. In the following, we introduce a partition

$$\mathcal{N} = \mathcal{N}_D \cup \mathcal{N}_N \cup \mathcal{N}_H \cup \mathcal{N}_F$$

of \mathcal{N} into Dirichlet and Neumann nodes, hanging nodes, and free nodes, respectively: First, let $\mathcal{N}_D := \{a \in \mathcal{N} \mid a \in \Gamma_D\}$ (resp., $\mathcal{N}_N := \{a \in \mathcal{N} \mid a \in \Gamma_N\}$) be the set of all nodes that belong to the Dirichlet boundary (resp., Neumann boundary). A node $a \in \mathcal{N} \setminus (\mathcal{N}_D \cup \mathcal{N}_N)$ is a hanging node, provided that there are elements $T_1, T_2 \in \mathcal{T}$ such that $a \in T_1 \cap T_2$ is a node of T_1 but not of T_2 . Let \mathcal{N}_H be the set of all hanging nodes. Finally, the set of free nodes is $\mathcal{N}_F := \mathcal{N} \setminus (\mathcal{N}_D \cup \mathcal{N}_N \cup \mathcal{N}_H)$. For an element $T \in \mathcal{T}$, we denote with \mathcal{N}_T the set of nodes of T , i.e., $|\mathcal{N}_T| = 3$ for T being a triangle and $|\mathcal{N}_T| = 4$ for T being a rectangle, respectively.

Edges. For the edges, we introduce a partition

$$\mathcal{E} = \mathcal{E}_D \cup \mathcal{E}_N \cup \mathcal{E}_H \cup \mathcal{E}_E$$

into Dirichlet and Neumann edges, nonelementary edges, and interior elementary edges, respectively: First, we define $\mathcal{E}_D := \{E \in \mathcal{E} \mid E \subseteq \Gamma_D\}$ and $\mathcal{E}_N := \{E \in \mathcal{E} \mid E \subseteq \overline{\Gamma}_N\}$. Second, an interior edge $E \in \mathcal{E}$ is nonelementary if there are pairwise different nodes $x, y, z \in \mathcal{N}$ such that $E = \text{conv}\{x, y\}$ and $z \in E$; i.e., there is a hanging node z in the interior of E . The set of all nonelementary edges is denoted by \mathcal{E}_H . Contrarily, $\mathcal{E} \setminus \mathcal{E}_H$ denotes the set of all elementary edges, which is split into boundary edges $\mathcal{E}_D \cup \mathcal{E}_N$ and interior elementary edges $\mathcal{E}_E := \mathcal{E} \setminus (\mathcal{E}_H \cup \mathcal{E}_D \cup \mathcal{E}_N)$. Moreover, we define the set

$$\mathcal{E}_0 := \{E \in \mathcal{E}_E \mid \nexists E' \in \mathcal{E}_H \quad E \subsetneq E'\}$$

of all interior elementary edges which are not part of a nonelementary edge. Finally, for an element $T \in \mathcal{T}$, we denote with $\mathcal{E}_T \subset \mathcal{E}$ the set of all edges of T , i.e.,

$$\mathcal{E}_T := \{E \in \mathcal{E} \mid E \subseteq \partial T \quad \text{for all } E' \in \mathcal{E} \quad (E \subsetneq E' \Rightarrow E' \not\subseteq \partial T)\}.$$

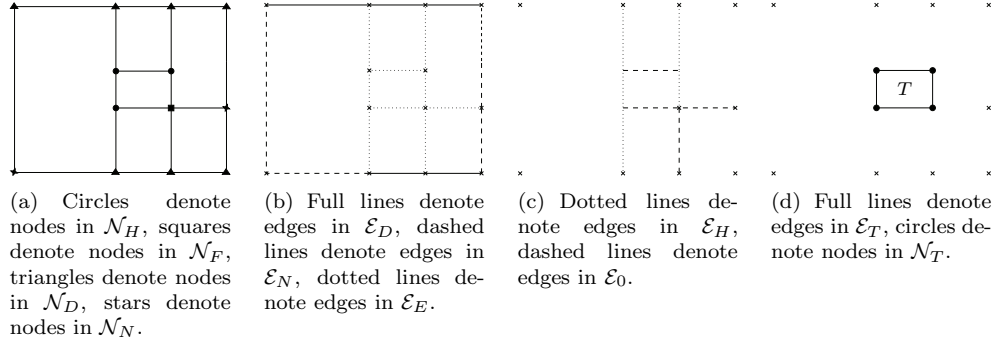


FIG. 2.1. The sets of edges and nodes for a simple (almost regular) triangulation, which consists of six rectangular elements.

Almost regular triangulations. We say that the triangulation \mathcal{T} is almost regular if the following hold:

- (i) the mixed boundary conditions are resolved; i.e., each edge $E \in \mathcal{E}$ with $E \cap \Gamma \neq \emptyset$ satisfies either $E \in \mathcal{E}_D$ or $E \in \mathcal{E}_N$;
- (ii) the intersection $T_1 \cap T_2$ of two elements $T_1, T_2 \in \mathcal{T}$ with $T_1 \neq T_2$ is either empty or a node or an edge;
- (iii) each nonelementary edge $E \in \mathcal{E}_H$ is the finite union of elementary edges; i.e., there are finitely many elementary edges $E_1, \dots, E_n \in \mathcal{E}_E$ such that $E = \bigcup_{i=1}^n E_i$.

With respect to regular triangulations in the sense of Ciarlet, the only difference is that in (ii) the intersection $T_1 \cap T_2$ may be, for instance, a node (or an edge) of T_1 but not of T_2 ; see Figure 2.1. However, in the case of $E := T_1 \cap T_2$ being an edge, (iii) implies that there holds at least either $E \in \mathcal{E}_{T_1}$ or $E \in \mathcal{E}_{T_2}$. From now on, we assume that all triangulations are at least almost regular (or even regular).

2.2. Normal and tangential vectors. For each edge $E \in \mathcal{E}$, we fix a normal vector \mathbf{n}_E as follows: For $E \in \mathcal{E}_D \cup \mathcal{E}_N$, let \mathbf{n}_E point outwards of Ω . For an edge $E \in \mathcal{E}_H$, there is a unique element $T \in \mathcal{T}$ with $E \in \mathcal{E}_T$, and we choose \mathbf{n}_E to point into T . For each elementary edge $E' \in \mathcal{E}_E$ with $E' \subset E$, we define $\mathbf{n}_{E'} := \mathbf{n}_E$. For the remaining edges, namely $E \in \mathcal{E}_0$, we may choose the orientation of \mathbf{n}_E arbitrarily.

In section 3, we shall use the following notational convention: For each elementary edge $E \in \mathcal{E}_E$, there are unique elements $T_{W,E}$ and $T_{E,E}$ such that $E \subseteq T_{W,E} \cap T_{E,E}$ and such that \mathbf{n}_E points from $T_{W,E}$ to $T_{E,E}$ (i.e., from west to east). For $E \in \mathcal{E}_D \cup \mathcal{E}_N$, there is a unique element $T_{W,E}$ with $E \subset \partial T_{W,E}$. If the edge E is clear from the context, we omit the additional subscript and simply write, e.g., $T_W = T_{W,E}$.

Moreover, for each element $T \in \mathcal{T}$ and an edge $E \in \mathcal{E}$ with $E \subset \partial T$, we define the sign

$$\sigma_{T,E} = \begin{cases} +1, & \text{provided } T = T_{W,E}, \\ -1 & \text{else;} \end{cases}$$

i.e., $\sigma_{T,E} \mathbf{n}_E$ is the outer normal vector $\mathbf{n}_T|_E$ of T restricted to the edge E .

Finally, the tangential vector \mathbf{t}_E of an edge $E \in \mathcal{E}$ is chosen orthogonal to \mathbf{n}_E in the mathematical positive sense. We note that $\sigma_{T,E} \mathbf{t}_E$ is the tangential vector $\mathbf{t}_T|_E$ of an element $T \in \mathcal{T}$ restricted to the edge E ; see Figure 2.2.

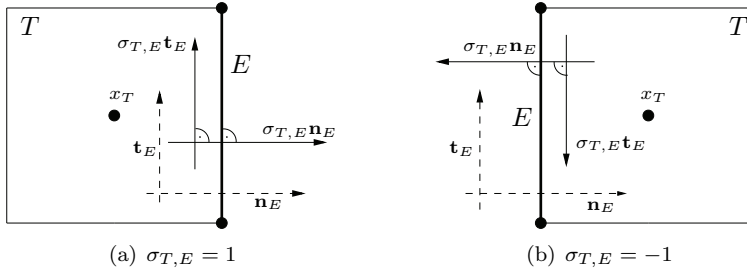


FIG. 2.2. The dashed lines show the a priori chosen normal vector \mathbf{n}_E (resp., tangential vector \mathbf{t}_E) on the edge E , whereas the full lines are the outer normal vector $\mathbf{n}_T|_E = \sigma_{T,E}\mathbf{n}_E$ of T (resp., $\mathbf{t}_T|_E = \sigma_{T,E}\mathbf{t}_E$) with respect to the edge E .

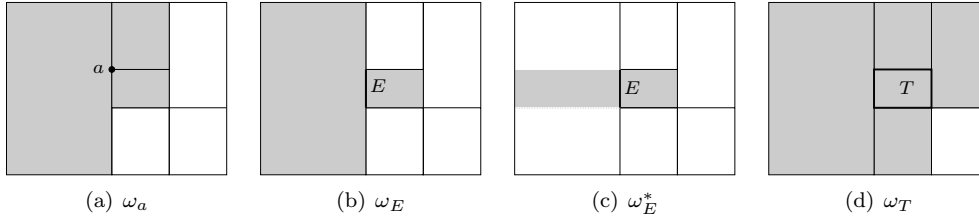


FIG. 2.3. The four patches introduced in section 2.3.

2.3. Patches. We recall the definition of the patches which are well known from finite element analysis. Additionally, we introduce the elementary patch of an edge which is needed for the handling of the hanging nodes in our a posteriori error analysis; see Figure 2.3.

Patch of a node. For $a \in \mathcal{N}$, the patch is given by

$$\omega_a = \bigcup_{T \in \tilde{\omega}_a} T, \quad \text{where } \tilde{\omega}_a := \{T \in \mathcal{T} \mid a \subseteq \partial T\}.$$

Patch of an edge. For an elementary edge $E \in \mathcal{E} \setminus \mathcal{E}_H$, the patch is given by

$$\omega_E := \bigcup_{T \in \tilde{\omega}_E} T, \quad \text{where } \tilde{\omega}_E := \{T \in \mathcal{T} \mid E \subseteq \partial T\}.$$

For a nonelementary edge $E \in \mathcal{E}_H$ and $E_1, \dots, E_n \in \mathcal{E}_E$ with $E = \bigcup_{i=1}^n E_i$, we define

$$\omega_E := \bigcup_{T \in \tilde{\omega}_E} T = \bigcup_{i=1}^n \omega_{E_i}, \quad \text{where } \tilde{\omega}_E := \bigcup_{i=1}^n \tilde{\omega}_{E_i}.$$

Elementary patch of an edge. Let us consider a nonelementary edge $E \in \mathcal{E}_H$ and $E_1, \dots, E_n \in \mathcal{E}_E$ with $E = \bigcup_{i=1}^n E_i$. Then there is a unique element $T_E \in \mathcal{T}$ with $E \in \mathcal{E}_{T_E}$. Moreover, there are unique elements $T_i \in \mathcal{T}$ such that $E_i \in \mathcal{E}_{T_i}$. We denote by $a_1, \dots, a_{n-1} \in \mathcal{N}_H$ the hanging nodes, which are on E . Moreover, let a_0 and a_n be the nodes of E . Without loss of generality, a_{i-1}, a_i are the nodes of E_i , i.e.,

$$E_i = \text{conv}\{a_{i-1}, a_i\}.$$

If $T_E = \text{conv}\{a_0, a_n, b\}$ is a triangle, we define triangles $\tilde{T}_i := \text{conv}\{a_{i-1}, a_i, b\}$ and note that

$$(2.1) \quad T_E = \bigcup_{i=1}^n \tilde{T}_i \quad \text{and} \quad \text{int}(\tilde{T}_i) \cap \text{int}(\tilde{T}_j) = \emptyset \quad \text{for } i \neq j,$$

where $\text{int}(\cdot)$ denotes the topological interior of a set. We stress that the triangles \tilde{T}_i cannot be elements of the triangulation \mathcal{T} . For T_E a rectangle, we can construct rectangles \tilde{T}_i with (2.1). For each of the elementary edges E_i , we may then define the elementary patch

$$\omega_{E_i}^* := T_i \cup \tilde{T}_i \quad \text{and} \quad \tilde{\omega}_{E_i}^* := \{T_i, \tilde{T}_i\}.$$

So far, we have defined the patch ω_E^* for all edges $E \in \mathcal{E}$ which are contained in a nonelementary edge. For the remaining edges $E \in \mathcal{E}$, we define $\omega_E^* := \omega_E$ and $\tilde{\omega}_E^* := \tilde{\omega}_E$.

Patch of an element. The patch of an element $T \in \mathcal{T}$ is defined by

$$\omega_T := \bigcup_{T' \in \tilde{\omega}_T} T', \quad \text{where} \quad \tilde{\omega}_T := \{T' \in \mathcal{T} \mid T \cap T' \in \mathcal{E}\}.$$

2.4. Jump terms. For $T \in \mathcal{T}$, $E \subseteq \partial T$, and $\varphi \in H^1(T)$, let $\varphi|_{E,T}$ denote the trace of φ on E . Now let $E \in \mathcal{E}_E$ be an interior elementary edge and T_E and T_W the unique elements with $E = T_E \cap T_W$. For a $\{T_E, T_W\}$ -piecewise H^1 function φ , the jump of φ on E is defined by

$$[[\varphi]]_E := \varphi|_{E,T_E} - \varphi|_{E,T_W}.$$

Note that $[[\varphi]]_E = 0$, provided $\varphi \in H^1(T_E \cup T_W)$. Moreover, for a $\{T_E, T_W\}$ -piecewise polynomial φ , the jump on E reads

$$[[\varphi]]_E(x) := \lim_{t \rightarrow 0_+} \varphi(x + t\mathbf{n}_E) - \lim_{t \rightarrow 0_+} \varphi(x - t\mathbf{n}_E) \quad \text{for all } x \in E.$$

For each nonelementary edge $E \in \mathcal{E}_H$, we define the jump $[[\varphi]]_E$ by

$$[[\varphi]]_E(x) := [[\varphi]]_{E_i}(x) \quad \text{for all } x \in E_i,$$

where $E = \bigcup_{i=1}^n E_i$ with $E_1, \dots, E_n \in \mathcal{E}_E$.

3. Cell-centered finite volume method. This section summarizes the discretization for the cell-centered finite volume method for our model problem. It especially points out the difference between the approximation of the diffusive flux on an admissible mesh and an almost regular mesh.

3.1. Discretization ansatz. We integrate the strong form (1.1) over a control volume $T \in \mathcal{T}$ and use the Gauss divergence theorem to obtain

$$\int_T f \, dx = - \int_T \Delta u \, dx = - \int_{\partial T} \frac{\partial u}{\partial \mathbf{n}_T} \, ds = - \sum_{E \in \mathcal{E}_T} \sigma_{T,E} \int_E \frac{\partial u}{\partial \mathbf{n}_E} \, ds \quad \text{for all } T \in \mathcal{T}.$$

With the diffusive flux $\Phi_E^D(u) = \int_E \partial u / \partial \mathbf{n}_E \, ds$, we get the so-called balance equation

$$(3.1) \quad - \sum_{E \in \mathcal{E}_T} \sigma_{T,E} \Phi_E^D(u) = \int_T f \, dx \quad \text{for all } T \in \mathcal{T}.$$

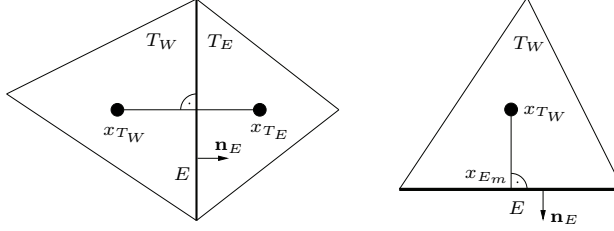


FIG. 3.1. The orthogonality condition for $E \in \mathcal{E}_E$ (left) (resp., $E \in \mathcal{E}_D$ (right)) for an admissible mesh in the sense of [10].

For the cell-centered finite volume method, one replaces the continuous diffusion flux $\Phi_E^D(u)$ by a discrete diffusion flux $F_E^D(u_h)$, which is discussed in section 3.2. Here $u_h \in \mathcal{P}_0(\mathcal{T})$ is a piecewise constant approximation of u , namely $u_T := u_h|_T \approx u(x_T)$, where x_T denotes the center of an element $T \in \mathcal{T}$. The discrete problem thus reads as follows: Find $u_h \in \mathcal{P}_0(\mathcal{T})$ such that

$$-\sum_{E \in \mathcal{E}_T} \sigma_{T,E} F_E^D(u_h) = \int_T f \, dx \quad \text{for all } T \in \mathcal{T}.$$

3.2. Discretization of diffusion flux. Note that $\Phi_E^D(u_h) = \int_E g \, ds$ is known for a Neumann edge $E \in \mathcal{E}_N$. One therefore defines

$$F_E^D(u_h) := \Phi_E^D(u_h) = \int_E g \, ds \quad \text{for } E \in \mathcal{E}_N.$$

Moreover, for a nonelementary edge with $E = \bigcup_{i=1}^n E_i$ and $E_i \in \mathcal{E}_E$, there holds $\Phi_E^D(u) = \sum_{i=1}^n \Phi_{E_i}^D(u)$, which leads to the definition

$$(3.2) \quad F_E^D(u_h) := \sum_{i=1}^n F_{E_i}^D(u_h) \quad \text{for all } E_1, \dots, E_n \in \mathcal{E}_E \quad \text{and} \quad E = \bigcup_{i=1}^n E_i \in \mathcal{E}_H.$$

Therefore, it remains only to define $F_E^D(u_h)$ for $E \in \mathcal{E}_E \cup \mathcal{E}_D$.

Admissible meshes. For an admissible mesh in the sense of [10, Definition 9.1], a first-order difference scheme leads to

$$(3.3) \quad \Phi_E^D(u) \approx F_E^D(u_h) := \begin{cases} \frac{u_{T_E} - u_{T_W}}{|x_{T_E} - x_{T_W}|} h_E & \text{if } E \in \mathcal{E}_E \text{ and } E = T_W \cap T_E, \\ \frac{u_{E_m} - u_{T_W}}{|x_{E_m} - x_{T_W}|} h_E & \text{if } E \in \mathcal{E}_D \text{ and } E = T_W \cap \Gamma_D, \end{cases}$$

with $u_{T_W} = u_h|_{T_W} \approx u(x_{T_W})$ and $u_{T_E} \approx u(x_{T_E})$ as well as for $E \in \mathcal{E}_D$, $u_{E_m} \approx u_D(x_{E_m})$.

The admissibility of the mesh \mathcal{T} allows one to choose the centers x_T for $T \in \mathcal{T}$ in a way that the edges $E = T_W \cap T_E$ for any $T_W, T_E \in \mathcal{T}$ are orthogonal to the directions $x_{T_E} - x_{T_W}$; see Figure 3.1. For general meshes, it is not possible to choose the centers x_T appropriately, and the approximation (3.3) is not consistent [10].

Remark 3.1. Even if a triangular mesh is admissible in the sense of [10, Definition 9.1], local mesh refinement is nontrivial: One has to guarantee that all angles are strictly less than $\pi/2$; i.e., one cannot avoid remeshing of the domain. For rectangular meshes, local mesh refinement cannot avoid hanging nodes. This, however, contradicts the admissibility condition.

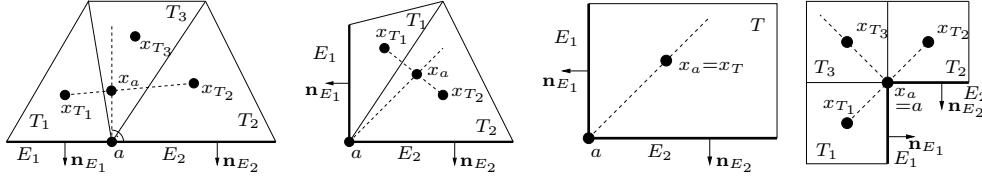


FIG. 3.2. The different cases for calculating u_a with $a \in \mathcal{N}_N$ and $E_1, E_2 \in \mathcal{E}_N$.

Diamond path method. A possible choice of $F_E^D(u_h)$ for general meshes is the so-called *diamond path method*, which has been mathematically analyzed in [6, 7] for rectangular meshes with a maximum of one hanging node per edge. For each node $a \in \mathcal{N}$, we define

$$(3.4) \quad u_a = \begin{cases} \sum_{T \in \tilde{\omega}_a} \psi_T(a) u_T & \text{for all } a \in \mathcal{N}_F \cup \mathcal{N}_H, \\ u_D(a) & \text{for all } a \in \mathcal{N}_D, \\ \bar{u}_a + \bar{g}_a & \text{for all } a \in \mathcal{N}_N \end{cases}$$

for certain weights $\{\psi_T(a) \mid T \in \mathcal{T}, a \in \mathcal{N}_T\}$. For details on the computation of the weights, the reader is referred to [5, 6, 7, 9]. We stress that the computation can be done in linear complexity with respect to the number $\#\mathcal{T}$ of elements.

We remark only on the computation of \bar{u}_a and \bar{g}_a in the case of a Neumann node $a \in \mathcal{N}_N$; see Figure 3.2: Two edges $E_1, E_2 \in \mathcal{E}_N$ correspond to $a \in \mathcal{N}_N$ such that $\{a\} = E_1 \cap E_2$. Let \mathbf{n}_j denote the normal vector of E_j . In the case of $\#\tilde{\omega}_a > 1$, let $T_1, T_2 \in \tilde{\omega}_a$ with $T_1 \neq T_2$. We define x_a as the intersection of the line $\gamma_1(s) = a + s(\mathbf{n}_1 + \mathbf{n}_2)/2$ and the line $\gamma_2(t) = t(x_{T_1} - x_{T_2})$. Moreover, provided $\#\tilde{\omega}_a > 2$, we assume that $|x_a - a|$ is minimized over all pairs $T_1, T_2 \in \tilde{\omega}_a$. Then $\bar{u}_a \approx u(x_a)$ is interpolated linearly from u_{T_1} and u_{T_2} :

$$\bar{u}_a = \frac{u_{T_2} - u_{T_1}}{|x_{T_2} - x_{T_1}|} |x_a - x_{T_1}| + u_{T_1}.$$

For $\mathbf{n}_1 = \mathbf{n}_2$, we choose

$$\bar{g}_a = |x_a - a| \left(\frac{1}{|E_1|} \int_{E_1} g \, ds + \frac{1}{|E_2|} \int_{E_2} g \, ds \right) / 2,$$

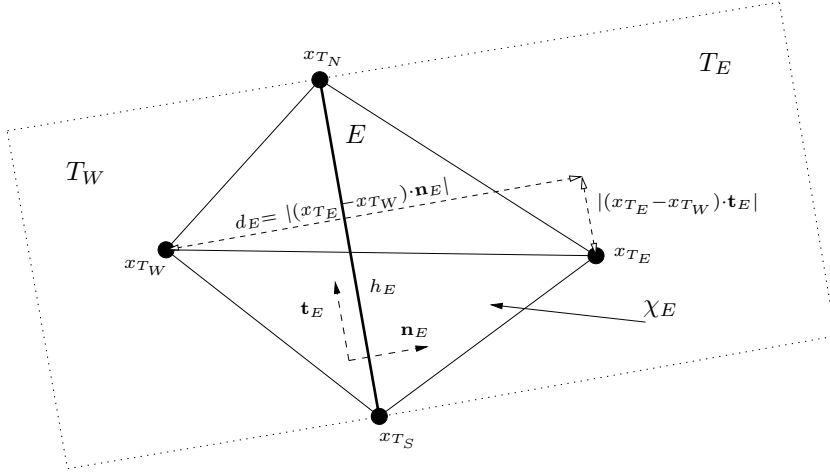
and, finally, for $\mathbf{n}_1 \neq \mathbf{n}_2$, we choose

$$\bar{g}_a = \lambda \frac{1}{|E_1|} \int_{E_1} g \, ds + \mu \frac{1}{|E_2|} \int_{E_2} g \, ds,$$

where $\lambda, \mu \in \mathbb{R}$ are calculated from the linear equation $a - x_a = \lambda \mathbf{n}_1 + \mu \mathbf{n}_2$. In the case $\tilde{\omega}_a = \{T\}$, i.e., a is the node of only one element $T \in \mathcal{T}$, we choose $x_a = x_T$ and $\bar{u}_a = u_T$, whereas \bar{g}_a is computed as before.

Remark 3.2. Provided $x_a = a$, we obtain $a - x_a = 0$, $\lambda = \mu = 0$, and $\bar{g}_a = 0$.

With the notation from Figure 3.3, where x_{T_S} and x_{T_N} are the starting and end

FIG. 3.3. Diamond path with domain χ_E .

points of $E \in \mathcal{E}_E \cup \mathcal{E}_D$, we compute $F_E^D(u_h)$. For an elementary edge $E \in \mathcal{E}_E$,

$$(3.5) \quad F_E^D(u_h) := h_E \left(\frac{u_{T_E} - u_{T_W}}{d_E} - \alpha_E \frac{u_{T_N} - u_{T_S}}{h_E} \right)$$

$$\text{with} \quad \alpha_E = \frac{(x_{T_E} - x_{T_W}) \cdot \mathbf{t}_E}{(x_{T_E} - x_{T_W}) \cdot \mathbf{n}_E}, \quad d_E = (x_{T_E} - x_{T_W}) \cdot \mathbf{n}_E.$$

Here the additional unknowns u_{T_N} and u_{T_S} are located at the nodes x_{T_N} and x_{T_S} and are computed by (3.4). For a boundary edge $E \in \mathcal{E}_D$, we compute $F_E^D(u_h)$ by (3.5), where x_{T_E} is now replaced by the midpoint x_{E_m} of E and u_{T_E} becomes $u_D(x_{E_m})$.

4. Morley interpolant. Let $u_h \in \mathcal{P}_0(\mathcal{T})$ be the computed discrete solution. In this section, we define an interpolant $\mathcal{I}u_h$ which is appropriate for the a posteriori error analysis, the definition of which is an extension of the definition in [11, section 5] to the case of hanging nodes and Neumann nodes.

Triangular Morley element. Let $T = \text{conv}\{a_1, a_2, a_3\} \subset \mathbb{R}^2$ be a nondegenerate triangle with edges $E_j = \text{conv}\{a_j, a_{j+1}\}$, where $a_4 := a_1$. The standard Morley element $(T, \mathcal{P}_T, \Sigma_T)$ is given by $\mathcal{P}_T = \mathcal{P}_2$ and $\Sigma_T = (S_1, \dots, S_6)$, where

$$S_j(p) = p(a_j) \quad \text{and} \quad S_{j+3}(p) = \int_{E_j} \frac{\partial p}{\partial \mathbf{n}_{T, E_j}} ds \quad \text{for } j = 1, \dots, 3 \quad \text{and } p \in \mathcal{P}_2.$$

Note that $S_{j+3}(p) = h_{E_j} \partial p(m_j) / \partial \mathbf{n}_{T, E_j}$, where $m_j := (a_j + a_{j+1})/2$ denotes the midpoint of E_j , so that this definition is consistent with [2, section 8.3].

Rectangular Morley element. Let $T = \text{conv}\{a_1, a_2, a_3, a_4\} \subset \mathbb{R}^2$ be a nondegenerate rectangle with edges E_j . A Morley-type element $(T, \mathcal{P}_T, \Sigma_T)$ is then given by $\mathcal{P}_T = \mathcal{P}_2 \oplus \text{span}\{x^3 - 3xy^2, y^3 - 3yx^2\}$ and $\Sigma_T = (S_1, \dots, S_8)$, where

$$S_j(p) = p(a_j) \quad \text{and} \quad S_{j+4}(p) = \int_{E_j} \frac{\partial p}{\partial \mathbf{n}_{T, E_j}} ds \quad \text{for } j = 1, \dots, 4 \quad \text{and } p \in \mathcal{P}_T;$$

cf. [11, section 4.2]. Note that the polynomials $x^3 - 3xy^2$ and $y^3 - 3yx^2$, which enrich the ansatz space, are harmonic.

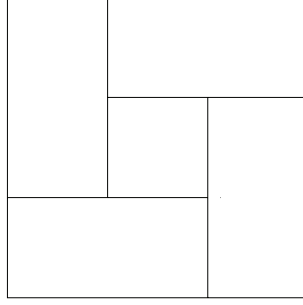


FIG. 4.1. An almost regular triangulation, where the elementwise and recursive computation of $\mathcal{I}u_h$ does not stop.

The Morley interpolant. In either of the cases, that T is a nondegenerate triangle or rectangle, the Morley element $(T, \mathcal{P}_T, \Sigma_T)$ is a nonconforming finite element. The Morley interpolant $\mathcal{I}u_h$ satisfies elementwise $(\mathcal{I}u_h)|_T \in \mathcal{P}_T$ for all $T \in \mathcal{T}$ defined by the following properties (4.1)–(4.3): For each free node $a \in \mathcal{N}_T \cap \mathcal{N}_F$, the value $\mathcal{I}u_h(a)$ satisfies

$$(4.1) \quad (\mathcal{I}u_h)|_T(a) = \sum_{T_a \in \tilde{\omega}_a} \psi_{T_a}(a) u_h|_{T_a},$$

where the weights $\psi_{T_a}(a)$ are the same as for the computation of u_h by the use of the diamond cell method. For each boundary node, the value $\mathcal{I}u_h(a)$ is prescribed:

$$(4.2) \quad (\mathcal{I}u_h)|_T(a) = \begin{cases} u_D(a) & \text{for } a \in \mathcal{N}_T \cap \mathcal{N}_D, \\ \bar{u}_a + \bar{g}_a & \text{for } a \in \mathcal{N}_T \cap \mathcal{N}_N, \end{cases}$$

where the calculation of \bar{u}_a and \bar{g}_a was discussed in section 3. For each hanging node $a \in \mathcal{N}_T \cap \mathcal{N}_H$, there holds

$$(4.3) \quad (\mathcal{I}u_h)|_T(a) = (\mathcal{I}u_h)|_{T_a}(a),$$

where $T_a \in \mathcal{T}$ is the unique element with $a \in \text{int}(E)$ for some (nonelementary) edge $E \in \mathcal{E}_{T_a}$. For each edge $E \in \mathcal{E}_T$, there holds

$$(4.4) \quad \int_E \frac{\partial(\mathcal{I}u_h)|_T}{\partial \mathbf{n}_E} ds = F_E^D(u_h),$$

where $F_E^D(u_h)$ is the numerical flux from section 3.2.

LEMMA 4.1. *The Morley interpolant $\mathcal{I}u_h$ is uniquely defined by (4.1)–(4.4). Moreover, $\mathcal{I}u_h$ is continuous in all nodes $a \in \mathcal{N}$ but not globally continuous in Ω .*

Proof. For an element $T \in \mathcal{T}$ without hanging nodes, i.e., $\mathcal{N}_T \cap \mathcal{N}_H = \emptyset$, the interpolant $(\mathcal{I}u_h)|_T$ is uniquely defined by (4.1)–(4.3) and (4.4) since $(T, \mathcal{P}_T, \Sigma_T)$ is a finite element. \square

Remark 4.1. The computation of $\mathcal{I}u_h$ can be performed by solving a large system of linear equations which is coupled through the hanging nodes. However, normally $\mathcal{I}u_h$ can be computed locally by solving a 6×6 (resp., 8×8) system for each element $T \in \mathcal{T}$. For an element $T \in \mathcal{T}$ without hanging nodes, the interpolant $(\mathcal{I}u_h)|_T$ is uniquely determined by (4.1)–(4.2) and (4.4). For an element with hanging nodes, we have to compute $(\mathcal{I}u_h)|_{T_a}$ first; cf. (4.3). This leads to a recursive algorithm. Figure 4.1 shows an almost regular triangulation, where the proposed recursion would not stop. Instead, one has to solve a global linear system to compute $\mathcal{I}u_h$.

Properties of Morley interpolant. From the definition of the discrete scheme and the property (4.3), we obtain an additional orthogonality property of $\mathcal{I}u_h$.

LEMMA 4.2. *The residual $R := f + \Delta(\mathcal{I}u_h)$ is L^2 -orthogonal to $\mathcal{P}_0(\mathcal{T})$, i.e.,*

$$(4.5) \quad \int_T (f + \Delta(\mathcal{I}u_h)) dx = 0 \quad \text{for all } T \in \mathcal{T}.$$

In particular, the residual satisfies $R = f - f_{\mathcal{T}}$.

Proof. From integration by parts and the definition of the balance equation (3.1), we infer

$$\int_T \Delta(\mathcal{I}u_h)|_T dx = \int_{\partial T} \frac{\partial(\mathcal{I}u_h)|_T}{\partial \mathbf{n}_T} ds = \sum_{E \in \mathcal{E}_T} \sigma_{T,E} F_E^D(u_h) = - \int_T f(x) dx,$$

where we have used (4.4) in the second equality. In particular, there holds $R_{\mathcal{T}}|_T := |T|^{-1} \int_T R dx = 0$. With $\Delta_{\mathcal{T}}(\mathcal{I}u_h) \in \mathcal{P}_0(\mathcal{T})$, we obtain $R = R - R_{\mathcal{T}} = f - f_{\mathcal{T}}$. \square

According to the definition of $\mathcal{I}u_h$ on the Dirichlet and Neumann boundaries, namely (4.2) and (4.3), we obtain corresponding orthogonalities.

LEMMA 4.3. *For boundary edges, there hold*

$$(4.6) \quad \int_E \frac{\partial(u - \mathcal{I}u_h)}{\partial \mathbf{t}_E} ds = 0 \quad \text{for all } E \in \mathcal{E}_D$$

as well as

$$(4.7) \quad \int_E \frac{\partial(u - \mathcal{I}u_h)}{\partial \mathbf{n}_E} ds = 0 \quad \text{for all } E \in \mathcal{E}_N.$$

Proof. For $E \in \mathcal{E}_D$, let a_S and a_N be the starting and end points of E , respectively. Then

$$\int_E \frac{\partial(u - \mathcal{I}u_h)}{\partial \mathbf{t}_E} ds = (u - \mathcal{I}u_h)(a_N) - (u - \mathcal{I}u_h)(a_S) = 0$$

by the use of (4.2). To prove (4.7), we use (4.4) and the definition of the finite volume scheme:

$$\int_E \frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{n}_E} ds = F_E^D(u_h) = \int_E g ds,$$

where $g = \partial u / \partial \mathbf{n}$. \square

Finally, we note some orthogonality relations of the normal and tangential jumps of $\mathcal{I}u_h$ which again follow from (4.4) (in combination with (3.2)) and from the nodal values (4.1)–(4.3) of $\mathcal{I}u_h$.

LEMMA 4.4. *For the interior edges, there hold*

$$(4.8) \quad \int_E \left[\left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{n}_E} \right] \right] ds = 0 \quad \text{for all } E \in \mathcal{E}_0 \cup \mathcal{E}_H$$

as well as

$$(4.9) \quad \int_E \left[\left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{t}_E} \right] \right] ds = 0 \quad \text{for all } E \in \mathcal{E}_E.$$

Proof. We first prove (4.8) for $E \in \mathcal{E}_H$. There holds $E = \bigcup_{i=1}^n E_i$ with $E_1, \dots, E_n \in \mathcal{E}_E$, $E_i = T_{W_i} \cap T_E$ and \mathbf{n}_E shows from element T_{W_i} to T_E . Therefore, the definition (3.2) of the discrete flux on nonelementary edges implies

$$\begin{aligned} \int_E \left[\left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{n}_E} \right] \right] ds &= \int_E \frac{\partial(\mathcal{I}u_h)|_{T_E}}{\partial \mathbf{n}_E} ds - \sum_{i=1}^n \int_{E_i} \frac{\partial(\mathcal{I}u_h)|_{T_{W_i}}}{\partial \mathbf{n}_E} ds \\ &= F_E^D(u_h) - \sum_{i=1}^n F_{E_i}^D(u_h) = 0. \end{aligned}$$

For $E \in \mathcal{E}_0$, the proof of (4.8) works analogously with $n = 1$. To prove (4.9), let a_S and a_N be the starting and end points of $E \in \mathcal{E}_E$, respectively. Note that $[[\mathcal{I}u_h]]_E(a_N) = 0 = [[\mathcal{I}u_h]]_E(a_S)$ because of the continuity of $\mathcal{I}u_h$ in all nodes. Therefore,

$$\int_E \left[\left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{t}_E} \right] \right] ds = \int_{a_S}^{a_N} \left[\left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{t}_E} \right] \right] ds = [[\mathcal{I}u_h]]_E(a_N) - [[\mathcal{I}u_h]]_E(a_S) = 0,$$

which concludes the proof. \square

5. A posteriori error estimate. In this section, we provide a residual-based a posteriori error analysis for the error $u - \mathcal{I}u_h$, where \mathcal{I} denotes the Morley interpolant from section 4. The idea goes back to [11] and is now extended to almost regular triangulations and mixed boundary conditions. The mathematical techniques follow the a posteriori error analysis for nonconforming finite elements. Throughout, $\nabla_{\mathcal{T}}$ and $\Delta_{\mathcal{T}}$ denote the \mathcal{T} -piecewise gradient and Laplacian, respectively.

5.1. Residual-based error estimator. For each element $T \in \mathcal{T}$, we define the refinement indicator

$$\begin{aligned} \eta_T^2 &:= h_T^2 \|f - f_{\mathcal{T}}\|_{L^2(T)}^2 + \sum_{E \in \{E \in \mathcal{E}_E \mid E \subset \partial T\}} h_E^* \|[[\nabla_{\mathcal{T}}(\mathcal{I}u_h)]]\|_{L^2(E)}^2 \\ (5.1) \quad &+ \sum_{E \in \mathcal{E}_T \cap \mathcal{E}_N} h_E \left\| \frac{\partial(u - \mathcal{I}u_h)}{\partial \mathbf{n}_E} \right\|_{L^2(E)}^2 + \sum_{E \in \mathcal{E}_T \cap \mathcal{E}_D} h_E \left\| \frac{\partial(u - \mathcal{I}u_h)}{\partial \mathbf{t}_E} \right\|_{L^2(E)}^2. \end{aligned}$$

Here $f_{\mathcal{T}}$ denotes the \mathcal{T} -piecewise integral mean, i.e., $f_{\mathcal{T}}|_T := |T|^{-1} \int_T f dx$. Moreover, the length h_E^* of an edge $E \in \mathcal{E}_E$ is defined by

$$h_E^* := \begin{cases} h_{E'} & \text{if } E \subset E' \text{ for some } E' \in \mathcal{E}_H, \\ h_E & \text{else, i.e., } E \in \mathcal{E}_0. \end{cases}$$

The residual-based error estimator is then given by the ℓ_2 -sum $\eta = \left(\sum_{T \in \mathcal{T}} \eta_T^2 \right)^{1/2}$ of all refinement indicators. In the following sections, we prove that η is (up to terms of higher order) a lower and upper bound of the error $\|\nabla_{\mathcal{T}}(u - \mathcal{I}u_h)\|_{L^2(\Omega)}$ in the energy norm.

5.2. Reliability of error estimator.

THEOREM 5.1. *There is a constant $c_1 > 0$ which depends only on the shape of the elements in \mathcal{T} but neither on the size nor the number of elements such that*

$$\|\nabla_{\mathcal{T}}(u - \mathcal{I}u_h)\|_{L^2(\Omega)} \leq c_1 \left(\sum_{T \in \mathcal{T}} \eta_T^2 \right)^{1/2}.$$

Proof. To abbreviate notation, we use the symbol \lesssim if an estimate holds up to a multiplicative constant that depends only on the shape of the elements in \mathcal{T} . For $e := u - \mathcal{I}u_h$, the Helmholtz decomposition, e.g., from [4, Lemma 2.1], provides $v \in H^1(\Omega)$ and $w \in H^1(\Omega)$ such that

$$\nabla_{\mathcal{T}} e = \nabla v + \operatorname{curl} w \quad \text{and} \quad v|_{\Gamma_D} = 0 \quad \text{as well as} \quad \operatorname{curl} w \cdot \mathbf{n}|_{\Gamma_N} = 0.$$

Moreover, there holds

(5.2)

$$\|\nabla v\|_{L^2(\Omega)}^2 + \|\operatorname{curl} w\|_{L^2(\Omega)}^2 = \|\nabla_{\mathcal{T}} e\|_{L^2(\Omega)}^2 = \int_{\Omega} \nabla_{\mathcal{T}} e \cdot \nabla v \, dx + \int_{\Omega} \nabla_{\mathcal{T}} e \cdot \operatorname{curl} w \, dx.$$

Note that $0 = \operatorname{curl} w \cdot \mathbf{n} = \partial w / \partial \mathbf{t}$ on Γ_N . Since Γ_N is connected, we infer that w is constant on Γ_N . Subtracting a constant, we may therefore guarantee $w|_{\Gamma_N} = 0$. We now estimate the two addends on the right-hand side separately. The first term reads

$$\int_{\Omega} \nabla_{\mathcal{T}} e \cdot \nabla v \, dx = \sum_{T \in \mathcal{T}} \int_T Rv \, dx + \int_{\Gamma_N} gv \, ds - \sum_{T \in \mathcal{T}} \int_{\partial T} \frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{n}_T} v \, ds$$

according to elementwise integration by parts and the definition of the residual $R := f + \Delta_{\mathcal{T}}(\mathcal{I}u_h)$. We now consider the sum over the boundary integrals, namely

$$\sum_{T \in \mathcal{T}} \int_{\partial T} \frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{n}_T} v \, ds = \sum_{T \in \mathcal{T}} \sum_{E \in \mathcal{E}_T} \sigma_{T,E} \int_E \frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{n}_E} v \, ds.$$

For $E \in \mathcal{E}_D$, the boundary integral vanishes due to $v|_{\Gamma_D} = 0$. The Neumann edges $E \in \mathcal{E}_N$ are combined with the boundary integral $\int_{\Gamma_N} g \, ds$. Each edge $E \in \mathcal{E}_0$ appears twice for associated elements T_W and T_E , respectively. The normal vectors $\sigma_{T_W,E} \mathbf{n}_E$ and $\sigma_{T_E,E} \mathbf{n}_E$ differ only in the sign so that we obtain the jump of the normal derivative on E . For a nonelementary edge $E \in \mathcal{E}_H$ with $E = \bigcup_{i=1}^n E_i$ and $E_i \in \mathcal{E}_E$, both E as well as the elementary edges E_i appear only once in the sum. Similarly to the prior arguments we are led to the jump of the normal derivative on E , where we make use of $\mathbf{n}_E = -\mathbf{n}_{E_i}$ for all $i = 1, \dots, n$. Altogether, we obtain

$$\begin{aligned} \int_{\Omega} \nabla_{\mathcal{T}} e \cdot \nabla v \, dx &= \sum_{T \in \mathcal{T}} \int_T Rv \, dx - \sum_{E \in \mathcal{E}_0 \cup \mathcal{E}_H} \int_E \left[\left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{n}_E} \right] \right] v \, ds \\ &\quad + \sum_{E \in \mathcal{E}_N} \int_E \left(g - \frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{n}_E} \right) v \, ds \\ &= \sum_{T \in \mathcal{T}} \int_T R(v - v_T) \, dx - \sum_{E \in \mathcal{E}_0 \cup \mathcal{E}_H} \int_E \left[\left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{n}_E} \right] \right] (v - v_E) \, ds \\ &\quad + \sum_{E \in \mathcal{E}_N} \int_E \frac{\partial e}{\partial \mathbf{n}_E} (v - v_E) \, ds, \end{aligned}$$

where we have applied the orthogonalities (4.5), (4.7), and (4.8) for the integral means $v_T = |T|^{-1} \int_T v \, dx$ and $v_E := h_E^{-1} \int_E v \, ds$, respectively. We now apply the Cauchy inequality combined with a Poincaré inequality $\|v - v_T\|_{L^2(T)} \lesssim h_T \|\nabla v\|_{L^2(T)}$ for the

first sum and a trace inequality $\|v - v_E\|_{L^2(E)} \lesssim h_E^{1/2} \|\nabla v\|_{L^2(T_E)}$ for the remaining sums, where $T_E \in \mathcal{T}$ is an arbitrary element with $E \in \mathcal{E}_{T_E}$. This leads to

$$\begin{aligned}
\int_{\Omega} \nabla_{\mathcal{T}} e \cdot \nabla v \, dx &\lesssim \left(\sum_{T \in \mathcal{T}} h_T^2 \|R\|_{L^2(T)}^2 \right)^{1/2} \left(\sum_{T \in \mathcal{T}} \|\nabla v\|_{L^2(T)}^2 \right)^{1/2} \\
&\quad + \left(\sum_{E \in \mathcal{E}_0 \cup \mathcal{E}_H} h_E \left\| \left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{n}_E} \right] \right\|_{L^2(E)}^2 \right)^{1/2} \left(\sum_{E \in \mathcal{E}_0 \cup \mathcal{E}_H} \|\nabla v\|_{L^2(T_E)}^2 \right)^{1/2} \\
&\quad + \left(\sum_{E \in \mathcal{E}_N} h_E \left\| \frac{\partial e}{\partial \mathbf{n}_E} \right\|_{L^2(E)}^2 \right)^{1/2} \left(\sum_{E \in \mathcal{E}_N} \|\nabla v\|_{L^2(T_E)}^2 \right)^{1/2} \\
&\lesssim \left[\left(\sum_{T \in \mathcal{T}} h_T^2 \|R\|_{L^2(T)}^2 \right)^{1/2} + \left(\sum_{E \in \mathcal{E}_E} h_E^* \left\| \left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{n}_E} \right] \right\|_{L^2(E)}^2 \right)^{1/2} \right. \\
(5.3) \quad &\quad \left. + \left(\sum_{E \in \mathcal{E}_N} h_E \left\| \frac{\partial e}{\partial \mathbf{n}_E} \right\|_{L^2(E)}^2 \right)^{1/2} \right] \|\nabla v\|_{L^2(\Omega)}.
\end{aligned}$$

For the second integral in (5.2), we proceed in the same manner: Elementwise integration by parts yields

$$\int_{\Omega} \nabla_{\mathcal{T}} e \cdot \operatorname{curl} w \, dx = - \sum_{T \in \mathcal{T}} \int_{\partial T} \frac{\partial e}{\partial \mathbf{t}_T} w \, ds = - \sum_{T \in \mathcal{T}} \sum_{E \in \mathcal{E}_T \setminus \mathcal{E}_N} \sigma_{T,E} \int_E \frac{\partial e}{\partial \mathbf{t}_E} w \, ds,$$

since $w|_{\Gamma_N} = 0$. Treating the interior edges as before, we obtain

$$\sum_{T \in \mathcal{T}} \sum_{E \in \mathcal{E}_T} \sigma_{T,E} \int_E \frac{\partial e}{\partial \mathbf{t}_E} w \, ds = \sum_{E \in \mathcal{E}_E} \int_E \left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{t}_E} \right] w \, ds + \sum_{E \in \mathcal{E}_D} \int_E \frac{\partial e}{\partial \mathbf{t}_E} w \, ds,$$

where we have used that, for an interior edge E , the tangential jump of an H^1 -function vanishes, i.e., $[\partial u / \partial \mathbf{t}_E]_E = 0$. With the orthogonalities (4.9) and (4.6), we prove

$$\int_{\Omega} \nabla_{\mathcal{T}} e \cdot \operatorname{curl} w \, dx = - \sum_{E \in \mathcal{E}_E} \int_E \left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{t}_E} \right] (w - w_E) \, ds - \sum_{E \in \mathcal{E}_D} \int_E \frac{\partial e}{\partial \mathbf{t}_E} (w - w_E) \, ds$$

for the integral mean $w_E := h_E^{-1} \int_E w \, ds$. As before, the application of the Cauchy inequality and the trace inequality yields

$$\begin{aligned}
(5.4) \quad &\int_{\Omega} \nabla_{\mathcal{T}} e \cdot \operatorname{curl} w \, dx \\
&\lesssim \left[\sum_{E \in \mathcal{E}_E} h_E^* \left\| \left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{t}_E} \right] \right\|_{L^2(E)}^2 + \sum_{E \in \mathcal{E}_D} h_E \left\| \frac{\partial e}{\partial \mathbf{t}_E} \right\|_{L^2(E)}^2 \right]^{1/2} \|\nabla w\|_{L^2(\Omega)}.
\end{aligned}$$

If we finally combine (5.2)–(5.4), we prove

$$\begin{aligned}
\|\nabla_{\mathcal{T}} e\|_{L^2(\Omega)} &\lesssim \left[\sum_{T \in \mathcal{T}} h_T^2 \|R\|_{L^2(T)}^2 + \sum_{E \in \mathcal{E}_E} h_E^* \left\| \left[\nabla \mathcal{I}u_h \right] \right\|_{L^2(E)}^2 \right. \\
&\quad \left. + \sum_{E \in \mathcal{E}_N} h_E \left\| \frac{\partial e}{\partial \mathbf{n}_E} \right\|_{L^2(E)}^2 + \sum_{E \in \mathcal{E}_D} h_E \left\| \frac{\partial e}{\partial \mathbf{t}_E} \right\|_{L^2(E)}^2 \right]^{1/2},
\end{aligned}$$

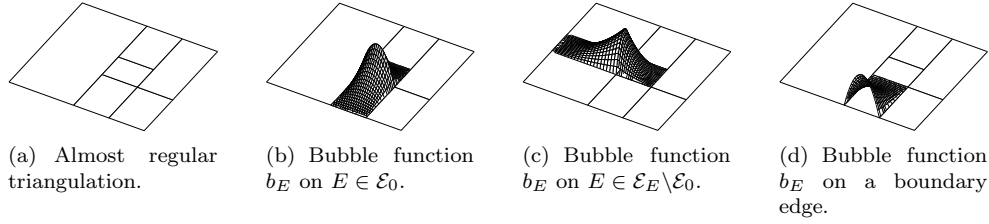


FIG. 5.1. The different types of the edge-bubble functions on an almost regular triangulation.

where we have used that $\{\mathbf{t}_E, \mathbf{n}_E\}$ is an orthonormal basis of \mathbb{R}^2 and that $\|\nabla v\|_{L^2} \leq \|\nabla_{\mathcal{T}} e\|_{L^2}$ as well as $\|\nabla w\|_{L^2} = \|\operatorname{curl} w\|_{L^2} \leq \|\nabla_{\mathcal{T}} e\|_{L^2}$. This and $R = f - f_{\mathcal{T}}$ conclude the proof. \square

5.3. Local efficiency of error estimator. To prove the efficiency of the proposed error estimator, we need to control the constant $c_2 > 0$ in the estimate $h_E \leq h_E^* \leq c_2 h_E$ uniformly for all $E \in \mathcal{E}_E$.

THEOREM 5.2. *There is a constant $c_3 > 0$ which depends only on c_2 and the shape of the elements in \mathcal{T} but neither on the size nor the number of elements such that*

$$(5.5) \quad \eta_T^2 \leq c_3 (\|\nabla_{\mathcal{T}}(u - \mathcal{I}u_h)\|_{L^2(\omega_T)}^2 + h_T^2 \|f - f_{\mathcal{T}}\|_{L^2(\omega_T)}^2) \quad \text{for all } T \in \mathcal{T}.$$

The proof follows along the arguments of Verfürth [12] by use of appropriate bubble functions b_E (see Figure 5.1) and an edge lifting operator F_{ext} . The elementary edge patch ω_E^* is defined in a way that it belongs to a locally regular triangulation. For $E \in \mathcal{E}_E$, we may therefore adopt the notation of b_E and F_{ext} from the literature [12].

LEMMA 5.3. *For each edge $E \in \mathcal{E}_E \cup \mathcal{E}_D \cup \mathcal{E}_N$, there is a $\tilde{\omega}_E^*$ -piecewise polynomial bubble function $b_E \in H^1(\omega_E^*)$ with $0 \leq b_E \leq 1$ such that, for all $w \in \mathcal{P}_p(E)$, there holds*

$$(5.6) \quad c_4 \|w\|_{L^2(E)} \leq \|w b_E^{1/2}\|_{L^2(E)} \leq \|w\|_{L^2(E)}.$$

The constant $c_4 > 0$ depends only on the shape of the elements of \mathcal{T} and the polynomial degree p . Moreover, for $E \in \mathcal{E}_E$, the bubble function satisfies $b_E \in H_0^1(\omega_E^*)$, whereas for a boundary edge $E \in \mathcal{E}_D \cup \mathcal{E}_N$ there holds $b_E|_{\partial\omega_E^* \setminus E} = 0$.

LEMMA 5.4. *For each edge $E \in \mathcal{E}_E \cup \mathcal{E}_D \cup \mathcal{E}_N$, there is a lifting operator $F_{\text{ext}} : \mathcal{P}_p(E) \rightarrow H^1(\omega_E^*)$ such that $F_{\text{ext}}(w)|_E = w$, for $w \in \mathcal{P}_p(E)$, as well as*

$$(5.7) \quad c_5 h_E^{1/2} \|w\|_{L^2(E)} \leq \|F_{\text{ext}}(w) b_E\|_{L^2(\omega_E^*)} \leq c_6 h_E^{1/2} \|w\|_{L^2(E)}$$

and

$$(5.8) \quad \|\nabla(F_{\text{ext}}(w) b_E)\|_{L^2(\omega_E^*)} \leq c_7 h_E^{-1/2} \|w\|_{L^2(E)}.$$

The constants $c_5, c_6, c_7 > 0$ depend only on the shape of the elements in \mathcal{T} and the polynomial degree p . Here b_E denotes the bubble function from Lemma 5.3.

The proof of Theorem 5.2 is now split into four claims which dominate the different edge contributions of η_T^2 separately. Throughout the proofs, we adopt the foregoing notation for $e = u - \mathcal{I}u_h$, $R = f + \Delta_{\mathcal{T}}(\mathcal{I}u_h)$, and \lesssim .

CLAIM 1. *There holds $h_E \left\| \left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{n}_E} \right] \right\|_{L^2(E)}^2 \lesssim \|\nabla_{\mathcal{T}} e\|_{L^2(\omega_E^*)}^2 + h_E^2 \|R\|_{L^2(\omega_E^*)}^2$ for each $E \in \mathcal{E}_E$.*

Proof. We first stress that $u \in H^1(\Omega)$ implies $\left[\left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{n}_E}\right]\right]_E = -\left[\left[\frac{\partial e}{\partial \mathbf{n}_E}\right]\right]_E$. With $b_E \in H_0^1(\omega_E^*)$ the corresponding edge-bubble function, (5.6) yields

$$\left\| \left[\left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{n}_E} \right] \right] \right\|_{L^2(E)}^2 \lesssim \int_E \left[\left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{n}_E} \right] \right]_E v \, ds \quad \text{with} \quad v := F_{\text{ext}} \left(\left[\left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{n}_E} \right] \right]_E \right) b_E \in H_0^1(\omega_E^*).$$

We rewrite the right-hand side and use integration by parts to prove

$$\int_E \left[\left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{n}_E} \right] \right]_E v \, dx = - \sum_{T \in \tilde{\omega}_E^*} \int_{\partial T} \frac{\partial e}{\partial \mathbf{n}_T} v \, dy = - \sum_{T \in \tilde{\omega}_E^*} \left(\int_T \nabla e \cdot \nabla v \, dx - \int_T Rv \, dx \right).$$

With the help of (5.7)–(5.8), the Cauchy inequality proves

$$\left\| \left[\left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{n}_E} \right] \right] \right\|_{L^2(E)}^2 \lesssim (h_E^{-1} \|\nabla_{\mathcal{T}} e\|_{L^2(\omega_E^*)}^2 + h_E \|R\|_{L^2(\omega_E^*)}^2)^{1/2} \left\| \left[\left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{n}_E} \right] \right] \right\|_{L^2(E)}. \quad \square$$

CLAIM 2. *There holds $h_E \left\| \left[\left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{t}_E} \right] \right] \right\|_{L^2(E)}^2 \lesssim \|\nabla_{\mathcal{T}} e\|_{L^2(\omega_E^*)}^2$ for each $E \in \mathcal{E}_E$.*

Proof. With $b_E \in H_0^1(\omega_E^*)$ the corresponding edge-bubble function, we observe

$$\left\| \left[\left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{t}_E} \right] \right] \right\|_{L^2(E)}^2 \lesssim \int_E \left[\left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{t}_E} \right] \right]_E v \, ds \quad \text{with} \quad v := F_{\text{ext}} \left(\left[\left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{t}_E} \right] \right]_E \right) b_E \in H_0^1(\omega_E^*).$$

As before, we rewrite the right-hand side and use integration by parts to prove

$$\int_E \left[\left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{t}_E} \right] \right]_E v \, ds = \sum_{T \in \tilde{\omega}_E^*} \int_{\partial T} \frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{t}_T} v \, dx = - \sum_{T \in \tilde{\omega}_E^*} \int_T \nabla(\mathcal{I}u_h) \cdot \text{curl} \, v \, dx.$$

Together with $\int_{\omega_E^*} \nabla u \cdot \text{curl} \, v \, dx = 0$ and (5.8), we obtain

$$\left\| \left[\left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{t}_E} \right] \right] \right\|_{L^2(E)}^2 \lesssim \int_{\omega_E^*} \nabla_{\mathcal{T}} e \cdot \text{curl} \, v \, dx \lesssim h_E^{-1/2} \|\nabla_{\mathcal{T}} e\|_{L^2(\omega_E^*)} \left\| \left[\left[\frac{\partial(\mathcal{I}u_h)}{\partial \mathbf{t}_E} \right] \right] \right\|_{L^2(E)},$$

where we used $\|\text{curl} \, v\|_{L^2} = \|\nabla v\|_{L^2}$. \square

CLAIM 3. *For $E \in \mathcal{E}_D$, there holds $h_E \left\| \frac{\partial e}{\partial \mathbf{t}_E} \right\|_{L^2(E)}^2 \lesssim \|\nabla e\|_{L^2(T)}^2$.*

Proof. For $E \in \mathcal{E}_D$, there is a unique element $\omega_E = T \in \mathcal{T}$ with $E \in \mathcal{E}_T$. The corresponding edge-bubble function $b_E \in H^1(T)$ satisfies $b_E|_{\partial T \setminus E} = 0$. We consider $v := F_{\text{ext}}(\partial e / \partial \mathbf{t}_E) b_E \in H^1(T)$ and note that $v|_{\partial T \setminus E} = 0$ as well as $\mathbf{t}_T|_E = \mathbf{t}_E$. Therefore,

$$\left\| \frac{\partial e}{\partial \mathbf{t}_E} \right\|_{L^2(E)}^2 \lesssim \int_E \frac{\partial e}{\partial \mathbf{t}_E} v \, ds = - \int_T \nabla e \cdot \text{curl} \, v \, dx \leq h_E^{-1/2} \|\nabla e\|_{L^2(T)} \left\| \frac{\partial e}{\partial \mathbf{t}_E} \right\|_{L^2(E)},$$

where we finally used the Cauchy inequality together with (5.8). \square

CLAIM 4. *For $E \in \mathcal{E}_N$, there holds $h_E \left\| \frac{\partial e}{\partial \mathbf{n}_E} \right\|_{L^2(E)}^2 \lesssim \|\nabla e\|_{L^2(T)}^2 + h_E^2 \|R\|_{L^2(T)}^2$.*

Proof. As in Claim 3, let $T \in \mathcal{T}$ be the unique element with $\omega_E = T$ for a fixed edge $E \in \mathcal{E}_N$ and let $b_E \in H^1(T)$ be the associated edge bubble function. With $v := F_{\text{ext}}(\partial e / \partial \mathbf{n}_E) b_E \in H^1(T)$ and integration by parts, there holds

$$\left\| \frac{\partial e}{\partial \mathbf{n}_E} \right\|_{L^2(E)}^2 \lesssim \int_{\partial T} \frac{\partial e}{\partial \mathbf{n}_T} v \, ds = \int_T \nabla e \cdot \nabla v \, dx - \int_T Rv \, dx.$$

The proof now follows as in Claim 1. \square

Proof of Theorem 5.2. According to Claims 1 and 2, there holds

$$\begin{aligned} & \sum_{E \in \{E \in \mathcal{E}_E \mid E \subset \partial T\}} h_E^* \|\llbracket \nabla_{\mathcal{T}}(\mathcal{I}u_h) \rrbracket\|_{L^2(E)}^2 \\ & \lesssim \sum_{E \in \{E \in \mathcal{E}_E \mid E \subset \partial T\}} (\|\nabla_{\mathcal{T}} e\|_{L^2(\omega_E^*)}^2 + h_E^2 \|R\|_{L^2(\omega_E^*)}^2) \\ & \lesssim \|\nabla_{\mathcal{T}} e\|_{L^2(\omega_T)}^2 + h_E^2 \|R\|_{L^2(\omega_T)}^2. \end{aligned}$$

With Claim 3, there holds

$$\sum_{E \in \mathcal{E}_T \cap \mathcal{E}_D} h_E \left\| \frac{\partial e}{\partial \mathbf{t}_E} \right\|_{L^2(E)}^2 \lesssim \sum_{E \in \mathcal{E}_T \cap \mathcal{E}_D} \|\nabla e\|_{L^2(T)}^2 \leq 4 \|\nabla e\|_{L^2(T)}^2.$$

With Claim 4, there holds

$$\begin{aligned} \sum_{E \in \mathcal{E}_T \cap \mathcal{E}_N} h_E \left\| \frac{\partial e}{\partial \mathbf{n}_E} \right\|_{L^2(E)}^2 & \lesssim \sum_{E \in \mathcal{E}_T \cap \mathcal{E}_N} (\|\nabla e\|_{L^2(T)}^2 + h_E^2 \|R\|_{L^2(T)}^2) \\ & \leq 4(\|\nabla e\|_{L^2(T)}^2 + h_E^2 \|R\|_{L^2(T)}^2). \end{aligned}$$

Finally, this and $R = f - f_{\mathcal{T}}$ prove (5.5). \square

6. Numerical experiments. In this section, we study the accuracy of the derived a posteriori error estimate from section 5 as well as the performance of an adaptive mesh-refining algorithm which is steered by the local refinement indicators η_T from (5.1). All computations are done in MATLAB. Throughout, we run the following standard algorithm, where we use $\theta = 1$ for uniform and $\theta = 0.5$ for adaptive mesh refinement, respectively.

ALGORITHM 6.1. *Given an initial mesh $\mathcal{T}^{(0)}$, $k = 0$, and $0 \leq \theta \leq 1$, do the following:*

1. *Compute the discrete solution $u_h \in \mathcal{P}_0(\mathcal{T}^{(k)})$ for the current mesh $\mathcal{T}^{(k)} = \{T_1, \dots, T_N\}$.*
2. *Compute the Morley interpolant $\mathcal{I}u_h$.*
3. *Compute the refinement indicators η_{T_j} for all elements $T_j \in \mathcal{T}^{(k)}$.*
4. *Construct a minimal subset $\mathcal{M}^{(k)}$ of $\mathcal{T}^{(k)}$ such that*

$$(6.1) \quad \theta \sum_{T \in \mathcal{T}^{(k)}} \eta_T^2 \leq \sum_{T \in \mathcal{M}^{(k)}} \eta_T^2$$

and mark all elements in $\mathcal{M}^{(k)}$ for refinement.

5. *Refine at least all marked elements $T \in \mathcal{M}^{(k)}$ and generate a new mesh $\mathcal{T}^{(k+1)}$.*

6. *Update $k \mapsto k + 1$ and go to (1).*

Remark 6.1. The marking criterion (6.1) was introduced in [8] to prove convergence of an adaptive algorithm for some P1-FEM for the Laplace problem. Despite convergence, even the question of optimal convergence rates of the adaptive FEM based on residual error estimators is well understood; cf. the recent work [3] for a precise statement of optimality and the history of mathematical arguments.

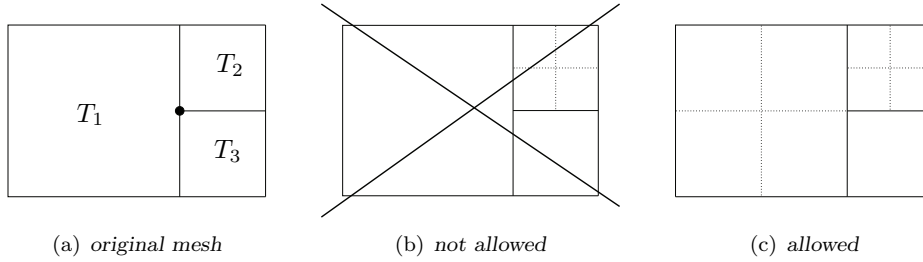


FIG. 6.1. To bound the constant c_2 which enters the efficiency estimate of Theorem 5.2, we allow only one hanging node per edge: If in configuration (a) the element T_2 is marked for refinement, we mark element T_1 for refinement as well. This leads to configuration (c) instead of (b) after refinement.

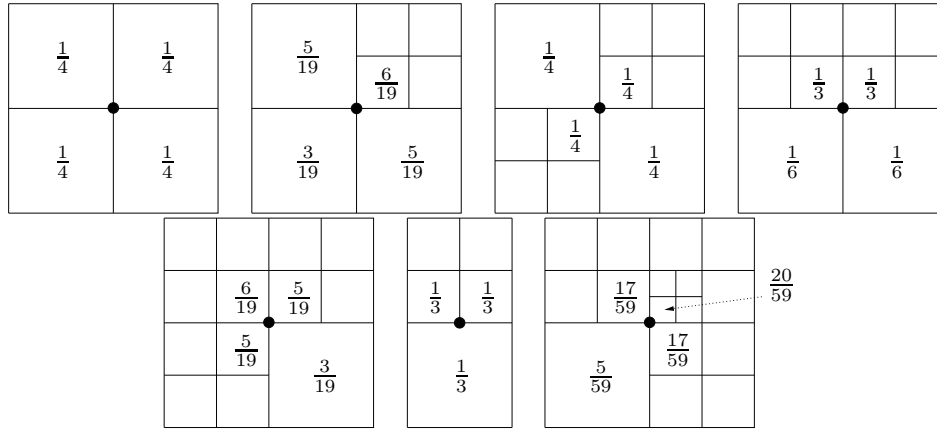


FIG. 6.2. A priori computed weights ψ_T for the special mesh of squares with at most one hanging node per edge.

In all experiments, the initial mesh $\mathcal{T}^{(0)}$ is a uniform and regular triangulation, where all of the elements are either triangles or squares. In the case of triangular elements, we use a red-green-blue strategy to obtain $\mathcal{T}^{(k+1)}$ from $\mathcal{T}^{(k)}$; i.e., marked elements are uniformly refined, and the obtained mesh is regularized by a green-blue closure [12]. In the case of square elements, a marked element is uniformly refined, and we allow hanging nodes. However, we do some additional marking to ensure the following assumption.

Assumption 6.1. For all almost regular meshes consisting of squares, there is at most one step of refinement between two neighboring cells; see Figure 6.1.

Note that under this assumption, there are only seven possible geometrical configurations for triangulations with square elements. This allows the a priori computation of the weights $\psi_T(a)$ in (3.4) and (4.1) which is shown in Figure 6.2.

Throughout, we compute and compare the following numerical quantities for uniform and adaptive mesh refinement: First, we have the Morley error

$$E_{\mathcal{T}} := \|\nabla_{\mathcal{T}}(u - \mathcal{I}u_h)\|_{L^2(\Omega)},$$

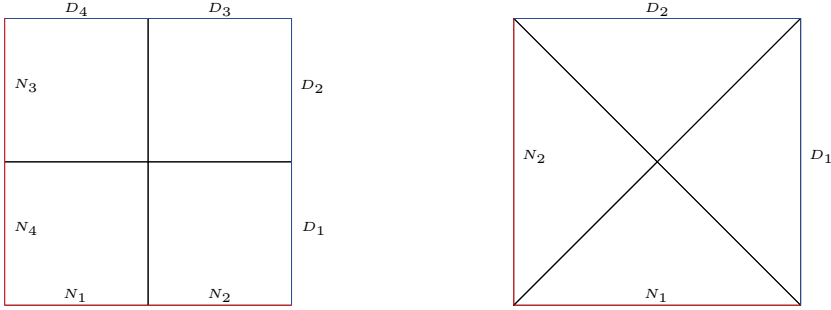


FIG. 6.3. Domain $\Omega = (0, 1)^2$ as well as Dirichlet and Neumann boundary conditions in the example in section 6.1. The initial mesh $\mathcal{T}^{(0)}$ consists of four squares (left) and four triangles (right), respectively.

where \mathcal{I} denotes the Morley interpolant. Second, we have the corresponding residual-based error estimator

$$\eta := \left(\sum_{T \in \mathcal{T}} \eta_T^2 \right)^{1/2},$$

where η_T are the refinement indicators of (5.1). Finally, we have the discretization error in the discrete H^1 -norm

$$E_h := \|u - u_h\|_{1,h} := \left(\|u - u_h\|_{L^2(\Omega)}^2 + |u_{\mathcal{T}} - u_h|_{1,h}^2 \right)^{1/2},$$

where $u_{\mathcal{T}} \in \mathcal{P}_0(\mathcal{T})$ is the \mathcal{T} -piecewise integral mean of u , i.e., $u_{\mathcal{T}}|_T = |T|^{-1} \int_T u \, dx$, and where the discrete H^1 -seminorm is defined by

$$|v_h|_{1,h} = \left(\sum_{E \in \mathcal{E}_E \cup \mathcal{E}_D} \left| \frac{v_{T_E} - v_{T_W}}{d_E} \right|^2 h_E d_E \right)^{1/2}$$

for any \mathcal{T} -piecewise constant function $v_h \in \mathcal{P}_0(\mathcal{T})$. According to [7], the diamond path method satisfies $E_h = \mathcal{O}(h)$ with $h = \max_{T \in \mathcal{T}} h_T$, provided $u \in H^2(\Omega)$. We stress that this, however, is proven only for locally refined Cartesian meshes, i.e., meshes consisting of rectangular elements with at most one hanging node per edge. In the case of triangular meshes, the proof still seems to be open.

6.1. Example with smooth solution. We consider the Laplace problem (1.1) on the unit square $\Omega = (0, 1)^2$ with prescribed exact solution

$$(6.2) \quad u(x, y) = \sinh(\pi x) \cos(\pi y) \quad \text{for } (x, y) \in \Omega.$$

Note that u is smooth and satisfies $f := \Delta u = 0$ so that the data oscillation term $\|h(f - f_{\mathcal{T}})\|_{L^2(\Omega)}$ of the error estimator η vanishes. We consider mixed boundary conditions, where the Dirichlet and Neumann data on

$$(6.3) \quad \Gamma_D = \{1\} \times [0, 1] \cup [0, 1] \times \{1\} \quad \text{and} \quad \Gamma_N = (0, 1) \times \{0\} \cup \{0\} \times (0, 1)$$

are computed from the given exact solution. The initial mesh $\mathcal{T}^{(0)}$ consists of either four squares or four triangles; see Figure 6.3.

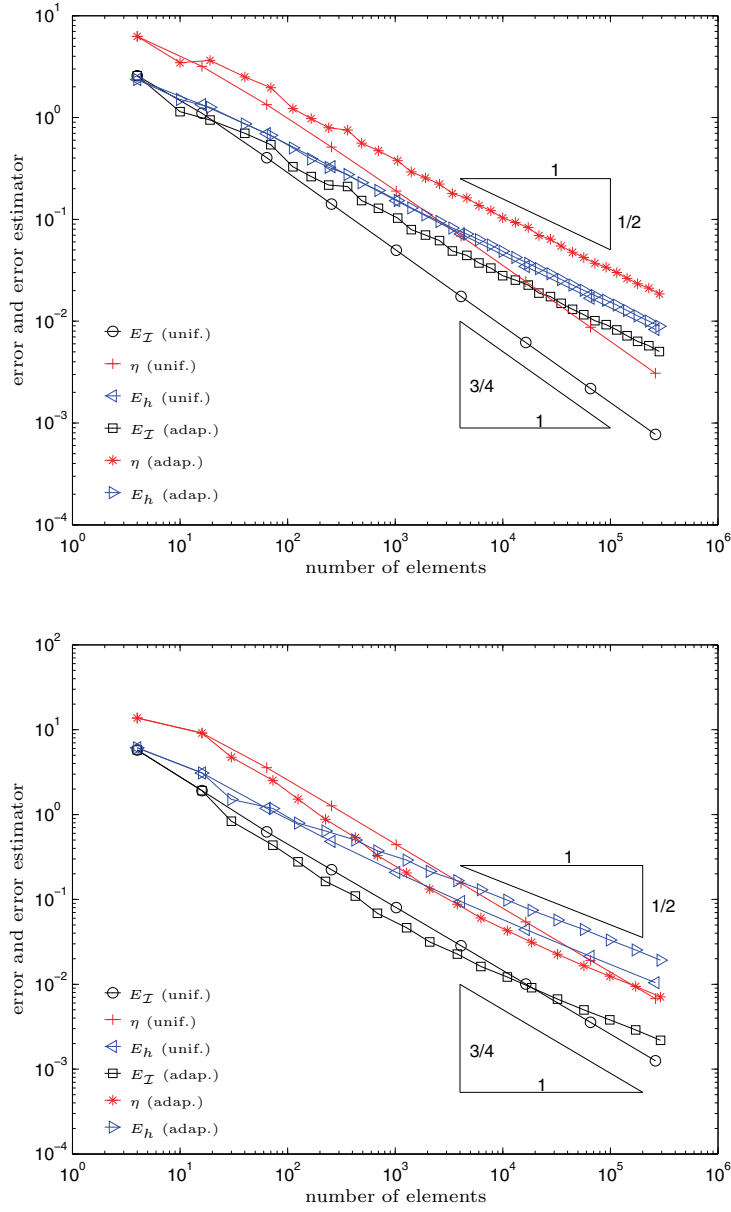


FIG. 6.4. Morley error $E_{\mathcal{I}} = \|\nabla_{\mathcal{T}}(u - \mathcal{I}u_h)\|_{L^2(\Omega)}$ and corresponding error estimator η as well as energy error $E_h = \|u - u_h\|_{1,h}$ in the example in section 6.1 for uniform and adaptive mesh refinement and triangulations consisting of squares (top) and triangles (bottom), respectively.

Figure 6.4 shows the curves of the errors $E_{\mathcal{I}} = \|\nabla_{\mathcal{T}}(u - \mathcal{I}u_h)\|_{L^2(\Omega)}$ and $E_h = \|u - u_h\|_{1,h}$ as well as the curve of the error estimator η with respect to uniform and adaptive mesh refinement. We plot the experimental results over the number of elements, where both axes are scaled logarithmically. Therefore, a straight line g with slope $-\alpha$ corresponds to a dependence $g = \mathcal{O}(N^{-\alpha})$, where $N = \#\mathcal{T}$ denotes

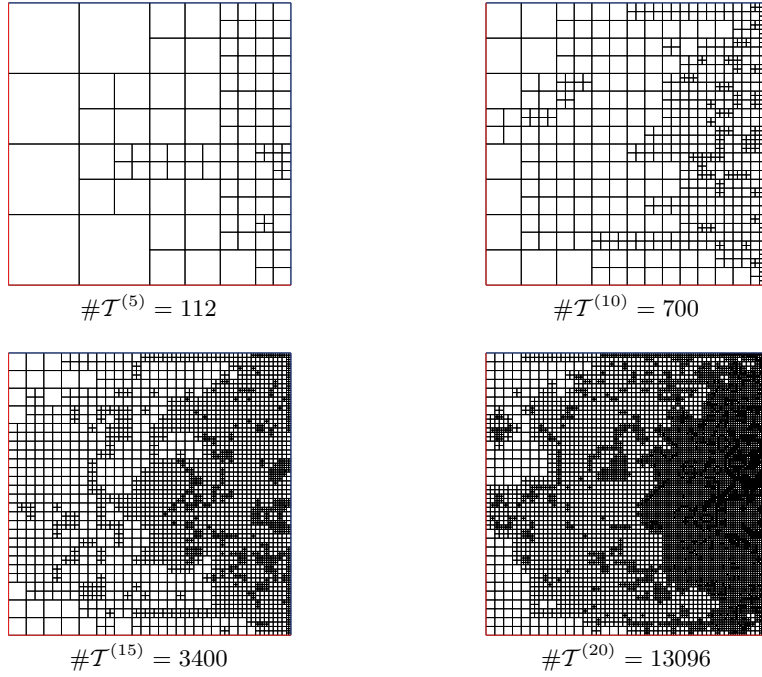


FIG. 6.5. Adaptively generated meshes $\mathcal{T}^{(k)}$ for $k = 5, 10, 15, 20$ with square elements in the example in section 6.1.

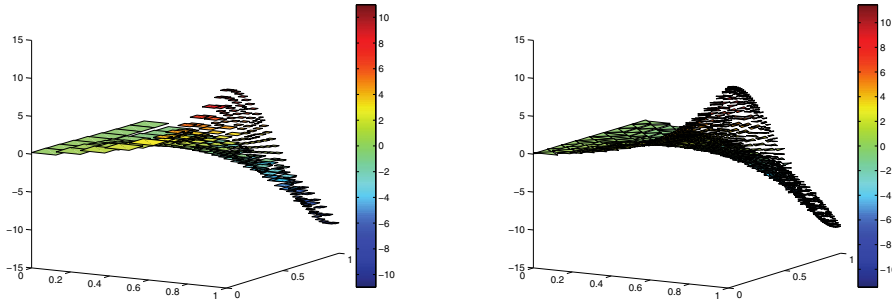


FIG. 6.6. \mathcal{T} -piecewise constant solutions u_h in the example in section 6.1 with respect to adaptively generated meshes $\mathcal{T}^{(8)}$ consisting of $\#\mathcal{T}^{(8)} = 361$ squares (left) and $\#\mathcal{T}^{(8)} = 1279$ triangles (right), respectively.

the number of elements. Note that, for uniform mesh refinement, the order $\mathcal{O}(N^{-\alpha})$ with respect to N corresponds to $\mathcal{O}(h^{2\alpha})$ with respect to the maximal mesh size $h := \max_{T \in \mathcal{T}} h_T$.

Because of $u \in H^2(\Omega)$, theory predicts the optimal order of convergence $E_h = \mathcal{O}(N^{-1/2})$ in the case of uniform mesh refinement and square elements. This is, in fact, observed. Moreover, in the case of square elements, the curves of E_h for uniform and adaptive mesh refinement almost coincide. However, the adaptive algorithm does not lead to uniformly refined meshes. Instead, the adaptive meshes plotted in Figure 6.5 show a certain refinement towards the edge $x = 1$ since the gradient of $u(x, y)$ is increasing with $x \rightarrow 1$ (see Figure 6.6), where we visualize some computed

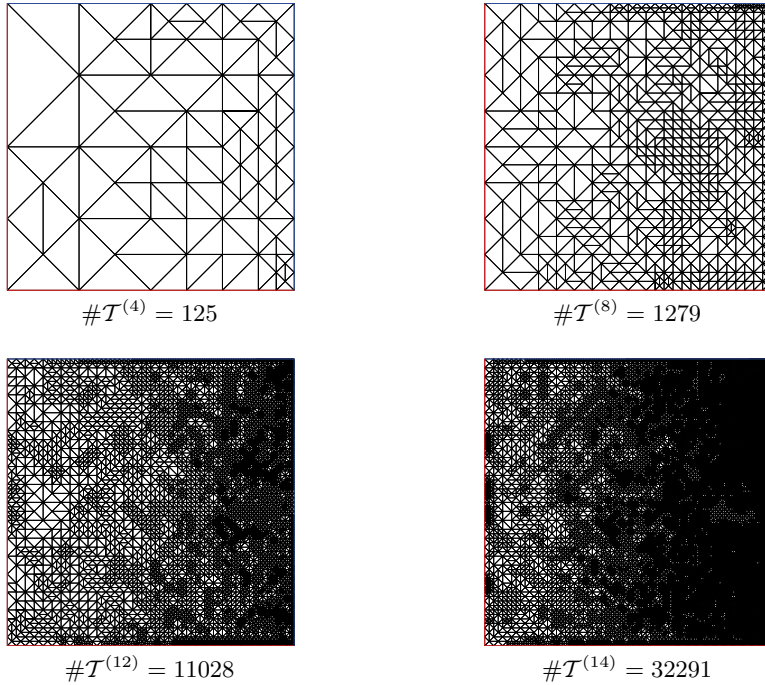


FIG. 6.7. *Adaptively generated meshes $\mathcal{T}^{(k)}$ for $k = 4, 8, 12, 14$ with triangular elements in the example in section 6.1.*

discrete solutions u_h . Although the Dirichlet and Neumann boundaries are not chosen symmetrically, the adaptive meshes appear to be almost symmetric with respect to the line $y = 1/2$, which corresponds to the symmetry $|\nabla u(x, 1/2 - y)| = |\nabla u(x, 1/2 + y)|$ of the exact solution.

For triangular elements, we observe the order $\mathcal{O}(N^{-1/2})$ for both uniform and adaptive mesh refinement. However, the absolute values of E_h are better in the case of uniform mesh refinement. As in the case of rectangular elements, we observe a certain refinement of the adaptively generated meshes towards the right edge $x = 1$ in Figure 6.7, and again they are almost symmetric related to the line $y = 1/2$.

For the Morley error $E_{\mathcal{T}}$ and uniform mesh refinement, we experimentally observe some superconvergence of order $3/4$ for both square and triangular elements in Figure 6.4. This superconvergence is destroyed by the use of adaptive mesh refinement, where we observe only a convergence order $1/2$. Independently of the mesh-refining strategy and the type of elements, we observe the theoretically predicted reliability and efficiency of the error estimator η : The curves of the Morley error $E_{\mathcal{T}}$ and the corresponding error estimator η are parallel up to a certain range.

6.2. Laplace problem with generic singularity. We consider the Laplace problem (1.1) on the L-shaped domain

$$(6.4) \quad \Omega = (-1, 1)^2 \setminus ([0, 1] \times [-1, 0])$$

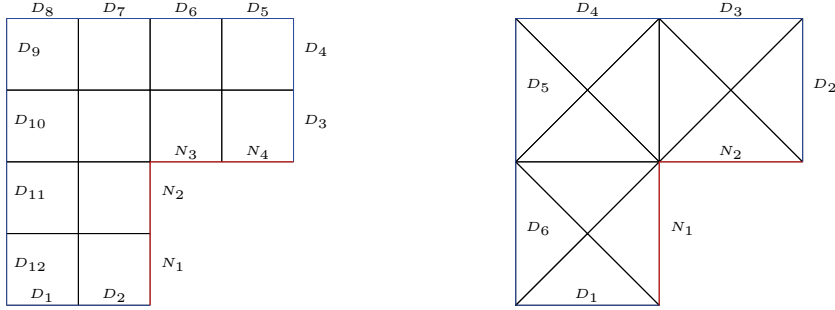


FIG. 6.8. *L-shaped domain as well as Dirichlet and Neumann boundary conditions in the Laplace problem in section 6.2. The initial mesh $\mathcal{T}^{(0)}$ consists of 12 squares (left) and 12 triangles (right), respectively.*

as shown in Figure 6.8. The given exact solution is the harmonic function $u(x, y) = \text{Im}((x + iy)^{2/3})$ and reads in polar coordinates

$$u(x, y) = r^{2/3} \sin(2\varphi/3) \quad \text{with} \quad (x, y) = r(\cos \varphi, \sin \varphi).$$

Note that u has a generic singularity at the reentrant corner $(0, 0)$, which leads to $u \in H^{1+2/3-\varepsilon}(\Omega)$ for all $\varepsilon > 0$. Therefore, a conforming finite element method with polynomial ansatz space leads to convergence of order $\mathcal{O}(h^{2/3})$ for the finite element error in the H^1 -norm, where h denotes the uniform mesh size. This corresponds to order $\mathcal{O}(N^{-1/3})$ with respect to the number of elements.

For the numerical computation, we prescribe the exact Neumann and Dirichlet data, where

$$\Gamma_D = \Gamma \setminus \Gamma_N \quad \text{and} \quad \Gamma_N := \{0\} \times (-1, 0) \cup (0, 1) \times \{0\}.$$

The initial meshes as well as Γ_D and Γ_N are shown in Figure 6.8. Note that Γ_N includes the reentrant corner, where the normal derivative $\partial u / \partial \mathbf{n}$ is singular.

Figure 6.9 plots the experimental results for the energy error E_h as well as for the Morley error $E_{\mathcal{I}}$ and the corresponding error estimator η over the number of elements. For uniform mesh refinement, the energy error E_h converges with a suboptimal order slightly better than $\mathcal{O}(N^{-1/3})$ for square elements and $\mathcal{O}(N^{-1/3})$ for triangular elements. The proposed adaptive strategy regains the optimal order of convergence $\mathcal{O}(N^{-1/2})$.

As can be expected from the finite element method, the Morley error $E_{\mathcal{I}}$ decreases like $\mathcal{O}(N^{-1/3})$ for uniform mesh refinement. The adaptive algorithm leads to an improved order of convergence $\mathcal{O}(N^{-1/2})$. For both mesh-refining strategies as well as for square and triangular elements, the error estimator η is observed to be reliable and efficient. For a sequence of adaptively generated meshes for both square and triangular elements, see Figures 6.10 and 6.11, respectively.

6.3. Laplace problem with inhomogeneous right-hand side. Finally, we consider the Laplace problem (1.1) on the L-shaped domain (6.4) from the previous experiment. The exact solution is prescribed by $u(x, y) = \text{Im}((x + iy)^{2/3}) + (x^2 + y^2)^{3/2}$ and reads in polar coordinates

$$u(x, y) = r^{2/3} \sin(2\varphi/3) + r^3 \quad \text{with} \quad (x, y) = r(\cos \varphi, \sin \varphi).$$

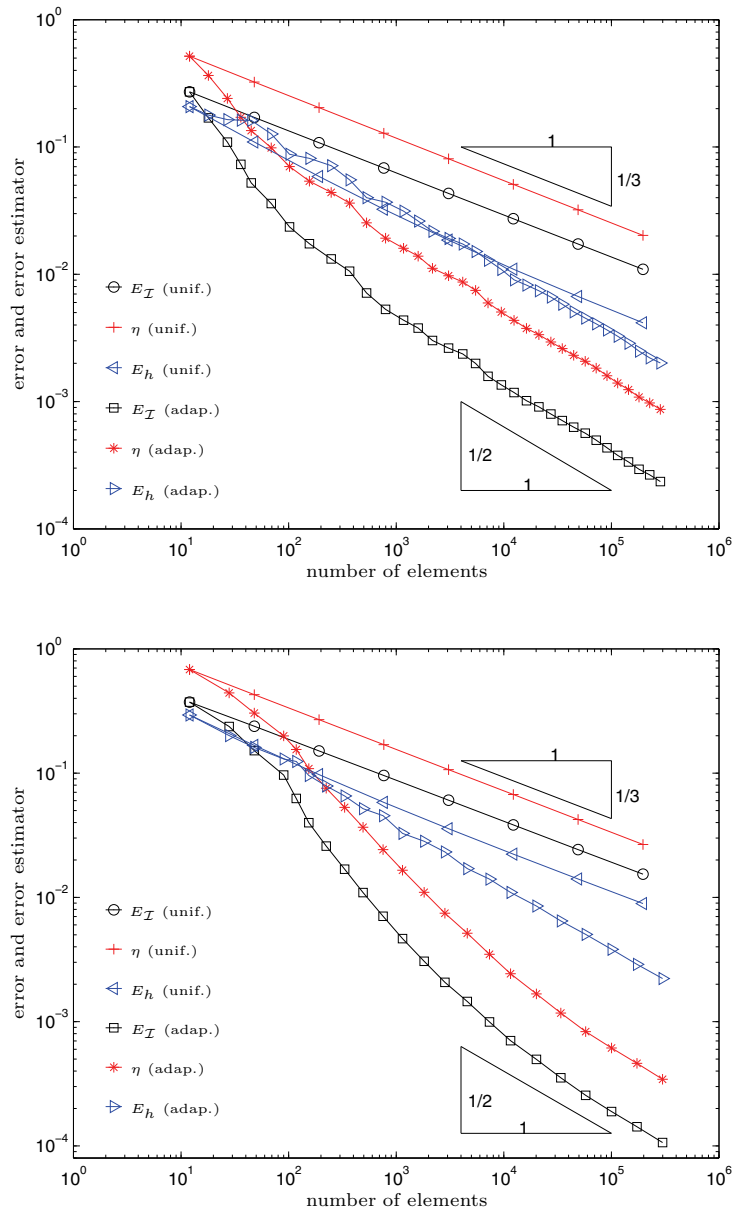


FIG. 6.9. Morley error $E_{\mathcal{I}} = \|\nabla_{\mathcal{T}}(u - \mathcal{I}u_h)\|_{L^2(\Omega)}$ and corresponding error estimator η as well as energy error $E_h = \|u - u_h\|_{1,h}$ in the Laplace problem in section 6.2 for uniform and adaptive mesh refinement and triangulations consisting of squares (top) and triangles (bottom), respectively.

Note that $f = -\Delta u$ reads $f(x, y) = -9(x^2 + y^2)^{1/2}$ (resp., $f(x, y) = -9r$) with respect to polar coordinates. We consider mixed boundary conditions with Γ_D and Γ_N as in the previous experiment. Figure 6.12 shows the numerical results of our computation. Despite a preasymptotic phase, where the f has to be resolved, we observe the same behavior as in the example in section 6.2.

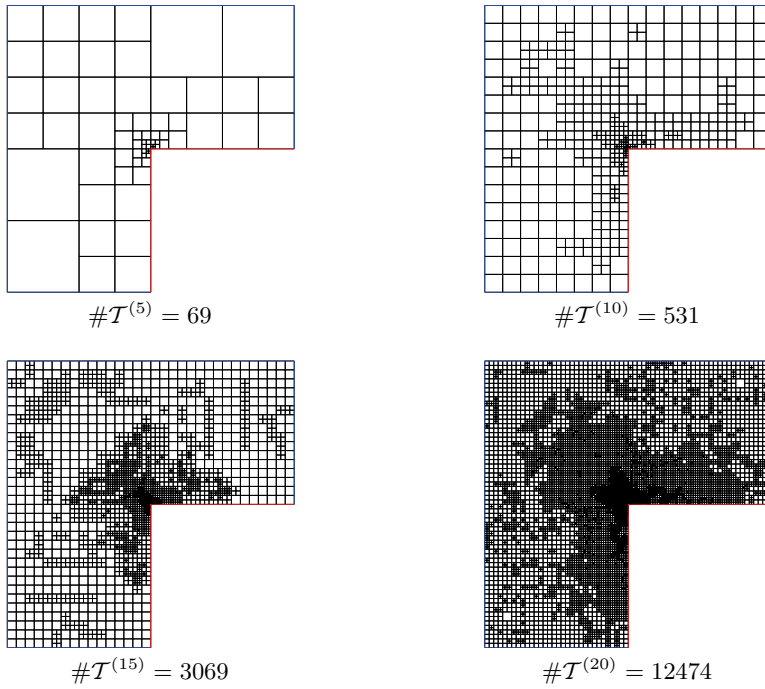


FIG. 6.10. Adaptively generated meshes $\mathcal{T}^{(k)}$ for $k = 5, 10, 15, 20$ with square elements in the Laplace problem in section 6.2.

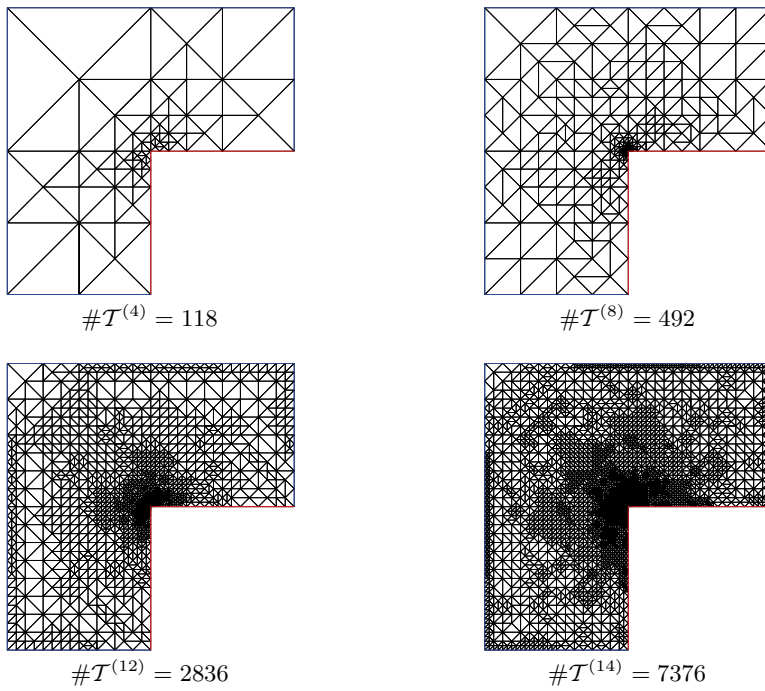


FIG. 6.11. Adaptively generated meshes $\mathcal{T}^{(k)}$ for $k = 4, 8, 12, 14$ with triangular elements in the Laplace problem in section 6.2.

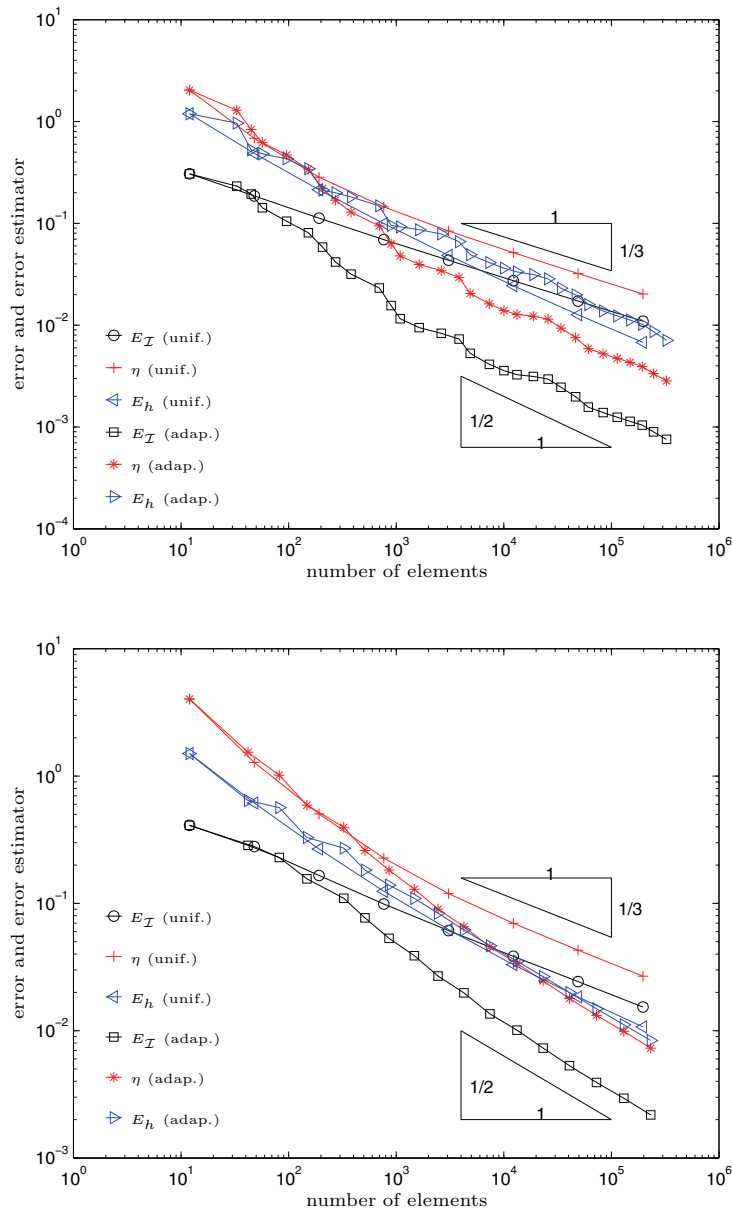


FIG. 6.12. Morley error $E_{\mathcal{T}} = \|\nabla_{\mathcal{T}}(u - \mathcal{I}u_h)\|_{L^2(\Omega)}$ and corresponding error estimator η as well as energy error $E_h = \|u - u_h\|_{1,h}$ in the Laplace problem in section 6.3 for uniform and adaptive mesh refinement and triangulations consisting of squares (top) and triangles (bottom), respectively.

REFERENCES

- [1] M. AINSWORTH AND J. T. ODEN, *A Posteriori Error Estimation in Finite Element Analysis*, John Wiley and Sons, New York, 2000.
- [2] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, Springer, New York, 1994.

- [3] J. M. CASCON, C. KREUZER, R. H. NOCHETTO, AND K. G. SIEBERT, *Quasi-optimal convergence rate for an adaptive finite element method*, SIAM J. Numer. Anal., 46 (2008), pp. 2524–2550.
- [4] S. COCHEZ-DHONDT AND S. NICAISE, *Equilibrated error estimators for discontinuous Galerkin methods*, Numer. Methods Partial Differential Equations, 24 (2008), pp. 1236–1252.
- [5] W. J. COIRIER, *An Adaptively-Refined, Cartesian, Cell-Based Scheme for the Euler and Navier-Stokes Equations*, Ph.D. thesis, University of Michigan, Ann Arbor, MI, 1994.
- [6] Y. COUDIÈRE, J. P. VILA, AND P. VILLEDIEU, *Convergence rate of a finite volume scheme for a two dimensional convection-diffusion problem*, M2AN Math. Model. Numer. Anal., 33 (2000), pp. 493–516.
- [7] Y. COUDIÈRE AND P. VILLEDIEU, *Convergence rate of a finite volume scheme for the linear convection-diffusion equation on locally refined meshes*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 1123–1149.
- [8] W. DÖRFLER, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.
- [9] C. ERATH, *Adaptive Finite Volumen Methode*, Diploma thesis, Vienna University of Technology, Vienna, Austria, 2005.
- [10] R. EYMARD, T. GALLOUËT, AND R. HERBIN, *Finite volume methods*, in Handbook of Numerical Analysis, Vol. 7, North-Holland, Amsterdam, 2000, pp. 713–1020.
- [11] S. NICAISE, *A posteriori error estimations of some cell-centered finite volume methods*, SIAM J. Numer. Anal., 43 (2005), pp. 1481–1503.
- [12] R. VERFÜRTH, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley-Teubner, New York, 1996.

A BDDC METHOD FOR MORTAR DISCRETIZATIONS USING A TRANSFORMATION OF BASIS*

HYEA HYUN KIM[†], MAKSYMILIAN DRYJA[‡], AND OLOF B. WIDLUND[§]

Abstract. A BDDC (balancing domain decomposition by constraints) method is developed for elliptic equations, with discontinuous coefficients, discretized by mortar finite element methods for geometrically nonconforming partitions in both two and three space dimensions. The coarse component of the preconditioner is defined in terms of one mortar constraint for each edge/face, which is the intersection of the boundaries of a pair of subdomains. A condition number bound of the form $C \max_i \{(1 + \log(H_i/h_i))^2\}$ is established under certain assumptions on the geometrically nonconforming subdomain partition in the three-dimensional case. Here H_i and h_i are the subdomain diameters and the mesh sizes, respectively. In the geometrically conforming case and the geometrically nonconforming cases in two dimensions, no assumptions on the subdomain partition are required. This BDDC preconditioner is also shown to be closely related to the Neumann–Dirichlet version of the FETI-DP algorithm. The results are illustrated by numerical experiments which confirm the theoretical results.

Key words. elliptic problems, finite elements, mortar methods, parallel algorithms, preconditioner, BDDC, FETI–DP, change of basis

AMS subject classifications. 65F10, 65N30, 65N55

DOI. 10.1137/070697859

1. Introduction. This study concerns a scalable BDDC (balancing domain decomposition by constraints) method for solving linear systems arising from mortar finite element discretizations of elliptic problems with discontinuous coefficients. BDDC methods were first introduced by Dohrmann [5] as an alternative to and an improvement of the balancing Neumann–Neumann methods. These more recent methods use different and more flexible coarse finite element spaces which lead to sparser linear systems. Additionally, as in the dual-primal finite element tearing and interconnecting (FETI-DP) methods, all linear systems actually solved have symmetric, positive definite coefficient matrices.

The coarse basis functions are related to a relatively small set of continuity constraints, across the interface between the subdomains, which are enforced throughout the iteration. In the standard, conforming finite element case, these constraints are given in terms of common values at subdomain vertices and/or common values of averages computed over subdomain edges and/or faces. We will refer to these as *primal* constraints and the corresponding subspace as the primal space of displacements. We

*Received by the editors July 20, 2007; accepted for publication (in revised form) July 16, 2008; published electronically October 24, 2008.

<http://www.siam.org/journals/sinum/47-1/69785.html>

[†]Department of Mathematics, Chonnam National University, Youngbong-dong, Buk-gu, Gwangju 500-757, Korea (hyeahyun@gmail.com or hkim@chonnam.ac.kr). This author’s research was supported in part by the U.S. Department of Energy under contract DE-FC02-01ER25482 and in part by the Post-doctoral Fellowship Program of Korea Science and Engineering Foundation (KOSEF).

[‡]Department of Mathematics, Warsaw University, Banacha 2, 02-097 Warsaw, Poland (dryja@mimuw.edu.pl). This author’s research was supported in part by the Polish Science Foundation under grant NN201006933.

[§]Department of Mathematics, Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012 (widlund@cs.nyu.edu, <http://cs.nyu.edu/cs/faculty/widlund/index.html>). This author’s research was supported in part by the U.S. Department of Energy under contracts DE-FG02-06ER25718 and DE-FC02-01ER25482 and in part by the National Science Foundation under grant NSF-DMS-0513251.

note that the theory (and practice) of BDDC methods for conforming finite elements is by now quite well developed; see [23, 26, 25].

In a FETI-DP method, a linear system, formulated for a set of Lagrange multipliers, is solved after eliminating the displacement variables. The resulting linear system, in itself, contains a coarse problem, which also is directly related to the primal constraints discussed above, i.e., they are given by matching conditions on averages over edges/faces and/or by enforcing continuity of the solutions at vertices. Its preconditioner, on the other hand, is built only from subdomain problems, while for a BDDC method a linear system in the original degrees of freedom is solved in an iteration with a preconditioner that has both coarse and subdomain components. This appears to provide the BDDC methods with more flexibility, e.g., in allowing for the use of inexact coarse problems. Thus, for standard finite element problems an inexact coarse problem can be introduced by applying the BDDC method recursively to the coarse problem; see Tu [30, 29] and a recent conference paper by Mandel, Sousedík, and Dohrmann [27]. The use of inexact local problems for the BDDC preconditioners has also been considered by Li and Widlund [24]. We also note that Klawonn and Rheinbach [17] have developed and extensively tested algorithms which use inexact solvers for the coarse problem of FETI-DP methods.

There are a number of articles on solving the algebraic problems given by the mortar discretizations considered in this paper; see [32] and the literature cited therein. Most of them concern the simpler case of geometrically conforming partitioning of the original region Ω ; see, however, Achdou, Maday, and Widlund [1], where some iterative substructuring methods are developed and analyzed for problems in two dimensions in the geometrically nonconforming case, and Kim and Widlund [13], where an additive Schwarz method with overlap is designed and analyzed. Among the papers on the geometrically conforming case that are related to this paper, we mention [14, 12], where a Neumann–Dirichlet version of a FETI-DP method is analyzed. In [6], a FETI-DP method is considered, which is a generalization of a variant known for the standard conforming discretization. To the best of our knowledge, BDDC methods for the mortar discretization have not previously been discussed in the literature even for the geometrically conforming case.

A condition number bound of the form $C(1 + \log(H/h))^2$ was first given for the BDDC operator by Mandel and Dohrmann [26] for a standard conforming discretization. This bound is of the same quality as the FETI-DP methods. In fact, the BDDC methods have been shown to be closely related to the FETI-DP methods. Thus, Mandel, Dohrmann, and Tezaur [25] have shown that the eigenvalues of the FETI-DP and BDDC operators are the same except possibly for eigenvalues equal to 0 and 1. More recently, a new formulation of the BDDC method was given by Li and Widlund [23]. They introduced a change of variables as well as an average operator for the BDDC method closely related to the jump operator used in [19] in the analysis of FETI-DP methods. The change of variables greatly simplifies the analysis; it has also led to a successful and robust implementation of FETI-DP methods; see [16, 18]. We note that the idea of changing the variables for FETI-DP algorithms was discussed already in [20]. We also note that FETI-DP algorithms have also been implemented using enough point constraints to assure that there are no floating subdomains. In addition, *optional admissible* primal constraints (e.g., averages over edges or faces) are added to enhance the rate of convergence of the iterations; see [9]. These constraints are then handled by a separate set of Lagrange multipliers. We note that in our context, we often have no point constraints, and therefore this second approach cannot be used.

In this paper, we will describe a BDDC method for mortar discretizations, after

a brief introduction to mortar methods. We will use a change of variables, as in [23] and Klawonn and Widlund [21], which is related to the primal constraints over edges/faces. We will consider quite general geometrically nonconforming partitions, i.e., we will not make any assumptions that the intersection of the boundaries of a pair of subdomains is a full face, edge, or a subdomain vertex.

We will work with mortar methods without any continuity constraints at subdomain vertices. Our results are valid for the traditional mortar methods as well as the dual basis mortar methods first introduced by Wohlmuth [31, 32]. We propose a preconditioner with a certain matrix of weights D and obtain the condition number bound, $C \max_i \{(1 + \log(H_i/h_i))^2\}$, under some assumptions on the geometrically nonconforming subdomain partition in three dimensions. When the algorithm is applied to a geometrically conforming partition in three dimensions or a geometrically nonconforming partition in two dimensions, we obtain the same bound without any assumption on the partition. The subdomain partition can have interfaces that are narrow faces and our bounds can be established for such quite general cases. Section 4 is devoted to proving our condition number bound in terms of a bound of an average operator E_D in an appropriate norm.

In section 5, we show that our BDDC preconditioner is closely connected to the Neumann–Dirichlet preconditioner for the FETI-DP methods given in [14, 12]. Connections are established between the average and jump operators, and the spectra of the BDDC and FETI-DP methods are then shown to be the same except possibly for an eigenvalue equal to 1.

Results of numerical experiments are reported in the final section and show that the FETI-DP and BDDC methods perform well and very similarly when the same set of primal constraints is selected.

Throughout this paper, C denotes a generic constant that depends neither on the mesh parameters nor on the coefficients of the elliptic problems.

We note that this paper originated from two projects developed separately by the first and second authors; the contribution of the third began with a suggestion that a theory could be developed for the geometrically nonconforming case using tools similar to those of [23].

2. Finite element spaces and mortar matching constraints.

2.1. A model problem and the mortar methods. We consider a model elliptic problem in a polygonal/polyhedral domain $\Omega \subset \mathbb{R}^2$ (\mathbb{R}^3): find $u \in H_0^1(\Omega)$ such that

$$(2.1) \quad \int_{\Omega} \rho(x) \nabla u(x) \cdot \nabla v(x) \, dx = \int_{\Omega} f(x) v(x) \, dx \quad \forall v \in H_0^1(\Omega),$$

where $\rho(x) \geq \rho_0 > 0$ and $f(x) \in L^2(\Omega)$.

We partition Ω into disjoint polygonal/polyhedral subdomains

$$\bar{\Omega} = \bigcup_{i=1}^N \bar{\Omega}_i.$$

As previously noted, the partition can be geometrically nonconforming; see further discussion below. We assume that $\rho(x) = \rho_i$, $x \in \Omega_i$ for some positive constant ρ_i .

We denote by X_i the P_1 -conforming finite element space on a quasi-uniform triangulation of the subdomain Ω_i . The finite element meshes typically do not align

across subdomain interfaces. The trace space of X_i on $\partial\Omega_i$ is denoted by W_i . We will use the product spaces

$$X := \prod_{i=1}^N X_i, \quad W := \prod_{i=1}^N W_i.$$

For functions in these spaces, we will impose mortar matching conditions across the interfaces using suitable spaces of Lagrange multipliers. Some of these matching conditions will be enforced throughout the iteration; they are directly related to the primal subspace.

In a geometrically nonconforming partition, the intersection of the boundaries of neighboring subdomains may be only part of an edge/face of a subdomain. We define the entire interface by

$$\Gamma = \left(\bigcup_{ij} \partial\Omega_i \cap \partial\Omega_j \right) \setminus \partial\Omega.$$

Among the subdomain edges/faces, we select nonmortar (slave) edges/faces F_l such that

$$\bigcup_l \overline{F_l} = \overline{\Gamma}, \quad F_l \cap F_k = \emptyset, \quad l \neq k;$$

see Figure 1 for an example of the selection of the nonmortar edges. For the case when $\rho(x)$ are very different across the interface, it is beneficial to select the part with smaller ρ_i as the nonmortar; see Assumption 4.2.

Since the subdomain partition can be geometrically nonconforming, a single nonmortar edge/face $F_l \subset \partial\Omega_i$ may intersect the boundaries of several other subdomains Ω_j . This provides F_l with a partition

$$\overline{F_l} = \bigcup_j \overline{F_{ij}}, \quad F_{ij} = \partial\Omega_i \cap \partial\Omega_j;$$

see Figure 1 for the mortar counter parts of the nonmortar edge F_l . A dual or standard Lagrange multiplier space $M(F_l)$ is introduced for each nonmortar edge/face F_l . We require $M(F_l)$ to have the same dimension as the space

$$(2.2) \quad \mathring{W}(F_l) := W_i|_{F_l} \cap H_0^1(F_l),$$

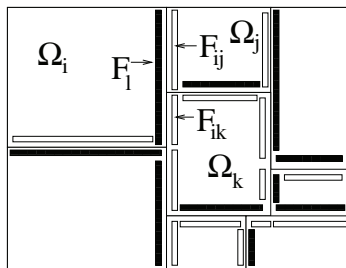


FIG. 1. Nonmortar edges (black) and mortar edges (white) in a geometrically nonconforming partition.

that it is nonempty, and that it contains the constants. Constructions of such Lagrange multiplier spaces are given in [2, 3] using standard Lagrange multiplier spaces, and in [31, 32] using dual Lagrange multiplier spaces; see also [11].

For $(w_1, \dots, w_N) \in W$, $w_i \in W_i$, we define $\phi_l \in L^2(F_l)$ by $\phi_l = w_j$ on $F_{ij} \subset F_l$. The mortar matching condition for the geometrically nonconforming partition is given by

$$(2.3) \quad \int_{F_l} (w_i - \phi_l) \lambda \, ds = 0 \quad \forall \lambda \in M(F_l), \forall F_l.$$

The mortar finite element method for problem (2.1) amounts to approximating the solution of the continuous problem by a Galerkin method using the mortar finite element space

$$\widehat{X} := \{v \in X : v|_\Gamma \text{ satisfies the mortar matching condition (2.3)}\},$$

where $v|_\Gamma$ is the restriction of v to the interface Γ . We introduce the space \widehat{W} as the restriction of \widehat{X} to Γ ,

$$\widehat{W} := \left\{ w : w = v|_\Gamma \, \forall v \text{ in } \widehat{X} \right\}.$$

2.2. Finite element spaces and a change of variables. In this subsection, we introduce a change of variables for some of the unknowns in the space W . It is based on the primal constraints that will be specified for our BDDC method. In mortar discretizations, we may consider the following sets of primal constraints: vertex constraints; vertex and edge average constraints, or edge average constraints only, for two dimensions; and vertex constraints and face average constraints, or face average constraints only, for three dimensions. We note that vertex constraints are appropriate only for the first generation of the mortar methods, in which case the subdomain vertex values are constrained to be continuous. In order to reduce the number of primal constraints, we can also select only some edges/faces as primal. Such choices have been considered for the FETI-DP methods and conforming finite elements in [21], and for mortar finite elements in [15].

In our BDDC formulation, we will select primal constraints over edges/faces from the set of mortar matching constraints (2.3). We consider $\{\lambda_{ij,k}\}_k$, the basis functions of $M(F_l)$ that are supported in $\overline{F}_{ij} \subset \overline{F}_l$, and define

$$(2.4) \quad \lambda_{ij} = \sum_k \lambda_{ij,k}.$$

We assume that at least one such basis function $\lambda_{ij,k}$ exists for each F_{ij} .

We now introduce one primal constraint over each interface $F_{ij} \subset F_l$ and for all edges/faces F_l ,

$$(2.5) \quad \int_{F_{ij}} (w_i - w_j) \lambda_{ij} \, ds = 0,$$

and define

$$(2.6) \quad \widetilde{W} = \{w \in W : w \text{ satisfies the primal constraints (2.5)}\}.$$

We note that $\widehat{W} \subset \widetilde{W} \subset W$, where \widehat{W} is the restriction of \widehat{X} to Γ . For the case of a geometrically conforming partition, i.e., when each F_{ij} is a full edge/face of two subdomains, these constraints are edge/face average matching conditions because $\lambda_{ij} = 1$. In addition to these constraints, vertex constraints can be considered but only if the partition is geometrically conforming.

Throughout this paper, we use hats for functions and function spaces that satisfy all of the mortar matching conditions. We use tildes for functions and function spaces that satisfy only the primal constraints across the subdomain interface.

Following Li and Widlund [23], we now introduce a change of variables based on the primal constraints. We provide details for the two-dimensional case but note that this approach can be extended to the three-dimensional case without any difficulty.

We recall that $F_l \subset \partial\Omega_i$, denoted from now on by F , is a nonmortar edge/face and that $\{F_{ij}\}_j$ is a partition of F given by $F_{ij} = F \cap \partial\Omega_j$, a mortar edge/face of Ω_j . We denote by $\{v_k\}_{k=1}^L$ the values of the unknowns of $w_i \in W_i$ at the nodes on F_{ij} , with nodal basis functions that are supported in \overline{F}_{ij} , and by $\{\eta_k\}_{k=1}^p$ the other unknowns on \overline{F}_{ij} . We will now define a transformation that retains the unknowns $\{\eta_k\}_{k=1}^p$ and changes $\{v_k\}_{k=1}^L$ into $\{\xi_k\}_{k=1}^L$ as follows: we pick one unknown ξ_m among $\{\xi_k\}_{k=1}^L$ and build a transformation $T_{F_{ij}}$ so that

$$(2.7) \quad \begin{pmatrix} \eta \\ v \end{pmatrix} = T_{F_{ij}} \begin{pmatrix} \eta \\ \xi \end{pmatrix}, \quad \xi_m = \frac{\int_{F_{ij}} w_i \lambda_{ij} ds}{\int_{F_{ij}} \lambda_{ij} ds}.$$

Here η , v , and ξ denote vectors of the unknowns $\{\eta_k\}_{k=1}^p$, $\{v_k\}_{k=1}^L$, and $\{\xi_k\}_{k=1}^L$, respectively.

Let

$$A_{\eta_k} = \frac{\int_{F_{ij}} \phi_{\eta_k} \lambda_{ij} ds}{\int_{F_{ij}} \lambda_{ij} ds}, \quad A_{v_k} = \frac{\int_{F_{ij}} \phi_{v_k} \lambda_{ij} ds}{\int_{F_{ij}} \lambda_{ij} ds},$$

where ϕ_{η_k} and ϕ_{v_k} are the nodal basis functions of the unknowns η_k and v_k , respectively. To make the presentation simpler, we assume that $p = 2$, but what follows can be generalized to any p . We will use the following transformation $T_{F_{ij}}$:

$$\begin{pmatrix} \eta_1 \\ \eta_2 \\ v_1 \\ \vdots \\ v_{m-1} \\ v_m \\ v_{m+1} \\ \vdots \\ v_L \end{pmatrix} = T_{F_{ij}} \begin{pmatrix} \eta_1 \\ \eta_2 \\ \xi_1 \\ \vdots \\ \xi_{m-1} \\ \xi_m \\ \xi_{m+1} \\ \vdots \\ \xi_L \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 & A & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & A & 0 & \cdots & 0 \\ c_1 & c_2 & r_1 & \cdots & r_{m-1} & A & r_{m+1} & \cdots & r_L \\ 0 & 0 & 0 & \cdots & 0 & A & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & A & 0 & \cdots & 1 \end{pmatrix}$$

$$\begin{pmatrix} \eta_1 \\ \eta_2 \\ \xi_1 \\ \vdots \\ \xi_{m-1} \\ \xi_m \\ \xi_{m+1} \\ \vdots \\ \xi_L \end{pmatrix} = A \begin{pmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 1 \end{pmatrix} \xi_m + \begin{pmatrix} \eta_1 \\ \eta_2 \\ \xi_1 \\ \vdots \\ \xi_{m-1} \\ \xi_0 \\ \xi_{m+1} \\ \vdots \\ \vdots \\ \vdots \\ \xi_L \end{pmatrix},$$

where

$$\xi_0 = c_1 \eta_1 + c_2 \eta_2 + r_1 \xi_1 + \cdots + r_{m-1} \xi_{m-1} + r_{m+1} \xi_{m+1} + \cdots + r_L \xi_L$$

and

$$A = \frac{\int_{F_{ij}} \lambda_{ij} ds}{\sum_{k=1}^L A_{v_k}}, \quad c_1 = -\frac{A_{\eta_1}}{A_{v_m}}, \quad c_2 = -\frac{A_{\eta_2}}{A_{v_m}}, \quad r_k = -\frac{A_{v_k}}{A_{v_m}}, \quad k \neq m.$$

We can then see that this transformation satisfies the (2.7) requirement. The transformation $T_{F_{ij}}$ can be applied to each face $F_{ij} \subset F$ independently, since it does not change any nodal values other than $\{v_k\}_{k=1}^L$, which are associated with the unknowns of the nodes interior to F_{ij} .

On the other side, the mortar side, of the interface F_{ij} , i.e., $F_{ij} \subset \partial\Omega_j$, we perform a change of basis to the unknowns in finite element space W_j . In this case, we introduce another set of unknowns $\{v_k\}_{k=1}^J$ and $\{\eta_k\}_{k=1}^p$. The unknowns $\{v_k\}_{k=1}^J$ are related to the nodes on F_{ij} with nodal basis functions, which belong to W_j and are supported in $\overline{F_{ij}}$. The unknowns $\{\eta_k\}_{k=1}^p$ are the remaining unknowns on F_{ij} . The transformation $T_{F_{ij}}$ is then defined for these unknowns similarly as for a nonmortar interface.

Using the transforms $T_{F_{ij}}$, we represent the Schur complement of the local stiffness and the mortar matching matrices, and the local force vector in the space of the new unknowns by

$$T^{(i)t} S^{(i)} T^{(i)}, \quad B^{(i)} T^{(i)}, \quad T^{(i)t} g^{(i)}.$$

Here $S^{(i)}$ is the reduced matrix obtained after eliminating all variables associated with only the subdomain Ω_i , and $T^{(i)}$ designates the transform of the original unknowns into the new unknowns of the subdomain boundary $\partial\Omega_i$. In the following, we will use the same notation, $S^{(i)}$, $B^{(i)}$, and $g^{(i)}$, for the matrices and vectors obtained after the change of unknowns, to simplify the notation. We will also use the notation W_i for the space of the new unknowns.

The unknowns ξ_m in (2.7), representing certain weighted averages over the edges, are the primal variables. Using the new variables, the space \widetilde{W} , defined in (2.6), can be represented as

$$(2.8) \quad \widetilde{W} = W_\Delta \oplus \widehat{W}_\Pi,$$

where W_Δ consists of the vectors of unknowns which are not primal unknowns, and \widetilde{W}_Π consists of the vectors of global, primal unknowns.

We now derive the matrix representation of the mortar matching condition (2.3) in the space \widetilde{W} of the new unknowns. The mortar matching condition (2.3) is redundant when enforced for the functions in the space \widetilde{W} . We recall that $\{\lambda_{ij,k}\}_k$ are the Lagrange multiplier basis elements supported in F_{ij} . To make the mortar matching condition nonredundant, we eliminate one basis element among $\{\lambda_{ij,k}\}_k$ for each $F_{ij} \subset F_l$, and we denote the reduced Lagrange multiplier space by $\overline{M}(F_l)$. The entire nonredundant Lagrange multiplier space is then defined as

$$\overline{M} = \prod_l \overline{M}(F_l).$$

The remaining nonprimal, mortar matching conditions of (2.3) are enforced using the reduced space $\overline{M}(F_l)$. In matrix form, this can be written as

$$(2.9) \quad B_\Delta w_\Delta + B_\Pi w_\Pi = 0.$$

The space W_Δ can be split into

$$W_\Delta = W_{\Delta,n} \oplus W_{\Delta,m},$$

where n and m denote unknowns in the interior of the nonmortar edges/faces and the remaining unknowns, respectively. The mortar matching conditions can then be written as

$$(2.10) \quad B_n w_n + B_m w_m + B_\Pi w_\Pi = 0.$$

Since these equations are obtained using only the nonredundant Lagrange multiplier space \overline{M} , the matrix B_n is invertible.

After a symmetric permutation, we can write the local Schur complement and the local Schur complement vector as

$$S^{(i)} = \begin{pmatrix} S_{\Delta\Delta}^{(i)} & S_{\Delta\Pi}^{(i)} \\ S_{\Pi\Delta}^{(i)} & S_{\Pi\Pi}^{(i)} \end{pmatrix}, \quad g^{(i)} = \begin{pmatrix} g_\Delta^{(i)} \\ g_\Pi^{(i)} \end{pmatrix},$$

and define a partially subassembled matrix and two vectors by

$$(2.11) \quad \tilde{S} = \begin{pmatrix} S_{\Delta\Delta} & S_{\Delta\Pi} \\ S_{\Pi\Delta} & S_{\Pi\Pi} \end{pmatrix}, \quad g_\Delta = \begin{pmatrix} g_\Delta^{(1)} \\ \vdots \\ g_\Delta^{(N)} \end{pmatrix}, \quad g_\Pi = \sum_{i=1}^N R_\Pi^{(i)t} g_\Pi^{(i)},$$

where

$$(2.12) \quad \begin{aligned} S_{\Delta\Delta} &= \text{diag}_{i=1}^N \left(S_{\Delta\Delta}^{(i)} \right), \\ S_{\Pi\Delta} &= \left(R_\Pi^{(1)t} S_{\Pi\Delta}^{(1)} \quad \cdots \quad R_\Pi^{(N)t} S_{\Pi\Delta}^{(N)} \right), \quad S_{\Delta\Pi} = S_{\Pi\Delta}^t, \\ S_{\Pi\Pi} &= \sum_{i=1}^N R_\Pi^{(i)t} S_{\Pi\Pi}^{(i)} R_\Pi^{(i)}. \end{aligned}$$

Here $R_\Pi^{(i)}$ is the restriction of the global primal unknowns to the subdomain primal unknowns. The matrix \tilde{S} is central to the description of our BDDC algorithm.

3. A BDDC method for the mortar discretizations. In this section, we will define a BDDC operator for the discrete elliptic problem described in section 2.1. We consider the same finite element space and subdomain partition as in section 2.1 and, as in section 2.2, we will work with the unknowns obtained after the change of variables.

Since the matrix B_n of (2.10) is invertible, we can solve for w_n ,

$$w_n = -B_n^{-1}(B_m w_m + B_\Pi w_\Pi).$$

We next define the matrix

$$(3.1) \quad R_\Gamma = \begin{pmatrix} -B_n^{-1}B_m & -B_n^{-1}B_\Pi \\ I & 0 \\ 0 & I \end{pmatrix},$$

which maps $(w_m^t, w_\Pi^t)^t$ into a vector $(w_n^t, w_m^t, w_\Pi^t)^t$ that satisfies the mortar matching condition (2.10). The mortar finite element space of section 2.1 can then be characterized as

$$\widehat{W} = \left\{ w \in \widetilde{W} : (w_n, w_m, w_\Pi) \text{ satisfies (2.10)} \right\}.$$

In the BDDC method, we work with the following discrete problem:

$$(3.2) \quad R_\Gamma^t \widetilde{S} R_\Gamma \begin{pmatrix} w_m \\ w_\Pi \end{pmatrix} = R_\Gamma^t \begin{pmatrix} g_m \\ g_\Pi \end{pmatrix},$$

where g_m is the component of the vector g_Δ in (2.11) not related to the nonmortar part.

Let us now define, with R_Γ given by (3.1),

$$(3.3) \quad R_{D,\Gamma} = D R_\Gamma = \begin{pmatrix} D_{nn} & & \\ & D_{mm} & \\ & & D_{\Pi\Pi} \end{pmatrix} R_\Gamma,$$

where the scaling matrices are selected to be

$$(3.4) \quad D_{nn} = 0, \quad D_{mm} = I, \quad D_{\Pi\Pi} = I.$$

We now propose the following preconditioner:

$$(3.5) \quad M^{-1} = R_{D,\Gamma}^t \widetilde{S}^{-1} R_{D,\Gamma}$$

for problem (3.2). Using the block Cholesky decomposition of \widetilde{S} as in Li and Widlund [23], we have

$$\widetilde{S}^{-1} = \begin{pmatrix} S_{\Delta\Delta}^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \Psi^t F_{\Pi\Pi}^{-1} \Psi,$$

where

$$F_{\Pi\Pi} = \sum_{i=1}^N \left(R_\Pi^{(i)} \right)^t \left(S_{\Pi\Pi}^{(i)} - S_{\Pi\Delta}^{(i)} S_{\Delta\Delta}^{(i)-1} S_{\Delta\Pi}^{(i)} \right) R_\Pi^{(i)},$$

$$\Psi^t = R_\Pi^t - \sum_{i=1}^N \left(R_\Delta^{(i)} \right)^t \left(S_{\Delta\Delta}^{(i)} \right)^{-1} S_{\Delta\Pi}^{(i)} R_\Pi^{(i)}.$$

Here $R_{\Pi}^{(i)} : \widehat{W}_{\Pi} \rightarrow W_{\Pi}^{(i)}$ is the restriction of the global primal variables to those of the subdomain Ω_i , and $R_{\Pi}^t : \widehat{W}_{\Pi} \rightarrow W_{\Delta} \oplus \widehat{W}_{\Pi}$ and $(R_{\Delta}^{(i)})^t : W_{\Delta}^{(i)} \rightarrow W_{\Delta} \oplus \widehat{W}_{\Pi}$ provide extensions by zero. The columns of the matrix Ψ are coarse basis functions of minimal energy with the value 1 at one of the primal unknowns and vanishing at the other primal unknowns; see [5].

The BDDC operator of the problem, given in (3.2), with the preconditioner M^{-1} , given in (3.5), is then given by

$$(3.6) \quad B_{DDC} = R_{D,\Gamma}^t \widetilde{S}^{-1} R_{D,\Gamma} R_{\Gamma}^t \widetilde{S} R_{\Gamma}.$$

4. Condition number analysis using a bound on E_D . In this section, we will estimate the condition number of the BDDC operator by using the approach introduced in [22]. A bound for the average operator E_D in the \widetilde{S} -norm is central in the analysis; see below. For definitions of R_{Γ} and $R_{D,\Gamma}$, see (3.1) and (3.3), respectively. The operator E_D is defined by

$$(4.1) \quad E_D = R_{\Gamma} R_{D,\Gamma}^t.$$

In the following, we will show that the weight matrix D has been chosen so that

$$(P1) \quad R_{\Gamma}^t R_{D,\Gamma} = R_{D,\Gamma}^t R_{\Gamma} = I,$$

$$(P2) \quad |E_D w|_{\widetilde{S}}^2 \leq C \max_i \left\{ \left(1 + \log \frac{H_i}{h_i} \right)^2 \right\} |w|_{\widetilde{S}}^2.$$

Here $|w|_{\widetilde{S}}^2 = \langle \widetilde{S} w, w \rangle$. We then consider

$$R_{\Gamma}^t R_{D,\Gamma} \begin{pmatrix} w_m \\ w_{\Pi} \end{pmatrix} = \begin{pmatrix} -B_m^t (B_n^t)^{-1} D_{nm} z_n + D_{mm} w_m \\ -B_{\Pi}^t (B_n^t)^{-1} D_{n\Pi} z_n + D_{\Pi\Pi} w_{\Pi} \end{pmatrix},$$

where

$$z_n = -B_n^{-1} (B_m w_m + B_{\Pi} w_{\Pi}).$$

We recall the scaling factors of the weight matrix D in (3.4) and we can easily see that these weights give the (P1) property.

Remark 4.1. The weights above lead to an operator E_D of the form

$$E_D \begin{pmatrix} w_n \\ w_m \\ w_{\Pi} \end{pmatrix} = \begin{pmatrix} -B_n^{-1} (B_m w_m + B_{\Pi} w_{\Pi}) \\ w_m \\ w_{\Pi} \end{pmatrix}.$$

In contrast to the case of conforming finite elements, this does not involve any averaging across the interface. We will still call E_D the average operator, borrowing the name from the conforming case.

We will now show that the average operator E_D satisfies the (P2) property for the weight matrix D just given. As a preparation, we need to establish an estimate for the mortar projection of a function w in \widetilde{W} in the $H_{00}^{1/2}(F)$ -norm. For an edge/face $F \subset \partial\Omega_i$, the space $H_{00}^{1/2}(F)$ consists of the functions for which the zero extension to the whole boundary $\partial\Omega_i$ belongs to the Sobolev space $H^{1/2}(\partial\Omega_i)$. It is equipped with the norm

$$\|w\|_{H_{00}^{1/2}(F)}^2 = |w|_{H^{1/2}(F)}^2 + \int_F \frac{|w(x)|^2}{\text{dist}(x, \partial F)} ds(x).$$

This norm has the well-known property that

$$(4.2) \quad c|\tilde{w}|_{H^{1/2}(\partial\Omega_i)} \leq \|w\|_{H_0^{1/2}(F)} \leq C|\tilde{w}|_{H^{1/2}(\partial\Omega_i)},$$

where \tilde{w} is the zero extension of w to $\partial\Omega_i \setminus F$; see [10, Lemma 1.3.2.6].

We recall that a nonmortar edge/face F of $\partial\Omega_i$ is a union of mortar interfaces F_{ij} common to $\partial\Omega_i$ and $\partial\Omega_j$. We recall that ϕ is a function defined on F with $\phi = w_j$ on each $F_{ij} \subset F$, and with $w_j \in W_j$, the finite element space provided for $\partial\Omega_j$. We then have $\phi \in H^{1/2-\epsilon}(F)$ for any $\epsilon > 0$. Because of the slightly weaker regularity of the function ϕ , caused by the geometrically nonconforming partition, we have some difficulty obtaining the condition number bound with only two logarithmic factors for geometrically nonconforming partitions in three dimensions. We will overcome this difficulty by using an additional finite element space for the interface F_{ij} and an L^2 -projection onto this space. This will result in a condition number bound with two logarithmic factors under some assumptions on the geometry of the subdomain partition; see Assumption 4.3 below.

We also need the following assumption on the coefficients of the elliptic problem. We note that this assumption basically reflects a weakness of the mortar methods in the case of geometrically nonconforming partitions.

ASSUMPTION 4.2. *The coefficients satisfy*

$$\rho_i \leq C\rho_j,$$

where Ω_i and Ω_j correspond to the nonmortar and mortar side of the common set $F_{ij} = \partial\Omega_i \cap \partial\Omega_j$, respectively.

We also will use the following assumption.

ASSUMPTION 4.3. *A geometrically nonconforming partition $\{\Omega_i\}_i$ in three dimensions satisfies the following three assumptions.*

1. *The subdomains are polytopes.*
2. *A quasi-uniform triangulation, with a mesh size comparable to h_i , is possible for the interface F_{ij} .*
3. *Any subdomain has a diameter comparable to those of its neighbors.*

We recall that the finite element space $\mathring{W}(F)$, given in (2.2), and a Lagrange multiplier space $M(F)$ are provided for the nonmortar edge/face F . We now define the mortar projection.

DEFINITION 4.4. *The mortar projection $\pi_F : L^2(F) \rightarrow \mathring{W}(F)$ of the nonmortar edge/face F is defined by*

$$\int_F (v - \pi_F(v))\lambda \, ds = 0 \quad \forall \lambda \in M(F).$$

This mortar projection has been shown to be stable in the L^2 - and $H_0^{1/2}$ -norms in [3, 2, 32].

LEMMA 4.5. *Under Assumptions 4.2 and 4.3 and with $w = (w_1, \dots, w_N) \in \widetilde{W}$, we have*

$$\rho_i \|\pi_F(\phi - w_i)\|_{H_0^{1/2}(F)}^2 \leq C \left(1 + \log \frac{H_i}{h_i}\right)^2 \sum_{k \in I(F)} \langle S^{(k)} w_k, w_k \rangle.$$

Here $F \subset \partial\Omega_i$ is an edge/face, $\phi = w_j$ on $F_{ij} \subset F$, and $I(F)$ is the set of indices of the subdomains with boundaries that intersect F .

Proof. We will prove the result for a geometrically nonconforming partition in three dimensions under Assumption 4.3. In the case of a geometrically conforming partition in three dimensions and for any partition in two dimensions, the same result can be obtained straightforwardly without any assumption on the partition.

For each interface F_{ij} , we define a characteristic function $\chi_{ij} \in L^2(F)$ with the value 1 on F_{ij} and the value 0 on $F \setminus F_{ij}$. In addition, we introduce a quasi-uniform finite element space $U(F_{ij})$ on the interface F_{ij} with a mesh size comparable to h_i , that of the finite element space W_i of the subdomain Ω_i of the nonmortar side. The L^2 -projection onto $U(F_{ij})$ is denoted by Q_{ij} and it satisfies the following properties (see [4, Chapter II]: $\forall w \in H^{1/2}(F_{ij})$):

$$(4.3) \quad \|w - Q_{ij}w\|_{L^2(F_{ij})}^2 \leq Ch_i |w|_{H^{1/2}(F_{ij})}^2, \quad \|Q_{ij}w\|_{H^{1/2}(F_{ij})}^2 \leq C \|w\|_{H^{1/2}(F_{ij})}^2,$$

where the L^2 -term in the $H^{1/2}$ -norm is scaled by $1/|F_{ij}|$. Here $|F_{ij}|$ is the diameter of F_{ij} .

Then, on F , consider

$$\begin{aligned} w_i - \phi &= \sum_j \chi_{ij}(w_i - w_j) \\ &= \sum_j \chi_{ij}((w_i - c_{ij}) - (w_j - c_{ij})). \end{aligned}$$

Here c_{ij} denotes the common average value of w_i and w_j defined by

$$c_{ij} = \frac{\int_{F_{ij}} w_i \lambda_{ij} ds}{\int_{F_{ij}} \lambda_{ij} ds} = \frac{\int_{F_{ij}} w_j \lambda_{ij} ds}{\int_{F_{ij}} \lambda_{ij} ds},$$

where λ_{ij} are defined in (2.4); c_{ij} is closely related to the primal mortar matching condition (2.5).

It suffices to show that

$$(4.4) \quad \|\pi_F(\chi_{ij}(w_j - c_{ij}))\|_{H_0^{1/2}(F)}^2 \leq C \left(1 + \log \frac{H_i}{h_i}\right)^2 |w_j|_{H^{1/2}(\partial\Omega_j)}^2,$$

and to give a similar estimate for $w_i - c_{ij}$. We will prove (4.4) but leave out the estimate for $w_i - c_{ij}$, which is quite similar. The required estimate then follows from Assumption 4.2 and the fact that $|w_j|_{H^{1/2}(\partial\Omega_j)}^2$ is spectrally equivalent to $(1/\rho_j)\langle S^{(j)}w_j, w_j \rangle$.

Let

$$z = w_j - c_{ij}.$$

We decompose $Q_{ij}(z)$ into

$$(4.5) \quad Q_{ij}(z) = I_{F_{ij}}(Q_{ij}(z)) + I_{\partial F_{ij}}(Q_{ij}(z)),$$

where the first term is equal to $Q_{ij}(z)$ at all interior nodal points of F_{ij} and vanishes on ∂F_{ij} while the second term is equal to $Q_{ij}(z)$ at the nodal points of ∂F_{ij} and vanishes at the remaining nodal points of F_{ij} . We have

$$(4.6) \quad \begin{aligned} \|\pi_F(\chi_{ij}(w_j - c_{ij}))\|_{H_0^{1/2}(F)}^2 &= \|\pi_F(\chi_{ij}z)\|_{H_0^{1/2}(F)}^2 \\ &\leq 2\|\pi_F(\chi_{ij}(z - Q_{ij}(z)))\|_{H_0^{1/2}(F)}^2 \\ &\quad + 2\|\pi_F(\chi_{ij}Q_{ij}(z))\|_{H_0^{1/2}(F)}^2. \end{aligned}$$

The first term above is estimated by

$$\begin{aligned}
\|\pi_F(\chi_{ij}(z - Q_{ij}(z)))\|_{H_{00}^{1/2}(F)}^2 &\leq Ch_i^{-1} \|\chi_{ij}(z - Q_{ij}(z))\|_{L^2(F)}^2 \\
&= Ch_i^{-1} \|z - Q_{ij}(z)\|_{L^2(F_{ij})}^2 \\
&\leq C|z|_{H^{1/2}(F_{ij})}^2 \\
(4.7) \qquad \qquad \qquad &\leq C|w_j|_{H^{1/2}(\partial\Omega_j)}.
\end{aligned}$$

We have used an inverse inequality, the L^2 -stability of π_F , and the properties of $Q_{ij}(z)$ given in (4.3).

There remains for us to estimate the second term of (4.6). By Assumption 4.3, the subdomain interfaces F_{ij} are polygonal regions. For a geometrically nonconforming partition, the area of the interface F_{ij} might be comparable to that of F_j , the face of Ω_j such that $F_j \cap \partial\Omega_i = F_{ij}$. In the other case, when F_{ij} is only a small part of F_j , it could be a narrow strip, e.g., $[0, H] \times [0, \delta]$, or a rectangular region with its area comparable to $[0, \delta] \times [0, \delta]$, where δ is comparable to the mesh size h .

We will first consider the second term in (4.6) when the area of the interface F_{ij} is comparable to that of F_j . Using (4.5), we have

$$\begin{aligned}
\|\pi_F(\chi_{ij}Q_{ij}(z))\|_{H_{00}^{1/2}(F)}^2 &= \|\pi_F(\chi_{ij}(I_{F_{ij}}Q_{ij}(z) + I_{\partial F_{ij}}Q_{ij}(z)))\|_{H_{00}^{1/2}(F)}^2 \\
&\leq C \left(\|\tilde{I}_{F_{ij}}(Q_{ij}(z))\|_{H_{00}^{1/2}(F)}^2 + h_i^{-1} \|\tilde{I}_{\partial F_{ij}}Q_{ij}(z)\|_{L^2(F)}^2 \right) \\
(4.8) \qquad \qquad \qquad &\leq C \left(\|I_{F_{ij}}(Q_{ij}(z))\|_{H_{00}^{1/2}(F_{ij})}^2 + \|I_{\partial F_{ij}}Q_{ij}(z)\|_{L^2(\partial F_{ij})}^2 \right),
\end{aligned}$$

where $\tilde{I}_{F_{ij}}(v)$ and $\tilde{I}_{\partial F_{ij}}(v)$ are the extensions of $I_{F_{ij}}(v)$ and $I_{\partial F_{ij}}(v)$ by zero, respectively. Here, we have used an inverse inequality, the stability of π_F in the L^2 - and $H_{00}^{1/2}$ -norms, and the following inequalities:

$$\begin{aligned}
\|\tilde{I}_{F_{ij}}(Q_{ij}(z))\|_{H_{00}^{1/2}(F)} &\leq \|I_{F_{ij}}(Q_{ij}(z))\|_{H_{00}^{1/2}(F_{ij})}, \\
\|\tilde{I}_{\partial F_{ij}}Q_{ij}(z)\|_{L^2(F)}^2 &\leq Ch_i \|I_{\partial F_{ij}}Q_{ij}(z)\|_{L^2(\partial F_{ij})}^2.
\end{aligned}$$

By applying Lemmas 4.17, 4.19, and 4.24 of [28] to the terms of (4.8), and using (4.3) and the Poincaré inequality, we obtain

$$(4.9) \qquad \|\pi_F(\chi_{ij}Q_{ij}(z))\|_{H_{00}^{1/2}(F)}^2 \leq C \left(1 + \log \frac{H_{ij}}{h_i} \right)^2 |w_j|_{H^{1/2}(\partial\Omega_j)}^2,$$

where H_{ij} is the diameter of F_{ij} , which satisfies $H_{ij} \leq H_i$.

We now consider the second term in (4.6) for the case when F_{ij} is only a small part of F_j . Then,

$$\begin{aligned}
\|\pi_F(\chi_{ij}Q_{ij}(z))\|_{H_{00}^{1/2}(F)}^2 &\leq Ch_i^{-1} \|\pi_F(\chi_{ij}Q_{ij}(z))\|_{L^2(F)}^2 \\
&\leq Ch_i^{-1} \|z\|_{L^2(F_{ij})}^2 = Ch_i^{-1} \|w_j - c_{ij}\|_{L^2(F_{ij})}^2 \\
&\leq Ch_i^{-1} \|w_j\|_{L^2(F_{ij})}^2 \\
&\leq Ch_i^{-1} \delta (1 + \log(H_j/\delta)) \|w_j\|_{H^{1/2}(F_j)}^2 \\
(4.10) \qquad \qquad \qquad &\leq C (1 + \log(H_i/h_i)) |w_j|_{H^{1/2}(\partial\Omega_j)}^2.
\end{aligned}$$

Here we have used an inverse inequality, the stability of π_F and Q in the L^2 -norm, the inequality

$$\|c_{ij}\|_{L^2(F_{ij})}^2 \leq C \|w_j\|_{L^2(F_{ij})}^2,$$

Lemma 3.4 of Dryja and Widlund [7] for the fourth inequality, the Poincaré inequality in the last inequality, and that δ is comparable to the mesh size h_i . We note that we have only one log factor in this case.

Therefore, (4.6) combined with (4.7) and (4.9) or (4.10) proves the desired bound (4.4). \square

With the help of Lemma 4.5, we can establish property (P2) for the operator E_D .

LEMMA 4.6. *With Assumptions 4.2 and 4.3, the operator E_D satisfies*

$$|E_D w|_{\tilde{S}}^2 \leq C \max_i \left\{ \left(1 + \log \frac{H_i}{h_i} \right)^2 \right\} |w|_{\tilde{S}}^2 \quad \text{for any } w \in \widetilde{W},$$

where \tilde{S} is defined in (2.11).

Proof. Using the weight matrix D of (3.4), the average operator E_D , given by (4.1), satisfies

$$E_D \begin{pmatrix} w_n \\ w_m \\ w_\Pi \end{pmatrix} = \begin{pmatrix} w_n - B_n^{-1}(B_n w_n + B_m w_m + B_\Pi w_\Pi) \\ w_m \\ w_\Pi \end{pmatrix},$$

as in Remark 4.1. Here $w = (w_n, w_m, w_\Pi) \in \widetilde{W}$. Let

$$\widehat{w}_n = w_n - B_n^{-1}(B_n w_n + B_m w_m + B_\Pi w_\Pi),$$

and construct \widehat{w}_i by restricting the unknowns $(\widehat{w}_n, w_m, w_\Pi)$ to the subdomain Ω_i . Similarly, we construct w_i from (w_n, w_m, w_Π) . We note that (w_1, \dots, w_N) satisfies the primal constraints on the edges/faces. By definition, $\widehat{w} = (\widehat{w}_1, \dots, \widehat{w}_N) \in \widetilde{W}$; i.e., \widehat{w} satisfies all of the mortar matching conditions, and each \widehat{w}_i is of the form

$$\widehat{w}_i = w_i - \sum_{F \subset \partial\Omega_i} \widetilde{\pi}_F(w_i - \phi),$$

where F is a nonmortar edge/face of $\partial\Omega_i$, $\widetilde{\pi}_F(w_i - \phi)$ is the zero extension of $\pi_F(w_i - \phi)$ to all of $\partial\Omega_i \setminus F$, and $\phi = w_j$ on $F_{ij} := \partial\Omega_j \cap \partial\Omega_i \subset F$. We then obtain

$$\begin{aligned} |E_D w|_{\tilde{S}}^2 &= \sum_{i=1}^N \langle S^{(i)} \widehat{w}_i, \widehat{w}_i \rangle \\ &\leq C \sum_{i=1}^N \left(\langle S^{(i)} w_i, w_i \rangle + \sum_{F \subset \partial\Omega_i} \langle S^{(i)} \widetilde{\pi}_F(\phi - w_i), \widetilde{\pi}_F(\phi - w_i) \rangle \right) \\ &\leq C \left(\sum_{i=1}^N \langle S^{(i)} w_i, w_i \rangle + \sum_{i=1}^N \sum_{F \subset \partial\Omega_i} \rho_i \|\pi_F(\phi - w_i)\|_{H_{00}^{1/2}(F)}^2 \right) \\ &\leq C \max_i \left\{ \left(1 + \log \frac{H_i}{h_i} \right)^2 \right\} \sum_{i=1}^N \langle S^{(i)} w_i, w_i \rangle \\ &= C \max_i \left\{ \left(1 + \log \frac{H_i}{h_i} \right)^2 \right\} \langle \tilde{S} w, w \rangle. \end{aligned}$$

Here we have used that $\langle S^{(i)} w_i, w_i \rangle \simeq \rho_i |w_i|_{H^{1/2}(\partial\Omega_i)}^2$, the bounds in (4.2), and Lemma 4.5. \square

By using the properties (P1) and (P2), we can show the following condition number bound for the BDDC operator (3.6). A proof for a quite similar case is given in Li and Widlund [22] in their analysis of a BDDC method for the Stokes problem with conforming meshes. We do not include a proof, which would be almost identical to that of [22].

THEOREM 4.7. *With Assumptions 4.2 and 4.3, we have the condition number bound*

$$\kappa(B_{DDC}) \leq C \max_i \left\{ \left(1 + \log \frac{H_i}{h_i} \right)^2 \right\}.$$

Remark 4.8. For a geometrically nonconforming partition, the number of primal constraints tends to be larger than for a conforming partition if only edge/face constraints are used. We note that there are several previous studies which explore the possibility of selecting primal constraints for only some of the edges/faces; see [15, 21, 19].

5. A connection between the FETI-DP and BDDC methods. In this section, we will show that the BDDC method developed in the previous sections is closely connected to the FETI-DP method developed by the first author in [14, 15] and jointly with Lee in [12]. We will show that the two methods share the same spectra except possibly for an eigenvalue equal to 1.

As previously noted, a comparison of the spectra of the BDDC method to that of the FETI-DP method was made by Mandel, Dohrmann, and Tezaur [25] for conforming finite elements. They showed that the two algorithms have the same set of eigenvalues except possibly for eigenvalues equal to 1. A simpler proof of this fact was given more recently by Li and Widlund [23]. They formulated the BDDC operators, as well as the FETI-DP operators, using a change of variables and introducing certain projections and average operators. These projections and average operators provide an important connection between the FETI-DP and the BDDC operators.

We now formulate an FETI-DP operator after the same change of variables as in section 2.2. We then show that the FETI-DP operator has essentially the same spectrum as the BDDC operator by establishing several properties of the projections and average operators that were used by Li and Widlund [23].

After the change of variables, the linear system considered in the FETI-DP formulation is given by

$$(5.1) \quad \begin{pmatrix} S_{\Delta\Delta} & S_{\Delta\Pi} & B_{\Delta}^t \\ S_{\Pi\Delta} & S_{\Pi\Pi} & B_{\Pi}^t \\ B_{\Delta} & B_{\Pi} & 0 \end{pmatrix} \begin{pmatrix} u_{\Delta} \\ u_{\Pi} \\ \lambda \end{pmatrix} = \begin{pmatrix} g_{\Delta} \\ g_{\Pi} \\ 0 \end{pmatrix},$$

where the matrices $S_{\Delta\Delta}$, $S_{\Delta\Pi}$, $S_{\Pi\Delta}$, and $S_{\Pi\Pi}$ are defined in (2.12) and the matrices B_{Δ} and B_{Π} are obtained from the mortar matching condition (2.9). We recall that the subscripts Π and Δ stand for the unknowns or submatrices related to the primal variables and the remaining part, respectively, and that $\lambda \in \overline{M}$, the reduced, nonredundant Lagrange multiplier space.

After eliminating the unknowns u_{Δ} and u_{Π} , we obtain an equation for $\lambda \in \overline{M}$:

$$(5.2) \quad B_{\Gamma} \tilde{S}^{-1} B_{\Gamma}^t \lambda = d,$$

where

$$(5.3) \quad B_\Gamma = (B_\Delta \quad B_\Pi), \quad \tilde{S} = \begin{pmatrix} S_{\Delta\Delta} & S_{\Delta\Pi} \\ S_{\Pi\Delta} & S_{\Pi\Pi} \end{pmatrix},$$

and d is also the result of the Gaussian elimination.

We will now express the Neumann–Dirichlet preconditioner considered in [14, 15, 12] using the new unknowns. The Neumann–Dirichlet preconditioner M_{DP}^{-1} is defined by

$$(5.4) \quad \langle M_{DP}\lambda, \lambda \rangle = \max_{w_{\Delta,n} \in W_{\Delta,n}} \frac{\langle B_\Gamma \mathcal{E}(w_{\Delta,n}), \lambda \rangle^2}{\langle \tilde{S} \mathcal{E}(w_{\Delta,n}), \mathcal{E}(w_{\Delta,n}) \rangle},$$

where $\mathcal{E}(w_{\Delta,n})$ is the extension by zero of $w_{\Delta,n} \in W_{\Delta,n}$ to elements in the space $\widetilde{W} = W_{\Delta,n} \oplus W_{\Delta,m} \oplus \widehat{W}_\Pi$.

We recall that the matrix B_Δ is partitioned into

$$B_\Delta = (B_n \quad B_m),$$

where n denotes the columns of the nonmortar unknowns and m those that remain. The formula (5.4) can then be written as

$$(5.5) \quad \langle M_{DP}\lambda, \lambda \rangle = \max_{w_{\Delta,n} \in W_{\Delta,n}} \frac{\langle B_n w_{\Delta,n}, \lambda \rangle^2}{\langle S_{nn} w_{\Delta,n}, w_{\Delta,n} \rangle},$$

where S_{nn} is the submatrix of $S_{\Delta\Delta}$ in (5.1) corresponding to the nonmortar part. We see that $S_{nn} : W_{\Delta,n} \rightarrow W'_{\Delta,n}$ and $B_n^t : \overline{M} \rightarrow W'_{\Delta,n}$ are invertible. Here $W'_{\Delta,n}$ is the space dual to $W_{\Delta,n}$. The maximum in (5.5) occurs when $S_{nn} w_{\Delta,n} = B_n^t \lambda$, and hence it follows that

$$M_{DP}^{-1} = (B_n^t)^{-1} S_{nn} B_n^{-1}.$$

Furthermore, this matrix can be written as

$$(5.6) \quad M_{DP}^{-1} = B_{\Sigma,\Gamma}^{-1} \tilde{S} B_{\Sigma,\Gamma}^t,$$

where

$$B_{\Sigma,\Gamma}^t = \begin{pmatrix} \Sigma_{nn} & & \\ & \Sigma_{mm} & \\ & & \Sigma_{\Pi\Pi} \end{pmatrix} \begin{pmatrix} B_n^t \\ B_m^t \\ B_\Pi^t \end{pmatrix}$$

with the weights given by

$$\Sigma_{nn} = (B_n^t B_n)^{-1}, \quad \Sigma_{mm} = 0, \quad \Sigma_{\Pi\Pi} = 0.$$

Therefore, the FETI-DP operator with the Neumann–Dirichlet preconditioner M_{DP}^{-1} is given by

$$M_{DP}^{-1} F_{DP} = B_{\Sigma,\Gamma}^{-1} \tilde{S} B_{\Sigma,\Gamma}^t B_\Gamma \tilde{S}^{-1} B_\Gamma^t,$$

while the preconditioned BDDC operator is given by

$$B_{DDC} = R_{D,\Gamma}^t \tilde{S}^{-1} R_{D,\Gamma} R_\Gamma^t \tilde{S} R_\Gamma.$$

Let us now define the following jump and average operators:

$$P_\Sigma = B_{\Sigma,\Gamma}^t B_\Gamma, \quad E_D = R_\Gamma R_{D,\Gamma}^t.$$

The following results are provided in [23, section 5].

THEOREM 5.1. *Assume that P_Σ and E_D satisfy*

1. $E_D + P_\Sigma = I$,
2. $E_D^2 = E_D$, $P_\Sigma^2 = P_\Sigma$, and
3. $E_D P_\Sigma = P_\Sigma E_D = 0$.

Then the operators $M_{DP}^{-1} F_{DP}$ and B_{DDC} have the same eigenvalues except possibly for an eigenvalue equal to 1.

We will now show that the assumptions of Theorem 5.1 hold for the operators P_Σ and E_D . We recall the definition of the space of functions satisfying the primal constraints

$$\widetilde{W} = \left\{ (w_n^t, w_m^t, w_\Pi^t)^t : \forall w_n \in W_{\Delta,n}, w_m \in W_{\Delta,m}, w_\Pi \in \widehat{W}_\Pi \right\},$$

and the mortar finite element space

$$\widehat{W} = \{w \in \widetilde{W} : B_m w_m + B_\Pi w_\Pi + B_n w_n = 0\}.$$

We note that P_Σ and E_D are operators defined on the space \widetilde{W} .

LEMMA 5.2. *The operators P_Σ and E_D satisfy the assumptions of Theorem 5.1.*

Proof. From

$$\begin{aligned} \Sigma_{mm} &= 0, & \Sigma_{\Pi\Pi} &= 0, & \Sigma_{nn} &= (B_n^t B_n)^{-1}, \\ D_{mm} &= I, & D_{\Pi\Pi} &= I, & D_{nn} &= 0, \end{aligned}$$

we have

$$\begin{aligned} P_\Sigma w &= \begin{pmatrix} B_n^{-1}(B_m w_m + B_\Pi w_\Pi + B_n w_n) \\ 0 \\ 0 \end{pmatrix}, \\ E_D w &= \begin{pmatrix} -B_n^{-1}(B_m w_m + B_\Pi w_\Pi) \\ w_m \\ w_\Pi \end{pmatrix}. \end{aligned}$$

Hence,

$$(5.7) \quad E_D + P_\Sigma = I.$$

From $E_D w = w$ and $P_\Sigma w = 0$ for all $w \in \widehat{W}$, and from $\text{Range}(E_D) \subset \widehat{W}$, we obtain

$$(5.8) \quad E_D^2 = E_D, \quad P_\Sigma E_D = 0.$$

From (5.7), we have the identities

$$E_D(E_D + P_\Sigma) = E_D, \quad P_\Sigma(E_D + P_\Sigma) = P_\Sigma,$$

and combining them with (5.8), we obtain

$$E_D P_\Sigma = 0, \quad P_\Sigma^2 = P_\Sigma. \quad \square$$

Remark 5.3. Other FETI-DP preconditioners in two dimensions with different weights

$$\Sigma = \begin{pmatrix} \Sigma_{nn} & & \\ & \Sigma_{mm} & \\ & & \Sigma_{\Pi\Pi} \end{pmatrix},$$

with nonzero weights Σ_{mm} and Σ_{III} , have been developed and shown to give condition number bounds of the form

$$C \max_i \{ (1 + \log(H_i/h_i))^2 \}$$

for geometrically conforming partitions; see [8, 6]. We have not found any weight matrix D that results in $E_D + P_\Sigma = I$ for such a choice of Σ .

6. Numerical results. In this section, we present numerical results. We first compare the BDDC and the FETI-DP methods with the suggested preconditioners, for geometrically conforming cases, and we then illustrate the performance of our BDDC methods for some geometrically nonconforming partitions. We solve an elliptic problem with the exact solution $u(x, y) = \sin(\pi x)(1 - y)y$,

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega, \\ u &= 0 \text{ on } \partial\Omega, \end{aligned}$$

where Ω is the unit square in \mathbb{R}^2 . The conjugate gradient iteration is halted when the ℓ_2 -norm of the relative residual has been reduced by a factor of 10^6 .

In the first series of experiments, the domain Ω is divided into uniform square subdomains, as in Figure 2, that are geometrically conforming. Common values at the subdomain vertices are selected as the primal constraints for this case. Each subdomain has either a nonuniform mesh or a uniform mesh with n nodes on each subdomain edge. The meshes do not match and have comparable mesh sizes across the interface as in Figure 2.

In Table 1, we show the performance of the two algorithms when Ω is partitioned into $N = 4 \times 4$ subdomains (see Figure 2) and with the local problem size n increasing. In this case, the upper and the right edges of each subdomain are selected to be nonmortar edges; see Figure 2. We provide the L^2 - and H^1 -errors between the exact solution and the solution of the iterative method, the number of conjugate gradient iterations, and the minimum and the maximum eigenvalues of the BDDC and the FETI-DP methods. For the H^1 -error, we use the broken H^1 -norm given by the subdomain partition. Table 2 shows the numerical results when we fix the local problem size to $n - 1 = 4$ and increase N , the number of subdomains to $N = 8 \times 8$, 16×16 , and 32×32 , and divide Ω into square subdomains in the same manner as for $N = 4 \times 4$. We observe that the two methods give the same L^2 - and H^1 -errors. The minimum eigenvalue of the BDDC operator is always equal to 1 while that of the

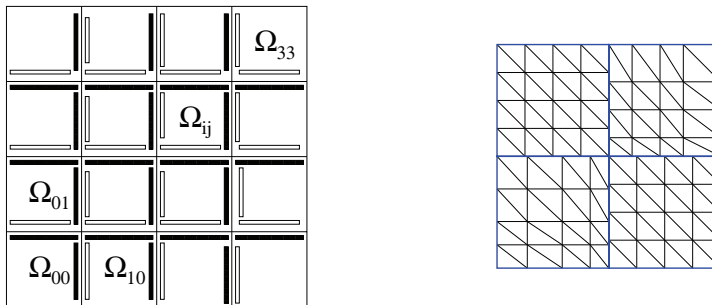


FIG. 2. A subdomain partition (left: white edges are mortar and black edges are nonmortar) with $N = 4 \times 4$ and nonmatching comparable meshes with the local problem size $n - 1 = 4$.

TABLE 1

Comparison of FETI-DP and BDDC methods where n , the local problem size, increases with a fixed subdomain partition ($N = 4 \times 4$).

$n - 1$	$\ u - u^h\ _0$	$\ u - u^h\ _1$	$M_{DP}^{-1}F_{DP}$			B_{DDC}		
			Iter	λ_{\min}	λ_{\max}	Iter	λ_{\min}	λ_{\max}
4	5.0850e-4	6.0126e-2	10	1.40	4.09	12	1.00	4.09
8	1.2865e-4	3.0128e-2	13	1.01	5.72	15	1.00	5.72
16	3.2231e-5	1.5072e-2	15	1.00	7.72	16	1.00	7.72
32	8.0621e-6	7.5374e-3	16	1.01	1.00e+1	17	1.00	1.00e+1
64	2.0134e-6	3.7688e-3	17	1.01	1.28e+1	19	1.00	1.28e+1

TABLE 2

Comparison of FETI-DP and BDDC methods when N , the number of subdomains, increases with a fixed local problem size ($n - 1 = 4$).

N	$\ u - u^h\ _0$	$\ u - u^h\ _1$	$M_{DP}^{-1}F_{DP}$			B_{DDC}		
			Iter	λ_{\min}	λ_{\max}	Iter	λ_{\min}	λ_{\max}
4×4	5.0850e-4	6.0126e-2	10	1.40	4.09	12	1.00	4.09
8×8	1.1744e-4	2.9900e-2	11	1.37	4.41	12	1.00	4.41
16×16	2.9743e-5	1.4980e-2	12	1.32	4.49	13	1.00	4.49
32×32	7.4317e-6	7.4917e-3	12	1.30	4.57	13	1.00	4.62

FETI-DP operator is greater than 1. The maximum eigenvalues of both operators are almost the same; the eigenvalues are estimated by using the parameters of the conjugate gradient iteration. We note that the minimum eigenvalue of the FETI-DP operator converges to 1 when the number of nodes increases; see Table 1. The two algorithms perform quite similarly with good scalability in terms of the local problem size and the number of subdomains.

We next illustrate the performance of the BDDC method for geometrically non-conforming partitions. We divide the unit square Ω into rectangular subdomains that are geometrically nonconforming. For a given N , we first divide Ω into N uniform vertical strips and then each strip into N or $N + 1$ rectangles, in succession; see Figure 3 for $N = 4$. Each subdomain has a uniform mesh with a number of nodes across the subdomain equal to n , $n + 2$, or $n + 4$; see Figure 3. We consider the case when the coefficient $\rho(x) = 1$ in Ω and the case when the coefficient $\rho(x)$ has jumps across the subdomain interfaces; i.e., $\rho(x) = \rho_i$, with different constants in different subdomains Ω_i . See Figure 3 for the distribution of the ρ_i with the values 1, 10, 100, and 1000 in a partition with $N = 4$, and for the selection of nonmortar and mortar edges which satisfies Assumption 4.2 with C less than 1. For the uniform case with $\rho(x) = 1$, we use the same selection of nonmortar and mortar edges. For a larger N , we copy the same pattern periodically. We run the BDDC method with increasing numbers of nodes in a fixed subdomain partition and with an increase of the number of subdomains with a fixed local problem size.

Table 3 presents the condition numbers and the number of iterations for both continuous and discontinuous $\rho(x)$. Since the subdomain partitions are geometrically nonconforming, we have chosen

$$\int_{F_{ij}} (v_i - v_j) \lambda_{ij} ds = 0$$

as the primal constraints for each face $F_{ij} = \partial\Omega_i \cap \partial\Omega_j$. Here λ_{ij} is the sum of the Lagrange multiplier basis functions that are supported in $\overline{F_{ij}}$. We observe good scalability in terms of the number of subdomains and the local problem size for both

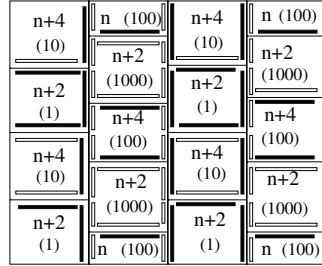


FIG. 3. A geometrically nonconforming partition with $N = 4$ and the number of nodes for each subdomain edge for a given n , and the values of ρ_i (in parentheses) in the jump coefficient case: nonmortar edges (black) and mortar edges (white) which satisfy Assumption 4.2 for the given ρ_i with C less than 1.

TABLE 3

Performance of the BDDC algorithm with an increase of N with a fixed local problem size ($n=6$) and with an increase of the local problem size, n , in a geometrically nonconforming partition with $N=4$. Cond (the condition number) and Iter (the number of iterations) are provided.

$\rho(x) = 1$						Jump coefficient ρ_i					
N	Cond	Iter	n	Cond	Iter	N	Cond	Iter	n	Cond	Iter
16	12.36	23	6	11.57	20	16	6.68	15	6	6.67	14
32	12.37	24	12	14.85	22	32	6.68	15	12	7.94	15
48	12.40	24	24	18.54	23	48	6.68	15	24	9.52	17
64	12.41	24	48	22.69	26	64	6.69	15	48	11.37	18

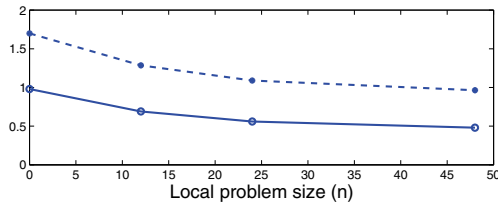


FIG. 4. Plot of the values, $\text{Cond}/(1 + \log n)^2$, with an increase of the local problem size, n , in a fixed geometrically nonconforming subdomain partition with $N = 4$; the dashed line is for the case $\rho(x) = 1$ and the solid line for the case with a jump coefficient ρ_i .

cases. In addition, the behavior of the condition number with an increase of the local problem size shows that the condition number bound $(1 + \log(H/h))^2$ appears to be optimal; see Figure 4.

REFERENCES

- [1] Y. ACHDOU, Y. MADAY, AND O. B. WIDLUND, *Iterative substructuring preconditioners for mortar element methods in two dimensions*, SIAM J. Numer. Anal., 36 (1999), pp. 551–580.
- [2] F. BEN BELGACEM AND Y. MADAY, *The mortar element method for three-dimensional finite elements*, RAIRO Modél. Math. Anal. Numér., 31 (1997), pp. 289–302.
- [3] C. BERNARDI, Y. MADAY, AND A. T. PATERA, *A new nonconforming approach to domain decomposition: The mortar element method*, in Nonlinear Partial Differential Equations and their Applications. Collège de France Seminar, Vol. XI (Paris, 1989–1991), Pitman

- Res. Notes Math. Ser. 299, Longman Sci. Tech., Harlow, UK, 1994, pp. 13–51.
- [4] D. BRAESS, *Finite Elements*, in Theory, fast solvers, and applications in solid mechanics, Translated from the 1992 German original by Larry L. Schumaker, Cambridge University Press, Cambridge, 1997.
 - [5] C. R. DOHRMANN, *A preconditioner for substructuring based on constrained energy minimization*, SIAM J. Sci. Comput., 25 (2003), pp. 246–258.
 - [6] N. DOKEVA, M. DRYJA, AND W. PROSKUROWSKI, *A FETI-DP preconditioner with special scaling for mortar discretization of elliptic problems with discontinuous coefficients*, SIAM J. Numer. Anal., 44 (2006), pp. 283–299.
 - [7] M. DRYJA AND O. B. WIDLUND, *Domain decomposition algorithms with small overlap*, SIAM J. Sci. Comput., 15 (1994), pp. 604–620.
 - [8] M. DRYJA AND O. B. WIDLUND, *A generalized FETI-DP method for a mortar discretization of elliptic problems*, in Domain Decomposition Methods in Science and Engineering, I. Herrera, D. E. Keyes, O. B. Widlund, and R. Yates, eds., National Auton. Universidad de Mexico, México, 2003, pp. 27–38.
 - [9] C. FARHAT, M. LESOINNE, AND K. PIERSON, *A scalable dual-primal domain decomposition method*, Numer. Linear Algebra Appl., 7 (2000), pp. 687–714.
 - [10] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, in Monographs and Studies in Mathematics 24, Pitman (Advanced Publishing Program), Boston, 1985.
 - [11] C. KIM, R. D. LAZAROV, J. E. PASCIAK, AND P. S. VASSILEVSKI, *Multiplier spaces for the mortar finite element method in three dimensions*, SIAM J. Numer. Anal., 39 (2001), pp. 519–538.
 - [12] H. H. KIM AND C.-O. LEE, *A preconditioner for the FETI-DP formulation with mortar methods in two dimensions*, SIAM J. Numer. Anal., 42 (2005), pp. 2159–2175.
 - [13] H. H. KIM AND O. B. WIDLUND, *Two-level Schwarz algorithms with overlapping subregions for mortar finite elements*, SIAM J. Numer. Anal., 44 (2006), pp. 1514–1534.
 - [14] H. H. KIM, *A FETI-DP preconditioner for mortar methods in three dimensions*, Electron. Trans. Numer. Anal., 26 (2007), pp. 103–120.
 - [15] H. H. KIM, *A FETI-DP formulation of three dimensional elasticity problems with mortar discretization*, SIAM J. Numer. Anal., 46 (2008), pp. 2346–2370.
 - [16] A. KLAWONN AND O. RHEINBACH, *A parallel implementation of dual-primal FETI methods for three-dimensional linear elasticity using a transformation of basis*, SIAM J. Sci. Comput., 28 (2006), pp. 1886–1906.
 - [17] A. KLAWONN AND O. RHEINBACH, *Inexact FETI-DP methods*, Internat. J. Numer. Methods Engrg., 69 (2007), pp. 284–307.
 - [18] A. KLAWONN AND O. RHEINBACH, *Robust FETI-DP methods for heterogeneous three dimensional elasticity problems*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 1400–1414.
 - [19] A. KLAWONN, O. B. WIDLUND, AND M. DRYJA, *Dual-primal FETI methods for three-dimensional elliptic problems with heterogeneous coefficients*, SIAM J. Numer. Anal., 40 (2002), pp. 159–179.
 - [20] A. KLAWONN AND O. B. WIDLUND, *Dual and dual-primal FETI methods for elliptic problems with discontinuous coefficients in three dimensions*, in Domain Decomposition Methods in Sciences and Engineering (Chiba, 1999), DDM.org, Augsburg, Germany, 2001, pp. 29–39.
 - [21] A. KLAWONN AND O. B. WIDLUND, *Dual-primal FETI methods for linear elasticity*, Comm. Pure Appl. Math., 59 (2006), pp. 1523–1572.
 - [22] J. LI AND O. WIDLUND, *BDDC algorithms for incompressible Stokes equations*, SIAM J. Numer. Anal., 44 (2006), pp. 2432–2455.
 - [23] J. LI AND O. B. WIDLUND, *FETI-DP, BDDC, and block Cholesky methods*, Internat. J. Numer. Methods Engrg., 66 (2006), pp. 250–271.
 - [24] J. LI AND O. WIDLUND, *On the use of inexact subdomain solvers for BDDC algorithms*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 1415–1428.
 - [25] J. MANDEL, C. R. DOHRMANN, AND R. TEZAUER, *An algebraic theory for primal and dual substructuring methods by constraints*, Appl. Numer. Math., 54 (2005), pp. 167–193.
 - [26] J. MANDEL AND C. R. DOHRMANN, *Convergence of a balancing domain decomposition by constraints and energy minimization*, Numer. Linear Algebra Appl., 10 (2003), pp. 639–659.
 - [27] J. MANDEL, B. SOUSEDÍK, AND C. R. DOHRMANN, *On multilevel BDDC*, in Proceedings of the 17th International Conference on Domain Decomposition Methods in Science and Engineering, Strobl, Austria, 2006, U. Langer, M. Discacciati, D. Keyes, O. Widlund, and W. Zulehner, eds., Lect. Notes Comput. Sci. Engrg. 60, Springer-Verlag, Berlin, 2007, pp. 287–294.
 - [28] A. TOSELLI AND O. WIDLUND, *Domain decomposition methods—Algorithms and theory*,

- Springer Ser. Comput. Math. 34, Springer-Verlag, Berlin, 2005.
- [29] X. TU, *Three-level BDDC in three dimensions*, SIAM J. Sci. Comput., 29 (2007), pp. 1759–1780.
 - [30] X. TU, *Three-level BDDC in two dimensions*, Internat. J. Numer. Methods Engrg., 69 (2007), pp. 33–59.
 - [31] B. I. WOHLMUTH, *A mortar finite element method using dual spaces for the Lagrange multiplier*, SIAM J. Numer. Anal., 38 (2000), pp. 989–1012.
 - [32] B. I. WOHLMUTH, *Discretization Methods and Iterative Solvers Based on Domain Decomposition*, Lect. Notes Comput. Sci. Engrg. 17, Springer-Verlag, Berlin, 2001.

ON PRESSURE APPROXIMATION VIA PROJECTION METHODS FOR NONSTATIONARY INCOMPRESSIBLE NAVIER–STOKES EQUATIONS*

ANDREAS PROHL†

Abstract. Projection methods are an efficient tool to approximate strong solutions of the incompressible Navier–Stokes equations. As a major deficiency, these methods often suffer from reduced accuracy for pressure iterates caused by nonphysical boundary data, going along with suboptimal error estimates for pressure iterates. We verify a rigorous bound for arising boundary layers in Chorin’s scheme under realistic regularity assumptions. In a second step, the new Chorin–Penalty method is proposed, where optimal rate of convergence for pressure iterates is shown.

Key words. Navier–Stokes equations, incompressible fluid, quasi-compressibility method, projection method

AMS subject classifications. 65M12, 65M60, 35K55, 35Q35

DOI. 10.1137/07069609X

1. Introduction. Given an open bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, for $d = 2, 3$, and a time $T > 0$, we consider the time-dependent Navier–Stokes equations for incompressible, viscous ($\nu > 0$) Newtonian fluids,

$$(1.1) \quad \mathbf{u}_t - \nu \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla) \mathbf{u} + \nabla p = \mathbf{f} \quad \text{in } \Omega_T := (0, T) \times \Omega,$$

$$(1.2) \quad \operatorname{div} \mathbf{u} = 0 \quad \text{in } \Omega_T,$$

$$(1.3) \quad \mathbf{u} = \mathbf{0} \quad \text{on } \partial\Omega_T := (0, T) \times \partial\Omega,$$

$$(1.4) \quad \mathbf{u}(0, \cdot) = \mathbf{u}_0 \quad \text{in } \Omega.$$

Here, $\mathbf{u} : \Omega_T \rightarrow \mathbb{R}^d$ denotes the velocity field, $p : \Omega_T \rightarrow \mathbb{R}$ the scalar pressure of vanishing mean value, i.e., $\int_{\Omega} p(\cdot, \mathbf{x}) \, d\mathbf{x} = 0$, and a given force $\mathbf{f} : \Omega_T \rightarrow \mathbb{R}^d$ is driving the fluid flow, with initial velocity field $\mathbf{u}_0 : \Omega \rightarrow \mathbb{R}^d$.

To construct and analyze numerical schemes for (1.1)–(1.4), we benefit from well-known analytical results for the given problem, which we recall here for the convenience of the reader. For this purpose, we introduce some notation: let $L^p(\Omega)$, $H^r(\Omega)$, and $H_0^r(\Omega)$, for $r \in \mathbb{N}$ be usual Lebesgue and Sobolev spaces, which are endowed with standard scalar products and induced norms $\|\cdot\|_{H^r}$. We recall that $H^{-1}(\Omega) = [H_0^1(\Omega)]^*$. Let $L_0^p(\Omega) \subset L^p(\Omega)$ be the space of functions, whose elements have vanishing integrals. Spaces of vector-valued functions will be indicated with bold-face letters, e.g., $\mathbf{H}_0^1(\Omega) = [H_0^1(\Omega)]^d$, for $d = 2, 3$. We make frequent use of the spaces

$$\mathbf{J}_0(\Omega) = \{\mathbf{v} \in \mathbf{L}^2(\Omega) : \operatorname{div} \mathbf{v} = 0 \text{ in } \Omega, \langle \mathbf{v}, \mathbf{n} \rangle = 0 \text{ on } \partial\Omega\},$$

$$\mathbf{J}_1(\Omega) = \{\mathbf{v} \in \mathbf{H}_0^1(\Omega) : \operatorname{div} \mathbf{v} = 0 \text{ in } \Omega\},$$

where $\langle \cdot, \cdot \rangle$ denotes the standard scalar product in \mathbb{R}^d , and $\mathbf{n}(\mathbf{x}) \in \mathbb{S}^{d-1}$ is the unit vector field pointing outside Ω . For a Banach space X , let $L^p(0, T; X)$, and $W^{m,p}(0, T; X)$ denote standard Bochner spaces.

*Received by the editors July 3, 2007; accepted for publication (in revised form) May 13, 2008; published electronically October 29, 2008.

<http://www.siam.org/journals/sinum/47-1/69609.html>

†Mathematisches Institut, Universität Tübingen, Auf der Morgenstelle 10, D-72076 Tübingen, Germany (prohl@na.uni-tuebingen.de).

Let us recall the concept of weak solutions to (1.1)–(1.4), for $\mathbf{u}_0 \in \mathbf{J}_0(\Omega)$, and $\mathbf{f} \in L^2(0, T, \mathbf{J}_1^*(\Omega))$ from [20, Chapter 3]: A function $\mathbf{u} \in L^2(0, T; \mathbf{J}_1(\Omega)) \cap L^\infty(0, T; \mathbf{J}_0(\Omega))$ is called a weak solution of (1.1)–(1.4), if (1.1)–(1.2) hold in distributional sense, and boundary and initial data in (1.3) and (1.4) are attained. Moreover, we have the following further properties specific for dimensions $d = 2$ and $d = 3$:

- $d = 2$: weak solutions are unique, belong to $W^{1,2}(0, T; \mathbf{J}_1^*(\Omega))$, and hence to $C([0, T]; \mathbf{J}_0(\Omega))$, and satisfy for almost all $t \in [0, T]$ the energy identity (1.5)

$$\frac{1}{2} \|\mathbf{u}(t, \cdot)\|_{L^2}^2 + \nu \int_0^t \|\nabla \mathbf{u}(s, \cdot)\|_{L^2}^2 ds = \frac{1}{2} \|\mathbf{u}_0\|_{L^2}^2 + \int_0^t \langle \mathbf{f}(s, \cdot), \mathbf{u}(s, \cdot) \rangle_{\mathbf{J}_1^* \times \mathbf{J}_1} ds.$$

In addition, provided $\mathbf{u}_0 \in \mathbf{J}_1(\Omega)$, $\mathbf{f} \in L^2(0, T; \mathbf{J}_0(\Omega))$, and Ω has $C^{1,1}$ -boundary or is a convex polygonal domain, there holds

$$(\mathbf{u}, p) \in \left[L^2(0, T; \mathbf{H}^2(\Omega)) \cap C([0, T]; \mathbf{J}_1(\Omega)) \right] \times L^2(0, T; L_0^2(\Omega) \cap H^1(\Omega)).$$

- $d = 3$: weak solutions belong to $W^{1,4/3}(0, T; \mathbf{J}_1^*(\Omega))$, are weakly continuous mappings from $[0, T]$ to $\mathbf{J}_0(\Omega)$, and satisfy an inequality version of (1.5). They are locally strong, provided $\mathbf{u}_0 \in \mathbf{J}_1(\Omega)$, $\mathbf{f} \in L^2(0, T; \mathbf{J}_0(\Omega))$, and Ω has $C^{1,1}$ -boundary.

In below, we always suppose that the data of problems (1.1)–(1.4) satisfy

- (A1) (regularity of domain) The unique solution $\mathbf{w} \in \mathbf{J}_1(\Omega)$ of the stationary, incompressible Stokes problem $-\nu \Delta \mathbf{w} + \nabla \pi = \mathbf{g}$ in $\Omega \subset \mathbb{R}^d$ is already in $\mathbf{J}_1(\Omega) \cap \mathbf{H}^2(\Omega)$, provided $\mathbf{g} \in \mathbf{L}^2(\Omega)$, and satisfies $\|\mathbf{w}\|_{H^2} \leq C \|\mathbf{g}\|_{L^2}$.
- (A2) (regularity of data) For any $T > 0$, let $\mathbf{u}_0 \in \mathbf{J}_1(\Omega) \cap \mathbf{H}^2(\Omega)$, and $\mathbf{f} \in W^{2,\infty}(0, T; \mathbf{L}^2(\Omega))$.

In order to approximate weak solutions of (1.1)–(1.4) by using a general Galerkin method, one proper temporal discretization strategy is the implicit Euler method, where iterates satisfy the (damped) discrete energy law ($M > 0$); see, e.g., [20, Chapter 3],

$$(1.6) \quad \frac{1}{2} \|\mathbf{u}^M\|_{L^2}^2 + \frac{k^2}{2} \sum_{m=1}^M \|d_t \mathbf{u}^m\|_{L^2}^2 + \nu k \sum_{m=1}^M \|\nabla \mathbf{u}^m\|_{L^2}^2 = \frac{1}{2} \|\mathbf{u}_0\|_{L^2}^2 + k \sum_{m=1}^M \langle \mathbf{f}(t_m, \cdot), \mathbf{u}^m \rangle_{\mathbf{J}_1^* \times \mathbf{J}_1}.$$

Here, we denote $d_t \phi^{m+1} := \frac{1}{k} \{\phi^{m+1} - \phi^m\}$, where $k = t_{m+1} - t_m > 0$ is the time-step. For given $\mathbf{f} \in L^2(0, T; \mathbf{J}_0(\Omega))$, the uniform bound (1.6) is then the key to conclude (subsequence) convergence of iterates against weak solutions $\mathbf{u} : \Omega_T \rightarrow \mathbb{R}^d$ of (1.1)–(1.4), for $k \rightarrow 0$.

The practical disadvantage of implicit discretization strategies is the significant computational effort implied from the necessity to solve coupled nonlinear algebraic problems to determine (Galerkin approximations) (\mathbf{u}^m, p^m) at every time-step given by $1 \leq m \leq M$. As a consequence, splitting algorithms were developed to reduce complexity of actual computations; among them, and one of the first, is Chorin's projection method [2, 3, 19], where iterates for velocity field and pressure are independently obtained at every time-step. Below, let $\mathbf{f}^m := \mathbf{f}(t_m, \cdot)$, and suppose that $\mathbf{u}^0 \in \mathbf{J}_1(\Omega)$ is given.

ALGORITHM A. 1. Let $m \geq 0$. Given $\mathbf{u}^m \in \mathbf{J}_0(\Omega)$, find $\tilde{\mathbf{u}}^{m+1} \in \mathbf{H}_0^1(\Omega)$ that satisfies

$$(1.7) \quad \frac{1}{k} \{ \tilde{\mathbf{u}}^{m+1} - \mathbf{u}^m \} - \nu \Delta \tilde{\mathbf{u}}^{m+1} + (\mathbf{u}^m \cdot \nabla) \tilde{\mathbf{u}}^{m+1} = \mathbf{f}^{m+1} \quad \text{in } \Omega.$$

2. Given $\tilde{\mathbf{u}}^{m+1} \in \mathbf{H}_0^1(\Omega)$, compute $(\mathbf{u}^{m+1}, p^{m+1}) \in \mathbf{J}_0(\Omega) \times [L_0^2(\Omega) \cap H^1(\Omega)]$ from

$$(1.8) \quad \frac{1}{k} \{ \mathbf{u}^{m+1} - \tilde{\mathbf{u}}^{m+1} \} + \nabla p^{m+1} = \mathbf{0}, \quad \operatorname{div} \mathbf{u}^{m+1} = 0 \quad \text{in } \Omega,$$

$$(1.9) \quad \langle \mathbf{u}^{m+1}, \mathbf{n} \rangle = 0 \quad \text{on } \partial\Omega.$$

The latter step can be reformulated as a problem for the pressure function only,

$$(1.10) \quad -\Delta p^{m+1} = -\frac{1}{k} \operatorname{div} \tilde{\mathbf{u}}^{m+1} \quad \text{in } \Omega, \quad \partial_{\mathbf{n}} p^{m+1} = 0 \quad \text{on } \partial\Omega.$$

Hence, each step consists of (1.7), (1.10), and the algebraic update (1.8) to obtain $(\mathbf{u}^{m+1}, p^{m+1})$.

In order to understand error effects inherent to temporal discretization, and operator splitting in Chorin's scheme, we shift the index in (1.8) back, and add the equation to (1.7); together with (1.10), we obtain

$$(1.11) \quad d_t \tilde{\mathbf{u}}^{m+1} - \nu \Delta \tilde{\mathbf{u}}^{m+1} + (\mathbf{u}^m \cdot \nabla) \tilde{\mathbf{u}}^{m+1} + \nabla p^m = \mathbf{f}^{m+1} \quad \text{in } \Omega,$$

$$(1.12) \quad \operatorname{div} \tilde{\mathbf{u}}^{m+1} - k \Delta p^{m+1} = 0 \quad \text{in } \Omega,$$

$$(1.13) \quad \partial_{\mathbf{n}} p^{m+1} = 0 \quad \text{on } \partial\Omega.$$

We make the following crucial observations implied from Chorin's decoupling strategy, for every $0 \leq m \leq M$:

- (i) The velocity field $\tilde{\mathbf{u}}^{m+1} : \Omega \rightarrow \mathbb{R}^d$ is *not divergence-free* any more, but satisfies a "quasi-compressibility equation" (1.12), with a penalization parameter equal to the time-step, and a penalization term that requires $p^{m+1} \in [L_0^2(\Omega) \cap H^1(\Omega)]$.
- (ii) Iterates of the pressure satisfy a *homogeneous Neumann boundary condition*, which is in contrast to the pressure $p : \Omega_T \rightarrow \mathbb{R}$ that satisfies (1.1)–(1.4).
- (iii) The pressure iterate in (1.11) is used in an explicit fashion, which *rules out an immediate discrete energy law*, where test functions \mathbf{u}^{m+1} and p^{m+1} in (1.11) and (1.12) are used.

As a consequence of the lack of a discrete energy law, we need not hope to construct weak solutions of (1.1)–(1.4) as proper limits of iterates from Chorin's scheme (e.g., in the sense of weak subsequence convergence). Instead, given that strong solutions to (1.1)–(1.4) exist, we may exploit their improved regularity properties to establish convergence of iterates of Algorithm A at an optimal rate. As already mentioned, strong solutions exist globally in time in 2D, and local existence is known for $d = 3$.

The convergence analysis of Algorithm A has a long history, which started with first studies by Temam [19], and continued with a series of interesting works of Shen, and W. E & J.G. Liu (see, e.g., [17, 6] and [4, 5]), where solutions to (1.1)–(1.4) were assumed to be smooth. Unfortunately, solutions to (1.1)–(1.4) suffer a breakdown of regularity for $t \rightarrow 0$ even for smooth initial data, which is due to an incompatibility of (1.2) and the prescribed data [8], which restricts the applicability of these results. A major step towards getting optimal error estimates in the context of existing strong solutions has been done by Rannacher [16], where the different error effects in Chorin's

method as a semi-implicit pressure-stabilization method (1.11)–(1.13) were pointed out, leading to the following result which is first proved in [13, Theorem 6.1].

THEOREM 1.1. *Let $\{(\tilde{\mathbf{u}}^m, p^m)\}_{m=0}^M$ be the solution of Chorin’s method (1.7)–(1.9), and let (\mathbf{u}, p) be a strong solution of (1.1)–(1.4) up to $t_M = T$. Suppose that*

$$\|\mathbf{u}^0 - \mathbf{u}_0\|_{L^2} + \sqrt{k} \|\mathbf{u}^0 - \mathbf{u}_0\|_{H^1} \leq Ck.$$

For sufficiently small time-steps $k \leq k_0(T)$, there exists a constant $C = C(T) > 0$, such that

$$\begin{aligned} \text{(a)} \quad & \max_{1 \leq m \leq M} \left[\|\mathbf{u}(t_m, \cdot) - \tilde{\mathbf{u}}^m\|_{L^2} + \tau_m \|p(t_m, \cdot) - p^m\|_{H^{-1}} \right] \leq Ck, \\ \text{(b)} \quad & \max_{1 \leq m \leq M} \left[\|\mathbf{u}(t_m, \cdot) - \tilde{\mathbf{u}}^m\|_{H^1} + \sqrt{\tau_m} \|p(t_m, \cdot) - p^m\|_{L^2} \right] \leq C\sqrt{k}, \end{aligned}$$

where $\tau_m = \min\{1, t_m\}$.

Corresponding results hold for $\{\mathbf{u}^m\}_{m=1}^M \subset \mathbf{J}_0(\Omega)$ from Step 2 in Algorithm A, thanks to well-known stability properties of the Helmholtz projection $\mathbf{P}_{\mathbf{J}_0} : \mathbf{L}^2(\Omega) \rightarrow \mathbf{J}_0(\Omega)$; cf. [20].

The proof of Theorem 1.1 in [13] is split into three steps: in a first step, optimal error estimates for the implicit Euler discretization from [8] in the presence of strong solutions of (1.1)–(1.4) are recalled to control time-discretization effects. In a second step, a modified version of (1.11)–(1.13) is studied, where the pressure iterate p^m in (1.11) is shifted to p^{m+1} . We remark that this pressure-stabilization method is of its own interest, since it allows for more finite element pairings [9, 1], where otherwise the discrete LBB condition restricts stable finite element pairings. Optimal error estimates which control perturbation effects due to (1.12) and (1.13) are the key results of the analysis, and provide k -independent a priori bounds for velocity and pressure iterates in strong norms. The latter bounds are then necessary for an optimal error estimate between this auxiliary problem, and (1.11)–(1.13) in the last step, which closes the proof.

Remark 1.1. An extension of this result to a fully discrete (LLB-stable) finite element discretization of Algorithm A is easily possible in the proof in [13]. Moreover, the stabilization effect in Algorithm A allows for equal order finite elements which violate the discrete LBB condition for choices $k \geq Ch^2$ [9]; see [13] for further details.

These $L^2(\Omega)$ -error bounds show optimal convergence behavior for velocity fields computed from Algorithm A, and only suboptimal convergence behavior for pressure iterates, which become optimal in the negative norm $H^{-1}(\Omega) = [H_0^1(\Omega)]^*$. This observation reflects observed boundary layers in the computed pressure iterates, which are caused by the nonphysical boundary condition in (1.10). In [16], it is conjectured that the thickness of the boundary layer is of order $\mathcal{O}(\sqrt{k} |\log(k)|)$, and first order of convergence holds on compact subdomains $\Omega_\delta \Subset \Omega$, where $\text{dist}(\Omega_\delta, \partial\Omega) \geq \delta$, for $\delta = \sqrt{k} |\log(k)|$. The conjecture in [16] is based on a heuristic argument to control the error due to pressure stabilization in the case of the stationary Stokes problem: as has been pointed out above, the perturbation of the incompressibility constraint, and prescription of nonphysical boundary data for the pressure in Algorithm A are accounted for by considering a fully implicit version of (1.11)–(1.13), where the key is again to study the following stationary problem, with $\varepsilon = k$: Find $(\mathbf{u}^\varepsilon, p^\varepsilon) \in [\mathbf{H}_0^1(\Omega) \cap \mathbf{H}^2(\Omega)] \times [L_0^2(\Omega) \cap H^1(\Omega)]$ such that

$$(1.14) \quad -\nu \Delta \mathbf{u}^\varepsilon + \nabla p^\varepsilon = \mathbf{f}, \quad \text{div } \mathbf{u}^\varepsilon - \varepsilon \Delta p^\varepsilon = 0 \quad \text{in } \Omega,$$

$$(1.15) \quad \partial_{\mathbf{n}} p^\varepsilon = 0 \quad \text{on } \partial\Omega.$$

The following equations control errors $\mathbf{e} := \mathbf{u} - \mathbf{u}^\varepsilon$ and $\eta = p - p^\varepsilon$, where (\mathbf{u}, p) is the strong solution of the stationary incompressible Stokes problem,

$$(1.16) \quad -\nu \Delta \mathbf{e} + \nabla \eta = \mathbf{0}, \quad \operatorname{div} \mathbf{e} - \varepsilon \Delta \eta = -\varepsilon \Delta p \quad \text{in } \Omega,$$

$$(1.17) \quad \partial_{\mathbf{n}} \eta = \partial_{\mathbf{n}} p \quad \text{on } \partial \Omega.$$

Thanks to $\mathbf{e} = \frac{1}{\nu} \Delta_D^{-1} \nabla \eta$, the second identity in (1.16) may be replaced as an equation only for the pressure error $\eta : \Omega \rightarrow \mathbb{R}$, which involves a pseudo-differential operator of order zero, such that

$$(1.18) \quad \frac{1}{\nu} \operatorname{div} \Delta_D^{-1} \nabla \eta - \varepsilon \Delta \eta = -\varepsilon \Delta p \quad \text{in } \Omega, \quad \partial_{\mathbf{n}} \eta = \partial_{\mathbf{n}} p \quad \text{on } \partial \Omega.$$

Here, we denote $\boldsymbol{\psi} := \Delta_D^{-1} \mathbf{w}$ as the solution of $\Delta \boldsymbol{\psi} = \mathbf{w}$ in Ω , and $\boldsymbol{\psi} = \mathbf{0}$ on $\partial \Omega$. In [16], the operator $\operatorname{div} \Delta_D^{-1} \nabla$ is replaced by the identity operator, and the above control of boundary layers is then derived. Our first result in this work is a rigorous derivation of arising boundary layers caused by Chorin's method for existing strong solutions. We use the following notation, with $0 < \delta < \frac{1}{2} \operatorname{diam}(\Omega)$,

$$\Omega_\delta = \{ \mathbf{x} \in \Omega : \operatorname{dist}(\mathbf{x}, \partial \Omega) > \delta \} \Subset \Omega \subset \mathbb{R}^d.$$

THEOREM 1.2. *Suppose that (\mathbf{u}, p) is a strong solution of (1.1)–(1.4) up to time $t_M = T$, with $\mathbf{f} \in C([0, T]; \mathbf{L}^{2r}(\Omega))$, $r \geq 1$, and $\Omega \subset \mathbb{R}^d$ of class $C^{2,\alpha}$, for $0 < \alpha < 1$. Let $\{(\mathbf{u}^m, p^m)\}_{m=1}^M$ be iterates from Algorithm A. Let $\Omega_\delta \subset \mathbb{R}^d$, for $d = 2, 3$. For sufficiently small $k \leq k_0(T)$, and $r^{-1} + (r')^{-1} = 1$, there holds*

$$\begin{aligned} & \max_{1 \leq m \leq M} [\tau_m \|p(t_m, \cdot) - p^m\|_{L^2(\Omega_\delta)}] \\ & \leq C \sqrt{k} \left[\sqrt{k} + \left(\frac{\sqrt{k}}{2r'} \right)^{\frac{1}{2r'}} \|\mathbf{f}\|_{L^\infty(L^{2r})} + \exp\left(-\frac{\delta}{\sqrt{k}}\right) \right]. \end{aligned}$$

The result is verified in section 2, where the key observation is a corresponding bound for errors $\eta \in W^{1,2r}(\Omega)$ which solve (1.18); cf. Theorem 2.1. We remark that the error analysis is of independent interest to, e.g., control the quantitative behavior of errors close to the boundary for pressure stabilization methods [9], where $\varepsilon = \mathcal{O}(h^2)$.

Remark 1.2. 1. The decay property has first been studied by W. E and J.G. Liu in [4] via asymptotic analysis for a restricted model problem, where first order of convergence is established on compact subdomains.

2. Regularity of $\Omega \subset \mathbb{R}^d$ is required to use L^p -theory for strong solutions of the stationary incompressible Stokes problem; cf. [20, Prop. 2.2].

3. This result evidences a boundary layer of order $\delta = \sqrt{k} |\log(k)|$, with improved rate of convergence of almost $\frac{3}{4}$ on subdomains Ω_κ , $\kappa \geq \delta$.

Apparent boundary layers of the projection method cannot be accepted when accurate data for the pressure or the velocity gradient close to the boundary are needed; undesirable consequences include pollution effects to involved quantities in more complex fluid flow problems (e.g., physicochemical hydrodynamics, or magnetohydrodynamics) that cannot be avoided in general [14, 15]. Hence, it is necessary to develop projection methods of comparable computational effort that are exempted from this deficiency. In [13], the *Chorin–Uzawa scheme* ($\beta = 0$) is proposed that avoids this drawback of Chorin's projection method. Let $\beta \geq 0$. The

method again splits each iteration step into several substeps, and starts with initial data $(\mathbf{u}^0, \tilde{\mathbf{u}}^0, p^0, \tilde{p}^0) \in \mathbf{J}_1(\Omega) \times \mathbf{H}_0^1(\Omega) \times [L_0^2(\Omega)]^2$.

ALGORITHM B. 1. For $0 \leq m \leq M$, let $(\mathbf{u}^m, \tilde{\mathbf{u}}^m, p^m, \tilde{p}^m) \in \mathbf{J}_0(\Omega) \times \mathbf{H}_0^1(\Omega) \times [L_0^2(\Omega)]^2$ be given. Find $\tilde{\mathbf{u}}^{m+1} \in \mathbf{H}_0^1(\Omega)$ such that

$$(1.19) \quad \frac{1}{k} \{ \tilde{\mathbf{u}}^{m+1} - \mathbf{u}^m \} - \beta \nabla \operatorname{div} d_t \tilde{\mathbf{u}}^{m+1} - \nu \Delta \tilde{\mathbf{u}}^{m+1} + (\mathbf{u}^m \cdot \nabla) \tilde{\mathbf{u}}^{m+1} + \nabla \{ p^m - \tilde{p}^m \} = \mathbf{f}^{m+1} \quad \text{in } \Omega.$$

2. Find $(\mathbf{u}^{m+1}, \tilde{p}^{m+1}) \in \mathbf{J}_0(\Omega) \times L_0^2(\Omega)$ that solves

$$(1.20) \quad \frac{1}{k} \{ \mathbf{u}^{m+1} - \tilde{\mathbf{u}}^{m+1} \} + \nabla \tilde{p}^{m+1} = 0, \quad \operatorname{div} \mathbf{u}^{m+1} = 0 \quad \text{in } \Omega,$$

$$(1.21) \quad \langle \mathbf{u}^{m+1}, \mathbf{n} \rangle = 0 \quad \text{on } \partial\Omega.$$

3. Determine $p^{m+1} \in L_0^2(\Omega)$ from

$$(1.22) \quad p^{m+1} = p^m - \alpha \operatorname{div} \tilde{\mathbf{u}}^{m+1} \quad \text{in } \Omega, \quad 0 < \alpha < 1.$$

Again, (1.20) may be reformulated as a Poisson problem for the pressure $\tilde{p}^{m+1} : \Omega \rightarrow \mathbb{R}$.

Step 1 in Algorithm B leads to a coupled computation of components of the velocity field, which is due to the second term in (1.19), and which is the price we pay to better enforce the incompressibility constraint. However, other advantages of projection methods are still valid, since velocity and pressure iterates are computed independently.

By eliminating \tilde{p}^{m+1} from the scheme, we easily obtain the following reformulation of the Chorin–Uzawa method as a semiexplicit “artificial compressibility method” [20, 16, 13],

$$(1.23) \quad d_t \tilde{\mathbf{u}}^{m+1} - \beta \nabla \operatorname{div} d_t \tilde{\mathbf{u}}^{m+1} - \nu \Delta \tilde{\mathbf{u}}^{m+1} + (\mathbf{u}^m \cdot \nabla) \tilde{\mathbf{u}}^{m+1} + \nabla p^m = \mathbf{f}^{m+1} \quad \text{in } \Omega,$$

$$(1.24) \quad \operatorname{div} \tilde{\mathbf{u}}^{m+1} + \frac{k}{\alpha} d_t p^{m+1} = 0 \quad \text{in } \Omega,$$

$$(1.25) \quad \tilde{\mathbf{u}}^{m+1} = \mathbf{0} \quad \text{on } \partial\Omega.$$

Since no unphysical boundary conditions are involved any more, and motivated by numerical experiments in [13], we conjecture accurate approximations of the pressure up to the boundary $\partial\Omega$. In fact, the following result is taken from [13, Theorem 8.2].

THEOREM 1.3. Suppose that (\mathbf{u}, p) is a strong solution of (1.1)–(1.4). For $0 < t_{m_1} = \mathcal{O}(1)$, let initial data $\tilde{p}^{m_1} = 0$, and $(\mathbf{u}^{m_1}, \tilde{\mathbf{u}}^{m_1}, p^{m_1}) \in \mathbf{J}_0(\Omega) \times \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ are such that

$$(1.26) \quad \|\mathbf{u}^{m_1} - \mathbf{u}(t_{m_1}, \cdot)\|_{L^2} + \|\tilde{\mathbf{u}}^{m_1} - \mathbf{u}(t_{m_1}, \cdot)\|_{L^2} + \sqrt{k} \|p^{m_1} - p(t_{m_1}, \cdot)\|_{L^2} \leq Ck.$$

Then, iterates $\{(\tilde{\mathbf{u}}^m, p^m)\}_{m=m_1+1}^M$ solving (1.19)–(1.22), for $\beta = 0$, satisfy for $k \leq k_0(T)$,

$$(1.27) \quad \max_{m_1 \leq m \leq M} \left[\|\tilde{\mathbf{u}}^m - \mathbf{u}(t_m, \cdot)\|_{L^2} + \sqrt{k} \|p^m - p(t_m, \cdot)\|_{L^2} \right] + \left(k \sum_{m=m_1}^M \|\tilde{\mathbf{u}}^m - \mathbf{u}(t_m, \cdot)\|_{H^1}^2 \right)^{\frac{1}{2}} \leq C \left(1 + \log \frac{1}{k} \right) k.$$

The Chorin–Uzawa scheme suffers from the need of accurate initial data for the pressure function and additional regularity requirements for strong solutions of (1.1)–(1.4), and computational experiments are reported in [13] where rates of convergence deteriorate if one of the requirements is violated; see also the computational experiments reported in section 5. However, both drawbacks can be avoided, if stretched time-grids $m \mapsto k_m = \min\{mk_0^2, k_0\}$ are used throughout the calculation that refine near the origin to attribute a singular weight to iterates as $t \rightarrow 0$. Obviously, this strategy asymptotically requires the same computational costs; for further details on the *revised Chorin–Uzawa scheme*, we refer to [13, Chapter 10].

Despite this improvement over the original Chorin–Uzawa method, and improved rates of convergence for gradients of computed velocity fields $\{\nabla \mathbf{u}^m\}$, no improved error statements for pressure iterates over those of Theorem 1.3 are known so far. Our goal here is to construct a scheme that does so; for this purpose, we come back to Algorithm B, with positive β , and change (1.22) to

$$(1.28) \quad p^{m+1} = -\frac{1}{k} \operatorname{div} \tilde{\mathbf{u}}^{m+1}.$$

In the sequel, we refer to (1.19)–(1.21), (1.28) as the *Chorin–Penalty scheme*. The following result will be shown in section 3, which verifies optimal order of convergence for iterates $\{p^m\} \subset L_0^2(\Omega)$ of Algorithm B. As will be shown in section 3.3, choices $\beta \geq 1$ are sufficient to effectively account for the decoupling in the Chorin–Penalty scheme; see also Remarks 3.1 and 3.2 below.

THEOREM 1.4. *Suppose that initial data $(\mathbf{u}^0, \tilde{\mathbf{u}}^0, p^0, \tilde{p}^0) \in [\mathbf{H}_0^1(\Omega)]^2 \times [L_0^2(\Omega)]^2$ satisfy*

$$\|\mathbf{u}^0 - \mathbf{u}_0\|_{L^2} + \|\tilde{\mathbf{u}}^0 - \mathbf{u}_0\|_{L^2} \leq Ck, \quad \tilde{p}^0 = p^0 = 0.$$

Let $\{(\tilde{\mathbf{u}}^m, p^m)\}_{m=1}^M$ solve (1.19)–(1.21), (1.28), for $\beta \geq 1$, and let (\mathbf{u}, p) be strong solution to (1.1)–(1.4). There exists $C = C(T) > 0$, such that

$$\max_{1 \leq m \leq M} \left[\|\tilde{\mathbf{u}}^m - \mathbf{u}(t_m, \cdot)\|_{L^2} + \sqrt{\tau_m} \|\tilde{\mathbf{u}}^m - \mathbf{u}(t_m, \cdot)\|_{H^1} + \tau_m \|p^m - p(t_m, \cdot)\|_{L^2} \right] \leq Ck.$$

Remark 1.3. 1. The Chorin–Penalty method can be reformulated as a semi-explicit penalty method; see (3.1)–(3.2). This stationary quasi-compressibility method has been analyzed in [13, Chapter 3].

2. No additional regularity requirements for strong solutions of (1.1)–(1.4) are needed, and $p^0 \equiv \tilde{p}^0 \equiv 0$ is convenient.

Chorin’s original method is by no means the only existing projection method to solve (1.1)–(1.4): over the last four decades, many different further projection schemes have been developed, which use modified splitting strategies (e.g., the Gauge method; see, e.g., [5, 11, 12], modified boundary conditions for pressure iterates), or variants which use higher order temporal discretization, in combination with different projection (or quasi-compressibility) strategies. Recently, an interesting splitting strategy is proposed in [10] to implement boundary conditions consistently; however, the approach requires C^1 -finite elements for the velocity field, and to prove qualitative convergence (without rates) of iterates towards strong solutions of (1.1)–(1.4) requires bounded domains $\Omega \subset \mathbb{R}^d$ ($d = 2, 3$) with C^3 -boundary. Another strategy to avoid artificial boundary layers, and currently studied in the literature, are “velocity-correction methods” [7], where the error analysis requires accurate initial data for the pressure function (see discussion above). All these schemes share the goal to

TABLE 1.1

Comparison of different first-order projection methods: additional regularity requirements for strong solutions of (1.1)–(1.4) needed for convergence analysis are displayed in the first column. The subsequent two columns display convergence rates for different quantities, reflecting presence/absence of boundary layers. The last column indicates the corresponding quasi-compressibility method (QCM).

Method	additional requirements	$(k \sum_{m=1}^M \ \bar{\mathbf{u}}^m - \mathbf{u}(t_m, \cdot)\ _{H^1}^2)^{1/2}$	$\ p^M - p(t_M, \cdot)\ _{L^2}$	Related QCM
Chorin	no	$\leq C \sqrt{k}$	$\leq C \frac{1}{\sqrt{\tau_M}} \sqrt{k}$	pressure stabilization
Chorin–Uzawa	yes	$\leq C k$	$\leq C (1 + \log k) \sqrt{k}$	artificial compressibility
Chorin–Penalty	no	$\leq C k$	$\leq C \frac{1}{\tau_M} k$	penalty

circumvent the drawbacks of Chorin’s original projection scheme, and we refer to [6] for a review of the current state of the art. However, most numerical analyses of these schemes are based on the assumption that solutions of (1.1)–(1.4) are smooth, which leaves unclear whether these results apply to strong solutions which are known to exist. Hence, for general applicability, it is our goal in this work to validate the results discussed above on solid analytical grounds.

The remainder of this work is organized as follows: In section 2, we quantify arising boundary layers due to pressure-stabilization methods for the stationary Stokes problem in the context of strong solutions; see Theorem 2.1. Theorem 1.4 then follows from this result. The main results and properties for the Chorin–Penalty method (i.e., Algorithm B, for $\beta \geq 1$) are given and compared with other methods discussed above in Table 1.1; in section 3, we validate optimal convergence for pressure iterates to strong solutions of (1.1)–(1.4), as stated in Theorem 1.4. Comparative computational studies for Chorin, Chorin–Uzawa, and Chorin–Penalty schemes are reported in section 4. A conclusion is given in section 5.

2. Boundary layers in Chorin’s projection method. As is already discussed in the introduction, Chorin’s projection method, i.e., Algorithm A, suffers from marked boundary layers for the pressure error, which are bounded in Theorem 1.2. In this section, we verify this theorem by first studying a corresponding effect for strong solutions of the stationary, stabilized Stokes problem: For given $\mathbf{f} \in \mathbf{L}^2(\Omega)$, and $\varepsilon > 0$, find solutions $(\mathbf{u}^\varepsilon, p^\varepsilon) \in [\mathbf{H}_0^1(\Omega) \cap \mathbf{H}^2(\Omega)] \times [L_0^2(\Omega) \cap H^1(\Omega)]$ of $(\nu > 0)$

$$(2.1) \quad -\nu \Delta \mathbf{u}^\varepsilon + \nabla p^\varepsilon = \mathbf{f}, \quad \operatorname{div} \mathbf{u}^\varepsilon - \varepsilon \Delta p^\varepsilon = 0 \quad \text{in } \Omega,$$

$$(2.2) \quad \partial_{\mathbf{n}} p^\varepsilon = 0 \quad \text{on } \partial\Omega.$$

This problem is a perturbation of the incompressible Stokes equation, where strong solutions $(\mathbf{u}, p) \in [\mathbf{J}_1(\Omega) \cap \mathbf{H}^2(\Omega)] \times [L_0^2(\Omega) \cap H^1(\Omega)]$ solve $-\nu \Delta \mathbf{u} + \nabla p = \mathbf{f}$ in Ω . Below, we consider compactly contained subsets Ω_δ of Ω ,

$$\Omega_\delta = \{ \mathbf{x} \in \Omega : \operatorname{dist}(\mathbf{x}, \partial\Omega) > \delta \}, \quad \text{for } 0 < \delta < \frac{1}{2} \operatorname{diam}(\Omega).$$

THEOREM 2.1. *Let $\Omega \subset \mathbb{R}^d$, for $d \geq 2$, and $\mathbf{f} \in \mathbf{L}^2(\Omega)$, such that $\operatorname{div} \mathbf{f} \in L^2(\Omega)$. Suppose that (\mathbf{u}, p) is a strong solution of the stationary, incompressible Stokes equation in $\Omega \subset \mathbb{R}^d$, such that $p \in W^{1,2r}(\Omega)$, $r \geq 1$, and $(\mathbf{u}^\varepsilon, p^\varepsilon)$ is a strong solution*

of (2.1) and (2.2). There holds

$$\|p - p^\varepsilon\|_{L^2(\Omega_\delta)} \leq C\sqrt{\varepsilon\nu} \left[\sqrt{\varepsilon\nu} \|\Delta p\|_{L^2} + \left(\frac{\sqrt{\varepsilon\nu}}{2r'}\right)^{\frac{1}{2r'}} \|\nabla p\|_{L^{2r}} + \exp\left(-\frac{\delta}{\sqrt{\varepsilon\nu}}\right) \|\nabla p\|_{L^2} \right].$$

Proof. The equations for the error $(\mathbf{e}, \eta) := (\mathbf{u} - \mathbf{u}^\varepsilon, p - p^\varepsilon)$ are

$$(2.3) \quad \begin{aligned} -\nu\Delta\mathbf{e} + \nabla\eta &= \mathbf{0}, & \operatorname{div}\mathbf{e} - \varepsilon\Delta\eta &= -\varepsilon\Delta p & \text{in } \Omega, \\ \partial_{\mathbf{n}}\eta &= \partial_{\mathbf{n}}p & & & \text{on } \partial\Omega. \end{aligned}$$

Let

$$\sigma(\mathbf{x}) = \exp\left(-\frac{\delta}{\sqrt{\varepsilon\nu}}\right) \min\left[\exp\left(\frac{d(\mathbf{x})}{\sqrt{\varepsilon\nu}}\right), \exp\left(\frac{\delta}{\sqrt{\varepsilon\nu}}\right)\right], \quad \text{where } d(\mathbf{x}) = \operatorname{dist}(\mathbf{x}, \partial\Omega).$$

We employ a duality argument: Find $(\mathbf{w}, q) \in [\mathbf{H}_0^1(\Omega) \cap \mathbf{H}^2(\Omega)] \times [L_0^2(\Omega) \cap H^1(\Omega)]$, such that

$$(2.4) \quad -\nu\Delta\mathbf{w} + \nabla q = 0, \quad \operatorname{div}\mathbf{w} - \varepsilon\Delta q = \sigma\eta \quad \text{in } \Omega,$$

$$(2.5) \quad \partial_{\mathbf{n}}q = 0 \quad \text{on } \partial\Omega.$$

The following stability bound for solutions of (2.4) and (2.5) is easy to verify:

$$(2.6) \quad \frac{1}{\nu} \|q\|_{L^2}^2 + \nu \|\nabla\mathbf{w}\|_{L^2}^2 + \varepsilon \|\nabla q\|_{L^2}^2 \leq C\nu \|\sigma\eta\|_{L^2}^2.$$

By testing (2.4) with (\mathbf{e}, η) , and (2.3) with (\mathbf{w}, q) , we find

$$(2.7) \quad \begin{aligned} \|\sqrt{\sigma}\eta\|_{L^2}^2 &= \varepsilon(\nabla q, \nabla\eta) + (\operatorname{div}\mathbf{w}, \eta) - (\operatorname{div}\mathbf{e}, q) - \varepsilon(\nabla\eta, \nabla q) + \varepsilon(\nabla q, \nabla p) \\ &\quad - \nu(\nabla\mathbf{w}, \nabla\mathbf{e}) + (\operatorname{div}\mathbf{e}, q) + \nu(\nabla\mathbf{w}, \nabla\mathbf{e}) - (\operatorname{div}\mathbf{w}, \eta) \\ &= \varepsilon(\nabla p, \nabla q) = -\varepsilon(\Delta p, q) + \varepsilon\langle \partial_{\mathbf{n}}p, q \rangle_{\partial\Omega}. \end{aligned}$$

Since $\exp(-\frac{d(\mathbf{x})}{\sqrt{\varepsilon\nu}}) = 1$ on $\partial\Omega$, we further conclude for $\tilde{\sigma}(\mathbf{x}) = \max[\exp(-\frac{d(\mathbf{x})}{\sqrt{\varepsilon\nu}}), \exp(-\frac{\delta}{\sqrt{\varepsilon\nu}})]$,

$$(2.8) \quad \varepsilon\langle \partial_{\mathbf{n}}p, \tilde{\sigma}q \rangle_{\partial\Omega} = \varepsilon(\Delta p, \tilde{\sigma}q) + \varepsilon(\nabla p, \nabla(\tilde{\sigma}q)).$$

There remains to bound the last term in (2.8). By $|\nabla\tilde{\sigma}| \leq \frac{\tilde{\sigma}}{\sqrt{\varepsilon\nu}}$, we conclude

$$\begin{aligned} &\varepsilon(\nabla p, \tilde{\sigma}\nabla q) + \varepsilon(\nabla p, q\nabla\tilde{\sigma}) \\ &\leq C \left[\varepsilon\nu \|\tilde{\sigma}\nabla p\|_{L^2}^2 + \sqrt{\frac{\varepsilon}{\nu}} \int_{\mathcal{B}} |\nabla p| |q| \tilde{\sigma} \, d\mathbf{x} \right] + \frac{\varepsilon}{4\nu} \|\nabla q\|_{L^2}^2 \\ &\leq C\varepsilon\nu \left[\|\tilde{\sigma}\nabla p\|_{L^2(\Omega \setminus \mathcal{B})}^2 + 2\|\tilde{\sigma}\nabla p\|_{L^2(\mathcal{B})}^2 \right] + \frac{1}{4} \left[\frac{\varepsilon}{\nu} \|\nabla q\|_{L^2}^2 + \frac{1}{\nu^2} \|q\|_{L^2}^2 \right], \end{aligned}$$

where $\mathcal{B} := \operatorname{supp}|\nabla\tilde{\sigma}|$, which has d -dimensional Lebesgue measure $\mathcal{L}^d(\mathcal{B}) = \mathcal{O}(\delta)$. Hence,

$$\int_{\mathcal{B}} \exp\left(-\frac{d(\mathbf{x})}{\sqrt{\varepsilon\nu}}\right) \, d\mathbf{x} = \mathcal{O}(\sqrt{\varepsilon\nu}),$$

and we may conclude from (2.7), (2.6) as follows, for $r, r' \geq 1$, such that $\frac{1}{r} + \frac{1}{r'} = 1$,

$$\begin{aligned}
\|\sqrt{\sigma}\eta\|_{L^2} &\leq C\varepsilon\nu\|\Delta p\|_{L^2} \\
&\quad + C\sqrt{\varepsilon\nu}\left[\exp\left(-\frac{\delta}{\sqrt{\varepsilon\nu}}\right)\|\nabla p\|_{L^2(\Omega\setminus\mathcal{B})} + \|\exp\left(-\frac{d(\mathbf{x})}{\sqrt{\varepsilon\nu}}\right)\nabla p\|_{L^2(\mathcal{B})}\right] \\
&\leq C\varepsilon\nu\|\Delta p\|_{L^2} \\
&\quad + C\sqrt{\varepsilon\nu}\left[\exp\left(-\frac{\delta}{\sqrt{\varepsilon\nu}}\right)\|\nabla p\|_{L^2} + \|\nabla p\|_{2r}\left(\int_{\mathcal{B}}\exp\left(-\frac{2r'd(\mathbf{x})}{\sqrt{\varepsilon\nu}}\right)dx\right)^{\frac{1}{2r'}}\right] \\
&\leq C\varepsilon\nu\|\Delta p\|_{L^2} \\
&\quad + C\sqrt{\varepsilon\nu}\left[\exp\left(-\frac{\delta}{\sqrt{\varepsilon\nu}}\right)\|\nabla p\|_{L^2} + \left(\frac{\sqrt{\varepsilon\nu}}{2r'}\right)^{\frac{1}{2r'}}\|\nabla p\|_{L^{2r}}\right].
\end{aligned}$$

Thanks to $\|p\|_{L^2(\Omega_\delta)} \leq \|\sigma p\|_{L^2}$, this proves the assertion of the lemma. \square

Let $1 \leq r < \infty$. By [20, Prop. 2.2], solutions of the incompressible Stokes equation with $\mathbf{f} \in \mathbf{L}^{2r}(\Omega)$ are strong, and satisfy $(\mathbf{u}, p) \in \mathbf{W}^{2,2r}(\Omega) \times W^{1,2r}(\Omega)$, provided the open bounded set $\Omega \subset \mathbb{R}^d$ is of class $C^{2,\alpha}$, for $0 < \alpha < 1$. Then, for large values $r \rightarrow \infty$, the above lemma motivates an $L^2(\Omega_\delta)$ -error decay behavior for the pressure, which is almost order $\frac{3}{4}$ in the interior, and deteriorates to order $\frac{1}{2}$ if errors on a boundary layer of width $\delta = \sqrt{\nu\varepsilon}|\log(\nu\varepsilon)|$ are included.

We use Theorem 2.1 to show Theorem 1.2. As has already be pointed out in the introduction, an error analysis for Algorithm A to optimally bound time-discretization, perturbation, and decoupling error effects is split into three steps; the most critical step to bound arising errors is the one where the pressure stabilization effect is accounted for. As a consequence, we consider the following auxiliary problem: Let $\mathbf{u}_k^0 = \mathbf{u}_0$ be given. For every $0 \leq m \leq M$, find $(\mathbf{u}_k^{m+1}, p_k^{m+1}) \in [\mathbf{H}_0^1(\Omega) \cap \mathbf{H}^2(\Omega)] \times [L_0^2(\Omega) \cap H^1(\Omega)]$ such that

$$(2.9) \quad d_t \mathbf{u}_k^{m+1} - \nu \Delta \mathbf{u}_k^{m+1} + (\mathbf{P}_{\mathbf{J}_0} \mathbf{u}_k^m \cdot \nabla) \mathbf{u}_k^{m+1} + \nabla p_k^{m+1} = \mathbf{f}^{m+1} \quad \text{in } \Omega,$$

$$(2.10) \quad \operatorname{div} \mathbf{u}_k^{m+1} - k \Delta p_k^{m+1} = 0 \quad \text{in } \Omega,$$

$$(2.11) \quad \partial_{\mathbf{n}} p_k^{m+1} = 0 \quad \text{on } \partial\Omega.$$

The following uniform estimates for solutions $\{(\mathbf{v}^{m+1}, \pi^{m+1})\}_{m=0}^M \subset [\mathbf{J}_1(\Omega) \cap \mathbf{H}^2(\Omega)] \times [L_0^2(\Omega) \cap H^1(\Omega)]$ of (2.9) will be useful below; we refer to [16, 13] for a proof.

LEMMA 2.1. *Let (A1), (A2) be valid. Then, iterates $\{(\mathbf{v}^{m+1}, \pi^{m+1})\}_{m=0}^M \subset [\mathbf{J}_1(\Omega) \cap \mathbf{H}^2(\Omega)] \times [L_0^2(\Omega) \cap H^1(\Omega)]$ of (2.9) satisfy*

i) *the following uniform bounds for values $i \in \{0, 1, 2\}$, and $r \in \{1, 2, 3\}$,*

$$\max_{r \leq m \leq M} \left[\tau_{m-1}^{r-1+\frac{i}{2}} \|d_t^r \mathbf{v}^m\|_{W^{i,2}} \right] + \left(k \sum_{m=r}^M \tau_{m-1}^{[2(r-1)+i-1]_+} \|d_t^r \mathbf{v}^m\|_{W^{i,2}}^2 \right)^{1/2} \leq C,$$

where $[x]_+ := \max\{x, 0\}$. For values $i \in \{0, 1\}$, and $r \in \{1, 2, 3\}$, there holds

$$\max_{r+1 \leq m \leq M} \left[\tau_{m-1}^{r+\frac{i-1}{2}} \|d_t^r \pi^m\|_{W^{i,2}} \right] + \left(k \sum_{m=r+1}^M \tau_m^{2(r-1)+i} \|d_t^r \pi^m\|_{W^{i,2}}^2 \right)^{1/2} \leq C.$$

ii) the following error estimates, for (\mathbf{u}, p) , a strong solution of (1.1)–(1.4),

$$\begin{aligned} \max_{1 \leq m \leq M} [\|\mathbf{u}(t_m, \cdot) - \mathbf{v}^m\|_{L^2} + \sqrt{\tau_m} [\|p(t_m, \cdot) - \pi^m\|_{H^{-1}} \\ + \|\mathbf{u}(t_m, \cdot) - \mathbf{v}^m\|_{H^1}] + \tau_m \|p(t_m, \cdot) - \pi^m\|_{L^2}] \leq Ck. \end{aligned}$$

We are now in a position to sketch the proof of Theorem 1.2.

Proof. Step 1: consistency error. This error contribution in Algorithm A is accounted for by introducing the semi-implicit Euler scheme as a first auxiliary problem; then, Lemma 2.1 provides both optimal error estimates and stability results for the iterates.

Step 2: quasi-compressibility constraint. Problem (2.9)–(2.11) is the second auxiliary problem to account for error effects due to violating the incompressibility constraint; this is the step where we employ Theorem 2.1 by first considering the following quasi-stationary auxiliary problem, for every $0 \leq m \leq M$: Find $(\mathbf{U}_k^{m+1}, \Pi_k^{m+1}) \in [\mathbf{H}_0^1(\Omega) \cap \mathbf{H}^2(\Omega)] \times [L_0^2(\Omega) \cap H^1(\Omega)]$ such that

$$(2.12) \quad -\Delta \mathbf{U}_k^{m+1} + \nabla \Pi_k^{m+1} = \mathbf{F}^{m+1} \quad \text{in } \Omega,$$

$$(2.13) \quad \operatorname{div} \mathbf{U}_k^{m+1} - k \Delta \Pi_k^{m+1} = 0 \quad \text{in } \Omega,$$

$$(2.14) \quad \partial_{\mathbf{n}} \Pi_k^{m+1} = 0 \quad \text{in } \partial\Omega,$$

for $\mathbf{F}^{m+1} := \mathbf{f}^{m+1} - d_t \mathbf{v}^{m+1} - (\mathbf{v}^m \cdot \nabla) \mathbf{v}^{m+1}$, and where $\{(\mathbf{v}^{m+1}, \pi^{m+1})\}_{m=0}^M \subset [\mathbf{J}_1(\Omega) \cap \mathbf{H}^2(\Omega)] \times [L_0^2(\Omega) \cap H^1(\Omega)]$ solves (2.9). In order to apply Theorem 2.1, we need $\mathbf{F}^{m+1} \in \mathbf{L}^{2r}(\Omega)$ to apply [20, Prop. 2.2]. Thanks to the first result in Lemma 2.1, and Sobolev's inequality for $d = 2, 3$, we easily obtain uniform bounds for $\tau_{m+1} \|\mathbf{F}^{m+1}\|_{L^{2r}}$, with $1 \leq r < \infty$, and hence the right-hand side of (2.12)–(2.14),

$$\begin{aligned} & \|\pi^m - \Pi_k^m\|_{L^2(\Omega_\delta)} \\ & \leq C\sqrt{k\nu} \left[\sqrt{k\nu} \|\Delta \pi^m\|_{L^2} + \left(\frac{\sqrt{k\nu}}{2r'} \right)^{\frac{1}{2r'}} \|\pi^m\|_{W^{1,2r}} + \exp\left(-\frac{\delta}{\sqrt{k\nu}}\right) \|\pi^m\|_{W^{1,2}} \right], \end{aligned}$$

can be controlled uniformly if a time-weight τ_m is used. Next, we employ the bound

$$\max_{1 \leq m \leq M+1} \|\Pi_k^m - p_k^m\|_{L^2(\Omega)} \leq Ck,$$

which is already known from [13, section 6.2], and easily follows from the fact that both p_k^{m+1} and Π_k^{m+1} satisfy (2.13) and (2.14).

Step 3: splitting error. This step controls errors between solutions of systems (2.9)–(2.11) and (1.11)–(1.13). This error analysis leads to first-order estimates for all iterates in the considered norms; see [13, section 6.4] for a detailed analysis. \square

3. Algorithm B: The Chorin–Penalty projection method. In order to analyze the Chorin–Penalty scheme (1.19), (1.20), and (1.28), we use its reformulation as a semi-explicit penalty method,

$$(3.1) \quad \begin{aligned} d_t (\tilde{\mathbf{u}}_k^{m+1} - \beta \nabla \operatorname{div} \tilde{\mathbf{u}}_k^{m+1}) - \nu \Delta \tilde{\mathbf{u}}_k^{m+1} \\ + (\mathbf{P}_{\mathbf{J}_0} \tilde{\mathbf{u}}_k^m \cdot \nabla) \tilde{\mathbf{u}}_k^{m+1} + \nabla p_k^m = \mathbf{f}^{m+1} \quad \text{in } \Omega, \end{aligned}$$

$$(3.2) \quad \operatorname{div} \tilde{\mathbf{u}}_k^{m+1} + k p_k^{m+1} = 0 \quad \text{in } \Omega.$$

The main part of the subsequent analysis focuses on the fully implicit modification of (3.1),

$$(3.3) \quad d_t (\tilde{\mathbf{u}}_k^{m+1} - \beta \nabla \operatorname{div} \tilde{\mathbf{u}}_k^{m+1}) - \nu \Delta \tilde{\mathbf{u}}_k^{m+1} + (\mathbf{P}_{\mathbf{J}_0} \tilde{\mathbf{u}}_k^m \cdot \nabla) \tilde{\mathbf{u}}_k^{m+1} + \nabla q_k^{m+1} = \mathbf{f}^{m+1} \quad \text{in } \Omega,$$

and strong solutions $(\tilde{\mathbf{u}}_k^{m+1}, q_k^{m+1})_{m=0}^M \subset [\mathbf{H}_0^1(\Omega) \cap \mathbf{H}^2(\Omega)] \times [L_0^2(\Omega) \cap H^1(\Omega)]$, with $\tilde{\mathbf{u}}_k^0 = \mathbf{u}_0$. In the sequel, we use the sequence $\{(\mathbf{v}^{m+1}, \pi^{m+1})\}_{m=0}^M \subset [\mathbf{J}_1(\Omega) \cap \mathbf{H}^2(\Omega)] \times [L_0^2(\Omega) \cap H^1(\Omega)]$, with $\mathbf{v}^0 = \mathbf{u}_0$, which solves (2.9). We independently bound errors in (3.3) and (3.2), which are due to the perturbation of the incompressibility constraint in the linear case (section 3.1), from those which are introduced by the nonlinear term (section 3.2), as well as those due to the decoupling strategy in Algorithm B (section 3.3).

Remark 3.1. On putting $\varepsilon = k$, system (3.1)–(3.2) may be considered as a semi-implicit temporal discretization of $(\tilde{\beta} \geq 0)$

$$(3.4) \quad \mathbf{v}_t - \tilde{\beta} \nabla \operatorname{div} \mathbf{v}_t - \nu \Delta \mathbf{v} + (\mathbf{P}_{\mathbf{J}_0} \mathbf{v} \cdot \nabla) \mathbf{v} + \nabla \pi = \mathbf{f} \quad \text{in } \Omega_T,$$

$$(3.5) \quad \operatorname{div} \mathbf{v} + \varepsilon \pi = 0 \quad \text{in } \Omega_T,$$

together with $\mathbf{v}(0, \cdot) = \mathbf{u}_0$ on Ω . For $\tilde{\beta} = 0$, this formulation is known as a penalty method, which is studied in [18], and [13, section 3.2]. Hence, (3.4) is a modification thereof, which uses the additional term $-\tilde{\beta} \nabla \operatorname{div} \mathbf{v}_t$ to additionally enforce the incompressibility constraint for $\tilde{\beta} > 0$.

Another interpretation of system (3.1)–(3.2) for $\beta = 1$ comes from its reformulation

$$(3.6) \quad d_t \tilde{\mathbf{u}}_k^{m+1} - \frac{1}{k} \nabla \operatorname{div} \tilde{\mathbf{u}}_k^{m+1} - \nu \Delta \tilde{\mathbf{u}}_k^{m+1} + (\mathbf{P}_{\mathbf{J}_0} \tilde{\mathbf{u}}_k^m \cdot \nabla) \tilde{\mathbf{u}}_k^{m+1} = \mathbf{f}^{m+1}.$$

Hence, iterates $\{\tilde{\mathbf{u}}_k^m\}_m \subset \mathbf{H}_0^1(\Omega)$ from Algorithm B solve an implicit temporal discretization of the penalty formulation (3.4)–(3.5), with $\varepsilon = k$, and $\tilde{\beta} = 0$. Moreover, a discrete energy law similar to (1.6) holds for solutions of (3.6), which justifies (subsequence) convergence to weak solutions of (1.1)–(1.4) for $k \rightarrow 0$.

For $\beta \geq 1$, system (3.1)–(3.2) combines different stabilizing mechanisms to enforce the incompressibility constraint for $\beta \geq 1$ for iterates $\{\tilde{\mathbf{u}}_k^m\}$. Note that (3.6) is an implicit discretization which effectively describes iterates $\{\tilde{\mathbf{u}}_k^m\}$ in this case, and does not require one to prescribe initial data for the pressure; this is in contrast to Algorithm B, due to the (decoupling) projection step to obtain $\{\mathbf{u}^m\}_m \subset \mathbf{J}_0$.

3.1. Perturbation analysis for the penalized formulation (3.3), (3.2);

Part I: The linear case. Let $\mathbf{w}_k^0 = \mathbf{u}_0$. For every $0 \leq m \leq M$, let $(\mathbf{w}_k^{m+1}, b_k^{m+1}) \in [\mathbf{H}_0^1(\Omega) \cap \mathbf{H}^2(\Omega)] \times [L_0^2(\Omega) \cap H^1(\Omega)]$ be the strong solution of

$$(3.7) \quad d_t (\mathbf{w}_k^{m+1} - \beta \nabla \operatorname{div} \mathbf{w}_k^{m+1}) - \nu \Delta \mathbf{w}_k^{m+1} + \nabla b_k^{m+1} = \mathbf{F}^{m+1} \quad \text{in } \Omega,$$

$$(3.8) \quad \operatorname{div} \mathbf{w}_k^{m+1} + k b_k^{m+1} = 0 \quad \text{in } \Omega,$$

where $\mathbf{F}^{m+1} = \mathbf{f}^{m+1} - (\mathbf{v}^m \cdot \nabla) \mathbf{v}^{m+1}$, for every $0 \leq m \leq M$. By Lemma 2.1, we have

$$\max_{0 \leq m \leq M} [\|\mathbf{F}^{m+1}\|_{L^2} + \|d_t \mathbf{v}^{m+1}\|_{L^2}] \leq C.$$

Problem (3.7)–(3.8) is a semidiscretization in time of a penalized version of the nonstationary, incompressible Stokes equations. In order to verify optimal rates of convergence with respect to $k > 0$ towards strong solutions of the nonstationary, incompressible Stokes equations, we need to study the following auxiliary problem

first: For every $0 \leq m \leq M$, let $(\mathbf{W}_k^{m+1}, B_k^{m+1}) \in [\mathbf{H}_0^1(\Omega) \cap \mathbf{H}^2(\Omega)] \times [L_0^2(\Omega) \cap H^1(\Omega)]$ be the strong solution of

$$(3.9) \quad -\nu \Delta \mathbf{W}_k^{m+1} + \nabla B_k^{m+1} = \mathbf{F}^{m+1} - d_t \mathbf{v}^{m+1} \quad \text{in } \Omega,$$

$$(3.10) \quad \operatorname{div} \mathbf{W}_k^{m+1} + k B_k^{m+1} = 0 \quad \text{in } \Omega.$$

The following convergence properties have been shown in [13, section 3.2],

$$(3.11) \quad \|\mathbf{W}_k^m - \mathbf{v}^m\|_{H^1} + \|B_k^m - \pi^m\|_{L^2} \leq C k \|\pi^m\|_{L^2},$$

for every $1 \leq m \leq M$. By linearity of the problem (3.9)–(3.10), and Lemma 2.1, we easily obtain for $r \in \{1, 2, 3\}$,

$$(3.12) \quad \begin{aligned} & \max_{r \leq m \leq M} \tau_m^{r-1/2} [\|d_t^r (\mathbf{W}_k^m - \mathbf{v}^m)\|_{H^1} + \|d_t^r (B_k^m - \pi^m)\|_{L^2}] \\ & + \left(k \sum_{m=r}^M \tau_{m-1}^{2(r-1)} \|d_t^r (\mathbf{W}_k^m - \mathbf{v}^m)\|_{H^1}^2 \right)^{1/2} \leq C k. \end{aligned}$$

In the next step, we bound errors $(\mathbf{e}_k^{m+1}, \eta_k^{m+1}) := (\mathbf{w}_k^{m+1} - \mathbf{W}_k^{m+1}, b_k^{m+1} - B_k^{m+1})$ between solutions of (3.7)–(3.8) and (3.9)–(3.10). We have the following identities ($0 \leq m \leq M$),

$$(3.13) \quad d_t (\mathbf{e}_k^{m+1} - \beta \nabla \operatorname{div} \mathbf{e}_k^{m+1}) - \nu \Delta \mathbf{e}_k^{m+1} + \nabla \eta_k^{m+1} = (\operatorname{Id} - \beta \nabla \operatorname{div}) d_t (\mathbf{v}^{m+1} - \mathbf{W}_k^{m+1}),$$

$$(3.14) \quad \operatorname{div} \mathbf{e}_k^{m+1} + k \eta_k^{m+1} = 0,$$

with $\mathbf{e}_k^{m+1} = \mathbf{0}$ on $\partial\Omega$, and $\mathbf{e}_k^0 = \mathbf{0}$ on Ω . Let $\mathbf{W}_k^0 := \mathbf{u}^0$. By testing (3.13)–(3.14) with $(\mathbf{e}_k^{m+1}, \eta_k^{m+1})$, and using (3.12) we arrive at

$$(3.15) \quad \begin{aligned} & \frac{1}{2} \max_{1 \leq m \leq M} [\|\mathbf{e}_k^m\|_{L^2}^2 + \beta \|\operatorname{div} \mathbf{e}_k^m\|_{L^2}^2] + \frac{k^2}{2} \sum_{m=1}^M [\|d_t \mathbf{e}_k^m\|_{L^2}^2 + \beta \|\operatorname{div} d_t \mathbf{e}_k^m\|_{L^2}^2] \\ & + k \sum_{m=1}^M [\nu \|\nabla \mathbf{e}_k^m\|_{L^2}^2 + k \|\eta_k^m\|_{L^2}^2] \leq C k^2. \end{aligned}$$

This result, together with (3.11), establishes optimal convergence behavior for the velocity field obtained from (3.7)–(3.8) in $\ell^\infty(0, t_M; \mathbf{L}^2)$.

In the next step, we want to verify error bounds for the velocity gradient in $\ell^\infty(0, t_M; \mathbf{L}^2)$. For this purpose, we make r times “discrete derivatives” of (3.13), test with $\tau_{m+1}^{2r} d_t^r \mathbf{e}_k^{m+1}$, $r \in \{1, 2\}$, and use (3.15) to find

$$(3.16) \quad \begin{aligned} & \max_{r \leq m \leq M} \tau_m^{2r} [\|d_t^r \mathbf{e}_k^m\|_{L^2}^2 + \beta \|\operatorname{div} d_t^r \mathbf{e}_k^m\|_{L^2}^2] \\ & + k \sum_{m=r}^M \tau_m^{2r} [\nu \|\nabla d_t^r \mathbf{e}_k^m\|_{L^2}^2 + k \|d_t^r \eta_k^m\|_{L^2}^2] \\ & \leq C k^2 + C k \sum_{m=r}^M \tau_m^{2r-1} \left[\tau_m \|d_t^{r+1} (\mathbf{v}^m - \mathbf{W}_k^m)\|_{L^2}^2 \right. \\ & \left. + (\|d_t^r \mathbf{e}_k^m\|_{L^2}^2 + \beta \|\operatorname{div} d_t^r \mathbf{e}_k^m\|_{L^2}^2) \right]. \end{aligned}$$

A similar argument, together with (3.15) leads to ($r \geq 1$)

$$\begin{aligned}
& k \sum_{m=r}^M \tau_m^{2r-1} \left[\|d_t^r \mathbf{e}_k^m\|_{L^2}^2 + \beta \|\operatorname{div} d_t^r \mathbf{e}_k^m\|_{L^2}^2 \right] \\
& \quad + \max_{r \leq m \leq M} \tau_m^{2r-1} \left[\nu \|\nabla d_t^{r-1} \mathbf{e}_k^m\|_{L^2}^2 + k \|d_t^{r-1} \eta_k^m\|_{L^2}^2 \right] \\
(3.17) \quad & \leq C k^2 + C k \sum_{m=r}^M \left[\tau_m^{2r-1} \|d_t^r (\mathbf{v}^m - \mathbf{W}_k^m)\|_{L^2}^2 \right. \\
& \quad \left. + \tau_m^{2(r-1)} \left(\|\nabla d_t^{r-1} \mathbf{e}_k^m\|_{L^2}^2 + k \|d_t^{r-1} \eta_k^m\|_{L^2}^2 \right) \right].
\end{aligned}$$

We can now combine (3.16), (3.17), and use (3.15) to verify the following bound, for $r \in \{1, 2\}$,

$$\begin{aligned}
& \max_{r \leq m \leq M} \tau_m^{2r-1} \left[\tau_m \left[\|d_t^r \mathbf{e}_k^m\|_{L^2}^2 + \beta \|\operatorname{div} d_t^r \mathbf{e}_k^m\|_{L^2}^2 \right] \right. \\
& \quad \left. + \nu \|\nabla d_t^{r-1} \mathbf{e}_k^m\|_{L^2}^2 + k \|d_t^{r-1} \eta_k^m\|_{L^2}^2 \right] \\
(3.18) \quad & + k \sum_{m=r}^M \tau_m^{2r-1} \left[\nu \tau_m \|\nabla d_t^r \mathbf{e}_k^m\|_{L^2}^2 \right. \\
& \quad \left. + \beta \|\operatorname{div} d_t^r \mathbf{e}_k^m\|_{L^2}^2 + k \tau_m \|d_t^r \eta_k^m\|_{L^2}^2 \right] \leq C k^2.
\end{aligned}$$

Then, Lemma 2.1, (3.11), (3.12), and a stability result for the *div*-operator yields to

$$(3.19) \quad \max_{0 \leq m \leq M} \left[\|\mathbf{u}(t_m, \cdot) - \mathbf{w}_k^m\|_{L^2} + \sqrt{\tau_m} \|\mathbf{u}(t_m, \cdot) - \mathbf{w}_k^m\|_{H^1} + \tau_m \|p(t_m, \cdot) - b_k^m\|_{L^2} \right] \leq C k,$$

with the latter result being a consequence of a stability result for the divergence operator.

3.2. Perturbation analysis for the penalized formulation (3.3), (3.2);

Part II: Extension to the nonlinear case. Because of (3.19), there remains to bound errors $(\boldsymbol{\xi}^{m+1}, \chi^{m+1}) := (\tilde{\mathbf{u}}_k^{m+1} - \mathbf{w}_k^{m+1}, q_k^{m+1} - b_k^{m+1})$ to estimate the error between strong solutions of (1.1)–(1.4), and (3.3), (3.2). We subtract the equations (3.7)–(3.8) of the linear auxiliary problem from the corresponding ones (3.3)–(3.2). For every $0 \leq m \leq M$, there holds

$$(3.20) \quad d_t (\boldsymbol{\xi}^{m+1} - \beta \nabla \operatorname{div} \boldsymbol{\xi}^{m+1}) - \nu \Delta \boldsymbol{\xi}^{m+1} + \nabla \chi^{m+1} = (\mathbf{v}^m \cdot \nabla) \mathbf{v}^{m+1} - (\mathbf{P}_{\mathbf{J}_0} \tilde{\mathbf{u}}_k^m \cdot \nabla) \tilde{\mathbf{u}}_k^{m+1},$$

$$(3.21) \quad \operatorname{div} \boldsymbol{\xi}^{m+1} + k \chi^{m+1} = 0.$$

We compute

$$\begin{aligned}
& -(\mathbf{v}^m \cdot \nabla) \mathbf{v}^{m+1} + (\mathbf{P}_{\mathbf{J}_0} \mathbf{u}_k^m \cdot \nabla) \mathbf{u}_k^{m+1} = (\mathbf{P}_{\mathbf{J}_0} \boldsymbol{\xi}^m \cdot \nabla) \mathbf{v}^{m+1} + (\mathbf{P}_{\mathbf{J}_0} \tilde{\mathbf{u}}_k^m \cdot \nabla) \boldsymbol{\xi}^{m+1} \\
& \quad - (\mathbf{P}_{\mathbf{J}_0} \tilde{\mathbf{u}}_k^m \cdot \nabla) (\mathbf{v}^{m+1} - \mathbf{w}_k^{m+1}) - (\mathbf{P}_{\mathbf{J}_0} (\mathbf{v}^m - \mathbf{w}_k^m) \cdot \nabla) \mathbf{v}^{m+1}.
\end{aligned}$$

This observation, the skew-symmetricity property $((\mathbf{P}_{\mathbf{J}_0} \boldsymbol{\phi} \cdot \nabla) \boldsymbol{\psi}, \boldsymbol{\psi}) = 0$ for $\boldsymbol{\psi} \in \mathbf{H}_0^1(\Omega)$, and the \mathbf{H}^1 -stability of $\mathbf{P}_{\mathbf{J}_0}$ then lead to

$$\begin{aligned}
& \frac{1}{2} \max_{1 \leq m \leq M} \left[\|\boldsymbol{\xi}^m\|_{L^2}^2 + \beta \|\operatorname{div} \boldsymbol{\xi}^m\|_{L^2}^2 \right] + \frac{k^2}{2} \sum_{m=1}^M \left[\|d_t \boldsymbol{\xi}^m\|_{L^2}^2 + \beta \|\operatorname{div} d_t \boldsymbol{\xi}^m\|_{L^2}^2 \right] \\
& + k \sum_{m=1}^M \left[\frac{\nu}{2} \|\nabla \boldsymbol{\xi}^m\|_{L^2}^2 + k \|\chi^m\|_{L^2}^2 \right] \\
& \leq \left[\|\boldsymbol{\xi}^0\|_{L^2}^2 + \|\operatorname{div} \boldsymbol{\xi}^0\|_{L^2}^2 \right] + C k \sum_{m=1}^M \|\Delta \mathbf{v}^m\|_{L^2}^2 \left[\|\boldsymbol{\xi}^m\|_{L^2}^2 + \|\mathbf{v}^m - \mathbf{w}_k^m\|_{L^2}^2 \right] \\
& + \left| \left((\mathbf{P}_{\mathbf{J}_0} \tilde{\mathbf{u}}_k^{m-1} \cdot \nabla) (\mathbf{v}^m - \mathbf{w}_k^m), \boldsymbol{\xi}^m \right) \right| \\
& \leq C k^2 + k \sum_{m=1}^M \|\boldsymbol{\xi}^m\|_{L^2}^2 + k \sum_{m=1}^M \left[\|\nabla \boldsymbol{\xi}^{m-1}\|_{L^2} + \|\nabla (\mathbf{v}^{m-1} - \mathbf{w}_k^{m-1})\|_{L^2} \right] \\
& \quad \times \|\nabla (\mathbf{v}^m - \mathbf{w}_k^m)\|_{L^2} \|\boldsymbol{\xi}^m\|_{L^2}^{1/2} \|\nabla \boldsymbol{\xi}^m\|_{L^2}^{1/2} + k \sum_{m=1}^M \left[\|\boldsymbol{\xi}^m\|_{L^2}^2 + \|\mathbf{v}^m - \mathbf{w}_k^m\|_{L^2}^2 \right].
\end{aligned}$$

We use Lemma 2.1, ii), and (3.19) to conclude with the discrete Gronwall's inequality that

$$\begin{aligned}
& \max_{1 \leq m \leq M} \left[\|\boldsymbol{\xi}^m\|_{L^2}^2 + \beta \|\operatorname{div} \boldsymbol{\xi}^m\|_{L^2}^2 \right] + \frac{k^2}{2} \sum_{m=1}^M \left[\|d_t \boldsymbol{\xi}^m\|_{L^2}^2 + \beta \|\operatorname{div} d_t \boldsymbol{\xi}^m\|_{L^2}^2 \right] \\
(3.22) \quad & + k \sum_{m=1}^M \left[\nu \|\nabla \boldsymbol{\xi}^m\|_{L^2}^2 + k \|\chi^m\|_{L^2}^2 \right] \leq C k^2.
\end{aligned}$$

Next, making “discrete time-derivatives” in (3.20), (3.21) with respect to time, and then test the system with $(\tau_{m+1}^2 d_t \boldsymbol{\xi}^{m+1}, \tau_{m+1}^2 d_t \chi_{m+1})$,

$$\begin{aligned}
& \max_{2 \leq m \leq M} \tau_m^2 \left[\|d_t \boldsymbol{\xi}^m\|_{L^2}^2 + \beta \|\operatorname{div} d_t \boldsymbol{\xi}^m\|_{L^2}^2 \right] \\
& + k^2 \sum_{m=2}^M \tau_m^2 \left[\|d_t^2 \boldsymbol{\xi}^m\|_{L^2}^2 + \beta \|\operatorname{div} d_t \boldsymbol{\xi}^m\|_{L^2}^2 \right] \\
(3.23) \quad & + k \sum_{m=2}^M \tau_m^2 \left[\nu \|\nabla d_t \boldsymbol{\xi}^m\|_{L^2}^2 + k \|d_t \chi^m\|_{L^2}^2 \right] \\
& \leq C k \sum_{m=2}^M \left[\tau_m^2 |(\text{NLT}_A^m, d_t \boldsymbol{\xi}^m)| + \tau_m \left(\|d_t \boldsymbol{\xi}^m\|_{L^2}^2 + \|\operatorname{div} d_t \boldsymbol{\xi}^m\|_{L^2}^2 \right) \right],
\end{aligned}$$

where

$$\begin{aligned}
\text{NLT}_A^{m+1} & = (\mathbf{P}_{\mathbf{J}_0} d_t \boldsymbol{\xi}^m \cdot \nabla) \mathbf{v}^{m+1} + (\mathbf{P}_{\mathbf{J}_0} \boldsymbol{\xi}^{m-1} \cdot \nabla) d_t \mathbf{v}^{m+1} + (\mathbf{P}_{\mathbf{J}_0} d_t \tilde{\mathbf{u}}_k^m \cdot \nabla) \boldsymbol{\xi}^{m+1} \\
(3.24) \quad & + (\mathbf{P}_{\mathbf{J}_0} \tilde{\mathbf{u}}_k^{m-1} \cdot \nabla) d_t \boldsymbol{\xi}^{m+1} - (\mathbf{P}_{\mathbf{J}_0} d_t \tilde{\mathbf{u}}_k^{m-1} \cdot \nabla) (\mathbf{v}^{m+1} - \mathbf{w}_k^{m+1}) \\
& - (\mathbf{P}_{\mathbf{J}_0} \tilde{\mathbf{u}}_k^{m-1} \cdot \nabla) d_t (\mathbf{v}^{m+1} - \mathbf{w}_k^{m+1}) \\
& - (\mathbf{P}_{\mathbf{J}_0} d_t (\mathbf{v}^m - \mathbf{w}_k^m) \cdot \nabla) \mathbf{v}^{m+1} - (\mathbf{P}_{\mathbf{J}_0} (\mathbf{v}^{m-1} - \mathbf{w}_k^{m-1}) \cdot \nabla) d_t \mathbf{v}^{m+1}.
\end{aligned}$$

Thanks to Lemma 2.1, i), the first two terms on the right-hand side of (3.24) in (3.23) can be easily controlled, as well as the fourth term. (Note that $\{d_t \tilde{\mathbf{u}}_k^{m+1}\}$ is uniformly bounded in $\ell^\infty(0, t_{M+1}; \mathbf{L}^2(\Omega))$.) By Lemma 2.1, i), and (3.18), the fifth, seventh, and eighth term on the right-hand side of (3.24) can be handled in a standard way. To bound the third and sixth term on the right-hand side of (3.24), we use the following reformulation:

$$(3.25) \quad \begin{aligned} (\mathbf{P}_{\mathbf{J}_0} d_t \tilde{\mathbf{u}}_k^m \cdot \nabla) \boldsymbol{\xi}^{m+1} &= (\mathbf{P}_{\mathbf{J}_0} d_t (\tilde{\mathbf{u}}_k^m - \mathbf{v}^m) \cdot \nabla) \boldsymbol{\xi}^{m+1} + (\mathbf{P}_{\mathbf{J}_0} d_t \mathbf{v}^m \cdot \nabla) \boldsymbol{\xi}^{m+1}, \\ (\mathbf{P}_{\mathbf{J}_0} \tilde{\mathbf{u}}_k^{m-1} \cdot \nabla) d_t (\mathbf{v}^{m+1} - \mathbf{w}_k^{m+1}) &= (\mathbf{P}_{\mathbf{J}_0} \boldsymbol{\xi}^{m-1} \cdot \nabla) d_t (\mathbf{v}^{m+1} - \mathbf{w}_k^{m+1}) \\ &+ (\mathbf{P}_{\mathbf{J}_0} \mathbf{w}_k^{m-1} \cdot \nabla) d_t (\mathbf{v}^{m+1} - \mathbf{w}_k^{m+1}). \end{aligned}$$

We use Lemma 2.1, i), (3.12), and (3.18) to obtain the uniform estimate

$$(3.26) \quad \max_{1 \leq m \leq M} \left[\|\nabla \mathbf{w}_k^m\|_{L^2} + \sqrt{\tau_m} \| \nabla d_t \mathbf{w}_k^m \|_{L^2} \right] \leq C.$$

By inserting these bounds in (3.23) yields to

$$(3.27) \quad \begin{aligned} &\max_{2 \leq m \leq M} \tau_m^2 \left[\|d_t \boldsymbol{\xi}^m\|_{L^2}^2 + \beta \|\operatorname{div} d_t \boldsymbol{\xi}^m\|_{L^2}^2 \right] \\ &+ k^2 \sum_{m=2}^M \tau_m^2 \left[\|d_t^2 \boldsymbol{\xi}^m\|_{L^2}^2 + \beta \|\operatorname{div} d_t^2 \boldsymbol{\xi}^m\|_{L^2}^2 \right] \\ &+ k \sum_{m=2}^M \tau_m^2 \left[\nu \|\nabla d_t \boldsymbol{\xi}^m\|_{L^2}^2 + k \|d_t \chi^m\|_{L^2}^2 \right] \\ &\leq C k^2 + k \sum_{m=2}^M \tau_m \left[\|d_t \boldsymbol{\xi}^m\|_{L^2}^2 + \|\operatorname{div} d_t \boldsymbol{\xi}^m\|_{L^2}^2 \right]. \end{aligned}$$

The last term in (3.27) can be controlled, if we test (3.20) by $d_t \boldsymbol{\xi}^{m+1}$ and the ‘‘discrete time-derivative’’ version of (3.21) by χ^{m+1} ,

$$(3.28) \quad \begin{aligned} &\|d_t \boldsymbol{\xi}^{m+1}\|_{L^2}^2 + \beta \|\operatorname{div} d_t \boldsymbol{\xi}^{m+1}\|_{L^2}^2 + \frac{\nu}{2} d_t \|\nabla \boldsymbol{\xi}^{m+1}\|_{L^2}^2 \\ &+ \frac{\nu k}{2} \|\nabla d_t \boldsymbol{\xi}^{m+1}\|_{L^2}^2 + \frac{k}{2} d_t \|\chi^{m+1}\|_{L^2}^2 + \frac{k^2}{2} \|d_t \chi^{m+1}\|_{L^2}^2 \\ &\leq C \left[\|\nabla \boldsymbol{\xi}^{m+1}\|_{L^2}^2 + \|\nabla (\mathbf{v}^m - \mathbf{w}_k^m)\|_{L^2}^2 \right] + \frac{C}{\delta \kappa} \|d_t \boldsymbol{\xi}^{m+1}\|_{L^2}^2 + \frac{\kappa}{\delta} \|\nabla d_t \boldsymbol{\xi}^{m+1}\|_{L^2}^2 \\ &+ \frac{\delta}{\tau_{m+1}^{1/2}} \|\nabla \mathbf{u}_k^m\|_{L^2}^2 \left[\|\nabla \boldsymbol{\xi}^{m+1}\|_{L^2}^2 + \|\nabla (\mathbf{v}^{m+1} - \mathbf{w}_k^{m+1})\|_{L^2}^2 \right], \end{aligned}$$

for $\delta, \kappa \geq 1$. If we choose $\kappa = \frac{1}{\sqrt{\delta}}$ and take δ sufficiently large, the last but one term can be absorbed on the left-hand side, and we obtain

$$\begin{aligned}
& k \sum_{m=1}^M \tau_m \left[\|d_t \boldsymbol{\xi}^m\|_{L^2}^2 + \beta \|\operatorname{div} d_t \boldsymbol{\xi}^m\|_{L^2}^2 \right] \\
& + \max_{1 \leq m \leq M} \tau_m \left[\nu \|\nabla \boldsymbol{\xi}^m\|_{L^2}^2 + k \|\chi^m\|_{L^2}^2 \right] \\
(3.29) \quad & + k^2 \sum_{m=1}^M \left[\nu \tau_m \|\nabla d_t \boldsymbol{\xi}^m\|_{L^2}^2 + k \|d_t \chi^m\|_{L^2}^2 \right] \\
& \leq Ck \sum_{m=1}^M \left[\nu \|\nabla \boldsymbol{\xi}^m\|_{L^2}^2 + k \|\chi^m\|_{L^2}^2 + \nu \|\nabla(\mathbf{v}^m - \mathbf{w}_k^m)\|_{L^2}^2 \right. \\
& \quad \left. + \frac{\nu}{\delta^{3/2}} \tau_m^2 \|\nabla d_t \boldsymbol{\xi}^m\|_{L^2}^2 \right].
\end{aligned}$$

As a consequence, (3.29) and (3.27), in combination with a stability result for the divergence operator, give the desired bound

$$(3.30) \quad \max_{1 \leq m \leq M} \left[\|\mathbf{u}(t_m, \cdot) - \mathbf{u}_k^m\|_{L^2} + \sqrt{\tau_m} \|\mathbf{u}(t_m, \cdot) - \mathbf{u}_k^m\|_{H^1} + \tau_m \|p(t_m) - q_k^m\|_{L^2} \right] \leq Ck.$$

We finish this part with useful uniform bounds for $\{(\tilde{\mathbf{u}}_k^{m+1}, q_k^{m+1})\}_{m=0}^M$: make $r-1$ “discrete time-derivatives” in (3.3), (3.2), for $r \in \{1, 2\}$, and test with $\tau_{m+1}^{2(r-1)} d_t^r u_k^{m+1}$; in a second step, we make r “discrete time-derivatives,” and test with $\tau_{m+1}^{2r-1} d_t^r u_k^{m+1}$. A simple calculation then yields

$$\begin{aligned}
(3.31) \quad & \max_{r \leq m \leq M} \tau_m^{2r-1} \left[\|d_t^r \tilde{\mathbf{u}}_k^m\|_{L^2}^2 + \beta \|\operatorname{div} d_t^r \tilde{\mathbf{u}}_k^m\|_{L^2}^2 \right] \\
& + k \sum_{m=r}^M \tau_m^{2(r-1)} \left[\|d_t^r \tilde{\mathbf{u}}_k^m\|_{L^2}^2 + \beta \|\operatorname{div} d_t^r \tilde{\mathbf{u}}_k^m\|_{L^2}^2 \right] \\
& + k \sum_{m=r}^M \tau_m^{2r-1} \left[\nu \|\nabla d_t^r \tilde{\mathbf{u}}_k^m\|_{L^2}^2 + k \|d_t^r q_k^m\|_{L^2}^2 \right] \leq C,
\end{aligned}$$

thanks to (3.27) and (3.30). Also, we may use Lemma 2.1, i), (3.12), (3.16), and (3.27) together with (3.21) to find

$$(3.32) \quad \max_{1 \leq m \leq M} \left[\tau_m^2 \|d_t q_k^m\|^2 \right] + k \sum_{m=1}^M \tau_m \|d_t q_k^m\|^2 \leq C, \quad k^2 \sum_{m=2}^M \tau_m^2 \|d_t^2 q_k^m\|^2 \leq C.$$

3.3. Perturbation analysis for the Chorin–Penalty scheme (3.1)–(3.2): Transition from the penalty method to the projection method. In this section, let $\{(\mathbf{u}_k^{m+1}, q_k^{m+1})\}_{m=0}^M$ denote the solution of (3.3), (3.2), and $\{(\tilde{\mathbf{u}}_k^{m+1}, p_k^{m+1})\}_{m=0}^M$ solves (3.1)–(3.2)—where the latter is the reformulation of Algorithm B. The last step to verify Theorem 1.4 consists in bounding the error $(\mathbf{E}^m, \Pi^m) := (\mathbf{u}_k^m - \tilde{\mathbf{u}}_k^m, q_k^m - p_k^m)$,

which satisfies for every $1 \leq m \leq M$,

$$(3.33) \quad d_t(\mathbf{E}^m - \beta \nabla \operatorname{div} \mathbf{E}^m) - \nu \Delta \mathbf{E}^m + \nabla \Pi^{m-1} = -k \nabla d_t q_k^m - \mathcal{Q}(\mathbf{E}^m),$$

$$(3.34) \quad \operatorname{div} \mathbf{E}^m + k \Pi^m = 0,$$

with $\mathcal{Q}(\mathbf{E}^{m+1}) := (\mathbf{P}_{\mathbf{J}_0} \mathbf{E}^m \cdot \nabla) \mathbf{u}_k^{m+1} + (\mathbf{P}_{\mathbf{J}_0} [\mathbf{u}_k^m - \mathbf{E}^m] \cdot \nabla) \mathbf{E}^{m+1}$. In the next step, we test (3.33)–(3.34) with (\mathbf{E}^m, Π^m) ; for this purpose, we first calculate

$$(3.35) \quad \begin{aligned} (\nabla \Pi^{m-1}, \mathbf{E}^m) &= (\nabla \Pi^m, \mathbf{E}^m) - k (\nabla d_t \Pi^m, \mathbf{E}^m) \\ &= k \|\Pi^m\|_{L^2}^2 - k^2 (d_t \Pi^m, \Pi^m) \\ &\geq k \left[1 - \frac{2}{3} \right] \|\Pi^m\|_{L^2}^2 - \frac{3k^3}{8} \|d_t \Pi^m\|_{L^2}^2, \end{aligned}$$

and

$$k (\nabla d_t q_k^m, \mathbf{E}^m) = k^2 (d_t q_k^m, \Pi^m),$$

as well as

$$\begin{aligned} |(\mathcal{Q}(\mathbf{E}^m), \mathbf{E}^m)| &\leq C \|\mathbf{E}^{m-1}\|_{L^3} \|\nabla \mathbf{u}_k^m\|_{L^2} \|\mathbf{E}^m\|_{L^6} \\ &\leq C \|\mathbf{E}^{m-1}\|_{L^2}^{1/2} \|\nabla \mathbf{E}^{m-1}\|_{L^2}^{1/2} \|\nabla \mathbf{u}_k^m\|_{L^2} \|\nabla \mathbf{E}^m\|_{L^2}. \end{aligned}$$

Therefore, we arrive at the estimate

$$(3.36) \quad \begin{aligned} &\frac{1}{2} \max_{1 \leq m \leq M} \left[\|\mathbf{E}^m\|_{L^2}^2 + \beta \|\operatorname{div} \mathbf{E}^m\|_{L^2}^2 \right] \\ &\quad + \frac{k^2}{2} \sum_{m=1}^M \left[\|d_t \mathbf{E}^m\|_{L^2}^2 + \beta \|\operatorname{div} d_t \mathbf{E}^m\|_{L^2}^2 \right] \\ &\quad + k \sum_{m=1}^M \left[\frac{\nu}{2} \|\nabla \mathbf{E}^m\|_{L^2}^2 + \left[\frac{1}{3} - \frac{1}{\delta} \right] k \|\Pi^m\|_{L^2}^2 \right] \\ &\leq k^4 \sum_{m=1}^M \left[\frac{3}{8} \|d_t \Pi^m\|_{L^2}^2 + C_\delta \|d_t q_k^m\|_{L^2}^2 \right] \\ &\quad + \frac{k}{\nu^3} \sum_{m=1}^M \|\mathbf{E}^m\|_{L^2}^2 \|\nabla \mathbf{u}_k^{m+1}\|_{L^2}^4, \quad \delta > 4. \end{aligned}$$

Thanks to (3.34), and $\beta \geq 1$, the first term on the right-hand side can be absorbed on the left-hand side. The second term can be bounded by $C k^2$, by means of (3.31) and (3.32).

To control the error for the pressure, we “make a time-derivative” of (3.33), test with $\tau_m^2 d_t \mathbf{E}^m$, and note (3.31), (3.32)₂.

$$\begin{aligned}
(3.37) \quad & \frac{1}{2} \max_{2 \leq m \leq M} \tau_m^2 \left[\|d_t \mathbf{E}^m\|_{L^2}^2 + \beta \|\operatorname{div} d_t \mathbf{E}^m\|_{L^2}^2 \right] \\
& + \frac{k^2}{2} \sum_{m=2}^M \tau_m^2 \left[\|d_t^2 \mathbf{E}^m\|_{L^2}^2 + \left[\beta - \frac{3}{4} \right] \|\operatorname{div} d_t^2 \mathbf{E}^m\|_{L^2}^2 \right] \\
& + k \sum_{m=2}^M \tau_m^2 \left[\nu \|\nabla d_t \mathbf{E}^m\|_{L^2}^2 + \left[\frac{1}{3} - \frac{1}{\delta} \right] k \|d_t \Pi^m\|_{L^2}^2 \right] \\
& \leq Ck^2 + C_\delta k^4 \sum_{m=2}^M \tau_m^2 \|d_t^2 q_k^m\|_{L^2}^2 \\
& + k \sum_{m=2}^M \tau_m \left[\|d_t \mathbf{E}^m\|_{L^2}^2 + \beta \|\operatorname{div} d_t \mathbf{E}^m\|_{L^2}^2 \right] \\
& + k \sum_{m=2}^M \tau_m^2 |\operatorname{NLT}_B^m|.
\end{aligned}$$

Here, we used (3.34) to conclude that

$$\begin{aligned}
\tau_m^2 (\nabla d_t \Pi^{m-1}, d_t \mathbf{E}^m) &= \tau_m^2 (\nabla d_t \Pi^m, d_t \mathbf{E}^m) - k \tau_m^2 (\nabla d_t^2 \Pi^m, d_t \mathbf{E}^m) \\
&\geq k \tau_m^2 \|d_t \Pi^m\|_{L^2}^2 - k^2 \tau_m^2 (d_t^2 \Pi^m, d_t \Pi^m) \\
&\geq \left[1 - \frac{2}{3} \right] k \tau_m^2 \|d_t \Pi^m\|_{L^2}^2 - \frac{3}{8} k^3 \tau_m^2 \|d_t^2 \Pi^m\|_{L^2}^2 \\
&\geq \left[1 - \frac{2}{3} \right] k \tau_m^2 \|d_t \Pi^m\|_{L^2}^2 - \frac{3}{8} k \tau_m^2 \|\operatorname{div} d_t^2 \mathbf{E}^m\|_{L^2}^2.
\end{aligned}$$

We skip the detailed study of NLT_B^m , since it does not involve further difficulties superior to those detailed in subsection 3.2 at this place. The crucial term, however, is the last but one term on the right-hand side of (3.37).

For this purpose, we test (3.33) with $\tau_m d_t \mathbf{E}^m$. In a first step, we observe that

$$\begin{aligned}
\tau_m (\nabla \Pi^{m-1}, d_t \mathbf{E}^m) &= \tau_m (\nabla \Pi^m, d_t \mathbf{E}^m) + \tau_m k (d_t \Pi^m, \operatorname{div} d_t \mathbf{E}^m) \\
&\geq \frac{k}{2} \tau_m \left[d_t \|\Pi^m\|_{L^2}^2 + k \|d_t \Pi^m\|_{L^2}^2 \right] \\
&\quad - \frac{3}{4} \tau_m \|\operatorname{div} d_t \mathbf{E}^m\|_{L^2}^2 - \frac{k^2}{3} \tau_m \|d_t \Pi^m\|_{L^2}^2 \\
&= \frac{k}{2} d_t \left[\tau_m \|\Pi^m\|_{L^2}^2 \right] - \frac{k}{2} \|\Pi^{m-1}\|_{L^2}^2 \\
&\quad + \frac{k^2}{6} \tau_m \|d_t \Pi^m\|_{L^2}^2 - \frac{3}{4} \tau_m \|\operatorname{div} d_t \mathbf{E}^m\|_{L^2}^2,
\end{aligned}$$

as well as

$$\begin{aligned}
k \tau_m (\nabla d_t q_k^m, d_t \mathbf{E}^m) &= k^2 \tau_m (d_t \Pi^m, d_t q_k^m) \\
&\leq 3 \tau_m k^2 \|d_t q_k^m\|_{L^2}^2 + \frac{k^2}{12} \tau_m \|d_t \Pi^m\|_{L^2}^2.
\end{aligned}$$

By (3.31), (3.32), and (3.36), we then obtain

$$\begin{aligned}
 (3.38) \quad & k \sum_{m=1}^M \tau_m \left[\|d_t \mathbf{E}^m\|^2 + \left[\beta - \frac{3}{4} \right] \|\operatorname{div} d_t \mathbf{E}^m\|^2 \right] \\
 & + \frac{1}{2} \max_{1 \leq m \leq M} \tau_m \left[\nu \|\nabla \mathbf{E}^m\|^2 + k \|\Pi^m\|^2 \right] \\
 & + \frac{k^2}{12} \sum_{m=1}^M \tau_m \left[\|\nabla d_t \mathbf{E}^m\|^2 + k \|d_t \Pi^m\|^2 \right] \\
 & \leq C_\gamma k^2 + \frac{k}{\gamma^{3/2}} \sum_{m=1}^M \tau_m^2 \|\nabla d_t \mathbf{E}^m\|^2, \quad (\gamma > 1).
 \end{aligned}$$

We may now use (3.38) to control the last but one term on the right-hand side of (3.37); upon choosing $\gamma > 1$ sufficiently large, the last term in (3.38) may be absorbed by the corresponding term on the left-hand side of (3.37). A standard argumentation then establishes the bound $\max_{0 \leq m \leq M} \tau_m \|\Pi^m\| \leq C k$.

Together with (3.36) through (3.38) and the results from subsections 3.1 to 3.3, this settles the proof of Theorem 1.4.

Remark 3.2. Choosing $\beta \geq 0$ is sufficient to obtain all results in subsections 3.1 and 3.2. Values $\beta \geq 1$ are needed only in subsection 3.3 to properly deal with splitting errors in the Chorin–Penalty projection method; it is sufficient to justify the argument below inequality (3.36), where errors due to the splitting, and computed in (3.35), are absorbed by the term headed by β ; see also section 4 for computational evidence.

4. Computational experiments. We computationally compare Chorin’s method (Algorithm A) with its variants, i.e., Chorin–Uzawa method (Algorithm B, for $\alpha \in (0, 1)$) and Chorin–Penalty method ((1.19)–(1.21), (1.28), for $\beta \geq 1$). Our main focuses are

- (i) to compare possible boundary layers of $m \mapsto [p(t_m, \cdot) - p^m] \in L_0^2(\Omega)$, including their evolution in time,
- (ii) to compare possible transition layers caused by starting with initial pressure data p^0 for Chorin–Uzawa, and Chorin–Penalty scheme, where $\|p^0 - p(0, \cdot)\|_{L^2}$ is large, and
- (iii) to study convergence behavior of computed pressures for the Chorin–Uzawa and Chorin–Penalty methods, depending on different choices of α, β .

Example 1. Let $\Omega = (0, 1)^2 \subset \mathbb{R}^2$, and

$$\mathbf{u}(x, y, t) = \begin{pmatrix} x^2(1-x)^2(2y-6y^2+4y^3) \\ -y^2(1-y)^2(2x-6x^2+4x^3) \end{pmatrix}, \quad p(x, y, t) = \left(x^2 + y^2 - \frac{2}{3}\right)(1+t^2),$$

be the solution of the evolutionary Stokes problem; i.e., $\mathbf{f} : \Omega_T \rightarrow \mathbb{R}^2$ is computed from (1.1)–(1.4), where the nonlinear term in (1.1) is neglected. Let \mathcal{T}_h be an equidistant triangulation of Ω of mesh-size $h = 1/30$, and $k = 2^j/500$, $j = 0, 1, 2, \dots$, an equidistant time-step for the time interval $[0, 1]$. The LBB-stable MINI-Stokes element is used for spatial discretization of the three projection methods.

Snapshots for the pressure error in Figure 1 show marked boundary layers for the pressure (first line), as opposed to almost uniform errors for Chorin–Uzawa (middle line) and Chorin–Penalty (last line); comparative plots in the last line for corresponding L^2 -errors (before being dominated by errors due to spatial discretization) motivate that significant errors close to the boundary control the overall error in Chorin’s scheme. Corresponding profiles are obtained for plots of $m \mapsto \operatorname{div} \tilde{\mathbf{u}}^m$.

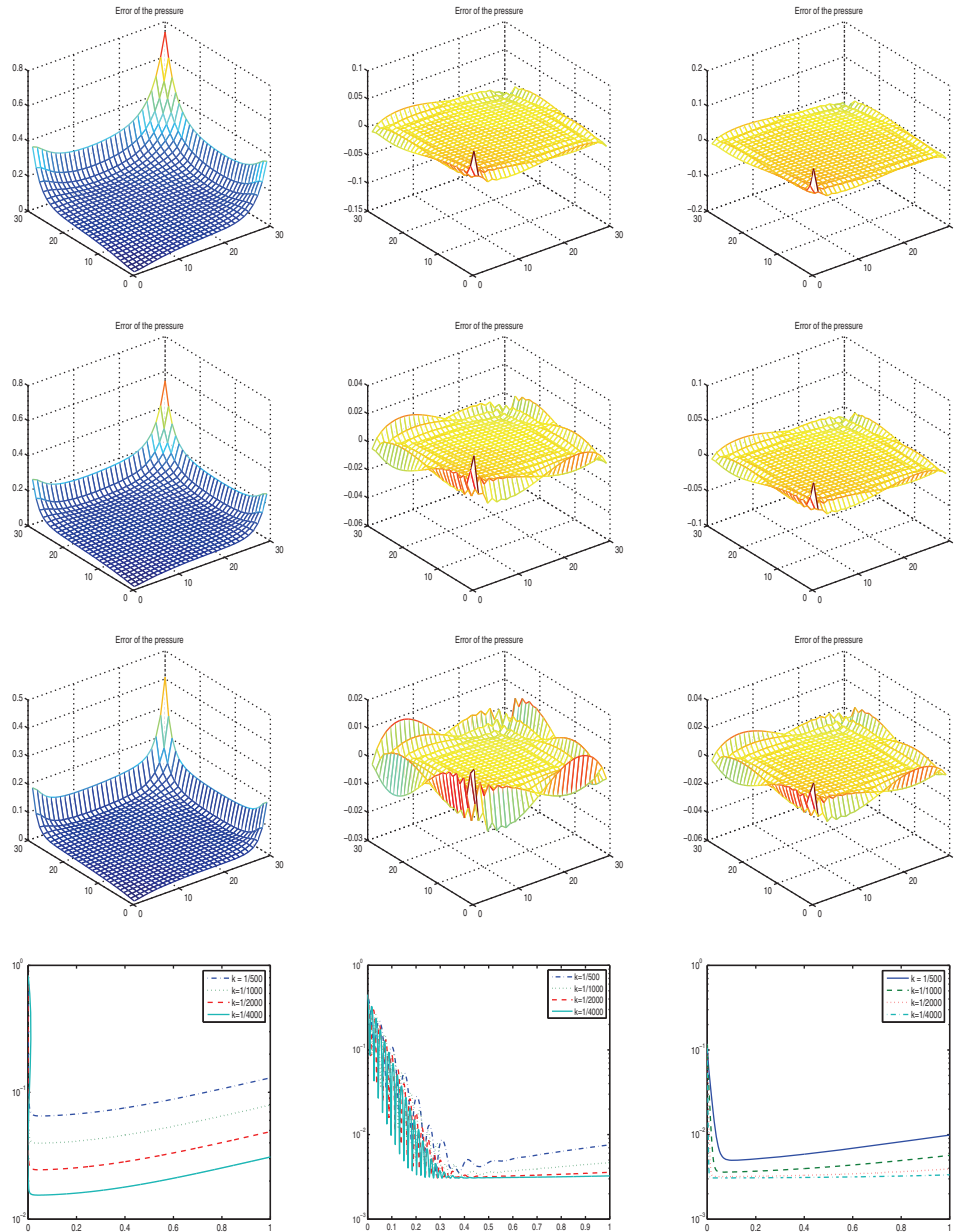


FIG. 1. Example 1: Errors of $p(t_m, \cdot) - p^m$ at time $t_m = 1$, for $k = 1/500$ (1st line), $k = 1/1000$ (2nd line), and $k = 1/2000$ (3rd line), and evolution plot of L^2 errors (4th line), for Chorin (left), Chorin-Uzawa (middle, $\alpha = 0,9$), and Chorin-Penalty (right, $\beta = 1.1$).

Both Chorin-Uzawa ($\alpha \in (0, 1)$) and Chorin-Penalty ($\beta \geq 1$) involve additional parameters; its dependence is studied in Figure 2, and failure of convergence of the Chorin-Penalty method is observed for some $\beta < 1$. The experiments suggest values $\alpha \approx 1$ and $\beta \approx 1$.

Over the last decade, different projection schemes have been developed and tested in academic examples, whose performance crucially relies on given initial functions

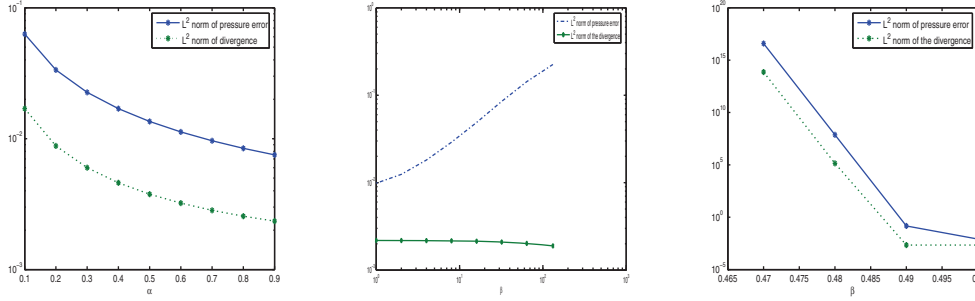


FIG. 2. Example 1: L^2 -errors of pressure at $t = 1$ for (i) Chorin–Uzawa with exact initial pressure, for different $\alpha \in (0, 1)$ (left). (ii) Chorin–Penalty for different $\beta \in (1, 100)$ (middle). (iii) Loss of convergence for Chorin–Penalty for $\beta < \frac{1}{2}$ (right). ($k = 1/500$).

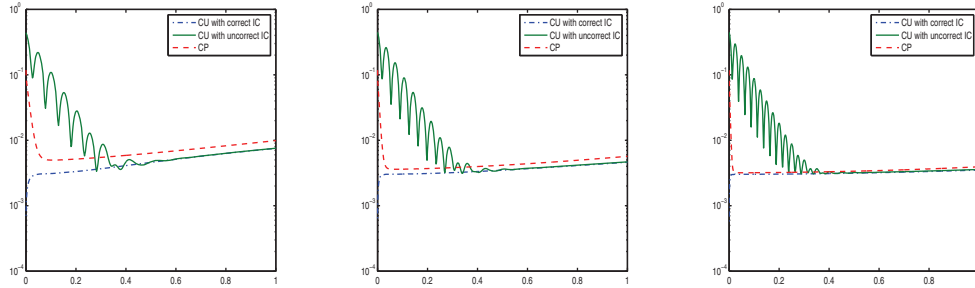


FIG. 3. Example 1: Correct vs. noncorrect initial pressures for Chorin–Uzawa ($\alpha = 0.9$), and comparison with Chorin–Penalty ($\beta = 1.1$) for $k = 1/500$ (left), $k = 1/1000$ (middle), and $k = 1/2000$ (right).

$q \equiv p(0, \cdot)$; cf. [6] for further details. In fact, to compute pressure initial functions in general amounts to solve the following problem (for nonstationary Stokes)

$$\Delta q = \operatorname{div} \mathbf{f}(0, \cdot) \quad \text{in } \Omega, \quad \partial_{\mathbf{n}} q = [\mathbf{f}(0, \cdot) + \nu \Delta \mathbf{u}_0] \cdot \mathbf{n} \quad \text{on } \partial \Omega,$$

where optimal convergence for (finite element) approximations $q_h \approx q$ is not clear. In our case, choosing accurate data $p^0 \approx p(0, \cdot)$ for the Chorin–Uzawa method has been pointed out to be crucial in Theorem 1.3; in contrast, Chorin–Penalty is designed in order to avoid boundary layers for the pressure error in space, and also perform optimally for zero initial pressure data (Theorem 1.4). Figure 3 supports these theoretical results: we observe marked transition layers for Chorin–Uzawa in the case of “noncorrect” initial pressures, while L^2 -errors of the pressure in Chorin–Penalty are almost instantaneously reduced to spatial discretization errors.

5. Conclusion. In recent papers [11, 12], the authors stress the importance to construct and analyze practical projection methods under realistic regularity assumptions—which is also the guideline in this paper. Over the last decade, several projection methods are studied in the literature where (i) smooth solutions to (1.1)–(1.4) and (ii) accurate initial pressure data are assumed, leaving serious doubts on the applicability of these results to more realistic situations of incompatible data and limited solution’s regularity.

Projection methods are efficient methods to approximate strong solutions of the nonstationary incompressible Navier–Stokes equations; the most well-known example

is Chorin's method, which suffers from marked pressure error boundary layers. We give a first rigorous analysis of its structure in the case of existing strong solutions of (1.1)–(1.4) (Theorem 1.2). The new Chorin–Penalty method is proposed, and optimal (i.e., first-order) rate of convergence for the pressure is proved (Theorem 1.4), which reflects uniform, optimal convergence behavior up to the boundary. Comparative computational studies illustrate that the Chorin–Penalty method is exempted from the deficiencies of the Chorin method, in the way that no significant pressure errors arise close to the boundary.

Acknowledgment. The author is grateful to E. Carelli (Universität Tübingen) for providing the computational tests.

REFERENCES

- [1] F. BREZZI AND M. FORTIN, *Mixed and hybrid finite element methods*, Springer, 1991.
- [2] A.J. CHORIN, *Numerical solution of the Navier–Stokes equations*, Math. Comp., 22 (1968), pp. 745–762.
- [3] A.J. CHORIN, *On the convergence of discrete approximations of the Navier–Stokes equations*, Math. Comp., 23 (1969), pp. 341–353.
- [4] W.E AND J.G. LIU, *Projection method I: Convergence and numerical boundary layers*, SIAM J. Numer. Anal., 32 (1995), pp. 1017–1057.
- [5] W.E AND J.G. LIU, *Gauge method for viscous incompressible flows*, Comm. Math. Sci., 1 (2003), pp. 317–332.
- [6] J.L. GUERMOND, P. MINEV, AND J. SHEN, *An overview of projection methods for incompressible flows*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 6011–6045.
- [7] J.L. GUERMOND, J. SHEN, AND X. YANG, *Error analysis of fully discrete velocity-correction methods for incompressible flows*, Math. Comp., 77 (2008), pp. 1387–1405.
- [8] J.G. HEYWOOD AND R. RANNACHER, *Finite element approximation of the nonstationary Navier–Stokes problem. I. Regularity of solutions and second order error estimates for spatial discretization*, SIAM J. Numer. Anal., 19 (1982), pp. 275–311.
- [9] T.J.R. HUGHES, L.P. FRANCA, AND M. BALESTRA, *A new finite element formulation of computational fluid dynamics: V. Circumventing the Babuska–Brezzi condition: A stable Petrov–Galerkin formulation of the Stokes problem accommodating equal-order interpolations*, Comput. Methods Appl. Mech. Engrg., 59 (1986), pp. 85–99.
- [10] J.-G. LIU, J. LIU, AND R. PEGO, *Stability and convergence of efficient Navier–Stokes solvers via a commutator estimate*, Comm. Pure Appl. Math., 60 (2007), pp. 1443–1487.
- [11] R.H. NOCETTO AND J.-H. PYO, *Error estimates for semi-discrete Gauge methods for the Navier–Stokes equations*, Math. Comp., 74 (2004), pp. 521–542.
- [12] R.H. NOCETTO AND J.-H. PYO, *The Gauge–Uzawa finite element method part I: The Navier–Stokes equations*, preprint, 2005.
- [13] A. PROHL, *Projection and Quasi-Compressibility Methods for Solving the incompressible Navier–Stokes Equations*, Teubner-Verlag, Stuttgart, 1997.
- [14] A. PROHL, *A first order projection-based time-splitting scheme for computing chemically reacting flows*, Numer. Math., 84 (2000), pp. 649–477.
- [15] A. PROHL, *On the pollution effect of quasi-compressibility methods in magneto-hydrodynamics and reactive flows*, Math. Methods Appl. Sci., 22 (1999), pp. 1555–1584.
- [16] R. RANNACHER, *On Chorin's projection method for the incompressible Navier–Stokes equations*, in LNM 1530, J.G. Heywood, K. Masuda, R. Rautmann, and S.A. Solonnikov, eds., The Navier–Stokes Equations II – Theory and Numerical Methods, Proc. Oberwolfach, 1991, pp. 167–183.
- [17] J. SHEN, *On error estimates of projection methods for the Navier–Stokes equations: First order schemes*, SIAM J. Numer. Anal., 29 (1992), pp. 57–77.
- [18] J. SHEN, *On error estimates of the penalty method for unsteady Navier–Stokes equations*, SIAM J. Numer. Anal., 32 (1995), pp. 386–403.
- [19] R. TEMAM, *Sur l'approximation de la solution des equations de Navier–Stokes par la methode des pas fractionnaires II*, Arch. Ration. Mech. Anal., 33 (1969), pp. 377–385.
- [20] R. TEMAM, *Navier–Stokes equations — Theory and numerical analysis*, AMS, Providence, RI, 2001.

B-SERIES ANALYSIS OF STOCHASTIC RUNGE–KUTTA METHODS THAT USE AN ITERATIVE SCHEME TO COMPUTE THEIR INTERNAL STAGE VALUES*

KRISTIAN DEBRABANT[†] AND ANNE KVÆRNØ[‡]

Abstract. In recent years, implicit stochastic Runge–Kutta (SRK) methods have been developed both for strong and weak approximations. For these methods, the stage values are only given implicitly. However, in practice these implicit equations are solved by iterative schemes such as simple iteration, modified Newton iteration or full Newton iteration. We employ a unifying approach for the construction of stochastic B-series which is valid both for Itô- and Stratonovich-stochastic differential equations (SDEs) and applicable both for weak and strong convergence to analyze the order of the iterated Runge–Kutta method. Moreover, the analytical techniques applied in this paper can be of use in many other similar contexts.

Key words. stochastic Runge–Kutta method, composite method, stochastic differential equation, iterative scheme, order, internal stage values, Newton’s method, weak approximation, strong approximation, growth functions, stochastic B-series

AMS subject classifications. 65C30, 60H35, 65C20, 68U20

DOI. 10.1137/070704307

1. Introduction. In many applications, e.g., in epidemiology and financial mathematics, taking stochastic effects into account when modeling dynamical systems often leads to stochastic differential equations (SDEs). Therefore, in recent years, the development of numerical methods for the approximation of SDEs has become a field of increasing interest; see, e.g., [16, 22] and references therein. Many stochastic schemes fall into the class of stochastic Runge–Kutta (SRK) methods. SRK methods have been studied both for strong approximation [1, 10, 11, 16], where one is interested in obtaining good pathwise solutions, and for weak approximation [8, 9, 16, 19, 21, 32], which focuses on the expectation of functionals of solutions. Order conditions for these methods are found by comparing series of the exact and the numerical solutions. In this paper, we will concentrate on the use of B-series and rooted trees. Such series are surprisingly general; as formal series they are independent of the choice of the stochastic integral, Itô or Stratonovich, or whether weak or strong convergence is considered. This is demonstrated in section 2. For solving SDEs which are stiff, implicit SRK methods have to be considered, which also has been done both for strong [4, 11, 12] and weak [7, 12, 17] approximation. For these methods, the stage values are only given implicitly. However, in practice these implicit equations are solved by iterative schemes like simple iteration or some kind of Newton iterations. For the numerical solution of ODEs such iterative schemes have been studied [13, 14], and it was shown that certain growth functions defined on trees give a precise description of the development of the iterations. Exactly the same growth functions apply to SRKs, as we prove in section 3. Only when these results

*Received by the editors October 2, 2007; accepted for publication (in revised form) June 9, 2008; published electronically October 29, 2008.

<http://www.siam.org/journals/sinum/47-1/70430.html>

[†]Fachbereich Mathematik, Technische Universität Darmstadt, Schloßgartenstr.7, D-64289 Darmstadt, Germany (debrabant@mathematik.tu-darmstadt.de).

[‡]Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway (anne@math.ntnu.no).

are interpreted in terms of the order of the overall scheme, we distinguish Itô from Stratonovich SDEs, weak from strong convergence. This is discussed in sections 4 and 5.

When considering strong convergence, it is difficult to implement fully implicit SRK methods in combination with Newton iterations due to the possible singularity of the numerical procedure. Therefore, various techniques have been developed to circumvent this problem [4]. One possibility is the use of so-called truncated random variables, which have finite distribution and can approximate the increment of Wiener processes to a chosen order [4, 23]. As the concrete choice of random variables in the numerical methods is not specified in this paper, all considerations are without any change also valid for SRK methods with such modified random variables.

Another possibility is to use composite methods [31], which are combinations of a semi-implicit SRK and an implicit SRK. Based on the results for conventional SRK methods, convergence results for iterated composite methods are given in section 6. Finally, in section 7 we present two numerical examples.

Let $(\Omega, \mathcal{A}, \mathcal{P})$ be a probability space. We denote by $(X(t))_{t \in I}$ the stochastic process which is the solution of either a d -dimensional Itô or Stratonovich SDE defined by

$$(1.1) \quad X(t) = x_0 + \int_{t_0}^t g_0(X(s)) ds + \sum_{l=1}^m \int_{t_0}^t g_l(X(s)) \star dW_l(s)$$

with an m -dimensional Wiener process $(W(t))_{t \geq 0}$ and $I = [t_0, T]$. The integral w.r.t. the Wiener process has to be interpreted either as Itô integral with $\star dW_l(s) = dW_l(s)$ or as Stratonovich integral with $\star dW_l(s) = \circ dW_l(s)$. We assume that the Borel-measurable coefficients $g_l : \mathbb{R}^d \rightarrow \mathbb{R}^d$ are sufficiently differentiable and satisfy a Lipschitz and a linear growth condition. For Stratonovich SDEs, we require in addition that the g_l are differentiable and that also the vectors $g'_l g_l$ satisfy a Lipschitz and a linear growth condition. Then the existence and uniqueness theorem [15] applies.

To simplify the presentation, we define $W_0(s) = s$, so that (1.1) can be written as

$$(1.2) \quad X(t) = x_0 + \sum_{l=0}^m \int_{t_0}^t g_l(X(s)) \star dW_l(s).$$

In the following we denote by Ξ a set of families of measurable mappings,

$$\Xi := \{ \{ \varphi(h) \}_{h \geq 0} : \varphi(h) : \Omega \rightarrow \mathbb{R} \text{ is } \mathcal{A} - \mathcal{B} \text{-measurable } \forall h \geq 0 \},$$

and by Ξ_0 the subset of Ξ defined by

$$\Xi_0 := \{ \{ \varphi(h) \}_{h \geq 0} \in \Xi : \varphi(0) \equiv 0 \}.$$

Let a discretization $I^h = \{t_0, t_1, \dots, t_N\}$ with $t_0 < t_1 < \dots < t_N = T$ of the time interval I with step sizes $h_n = t_{n+1} - t_n$ for $n = 0, 1, \dots, N-1$ be given. Now, we consider the general class of s -stage SRK methods given by $Y_0 = x_0$ and

$$(1.3a) \quad Y_{n+1} = Y_n + \sum_{l=0}^m \sum_{\nu=0}^M \left(z^{(l,\nu)\top} \otimes I_d \right) g_l \left(H^{(l,\nu)} \right)$$

for $n = 0, 1, \dots, N-1$ with $Y_n = Y(t_n)$, $t_n \in I^h$, $I_d \in \mathbb{R}^{d,d}$ the identity matrix, and

$$(1.3b) \quad H^{(l,\nu)} = \mathbb{1}_s \otimes Y_n + \sum_{r=0}^m \sum_{\mu=0}^M \left(Z^{(l,\nu)(r,\mu)} \otimes I_d \right) g_r \left(H^{(r,\mu)} \right)$$

for $l = 0, \dots, m$ and $\nu = 0, \dots, M$ with $\mathbf{1}_s = (1, \dots, 1)^\top \in \mathbb{R}^s$,

$$g_l \left(H^{(l,\nu)} \right) = \left(g_l \left(H_1^{(l,\nu)} \right)^\top, \dots, g_l \left(H_s^{(l,\nu)} \right)^\top \right)^\top$$

and

$$z^{(l,\nu)} \in \Xi_0^s, \quad Z^{(l,\nu)(r,\mu)} \in \Xi_0^{s,s}$$

for $l, r = 0, \dots, m$, $\mu, \nu = 0, \dots, M$.

The formulation (1.3) is a slight generalization of the class considered in [27] and includes the classes of SRK methods considered in [4, 11, 18, 20, 28, 29, 30] as well as the SRK methods considered in [12, 16, 25]. It defines a d -dimensional approximation process Y with $Y_n = Y(t_n)$.

Application of an iterative scheme yields

$$\begin{aligned} H_{k+1}^{(l,\nu)} &= \mathbf{1}_s \otimes Y_n + \sum_{r=0}^m \sum_{\mu=0}^M \left(Z^{(l,\nu)(r,\mu)} \otimes I_d \right) g_r \left(H_k^{(r,\mu)} \right) \\ &+ \sum_{r=0}^m \sum_{\mu=0}^M \left(Z^{(l,\nu)(r,\mu)} \otimes I_d \right) J_k^{(r,\mu)} \left(H_{k+1}^{(r,\mu)} - H_k^{(r,\mu)} \right), \end{aligned} \tag{1.4a}$$

$$Y_{n+1,k} = Y_n + \sum_{l=0}^m \sum_{\nu=0}^M \left(z^{(l,\nu)}^\top \otimes I_d \right) g_l \left(H_k^{(l,\nu)} \right) \tag{1.4b}$$

with some approximation $J_k^{(r,\mu)}$ to the Jacobian of $g_r(H_k^{(r,\mu)})$ and a predictor $H_0^{(l,\nu)}$. In the following we assume that (1.4a) can be solved uniquely at least for small enough h_n . We consider simple iterations with $J_k^{(r,\mu)} = 0$ (i.e., predictor-corrector methods), modified Newton iterations with $J_k^{(r,\mu)} = I_s \otimes g'_r(x_0)$, and full Newton iterations.

2. Some notation, definitions, and preliminary results. In this section we introduce some necessary notation and provide a few definitions and preliminary results that will be used later.

2.1. Convergence and consistency. Here we will give the definitions for both weak and strong convergence and results which relate convergence to consistency.

Let $C_P^l(\mathbb{R}^d, \mathbb{R}^{\hat{d}})$ denote the space of all $g \in C^l(\mathbb{R}^d, \mathbb{R}^{\hat{d}})$ fulfilling a polynomial growth condition [16].

DEFINITION 1. *A time discrete approximation $Y = (Y(t))_{t \in I^h}$ converges weakly with order p to X as $h \rightarrow 0$ at time $t \in I^h$ if for each $f \in C_P^{2(p+1)}(\mathbb{R}^d, \mathbb{R})$ there exist a constant C_f and a finite $\delta_0 > 0$ such that*

$$|\mathbb{E}(f(Y(t))) - \mathbb{E}(f(X(t)))| \leq C_f h^p$$

holds for each $h \in]0, \delta_0[$.

Now, let $le_f(h; t, x)$ be the weak local error of the method starting at the point (t, x) with respect to the functional f and step size h , i.e.,

$$le_f(h; t, x) = \mathbb{E} \left(f(Y(t+h)) - f(X(t+h)) \mid Y(t) = X(t) = x \right).$$

The following theorem due to Milstein [22], which holds also in the case of general one-step methods, shows that, as in the deterministic case, consistency implies convergence.

THEOREM 1. *Suppose the following conditions hold:*

- *The coefficients g_l are continuous, satisfy a Lipschitz condition, and belong to $C_P^{2(p+1)}(\mathbb{R}^d, \mathbb{R}^d)$ for $l = 0, \dots, m$. For Stratonovich SDEs, we require in addition that the g_l are differentiable and that also the vectors $g'_l g_l$ satisfy a Lipschitz condition and belong to $C_P^{2(p+1)}(\mathbb{R}^d, \mathbb{R}^d)$ for $l = 0, \dots, m$.*
- *For sufficiently large r (see, e.g., [22] for details) the moments $\mathbb{E}(\|Y(t_n)\|^{2r})$ exist for $t_n \in I^h$ and are uniformly bounded with respect to N and $n = 0, 1, \dots, N$.*
- *Assume that for all $f \in C_P^{2(p+1)}(\mathbb{R}^d, \mathbb{R})$ there exists a $K \in C_P^0(\mathbb{R}^d, \mathbb{R})$ such that*

$$|le_f(h; t, x)| \leq K(x) h^{p+1}$$

is valid for $x \in \mathbb{R}^d$ and $t, t+h \in I^h$, i.e., the approximation is weak consistent of order p .

Then the method (1.3) is convergent of order p in the sense of weak approximation.

Whereas weak approximation methods are used to estimate the expectation of functionals of the solution, strong approximation methods approach the solution pathwise.

DEFINITION 2. *A time discrete approximation $Y = (Y(t))_{t \in I^h}$ converges strongly, respectively, in the mean square with order p to X as $h \rightarrow 0$ at time $t \in I^h$ if there exists a constant C and a finite $\delta_0 > 0$ such that*

$$\mathbb{E} \|Y(t) - X(t)\| \leq C h^p, \quad \text{respectively,} \quad \sqrt{\mathbb{E}(\|Y(t) - X(t)\|^2)} \leq C h^p$$

holds for each $h \in]0, \delta_0[$.

In this article we will consider convergence in the mean square sense. But as by Jensen's inequality we have

$$(\mathbb{E} \|Y(t) - X(t)\|)^2 \leq \mathbb{E}(\|Y(t) - X(t)\|^2),$$

mean square convergence implies strong convergence of the same order.

Now, let $le^m(h; t, x)$ and $le^{ms}(h; t, x)$, respectively, be the mean and mean square local error, respectively, of the method starting at the point (t, x) with respect to the step size h ; i.e.,

$$\begin{aligned} le^m(h; t, x) &= \mathbb{E} (Y(t+h) - X(t+h) | Y(t) = X(t) = x), \\ le^{ms}(h; t, x) &= \sqrt{\mathbb{E} ((Y(t+h) - X(t+h))^2 | Y(t) = X(t) = x)}. \end{aligned}$$

The following theorem due to Milstein [22] which holds also in the case of general one step methods shows that in the mean square convergence case we obtain order p if the mean local error is consistent of order p and the mean square local error is consistent of order $p - \frac{1}{2}$.

THEOREM 2. *Suppose the following conditions hold:*

- *The coefficients g_l are continuous and satisfy a Lipschitz condition for $l = 0, \dots, m$, and $\mathbb{E}(\|X(t_0)\|^2) < \infty$. For Stratonovich SDEs, we require in addition that the g_l are differentiable and that also the vectors $g'_l g_l$ satisfy a Lipschitz condition.*
- *There exists a constant K independent of h such that*

$$\|le^m(h; t, x)\| \leq K \sqrt{1 + \|x\|^2} h^{p_1}, \quad le^{ms}(h; t, x) \leq K \sqrt{1 + \|x\|^2} h^{p+\frac{1}{2}}$$

with $p \geq 0$, $p_1 \geq p + 1$ is valid for $x \in \mathbb{R}^d$, and $t, t + h \in I^h$; i.e., the approximation is consistent in the mean of order $p_1 - 1 \geq p$ and in the mean square of order $p - \frac{1}{2}$.

Then the SRK method (1.3) is convergent of order p in the sense of mean square approximation.

For Stratonovich SDEs, this result is also obtained by Burrage and Burrage [2].

2.2. Stochastic B-series. In this section we will develop stochastic B-series for the solution of (1.2) as well as for the numerical solution given by (1.3). B-series for deterministic ODEs were introduced by Butcher [6]. Today such series appear as a fundamental tool to do local error analysis on a wide range of problems. B-series for SDEs have been developed by Burrage and Burrage [1, 2] to study strong convergence in the Stratonovich case, by Komori, Mitsui, and Sugiura [20] and Komori [18] to study weak convergence in the Stratonovich case, and by Rößler [26, 27] to study weak convergence in both the Itô and the Stratonovich case. However, the distinction between the Itô and the Stratonovich integrals depends only on the definition of the integrals, not on how the B-series are constructed. Similarly, the distinction between weak and strong convergence depends only on the definition of the local error. Thus, we find it convenient to present a uniform and self-contained theory for the construction of stochastic B-series. We will present results and proofs in a certain detail, since some intermediate results will be used in later sections.

Following the idea of Burrage and Burrage, we introduce the set of colored, rooted trees related to the SDE (1.1), as well as the elementary differentials associated with each of these trees.

DEFINITION 3 (trees). *The set of $m + 1$ -colored, rooted trees*

$$T = \{\emptyset\} \cup T_0 \cup T_1 \cup \dots \cup T_m$$

is recursively defined as follows:

(a) *The graph $\bullet_l = [\emptyset]_l$ with only one vertex of color l belongs to T_l .*

Let $\tau = [\tau_1, \tau_2, \dots, \tau_\kappa]_l$ be the tree formed by joining the subtrees $\tau_1, \tau_2, \dots, \tau_\kappa$ each by a single branch to a common root of color l .

(b) *If $\tau_1, \tau_2, \dots, \tau_\kappa \in T$, then $\tau = [\tau_1, \tau_2, \dots, \tau_\kappa]_l \in T_l$.*

Thus, T_l is the set of trees with an l -colored root, and T is the union of these sets.

DEFINITION 4 (elementary differentials). *For a tree $\tau \in T$ the elementary differential is a mapping $F(\tau) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ defined recursively by*

- (a) $F(\emptyset)(x_0) = x_0$,
- (b) $F(\bullet_l)(x_0) = g_l(x_0)$,
- (c) *If $\tau = [\tau_1, \tau_2, \dots, \tau_\kappa]_l \in T_l$, then*

$$F(\tau)(x_0) = g_l^{(\kappa)}(x_0) (F(\tau_1)(x_0), F(\tau_2)(x_0), \dots, F(\tau_\kappa)(x_0)).$$

As will be shown in the following, both the exact and the numerical solutions, including the iterated solutions as we will see later, can formally be written in terms of B-series.

DEFINITION 5 (B-series). *Given a mapping $\phi : T \rightarrow \Xi$ satisfying*

$$\phi(\emptyset)(h) \equiv 1 \text{ and } \phi(\tau)(0) \equiv 0, \quad \forall \tau \in T \setminus \{\emptyset\}.$$

A (stochastic) B-series is then a formal series of the form

$$B(\phi, x_0; h) = \sum_{\tau \in T} \alpha(\tau) \cdot \phi(\tau)(h) \cdot F(\tau)(x_0),$$

where $\alpha : T \rightarrow \mathbb{Q}$ is given by

$$\alpha(\emptyset) = 1, \quad \alpha(\bullet_l) = 1, \quad \alpha(\tau = [\tau_1, \dots, \tau_\kappa]_l) = \frac{1}{r_1!r_2! \cdots r_q!} \prod_{j=1}^\kappa \alpha(\tau_j),$$

where r_1, r_2, \dots, r_q count equal trees among $\tau_1, \tau_2, \dots, \tau_\kappa$.

If $\phi : T \rightarrow \Xi^s$, then $B(\phi, x_0; h) = [B(\phi_1, x_0; h), \dots, B(\phi_s, x_0; h)]^\top$.

The next lemma proves that if $Y(h)$ can be written as a B-series, then $f(Y(h))$ can be written as a similar series, where the sum is taken over trees with a root of color f and subtrees in T . The lemma is fundamental for deriving B-series for the exact and the numerical solution. It will also be used for deriving weak convergence results.

LEMMA 3. *If $Y(h) = B(\phi, x_0; h)$ is some B-series and $f \in C^\infty(\mathbb{R}^d, \mathbb{R}^{\hat{d}})$, then $f(Y(h))$ can be written as a formal series of the form*

$$(2.1) \quad f(Y(h)) = \sum_{u \in U_f} \beta(u) \cdot \psi_\phi(u)(h) \cdot G(u)(x_0),$$

where U_f is a set of trees derived from T , by

- (a) $[\emptyset]_f \in U_f$, and if $\tau_1, \tau_2, \dots, \tau_\kappa \in T$, then $[\tau_1, \tau_2, \dots, \tau_\kappa]_f \in U_f$.
- (b) $G([\emptyset]_f)(x_0) = f(x_0)$ and $G(u = [\tau_1, \dots, \tau_\kappa]_f)(x_0) = f^{(\kappa)}(x_0)(F(\tau_1)(x_0), \dots, F(\tau_\kappa)(x_0))$.
- (c) $\beta([\emptyset]_f) = 1$ and $\beta(u = [\tau_1, \dots, \tau_\kappa]_f) = \frac{1}{r_1!r_2! \cdots r_q!} \prod_{j=1}^\kappa \alpha(\tau_j)$, where r_1, r_2, \dots, r_q count equal trees among $\tau_1, \tau_2, \dots, \tau_\kappa$.
- (d) $\psi_\phi([\emptyset]_f)(h) \equiv 1$ and $\psi_\phi(u = [\tau_1, \dots, \tau_\kappa]_f)(h) = \prod_{j=1}^\kappa \phi(\tau_j)(h)$.

Proof. Writing $Y(h)$ as a B-series, we have

$$\begin{aligned} f(Y(h)) &= f \left(\sum_{\tau \in T} \alpha(\tau) \cdot \phi(\tau)(h) \cdot F(\tau)(x_0) \right) \\ &= \sum_{\kappa=0}^\infty \frac{1}{\kappa!} f^{(\kappa)}(x_0) \left(\sum_{\tau \in T \setminus \{\emptyset\}} \alpha(\tau) \cdot \phi(\tau)(h) \cdot F(\tau)(x_0) \right)^\kappa \\ &= f(x_0) + \sum_{\kappa=1}^\infty \frac{1}{\kappa!} \sum_{\{\tau_1, \tau_2, \dots, \tau_\kappa\} \in T \setminus \{\emptyset\}} \frac{\kappa!}{r_1!r_2! \cdots r_q!} \\ &\quad \cdot \left(\prod_{j=1}^\kappa \alpha(\tau_j) \cdot \phi(\tau_j)(h) \right) f^{(\kappa)}(x_0)(F(\tau_1)(x_0), \dots, F(\tau_\kappa)(x_0)), \end{aligned}$$

where the last sum is taken over all possible unordered combinations of κ trees in T . For each set of trees $\tau_1, \tau_2, \dots, \tau_\kappa \in T$ we assign a $u = [\tau_1, \tau_2, \dots, \tau_\kappa]_f \in U_f$. The theorem is now proved by comparing term by term with (2.1). \square

Remark 1. For example, in the definition of weak convergence, just $f \in C_P^{2(p+1)}(\mathbb{R}^d, \mathbb{R})$ is required. Thus $f(Y(h))$ can be written only as a truncated B-series, with a remainder term. However, to simplify the presentation in the following we assume that all derivatives of f, g_0, \dots, g_l exist.

We will also need the following result.

LEMMA 4. *If $Y(h) = B(\phi_Y, x_0; h)$ and $Z(h) = B(\phi_Z, x_0; h)$ and $f \in C^\infty(\mathbb{R}^d, \mathbb{R}^{\hat{d}})$, then*

$$f'(Y(h))Z(h) = \sum_{u \in U_f} \beta(u) \cdot \Upsilon(u)(h) \cdot G(u)(x_0)$$

with

$$\Upsilon([\emptyset]_f)(h) \equiv 0, \quad \Upsilon([u = [\tau_1, \dots, \tau_\kappa]_f])(h) = \sum_{i=1}^{\kappa} \left(\prod_{\substack{j=1 \\ j \neq i}}^{\kappa} \phi_Y(\tau_j)(h) \right) \phi_Z(\tau_i)(h)$$

with $\beta(u)$ and $G(u)(x_0)$ given by Lemma 3. The proof is similar to the deterministic case (see [24]).

When Lemma 3 is applied to the functions g_l on the right-hand side of (1.2) we get the following result: If $Y(h) = B(\phi, x_0; h)$, then

$$(2.2) \quad g_l(Y(h)) = \sum_{\tau \in T_l} \alpha(\tau) \cdot \phi'_l(\tau)(h) \cdot F(\tau)(x_0)$$

in which

$$\phi'_l(\tau)(h) = \begin{cases} 1 & \text{if } \tau = \bullet_l, \\ \prod_{j=1}^{\kappa} \phi(\tau_j)(h) & \text{if } \tau = [\tau_1, \dots, \tau_\kappa]_l \in T_l. \end{cases}$$

THEOREM 5. *The solution $X(t_0 + h)$ of (1.2) can be written as a B-series $B(\varphi, x_0; h)$ with*

$$\varphi(\emptyset)(h) \equiv 1, \quad \varphi(\bullet_l)(h) = W_l(h), \quad \varphi(\tau = [\tau_1, \dots, \tau_\kappa]_l)(h) = \int_0^h \prod_{j=1}^{\kappa} \varphi(\tau_j)(s) \star dW_l(s).$$

Proof. Write the exact solution as some B-series $X(t_0 + h) = B(\varphi, x_0; h)$. By (2.2) the SDE (1.2) can be written as

$$\sum_{\tau \in T} \alpha(\tau) \cdot \varphi(\tau)(h) \cdot F(\tau)(x_0) = x_0 + \sum_{l=0}^m \sum_{\tau \in T_l} \alpha(\tau) \cdot \int_0^h \varphi'_l(\tau)(s) \star dW_l(s) \cdot F(\tau)(x_0).$$

Comparing term by term we get

$$\varphi(\emptyset)(h) \equiv 1, \quad \text{and} \quad \varphi(\tau)(h) = \int_0^h \varphi'(\tau)(s) \star dW_l(s) \quad \text{for } \tau \in T_l, \quad l = 0, 1, \dots, m.$$

The proof is completed by induction on τ . \square

The same result is given for the Stratonovich case in [2, 18], but it clearly also applies to the Itô case.

The definition of the order of the tree, $\rho(\tau)$, is motivated by the fact that $E W_l(h)^2 = h$ for $l \geq 1$ and $W_0(h) = h$.

DEFINITION 6 (order). *The order of a tree $\tau \in T$ is defined by*



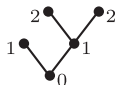
$$\rho(\emptyset) = 0, \quad \rho([\tau_1, \dots, \tau_\kappa]_f) = \sum_{i=1}^{\kappa} \rho(\tau_i)$$

and

$$\rho(\tau = [\tau_1, \dots, \tau_\kappa]_l) = \sum_{i=1}^{\kappa} \rho(\tau_i) + \begin{cases} 1 & \text{for } l = 0, \\ \frac{1}{2} & \text{otherwise.} \end{cases}$$

TABLE 2.1

Examples of trees and corresponding functions $\rho(\tau)$, $\alpha(\tau)$, and $\varphi(\tau)$. The integrals $\varphi(\tau)$ are also expressed in terms of multiple integrals $J_{(\dots)}$ for the Stratonovich (S) and $I_{(\dots)}$ for the Itô (I) cases; see [16] for their definition. In bracket notation, the trees will be written as \bullet_l , $[[\bullet_2]_0]_1$, $[\bullet_1, \bullet_1]_0$, and $[\bullet_1, [\bullet_2, \bullet_2]_1]_0$, respectively.

τ	$\rho(\tau)$	$\alpha(\tau)$	$\varphi(\tau)(h)$
\bullet_l	$\begin{cases} 1 & \text{if } l = 0 \\ \frac{1}{2} & \text{if } l \neq 0 \end{cases}$	1	$W_l(h) = \begin{cases} h & \text{if } l = 0 \\ J_{(l)} & \text{(S)} \\ I_{(l)} & \text{(I)} \end{cases}$
	2	1	$\int_0^h \int_0^{s_1} W_2(s_2) \star ds_2 \star dW_1(s_1) = \begin{cases} J_{(2,0,1)} & \text{(S)} \\ I_{(2,0,1)} & \text{(I)} \end{cases}$
	2	$\frac{1}{2}$	$\int_0^h W_1(s)^2 \star ds = \begin{cases} 2J_{(1,1,0)} & \text{(S)} \\ 2I_{(1,1,0)} + I_{(0,0)} & \text{(I)} \end{cases}$
	3	$\frac{1}{2}$	$\int_0^h W_1(s_1) (\int_0^{s_1} W_2(s_2)^2 \star dW_1(s_2)) \star ds_1$ $= \begin{cases} 4J_{(2,2,1,1,0)} + 2J_{(2,1,2,1,0)} + 2J_{(1,2,2,1,0)} & \text{(S)} \\ 4I_{(2,2,1,1,0)} + 2I_{(2,1,2,1,0)} + 2I_{(1,2,2,1,0)} \\ + 2I_{(0,1,1,0)} + 2I_{(2,2,0,0)} + I_{(1,0,1,0)} + I_{(0,0,0)} & \text{(I)} \end{cases}$

In Table 2.1 some trees and the corresponding values for the functions ρ , α , and φ are presented.

The following result is similar to results given in [1].

THEOREM 6. *If the coefficients $Z^{(l,\nu),(r,\mu)} \in \Xi_0^{s,s}$ and $z^{(l,\nu)} \in \Xi_0^s$, then the numerical solution Y_1 as well as the stage values can be written in terms of B-series*

$$H^{(l,\nu)} = B\left(\Phi^{(l,\nu)}, x_0; h\right), \quad Y_1 = B(\Phi, x_0; h)$$

for all l, ν , with

$$(2.3a) \quad \Phi^{(l,\nu)}(\emptyset)(h) \equiv \mathbb{1}_s, \quad \Phi^{(l,\nu)}(\bullet_r)(h) = \sum_{\mu=0}^M Z^{(l,\nu)(r,\mu)} \mathbb{1}_s,$$

$$(2.3b) \quad \Phi^{(l,\nu)}(\tau = [\tau_1, \dots, \tau_\kappa]_r)(h) = \sum_{\mu=0}^M Z^{(l,\nu)(r,\mu)} \prod_{j=1}^{\kappa} \Phi^{(r,\mu)}(\tau_j)(h)$$

and

$$(2.4a) \quad \Phi(\emptyset)(h) \equiv 1, \quad \Phi(\bullet_l)(h) = \sum_{\nu=0}^M z^{(l,\nu)} \mathbb{1}_s,$$

$$(2.4b) \quad \Phi(\tau = [\tau_1, \dots, \tau_\kappa]_l)(h) = \sum_{\nu=0}^M z^{(l,\nu)} \prod_{j=1}^{\kappa} \Phi^{(l,\nu)}(\tau_j)(h).$$

Proof. Write $H^{(l,\nu)}$ as a B-series, that is

$$H^{(l,\nu)} = \sum_{\tau \in T} \alpha(\tau) \left(\Phi^{(l,\nu)}(h) \otimes I_d \right) (\mathbb{1}_s \otimes F(\tau)(x_0)).$$

Use the definition of the method (1.3) together with (2.2) to obtain

$$H^{(l,\nu)} = \mathbb{1}_s \otimes x_0 + \sum_{r=0}^m \sum_{\mu=0}^M \sum_{\tau \in T_r} \alpha(\tau) \left(\left(Z^{(l,\nu)(r,\mu)} \cdot \left(\Phi^{(r,\mu)} \right)'_r(\tau)(h) \right) \otimes I_d \right) (\mathbb{1}_s \otimes F(\tau)(x_0))$$

with $(\Phi^{(r,\mu)})'_r(\tau)(h) = ((\Phi_1^{(r,\mu)})'_r(\tau)(h), \dots, (\Phi_s^{(r,\mu)})'_r(\tau)(h))^\top$. Comparing term-by-term gives the relations (2.3). The proof of (2.4) is similar. \square

To decide the weak order we will also need the B-series of the function f , evaluated at the exact and the numerical solution. From Theorem 5, Theorem 6, and Lemma 3 we obtain

$$f(X(t_0 + h)) = \sum_{u \in U_f} \beta(u) \cdot \psi_\varphi(u)(h) \cdot G(u)(x_0),$$

$$f(Y_1) = \sum_{u \in U_f} \beta(u) \cdot \psi_\Phi(u)(h) \cdot G(u)(x_0),$$

with

$$\psi_\varphi([\emptyset]_f)(h) \equiv 1, \quad \psi_\varphi(u = [\tau_1, \dots, \tau_\kappa]_f)(h) = \prod_{j=1}^\kappa \varphi(\tau_j)(h)$$

and

$$\psi_\Phi([\emptyset]_f)(h) \equiv 1, \quad \psi_\Phi(u = [\tau_1, \dots, \tau_\kappa]_f)(h) = \prod_{j=1}^\kappa \Phi(\tau_j)(h).$$

So, for the weak local error it follows

$$le_f(h; t, x) = \sum_{u \in U_f} \beta(u) \cdot \mathbb{E} [\psi_\Phi(u)(h) - \psi_\varphi(u)(h)] \cdot G(u)(x).$$

For the mean and mean square local error we obtain from Theorem 5 and Theorem 6,

$$le^{ms}(h; t, x) = \sqrt{\mathbb{E} \left(\sum_{\tau \in T} \alpha(\tau) \cdot (\Phi(\tau)(h) - \varphi(\tau)(h)) \cdot F(\tau)(x) \right)^2},$$

$$le^m(h; t, x) = \sum_{\tau \in T} \alpha(\tau) \cdot \mathbb{E} (\Phi(\tau)(h) - \varphi(\tau)(h)) \cdot F(\tau)(x).$$

With all the B-series in place, we can now present the order conditions for the weak and strong convergence for both the Itô and the Stratonovich case.¹ We have weak

¹As usual we assume that method (1.3) is constructed such that $E\psi_\varphi(u)(h) = \mathcal{O}(h^{\rho(u)}) \forall u \in U_f$ and $\varphi(\tau)(h) = \mathcal{O}(h^{\rho(\tau)}) \forall \tau \in T$, respectively, where especially in the latter expression the $\mathcal{O}(\cdot)$ -notation refers to the $L^2(\Omega)$ -norm and $h \rightarrow 0$. These conditions are fulfilled if for $i, j = 1, \dots, s$, $k \in \mathbb{N} = \{0, 1, \dots\}$ it holds that

$$(z_i^{(l,\nu)})^{2^k} = \begin{cases} \mathcal{O}(h^{(2^k)}) & l = 0 \\ \mathcal{O}(h^{(2^{k-1})}) & l > 0 \end{cases}, \quad (Z_{ij}^{(l,\nu)(r,\mu)})^{2^k} = \begin{cases} \mathcal{O}(h^{(2^k)}) & l = 0 \\ \mathcal{O}(h^{(2^{k-1})}) & l > 0 \end{cases}.$$

consistency of order p if and only if

$$(2.5) \quad \mathbb{E} \psi_{\Phi}(u)(h) = \mathbb{E} \psi_{\varphi}(u)(h) + \mathcal{O}(h^{p+1}) \quad \forall u \in U_f \text{ with } \rho(u) \leq p + \frac{1}{2},$$

where $\rho(u = [\tau_1, \dots, \tau_{\kappa}]_f) = \sum_{j=1}^{\kappa} \rho(\tau_j)$ ((2.5) slightly weakens conditions given in [27]), and mean square global order p if [4]

$$(2.6) \quad \Phi(\tau)(h) = \varphi(\tau)(h) + \mathcal{O}\left(h^{p+\frac{1}{2}}\right) \quad \forall \tau \in T \text{ with } \rho(\tau) \leq p,$$

$$(2.7) \quad \mathbb{E} \Phi(\tau)(h) = \mathbb{E} \varphi(\tau)(h) + \mathcal{O}(h^{p+1}) \quad \forall \tau \in T \text{ with } \rho(\tau) \leq p + \frac{1}{2}$$

and all elementary differentials $F(\tau)$ fulfill a linear growth condition. Instead of the last requirement it is also enough to claim that there exists a constant C such that $\|g'_j(y)\| \leq C \quad \forall y \in \mathbb{R}^m, j = 0, \dots, M$ (which implies the global Lipschitz condition) and all necessary partial derivatives exist [2].

3. B-series of the iterated solution and growth functions. In this section we will discuss how the iterated solution defined in (1.4) can be written in terms of B-series, that is,

$$H_k^{(l,\nu)} = B\left(\Phi_k^{(l,\nu)}, x_0; h\right) \quad \text{and} \quad Y_{1,k} = B(\Phi_k, x_0; h).$$

For notational convenience, in the following the h -dependency of the weight functions will be suppressed, so $\Phi(\tau)(h)$ will be written as $\Phi(\tau)$. Further, all results are valid for all $l = 0, \dots, m$ and $\nu = 0, \dots, M$.

Assume that the predictor can be written as a B-series,

$$H_0^{(l,\nu)} = B\left(\Phi_0^{(l,\nu)}, x_0; h\right),$$

satisfying $\Phi_0^{(l,\nu)}(\emptyset) = \mathbb{1}_s$ and $\Phi_0^{(l,\nu)}(\tau) = \mathcal{O}(h^{\rho(\tau)}) \forall \tau \in T$. The most common situation is the use of the trivial predictor $H^{(l,\nu)} = \mathbb{1}_s \otimes x_0$, for which $\Phi_0^{(l,\nu)}(\emptyset) = \mathbb{1}_s$ and $\Phi_0^{(l,\nu)}(\tau) = 0$ otherwise.

The iteration schemes we discuss here differ only in the choice of $J_k^{(r,\mu)}$ in (1.4). For all schemes, the following lemma applies. The proof follows directly from Lemma 3.

LEMMA 7. *If $H_k^{(l,\nu)} = B(\Phi_k^{(l,\nu)}, x_0; h)$, then $Y_{1,k} = B(\Phi_k, x_0; h)$ with*

$$\Phi_k(\emptyset) \equiv \mathbb{1}, \quad \Phi_k(\bullet_l) = \sum_{\nu=0}^M z^{(l,\nu)} \mathbb{1}_s, \quad \Phi_k(\tau = [\tau_1, \dots, \tau_{\kappa}]_l) = \sum_{\nu=0}^M z^{(l,\nu)} \prod_{j=1}^{\kappa} \Phi_k^{(l,\nu)}(\tau_j).$$

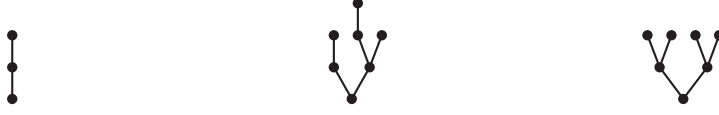
Further,

$$f(Y_{1,k}) = \sum_{u \in U_f} \beta(u) \cdot \psi_{\Phi_k}(u) \cdot G(u)(x_0)$$

with

$$\psi_{\Phi_k}([\emptyset]_f) = 1, \quad \psi_{\Phi_k}(u = [\tau_1, \dots, \tau_{\kappa}]_f) = \prod_{j=1}^{\kappa} \Phi_k(\tau_j),$$

where $\beta(u)$ and $G(u)(x_0)$ are given in Lemma 3.



$$\mathfrak{h}(\tau) = 3, \mathfrak{r}(\tau) = \mathfrak{d}(\tau) = 1; \quad \mathfrak{h}(\tau) = 4, \mathfrak{r}(\tau) = 3, \mathfrak{d}(\tau) = 2; \quad \mathfrak{h}(\tau) = \mathfrak{r}(\tau) = \mathfrak{d}(\tau) = 3$$

FIG. 3.1. Examples of trees and their growth functions for simple (\mathfrak{h}), modified Newton (\mathfrak{r}), and full Newton (\mathfrak{d}) iterations.

We are now ready to study each of the iteration schemes. In each case, we will first find the recurrence formula for $\Phi_k^{(l,\nu)}(\tau)$. From this we define a growth function $\mathfrak{g}(\tau)$.

DEFINITION 7 (growth function). A growth function $\mathfrak{g} : T \rightarrow \mathbb{N}$ is a function satisfying

$$(3.1) \quad \begin{aligned} \Phi_k^{(l,\nu)}(\tau) &= \Phi^{(l,\nu)}(\tau), \quad \forall \tau \in T, \quad \mathfrak{g}(\tau) \leq k \\ \Rightarrow \quad \Phi_{k+1}^{(l,\nu)}(\tau) &= \Phi^{(l,\nu)}(\tau), \quad \forall \tau \in T, \quad \mathfrak{g}(\tau) \leq k + 1, \end{aligned}$$

for all $k \geq 0$.

This result should be sharp in the sense that in general there exists $\tau \neq \emptyset$ with $\Phi_0^{(l,\nu)}(\tau) \neq \Phi^{(l,\nu)}(\tau)$ and $\Phi_k^{(l,\nu)}(\tau) \neq \Phi^{(l,\nu)}(\tau)$ when $k < \mathfrak{g}(\tau)$. From Lemma 7 we also have

$$(3.2) \quad \begin{aligned} \Phi_k(\tau) &= \Phi(\tau) & \forall \tau = [\tau_1, \dots, \tau_\kappa]_l \in T, & \quad \mathfrak{g}'(\tau) = \max_{j=1}^{\kappa} \mathfrak{g}(\tau_j) \leq k, \\ \psi_{\Phi_k}(u) &= \psi_{\Phi(\tau)} & \forall u = [\tau_1, \dots, \tau_\kappa]_f \in U_f, & \quad \mathfrak{g}'(u) = \max_{j=1}^{\kappa} \mathfrak{g}'(\tau_j) \leq k. \end{aligned}$$

The growth functions give a precise description of the development of the iterations. As we will see, the growth functions are exactly the same as in the deterministic case (see [13, 14]). However, to get applicable results, we will at the end need the relation between the growth functions and the order. Further, we will also take advantage of the fact that $E \Phi(\tau) = 0$ and $E \psi_{\Phi}(u) = 0$ for some trees. These aspects are discussed in the next sections. Examples of trees and the values of the growth functions for the three iteration schemes are given in Figure 3.1.

The simple iteration. Simple iterations are described by (1.4a) with $J_k^{(r,\mu)} = 0$, that is,

$$(3.3) \quad H_{k+1}^{(l,\nu)} = \mathbb{1}_s \otimes x_0 + \sum_{r=0}^m \sum_{\mu=0}^M \left(Z^{(l,\nu)(r,\mu)} \otimes I_d \right) g_r \left(H_k^{(r,\mu)} \right).$$

By (2.2) and Theorem 6 we easily get the next two results.

LEMMA 8. If $H_0^{(l,\nu)} = B(\Phi_0^{(l,\nu)}, x_0; h)$, then $H_k^{(l,\nu)} = B(\Phi_k^{(l,\nu)}, x_0; h)$, where

$$\Phi_{k+1}^{(l,\nu)}(\emptyset) \equiv \mathbb{1}_s, \quad \Phi_{k+1}^{(l,\nu)}(\tau = [\tau_1, \dots, \tau_\kappa]_r) = \sum_{\mu=0}^M Z^{(l,\nu)(r,\mu)} \prod_{j=1}^{\kappa} \Phi_k^{(r,\mu)}(\tau_j).$$

The corresponding growth function is given by

$$\mathfrak{h}(\emptyset) = 0, \quad \mathfrak{h}([\tau_1, \dots, \tau_\kappa]_l) = 1 + \max_{j=1}^{\kappa} \mathfrak{h}(\tau_j).$$

The function $\mathfrak{h}(\tau)$ is the height of τ , that is, the maximum number of nodes along one branch. The functions $\mathfrak{h}'(\tau)$ and $\mathfrak{h}'(u)$ are defined by (3.2), with \mathfrak{g} replaced by \mathfrak{h} .

The modified Newton iteration. In this subsection, we consider the modified Newton iteration (1.4a) with $J_k^{(r,\mu)} = I_s \otimes g'_r(x_0)$. The B-series for $H_k^{(l,\nu)}$ and the corresponding growth function can now be described by the following lemma.

LEMMA 9. *If $H_0^{(l,\nu)} = B(\Phi_0^{(l,\nu)}, x_0; h)$, then $H_k^{(l,\nu)} = B(\Phi_k^{(l,\nu)}, x_0; h)$ with*

$$(3.4) \quad \begin{aligned} \Phi_{k+1}^{(l,\nu)}(\emptyset) &\equiv \mathbb{1}_s, \\ \Phi_{k+1}^{(l,\nu)}(\tau) &= \begin{cases} \sum_{\mu=0}^M Z^{(l,\nu)(r,\mu)} \prod_{j=1}^{\kappa} \Phi_k^{(r,\mu)}(\tau_j) & \text{if } \tau = [\tau_1, \dots, \tau_\kappa]_r \in T \text{ and } \kappa \geq 2, \\ \sum_{\mu=0}^M Z^{(l,\nu)(r,\mu)} \Phi_{k+1}^{(r,\mu)}(\tau_1) & \text{if } \tau = [\tau_1]_r \in T. \end{cases} \end{aligned}$$

The corresponding growth function is given by

$$\mathfrak{r}(\emptyset) = 0, \quad \mathfrak{r}(\bullet_r) = 1, \quad \mathfrak{r}(\tau = [\tau_1, \dots, \tau_\kappa]_l) = \begin{cases} \mathfrak{r}(\tau_1) & \text{if } \kappa = 1, \\ 1 + \max_{j=1}^{\kappa} \mathfrak{r}(\tau_j) & \text{if } \kappa \geq 2. \end{cases}$$

The function $\mathfrak{r}(\tau)$ is one plus the maximum number of ramifications along any branch of the tree.

Proof. The iteration scheme (1.4a) can be rewritten in B-series notation as

$$(3.5) \quad \begin{aligned} \sum_{\tau \in T} \alpha(\tau) \cdot \Phi_{k+1}^{(l,\nu)}(\tau) \otimes F(\tau)(x_0) &= \mathbb{1} \otimes x_0 \\ + \sum_{r=0}^m \sum_{\tau \in T_r} \alpha(\tau) \cdot \left(\sum_{\mu=0}^M Z^{(l,\nu)(r,\mu)} \left(\Phi_k^{(r,\mu)} \right)'_r(\tau) \right) \otimes F(\tau)(x_0) \\ + \sum_{r=0}^m \sum_{\tau_1 \in T_r} \alpha(\tau_1) \cdot \left(\sum_{\mu=0}^M Z^{(l,\nu)(r,\mu)} \left(\Phi_{k+1}^{(r,\mu)}(\tau_1) - \Phi_k^{(r,\mu)}(\tau_1) \right) \right) \otimes (g'_r(x_0)F(\tau_1)(x_0)), \end{aligned}$$

where we have used (2.2). Clearly, $\Phi_{k+1}^{(l,\nu)}(\emptyset) \equiv \mathbb{1}_s$ for all $k \geq 0$ and

$$\Phi_{k+1}^{(l,\nu)}(\bullet_r) = \sum_{\mu=0}^M Z^{(l,\nu)(r,\mu)} \mathbb{1}_s,$$

proving the lemma for $\tau = \bullet_r = [\emptyset]_r$. Next, let $\tau = [\tau_1]_r$, where $\tau_1 \neq \emptyset$. Then $F(\tau)(x_0) = g'_r(x_0)F(\tau_1)$. Comparing equal terms on both sides of the equation, using $\alpha(\tau) = \alpha(\tau_1)$, we get

$$\Phi_{k+1}^{(l,\nu)}(\tau) = \sum_{\mu=0}^M Z^{(l,\nu)(r,\mu)} \left(\left(\Phi_k^{(r,\mu)} \right)'_r(\tau) + \Phi_{k+1}^{(r,\mu)}(\tau_1) - \Phi_k^{(r,\mu)}(\tau_1) \right).$$

Since $\left(\Phi_k^{(r,\mu)} \right)'_r(\tau) = \Phi_k^{(r,\mu)}(\tau_1)$ the lemma is proved for all $\tau = [\tau_1]_r$. For $\tau = [\tau_1, \dots, \tau_\kappa]_r$ with $\kappa \geq 2$ the last sum of (3.5) contributes nothing, thus

$$\Phi_{k+1}^{(l,\nu)}(\tau) = \sum_{\mu=0}^M Z^{(l,\nu)(r,\mu)} \left(\Phi_k^{(r,\mu)} \right)'_r(\tau),$$

which concludes the proof of (3.4).

The second statement of the lemma is obviously true for $\tau = \emptyset$. Let τ be any tree satisfying $\mathfrak{r}(\tau) \leq k + 1$. Then either $\tau = [\tau_1]_l$ with $\mathfrak{r}(\tau_1) \leq k + 1$ or $\tau = [\tau_1, \dots, \tau_\kappa]_l$ with $\kappa \geq 2$ and $\mathfrak{r}(\tau_i) \leq k$. In the latter case, we have by the hypothesis, by (3.4) and Theorem 6, that

$$\Phi_{k+1}^{(l,\nu)}(\tau) = \sum_{\mu=0}^M Z^{(l,\nu)(r,\mu)} \prod_{j=1}^{\kappa} \Phi^{(r,\mu)}(\tau_j) = \Phi^{(l,\nu)}(\tau).$$

In the first case, it follows easily by induction on τ that $\Phi_{k+1}^{(l,\nu)}(\tau) = \Phi^{(l,\nu)}(\tau)$ since $\Phi_{k+1}^{(l,\nu)}(\tau) = \sum_{\mu=0}^M Z^{(l,\nu)(r,\mu)} \Phi_{k+1}^{(r,\mu)}(\tau_1)$. \square

The full Newton iteration. In this subsection, we consider the full Newton iteration (1.4a) with

$$J_k^{(r,\mu)} = g_r' \left(H_k^{(r,\mu)} \right).$$

It follows that the B-series for $H_k^{(l,\nu)}$ and the corresponding growth function satisfy.

LEMMA 10. *If $H_0^{(l,\nu)} = B(\Phi_0^{(l,\nu)}, x_0; h)$, then $H_k^{(l,\nu)} = B(\Phi_k^{(l,\nu)}, x_0; h)$ with*

$$\begin{aligned} \Phi_{k+1}^{(l,\nu)}(\emptyset) &\equiv \mathbb{1}_s, \\ \Phi_{k+1}^{(l,\nu)}(\tau) &= \sum_{\mu=0}^M Z^{(l,\nu)(r,\mu)} \prod_{j=1}^{\kappa} \Phi_k^{(r,\mu)}(\tau_j) \\ &+ \sum_{\mu=0}^M Z^{(l,\nu)(r,\mu)} \sum_{i=1}^{\kappa} \left(\prod_{\substack{j=1 \\ j \neq i}}^{\kappa} \Phi_k^{(r,\mu)}(\tau_j) \right) \left(\Phi_{k+1}^{(r,\mu)}(\tau_i) - \Phi_k^{(r,\mu)}(\tau_i) \right), \end{aligned} \tag{3.6}$$

where $\tau = [\tau_1, \dots, \tau_\kappa]_r$ and the rightmost \prod is taken to be $\mathbb{1}_s$ if $\kappa = 1$. The corresponding growth function is given by

$$\begin{aligned} \mathfrak{d}(\emptyset) &= 0, & \mathfrak{d}(\bullet_l) &= 1, \\ \mathfrak{d}(\tau = [\tau_1, \dots, \tau_\kappa]_l) &= \begin{cases} \max_{j=1}^{\kappa} \mathfrak{d}(\tau_j) & \text{if } \gamma = 1, \\ \max_{j=1}^{\kappa} \mathfrak{d}(\tau_j) + 1 & \text{if } \gamma \geq 2, \end{cases} \end{aligned}$$

where γ is the number of subtrees in τ satisfying $\mathfrak{d}(\tau_i) = \max_{j=1}^{\kappa} \mathfrak{d}(\tau_j)$.

The function \mathfrak{d} is called the doubling index of τ .

Proof. Using (2.2) and Lemma 4 the scheme (1.4a) can be written as

$$\begin{aligned} \sum_{\tau \in \mathcal{T}} \alpha(\tau) \cdot \Phi_{k+1}^{(l,\nu)}(\tau) \otimes F(\tau)(x_0) &= \mathbb{1} \otimes x_0 \\ &+ \sum_{r=0}^m \sum_{\tau \in \mathcal{T}_r} \alpha(\tau) \cdot \left(\sum_{\mu=0}^M Z^{(l,\nu)(r,\mu)} \left(\Phi_k^{(r,\mu)} \right)'_r(\tau) \right) \otimes F(\tau)(x_0) \\ &+ \sum_{r=0}^m \sum_{u \in \mathcal{U}_{g_r}} \beta(u) \cdot \left(\sum_{\mu=0}^M Z^{(l,\nu)(r,\mu)} \Upsilon_k^{(r,\mu)}(u) \right) \otimes G(u)(x_0), \end{aligned} \tag{3.7}$$

where

$$\Upsilon_k^{(r,\mu)}(u = [\tau_1, \dots, \tau_\kappa]_{g_r}) = \sum_{i=1}^{\kappa} \left(\prod_{\substack{j=1 \\ j \neq i}}^{\kappa} \Phi_k^{(r,\mu)}(\tau_j) \right) \left(\Phi_{k+1}^{(r,\mu)}(\tau_i) - \Phi_k^{(r,\mu)}(\tau_i) \right).$$

From the definition of $F(\tau)$, $G(u = [\tau_1, \dots, \tau_\kappa]_{g_r})(x_0) = F(\tau = [\tau_1, \dots, \tau_\kappa]_r)(x_0)$. The sum over all $u \in U_{g_r}$ can be replaced by the sum over all $\tau \in T_r$, and the result is proved. Next, we will prove that $\mathfrak{d}(\tau)$ satisfies the implication (3.1) given in Definition 7. We will do so by induction on $n(\tau)$, the number of nodes in τ . Since \emptyset is the only tree satisfying $n(\tau) = 0$, and $\Phi_{k+1}^{(r,\mu)}(\emptyset) = \Phi^{(r,\mu)}(\emptyset) \equiv \mathbf{1}_s$, the conclusion of (3.1) is true for all $\tau \in T$ with $n(\tau) = 0$. Let $\bar{n} \geq 1$ and assume by the induction hypothesis that the conclusion of (3.1) holds for any tree satisfying $\mathfrak{d}(\tau) \leq k+1$ and $n(\tau) < \bar{n}$. We will show that $\Phi_{k+1}^{(r,\mu)}(\bar{\tau}) = \Phi^{(r,\mu)}(\bar{\tau})$ for all $\bar{\tau}$ satisfying $\mathfrak{d}(\bar{\tau}) \leq k+1$ and $n(\bar{\tau}) \leq \bar{n}$. Let $\bar{\tau} = [\tau_1, \dots, \tau_\kappa]_l$ where $n(\tau_j) < \bar{n}$ for $j = 1, \dots, \kappa$. Since $\mathfrak{d}(\bar{\tau}) \leq k+1$ there is at most one subtree τ_j satisfying $\mathfrak{d}(\tau_j) = k+1$, let us for simplicity assume this to be the last one. Thus $\mathfrak{d}(\tau_j) \leq k$ for $j = 1, \dots, \kappa-1$ and $\mathfrak{d}(\tau_\kappa) \leq k+1$. Consequently, $\Phi_k^{(r,\mu)}(\tau_j) = \Phi^{(r,\mu)}(\tau_j)$, $j = 1, \dots, \kappa-1$ by the hypothesis of (3.1), and $\Phi_{k+1}^{(r,\mu)}(\tau_j) = \Phi^{(r,\mu)}(\tau_j)$, $j = 1, \dots, \kappa$ by the induction hypothesis. By (3.6) and Theorem 6,

$$\begin{aligned} \Phi_{k+1}^{(l,\nu)}(\bar{\tau}) &= \sum_{\mu=0}^M Z^{(l,\nu)(r,\mu)} \prod_{j=1}^{\kappa} \Phi_k^{(r,\mu)}(\tau_j) \\ &\quad + \sum_{\mu=0}^M \sum_{i=1}^{\kappa} Z^{(l,\nu)(r,\mu)} \left(\prod_{\substack{j=1 \\ j \neq i}}^{\kappa} \Phi_k^{(r,\mu)}(\tau_j) \right) \left(\Phi_{k+1}^{(r,\mu)}(\tau_i) - \Phi_k^{(r,\mu)}(\tau_i) \right) \\ &= \sum_{\mu=0}^M Z^{(l,\nu)(r,\mu)} \left(\prod_{j=1}^{\kappa-1} \Phi_k^{(r,\mu)}(\tau_j) \right) \Phi_k^{(r,\mu)}(\tau_\kappa) \\ &\quad + \sum_{\mu=0}^M Z^{(l,\nu)(r,\mu)} \left(\prod_{j=1}^{\kappa-1} \Phi^{(r,\mu)}(\tau_j) \right) \left(\Phi^{(r,\mu)}(\tau_\kappa) - \Phi_k^{(r,\mu)}(\tau_\kappa) \right) \\ &= \Phi^{(l,\nu)}(\bar{\tau}), \end{aligned}$$

completing the induction proof. \square

4. General convergence results for iterated methods. Now we will relate the results of the previous section to the order of the overall scheme. In the following, we assume that the predictors satisfy the conditions

$$(4.1) \quad \begin{aligned} \Phi_0^{(l,\nu)}(\tau) &= \Phi^{(l,\nu)}(\tau) & \forall \tau \in T \text{ with } \mathfrak{g}(\tau) \leq \mathcal{G}_0, \\ \Phi_0^{(l,\nu)}(\tau) &\in \left\{ \Phi^{(l,\nu)}(\tau), 0 \right\} & \forall \tau \in T \text{ with } \mathfrak{g}(\tau) \leq \hat{\mathcal{G}}_0, \end{aligned}$$

where \mathcal{G}_0 and $\hat{\mathcal{G}}_0$ are chosen as large as possible. In particular, the trivial predictor satisfies $\mathcal{G}_0 = 0$ while $\hat{\mathcal{G}}_0 = \infty$. We assume further that in analogy to (3.1) we have

$$(4.2) \quad \begin{aligned} \Phi_k^{(l,\nu)}(\tau) &\in \left\{ \Phi^{(l,\nu)}(\tau), 0 \right\}, \quad \forall \tau \in T, \quad \mathfrak{g}(\tau) \leq k \\ \Rightarrow \quad \Phi_{k+1}^{(l,\nu)}(\tau) &\in \left\{ \Phi^{(l,\nu)}(\tau), 0 \right\}, \quad \forall \tau \in T, \quad \mathfrak{g}(\tau) \leq k+1, \end{aligned}$$

for all $k \geq 0$. By Lemmas 8, 9, and 10 this is guaranteed for the iteration schemes considered here.

It follows from (3.1), (3.2), and (4.2) that

$$(4.3) \quad \begin{aligned} \Phi_k(\tau) &= \Phi(\tau) & \forall \tau \in T \text{ with } \mathbf{g}'(\tau) \leq \mathcal{G}_0 + k, \\ \Phi_k(\tau) &\in \{\Phi(\tau), 0\} & \forall \tau \in T \text{ with } \mathbf{g}'(\tau) \leq \hat{\mathcal{G}}_0 + k \end{aligned}$$

as well as

$$(4.4) \quad \begin{aligned} \psi_{\Phi_k}(u) &= \psi_{\Phi}(u) & \forall u \in U_f \text{ with } \mathbf{g}'(u) \leq \mathcal{G}_0 + k, \\ \psi_{\Phi_k}(\tau) &\in \{\psi_{\Phi}(u), 0\} & \forall u \in U_f \text{ with } \mathbf{g}'(u) \leq \hat{\mathcal{G}}_0 + k. \end{aligned}$$

The next step is to establish the relation between the order and the growth function of a tree. We have chosen to do so by some maximum height functions, given by

$$(4.5) \quad \begin{aligned} \mathcal{G}_T(q) &= \max_{\tau \in T} \{\mathbf{g}'(\tau) : \rho(\tau) \leq q\}, & \mathcal{G}_{T,\varphi}(q) &= \max_{\tau \in T} \{\mathbf{g}'(\tau) : \mathbb{E} \varphi(\tau) \neq 0, \rho(\tau) \leq q\}, \\ \mathcal{G}_{U_f}(q) &= \max_{u \in U_f} \{\mathbf{g}'(u) : \rho(u) \leq q\}, & \mathcal{G}_{U_f,\psi_\varphi}(q) &= \max_{u \in U_f} \{\mathbf{g}'(u) : \mathbb{E} \psi_\varphi(u) \neq 0, \rho(u) \leq q\}. \end{aligned}$$

Note that the definition relates to the weights of the exact, not the numerical, solution. We are now ready to establish results on weak and strong convergence for the iterated solution.

Weak convergence. Let p be the weak order of the underlying scheme. The weak order of the iterated solution after k iterations is $\min(q_k, p)$ if

$$\mathbb{E} \psi_{\Phi_k}(u) = \mathbb{E} \psi_{\Phi}(u) \quad \forall u \in U_f, \quad \rho(u) \leq q_k + \frac{1}{2}.$$

If $q_k \leq p$ we can take advantage of the fact that $0 = \mathbb{E} \psi_\varphi(u) = \mathbb{E} \psi_{\Phi}(u) + \mathcal{O}(h^{p+1})$ for some u , and thereby relax the conditions to

$$(4.6) \quad \begin{aligned} \psi_{\Phi_k}(u) &= \psi_{\Phi}(u) & \forall u \in U_f \text{ with } \mathbb{E} \psi_\varphi(u) \neq 0, \\ \psi_{\Phi_k}(u) &\in \{\psi_{\Phi}(u), 0\} & \forall u \in U_f \text{ with } \mathbb{E} \psi_\varphi(u) = 0. \end{aligned}$$

By (4.4), this is fulfilled for all u of order $\rho(u) \leq \min(q_k, p)$ if

$$\mathcal{G}_{U_f,\Psi_\varphi} \left(q_k + \frac{1}{2} \right) \leq \mathcal{G}_0 + k \quad \text{and} \quad \mathcal{G}_{U_f} \left(q_k + \frac{1}{2} \right) \leq \hat{\mathcal{G}}_0 + k.$$

The results can then be summarized in the following theorem.

THEOREM 11. *The iterated method is of weak order $q_k \leq p$ after*

$$\max \left\{ \mathcal{G}_{U_f,\psi_\varphi} \left(q_k + \frac{1}{2} \right) - \mathcal{G}_0, \mathcal{G}_{U_f} \left(q_k + \frac{1}{2} \right) - \hat{\mathcal{G}}_0 \right\}$$

iterations.

Strong convergence. The strong convergence case can be treated similarly. Let p now be the mean square order of the underlying method. The iterated solution is of order $\min(p, q_k)$ if

$$(4.7) \quad \begin{aligned} \Phi_k(\tau) &= \Phi(\tau) & \forall \tau \in T \text{ with } \rho(\tau) \leq q_k, \\ \Phi_k(\tau) &= \Phi(\tau) & \forall \tau \in T \text{ with } \rho(\tau) = q_k + \frac{1}{2}, \quad \mathbb{E} \phi(\tau) \neq 0, \\ \Phi_k(\tau) &\in \{\Phi(\tau), 0\} & \forall \tau \in T \text{ with } \rho(\tau) = q_k + \frac{1}{2}, \quad \mathbb{E} \phi(\tau) = 0. \end{aligned}$$

According to (4.3) these are satisfied if all the conditions

$$\mathcal{G}_T(q_k) \leq \mathcal{G}_0 + k, \quad \mathcal{G}_T\left(q_k + \frac{1}{2}\right) \leq \hat{\mathcal{G}}_0 + k, \quad \text{and} \quad \mathcal{G}_{T,\varphi}\left(q_k + \frac{1}{2}\right) \leq \mathcal{G}_0 + k$$

are satisfied. We can summarize this by the following theorem.

THEOREM 12. *The iterated method is of mean square order $q_k \leq p$ after*

$$\max \left\{ \max \left\{ \mathcal{G}_T(q_k), \mathcal{G}_{T,\varphi}\left(q_k + \frac{1}{2}\right) \right\} - \mathcal{G}_0, \mathcal{G}_T\left(q_k + \frac{1}{2}\right) - \hat{\mathcal{G}}_0 \right\}$$

iterations.

5. Growth functions and order. In this section we will discuss the relation between the order of trees and the growth functions defined in section 3. Let us start with the following lemma.

LEMMA 13. *For $k \geq 1$,*

$$\begin{aligned} \mathfrak{h}'(\tau) = k &\Rightarrow \rho(\tau) \geq \frac{k}{2} + \frac{1}{2}, \\ \mathfrak{r}'(\tau) = k &\Rightarrow \rho(\tau) \geq k, \\ \mathfrak{d}'(\tau) = k &\Rightarrow \rho(\tau) \geq 2^{k-1}. \end{aligned}$$

The same result is valid for $\mathfrak{h}'(u)$, $\mathfrak{r}'(u)$, and $\mathfrak{g}'(u)$.

Proof. Let $\mathcal{T}_{\mathfrak{h},k}$, $\mathcal{T}_{\mathfrak{r},k}$, and $\mathcal{T}_{\mathfrak{d},k}$ be sets of trees of minimal order satisfying $\mathfrak{h}(\tau) = k \forall \tau \in \mathcal{T}_{\mathfrak{h},k}$, $\mathfrak{r}(\tau) = k \forall \tau \in \mathcal{T}_{\mathfrak{r},k}$, and $\mathfrak{d}(\tau) = k \forall \tau \in \mathcal{T}_{\mathfrak{d},k}$ (see Figure 5.1), and denote this minimal order by $\rho_{\mathfrak{h},k}$, $\rho_{\mathfrak{r},k}$ and $\rho_{\mathfrak{d},k}$. Minimal order trees are built up only by stochastic nodes. It follows immediately that $\mathcal{T}_{\mathfrak{h},1} = \mathcal{T}_{\mathfrak{r},1} = \mathcal{T}_{\mathfrak{d},1} = \{\bullet_l : l \geq 1\}$. Since $\rho(\bullet_l) = 1/2$ for $l \geq 1$, the results are proved for $k = 1$. It is easy to show by induction on k that

$$(5.1) \quad \begin{aligned} \mathcal{T}_{\mathfrak{h},k} &= \{[\tau]_l : \tau \in \mathcal{T}_{\mathfrak{h},k-1}, l \geq 1\}, & \rho_{\mathfrak{h},k} &= \rho_{\mathfrak{h},k-1} + \frac{1}{2} = \frac{k}{2}, \\ \mathcal{T}_{\mathfrak{r},k} &= \{[\bullet_{l_1}, \tau]_{l_2} : \tau \in \mathcal{T}_{\mathfrak{r},k-1}, l_1, l_2 \geq 1\}, & \rho_{\mathfrak{r},k} &= \rho_{\mathfrak{r},k-1} + 1 = k - \frac{1}{2}, \\ \mathcal{T}_{\mathfrak{d},k} &= \{[\tau_1, \tau_2]_l : \tau_1, \tau_2 \in \mathcal{T}_{\mathfrak{d},k-1}, l \geq 1\}, & \rho_{\mathfrak{d},k} &= 2\rho_{\mathfrak{d},k-1} + \frac{1}{2} = 2^{k-1} - \frac{1}{2}. \end{aligned}$$

For each \mathfrak{g} being either \mathfrak{h} , \mathfrak{r} , or \mathfrak{d} , the minimal order trees satisfying $\mathfrak{g}'(\tau'_{\mathfrak{g},k}) = k$, $\mathfrak{g}'(u_{\mathfrak{g},k}) = k$ are $\tau'_{\mathfrak{g},k} = [\tau_{\mathfrak{g},k}]_l$ with $\tau_{\mathfrak{g},k} \in \mathcal{T}_{\mathfrak{g},k}$ and $l \geq 1$, and $u_{\mathfrak{g},k} = [\tau'_{\mathfrak{g},k}]_f$. Both are of order $\rho(\tau_{\mathfrak{g},k}) + 1/2$. \square

Let $\mathcal{G}_T(q)$ and $\mathcal{G}_{U_f}(q)$ be defined by (4.5). Then the following corollary holds.

COROLLARY 14. *For $q \geq \frac{1}{2}$ we have*

$$\mathcal{G}_T(q) = \mathcal{G}_{U_f}(q) = \begin{cases} 2q - 1 & \text{for simple iterations,} \\ \lfloor q \rfloor & \text{for modified Newton iterations,} \\ \lfloor \log_2(q) \rfloor + 1 & \text{for full Newton iterations.} \end{cases}$$

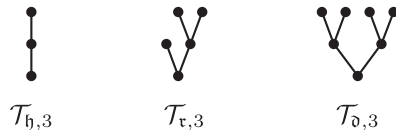


FIG. 5.1. Minimal order trees with $\mathfrak{g}(\tau) = 3$. The sets $\mathcal{T}_{\mathfrak{g},3}$ consist of all such trees with only stochastic nodes.

Proof. The minimal order trees are also the maximum height/ramification number/doubling index trees, in the sense that as long as $\rho(\tau'_{\mathbf{g},k}) \leq q < \rho(\tau'_{\mathbf{g},k+1})$ there are no trees of order q for which the growth function can exceed k . \square

Let $T^S \subset T$ and $U_f^S \subset U_f$ be the set of trees with an even number of each kind of stochastic nodes. Further, let $T^I \subset T_0$ and $U_f^I \subset U_f$ be the set of trees constructed from the root $(\bullet_0$ or $\bullet_f)$, by a finite number of steps of the form:

- (i) add one deterministic node, or
- (ii) add two equal stochastic nodes, neither of them being a father of the other.

Clearly $T^I \subset T^S$ and $U_f^I \subset U_f^S$. From [5, 26] we have

$$(5.2) \quad \begin{aligned} \mathbb{E} \varphi(\tau) = 0 & \quad \text{if } \tau \notin \begin{cases} T^S & \text{in the Stratonovich case,} \\ T^I & \text{in the It\^o case,} \end{cases} \\ \mathbb{E} \psi_\varphi(u) = 0 & \quad \text{if } u \notin \begin{cases} U_f^S & \text{in the Stratonovich case,} \\ U_f^I & \text{in the It\^o case.} \end{cases} \end{aligned}$$

Considering only trees for which $\mathbb{E} \varphi$ or $\mathbb{E} \psi_\varphi$ are different from zero, we get the following lemma.

LEMMA 15. For $k \geq 1$,

$$\begin{aligned} \mathfrak{h}'(\tau) = k & \Rightarrow \rho(\tau) \geq \begin{cases} \lceil \frac{k+1}{2} \rceil & \text{if } \tau \in T^S, \\ k+1 & \text{if } \tau \in T^I, \end{cases} \\ \mathfrak{r}'(\tau) = k & \Rightarrow \rho(\tau) \geq \begin{cases} k & \text{if } \tau \in T^S, \\ k+1 & \text{if } \tau \in T^I, \end{cases} \\ \mathfrak{d}'(\tau) = k & \Rightarrow \rho(\tau) \geq \begin{cases} 2^{k-1} & \text{if } \tau \in T^S, \\ 2^{k-1} + 1 & \text{if } \tau \in T^I. \end{cases} \end{aligned}$$

This result is also valid for $\mathfrak{h}'(u)$, $\mathfrak{r}'(u)$, and $\mathfrak{g}'(u)$, with T replaced by U_f .

Proof. In the Stratonovich case, we consider only trees of integer order, which immediately gives the results. In the It\^o case, let $\tau_{\mathbf{g},k}$, $\tau'_{\mathbf{g},k}$ be the minimal order trees used in the proof of Lemma 13. Unfortunately $\tau'_{\mathbf{g},k}$ has a stochastic root, so $\tau'_{\mathbf{g},k} \notin T^I$, and there are no trees $\tau \in T^I$ of order $\rho(\tau_{\mathbf{g},k}) + 1/2$ satisfying $\mathfrak{g}'(\tau) = k$. When \mathbf{g} is either \mathfrak{r} or \mathfrak{d} then the tree $[\tau_{\mathbf{g}}, \bullet_l]_0 \in T^I$ if all the stochastic nodes are of color $l \geq 1$. The order of this tree is $\rho(\tau_{\mathbf{g}}) + 3/2$, proving the result for $\mathfrak{r}'(\tau)$ and $\mathfrak{d}'(\tau)$. Let $\hat{\tau}'_{\mathfrak{h},k} \in T^I$ be a tree of minimal order satisfying $\mathfrak{h}'(\hat{\tau}'_{\mathfrak{h},k}) = k$. Clearly, $\hat{\tau}'_{\mathfrak{h},1}$ can be either $[\bullet_0]_0$ or $[\bullet_l, \bullet_l]_0$ with $l \geq 1$, both of order 2. From the construction of trees in T^I it is clear that the height of the tree can be increased only by one for each order, thus $\rho(\hat{\tau}'_{\mathfrak{h},k}) = k + 1$. The result for U_f^I follows immediately. \square

Let $\mathcal{G}_{T,\varphi}(q)$ and $\mathcal{G}_{U_f,\psi_\varphi}(q)$ be given by (4.5). Then the analogue of Corollary 14 is as follows.

COROLLARY 16. For $q \geq \frac{1}{2}$ we have in the Stratonovich case

$$\mathcal{G}_{T,\varphi}(q) = \mathcal{G}_{U_f,\psi_\varphi}(q) = \begin{cases} \max\{0, 2\lfloor q \rfloor - 1\} & \text{for simple iterations,} \\ \lfloor q \rfloor & \text{for modified Newton iterations,} \\ \lfloor \log_2(q) \rfloor + 1 & \text{for full Newton iterations,} \end{cases}$$

TABLE 5.1

Number of iterations needed to achieve order p when using the simple, modified, or full Newton iteration scheme in the Itô or Stratonovich case for strong or weak approximation.

p	Stratonovich			Itô					
	Strong/weak appr.			Weak appr.			Strong appr.		
	simple	mod.	full	simple	mod.	full	simple	mod.	full
$\frac{1}{2}$	1	1	1	0	0	0	0	0	0
1	1	1	1	0	0	0	1	1	1
$1\frac{1}{2}$	3	2	2	1	1	1	2	1	1
2	3	2	2	1	1	1	3	2	2
$2\frac{1}{2}$	5	3	2	2	2	1	4	2	2
3	5	3	2	2	2	1	5	3	2

and in the Itô case

$$\mathcal{G}_{T,\varphi}(q) = \mathcal{G}_{U_f,\psi_\varphi}(q) = \begin{cases} \max\{0, \lfloor q \rfloor - 1\} & \text{for simple iterations,} \\ \max\{0, \lfloor q \rfloor - 1\} & \text{for modified Newton iterations,} \\ \max\{0, \lfloor \log_2(q) \rfloor\} & \text{for full Newton iterations.} \end{cases}$$

For the trivial predictor, Table 5.1 gives the number of iterations needed to achieve a certain order of convergence. The results concerning the Stratonovich case when considering strong approximation and using the simple iteration scheme were already obtained by Burrage and Tian [3] analyzing predictor corrector methods.

6. Convergence results for composite methods. Composite methods have been introduced by Tian and Burrage [31]. At each step either a semi-implicit Runge–Kutta method or an implicit Runge–Kutta method is used in order to obtain better stability properties, which results in the method

$$(6.1a) \quad \begin{aligned} Y_{n+1} = Y_n + \lambda_n \sum_{l=0}^m \sum_{\nu=0}^M \left(z^{(1,l,\nu)\top} \otimes I_d \right) g_l \left(H^{(1,l,\nu)} \right) \\ + (1 - \lambda_n) \sum_{l=0}^m \sum_{\nu=0}^M \left(z^{(2,l,\nu)\top} \otimes I_d \right) g_l \left(H^{(2,l,\nu)} \right) \end{aligned}$$

for $n = 0, 1, \dots, N - 1$, $t_n \in I^h$, $\lambda_n \in \{0, 1\}$ and

$$(6.1b) \quad H^{(j,l,\nu)} = \mathbb{1}_s \otimes Y_n + \sum_{r=0}^m \sum_{\mu=0}^M \left(Z^{(j,l,\nu)(r,\mu)} \otimes I_d \right) g_r \left(H^{(j,r,\mu)} \right), \quad j = 1, 2.$$

Here, the method coefficients with superscripts 1 are those of the implicit SRK and the method coefficients with superscripts 2 are those of the semi-implicit SRK. Let $\Phi^{(1)}$, $\Phi^{(2)}$ be the corresponding weight-functions and $\Phi_k^{(1)}$, $\Phi_k^{(2)}$ be the corresponding weight-functions of the iterated methods. Then the weight-function Φ of the composite method is given by $\Phi = \lambda_1 \Phi^{(1)} + (1 - \lambda_1) \Phi^{(2)}$, and similarly we have for the iterated method $\Phi_k = \lambda_1 \Phi_k^{(1)} + (1 - \lambda_1) \Phi_k^{(2)}$. It follows that the convergence conditions (4.6) and (4.7), respectively, are satisfied if and only if they are satisfied as well for the underlying implicit SRK as for the semi-implicit SRK. Thus, an iterated composite method has the same order as the original composite method, if in each step the number of iterations is chosen according to Theorem 11 and Theorem 12, respectively. For the trivial predictor, the number of iterations needed to achieve a certain order of convergence is again given by Table 5.1.

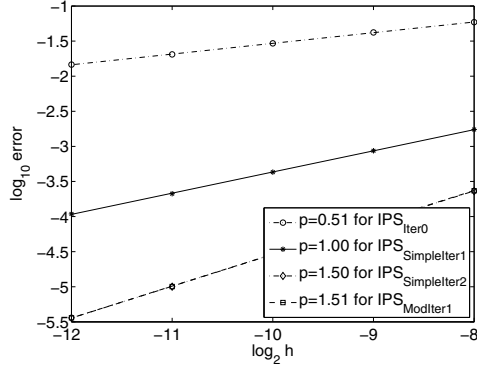


FIG. 7.1. Error of IPS applied to (7.1) without iteration, with one or two simple iterations, and with one modified Newton iteration (the last two results nearly coincide).

7. Numerical examples. In the following, we analyze numerically the order of convergence of three SRK methods in dependence on the kind and number of iterations. In each example, the solution is approximated with step sizes $2^{-8}, \dots, 2^{-12}$ and the sample average of $M = 20,000$ independent simulated realizations of the absolute error is calculated in order to estimate the expectation.

As a first example, we apply the drift implicit strong order 1.5 scheme due to Kloeden and Platen [16], implemented as a stiffly accurate SRK scheme with six stages and denoted by IPS; i.e., for one-dimensional Wiener processes

$$\begin{aligned}
 Y_{n+1} &= Y_n + \sum_{l=0}^1 \left(z^{(l)\top} \otimes I_d \right) g_l(H), & H &= \mathbf{1}_6 \otimes Y_n + \sum_{l=0}^1 \left(Z^{(l)} \otimes I_d \right) g_l(H), \\
 Z^{(0)} &= \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ h & 0 & 0 & 0 & 0 & 0 \\ h & 0 & 0 & 0 & 0 & 0 \\ h & 0 & 0 & 0 & 0 & 0 \\ h & 0 & 0 & 0 & 0 & 0 \\ \frac{h}{2} & a & -a & 0 & 0 & \frac{h}{2} \end{pmatrix}, & Z^{(1)} &= \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ \sqrt{h} & 0 & 0 & 0 & 0 & 0 \\ -\sqrt{h} & 0 & 0 & 0 & 0 & 0 \\ \sqrt{h} & \sqrt{h} & 0 & 0 & 0 & 0 \\ \sqrt{h} & -\sqrt{h} & 0 & 0 & 0 & 0 \\ I_{(1)} & b+c & b-c & d & -d & 0 \end{pmatrix}, \\
 z^{(0)} &= \left(\frac{h}{2}, a, -a, 0, 0, \frac{h}{2} \right)^\top, & z^{(1)} &= \left(I_{(1)}, b+c, b-c, d, -d, 0 \right)^\top, \\
 a &= \frac{I_{(1,0)} - \frac{1}{2}I_{(1)}h}{2\sqrt{h}}, & b &= \frac{I_{(0,1)}}{2h}, & c &= \frac{I_{(1,1)}}{2\sqrt{h}} - d, & d &= \frac{I_{(1,1,1)}}{2h},
 \end{aligned}$$

to the nonlinear SDE [16, 21]

$$(7.1) \quad dX(t) = \left(\frac{1}{2}X(t) + \sqrt{X(t)^2 + 1} \right) dt + \sqrt{X(t)^2 + 1} dW(t), \quad X(0) = 0,$$

on the time interval $I = [0, 1]$ with the solution $X(t) = \sinh(t + W(t))$.

The results at time $t = 1$ are presented in Figure 7.1, where the orders of convergence correspond to the slope of the regression lines. As predicted by Table 5.1 we observe strong order 0.5 without iteration, strong order 1.0 for one simple iteration, and strong order 1.5 for two simple or one modified Newton iteration.

As second example, we apply the diagonal implicit strong order 1.5 method DIRK4 which for one-dimensional Wiener processes is given by

$$Y_{n+1} = Y_n + \sum_{l=0}^1 \left(z^{(l)\top} \otimes I_d \right) g_l(H), \quad H = \mathbf{1}_3 \otimes Y_n + \sum_{l=0}^1 \left(Z^{(l)} \otimes I_d \right) g_l(H)$$

with coefficients²

$$\begin{aligned} z^0 &= h\alpha, & z^1 &= J_{(1)}\gamma^{(1)} + \frac{J_{(1,0)}}{h}\gamma^{(2)}, \\ Z^{(0)} &= hA, & Z^{(1)} &= J_{(1)}B^{(1)} + \frac{J_{(1,0)}}{h}B^{(2)} + \sqrt{h}B^{(3)}, \\ \alpha^\top &= (0.169775184, 0.297820839, 0.042159965, 0.490244012), \\ \gamma^{(1)\top} &= (-1.008751744, 0.285118644, 0.760818846, 0.962814254), \\ \gamma^{(2)\top} &= (1.507774621, 1.085932735, -1.458091242, -1.135616114), \\ A &= \begin{pmatrix} 0.240968725 & 0 & 0 & 0 \\ 0.167810317 & 0.160243373 & 0 & 0 \\ -0.002766912 & 0.473332751 & 0.178081733 & 0 \\ 0.415057712 & 0.115126049 & 0.020652745 & 0.130541130 \end{pmatrix}, \\ B_1 &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ -0.476890860 & 0 & 0 & 0 \\ 0.514160282 & 0.012424879 & 0 & 0 \\ -0.879966702 & 0.412866280 & 0.711524058 & 0 \end{pmatrix}, \\ B_2 &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1.287951512 & 0 & 0 & 0 \\ 0.665416412 & -0.686930244 & 0 & 0 \\ 0.703868780 & 0.876627859 & -0.321270197 & 0 \end{pmatrix}, \\ B_3 &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0.568300129 & -0.568300129 & 0 & 0 \\ 1.614193125 & -0.618659748 & -0.995533377 & 0 \\ 0.660721631 & -0.714401673 & -0.896487337 & 0.950167380 \end{pmatrix} \end{aligned}$$

to the corresponding Stratonovich version of (7.1). This method is constructed such that the regularity of the linear system which has to be solved in each modified Newton iteration step does not depend directly on $J_{(1)}$ and $J_{(1,0)}$.

The results at time $t = 1$ are presented in Figure 7.2. As predicted by Table 5.1 we observe no convergence without iteration, strong order 1.0 for one or two simple iterations or one modified Newton iteration, and strong order 1.5 in the case of three simple iterations or two modified Newton iterations.

²For typographical reasons, we restrict ourselves to an accuracy of $5 \cdot 10^{-10}$. A 16-digits version of the coefficients can be obtained on request from the authors.

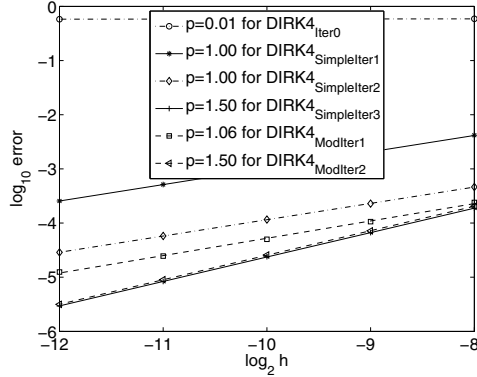


FIG. 7.2. Error of DIRK4 applied to the Stratonovich version of (7.1) without iteration, with one, two, or three simple iterations, and with one or two modified Newton iterations (the results for three simple iterations and two modified Newton iterations nearly coincide).

Finally, we apply the drift implicit strong order 1.0 scheme due to Kloeden and Platen [16], implemented as a stiffly accurate SRK scheme in the form

$$\begin{aligned}
 Y_{n+1} &= Y_n + \sum_{l=0}^m \sum_{\nu=0}^m \left(z^{(l,\nu)\top} \otimes I_d \right) g_l \left(H^{(\nu)} \right), \\
 H^{(\nu)} &= \mathbb{1}_2 \otimes Y_n + \sum_{l=0}^m \sum_{\mu=0}^m \left(Z^{(\nu)(l,\mu)} \otimes I_d \right) g_l \left(H^{(\mu)} \right), \quad \nu = 0, \dots, m, \\
 Z^{(0)(0,0)} &= \begin{pmatrix} 0 & 0 \\ \frac{h}{2} & \frac{h}{2} \end{pmatrix}, & Z^{(0)(0,\mu)} &= \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, & Z^{(0)(l,0)} &= \begin{pmatrix} 0 & 0 \\ I_{(l)} & 0 \end{pmatrix}, \\
 Z^{(0)(l,\mu)} &= \begin{pmatrix} 0 & 0 \\ -\frac{I_{(\mu,l)}}{\sqrt{h}} & \frac{I_{(\mu,l)}}{\sqrt{h}} \end{pmatrix}, & Z^{(j)(0,0)} &= \begin{pmatrix} 0 & 0 \\ h & 0 \end{pmatrix}, & Z^{(j)(l,l)} &= \begin{pmatrix} 0 & 0 \\ \sqrt{h} & 0 \end{pmatrix}, \\
 Z^{(j)(l,\mu)} &= \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \text{ for } l \neq \mu, & z^{(0,0)} &= \left(\frac{h}{2}, \frac{h}{2} \right)^\top, & z^{(0,\mu)} &= (0, 0)^\top, \\
 z^{(l,0)} &= (I_{(l)}, 0)^\top, & z^{(l,\mu)} &= \left(-\frac{I_{(\mu,l)}}{\sqrt{h}}, \frac{I_{(\mu,l)}}{\sqrt{h}} \right)^\top, & j, l, \mu &= 1, \dots, m,
 \end{aligned}$$

and denoted by IPS10 to the following nonlinear problem of dimension two driven by two Wiener processes in which there is no commutativity between the driving terms,

$$\begin{aligned}
 dX_1(t) &= \left(\frac{1}{2}X_1(t) + \sqrt{X_1(t)^2 + X_2(t)^2 + 1} \right) dt + \sqrt{X_2(t)^2 + 1} dW_1(t) \\
 (7.2a) \quad &+ \cos X_1(t) dW_2(t),
 \end{aligned}$$

$$\begin{aligned}
 (7.2b) \quad dX_2(t) &= \left(\frac{1}{2}X_1(t) + \sqrt{X_2(t)^2 + 1} \right) dt + \sqrt{X_1(t)^2 + 1} dW_1(t) + \sin X_2(t) dW_2(t), \\
 (7.2c) \quad X(0) &= 0.
 \end{aligned}$$

As here we don't know the exact solution, to approximate it we use IPS10 with two modified Newton iterations and a step size ten times smaller than the actual step size. The multiple Itô integrals are approximated as described in [16]. The results at time

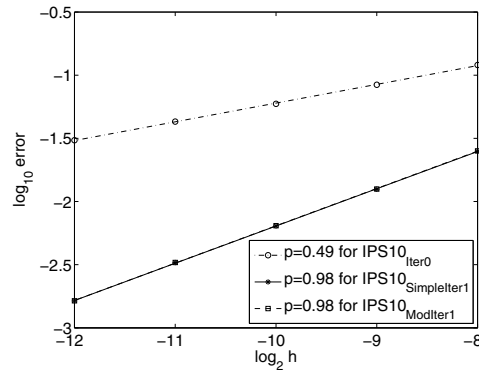


FIG. 7.3. Error of IPS10 applied to (7.2) without iteration, with one simple iteration and with one modified Newton iteration (the last two results nearly coincide).

$t = 1$ are presented in Figure 7.3. As predicted by Table 5.1 we observe strong order 0.5 without iteration and strong order 1.0 for one simple iteration or one modified Newton iteration.

8. Conclusion. For stochastic Runge–Kutta methods that use an iterative scheme to compute their internal stage values, we derived convergence results based on the order of the underlying Runge–Kutta method, the choice of the iteration method, the predictor, and the number of iterations. This was done by developing a unifying approach for the construction of stochastic B-series, which is valid both for Itô and Stratonovich-SDEs and can be used both for weak and strong convergence. We expect this to be useful also for the further development and analysis of stochastic Runge–Kutta type methods.

Acknowledgement. We thank the anonymous referees for their valuable comments and suggestions, which helped to improve this paper.

REFERENCES

- [1] K. BURRAGE AND P. M. BURRAGE, *High strong order explicit Runge–Kutta methods for stochastic ordinary differential equations*, Appl. Numer. Math., 22 (1996), pp. 81–101. Special issue celebrating the centenary of Runge–Kutta methods.
- [2] K. BURRAGE AND P. M. BURRAGE, *Order conditions of stochastic Runge–Kutta methods by B-series*, SIAM J. Numer. Anal., 38 (2000), pp. 1626–1646.
- [3] K. BURRAGE AND T. H. TIAN, *Predictor-corrector methods of Runge–Kutta type for stochastic differential equations*, SIAM J. Numer. Anal., 40 (2002), pp. 1516–1537.
- [4] K. BURRAGE AND T. H. TIAN, *Implicit stochastic Runge–Kutta methods for stochastic differential equations*, BIT, 44 (2004), pp. 21–39.
- [5] P. M. BURRAGE, *Runge–Kutta methods for stochastic differential equations*, Ph.D. thesis, The University of Queensland, Brisbane, 1999.
- [6] J. C. BUTCHER, *Coefficients for the study of Runge–Kutta integration processes*, J. Austral. Math. Soc., 3 (1963), pp. 185–201.
- [7] K. DEBRABANT AND A. RÖBLER, *Diagonally drift-implicit Runge–Kutta methods of weak order one and two for Itô SDEs and stability analysis*, Appl. Numer. Math., 2008, in press, doi:10.1016/j.apnum.2008.03.011.
- [8] K. DEBRABANT AND A. RÖBLER, *Families of efficient second order Runge–Kutta methods for the weak approximation of Itô stochastic differential equations*, Appl. Numer. Math., 2008, in press, doi:10.1016/j.apnum.2008.03.012.
- [9] K. DEBRABANT AND A. RÖBLER, *Classification of stochastic Runge–Kutta methods for the weak approximation of stochastic differential equations*, Math. Comput. Simulation, 77 (2008), pp. 408–420.

- [10] T. C. GARD, *Introduction to stochastic differential equations*, Monographs and Textbooks in Pure Appl. Math. 114, Marcel Dekker Inc., New York, 1988.
- [11] D. B. HERNÁNDEZ AND R. SPIGLER, *Convergence and stability of implicit Runge–Kutta methods for systems with multiplicative noise*, BIT, 33 (1993), pp. 654–669.
- [12] D. J. HIGHAM, *Mean-square and asymptotic stability of the stochastic theta method*, SIAM J. Numer. Anal., 38 (2000), pp. 753–769.
- [13] K. R. JACKSON, A. KVÆRNØ, AND S. P. NØRSETT, *The use of Butcher series in the analysis of Newton-like iterations in Runge–Kutta formulas*, Appl. Numer. Math., 15 (1994), pp. 341–356. International Conference on Scientific Computation and Differential Equations (Auckland, 1993).
- [14] K. R. JACKSON, A. KVÆRNØ, AND S. P. NØRSETT, *An analysis of the order of Runge–Kutta methods that use an iterative scheme to compute their internal stage values*, BIT, 36 (1996), pp. 713–765.
- [15] I. KARATZAS AND S. E. SHREVE, *Brownian motion and stochastic calculus*, Graduate Texts in Mathematics 113, 2nd ed., Springer-Verlag, New York, 1991.
- [16] P. E. KLOEDEN AND E. PLATEN, *Numerical solution of stochastic differential equations*, Applications of Mathematics 21, 2nd ed., Springer-Verlag, Berlin, 1999.
- [17] Y. KOMORI, *Weak first- or second-order implicit Runge–Kutta methods for stochastic differential equations with a scalar Wiener process*, J. Comput. Appl. Math., 217 (2008), pp. 166–179.
- [18] Y. KOMORI, *Multi-colored rooted tree analysis of the weak order conditions of a stochastic Runge–Kutta family*, Appl. Numer. Math., 57 (2007), pp. 147–165.
- [19] Y. KOMORI, *Weak second-order stochastic Runge–Kutta methods for non-commutative stochastic differential equations*, J. Comput. Appl. Math., 206 (2007), pp. 158–173.
- [20] Y. KOMORI, T. MITSUI, AND H. SUGIURA, *Rooted tree analysis of the order conditions of ROW-type scheme for stochastic differential equations*, BIT, 37 (1997), pp. 43–66.
- [21] V. MACKEVIČIUS AND J. NAVIKAS, *Second order weak Runge–Kutta type methods of Itô equations*, Math. Comput. Simulation, 57 (2001), pp. 29–34.
- [22] G. N. MILSTEIN, *Numerical integration of stochastic differential equations*, Mathematics and its Applications 313, Kluwer Academic Publishers Group, Dordrecht, 1995. Translated and revised from the 1988 Russian original.
- [23] G. N. MILSTEIN, Y. M. REPIN, AND M. V. TRETYAKOV, *Numerical methods for stochastic systems preserving symplectic structure*, SIAM J. Numer. Anal., 40 (2003), pp. 1583–1604.
- [24] S. P. NØRSETT AND A. WOLFBRANDT, *Order conditions for Rosenbrock type methods*, Numer. Math., 32 (1979), pp. 1–15.
- [25] W. P. PETERSEN, *A general implicit splitting for stabilizing numerical simulations of Itô stochastic differential equations*, SIAM J. Numer. Anal., 35 (1998), pp. 1439–1451.
- [26] A. RÖBLER, *Stochastic Taylor expansions for the expectation of functionals of diffusion processes*, Stoch. Anal. Appl., 22 (2004), pp. 1553–1576.
- [27] A. RÖBLER, *Rooted tree analysis for order conditions of stochastic Runge–Kutta methods for the weak approximation of stochastic differential equations*, Stoch. Anal. Appl., 24 (2006), pp. 97–134.
- [28] A. RÖBLER, *Runge–Kutta methods for Itô stochastic differential equations with scalar noise*, BIT, 46 (2006), pp. 97–110.
- [29] A. RÖBLER, *Second order Runge–Kutta methods for Itô stochastic differential equations*, Technical report Preprint 2479, TU Darmstadt, 2006.
- [30] A. RÖBLER, *Second order Runge–Kutta methods for Stratonovich stochastic differential equations*, BIT, 47 (2007), pp. 657–680.
- [31] T. H. TIAN AND K. BURRAGE, *Two-stage stochastic Runge–Kutta methods for stochastic differential equations*, BIT, 42 (2002), pp. 625–643.
- [32] A. TOCINO AND J. VIGO-AGUIAR, *Weak second order conditions for stochastic Runge–Kutta methods*, SIAM J. Sci. Comput., 24 (2002), pp. 507–523.

FINITE ELEMENT METHOD FOR THE SPACE AND TIME FRACTIONAL FOKKER–PLANCK EQUATION*

WEIHUA DENG[†]

Abstract. We develop the finite element method for the numerical resolution of the space and time fractional Fokker–Planck equation, which is an effective tool for describing a process with both traps and flights; the time fractional derivative of the equation is used to characterize the traps, and the flights are depicted by the space fractional derivative. The stability and error estimates are rigorously established, and we prove that the convergent order is $O(k^{2-\alpha} + h^\mu)$, where k is the time step size and h the space step size. Numerical computations are presented which demonstrate the effectiveness of the method and confirm the theoretical claims.

Key words. finite element method, fractional Fokker–Planck equation, Lévy flights, stability, convergence

AMS subject classifications. 65M60, 35S10, 65M12, 02.70.-c, 05.10.-a, 05.40.Fb

DOI. 10.1137/080714130

1. Introduction. The use of fractional calculus to deal with engineering and physical problems has become increasingly popular in recent years [2, 7, 21, 33, 36]. The fractional approach has become a powerful modeling methodology; it is widely applied in material and mechanics, signal processing and systems identification, anomalous diffusion, control and robotics, wave propagation, turbulence, seepage in fractal media, friction modeling, etc. [2]. At the same time, more and more fractional dynamical appearances are disclosed, for instance, viscoelasticity [22], colored noise [31], boundary layer effects in ducts [39], electromagnetic waves [17], fractional kinetics [23, 41], electrode-electrolyte polarization [19], synchronization of chaos [10], and multidirectional multiscroll attractors [9]. When the fractional differential equations describe the asymptotic behavior of continuous time random walks, their solutions correspond to the Lévy walks. The advantage of the fractional model basically lies in the straightforward way of including external force terms and of calculating boundary value problems. The complexity of these equations comes from involving pseudodifferential operators that are nonlocal and have the character of history dependence and universal mutuality.

Anomalous diffusion is one of the most ubiquitous phenomena in nature, being observed in various fields of physics, for instance, transport of fluid in porous media, surface growth, diffusion of plasma, diffusion at liquid surfaces, and two-dimensional rotating flow [5, 16, 25, 38]. The processes of anomalous diffusion are various, including processes with infinite mean-square displacement and processes with distributed orders of fraction leading to the mean-square displacement for power laws with a time-dependent exponent; at the same time the most often arisen anomalous diffusion has the power law form $\langle x^2(t) \rangle - \langle x(t) \rangle^2 \sim k_\alpha t^\gamma$ of the mean-square displacement, deviating from the well-known property $\langle x^2(t) \rangle - \langle x(t) \rangle^2 \sim k_\alpha t$ of Brownian motion.

*Received by the editors January 24, 2008; accepted for publication (in revised form) July 1, 2008; published electronically October 29, 2008. This work was supported by the National Natural Science Foundation of China under grant 10801067 and the Fundamental Research Fund for Physics and Mathematics of Lanzhou University under grant Lzu07001.

<http://www.siam.org/journals/sinum/47-1/71413.html>

[†]School of Mathematics and Statistics, Lanzhou University, Lanzhou 730000, People's Republic of China (dengwh@lzu.edu.cn, dengwhmath@yahoo.com.cn).

For Lévy flights, γ is larger than one (but typically smaller than two), which is called superdiffusion; $\gamma = 2$ corresponds to “ballistic” motion, for example, the particles of a bomb which is exploding, and it is called subdiffusion if γ is less than one, which in general corresponds to the divergence of microscopic time scales in random walk schemes, i.e., traps [5, 25, 34, 35, 38, 41]. A prominent characteristic feature of Lévy flights and traps is the probability distribution functions both on the step length for Lévy flights and on the waiting time between steps for traps that have a powerlike tail. There is no characteristic length or time scale in the case of flights or traps. Combining flights with traps is a more general way to describe an anomalous process. When a process has both flights and traps, then γ depends on the competition between flights and traps; that is, the process can be subdiffusion, superdiffusion, or normal diffusion (Brownian motion). Fractional derivatives play a key role in characterizing anomalous diffusion, including the space fractional Fokker–Planck (advection-dispersion) equation describing Lévy flights, the time fractional Fokker–Planck equation depicting traps, and the space and time fractional Fokker–Planck equation characterizing the competition between Lévy flights and traps [3, 4, 16, 18, 20, 23, 25, 28, 32, 34, 35, 38, 41].

The Fokker–Planck equation (named after Adriaan Fokker and Max Planck) describes the time evolution of the probability density function of the position and the velocity of a particle, which is one of the classical, widely used equations of statistical physics. For the numerical algorithms of the time fractional Fokker–Planck-type equation and the space fractional Fokker–Planck-type equation, there is already some progress covering the finite difference method, the finite element method, and some random approaches [11, 12, 13, 15, 18, 27, 28, 32, 40]. However, published papers on the numerical algorithm of the space and time fractional Fokker–Planck equation are very sparse. Using the Monte Carlo approach for this equation, Magdziarz and Weron [30] positively answer a question raised by Metzler and Klafter [34]: Can one see a competition between subdiffusion and Lévy flights in the framework of the fractional Fokker–Planck dynamics? Herein, we focus on developing the finite element method for the space and time fractional Fokker–Planck equation [30, 34, 35, 41], describing the competition between Lévy flights and traps under the influence of an external potential $U(x)$, given by

$$(1.1) \quad \frac{\partial}{\partial t} p(x, t) = {}_0D_t^{1-\alpha} \left[\frac{\partial}{\partial x} \frac{U'(x)}{\eta_\alpha} + \kappa_\alpha \nabla^\mu \right] p(x, t),$$

where $p(x, t)$ is the probability density, prime stands for the derivative w.r.t. the space coordinate, κ_α denotes the anomalous diffusion coefficient with physical dimension $[\text{m}^\mu \text{s}^{-\alpha}]$, η_α represents the generalized friction coefficient possessing the dimension $[\text{kg s}^{\alpha-2}]$, $\alpha \in (0, 1)$, and $\mu \in (1, 2)$ throughout this paper. Here, the operators

$${}_0D_t^{1-\alpha} p(x, t) = \frac{1}{\Gamma(\alpha)} \frac{\partial}{\partial t} \int_0^t (t - \tau)^{\alpha-1} p(x, \tau) d\tau$$

and $\nabla^\mu = \frac{1}{2} {}_aD_x^\mu + \frac{1}{2} {}_xD_b^\mu$; ${}_aD_x^\mu$ and ${}_xD_b^\mu$ are the left and right Riemann–Liouville space fractional derivatives of order μ , respectively, described by

$${}_aD_x^\mu p(x, t) = D^2 {}_aD_x^{-(2-\mu)} p(x, t) = \frac{1}{\Gamma(2-\mu)} \frac{d^2}{dx^2} \int_a^x (x - \xi)^{1-\mu} p(\xi, t) d\xi$$

and

$${}_xD_b^\mu p(x, t) = (-D)^2 {}_xD_b^{-(2-\mu)} p(x, t) = \frac{1}{\Gamma(2-\mu)} \frac{d^2}{dx^2} \int_x^b (\xi - x)^{1-\mu} p(\xi, t) d\xi.$$

Letting ${}_0D_t^{\alpha-1}$ perform on both sides of (1.1) and according to the attributes of the Riemann–Liouville and the Caputo derivatives [11], we obtain the equivalent form of (1.1) as

$$(1.2) \quad D_*^\alpha p(x, t) = \left[\frac{\partial}{\partial x} \frac{U'(x)}{\eta_\alpha} + \kappa_\alpha \nabla^\mu \right] p(x, t),$$

where D_*^α is the Caputo derivative and it is defined by

$$D_*^\alpha p(x, t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t (t-\tau)^{-\alpha} \frac{\partial p(x, \tau)}{\partial \tau} d\tau.$$

There are several ways to discrete the time fractional derivative and speed its computation [8, 27, 29, 36]. Here we use the one provided by Lin and Xu in [27] to discrete the left-hand side time Caputo derivative and exploit the finite element method to approximate the right-hand side space fractional derivative. This approach based on the temporal backward differentiation and the spatial finite element method obtains estimates of $(2-\alpha)$ -order convergence in time and μ -order convergence in space. The time-stepping scheme is shown to be unconditionally stable. The numerical example is provided, and the real physical cases are simulated, which illustrate the effectiveness of the method and support the theoretical claims.

In section 2, we introduce the temporal discretization of (1.2) and the abstract setting, fractional derivative spaces, for the analysis of the finite element approximation of the space and time fractional Fokker–Planck equation. Section 3 is devoted to the stability analysis of the time-stepping scheme and the detailed error analysis of semidiscretization on time and of full discretization. In section 4, numerical experiments are carried out, and some of their results are compared with the exact solution. Some concluding remarks are given in the last section.

2. Preliminaries: Discretization of the Caputo derivative and the fractional derivative spaces. For the completeness of the paper, in this section we introduce the scheme to discretize the temporal Caputo derivative and the fractional derivative spaces; for more detailed discussions see [27] and [12, 13], respectively.

For the finite difference of the Caputo derivative, let $t_m := mk$, $m = 0, 1, \dots, M$, where $k := \frac{T}{M}$ is the time-step length, so we have the following formulation of the Caputo derivative [27]:

$$(2.1) \quad \begin{aligned} D_*^\alpha p(x, t_{m+1}) &= \frac{1}{\Gamma(1-\alpha)} \sum_{j=0}^m \int_{t_j}^{t_{j+1}} (t_{m+1}-\tau)^{-\alpha} \frac{\partial p(x, \tau)}{\partial \tau} d\tau \\ &= \frac{1}{\Gamma(1-\alpha)} \sum_{j=0}^m \frac{p(x, t_{j+1}) - p(x, t_j)}{k} \int_{t_j}^{t_{j+1}} \frac{d\tau}{(t_{m+1}-\tau)^\alpha} + r_k^{m+1} \\ &= \frac{1}{\Gamma(2-\alpha)} \sum_{j=0}^m \frac{p(x, t_{m+1-j}) - p(x, t_{m-j})}{k^\alpha} d_j + r_k^{m+1}, \end{aligned}$$

where $d_j = (j+1)^{1-\alpha} - j^{1-\alpha}$. Furthermore,

$$(2.2) \quad r_k^{m+1} \leq \tilde{c}_p k^{2-\alpha},$$

where \tilde{c}_p is a constant depending only on p . Let us define the discrete fractional

differential operator L_t^α by

$$L_t^\alpha p(x, t_{m+1}) := \frac{1}{\Gamma(2-\alpha)} \sum_{j=0}^m d_j \frac{p(x, t_{m+1-j}) - p(x, t_{m-j})}{k^\alpha}.$$

Then (2.1) reads

$$(2.3) \quad D_*^\alpha p(x, t_{m+1}) = L_t^\alpha p(x, t_{m+1}) + r_k^{m+1}.$$

In what follows, we introduce the left, right, and symmetric fractional derivative spaces.

DEFINITION 2.1 (left fractional derivative [12, 24, 37]). *Let q be a function defined on R , $\beta > 0$, n be the smallest integer greater than β ($n-1 \leq \beta < n$), and $\sigma = n - \beta$. Then the left fractional derivative of order β is defined to be*

$$\mathbf{D}^\beta q := D_{-\infty}^n D_x^{-\sigma} q(x) = \frac{1}{\Gamma(\sigma)} \frac{d^n}{dx^n} \int_{-\infty}^x (x-\xi)^{\sigma-1} q(\xi) d\xi.$$

DEFINITION 2.2 (right fractional derivative [12, 24, 37]). *Let q be a function defined on R , $\beta > 0$, n be the smallest integer greater than β ($n-1 \leq \beta < n$), and $\sigma = n - \beta$. Then the right fractional derivative of order β is defined to be*

$$\mathbf{D}^{\beta*} q := (-D)^n {}_x D_\infty^{-\sigma} q(x) = \frac{(-1)^n}{\Gamma(\sigma)} \frac{d^n}{dx^n} \int_x^\infty (\xi-x)^{\sigma-1} q(\xi) d\xi.$$

Note. If $\text{supp}(q) \subset (a, b)$, then $\mathbf{D}^\beta q = {}_a D_x^\beta q$ and $\mathbf{D}^{\beta*} q = {}_x D_b^\beta q$, where ${}_a D_x^\beta q$ and ${}_x D_b^\beta q$ are the left and right Riemann–Liouville fractional derivatives, respectively, of order β defined as

$${}_a D_x^\beta q = \frac{1}{\Gamma(\sigma)} \frac{d^n}{dx^n} \int_a^x (x-\xi)^{\sigma-1} q(\xi) d\xi$$

and

$${}_x D_b^\beta q = \frac{(-1)^n}{\Gamma(\sigma)} \frac{d^n}{dx^n} \int_x^b (\xi-x)^{\sigma-1} q(\xi) d\xi.$$

DEFINITION 2.3 (left fractional derivative space [12]). *Let $\beta > 0$. Define the seminorm*

$$|q|_{J_L^\beta(R)} := \|\mathbf{D}^\beta q\|_{L^2(R)}$$

and the norm

$$\|q\|_{J_L^\beta(R)} := \left(\|q\|_{L^2(R)}^2 + |q|_{J_L^\beta(R)}^2 \right)^{1/2},$$

and let $J_L^\beta(R)$ denote the closure of $C^\infty(R)$ with respect to $\|\cdot\|_{J_L^\beta(R)}$.

DEFINITION 2.4 (right fractional derivative space [12]). *Let $\beta > 0$. Define the seminorm*

$$|q|_{J_R^\beta(R)} := \|\mathbf{D}^{\beta*} q\|_{L^2(R)}$$

and the norm

$$\|q\|_{J_R^\beta(R)} := \left(\|q\|_{L^2(R)}^2 + |q|_{J_R^\beta(R)}^2 \right)^{1/2},$$

and let $J_R^\beta(R)$ denote the closure of $C^\infty(R)$ with respect to $\|\cdot\|_{J_R^\beta(R)}$.

DEFINITION 2.5 (symmetric fractional derivative space [12]). Let $\beta > 0$, $\beta \neq n - 1/2$, $n \in \mathbb{N}$. Define the seminorm

$$|q|_{J_S^\beta(R)} := \left| (\mathbf{D}^\beta q, \mathbf{D}^{\beta*} q)_{L^2(R)} \right|^{1/2}$$

and the norm

$$\|q\|_{J_S^\beta(R)} := \left(\|q\|_{L^2(R)}^2 + |q|_{J_S^\beta(R)}^2 \right)^{1/2},$$

and let $J_S^\beta(R)$ denote the closure of $C^\infty(R)$ with respect to $\|\cdot\|_{J_S^\beta(R)}$.

DEFINITION 2.6 (see [12]). Let $\beta > 0$. Define the seminorm

$$|q|_{H^\beta(R)} := \|\omega^{|\beta|} \hat{q}\|_{L^2(R)}$$

and the norm

$$(2.4) \quad \|q\|_{H^\beta(R)} := \left(\|q\|_{L^2(R)}^2 + |q|_{H^\beta(R)}^2 \right)^{1/2},$$

and let $H^\beta(R)$ denote the closure of $C^\infty(R)$ with respect to $\|\cdot\|_{H^\beta(R)}$.

Note. In this paper, instead of (2.4), we prefer to use

$$(2.5) \quad \|q\|_{H^\beta(R)} := \left(\|q\|_{L^2(R)}^2 + \alpha_0 \kappa_\alpha \left| \cos\left(\frac{\beta}{2}\pi\right) \right| |q|_{H^\beta(R)}^2 \right)^{1/2},$$

where $\alpha_0 = \Gamma(2 - \alpha)k^\alpha$; it is well known that these two definitions are equivalent [1].

LEMMA 2.7 (see [12]). Let $\beta > 0$ be given. Then

$$(2.6) \quad (\mathbf{D}^\beta q, \mathbf{D}^{\beta*} q) = \cos(\beta\pi) \|\mathbf{D}^\beta q\|_{L^2(R)}^2.$$

LEMMA 2.8 (see [12]). Let $\beta > 0$. The spaces $J_L^\beta(R)$, $J_R^\beta(R)$, and $H^\beta(R)$ are equivalent, with equivalent seminorms and norms; and in fact

$$|\cdot|_{J_L^\beta(R)} = |\cdot|_{J_R^\beta(R)} = |\cdot|_{H^\beta(R)},$$

in particular, when $\beta \neq n - 1/2$, $n \in \mathbb{N}$. The spaces $J_L^\beta(R)$ ($J_R^\beta(R)$ or $H^\beta(R)$) and $J_S^\beta(R)$ are equivalent, with equivalent seminorms and norms, and

$$|\cdot|_{J_S^\beta(R)} = |\cos(\beta\pi)| |\cdot|_{J_L^\beta(R)}.$$

DEFINITION 2.9 (see [12]). Define the spaces $J_L^\beta(\Omega)$, $J_R^\beta(\Omega)$, and $J_S^\beta(\Omega)$ as the closures of $C^\infty(\Omega)$ under their respective norms.

DEFINITION 2.10 (see [12]). Define the spaces $J_{L,0}^\beta(\Omega)$, $J_{R,0}^\beta(\Omega)$, and $J_{S,0}^\beta(\Omega)$ as the closures of $C_0^\infty(\Omega)$ under their respective norms.

LEMMA 2.11 (see [12]). *Let $\beta > 0$. Then the spaces $J_{L,0}^\beta(\Omega)$, $J_{R,0}^\beta(\Omega)$, and $H_0^\beta(\Omega)$ are equivalent. Also, if $\beta \neq n - \frac{1}{2}$, $n \in \mathbb{N}$, the seminorms and norms of $J_{L,0}^\beta(\Omega)$, $J_{R,0}^\beta(\Omega)$, and $H_0^\beta(\Omega)$ are equivalent, and $J_{S,0}^\beta(\Omega)$ and $J_{L,0}^\beta(\Omega)$ ($J_{R,0}^\beta(\Omega)$ or $H_0^\beta(\Omega)$) are equivalent with equivalent seminorms and norms.*

LEMMA 2.12 (fractional Poincaré–Friedrichs [12]). *For $q \in H_0^\beta(\Omega)$, we have*

$$\|q\|_{L^2(\Omega)} \leq c|q|_{H_0^\beta(\Omega)},$$

and for $0 < s < \beta$, $s \neq n - \frac{1}{2}$, $n \in \mathbb{N}$,

$$|q|_{H_0^s(\Omega)} \leq c|q|_{H_0^\beta(\Omega)}.$$

LEMMA 2.13. *The left and right Riemann–Liouville fractional integral operators are adjoint w.r.t. the $L^2(a, b)$ inner product, i.e.,*

$$\left({}_a D_x^{-\beta} p, q \right)_{L^2(a,b)} = \left(p, {}_x D_b^{-\beta} q \right)_{L^2(a,b)} \quad \forall \beta > 0,$$

where ${}_a D_x^{-\beta}$ and ${}_x D_b^{-\beta}$ are defined by

$${}_a D_x^{-\beta} p = \frac{1}{\Gamma(\beta)} \int_a^x (x - \xi)^{\beta-1} p(\xi) d\xi$$

and

$${}_x D_b^{-\beta} q = \frac{1}{\Gamma(\beta)} \int_x^b (\xi - x)^{\beta-1} q(\xi) d\xi.$$

Proof. Interchanging the order of integration leads to

$$\begin{aligned} \left({}_a D_x^{-\beta} p, q \right)_{L^2(a,b)} &= \frac{1}{\Gamma(\beta)} \int_a^b \int_a^x (x - \xi)^{\beta-1} p(\xi) q(x) d\xi dx \\ &= \frac{1}{\Gamma(\beta)} \int_a^b p(\xi) \int_\xi^b (x - \xi)^{\beta-1} q(x) dx d\xi \\ &= \left(p, {}_x D_b^{-\beta} q \right)_{L^2(a,b)}. \quad \square \end{aligned}$$

3. Variational formulation and finite element approximation. Letting $T > 0$, $\Omega = (a, b)$, rewriting (1.2), and making it subject to the given initial and boundary conditions, we have

$$(3.1) \quad D_*^\alpha p(x, t) = \left[\frac{\partial}{\partial x} \frac{U'(x)}{\eta_\alpha} + \kappa_\alpha \nabla^\mu \right] p(x, t), \quad 0 < t \leq T, \quad x \in \Omega,$$

with initial and boundary conditions

$$(3.2) \quad p(x, 0) = g(x), \quad x \in \Omega,$$

$$(3.3) \quad p(a, t) = p(b, t) = 0, \quad 0 \leq t \leq T.$$

Because of (2.3), using $L_t^\alpha p(x, t_{m+1})$ as an approximation of $D_*^\alpha p(x, t_{m+1})$ leads to the following time-discrete scheme of (3.1):

$$(3.4) \quad L_t^\alpha p^{m+1}(x) = \left[\frac{\partial}{\partial x} \frac{U'(x)}{\eta_\alpha} + \kappa_\alpha \nabla^\mu \right] p^{m+1}(x), \quad m = 0, 1, \dots, M-1,$$

where $p^{m+1}(x)$ is an approximation of $p(x, t_{m+1})$.

The scheme (3.4) can be recast as, with simplification by omitting the dependence of $p^{m+1}(x)$ on x ,

$$(3.5) \quad p^{m+1} - \alpha_0 \left[\frac{\partial}{\partial x} \frac{U'(x)}{\eta_\alpha} + \kappa_\alpha \nabla^\mu \right] p^{m+1} = p^m - \sum_{j=0}^{m-1} d_{j+1} p^{m-j} + \sum_{j=1}^m d_j p^{m-j}.$$

For the first time step, that is, $m = 0$, the scheme simply reads

$$(3.6) \quad p^1 - \alpha_0 \left[\frac{\partial}{\partial x} \frac{U'(x)}{\eta_\alpha} + \kappa_\alpha \nabla^\mu \right] p^1 = p^0,$$

and for the remaining steps, the equivalent form of (3.5) is

$$(3.7) \quad \begin{aligned} & p^{m+1} - \alpha_0 \left[\frac{\partial}{\partial x} \frac{U'(x)}{\eta_\alpha} + \kappa_\alpha \nabla^\mu \right] p^{m+1} \\ &= (1 - d_1) p^m + \sum_{j=1}^{m-1} (d_j - d_{j+1}) p^{m-j} + d_m p^0, \quad m \geq 1. \end{aligned}$$

So the complete semidiscrete scheme of (3.1) is (3.6) and (3.7), together with the boundary conditions

$$(3.8) \quad p^{m+1}(a) = p^{m+1}(b) = 0, \quad m \geq 0,$$

and the initial condition

$$(3.9) \quad p^0(x) = g(x), \quad x \in \Omega.$$

The variational formulation of (3.7) subject to the boundary condition (3.8) reads as follows: Find $p^{m+1} \in H_0^{\frac{\mu}{2}}(\Omega)$ such that

$$(3.10) \quad \begin{aligned} & (p^{m+1}, q) - \alpha_0 \left(\left[\frac{\partial}{\partial x} \frac{U'(x)}{\eta_\alpha} + \kappa_\alpha \nabla^\mu \right] p^{m+1}, q \right) \\ &= (1 - d_1)(p^m, q) + \sum_{j=1}^{m-1} (d_j - d_{j+1})(p^{m-j}, q) + d_m(p^0, q) \quad \forall q \in H_0^{\frac{\mu}{2}}(\Omega). \end{aligned}$$

3.1. Stability analysis and error estimates for the semidiscrete scheme.

In this subsection, we discuss the stability and error estimates for the weak time-discrete problem. First, let us introduce the denotation

$$B(p, q) := - \left(\left[\frac{\partial}{\partial x} \frac{U'(x)}{\eta_\alpha} + \kappa_\alpha \nabla^\mu \right] p, q \right),$$

and we have the following lemma for $B(p, q)$.

LEMMA 3.1. *Let $U''(x) \leq 0$ for any $x \in \Omega$, $q \in H_0^{\frac{\mu}{2}}(\Omega)$. Then $B(q, q) \geq \kappa_\alpha |\cos(\frac{\mu}{2}\pi)| |q|_{H_0^{\frac{\mu}{2}}(\Omega)}^2$.*

Proof. To expand the expression, we obtain

$$\begin{aligned}
& B(q, q) \\
&= - \int_a^b \frac{U''(x)}{\eta_\alpha} q^2 dx - \int_a^b \frac{U'(x)}{\eta_\alpha} \frac{\partial q}{\partial x} q dx \\
&\quad - \frac{\kappa_\alpha}{2} \int_a^b \left(D_a^2 D_x^{-(2-\mu)} q \right) q dx - \frac{\kappa_\alpha}{2} \int_a^b \left(D_x^2 D_b^{-(2-\mu)} q \right) q dx \\
&= - \frac{1}{2} \int_a^b \frac{U''(x)}{\eta_\alpha} q^2 dx + \frac{\kappa_\alpha}{2} \left(\int_a^b \left({}_a D_x^{-(2-\mu)} Dq \right) Dq dx + \int_a^b \left({}_x D_b^{-(2-\mu)} Dq \right) Dq dx \right) \\
&\geq \frac{\kappa_\alpha}{2} \left(\int_a^b \left({}_a D_x^{-(2-\mu)} Dq \right) Dq dx + \int_a^b \left({}_x D_b^{-(2-\mu)} Dq \right) Dq dx \right) \\
&= \frac{\kappa_\alpha}{2} \left(\int_a^b \left({}_a D_x^{-(1-\frac{\mu}{2})} Dq \right) {}_x D_b^{-(1-\frac{\mu}{2})} Dq dx + \int_a^b \left({}_x D_b^{-(1-\frac{\mu}{2})} Dq \right) {}_a D_x^{-(1-\frac{\mu}{2})} Dq dx \right) \\
&= \kappa_\alpha \int_a^b \left({}_a D_x^{\frac{\mu}{2}} q \right) \cdot \left(-{}_x D_b^{\frac{\mu}{2}} q \right) dx \\
&= -\kappa_\alpha \cos\left(\frac{\mu}{2}\pi\right) |q|_{H_0^{\frac{\mu}{2}}(\Omega)}^2.
\end{aligned}$$

In the above calculations, Lemmas 2.7, 2.8, and 2.11 are used, and some of the properties of fractional operators [11, 26], such as ${}_a D_x^{-(2-\mu)} = {}_a D_x^{-(1-\frac{\mu}{2})} {}_a D_x^{-(1-\frac{\mu}{2})}$, ${}_a D_x^{-(2-\mu)} Dq = D_a D_x^{-(2-\mu)} q$, and ${}_x D_b^{-(2-\mu)} Dq = D_x D_b^{-(2-\mu)} q$, when $q \in H_0^{\frac{\mu}{2}}(\Omega)$, are also utilized. \square

The following theorem presents the stability results for the weak semidiscrete problem (3.10).

THEOREM 3.2. *Let $U''(x) \leq 0$ for any $x \in \Omega$. The weak semidiscrete scheme (3.10) is unconditionally stable in the sense that for any time-step length $k > 0$, it holds that*

$$\|p^{m+1}\|_{H_0^{\frac{\mu}{2}}(\Omega)} \leq \|p^0\|_{L^2}, \quad m = 0, 1, \dots, M-1.$$

Proof. The induction will be used to prove this theorem. When $m = 0$, we have

$$(p^1, q) - \alpha_0 \left(\left[\frac{\partial}{\partial x} \frac{U'(x)}{\eta_\alpha} + \kappa_\alpha \nabla^\mu \right] p^1, q \right) = (p^0, q) \quad \forall q \in H_0^{\frac{\mu}{2}}(\Omega).$$

Taking $q = p^1$ and using (2.5) and Lemma 3.1, we obtain

$$\|p^1\|_{H_0^{\frac{\mu}{2}}(\Omega)}^2 \leq (p^0, p^1).$$

Applying the inequality $\|p^1\|_{L^2(\Omega)} \leq \|p^1\|_{H_0^{\frac{\mu}{2}}(\Omega)}$ and the Schwarz inequality, we attain immediately

$$\|p^1\|_{H_0^{\frac{\mu}{2}}(\Omega)} \leq \|p^0\|_{L^2(\Omega)}.$$

Assume we have proven

$$(3.11) \quad \|p^j\|_{H_0^{\frac{\mu}{2}}(\Omega)} \leq \|p^0\|_{L^2(\Omega)}, \quad j = 1, 2, \dots, m;$$

we will prove $\|p^{m+1}\|_{H_0^{\frac{\mu}{2}}(\Omega)} \leq \|p^0\|_{L^2(\Omega)}$. Letting $q = p^{m+1}$ in (3.10) gives

$$\begin{aligned} & (p^{m+1}, p^{m+1}) - \alpha_0 \left(\left[\frac{\partial}{\partial x} \frac{U'(x)}{\eta_\alpha} + \kappa_\alpha \nabla^\mu \right] p^{m+1}, p^{m+1} \right) \\ &= (1 - d_1) (p^m, p^{m+1}) + \sum_{j=1}^{m-1} (d_j - d_{j+1}) (p^{m-j}, p^{m+1}) + d_m (p^0, p^{m+1}). \end{aligned}$$

Further using Lemma 3.1 and (3.11), we have

$$\begin{aligned} \|p^{m+1}\|_{H_0^{\frac{\mu}{2}}(\Omega)}^2 &\leq (1 - d_1) \|p^m\|_{L^2(\Omega)} \|p^{m+1}\|_{L^2(\Omega)} \\ &\quad + \sum_{j=1}^{m-1} (d_j - d_{j+1}) \|p^{m-j}\|_{L^2(\Omega)} \|p^{m+1}\|_{L^2(\Omega)} \\ &\quad + d_m \|p^0\|_{L^2(\Omega)} \|p^{m+1}\|_{L^2(\Omega)} \\ &\leq \left[(1 - d_1) + \sum_{j=1}^{m-1} (d_j - d_{j+1}) + d_m \right] \|p^0\|_{L^2(\Omega)} \|p^{m+1}\|_{H_0^{\frac{\mu}{2}}(\Omega)} \\ &= \|p^0\|_{L^2(\Omega)} \|p^{m+1}\|_{H_0^{\frac{\mu}{2}}(\Omega)}. \end{aligned}$$

Then we conclude

$$\|p^{m+1}\|_{H_0^{\frac{\mu}{2}}(\Omega)} \leq \|p^0\|_{L^2(\Omega)}. \quad \square$$

By virtue of (2.3), the weak semidiscrete scheme (3.10) is formally of $(2 - \alpha)$ -order accuracy. Now we carry out the rigorous error analysis. From now on, we denote by c (or $c(\alpha_0)$) if it depends on α_0 given in (2.5)) a generic constant which may not be the same at different occurrences.

THEOREM 3.3. *Let $U''(x) \leq 0$ for any $x \in \Omega$, p be the exact solution of (3.1)–(3.3), and $\{p^m\}_{m=0}^M$ be the solution of (3.10) with the boundary and initial conditions (3.8) and (3.9); then the error estimates are*

(1) when $0 < \alpha < 1$,

$$(3.12) \quad \|p(t_m) - p^m\|_{H_0^{\frac{\mu}{2}}(\Omega)} \leq c_{p,\alpha} T^\alpha k^{2-\alpha}, \quad m = 1, 2, \dots, M,$$

where $c_{p,\alpha} := c_p / (1 - \alpha)$, with constant c_p defined in (3.15);

(2) when $\alpha \rightarrow 1$,

$$(3.13) \quad \|p(t_m) - p^m\|_{H_0^{\frac{\mu}{2}}(\Omega)} \leq c_p T k, \quad m = 1, 2, \dots, M,$$

where c_p is defined in (3.15).

For the convenience of denotation in the proof, let us define the error term by

$$(3.14) \quad r^{m+1} := \alpha_0 r_k^{m+1};$$

then from (2.2) we have

$$(3.15) \quad |r^{m+1}| = \Gamma(2 - \alpha) k^\alpha |r_k^{m+1}| \leq \Gamma(2 - \alpha) \tilde{c}_p k^2 = c_p k^2.$$

Proof of Theorem 3.3. (1) The proof of the idea using the mathematical induction is similar to Theorem 3.2 of [27]. Let us first consider the case $0 \leq \alpha < 1$. We would like to prove the estimate

$$(3.16) \quad \|p(t_j) - p^j\|_{H_0^{\frac{\mu}{2}}(\Omega)} \leq c_p d_{j-1}^{-1} k^2, \quad j = 1, 2, \dots, M.$$

Denote $e^m = p(t_m) - p^m$. For $j = 1$, from (3.1), (3.6), and (3.15), we obtain the error equation

$$(3.17) \quad (e^1, q) - \alpha_0 \left(\left[\frac{\partial U'(x)}{\partial x} \frac{1}{\eta_\alpha} + \kappa_\alpha \nabla^\mu \right] e^1, q \right) = (r^1, q) \quad \forall q \in H_0^{\frac{\mu}{2}}(\Omega).$$

Taking $q = e^1$ and using Lemma 3.1 yields

$$\|e^1\|_{H_0^{\frac{\mu}{2}}(\Omega)}^2 \leq \|r^1\|_{L^2(\Omega)} \|e^1\|_{L^2(\Omega)}.$$

This, together with (3.15), gives

$$\|p(t_1) - p^1\|_{H_0^{\frac{\mu}{2}}(\Omega)} \leq c_p d_0^{-1} k^2.$$

So, (3.16) is verified for the case $j = 1$. Suppose now that (3.16) holds for all $j = 1, 2, \dots, m$; we prove it holds also for $j = m + 1$.

Combining (3.1), (3.10), and (3.15) leads to

$$(3.18) \quad \begin{aligned} & (e^{m+1}, q) - \alpha_0 \left(\left[\frac{\partial U'(x)}{\partial x} \frac{1}{\eta_\alpha} + \kappa_\alpha \nabla^\mu \right] e^{m+1}, q \right) \\ &= (1 - d_1)(e^m, q) + \sum_{j=1}^{m-1} (d_j - d_{j+1})(e^{m-j}, q) + (r^{m+1}, q) \quad \forall q \in H_0^{\frac{\mu}{2}}(\Omega). \end{aligned}$$

Taking $q = e^{m+1}$ in (3.18) and using Lemma 3.1 yields

$$\begin{aligned} \|e^{m+1}\|_{H_0^{\frac{\mu}{2}}(\Omega)}^2 &\leq (1 - d_1) \|e^m\|_{L^2(\Omega)} \|e^{m+1}\|_{L^2(\Omega)} \\ &\quad + \sum_{j=1}^{m-1} (d_j - d_{j+1}) \|e^{m-j}\|_{L^2(\Omega)} \|e^{m+1}\|_{L^2(\Omega)} \\ &\quad + \|r^{m+1}\|_{L^2(\Omega)} \|e^{m+1}\|_{L^2(\Omega)}. \end{aligned}$$

According to the induction assumption, we have

$$\begin{aligned} \|e^{m+1}\|_{H_0^{\frac{\mu}{2}}(\Omega)} &\leq \left((1 - d_1) d_{m-1}^{-1} + \sum_{j=1}^{m-1} (d_j - d_{j+1}) d_{m-j-1}^{-1} \right) c_p k^2 + c_p k^2 \\ &\leq \left((1 - d_1) + \sum_{j=1}^{m-1} (d_j - d_{j+1}) + d_m \right) c_p d_m^{-1} k^2 \\ &= c_p d_m^{-1} k^2. \end{aligned}$$

Therefore, the estimate (3.16) is proved.

A direct computation shows that $m^{-\alpha}d_{m-1}^{-1} = 1$ when $m = 1$, and $m^{-\alpha}d_{m-1}^{-1}$ increasingly tends to $\frac{1}{1-\alpha}$ as $1 < m \rightarrow +\infty$. Therefore,

$$(3.19) \quad m^{-\alpha}d_{m-1}^{-1} \leq \frac{1}{1-\alpha}, \quad m = 1, 2, \dots, M.$$

For all m such that $mk \leq T$, we obtain

$$\begin{aligned} \|p(t_m) - p^m\|_{H_0^{\frac{\mu}{2}}(\Omega)} &\leq c_p d_{m-1}^{-1} k^2 = c_p m^{-\alpha} d_{m-1}^{-1} m^\alpha k^2 \\ &\leq c_p \frac{1}{1-\alpha} (mk)^\alpha k^{2-\alpha} = c_{p,\alpha} T^\alpha k^{2-\alpha}. \end{aligned}$$

(2) Further consider the case $\alpha \rightarrow 1$. In this case, $c_{p,\alpha}$ tends to infinity as $\alpha \rightarrow 1$, so the estimate (3.12) has no meaning. We need an estimate of another form. The proof is similar to the procedure in case (1).

Noting the fact that $jk \leq T$ for all $j = 1, 2, \dots, M$, we want to prove

$$(3.20) \quad \|p(t_j) - p^j\|_{H_0^{\frac{\mu}{2}}(\Omega)} \leq c_p j k^2, \quad j = 1, 2, \dots, M.$$

When $j = 1$, the estimate holds from (3.16). Suppose now that (3.20) holds for $j = 1, 2, \dots, m$; we prove it also remains true for $j = m + 1$. Based on the induction assumption and the estimates of the right-hand side of the error equation, the details of which are omitted, we get

$$\|e^{m+1}\|_{H_0^{\frac{\mu}{2}}(\Omega)} \leq \left((1 - d_1) + \sum_{j=1}^{m-1} (d_j - d_{j+1}) + d_m \right) c_p (m+1) k^2 = c_p (m+1) k^2.$$

Then (3.20) holds, and (3.13) is proved. \square

3.2. Existence, uniqueness, and regularity of weak solutions for (3.10).

First, we further study the variational problem (3.10) for fixed m , where we suppose p^j , $j = 1, 2, \dots, m$, are known quantities and p^{m+1} is an unknown variable. We introduce the following denotations:

$$\begin{aligned} \tilde{B}(p^{m+1}, q) &:= (p^{m+1}, q) + \alpha_0 B(p^{m+1}, q), \\ f &:= (1 - d_1)p^m + \sum_{j=1}^{m-1} (d_j - d_{j+1})p^{m-j} + d_m p^0, \end{aligned}$$

and

$$\tilde{F}(q) := (f, q).$$

Then we can recast (3.10) as

$$(3.21) \quad \tilde{B}(p^{m+1}, q) = \tilde{F}(q) \quad \forall q \in H_0^{\frac{\mu}{2}}(\Omega).$$

We show that there exists a unique solution to (3.21). To do this, we need to establish the coercivity and continuity of $\tilde{B}(\cdot, \cdot)$.

LEMMA 3.4. *Let $U''(x) \leq 0$ for any $x \in \Omega$ and $q \in H_0^{\frac{\mu}{2}}(\Omega)$. The bilinear form $\tilde{B}(q, q) \geq \|q\|_{H_0^{\frac{\mu}{2}}(\Omega)}^2$; i.e., it is coercive over $H_0^{\frac{\mu}{2}}(\Omega)$.*

Proof. According to Lemma 3.1,

$$\begin{aligned}\tilde{B}(q, q) &= (q, q) + \alpha_0 B(q, q) \\ &\geq (q, q) + \alpha_0 \kappa_\alpha \left| \cos\left(\frac{\mu}{2}\pi\right) \right| \|q\|_{H_0^{\frac{\mu}{2}}(\Omega)}^2 \\ &= \|q\|_{H_0^{\frac{\mu}{2}}(\Omega)}^2.\end{aligned}$$

The proof is completed. \square

LEMMA 3.5. *The bilinear form $\tilde{B}(\cdot, \cdot)$ is continuous on $H_0^{\frac{\mu}{2}}(\Omega) \times H_0^{\frac{\mu}{2}}(\Omega)$; i.e., there exists a constant $c(\alpha_0)$ such that*

$$\left| \tilde{B}(p, q) \right| \leq c(\alpha_0) \|p\|_{H_0^{\frac{\mu}{2}}(\Omega)} \|q\|_{H_0^{\frac{\mu}{2}}(\Omega)} \quad \forall p, q \in H_0^{\frac{\mu}{2}}(\Omega).$$

Proof. From the definition of $|\tilde{B}(p, q)|$ we have

$$\begin{aligned}\left| \tilde{B}(p, q) \right| &= |(p, q) + \alpha_0 B(p, q)| \\ &= \left| (p, q) - \alpha_0 \int_a^b \frac{U''(x)}{\eta_\alpha} p q dx - \alpha_0 \int_a^b \frac{U'(x)}{\eta_\alpha} \frac{\partial p}{\partial x} q dx \right. \\ &\quad \left. - \alpha_0 \frac{\kappa_\alpha}{2} \int_a^b \left(D^2 {}_a D_x^{-(2-\mu)} p \right) q dx - \alpha_0 \frac{\kappa_\alpha}{2} \int_a^b \left(D^2 {}_x D_b^{-(2-\mu)} p \right) q dx \right| \\ &\leq |(p, q)| + \alpha_0 \left| \int_a^b \frac{U''(x)}{\eta_\alpha} p q dx \right| + \alpha_0 \left| \int_a^b \frac{U'(x)}{\eta_\alpha} \frac{\partial p}{\partial x} q dx \right| \\ &\quad + \alpha_0 \left| \frac{\kappa_\alpha}{2} \int_a^b (\mathbf{D}^\mu p) q dx \right| + \alpha_0 \left| \frac{\kappa_\alpha}{2} \int_a^b (\mathbf{D}^{\mu*} p) q dx \right|.\end{aligned}$$

Using the equivalence of norms, the fractional Poincaré–Friedrichs inequality, and Lemma 2.13,

$$\begin{aligned}\left| \int_a^b \frac{U'(x)}{\eta_\alpha} \frac{\partial p}{\partial x} q dx \right| &= \left| \int_a^b \mathbf{D}^{\frac{\mu}{2}} p \mathbf{D}^{(1-\frac{\mu}{2})*} \left(\frac{U'(x)}{\eta_\alpha} q \right) \right| \\ &\leq \|p\|_{J_{L,0}^{\frac{\mu}{2}}(\Omega)} \left| \frac{U'(x)}{\eta_\alpha} q \right|_{J_{R,0}^{1-\frac{\mu}{2}}(\Omega)} \\ &\leq c(\alpha_0) \|p\|_{H_0^{\frac{\mu}{2}}(\Omega)} \|q\|_{H_0^{\frac{\mu}{2}}(\Omega)}\end{aligned}$$

and

$$\begin{aligned}\left| \frac{\kappa_\alpha}{2} \int_a^b (\mathbf{D}^\mu p) q dx \right| &= \left| \frac{\kappa_\alpha}{2} \int_a^b \left(\mathbf{D}^{\frac{\mu}{2}} p \right) \mathbf{D}^{\frac{\mu}{2}*} q dx \right| \\ &\leq \frac{\kappa_\alpha}{2} \|p\|_{J_{L,0}^{\frac{\mu}{2}}(\Omega)} \|q\|_{J_{R,0}^{\frac{\mu}{2}}(\Omega)} \\ &\leq c(\alpha_0) \|p\|_{H_0^{\frac{\mu}{2}}(\Omega)} \|q\|_{H_0^{\frac{\mu}{2}}(\Omega)},\end{aligned}$$

and similarly

$$\left| \frac{\kappa_\alpha}{2} \int_a^b (\mathbf{D}^{\mu*} p) q dx \right| = \left| \frac{\kappa_\alpha}{2} \int_a^b (\mathbf{D}^{\frac{\mu}{2}*} p) \mathbf{D}^{\frac{\mu}{2}} q dx \right| \leq c(\alpha_0) \|p\|_{H_0^{\frac{\mu}{2}}(\Omega)} \|q\|_{H_0^{\frac{\mu}{2}}(\Omega)}.$$

Thus,

$$\begin{aligned} & \left| \tilde{B}(p, q) \right| \\ & \leq \|p\|_{L^2(\Omega)} \|q\|_{L^2(\Omega)} + \alpha_0 \left\| \frac{U''(x)}{\eta_\alpha} \right\|_\infty \|p\|_{L^2(\Omega)} \|q\|_{L^2(\Omega)} + c(\alpha_0) \|p\|_{H_0^{\frac{\mu}{2}}(\Omega)} \|q\|_{H_0^{\frac{\mu}{2}}(\Omega)} \\ & \leq c(\alpha_0) \|p\|_{H_0^{\frac{\mu}{2}}(\Omega)} \|q\|_{H_0^{\frac{\mu}{2}}(\Omega)}. \end{aligned}$$

The proof is completed. \square

LEMMA 3.6. *The linear functional $\tilde{F}(\cdot)$ is continuous over $H_0^{\frac{\mu}{2}}(\Omega)$.*

Proof. Using Theorem 3.2 and Minkowski's inequality leads to $f \in H_0^{-\frac{\mu}{2}}(\Omega) \subset H^{-\frac{\mu}{2}}(\Omega)$. Then the result of this lemma follows from

$$|\tilde{F}(q)| \leq \|f\|_{H^{-\frac{\mu}{2}}(\Omega)} \|q\|_{H^{\frac{\mu}{2}}(\Omega)}. \quad \square$$

According to Lemmas 3.4, 3.5, and 3.6, \tilde{B} and \tilde{F} satisfy the hypotheses of the Lax–Milgram theorem, and the solution p^{m+1} is bounded by f . So we have the following theorem.

THEOREM 3.7. *Let $U''(x) \leq 0$ for any $x \in \Omega$. There exists a unique solution $p^{m+1} \in H_0^{\frac{\mu}{2}}(\Omega)$ to (3.10) satisfying*

$$(3.22) \quad \|p^{m+1}\|_{H_0^{\frac{\mu}{2}}(\Omega)} \leq \|f\|_{H^{-\frac{\mu}{2}}(\Omega)}.$$

In order to establish the error estimate in the following subsection, we need to prove $\|p^{m+1}\|_{H_0^\mu(\Omega)} \leq \|f\|_{L^2(\Omega)}$. For the proof of this result, we need the following lemma.

LEMMA 3.8. *Let $U''(x) \leq 0$ for any $x \in \Omega$, $f \in L^2(\Omega)$, and p^{m+1} satisfy (3.7)–(3.9). If, in addition, for $\frac{\mu}{2} \leq s < 1$, $p^{m+1} \in H^s(\Omega) \cap H_0^{\frac{\mu}{2}}(\Omega)$ and satisfies*

$$(3.23) \quad \|p^{m+1}\|_{H^s(\Omega)} \leq c(\alpha_0) \|f\|_{L^2(\Omega)},$$

then $p^{m+1} \in H^{s+\frac{\mu}{2}-\frac{1}{2}}(\Omega)$ with estimate

$$(3.24) \quad \|p^{m+1}\|_{H^{s+\frac{\mu}{2}-\frac{1}{2}}(\Omega)} \leq c(\alpha_0) \|f\|_{L^2(\Omega)}.$$

Proof. Note that as $\frac{1}{2} < s < 1$, $0 < 2s - 1 < 1$, we can write

$$\mathbf{D}^{2s-1} p^{m+1} = {}_a D_x^{-\gamma} D p^{m+1}, \quad \text{with } \gamma = 2 - 2s.$$

Since p^{m+1} satisfies

$$\left(1 - \alpha_0 \frac{U''(x)}{\eta_\alpha}\right) p^{m+1} - \alpha_0 \left(\frac{U'(x)}{\eta_\alpha} \frac{\partial}{\partial x} + \frac{\kappa_\alpha}{2} \left(D^2 {}_a D_x^{-(2-\mu)} + D^2 {}_x D_b^{-(2-\mu)}\right)\right) p^{m+1} = f,$$

multiplying both sides by $\mathbf{D}^{2s-1} p^{m+1}$ and integrating over Ω , we have

$$\begin{aligned} & \frac{\alpha_0 \kappa_\alpha}{2} \left(\left(D^2 {}_a D_x^{-(2-\mu)} p^{m+1}, {}_a D_x^{-\gamma} D p^{m+1} \right) + \left(D^2 {}_x D_b^{-(2-\mu)} p^{m+1}, {}_a D_x^{-\gamma} D p^{m+1} \right) \right) \\ &= \left(\left(1 - \alpha_0 \frac{U''(x)}{\eta_\alpha} \right) p^{m+1}, {}_a D_x^{-\gamma} D p^{m+1} \right) + \alpha_0 \left(\frac{U'(x)}{\eta_\alpha} \frac{\partial p^{m+1}}{\partial x}, {}_a D_x^{-\gamma} D p^{m+1} \right) \\ & \quad - \left(f, {}_a D_x^{-\gamma} D p^{m+1} \right). \end{aligned}$$

First, we bound each of the terms on the right-hand side of the above equation. Using the Cauchy–Schwarz inequality and (3.23) leads to

$$\begin{aligned} \left| - \left(f, {}_a D_x^{-\gamma} D p^{m+1} \right) \right| &\leq \|f\|_{L^2(\Omega)} \left| P^{m+1} \right|_{J_L^{2s-1}(\Omega)} \\ &\leq c \|f\|_{L^2(\Omega)} \left| P^{m+1} \right|_{J_L^s(\Omega)}, \quad \text{since } 2s-1 < s, \\ &\leq c \|f\|_{L^2(\Omega)} \left| P^{m+1} \right|_{H^s(\Omega)} \\ &\leq c \|f\|_{L^2(\Omega)}. \end{aligned}$$

Because of Lemma 2.13, we have

$$\begin{aligned} & \left| \left(\frac{U'(x)}{\eta_\alpha} \frac{\partial p^{m+1}}{\partial x}, {}_a D_x^{-\gamma} D p^{m+1} \right) \right| \\ &= \left| \left({}_x D_b^{-\frac{\gamma}{2}} \left(\frac{U'(x)}{\eta_\alpha} \frac{\partial p^{m+1}}{\partial x} \right), {}_a D_x^{-\frac{\gamma}{2}} D p^{m+1} \right) \right| \\ &\leq \left\| {}_x D_b^{-\frac{\gamma}{2}} \left(\frac{U'(x)}{\eta_\alpha} \frac{\partial p^{m+1}}{\partial x} \right) \right\|_{L^2(\Omega)} \left\| {}_a D_x^{-\frac{\gamma}{2}} D p^{m+1} \right\|_{L^2(\Omega)} \\ &= \left\| {}_x D_b^{-\frac{\gamma}{2}} \left(D \left(\frac{U'(x)}{\eta_\alpha} p^{m+1} \right) - D \left(\frac{U'(x)}{\eta_\alpha} \right) p^{m+1} \right) \right\|_{L^2(\Omega)} \left\| {}_a D_x^{-\frac{\gamma}{2}} D p^{m+1} \right\|_{L^2(\Omega)} \\ &\leq \left(\left\| \mathbf{D}^{s*} \left(\frac{U'(x)}{\eta_\alpha} p^{m+1} \right) \right\|_{L^2(\Omega)} + \left\| D \left(\frac{U'(x)}{\eta_\alpha} \right) p^{m+1} \right\|_{L^2(\Omega)} \right) \left\| {}_a D_x^{-\frac{\gamma}{2}} D p^{m+1} \right\|_{L^2(\Omega)} \\ &\leq (c \|p^{m+1}\|_{H^s(\Omega)} + c \|p^{m+1}\|_{L^2(\Omega)}) \|p^{m+1}\|_{H^s(\Omega)} \\ &\leq c(\alpha_0) \|p^{m+1}\|_{H^s(\Omega)}. \end{aligned}$$

For the remaining term we have

$$\begin{aligned} \left| \left(\left(1 - \alpha_0 \frac{U''(x)}{\eta_\alpha} \right) p^{m+1}, {}_a D_x^{-\gamma} D p^{m+1} \right) \right| &\leq \left\| 1 - \alpha_0 \frac{U''(x)}{\eta_\alpha} \right\|_\infty \|p^{m+1}\|_{L^2(\Omega)} \left| p^{m+1} \right|_{J_L^{2s-1}(\Omega)} \\ &\leq c \|f\|_{L^2(\Omega)}. \end{aligned}$$

Further we bound the two terms of the left-hand side:

$$\begin{aligned}
& \left| \left(D^2 {}_a D_x^{-(2-\mu)} p^{m+1}, {}_a D_x^{-\gamma} D p^{m+1} \right) \right| \\
&= - \left(\mathbf{D}^{\mu-1} p^{m+1}, \mathbf{D}^{2s} p^{m+1} \right) \\
&= - \left(\mathbf{D}^{\mu-1} p^{m+1}, \mathbf{D}^{2s-\mu+1} \mathbf{D}^{\mu-1} p^{m+1} \right) \\
&= - \left(\mathbf{D}^{(s-\frac{\mu}{2}+\frac{1}{2})^*} \mathbf{D}^{\mu-1} p^{m+1}, \mathbf{D}^{s-\frac{\mu}{2}+\frac{1}{2}} \mathbf{D}^{\mu-1} p^{m+1} \right) \\
&= \left| \mathbf{D}^{\mu-1} p^{m+1} \right|_{J_s^{s-\frac{\mu}{2}+\frac{1}{2}}(\Omega)}^2 \\
&\geq c \left| \mathbf{D}^{\mu-1} p^{m+1} \right|_{J_L^{s-\frac{\mu}{2}+\frac{1}{2}}(\Omega)}^2 \\
&= c \left| p^{m+1} \right|_{J_L^{s+\frac{\mu}{2}-\frac{1}{2}}(\Omega)}^2.
\end{aligned}$$

Similarly, we obtain

$$\left| \left(D^2 {}_x D_b^{-(2-\mu)} p^{m+1}, {}_a D_x^{-\gamma} D p^{m+1} \right) \right| \geq c \left| p^{m+1} \right|_{J_R^{s+\frac{\mu}{2}-\frac{1}{2}}(\Omega)}^2.$$

Based on the above estimates and the equivalence of the spaces $J_L^{s+\frac{\mu}{2}-\frac{1}{2}}(\Omega)$, $J_R^{s+\frac{\mu}{2}-\frac{1}{2}}(\Omega)$, $H^{s+\frac{\mu}{2}-\frac{1}{2}}(\Omega)$, and $\left| p^{m+1} \right|_{L^2(\Omega)}^2 \leq c \left| p^{m+1} \right|_{J_R^{s+\frac{\mu}{2}-\frac{1}{2}}(\Omega)}^2$, the stated results follow. \square

THEOREM 3.9. *Let $U''(x) \leq 0$ for any $x \in \Omega$, $f \in L^2(\Omega)$, and p^{m+1} satisfy (3.7)–(3.9). Then $p^{m+1} \in H^\mu(\Omega)$ with*

$$(3.25) \quad \left\| p^{m+1} \right\|_{H^\mu(\Omega)} \leq c(\alpha_0) \|f\|_{L^2(\Omega)}, \quad \mu \neq \frac{3}{2},$$

$$(3.26) \quad \left\| p^{m+1} \right\|_{H^{\mu-\epsilon}(\Omega)} \leq c(\alpha_0) \|f\|_{L^2(\Omega)}, \quad \mu = \frac{3}{2}, \quad 0 < \epsilon < \frac{1}{2}.$$

Proof. First, we must establish an estimate for p^{m+1} in the H^1 norm. In order to do this, we repeatedly use Lemma 3.8 in an induction argument.

Suppose that $p^{m+1} \in H^{\frac{j\mu}{2}-\frac{j-1}{2}}(\Omega)$ for $j \in N$ such that $\frac{j\mu}{2} - \frac{j-1}{2} < 1$, with

$$\left\| p^{m+1} \right\|_{H^{\frac{j\mu}{2}-\frac{j-1}{2}}(\Omega)} \leq c(\alpha_0) \|f\|_{L^2(\Omega)}.$$

By Lemma 3.8, $p^{m+1} \in H^{\frac{(j+1)\mu}{2}-\frac{j}{2}}$ with

$$\left\| p^{m+1} \right\|_{H^{\frac{(j+1)\mu}{2}-\frac{j}{2}}(\Omega)} \leq c(\alpha_0) \|f\|_{L^2(\Omega)}.$$

For $j = 1$ we have that $p^{m+1} \in H_0^{\frac{\mu}{2}}(\Omega)$, so the initial step is valid. We repeat this process for $j = 1, j = 2, j = \dots$, until $j = K$, where

$$\frac{K\mu}{2} - \frac{K-1}{2} < 1 \quad \text{and} \quad \frac{(K+1)\mu}{2} - \frac{K}{2} \geq 1.$$

For this value of j , we have $p^{m+1} \in H_0^1(\Omega) \subset H^{\frac{(K+1)\mu}{2}-\frac{K}{2}}(\Omega)$ with the estimate

$$\left\| p^{m+1} \right\|_{H^1(\Omega)} \leq c(\alpha_0) \|f\|_{L^2(\Omega)}.$$

Now, multiplying both sides of (3.7) by $\mathbf{D}^\mu p^{m+1}$ and integrating over Ω , similar to the estimates in the proof of Lemma 3.8, we have

$$c(\alpha_0) \|p^{m+1}\|_{J_L^\mu(\Omega)}^2 + c(\alpha_0) \|p^{m+1}\|_{J_L^\mu(\Omega)} \|p^{m+1}\|_{J_R^\mu(\Omega)} \leq c(\alpha_0) \|f\|_{L^2(\Omega)} \|p^{m+1}\|_{J_L^\mu(\Omega)},$$

where the three $c(\alpha_0)$ terms generally are not equal.

When $\mu \neq \frac{3}{2}$, the three norms $J_L^\mu(\Omega)$, $J_R^\mu(\Omega)$, and $H^\mu(\Omega)$ are equivalent; then the stated result (3.25) follows. As $\mu = \frac{3}{2}$, because of the fact that $0 < \epsilon < \frac{1}{2}$, the following holds:

$$|p^{m+1}|_{H^{\mu-\epsilon}(\Omega)} \leq c |p^{m+1}|_{J_L^\mu(\Omega)} \quad \text{and} \quad |p^{m+1}|_{H^{\mu-\epsilon}(\Omega)} \leq c |p^{m+1}|_{J_R^\mu(\Omega)},$$

so the stated result (3.26) holds. \square

3.3. Finite element approximation and error estimates in the space and error estimates for full discretization. Denote $\{S_h\}$ to be a family of partitions of Ω with grid parameter h , and associated with S_h define the finite-dimensional subspace X_h to be the basis of the piecewise polynomials of order $n-1$, where $n \in \mathbb{N}$. Denote $I^h p^{m+1}$ to be the piecewise interpolation polynomial of p^{m+1} in S_h , being uniquely determined. We have the well-known approximation property.

LEMMA 3.10 (approximation property [6]). *Let $p^{m+1} \in H^l(\Omega)$, $0 < l \leq n$, and $0 \leq s \leq l$. Then there exists a constant $c(\alpha_0)$ depending on Ω such that*

$$(3.27) \quad \|p^{m+1} - I^h p^{m+1}\|_{H^s(\Omega)} \leq c(\alpha_0) h^{l-s} \|p^{m+1}\|_{H^l(\Omega)}.$$

Let \tilde{p}_h^{m+1} be the solution of the finite-dimensional variational problem

$$(3.28) \quad \tilde{B}(\tilde{p}_h^{m+1}, q_h) = \tilde{F}(q_h) \quad \forall q_h \in X_h.$$

We define the energy norm associated with (3.21) as

$$(3.29) \quad \|p^{m+1}\|_E := \tilde{B}(p^{m+1}, p^{m+1})^{\frac{1}{2}}.$$

According to Lemmas 3.4 and 3.5, we have the norm equivalence of $\|\cdot\|_E$ and $\|\cdot\|_{H_0^{\frac{\mu}{2}}(\Omega)}$.

THEOREM 3.11. *Let $U''(x) \leq 0$ for any $x \in \Omega$ and p^{m+1} be the solution to (3.21). There exists a unique solution to (3.28) satisfying the estimate*

$$(3.30) \quad \|p^{m+1} - \tilde{p}_h^{m+1}\|_E \leq c(\alpha_0) \inf_{q_h \in X_h} \|p^{m+1} - q_h\|_E \leq c(\alpha_0) \|p^{m+1} - I^h p^{m+1}\|_E.$$

Proof. Since X_h is a subset of the space $H_0^{\frac{\mu}{2}}(\Omega)$, (3.28) satisfies the hypothesis of the Lax–Milgram lemma over the finite-dimensional subspace. Then the existence and uniqueness hold for (3.28). The estimate (3.30) follows from Céa’s lemma. \square

Combining the previous results into an estimate for $p^{m+1} - \tilde{p}_h^{m+1}$ leads to the following corollary.

COROLLARY 3.12. *Let $U''(x) \leq 0$ for any $x \in \Omega$, $p^{m+1} \in H_0^{\frac{\mu}{2}}(\Omega) \cap H^l(\Omega)$ ($\frac{\mu}{2} \leq l \leq n$) solve (3.21), and \tilde{p}_h^{m+1} solve (3.28). Then there exists a constant $c(\alpha_0)$ such that*

$$(3.31) \quad \|p^{m+1} - \tilde{p}_h^{m+1}\|_{H_0^{\frac{\mu}{2}}(\Omega)} \leq c(\alpha_0) h^{l-\frac{\mu}{2}} \|p^{m+1}\|_{H^l(\Omega)}.$$

Applying the Aubin–Nitsche trick derives the convergence estimate in the L^2 norm.

THEOREM 3.13. *Let $U''(x) \leq 0$ for any $x \in \Omega$, $p^{m+1} \in H_0^{\frac{\mu}{2}}(\Omega) \cap H^l(\Omega)$ ($\frac{\mu}{2} \leq l \leq n$) solve (3.21), and \tilde{p}_h^{m+1} solve (3.28). Then there exists a constant $c(\alpha_0)$ such that*

$$(3.32) \quad \|p^{m+1} - \tilde{p}_h^{m+1}\|_{L^2(\Omega)} \leq c(\alpha_0)h^l \|p^{m+1}\|_{H^l(\Omega)}, \quad \mu \neq \frac{3}{2},$$

$$(3.33) \quad \|p^{m+1} - \tilde{p}_h^{m+1}\|_{L^2(\Omega)} \leq c(\alpha_0)h^{l-\epsilon} \|p^{m+1}\|_{H^l(\Omega)}, \quad \mu = \frac{3}{2}, \quad 0 < \epsilon < \frac{1}{2}.$$

Proof. Introduce the adjoint problem: Find $\omega \in H_0^{\frac{\mu}{2}}(\Omega)$ such that

$$(3.34) \quad \tilde{B}(\omega, q) = (p^{m+1} - \tilde{p}_h^{m+1}, q) \quad \forall q \in H_0^{\frac{\mu}{2}}(\Omega).$$

Theorem 3.9 implies that, for $\mu \neq \frac{3}{2}$, ω satisfies the regularity estimate

$$\|\omega\|_{H^\mu(\Omega)} \leq c(\alpha_0) \|p^{m+1} - \tilde{p}_h^{m+1}\|_{L^2(\Omega)}.$$

Substituting $q = p^{m+1} - \tilde{p}_h^{m+1}$ and using the Galerkin orthogonality, we obtain

$$\begin{aligned} \|p^{m+1} - \tilde{p}_h^{m+1}\|_{L^2(\Omega)}^2 &= \tilde{B}(\omega, p^{m+1} - \tilde{p}_h^{m+1}) \\ &= \tilde{B}(\omega - I^h\omega, p^{m+1} - \tilde{p}_h^{m+1}) \\ &\leq c(\alpha_0) \|\omega - I^h\omega\|_{H^{\frac{\mu}{2}}(\Omega)} \|p^{m+1} - \tilde{p}_h^{m+1}\|_{H^{\frac{\mu}{2}}(\Omega)} \\ &\leq c(\alpha_0)h^{\frac{\mu}{2}} \|\omega\|_{H^\mu(\Omega)} \|p^{m+1} - \tilde{p}_h^{m+1}\|_{H^{\frac{\mu}{2}}(\Omega)} \\ &\leq c(\alpha_0)h^{\frac{\mu}{2}} \|p^{m+1} - \tilde{p}_h^{m+1}\|_{L^2(\Omega)} \|p^{m+1} - \tilde{p}_h^{m+1}\|_{H^{\frac{\mu}{2}}(\Omega)}. \end{aligned}$$

Thus, we have the estimate

$$\|p^{m+1} - \tilde{p}_h^{m+1}\|_{L^2(\Omega)} \leq c(\alpha_0)h^{\frac{\mu}{2}} \|p^{m+1} - \tilde{p}_h^{m+1}\|_{H^{\frac{\mu}{2}}(\Omega)}.$$

Further applying (3.31), we get (3.32). Because of (3.26) in Theorem 3.9, we can similarly prove (3.33) with $\mu = \frac{3}{2}$. \square

Let p_h^j be the computational value at time t_j of full discretization; i.e., $\{p_h^j\}_{j=1}^M$ satisfy for all $q_h \in X_h$

$$(3.35) \quad \begin{aligned} &(p_h^{m+1}, q_h) - \alpha_0 \left(\left[\frac{\partial}{\partial x} \frac{U'(x)}{\eta_\alpha} + \kappa_\alpha \nabla^\mu \right] p_h^{m+1}, q_h \right) \\ &= \left((1 - d_1)p_h^m + \sum_{j=1}^{m-1} (d_j - d_{j+1})p_h^{m-j} + d_m p_h^0, q_h \right). \end{aligned}$$

Now we aim at deriving the estimates for $\|p(t_m) - p_h^m\|_{H_0^{\frac{\mu}{2}}(\Omega)}$ and $\|p(t_m) - p_h^m\|_{L^2(\Omega)}$, being given in the following theorem.

THEOREM 3.14. *Let $U''(x) \leq 0$ for any $x \in \Omega$, p be the exact solution of (3.1)–(3.3), and $\{p_h^m\}_{m=0}^M$ be the solution of problem (3.35) with the initial condition $p_h^0 = I^h g$. Suppose $p \in H^1([0, T], H_0^{\frac{\mu}{2}}(\Omega) \cap H^l(\Omega))$ ($\frac{\mu}{2} \leq l \leq n$), then we have*

(1) when $0 < \alpha < 1$,

$$\begin{aligned} \|p(t_m) - p_h^m\|_{H^{\frac{\mu}{2}}(\Omega)} &\leq \frac{c_p T^\alpha}{1-\alpha} \left(k^{2-\alpha} + c(\alpha_0) k^{-\alpha} h^{l-\frac{\mu}{2}} \|p\|_{L^\infty(H^l(\Omega))} \right), \\ \|p(t_m) - p_h^m\|_{L^2(\Omega)} &\leq \frac{c_p T^\alpha}{1-\alpha} \left(k^{2-\alpha} + c(\alpha_0) k^{-\alpha} h^l \|p\|_{L^\infty(H^l(\Omega))} \right), \quad \mu \neq \frac{3}{2}, \\ \|p(t_m) - p_h^m\|_{L^2(\Omega)} &\leq \frac{c_p T^\alpha}{1-\alpha} \left(k^{2-\alpha} + c(\alpha_0) k^{-\alpha} h^{l-\epsilon} \|p\|_{L^\infty(H^l(\Omega))} \right), \quad \mu = \frac{3}{2}, \quad 0 < \epsilon < \frac{1}{2}; \end{aligned}$$

(2) when $\alpha \rightarrow 1$,

$$\begin{aligned} \|p(t_m) - p_h^m\|_{H^{\frac{\mu}{2}}(\Omega)} &\leq c_p T \left(k + c(\alpha_0) k^{-1} h^{l-\frac{\mu}{2}} \|p\|_{L^\infty(H^l(\Omega))} \right), \\ \|p(t_m) - p_h^m\|_{L^2(\Omega)} &\leq c_p T \left(k + c(\alpha_0) k^{-1} h^l \|p\|_{L^\infty(H^l(\Omega))} \right), \quad \mu \neq \frac{3}{2}, \\ \|p(t_m) - p_h^m\|_{L^2(\Omega)} &\leq c_p T \left(k + c(\alpha_0) k^{-1} h^{l-\epsilon} \|p\|_{L^\infty(H^l(\Omega))} \right), \quad \mu = \frac{3}{2}, \quad 0 < \epsilon < \frac{1}{2}, \end{aligned}$$

where c_p is given in (3.15).

Proof. From (2.3), (3.4), (3.5), and (3.14), we know that $\{p(t_j)\}_{j=1}^M$ satisfy $\forall q \in H_0^{\frac{\mu}{2}}(\Omega)$

$$\begin{aligned} (3.36) \quad & (p(t_{m+1}), q) - \alpha_0 \left(\left[\frac{\partial}{\partial x} \frac{U'(x)}{\eta_\alpha} + \kappa_\alpha \nabla^\mu \right] p(t_{m+1}), q \right) \\ &= \left((1-d_1)p(t_m) + \sum_{j=1}^{m-1} (d_j - d_{j+1})p(t_{m-j}) + d_m p(t_0), q \right) + (r^{m+1}, q). \end{aligned}$$

Let $\pi_h^{\frac{\mu}{2}}$ be the $H^{\frac{\mu}{2}}$ -orthogonal projection operator from $H_0^{\frac{\mu}{2}}$ to X_h , associated with the energy norm defined in (3.29), i.e., for all $q_h \in X_h$,

$$\begin{aligned} (3.37) \quad & \left(\pi_h^{\frac{\mu}{2}} p(t_{m+1}), q_h \right) - \alpha_0 \left(\left[\frac{\partial}{\partial x} \frac{U'(x)}{\eta_\alpha} + \kappa_\alpha \nabla^\mu \right] \pi_h^{\frac{\mu}{2}} p(t_{m+1}), q_h \right) \\ &= \left((1-d_1)p(t_m) + \sum_{j=1}^{m-1} (d_j - d_{j+1})p(t_{m-j}) + d_m p(t_0), q_h \right) + (r^{m+1}, q_h). \end{aligned}$$

Let $\tilde{e}_h^{m+1} = \pi_h^{\frac{\mu}{2}} p(t_{m+1}) - p_h^{m+1}$ and $e_h^{m+1} = p(t_{m+1}) - p_h^{m+1}$; by subtracting (3.35) from (3.37), we obtain

$$\begin{aligned} (3.38) \quad & (\tilde{e}_h^{m+1}, q_h) - \alpha_0 \left(\left[\frac{\partial}{\partial x} \frac{U'(x)}{\eta_\alpha} + \kappa_\alpha \nabla^\mu \right] \tilde{e}_h^{m+1}, q_h \right) \\ &= \left((1-d_1)e_h^m + \sum_{j=1}^{m-1} (d_j - d_{j+1})e_h^{m-j} + d_m e_h^0, q_h \right) + (r^{m+1}, q_h). \end{aligned}$$

Taking $q_h = \tilde{e}_h^{m+1}$ in (3.38) and using the triangular inequality $\|e_h^{m+1}\|_{H^{\frac{\mu}{2}}(\Omega)} \leq \|\tilde{e}_h^{m+1}\|_{H^{\frac{\mu}{2}}(\Omega)} + \|p(t_{m+1}) - \pi_h^{\frac{\mu}{2}} p(t_{m+1})\|_{H^{\frac{\mu}{2}}(\Omega)}$, from Corollary 3.12 we have

$$\begin{aligned}
(3.39) \quad & \|e_h^{m+1}\|_{H^{\frac{\mu}{2}}(\Omega)} \leq (1-d_1)\|e_h^m\|_{L^2(\Omega)} + \sum_{j=1}^{m-1} (d_j - d_{j+1})\|e_h^{m-j}\|_{L^2(\Omega)} \\
& + d_m \|e_h^0\|_{L^2(\Omega)} + \|r^{m+1}\|_{L^2(\Omega)} + \left\| p(t_{m+1}) - \pi_h^{\frac{\mu}{2}} p(t_{m+1}) \right\|_{H^{\frac{\mu}{2}}(\Omega)} \\
& \leq (1-d_1)\|e_h^m\|_{L^2(\Omega)} + \sum_{j=1}^{m-1} (d_j - d_{j+1})\|e_h^{m-j}\|_{L^2(\Omega)} \\
& + d_m \|e_h^0\|_{L^2(\Omega)} + c_p k^2 + c(\alpha_0) h^{l-\frac{\mu}{2}} \|p(t_{m+1})\|_{H^1(\Omega)}.
\end{aligned}$$

Similarly, taking $q_h = \tilde{e}_h^{m+1}$ in (3.38) and using the triangular inequality $\|e_h^{m+1}\|_{L^2(\Omega)} \leq \|\tilde{e}_h^{m+1}\|_{L^2(\Omega)} + \|p(t_{m+1}) - \pi_h^{\frac{\mu}{2}} p(t_{m+1})\|_{L^2(\Omega)}$ and $\|\tilde{e}_h^{m+1}\|_{L^2(\Omega)} \leq \|\tilde{e}_h^{m+1}\|_{H^{\frac{\mu}{2}}(\Omega)}$, from Theorem 3.13 we have

$$\begin{aligned}
(3.40) \quad & \|e_h^{m+1}\|_{L^2(\Omega)} \leq (1-d_1)\|e_h^m\|_{L^2(\Omega)} + \sum_{j=1}^{m-1} (d_j - d_{j+1}) \|e_h^{m-j}\|_{L^2(\Omega)} \\
& + d_m \|e_h^0\|_{L^2(\Omega)} + \|r^{m+1}\|_{L^2(\Omega)} + \left\| p(t_{m+1}) - \pi_h^{\frac{\mu}{2}} p(t_{m+1}) \right\|_{L^2(\Omega)} \\
& \leq (1-d_1)\|e_h^m\|_{L^2(\Omega)} + \sum_{j=1}^{m-1} (d_j - d_{j+1}) \|e_h^{m-j}\|_{L^2(\Omega)} \\
& + d_m \|e_h^0\|_{L^2(\Omega)} + c_p k^2 + c(\alpha_0) h^{\tilde{l}} \|p(t_{m+1})\|_{H^1(\Omega)},
\end{aligned}$$

where $\tilde{l} = l$ if $\mu \neq \frac{3}{2}$ and $\tilde{l} = l - \epsilon$, $0 < \epsilon < \frac{1}{2}$, if $\mu = \frac{3}{2}$.

It is at this point that we distinguish two cases for α ; from (3.39) and (3.40) the stated results are obtained by following the same lines as in Theorem 3.3, so the remaining details are omitted. \square

4. Numerical experiments and simulation of physical systems. In this section, we confirm the theoretical analysis by doing numerical computations with a smooth initial condition and then simulate the real physical cases with the δ distribution as the initial condition. The trial space used in the space finite element approximation is taken as X_h to be the basis with piecewise linear polynomials, i.e., $n = 2$.

In this paper, both the space and time derivatives of (3.1) are fractional. It is worth noting that fractional derivatives are nonlocal operators. Consequently, when approximating the space derivative, a sparse coefficient matrix [14], having the characteristic of using a finite element basis for the classical differential equation, does not occur. For the time derivative, because of its nonlocal property, we need to do a summation operation, the right-hand side of (3.35), in each time step. But it is possible to reduce the computational cost, since the fractional derivative has the short memory principle [8].

First, for confirming the theoretical prediction, we consider the problem (3.1)–(3.3) but, without loss of generality, add a force term $h(x, t)$ on the right-hand side of (3.1). Now the problem has the analytical solution

$$p(x, t) = t^2(x-a)^2(x-b)^2,$$

if taking

$$U(x) = 3x, \quad \kappa_\alpha = \eta_\alpha = 1.$$

TABLE 4.1

Experimental error results with fixed time step size $k = 0.0001$, $a = -2$, $b = -2$.

h	$\ p(T) - p_h^M\ _{L^2(\Omega)}$	Convergent rate
1/2	$3.077979 \cdot 10^{-2}$	
1/4	$9.020087 \cdot 10^{-3}$	1.77
1/8	$2.603326 \cdot 10^{-3}$	1.79

TABLE 4.2

Computational error results with fixed space step length $h = 0.05$, $a = -2$, $b = -2$.

k	$\ p(T) - p_h^M\ _{L^2(\Omega)}$	Convergent rate
1/10	$1.646337 \cdot 10^{-2}$	
1/20	$8.455413 \cdot 10^{-3}$	0.96
1/40	$4.270158 \cdot 10^{-3}$	0.99

It can be checked that the corresponding initial condition and force term are, respectively,

$$g(x) = 0,$$

$$\begin{aligned} h(x, t) = & \frac{2\Gamma(2)}{\Gamma(3-\alpha)} t^{2-\alpha} (x-a)^2 (x-b)^2 - 6t^2 ((x-a)(x-b))^2 + (x-a)^2 (x-b) \\ & - \frac{1}{2} t^2 \left(\frac{\Gamma(5)}{\Gamma(5-\mu)} ((x-a)^{4-\mu} + (b-x)^{4-\mu}) - 2(b-a) \frac{\Gamma(4)}{\Gamma(4-\mu)} ((x-a)^{3-\mu} \right. \\ & \left. + (b-x)^{3-\mu}) + (b-a)^2 \frac{\Gamma(3)}{\Gamma(3-\mu)} ((x-a)^{2-\mu} + (b-x)^{2-\mu}) \right). \end{aligned}$$

We compute the errors $\|p(T) - p_h^M\|_{L^2(\Omega)}$ at time $T = 1$, with space fractional order $\mu = 1.8$ and time fractional order $\alpha = 0.8$. In Table 4.1, with fixed time-step size, choosing different space step lengths leads to the experimental convergent rate in space. In Table 4.2, we fix space step size; using different time step lengths leads to the numerical convergent rate in time.

Further we simulate the real physical cases for (3.1)–(3.3) with absorbing boundary conditions and still taking $U(x) = 3x$, $\kappa_\alpha = \eta_\alpha = 1$, but using the δ distribution as the initial condition. The noteworthy features are the appearance of cusps when the time fractional order $\alpha \neq 1.0$, and the tail of the probability density function p decays as a power law when the space fractional order $\mu \neq 2.0$. See Figure 1, where $p(x, t|0, 0)$ is the unnormalized probability density at (x, t) starting from $t = 0$ with $\delta(x)$ as the initial distribution, and $a = -2$, $b = -2$, $k = 0.01$, and $h = 0.02$.

5. Conclusions. So far, it seems that no other published research takes into account the numerical method and detailed numerical error analysis for the space and time fractional partial differential equation, besides the stochastic approach used in [30]. In this paper, we have developed and analyzed an efficient numerical method for a partial differential equation with fractional derivatives in both space and time, the space and time fractional Fokker-Planck equation, which has a strong physical background, for example, characterizing the anomalous diffusion with both traps and flights. First, based on the properties of the Riemann-Liouville and Caputo derivatives, the formulation of the original equation is transformed. The time fractional derivative is discretized by using a backward differentiation, having $(2 - \alpha)$ -order accuracy.

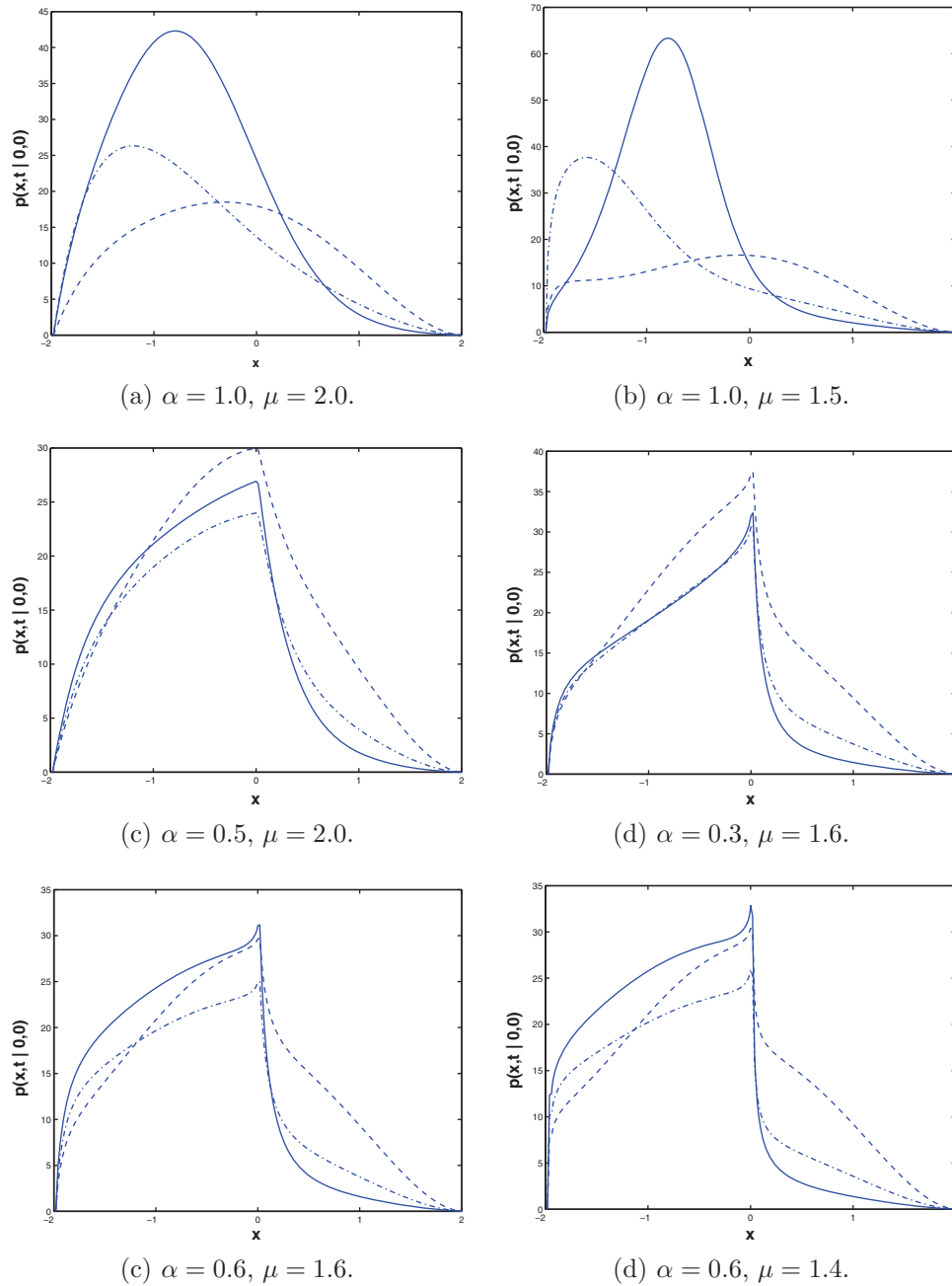


FIG. 1. The evolution of $p(x, t | 0, 0)$, where the solid line “—” stands for the solution when $t = 0.3$, the dashed-dotted line “-.” stands for the solution when $t = 0.6$, and the dashed line “--” stands for the solution when $t = 1.0$ ($p(x, t | 0, 0)$ is the unnormalized probability density at (x, t) starting from $t = 0$ with $\delta(x)$ as the initial distribution, and $a = -2$, $b = -2$, $k = 0.01$, and $h = 0.02$).

We then derived the variational formation of the semidiscrete scheme; its stability is rigorously proved. Use of the finite element method to approximate the space fractional derivative results in full discretization with μ -order convergence in space

and $(2 - \alpha)$ -order convergence in time. Our numerical experiments are in agreement with the theoretical analysis. The real physical cases are also simulated.

A key difference between fractional derivatives and the classical derivatives is that fractional derivatives are nonlocal operators. The matrix generated by finite element approximation in space is no longer sparse, and in the time direction, when advancing one step, we need to sum up all of the previous computed results' multiplied variational coefficients. So the computational cost, storage requirement, and time spent are expensive. Fortunately, although fractional derivatives have global dependence, they have the short memory principle. This principle works well for computing fractional ODEs [8]. Our future research along this direction is to investigate the effective ways of using the short memory principle for computing partial differential equations with both space and time fractional derivatives so as to reduce the computational cost.

REFERENCES

- [1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
- [2] O. P. AGRAWAL, J. A. TENREIRO MACHADO, AND J. SABATIER, *Introduction*, *Nonlinear Dynam.*, 38 (2004), pp. 1–2.
- [3] E. BARKAI, R. METZLER, AND J. KLAFTER, *From continuous time random walks to the fractional Fokker-Planck equation*, *Phys. Rev. E*, 61 (2000), pp. 132–138.
- [4] E. BARKAI, *Fractional Fokker-Planck equation, solution, and application*, *Phys. Rev. E*, 63 (2001), article 046118.
- [5] J. BOUCHAUD AND A. GEORGES, *Anomalous diffusion in disordered media: Statistical mechanisms, models and physical applications*, *Phys. Rep.*, 195 (1990), pp. 127–293.
- [6] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.
- [7] P. L. BUTZER AND U. WESTPHAL, *An Introduction to Fractional Calculus*, World Scientific, Singapore, 2000.
- [8] W. H. DENG, *Short memory principle and a predictor-corrector approach for fractional differential equations*, *J. Comput. Appl. Math.*, 206 (2007), pp. 174–188.
- [9] W. H. DENG AND J. H. LÜ, *Design of multi-directional multi-scroll chaotic attractors based on fractional differential systems via switching control*, *Chaos*, 16 (2006), article 043120.
- [10] W. H. DENG, *Generalized synchronization in fractional order systems*, *Phys. Rev. E*, 75 (2007), article 056201.
- [11] W. H. DENG, *Numerical algorithm for the time fractional Fokker-Planck equation*, *J. Comput. Phys.*, 227 (2007), pp. 1510–1522.
- [12] V. J. ERVIN AND J. P. ROOP, *Variational formulation for the stationary fractional advection dispersion equation*, *Numer. Methods Partial Differential Equations*, 22 (2005), pp. 558–576.
- [13] V. J. ERVIN, N. HEUER, AND J. P. ROOP, *Numerical approximation of a time dependent, nonlinear, space-fractional diffusion equation*, *SIAM J. Numer. Anal.*, 45 (2007), pp. 572–591.
- [14] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.
- [15] R. GORENFLO, F. MAINARDI, D. MORETTI, AND P. PARADISI, *Time fractional diffusion: A discrete random walk approach*, *Nonlinear Dynam.*, 29 (2002), pp. 129–143.
- [16] I. GOYCHUK, E. HEINSALU, M. PATRIARCA, G. SCHMID, AND P. HÄNGGI, *Current and universal scaling in anomalous transport*, *Phys. Rev. E*, 73 (2006), article 020101.
- [17] O. HEAVISIDE, *Electromagnetic Theory*, Chelsea, New York, 1971.
- [18] E. HEINSALU, M. PATRIARCA, I. GOYCHUK, G. SCHMID, AND P. HÄNGGI, *Fractional Fokker-Planck dynamics: Numerical algorithm and simulations*, *Phys. Rev. E*, 73 (2006), article 046133.
- [19] M. ICHISE, Y. NAGAYANAGI, AND T. KOJIMA, *An analog simulation of noninteger order transfer functions for analysis of electrode processes*, *J. Electroanal. Chem.*, 33 (1971), pp. 253–265.
- [20] G. JUMARIE, *A Fokker-Planck equation of fractional order with respect to time*, *J. Math. Phys.*, 33 (1992), pp. 3536–3542.
- [21] S. M. KENNETH AND R. BERTRAM, *An Introduction to the Fractional Calculus and Fractional Differential Equations*, Wiley-Interscience, New York, 1993.

- [22] R. C. KOELLER, *Application of fractional calculus to the theory of viscoelasticity*, J. Appl. Mech., 51 (1984), pp. 229–307.
- [23] D. KUSNEZOV, A. BULGAC, AND G. D. DANG, *Quantum levy processes and fractional kinetics*, Phys. Rev. Lett., 82 (1999), pp. 1136–1139.
- [24] J. L. LAVOIE, T. J. OSLER, AND R. TREMBLAY, *Fractional derivatives and special functions*, SIAM Rev., 18 (1976), pp. 240–268.
- [25] E. K. LENZI, R. S. MENDES, K. S. FA, AND L. C. MALACARNE, *Anomalous diffusion: Fractional Fokker-Planck equation and its solutions*, J. Math. Phys., 44 (2003), pp. 2179–2185.
- [26] C. P. LI AND W. H. DENG, *Remarks on fractional derivatives*, Appl. Math. Comput., 187 (2007), pp. 777–784.
- [27] Y. M. LIN AND C. J. XU, *Finite difference/spectral approximations for the time-fractional diffusion equation*, J. Comput. Phys., 225 (2007), pp. 1533–1552.
- [28] F. LIU, V. ANH, AND I. TURNER, *Numerical solution of the space fractional Fokker-Planck equation*, J. Comput. Appl. Math., 166 (2004), pp. 209–219.
- [29] C. LUBICH, *Discretized fractional calculus*, SIAM J. Math. Anal., 17 (1986), pp. 704–719.
- [30] M. MAGDZIARZ AND A. WERON, *Competition between subdiffusion and Lévy flights: A Monte Carlo approach*, Phys. Rev. E, 75 (2007), article 056702.
- [31] B. MANDELBROT, *Some noises with $1/f$ spectrum, a bridge between direct current and white noise*, IEEE Trans. Inform. Theory, 13 (1967), pp. 289–298.
- [32] M. M. MEERSCHAERT, H.-P. SCHEFFLER, AND C. TADJERAN, *Finite difference methods for two-dimensional fractional dispersion equation*, J. Comput. Phys., 211 (2006), pp. 249–261.
- [33] R. METZLER, E. BARKAI, AND J. KLAFTER, *Anomalous diffusion and relaxation close to thermal equilibrium: A fractional Fokker-Planck equation approach*, Phys. Rev. Lett., 82 (1999), pp. 3563–3567.
- [34] R. METZLER AND J. KLAFTER, *The random walk's guide to anomalous diffusion: A fractional dynamics approach*, Phys. Rep., 339 (2000), pp. 1–77.
- [35] R. METZLER AND T. F. NONNENMACHER, *Space- and time-fractional diffusion and wave equations, fractional Fokker-Planck equations, and physical motivation*, Chem. Phys., 284 (2002), pp. 67–90; see also references therein.
- [36] I. PODLUBNY, *Fractional Differential Equations*, Academic Press, New York, 1999.
- [37] S. SAMKO, A. KILBAS, AND O. MARICHEV, *Fractional Integrals and Derivatives: Theory and Applications*, Gordon and Breach, London, 1993.
- [38] T. H. SOLOMON, E. R. WEEKS, AND H. L. SWINNEY, *Observations of anomalous diffusion and Lévy flights in a 2-dimensional rotating flow*, Phys. Rev. Lett., 71 (1993), pp. 3975–3979.
- [39] N. SUGIMOTO, *Burgers equation with a fractional derivative: Hereditary effects on nonlinear acoustic waves*, J. Fluid Mech., 225 (1991), pp. 631–653.
- [40] S. B. YUSTE AND L. ACEDO, *An explicit finite difference method and a new von Neumann-type stability analysis for fractional diffusion equations*, SIAM J. Numer. Anal., 42 (2005), pp. 1862–1874.
- [41] G. M. ZASLAVSKY, *Chaos, fractional kinetics, and anomalous transport*, Phys. Rep., 371 (2002), pp. 461–580.

RAPID SOLUTION OF THE WAVE EQUATION IN UNBOUNDED DOMAINS*

L. BANJAI[†] AND S. SAUTER[†]

Abstract. In this paper we propose and analyze a new, fast method for the numerical solution of time domain boundary integral formulations of the wave equation. We employ Lubich’s convolution quadrature method for the time discretization and a Galerkin boundary element method for the spatial discretization. The coefficient matrix of the arising system of linear equations is a triangular block Toeplitz matrix. Possible choices for solving the linear system arising from the above discretization include the use of fast Fourier transform (FFT) techniques and the use of data-sparse approximations. By using FFT techniques, the computational complexity can be reduced substantially while the storage cost remains unchanged and is, typically, high. Using data-sparse approximations, the gain is reversed; i.e., the computational cost is (approximately) unchanged while the storage cost is substantially reduced. The method proposed in this paper combines the advantages of these two approaches. First, the discrete convolution (related to the block Toeplitz system) is transformed into the (discrete) Fourier image, thereby arriving at a decoupled system of discretized Helmholtz equations with complex wave numbers. A fast data-sparse (e.g., fast multipole or panel-clustering) method can then be applied to the transformed system. Additionally, significant savings can be achieved if the boundary data are smooth and time-limited. In this case the right-hand sides of many of the Helmholtz problems are almost zero, and hence can be disregarded. Finally, the proposed method is inherently parallel. We analyze the stability and convergence of these methods, thereby deriving the choice of parameters that preserves the convergence rates of the unperturbed convolution quadrature. We also present numerical results which illustrate the predicted convergence behavior.

Key words. wave equation, boundary element methods, convolution quadrature

AMS subject classifications. 65N38, 65R20, 35L05

DOI. 10.1137/070690754

1. Introduction. Boundary value problems governed by the wave equation

$$\partial_t^2 u - \Delta u = f$$

arise in many physical applications such as electromagnetic wave propagation or the computation of transient acoustic waves. Since such problems are typically formulated in unbounded domains, the method of integral equations is an elegant tool for transforming this partial differential equation (PDE) into an integral equation on the bounded surface of the scatterer.

Although this approach goes back to the early 1960s (cf. [19]), the development of fast numerical methods for integral equations in the field of hyperbolic problems is still in its infancy compared to the multitude of fast methods for elliptic boundary integral equations (cf. [38] and references therein). Existing numerical discretization methods include collocation methods with some stabilization techniques (cf. [7], [8], [14], [15], [16], [33], [37]), and Laplace–Fourier methods coupled with Galerkin boundary elements in space (see [3], [12], [17], [20]). Numerical experiments can be found, e.g., in [21].

*Received by the editors May 7, 2008; accepted for publication (in revised form) June 17, 2008; published electronically October 31, 2008. A short description of the results of this paper has been published in the Proceedings of the Waves 2007 Conference [6].

<http://www.siam.org/journals/sinum/47-1/69075.html>

[†]Institut für Mathematik, Universität Zürich, Winterthurerstr. 190, CH-8057 Zürich, Switzerland (lehel.banjai@math.uzh.ch, stas@math.uzh.ch). The first author gratefully acknowledges the support given by SNF 200021–113481/1.

In [18] a fast version of the *marching-on-in-time* (MOT) method is presented which is based on a suitable plane wave expansion of the arising potential, which reduces the storage and computational costs.

We here employ the convolution quadrature method for the time discretization and a Galerkin boundary element method in space. The convolution quadrature method for the time discretization has been developed in [29], [30], [31], [32]. It provides a straightforward way to obtain a stable time stepping scheme using the Laplace transform of the kernel function. For applications to problems such as viscoelastic and poroelastic continua see [40], [41], [42].

The coefficient matrix in the arising system of linear equation is a block-triangular Toeplitz matrix consisting of N blocks of dimension $M \times M$, where N denotes the number of time steps and M is the number of spatial degrees of freedom. Due to the nonlocalness of the arising boundary integral operators, the $M \times M$ matrix blocks are densely populated.

In the literature, there exist (at least) two alternatives for solving this system efficiently. In [24], fast Fourier transform (FFT) techniques are employed, which make use of the Toeplitz structure of the system matrix, and the computational complexity is reduced to $\mathcal{O}(M^2 N \log^2 N)$, while the storage complexity stays at $\mathcal{O}(NM^2)$. In [23], [22], [28], the $M \times M$ block matrices are approximated by data-sparse representations based on a cutoff and panel-clustering strategy. This leads to a significant reduction of the storage complexity. The computational complexity is reduced compared to the $\mathcal{O}(N^2 M^2)$ cost of the naive approach but increased compared to the computational cost of the FFT approach.

Also the classical Galerkin discretization of the retarded boundary integral equation (see [3], [20]), leads to a block Toeplitz system matrix, where the matrix blocks are of size $M \times M$ and sparse. More precisely, the number of nonzero entries in the block Toeplitz matrix is, for piecewise constant boundary elements, of order $\mathcal{O}(M^2)$ and, for piecewise linear boundary elements, of order $\mathcal{O}(M^{2+\frac{1}{3}})$ for this approach. Here, the total cost for the computation of a full Galerkin approximation sums up to $\mathcal{O}(M^2 N)$ for piecewise constant boundary elements and to $\mathcal{O}(N^2 M^{3/2})$ for piecewise linear boundary elements. A drawback of this approach, however, is that the numerical quadrature for computing the coefficients of the system matrix has to be carried out on the intersections of the boundary element mesh with the discrete light cone. The stable handling of these intersections and the implementation is especially complicated for curved panels.

In this paper, we propose a new approach which combines the advantages of the FFT technique with the sparse approximation. We transfer the block Toeplitz system to the Fourier image by the discrete Fourier transform and then face the problem of computing approximate solutions of Helmholtz problems at different (complex) wave numbers. These Helmholtz problems are fully decoupled, and hence can be efficiently solved on parallel computers. Relatively standard, fast methods (e.g., fast multipole method, hierarchical matrices) for the solution of frequency domain scattering can effectively be applied to these problems; see [9], [35], and [5]. It may also be possible to further reduce the computational cost of assembling the matrices by using the techniques for multifrequency analysis described in [27], [44]. Further, we also show that if the boundary data are sufficiently smooth and compatible and of limited time duration, instead of N , only $\mathcal{O}(N^\epsilon)$, for any fixed $\epsilon > 0$, Helmholtz systems need to be solved. Our method is similar and shares some properties (the need to solve a series of elliptic problems and the intrinsic parallelizability) of certain methods for

parabolic equations; see [26], [43]. A related, interesting variation of the convolution quadrature for convolution kernels whose Laplace transform is sectorial can be found in [39].

2. Integral formulation of the wave equation. Let $\Omega \subset \mathbb{R}^3$ be a Lipschitz domain with boundary Γ ; typically, e.g., in scattering problems, Ω is an unbounded domain. In this paper, we present efficient methods for numerically solving the homogeneous wave equation

$$(2.1a) \quad \partial_t^2 u - \Delta u = 0 \quad \text{in } \Omega \times (0, T)$$

with initial conditions

$$(2.1b) \quad u(\cdot, 0) = \partial_t u(\cdot, 0) = 0 \quad \text{in } \Omega$$

and boundary conditions

$$(2.1c) \quad u = g \quad \text{on } \Gamma \times (0, T)$$

on a time interval $(0, T)$ for some $T > 0$. For its solution, we employ an ansatz as a *single layer potential*

$$(2.2) \quad u(x, t) = \int_0^t \int_{\Gamma} k(x - y, t - \tau) \phi(y, \tau) d\Gamma_y d\tau, \quad (x, t) \in \Omega \times (0, T),$$

where $k(z, t)$ is the fundamental solution of the wave equation,

$$(2.3) \quad k(z, t) = \frac{\delta(t - \|z\|)}{4\pi\|z\|},$$

with $\delta(t)$ being the Dirac delta distribution. The ansatz (2.2) satisfies the homogeneous equation (2.1a) and the initial conditions (2.1b). The extension $x \rightarrow \Gamma$ is continuous, and hence the unknown density ϕ in (2.2) is determined via the boundary conditions (2.1c), $u(x, t) = g(x, t)$. This results in the boundary integral equation for ϕ ,

$$(2.4) \quad \int_0^t \int_{\Gamma} k(x - y, t - \tau) \phi(y, \tau) d\Gamma_y d\tau = g(x, t) \quad \forall (x, t) \in \Gamma \times (0, T).$$

Existence and uniqueness results for the solution of the continuous problem are proved in [31] and [3, Prop. 3].

3. Numerical discretization.

3.1. Time discretization via convolution quadrature. For the time discretization, we employ the convolution quadrature approach which has been developed by Lubich in [29], [30], [31] and Lubich and Schneider in [32]. We do not recall the theoretical framework here but directly apply the approach to the wave equation. We make use of the following notation for the time convolution:

$$V(\partial_t)\phi := \int_0^t v(t - \tau)\phi(\tau)d\tau,$$

where V denotes the Laplace transform of the operator v ; for the reasons behind using this notation see [29]. Note that, for the retarded single layer potential (2.2),

v is a parameter-dependent integral operator, i.e., $(v(t - \tau)\phi(\tau))(x) = \int_{\Gamma} k(x - y, t - \tau)\phi(\tau, y)d\Gamma_y$ (where we write $\phi(\tau, y)$ for $(\phi(\tau))(y)$) and $V(s)$ is the Laplace transform of v given by (3.4).

To discretize the time convolution we split the time interval $[0, T]$ into $N + 1$ time steps of equal length $\Delta t = T/N$ and compute an approximate solution at the discrete time levels $t_n = n\Delta t$. The continuous convolution operator $V(\partial_t)$ at the discrete times t_n is replaced by the discrete convolution operator, for $n = 0, 1, \dots, N$,

$$(3.1) \quad (V(\partial_t^{\Delta t})\phi^{\Delta t})(t_n) := \sum_{j=0}^n \omega_{n-j}^{\Delta t}(V)\phi^{\Delta t}(t_j).$$

The convolution weights $\omega_n^{\Delta t}(V)$ are defined below (see (3.3)); whenever the underlying operator v , respectively, V , is clear from the context, we will write $\omega_n^{\Delta t}$. The time-discrete problem is given as follows: Find $\phi_j(\cdot) = \phi^{\Delta t}(\cdot, t_j)$, such that

$$(3.2) \quad \sum_{j=0}^n (\omega_{n-j}^{\Delta t}\phi_j)(x) = g_n(x), \quad n = 1, \dots, N, \quad x \in \Gamma,$$

where $g_n(x)$ is some approximation to $g(x, t_n)$, or $g(x, t_n)$ itself.

For the derivation, the general framework, and various applications, we refer the reader to [29], [30], [31], and for our concrete problem to [23]. If the time discretization is related to the unconditionally stable backward difference formula of second order (BDF2) scheme, the convolution weights $\omega_n^{\Delta t}$ are implicitly defined by

$$(3.3) \quad V\left(\frac{\gamma(\zeta)}{\Delta t}\right) = \sum_{n=0}^{\infty} \omega_n^{\Delta t}\zeta^n, \quad |\zeta| < 1.$$

Here, $V(s) : H^{-1/2}(\Gamma) \rightarrow H^{1/2}(\Gamma)$, $\text{Re } s > 0$, is the single layer potential for the Helmholtz operator $\Delta U - s^2U = 0$,

$$(3.4) \quad (V(s)\varphi)(x) = \int_{\Gamma} K(\|x - y\|, s)\varphi(y)d\Gamma_y, \quad \text{where } K(d, s) := \frac{e^{-sd}}{4\pi d}.$$

Note that K is the Laplace transform of the original time domain kernel function (2.3). The function $\gamma(\zeta)$ is the quotient of the generating polynomials of the BDF2 scheme and is given by

$$\gamma(\zeta) = \frac{1}{2}(\zeta^2 - 4\zeta + 3).$$

3.2. A decoupled system of Helmholtz problems. As recommended in [29, 31], the convolution weights $\omega_j^{\Delta t}$ can be numerically computed by applying the trapezoidal rule to its representation as a contour integral,

$$(3.5) \quad \omega_j^{\Delta t}(V) = \frac{1}{2\pi i} \oint_C \frac{V(\gamma(\zeta)/\Delta t)}{\zeta^{j+1}} d\zeta,$$

where C can be chosen as a circle centered at the origin of radius $\lambda < 1$. The approximate convolution weights are then given by the scaled inverse discrete Fourier transform

$$\omega_j^{\Delta t, \lambda}(V) := \frac{\lambda^{-j}}{N+1} \sum_{l=0}^N V(s_l)\zeta_{N+1}^{lj}, \quad \text{where } \zeta_{N+1} = e^{\frac{2\pi i}{N+1}}, \quad s_l = \frac{\gamma(\lambda\zeta_{N+1}^{-l})}{\Delta t}.$$

Let us extend the above two formulae to negative indices $j < 0$; note that this implies $\omega_j^{\Delta t} = 0$ for $j < 0$. As $N \rightarrow \infty$ or $\lambda \rightarrow 0$, we have $\omega_j^{\Delta t} - \omega_j^{\Delta t, \lambda} = \mathcal{O}(\lambda^{N+1})$, $j = -N, \dots, N$; see Proposition 5.4. By extending the sum in (3.1) to $j = N$ and substituting the approximate weights in (3.2), we obtain the following new system of equations for the new unknown $\phi^{\Delta t, \lambda}$:

$$(3.6) \quad \left(V(\partial_t^{\Delta t, \lambda}) \phi^{\Delta t, \lambda} \right) (t_n) := \sum_{j=0}^N \omega_{n-j}^{\Delta t, \lambda} (V) \phi_j^\lambda = g_n, \quad n = 0, 1, \dots, N.$$

The effect of the approximation on the difference between $\phi^{\Delta t, \lambda}$ and $\phi^{\Delta t}$ is discussed later. Substituting the definition of $\omega^{\Delta t, \lambda}$ in (3.6), we obtain the system of equations

$$(3.7) \quad \frac{\lambda^{-n}}{N+1} \sum_{l=0}^N \left(V(s_l) \hat{\phi}_l \right) (x) \zeta_{N+1}^{nl} = g_n(x), \quad n = 0, 1, \dots, N,$$

where

$$\hat{\phi}_l := \sum_{j=0}^N \lambda^j \phi_j^\lambda \zeta_{N+1}^{-lj}.$$

Note that the inverse transform is given by

$$(3.8) \quad \phi_l^\lambda = \frac{\lambda^{-l}}{N+1} \sum_{j=0}^N \hat{\phi}_j \zeta_{N+1}^{lj}.$$

Now, notice that, after multiplying by λ^n , applying the discrete Fourier transform with respect to n to both sides gives $N+1$ decoupled problems as follows:

$$(3.9) \quad \left(V(s_l) \hat{\phi}_l \right) (x) = \hat{g}_l(x) \quad \forall x \in \Gamma,$$

where

$$\hat{g}_l(x) = \sum_{n=0}^N \lambda^n g_n(x) \zeta_{N+1}^{-ln}.$$

We have thereby reduced the problem of solving numerically the wave equation to a system of Helmholtz problems with complex wave numbers s_l , $l = 0, 1, \dots, N$. An example of the range of frequencies is given in Figure 1.

Remark 3.1. An important remark to make here is that

$$V \left(\partial_t^{\Delta t, \lambda} \right) \phi^{\Delta t, \lambda} = g \quad \text{implies} \quad \phi^{\Delta t, \lambda} = V^{-1} \left(\partial_t^{\Delta t, \lambda} \right) g.$$

This can be seen by applying the scaled discrete inverse Fourier transform (see (3.8)) to

$$\hat{\phi}_l = V^{-1}(s_l) \hat{g}_l,$$

thereby obtaining

$$\phi_n^\lambda = \frac{\lambda^{-n}}{N+1} \sum_{l=0}^N \hat{\phi}_l \zeta_{N+1}^{nl} = \frac{\lambda^{-n}}{N+1} \sum_{l=0}^N V^{-1}(s_l) \hat{g}_l \zeta_{N+1}^{nl} = \sum_{j=0}^N \omega_{n-j}^{\Delta t, \lambda} (V^{-1}) g_j.$$

The last step is obtained from the definition of \hat{g}_l and $\omega_n^{\Delta t, \lambda} (V^{-1})$; see also (3.6) and (3.7). This fact will help us in obtaining optimal error and stability estimates.

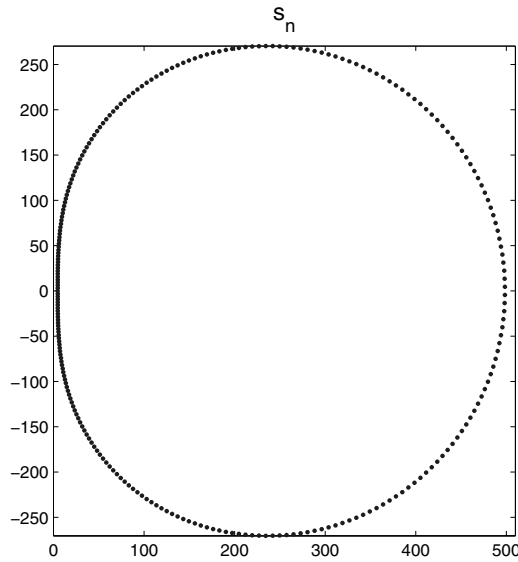


FIG. 1. A range of complex frequencies for $N = 256$, $T = 2$, and $\lambda^N = 10^{-4}$. For this example it holds that $\text{Re } s_n > 4.6$, $n = 0, 1, \dots, N$.

3.3. Spatial discretization. Galerkin boundary element methods. In the previous section we derived the following semidiscrete problem: For $n = 0, 1, \dots, N$, find $\phi_n^\lambda \in H^{-1/2}(\Gamma)$ such that

$$(3.10) \quad \sum_{j=0}^N \omega_{n-j}^{\Delta t, \lambda} \phi_j^\lambda = g_n, \quad n = 0, 1, \dots, N.$$

We have further shown that the above system is equivalent to a system of decoupled Helmholtz equations

$$(3.11) \quad \left(V(s_l) \hat{\phi}_l \right) (x) = \hat{g}_l(x) \quad \forall x \in \Gamma.$$

In this paper we use a Galerkin boundary element method for the spatial discretization. Let \mathcal{G} be a regular (in the sense of Ciarlet [11]) boundary element mesh on Γ consisting of shape regular, possibly curved, triangles. For a triangle $\tau \in \mathcal{G}$, the (regular) pull-back to the reference triangle $\hat{\tau} := \text{conv} \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\}$ is denoted by $\chi_\tau : \hat{\tau} \rightarrow \tau$. The space of piecewise constant, discontinuous functions is

$$S_{-1,0} := \{ u \in L^\infty(\Gamma) \quad : \quad \forall \tau \in \mathcal{G} : u|_\tau \in \mathbb{P}_0 \},$$

and, alternatively, we consider the space of continuous, piecewise linear functions

$$S_{0,1} := \{ u \in C^0(\Gamma) \quad : \quad \forall \tau \in \mathcal{G} : (u \circ \chi_\tau)|_\tau \in \mathbb{P}_1 \}$$

for the space discretization. As a basis for $S_{-1,0}$, we choose the characteristic functions for the panels $\tau \in \mathcal{G}$, while the basis for $S_{0,1}$ consists of the standard hat functions, lifted to the surface Γ . The general notation is S for the boundary element space and $(b_m)_{m=1}^M$ for the basis. The mesh width is given by

$$h := \max_{\tau \in \mathcal{G}} h_\tau, \quad \text{where} \quad h_\tau := \text{diam}(\tau).$$

For the space-time discrete solution at time t_n we employ the ansatz

$$(3.12) \quad \phi_n^{h,\lambda}(y) = \sum_{m=1}^M \phi_{n,m} b_m(y),$$

where $(\phi_{n,m})_{m=1}^M \in \mathbb{R}^M$ are the nodal values of the discrete solution at time step t_n . Therefore, for the Helmholtz problems (3.11), the corresponding ansatz is

$$(3.13) \quad \hat{\phi}_l^h(y) = \sum_{m=1}^M \hat{\phi}_{l,m} b_m(y),$$

where the relationship between $\hat{\phi}_{l,m}$ and $\phi_{n,m}$ is given by $\hat{\phi}_{l,m} = \sum_{n=0}^N \lambda^n \phi_{n,m} \zeta_{N+1}^{ln}$.

To solve for the coefficients $\hat{\phi}_{l,m}$ we impose the integral equations (3.11) not pointwise but in a weak form as follows: Find $\hat{\phi}_l^h \in S$ of the form (3.13) such that

$$(3.14) \quad \sum_{m=1}^M \hat{\phi}_{l,m} \int_{\Gamma} \int_{\Gamma} K(\|x-y\|, s_l) b_m(y) b_k(x) d\Gamma_y d\Gamma_x = \int_{\Gamma} \hat{g}_l(x) b_k(x) d\Gamma_x,$$

for $l = 0, 1, \dots, N$, $k = 1, 2, \dots, M$. Note that this is equivalent to imposing (3.10) in a weak form in order to compute $\phi_n^{h,\lambda}$.

4. Algorithmic realization and sparse approximation. Applying the Galerkin boundary element method to the time-discrete equations (3.1) obtained by convolution quadrature results in a block-triangular, block Toeplitz system, where each block is a dense Galerkin boundary element matrix; see [31] and [22]. This block system can be solved by using FFT techniques (see [24]), with computational complexity of $\mathcal{O}(M^2 N \log^2 N)$ and a storage complexity of $\mathcal{O}(M^2 N)$. Alternatively (see [28]), one can approximate the block matrices \mathbf{A}_n by a cutoff strategy and panel-clustering and directly solve the system without the FFT. This reduces the storage cost significantly, while the computational complexity is $\mathcal{O}(M^2 N^{1+s})$, where the small value of s depends on the chosen discretization. By rewriting (3.1) as a system of decoupled Helmholtz problems, we are able to combine the advantages of both approaches.

We note that also the classical Galerkin discretization of the retarded boundary integral equation leads to a block Toeplitz system. Solving this system (see [3], [20]) nevertheless results in suboptimal, higher than linear, computational complexity.

4.1. Reduction of the number of Helmholtz problems to be solved. A closer look at the Helmholtz problems tells us that only half of the problems need to be solved. Since $\hat{\phi}_l$, \hat{g}_l , and s_l are discrete Fourier transforms of real data, we know that they are, for $l = 1, 2, \dots, \lfloor \frac{N}{2} + 1 \rfloor$, the complex conjugates of $\hat{\phi}_j$, \hat{g}_j , s_j , for $j = \lceil \frac{N}{2} + 2 \rceil, \dots, N + 1$; for the case of s_l see Figure 1. Most importantly, for us this means that

$$(4.1) \quad \hat{\phi}_{N+2-j} = \overline{\hat{\phi}_j}, \quad j = 1, 2, \dots, \left\lceil \frac{N}{2} + 1 \right\rceil.$$

Depending on the properties of the right-hand side g , it is possible to avoid the solution of a much larger number of Helmholtz problems without destroying the accuracy of the overall approximation. A particularly favorable case arises if g as a function of time can be extended to \mathbb{R} as a smooth function with support contained in $[0, T]$.

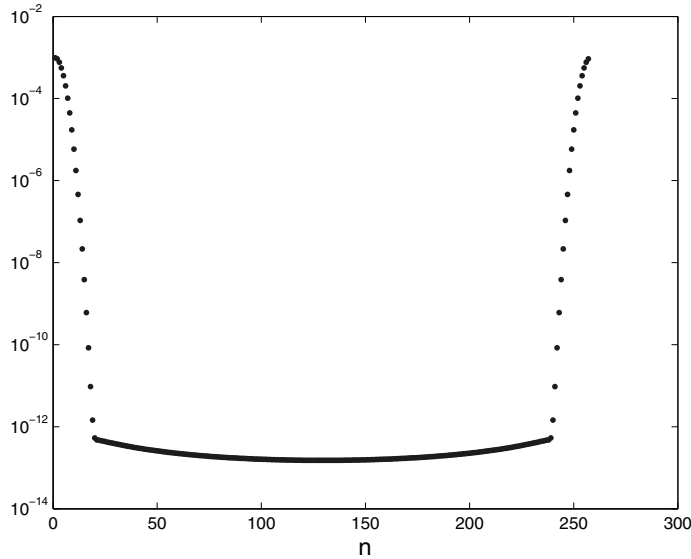


FIG. 2. We plot $\max_{\|x\|=1} |\hat{g}_n(x)|$ for $N = 256$, $T = 2$, $\lambda^N = 10^{-8}$, and where $g(x, t)$ is the Gaussian pulse given by (6.5). The solution to the n th Helmholtz problem, with n in the central plateau in the above plot, is accurately approximated by zero.

Let us assume that for some $x \in \Gamma$, $g(x, \cdot) \in C^\infty([0, T])$, and that

$$\partial_t^n g(x, 0) = \partial_t^n g(x, T) = 0 \quad \forall n \in \mathbb{N}_0.$$

Further, define $g_\lambda(x, t) := \lambda^{t/\Delta t} g(x, t)$. Then it is clear that $g_\lambda(x, \cdot) \in C^\infty([0, T])$ and that also all the partial derivatives with respect to time vanish at the end points of the time interval $[0, T]$. The reason for defining this function is that $\hat{g}_n(x)$ is an approximation of a Fourier coefficient of $g_\lambda(x, t)$, as we see next.

Let $g_\lambda(x, \cdot)$ be extended to the domain $[0, T + \Delta t]$ by zero (i.e., in a smooth way) and further extended to \mathbb{R} in a periodic way with period $T + \Delta t$. Let then

$$g_\lambda(x, t) = \sum_{j=-\infty}^{\infty} a_j e^{\frac{2\pi i j t}{T+\Delta t}}, \quad a_j = \frac{1}{T + \Delta t} \int_0^{T+\Delta t} g_\lambda(x, \tau) e^{\frac{-2\pi i j \tau}{T+\Delta t}} d\tau$$

be its Fourier expansion. Approximating the integral in the definition of the coefficients a_j by the trapezoidal rule, we obtain exactly the values $\frac{1}{N+1} \hat{g}_j(x)$, where, assuming N is even,

$$a_j \approx \frac{1}{N+1} \sum_{n=0}^N g_\lambda(x, t_n) e^{\frac{-2\pi i j n}{N+1}} = \frac{1}{N+1} \hat{g}_j(x) \quad \text{for } 0 \leq j \leq \frac{N}{2}.$$

See Figure 2 for an example of a right-hand side with the above properties and the decay of its Fourier coefficients. The solutions of Helmholtz problems with right-hand sides that are close to zero (i.e., all the right-hand sides on the central plateau in Figure 2) can be set to zero with no adverse affect on the accuracy of the overall method.

Remark 4.1. A right-hand side g with the above properties can be thought of as a smooth signal of finite durability. If g does not have these properties, it may still be possible to split the signal into a number of smooth and time-limited signals.

4.2. Data-sparse approximation. To find a solution to (3.9) we need to solve a number of dense linear systems, each of size $M \times M$. The cost of solving a single system by a direct method is $\mathcal{O}(M^3)$, and if a good preconditioner for an iterative method is available, this can be reduced to $\mathcal{O}(M^2)$. In both cases the storage costs are $\mathcal{O}(M^2)$. The cost of recovering the values $\phi_{j,m}$ from $\hat{\phi}_{l,m}$ is negligible since it can be done exactly (if we ignore errors due to finite precision arithmetic) and efficiently using the FFT in time $\mathcal{O}(MN \log N)$; see also Remark 5.11.

One possibility for reducing these costs is to use panel-clustering or fast multipole techniques. We explain the basic idea behind these methods.

Let A_n be the n th linear system to be solved in (3.9), i.e.,

$$(A_n)_{kj} = \int_{\Gamma} \int_{\Gamma} K(\|x - y\|, s_n) b_j(y) b_k(x) d\Gamma_y d\Gamma_x.$$

Further, we denote by \mathcal{I} the index set $\mathcal{I} := \{1, 2, \dots, M\}$, refer to subsets $\tau \subset \mathcal{I}$ as *clusters*, and define corresponding subsets of the boundary Γ by

$$\Gamma_{\tau} := \cup_{j \in \tau} \text{supp } b_j.$$

We call a pair of clusters $\tau \times \sigma$ a *block*. The corresponding block of the matrix A_n is then given by

$$(A_n|_{\tau \times \sigma})_{kj} = \begin{cases} (A_n)_{kj} & \text{if } k \in \tau \text{ and } j \in \sigma, \\ 0 & \text{otherwise.} \end{cases}$$

In the following definition, $B(c, r)$ denotes the ball centered at $c \in \mathbb{R}^3$ and radius $r > 0$.

DEFINITION 4.2. A block $b = \tau \times \sigma$ is said to be η -admissible, for some $\eta < 1$, if there exist $r_{\tau}, r_{\sigma} > 0$ and $c_{\tau}, c_{\sigma} \in \mathbb{R}^3$ such that

$$r_{\tau} + r_{\sigma} \leq \eta \|c_{\tau} - c_{\sigma}\| \quad \text{and} \quad \Gamma_{\tau} \subset B(c_{\tau}, r_{\tau}), \Gamma_{\sigma} \subset B(c_{\sigma}, r_{\sigma}).$$

For an admissible block, our goal is to find a *separable* approximation of the following fundamental solution:

$$(4.2) \quad K(\|x - y\|, s) \approx \sum_{l,k=1}^L u_k^{\tau}(x) s_{kl}^{\tau, \sigma} v_l^{\sigma}(y), \quad x \in \Gamma_{\tau}, y \in \Gamma_{\sigma}.$$

As indicated by the notation, we require that the basis functions $u_k^{\tau}(\cdot)$ (respectively, $v_l^{\sigma}(\cdot)$) depend only on the cluster τ (respectively, σ), and that the coefficients $s_{k,l}^b$ depend only on the block cluster $b = \tau \times \sigma$. Such an expansion allows us to approximate the block $A_n|_{\tau \times \sigma}$ of the matrix by a low rank matrix as follows:

$$(4.3) \quad A_n|_{\tau \times \sigma} \approx USV^{\top},$$

where

$$(4.4) \quad (U)_{kl} := \begin{cases} \int_{\Gamma_\tau} u_l^\tau(x) b_k(x) d\Gamma_x & \text{if } k \in \tau, l = 1, \dots, L, \\ 0 & \text{otherwise,} \end{cases}$$

$$(4.5) \quad (V)_{jl} := \begin{cases} \int_{\Gamma_\sigma} v_l^\sigma(y) b_j(y) d\Gamma_y & \text{if } j \in \sigma, l = 1, \dots, L, \\ 0 & \text{otherwise,} \end{cases}$$

and $(S)_{lm} := s_{lm}^{\tau, \sigma}$. Note that for $A_n|_{\tau \times \sigma}$ we need $O(|\tau||\sigma|)$ amount of storage, whereas for USV^\top we need $O(|\tau|L + |\sigma|L)$. If $L \ll \max\{|\tau|, |\sigma|\}$, it is significantly advantageous to use the low rank approximation of the block.

An extensive literature exists on the use of these methods to speed up the solution of the Helmholtz integral equations discretized by Galerkin boundary elements [2], [5], [13], [35], [36]. Most of this literature is, however, focused on the Helmholtz problem with a purely real wave number. For a purely real wave number the single layer potential representation is not always invertible; therefore certain stabilization methods need to be used. In our case the imaginary part of the wave number is strictly positive and we can use the single layer representation. The details of applying these “fast” methods to our case, together with algorithms and complexity estimates, will be given in a forthcoming paper. Here we investigate the effect of perturbations, due to the application of the fast methods, on the stability and accuracy. We assume that the kernel function $K(\cdot, s_l)$ in (3.9) is replaced by a separable approximation $K^{\text{pc}}(\cdot, s_l)$ such that

$$(4.6) \quad |K(d, s_l) - K^{\text{pc}}(d, s_l)| \leq \frac{\delta}{d} \quad \text{for some } \delta > 0.$$

The solution of the resulting perturbed system is denoted by $\hat{\phi}_{l,m}^{\text{pc}}$. To obtain a uniform approximation (4.6), the length of expansion L needs to depend both on the block cluster $b = \tau \times \sigma$ and on s_l . Typically L is chosen so that

$$(4.7) \quad L \geq C \left(\text{Re } s_l \|c_\tau - c_\sigma\| + \log \frac{1}{\delta} \right)^{d-1},$$

where C depends on the admissibility parameter η , and $d = 2, 3$ is the space dimension. Explicit and sharp estimates on the optimal choice of L are difficult to obtain, especially for complex wave numbers. In practice, one would estimate the error by a product of a Bessel function and a Hankel function; see, e.g., [1], [9]. Nevertheless, an important observation that can be made is that once L is greater than some threshold, the threshold depending on s_l , the convergence is exponential. This means that high accuracy can be obtained at little extra cost.

5. Error analysis. In the previous section we have introduced a method to reduce the numerical solution of the wave equation to a system of Helmholtz problems. We have also described two ways of reducing the cost of solving these systems by introducing further approximations. In this section we investigate the stability and convergence of both the basic method and the further approximations. This allows us to adjust the control parameters of these methods to the required accuracy in an optimal way.

Let the approximation to the unknown density $\phi(x, t_n)$ obtained by the pure Lubich’s method, i.e., with exact convolution weights, be given by $\phi_n^h \in S$. In [31] it is proved that if the data g are sufficiently smooth and compatible, then

$$(5.1) \quad \|\phi_n^h(\cdot) - \phi(\cdot, t_n)\|_{H^{-1/2}(\Gamma)} \leq C(\Delta t^2 + h^{m+3/2}),$$

where $m = 0$ for a piecewise constant basis and $m = 1$ for a piecewise linear basis. By “smooth and compatible” we mean that $g \in H_0^5([0, T]; H^{1/2}(\Gamma))$, where

$$H_0^r([0, T]; H^{1/2}(\Gamma)) := \left\{ g : \Gamma \times [0, T] \rightarrow \mathbb{R} : \text{there exists } g^* \in H^r(\mathbb{R}; H^{1/2}(\Gamma)) \right. \\ \left. \text{with } g = g^*|_{[0, T]} \text{ and } g^* \equiv 0 \text{ on } (-\infty, 0) \right\},$$

$$H^r(\mathbb{R}; H^{1/2}(\Gamma)) := \left\{ g : \Gamma \times \mathbb{R} \rightarrow \mathbb{R} : \int_{-\infty}^{\infty} (1 + |\omega|)^{2r} \|(\mathcal{F}g)(\cdot, \omega)\|_{H^{1/2}(\Gamma)}^2 d\omega < \infty \right\},$$

and \mathcal{F} denotes the integral Fourier transform with respect to the time variable $t \in \mathbb{R}$.

Our goal is to prove that the parameters in our method can be chosen so that convergence rates in (5.1) are preserved.

5.1. Errors due to the perturbation of $\omega_n^{\Delta t}$. Let $V_h(s) : S \rightarrow S$ be defined by

$$(V_h(s)\varphi, \psi)_{L^2(\Gamma)} := (V(s)\varphi, \psi)_{L^2(\Gamma)} \quad \forall \varphi, \psi \in S.$$

Whenever necessary, we will identify the inner product $(\cdot, \cdot)_{L^2(\Gamma)}$ with its extension to the dual pairing $H^{-1/2}(\Gamma) \times H^{1/2}(\Gamma)$. The solution by the convolution quadrature, i.e., with exact weights, is given by (see equation (5.5) in [31])

$$\phi^h = V_h^{-1}(\partial_t^{\Delta t})g^h,$$

whereas with the perturbed weights the solution is given by

$$\phi_n^{h, \lambda} = \left(V_h^{-1}(\partial_t^{\Delta t, \lambda})g^h \right)(t_n)$$

(see Remark 3.1), where $g^h \in S$ is the L^2 -projection of g on S as follows:

$$(g^h, \psi)_{L^2(\Gamma)} = (g, \psi)_{L^2(\Gamma)} \quad \forall \psi \in S.$$

For the remainder of the paper we will make use of the following notation:

$$(5.2) \quad \|\cdot\|_{+1} = \|\cdot\|_{H^{1/2}(\Gamma) \leftarrow H^{-1/2}(\Gamma)} \quad \text{and} \quad \|\cdot\|_{-1} = \|\cdot\|_{H^{-1/2}(\Gamma) \leftarrow H^{1/2}(\Gamma)}.$$

LEMMA 5.1. *Let $\text{Re } s \geq \sigma_0 > 0$. Then*

$$\|V_h^{-1}(s)\|_{-1} \leq \frac{C_{\text{stab}}}{\min(1, \sigma_0)} |s|^2.$$

Proof. The result follows immediately from the definition of $V_h(s)$ and the coercivity estimate for $V(s)$ as follows (see [3]):

$$\text{Re } (sV(s)\psi, \psi)_{L^2(\Gamma)} \geq C_{\text{stab}}^{-1} \frac{\min(1, \sigma_0)}{|s|} \|\psi\|_{H^{-1/2}(\Gamma)}^2. \quad \square$$

Remark 5.2. For $\omega \in \mathbb{R}$, there holds

$$\gamma(\lambda e^{i\omega}) = \frac{(\lambda + 3)(1 - \lambda) + 8(1 - \lambda)\lambda \sin^2 \frac{\omega}{2} + 8\lambda^2 \sin^4 \frac{\omega}{2}}{2} \\ - i\lambda \sin \omega \left(2(1 - \lambda) + \lambda \left(1 + 2 \sin^2 \frac{\omega}{2} \right) \right).$$

For the real part, we obtain the estimate

$$\operatorname{Re}(\gamma(\lambda e^{i\omega})/\Delta t) \geq \left(\frac{1-\lambda}{2} + 4\lambda^2 \sin^4 \frac{\omega}{2} \right) / \Delta t.$$

For $0 \leq \lambda < 1$, we have the uniform bound with respect to ω ,

$$\operatorname{Re} \frac{\gamma(\lambda e^{i\omega})}{\Delta t} \geq \operatorname{Re} \frac{\gamma(\lambda)}{\Delta t} = \frac{(3+\lambda)(1-\lambda)}{2\Delta t} \geq \frac{3(1-\lambda)}{2\Delta t}.$$

For the modulus, the (rough) upper estimate holds as follows:

$$\left| \frac{\gamma(\lambda e^{i\omega})}{\Delta t} \right| \leq \frac{C}{\Delta t} \quad \text{with} \quad C = 5^{3/2}.$$

LEMMA 5.3. *Let $W_h(s) := V_h^{-1}(s)/s^2$. Then,*

$$(5.3) \quad \|\omega_j^{\Delta t}(W_h)\|_{-1} \leq 2C_{\text{stab}}eT.$$

Further, for sufficiently smooth and compatible g , the identities

$$(5.4) \quad V_h^{-1}(\partial_t^{\Delta t})g = W_h(\partial_t^{\Delta t})((\partial_t^{\Delta t})^2g)$$

and, for $N \geq 4$,

$$(5.5) \quad V_h^{-1}(\partial_t^{\Delta t, \lambda})g = W_h(\partial_t^{\Delta t, \lambda})((\partial_t^{\Delta t})^2g),$$

hold, where $(\partial_t^{\Delta t})^2g$ denotes the twofold application of the multistep approximation, which in our case is the BDF2 scheme.

Proof. The bound for $\|\omega_j^{\Delta t}(W_h)\|_{-1}$ follows from the Cauchy estimate by choosing the circle with radius $e^{-\Delta t/T}$ as the integration contour in (3.5), Remark 5.2, and Lemma 5.1 as follows:

$$\begin{aligned} \|\omega_j^{\Delta t}(W_h)\|_{-1} &\leq e^{j\Delta t/T} \max_{\|z\|=1} \left\| W_h \left(\gamma(e^{-\Delta t/T}z)/\Delta t \right) \right\|_{-1} \\ &\leq \frac{C_{\text{stab}}}{\min(1, (1 - e^{-\Delta t/T})/(2\Delta t))} e^{j\Delta t/T} \leq 2C_{\text{stab}}Te^{j/N}. \end{aligned}$$

Applying the (scaled) inverse discrete Fourier transform to the identity $V_h^{-1}(s_l)\hat{g}_l = W_h(s_l)s_l^2\hat{g}_l$, we see that $V_h^{-1}(\partial_t^{\Delta t, \lambda})g^h = W_h(\partial_t^{\Delta t, \lambda})\widetilde{g}^h$, where

$$\widetilde{g}_n^h = \frac{\lambda^{-n}}{N+1} \sum_{l=0}^{N+1} \hat{g}_l^h s_l^2 \zeta_{N+1}^{ln}, \quad s_l = \frac{\gamma(\lambda \zeta_{N+1}^{-l})}{\Delta t}.$$

The inverse discrete Fourier transform of s_l^2 is

$$(5.6) \quad \frac{1}{N+1} \sum_{l=0}^{N+1} \left(\frac{\gamma(\lambda \zeta_{N+1}^{-l})}{\Delta t} \right)^2 \zeta_{N+1}^{lj} \approx \frac{\lambda^j}{2\pi i} \oint_C \frac{(\gamma(\lambda \zeta)/\Delta t)^2}{\zeta^{j+1}} d\zeta = \frac{\lambda^j}{\Delta t^2} \delta_j,$$

where

$$(\gamma(\zeta))^2 = \sum_{k=-\infty}^{\infty} \delta_k \zeta^k = \left(\frac{3}{2} - 2\zeta + \frac{1}{2}\zeta^2 \right)^2.$$

Since $(\gamma(\zeta))^2$ is a polynomial of order 4 and $N \geq 4$, the coefficients $\frac{\lambda^j}{\Delta t^2} \delta_j$ are reproduced exactly, without any quadrature error in (5.6). Therefore

$$\widetilde{g^h}_n = \frac{1}{\Delta t^2} \sum_{j=0}^n \delta_{n-j} g_j^h,$$

which is exactly the result of applying the BDF2 multistep method twice, where it is implicitly assumed that $g(t) = 0$ for $t \leq 0$. The result for $V_h^{-1}(\partial_t^{\Delta t})g^h$ is proved similarly, but with no restriction on N ; see also [31]. \square

PROPOSITION 5.4. *Let $0 < \lambda < 1$. Then*

$$\|V_h^{-1}(\partial_t^{\Delta t})g^h - V_h^{-1}(\partial_t^{\Delta t, \lambda})g^h\|_{H^{-1/2}(\Gamma)} \leq 2C_{\text{stab}}eT^2 \frac{\lambda^{N+1}}{1 - \lambda^{N+1}} \Delta t^{-1}.$$

Proof. Let $a_j := \lambda^j \omega_j^{\Delta t}(W_h)$, and let $\hat{a}_j := \lambda^j \omega_j^{\Delta t, \lambda}(W_h)$, $W_h(s) = V_h^{-1}(s)/s^2$. Then \hat{a}_j is the discrete Fourier transform approximation to a_j for $j = -N, \dots, N$ and (see [25])

$$\begin{aligned} \|a_j - \hat{a}_j\|_{-1} &= \left\| \sum_{l=1}^{\infty} a_{j+l(N+1)} + a_{j-l(N+1)} \right\|_{-1} \leq \sum_{l=1}^{\infty} \|a_{j+l(N+1)}\|_{-1} \\ &\leq \lambda^j \sum_{l=1}^{\infty} \lambda^{l(N+1)} \|\omega_{j+l(N+1)}^{\Delta t}\|_{-1} \leq 2C_{\text{stab}}eT\lambda^j \frac{\lambda^{N+1}}{1 - \lambda^{N+1}}, \end{aligned}$$

where we have used the bound (5.3). Therefore

$$\|\omega_j^{\Delta t}(W_h) - \omega_j^{\Delta t, \lambda}(W_h)\|_{-1} \leq 2C_{\text{stab}}T \frac{\lambda^{N+1}}{1 - \lambda^{N+1}},$$

and the result follows from the definition of the discrete convolution and identities (5.4) and (5.5). \square

THEOREM 5.5. *Let the exact solution $\phi(\cdot, t)$ be in $H^{m+1}(\Gamma)$ for any $t \in [0, T]$, data $g \in H_0^5([0, T]; H^{1/2}(\Gamma))$, $0 < \lambda < 1$, and let the boundary element space be $S = S_{m-1, m}$ for $m \in \{0, 1\}$. Then the discrete solution*

$$\phi_n^{h, \lambda} = \left(V_h^{-1}(\partial_t^{\Delta t, \lambda})g^h \right)(t_n)$$

satisfies the error estimate

$$\|\phi_n^{h, \lambda} - \phi(\cdot, t_n)\|_{H^{-1/2}(\Gamma)} \leq C_g \left(\frac{\lambda^{N+1}}{1 - \lambda^{N+1}} T^2 \Delta t^{-1} + \Delta t^2 + h^{m+3/2} \right),$$

where C_g depends on the right-hand side g , C_{stab} , and the time interval length T .

Proof. The result is a direct consequence of Proposition 5.4 and (5.1); see [31, Theorem 5.4]. \square

5.2. Error due to the perturbation of $V_h(s)$. We investigate the effect of perturbing $V_h(s)$, in particular the effect of approximate evaluation of the kernel $K(d, s)$ by separable expansions. If these perturbations could be chosen to be analytic in s , then a stability and error estimate from Lemma 5.5 in [31] could be used, in which there is no loss of powers of Δt . Unfortunately due to numerical stability issues

(see [5], [9], [34]), this is not the case for the problem at hand, i.e., different expansions need to be used for different values of s . Hence we will simply assume that

$$(5.7) \quad \|V_h^\varepsilon(s_l) - V_h(s_l)\|_{+1} \leq \varepsilon, \quad l = -N, -N+1, \dots, N-1, N,$$

and investigate how this perturbation affects the final solution.

LEMMA 5.6. *Let $\operatorname{Re} s > \sigma_0 > 0$ and $\varepsilon < \frac{1}{2}C_{\text{stab}}^{-1} \frac{\min(1, \sigma_0)}{|s|^2}$. Then $(V_h^\varepsilon(s))^{-1}$ is bounded and*

$$\|(V_h^\varepsilon(s))^{-1}\|_{-1} \leq 2C_{\text{stab}} \frac{|s|^2}{\min(1, \sigma_0)}.$$

Proof. Let us write

$$V_h^\varepsilon(s) = V_h(s) [I - V_h^{-1}(s)(V_h(s) - V_h^\varepsilon(s))].$$

From the estimate $\|V_h^{-1}(s)\|_{-1} \leq C_{\text{stab}}|s|^2/\min(1, \sigma_0)$ (see Lemma 5.1), we see that $\varepsilon < \frac{1}{2}C_{\text{stab}}^{-1} \min(1, \sigma_0)/|s|^2$ is sufficient for $(V_h^\varepsilon(s))^{-1}$ to exist and to be bounded as above. \square

LEMMA 5.7. *Let $\min_{l=0,1,\dots,N} \operatorname{Re} s_l > \sigma_0 > 0$ and $\varepsilon < \frac{1}{2}C_{\text{stab}} \frac{\min(1, \sigma_0)}{\max_{l=0,1,\dots,N} |s_l|^2}$. Then*

$$\|\omega_j^{\Delta t, \lambda}(Q_h) - \omega_j^{\Delta t, \lambda}(Q_h^\varepsilon)\|_{-1} \leq CT\lambda^{-j}\varepsilon\Delta t^{-1},$$

where

$$C = \left(\frac{C_{\text{stab}}}{\min(1, \sigma_0)} \right)^2, \quad Q_h(s) := \frac{V_h^{-1}(s)}{s^4}, \quad \text{and} \quad Q_h^\varepsilon(s) := \frac{(V_h^\varepsilon(s))^{-1}}{s^4}.$$

Proof. Using the fact that $Q_h^{-1}(s) = s^4 V_h(s)$, we obtain the bound

$$\|Q_h(s_l) - Q_h^\varepsilon(s_l)\|_{-1} = \|Q_h(s_l)(s_l^4 V_h^\varepsilon(s_l) - s_l^4 V_h(s_l))Q_h^\varepsilon(s_l)\|_{-1} \leq \left(\frac{C_{\text{stab}}}{\min(1, \sigma_0)} \right)^2 \varepsilon.$$

From this and the definition of the perturbed convolution weights, the result follows. \square

Let us define the solution of the ε -perturbed convolution equation by

$$\phi^{\lambda, h, \varepsilon} := (V_h^\varepsilon)^{-1}(\partial_t^{\Delta t, \lambda})g = Q_h^\varepsilon(\partial_t^{\Delta t, \lambda})((\partial_t^{\Delta t})^4 g)$$

and, as before,

$$\phi^{\lambda, h} := V_h^{-1}(\partial_t^{\Delta t, \lambda})g = Q_h(\partial_t^{\Delta t, \lambda})((\partial_t^{\Delta t})^4 g).$$

In the next result we estimate the difference between the two.

PROPOSITION 5.8. *Let $\min_{l=0,1,\dots,N} \operatorname{Re} s_l > \sigma_0 > 0$, let*

$$\varepsilon < \frac{1}{2}C_{\text{stab}} \frac{\min(1, \sigma_0)}{\max_{l=0,1,\dots,N} |s_l|^2},$$

and let the data g be sufficiently smooth and compatible. Then

$$\|\phi_n^{\lambda, h, \varepsilon} - \phi_n^{t, \lambda, h}\|_{H^{-1/2}(\Gamma)} \leq C\varepsilon T^2 \lambda^{-N} \Delta t^{-2},$$

with $C > 0$ as in Lemma 5.7.

Proof. The result is a direct consequence of the above lemma. \square

The above result, together with Remark 5.2, implies that to obtain optimal convergence it is sufficient to insure that $\varepsilon \leq C\lambda^N \Delta t^4$.

Let us now investigate the effect of perturbations on the kernel function $K(d, s)$. In order to do this, we assume

$$(5.8) \quad |K(\|x - y\|, s_l) - K^{pc}(\|x - y\|, s_l)| \leq \delta \frac{1}{\|x - y\|} \quad \forall x, y \in \Gamma$$

for $l = 0, 1, \dots, N$, and define the operator $V_h^{pc}(s) : S \rightarrow S$ by

$$(V_h^{pc}(s)\psi, \varphi)_{L^2(\Gamma)} = \int_{\Gamma} \int_{\Gamma} K^{pc}(\|x - y\|, s)\psi(y)\overline{\varphi(x)}d\Gamma_y d\Gamma_x.$$

PROPOSITION 5.9. *Let (5.8) hold. Then, there exists $C_0 > 0$ such that*

$$\|V_h^{pc}(s_l) - V_h(s_l)\|_{+1} \leq C_0 h^{-1} \delta.$$

Hence if $\delta \leq \frac{1}{2} C_0 C_{\text{stab}} h \frac{\min(1, \sigma_0)}{\max_l |s_l|^2} \leq Ch\Delta t^2$, the estimate

$$\|{}^{pc}\phi_n^{\lambda, h} - \phi_n^{\lambda, h}\|_{H^{-1/2}(\Gamma)} \leq C\delta T h^{-1} \lambda^{-N} \Delta t^{-2}$$

holds, where

$${}^{pc}\phi^{\lambda, h} = (V_h^{pc})^{-1} (\partial_t^{\Delta t, \lambda} g).$$

Proof. Let $\varphi \in S$. The well-known L^2 -continuity of the single layer potential for the Laplacian and a scaling inequality for boundary element functions lead to

$$\begin{aligned} & \| (V_h^{pc}(s_l) - V_h(s_l)) \varphi \|_{H^{1/2}(\Gamma)} \\ & \leq \delta \sup_{\substack{\psi \in S(\Gamma) \\ \|\psi\|_{H^{-1/2}(\Gamma)}=1}} \int_{\Gamma \times \Gamma} |\varphi(y)| |\psi(x)| \frac{1}{\|x - y\|} ds_x ds_y \\ & \leq C\delta \sup_{\substack{\psi \in S(\Gamma) \\ \|\psi\|_{H^{-1/2}(\Gamma)}=1}} \|\varphi\|_{L^2(\Gamma)} \|\psi\|_{L^2(\Gamma)} \leq Ch^{-1} \delta \|\varphi\|_{H^{-1/2}(\Gamma)}. \end{aligned}$$

The estimate of the error in the solution is then a direct consequence of Proposition 5.8 and Remark 5.2. \square

In the following result, the binary relation $A \lesssim B$ is used to denote the existence of a constant C independent of any discretization parameters such that $A \leq CB$. Further, $A \sim B$ implies $A \lesssim B$ and $B \lesssim A$.

COROLLARY 5.10. *Let the conditions of Theorem 5.5 be satisfied, let (5.8) hold, and let*

$$h^{m+3/2} \lesssim \Delta t^2, \quad \lambda^{N+1} \sim \Delta t^3, \quad \delta \lesssim \lambda^N h \Delta t^4 \lesssim h^{7m/2+25/4}.$$

Then the optimal rate of convergence is achieved,

$$\|{}^{pc}\phi_n^{\lambda, h} - \phi(\cdot, t_n)\|_{H^{-1/2}(\Gamma)} \leq C\Delta t^2,$$

where C depends on the data g .

Remark 5.11. According to the above result, λ should be chosen as $\lambda \sim \Delta t^{3/(N+1)} = e^{\frac{3}{N+1} \log \frac{T}{N}}$. Since the rounding errors, in the same manner as the errors due to panel-clustering, are magnified by λ^{-j} , λ should be chosen in the interval $\sqrt{\text{eps}} < \lambda^N < 1$, where eps is the machine accuracy. In IEEE double precision this is approximately 10^{-16} ; therefore the accuracy of the method is limited by the choice $\lambda > 10^{-8/N}$. This accuracy limit can, however, be improved if an n -trapezoidal rule is used to compute the weights $\omega_j^{\Delta t, \lambda}$ with $n = jN$, $j > 1$.

Remark 5.12. The condition on the accuracy of the panel-clustering approximation is rather stringent. However, since the convergence of the separable expansion is exponential for a large enough length of expansion L (see (4.7)), the computational costs of the panel-clustering method depend only logarithmically on the required accuracy. Therefore the overall computational cost is not significantly affected.

If we had assumed that $V_h^{\text{pc}}(s) - V_h(s)$ is analytic in s and could be bounded by $C|s|^2$, we could be obtained significantly better error estimates by using Lemma 5.5 in [31]. Unfortunately, due to the well-known numerical stability issues with the multipole expansions for the Helmholtz kernel [5], [9], [34], different types of expansions need to be used for different admissible blocks; the choice of the expansion depends on the wave number s_l and the size of the block. This restricts us from using the more favorable results of Lemma 5.5 in [31].

5.3. Error due to the reduction of the number of linear systems.

COROLLARY 5.13. *Let $0 \leq \lambda < 1$ and $\sigma_l = \text{Re } s_l$. Then*

$$\|\hat{\phi}_l^h\|_{H^{-1/2}(\Gamma)} \leq C_1(\Delta t)^{-2} \|\hat{g}_l\|_{H^{1/2}(\Gamma)},$$

where $C_1 = 5^3 \frac{C_{\text{stab}}}{\min(1, \sigma_l)}$.

Proof. The result is a direct consequence of Lemma 5.1 and Remark 5.2. □

Let $\mathbf{N}_z \subset \{0, 1, \dots, N\}$ determine the Helmholtz problems, the solution of which will be computed; the rest will be approximated by zero. Then we define the resulting approximation to $\phi^{h, \lambda}$ by

$$\emptyset \phi_n^{h, \lambda}(x) := \frac{\lambda^{-n}}{N+1} \sum_{l \in \mathbf{N}_z} \hat{\phi}_l^h(x) \zeta_{N+1}^{ln}.$$

COROLLARY 5.14. *Let $n \in \{0, 1, \dots, N\}$. If*

$$\max_{l \notin \mathbf{N}_z} \|\hat{g}_l\|_{H^{1/2}(\Gamma)} \leq C_1^{-1} \lambda^n (\Delta t)^4,$$

then we obtain optimal order convergence at time step t_n :

$$\|\emptyset \phi_n^{h, \lambda} - \phi_n^{h, \lambda}\|_{H^{-1/2}(\Gamma)} \leq \Delta t^2.$$

Proof. The proof follows directly from Corollary 5.13. □

Next we show that if the right-hand side is smooth and of finite duration, it is sufficient to solve only a few Helmholtz systems. Let us introduce the space of functions that are zero at both $t = 0$ and $t = T$ as follows:

$$H_{00}^r([0, T]; H^{1/2}(\Gamma)) := \left\{ g : \Gamma \times [0, T] \rightarrow \mathbb{R} : \text{there exists } g^* \in H^r(\mathbb{R}; H^{1/2}(\Gamma)) \right. \\ \left. \text{with } g = g^*|_{[0, T]} \text{ and } \text{supp } g^* \subset [0, T] \right\}.$$

THEOREM 5.15. *Let $g \in H_{00}^r([0, T]; H^{1/2}(\Gamma))$ for some $r \geq 3.5$, and let $\epsilon > 0$ be given. For any $N \in \mathbb{N}$ let $\lambda := \epsilon^{\frac{1}{N}}$. Then, \mathbf{N}_z can be chosen so that $\#\mathbf{N}_z \leq C\epsilon^{-\frac{1}{r+1/2}} N^{\frac{4}{r+1/2}}$ and the optimal order convergence is retained. The constant C depends on $r, (\log \epsilon)/T$, and g .*

Proof. Let $g_\lambda(x, t) := \lambda^{t/\Delta t} g(x, t) = \epsilon^{\frac{t}{T}} g(x, t) = e^{t \frac{\log \epsilon}{T}} g(x, t)$ on $t \in [0, T]$. Then we see that g_λ is independent of N and that $g_\lambda \in H_{00}^r([0, T]; H^{1/2}(\Gamma))$. Then for $\omega \in \mathbb{R}$, $\|(\mathcal{F}g_\lambda)(\cdot, \omega)\|_{H^{1/2}(\Gamma)} = o(|\omega|^{-r-1/2})$. Taking $\omega_j = 2\pi j/(T + \Delta t) = 2\pi jN/(T(N + 1))$, we define

$$a_j := \|(\mathcal{F}g_\lambda)(\cdot, \omega_j)\|_{H^{1/2}(\Gamma)} = o(j^{-r-1/2}), \quad j \in \mathbb{Z}.$$

Then using the aliasing formula (see [25]), we arrive at the following estimate for \hat{g}_n , $n = 1, \dots, N/2 - 1$:

$$\|\hat{g}_n\|_{H^{1/2}(\Gamma)} \leq a_n + \sum_{k>N/2} a_k = o(n^{-r-1/2} + N^{-r+1/2}) = o(n^{-r-1/2}).$$

The constants in the $o(\cdot)$ notation depend only on $r, (\log \epsilon)/T, \epsilon$, and g . The result now follows from Corollary 5.14. \square

6. Numerical experiments. In this section we present the results of numerical experiments. Except for one simple example, the experiments will be done in two dimensions. All the steps in the method remain the same in two dimensions, except that the fundamental solution for the wave equation is given by

$$(6.1) \quad k_{2D}(d, t) = \frac{H(t - d)}{2\pi\sqrt{t^2 - d^2}},$$

where H is the Heaviside function,

$$H(t) = \begin{cases} 0 & \text{for } t < 0, \\ 1 & \text{for } t > 0. \end{cases}$$

The Laplace transform $K_{2D}(d, s)$ is again the fundamental solution of the Helmholtz equation $\Delta U - s^2 U$ as follows:

$$(6.2) \quad K_{2D}(d, s) = \frac{i}{4} H_0^{(1)}(isd),$$

where $H_0^{(1)}(\cdot)$ is the zero order Hankel function of the first kind.

Let us consider the case of Γ being the unit ball in \mathbb{R}^2 or \mathbb{R}^3 and a right-hand side that is separable in the time and the spatial variables: $g(x, t) = g(t)e(x)$, where $e(x)$ is an eigenfunction of the single layer potential $V(s)$ with the eigenvalue $\lambda_l(s)$. In two dimensions the eigenfunctions are the complex exponentials $e^{il\theta}$ and $\lambda_l(s) = \frac{i\pi}{2} J_l(is)H_l(is)$, whereas in three dimensions these are the spherical harmonics $Y_l^m(\theta, \varphi)$ with $\lambda_l(s) = -sj_l(is)h_l(is)$; we have used the standard polar/spherical coordinates to describe the eigenfunctions. Here $J_l(\cdot)$ (respectively, $j_l(\cdot)$) are cylindrical (respectively, spherical) Bessel functions of order l , whereas $H_l^{(1)}(\cdot)$ (respectively, $h_l^{(1)}(\cdot)$) are the cylindrical (respectively, spherical) Hankel functions of the first kind and order l . The problem of finding the unknown density $\phi(x, t)$ can then be reduced to the single, time, dimension. This can be seen by replacing the fundamental solution k in the

single layer representation formula by the inverse Laplace transform of its Laplace transform K as follows:

$$\begin{aligned} g(t)e(x) &= \int_0^t \int_{\Gamma} k(t-\tau, \|x-y\|)\phi(\tau, x)d\Gamma_y d\tau \\ &= \frac{1}{2\pi i} \int_{\sigma-i\infty}^{\sigma+i\infty} \int_0^t e^{s\tau} \int_{\Gamma} K(s, \|x-y\|)\phi(t-\tau, y)d\Gamma_y d\tau ds \\ &= \frac{1}{2\pi i} \int_{\sigma-i\infty}^{\sigma+i\infty} \int_0^t e^{s\tau} (V(s)\phi(t-\tau, \cdot))(x) d\tau ds, \quad x \in \Gamma, \text{ for some } \sigma > 0. \end{aligned}$$

Therefore, we can use the ansatz $\phi(x, t) = \phi(t)e(x)$ to reduce the problem to finding $\phi(t)$ such that

$$g(t) = \frac{1}{2\pi i} \int_{\sigma-i\infty}^{\sigma+i\infty} \int_0^t e^{s\tau} \lambda_l(is)\phi(t-\tau) d\tau ds.$$

Hence we need to solve a convolution integral equation in one dimension as follows:

$$(6.3) \quad g(t) = \int_0^t \check{\lambda}_l(\tau)\phi(t-\tau) d\tau,$$

where $\check{\lambda}_l(\cdot)$ is the inverse Laplace transform of $\lambda_l(\cdot)$. The latter equation can then be solved by Lubich’s original method, which makes use of only $\lambda_l(\cdot)$ and not its inverse Laplace transform. The first few numerical examples will be of this type.

6.1. Radial solution of scattering by unit sphere. In this example we consider the three-dimensional case, $\Gamma = \mathbb{S}^2$. Let $g(x, t) = g(t)$ be constant for a fixed time t , i.e., $e(x) = 2\sqrt{\pi}Y_0^0 = 1$. In this particularly simple case it can be shown that

$$\phi(t) = 2g'(t), \quad t \in [0, 2].$$

The restriction to the interval $[0, 2]$ is a consequence of the fact that the diameter of the sphere is 2. For time $t > 2$ the expression for $\phi(t)$ is more complicated.

The right-hand side of the n th Helmholtz problem is a constant,

$$\hat{g}_n = \sum_{j=0}^N \lambda^j g(t_j) \zeta_{N+1}^{-nj},$$

and the solution of the Helmholtz problem is also a constant and is given by

$$\hat{\phi}_n = \frac{\hat{g}_n}{\lambda_0(s_n)}.$$

The approximation to the unknown density at time step t_n is given by

$$\phi_n := \frac{\lambda^{-n}}{N+1} \sum_{j=0}^N \hat{\phi}_j \zeta_{N+1}^{nj}.$$

If λ is chosen small enough, theoretical estimates predict the following behavior of the error:

$$\left(\sum_{n=0}^N \Delta t |\phi(t_n) - \phi_n|^2 \right)^{1/2} \leq C \Delta t^2.$$

TABLE 1

The results for scattering by unit sphere with $g(x, t) = \sin^5(t)$ and $\lambda = \Delta t^{3/N}$.

N	Error	Rate
4	1.44	-/-
8	0.45	1.68
16	0.12	1.90
32	0.032	1.94
64	0.0081	1.96
128	0.0020	1.99
256	0.00051	1.99
512	0.00013	2.00
1024	3.2×10^{-5}	2.00

One more detail needs to be fixed before the experiments can be started, namely, the choice of λ . Recall that λ needs to be chosen small enough to insure stability and accuracy (see Theorem 5.5) but also large enough to avoid numerical instability issues (see Remark 5.11). As suggested in Remark 5.11, we make the choice

$$(6.4) \quad \lambda = \max(\Delta t^{3/N}, \text{eps}^{\frac{1}{2N}}).$$

Numerical results for the scattering by unit sphere are given in Table 1 and show that our theoretical estimates are sharp for this example.

6.2. A nonradial example. In this example we consider the two-dimensional case. We pick the right-hand side to be $g(x, t) = h(t) \cos(l\theta)$, where for the space variable we use the polar coordinate system $r \in \mathbb{R}_{\geq 0}$, $\theta \in [0, 2\pi)$. Since $\cos(l\theta)$ is an eigenfunction of the single layer potential $V(s_n)$, the Helmholtz problems can be solved exactly. However, to investigate the effect of spatial discretization we solve the problems using the Galerkin method, and hence obtain an approximation $\phi^{h,\lambda}(t_n, \theta)$ of the unknown density. To investigate the error, we use the fact that $\phi(\theta, t) = \phi(t) \cos(l\theta)$ and solve with high accuracy for $\phi(t)$ by applying Lubich’s method to the one-dimensional problem (6.3). The error measure we use is the following:

$$\|\phi - \phi^{h,\lambda}\|_{-1/2,l^2} := \left(\sum_{n=0}^N \Delta t \|\phi(t_n) \cos(l\cdot) - \phi^{h,\lambda}(t_n, \cdot)\|_{H^{-1/2}(\Gamma)}^2 \right)^{1/2}.$$

The theory predicts the above error to be proportional to $h^{m+3/2} + \Delta t^2$, where $m = 0$ for the Galerkin basis of piecewise constant functions and $m = 1$ for the basis of piecewise linear functions. In all the experiments, we choose λ as in (6.4). To see if the spatial discretization has introduced significant errors, we compute the error obtained when the Helmholtz problems are solved exactly. The results are given in the following table:

N	4	8	16	32	64	128
$\ \phi - \phi^{h,\lambda}\ _{-1/2,l^2}$	0.61	0.24	0.077	0.022	0.0057	0.0015

Comparing these results to Tables 2 and 3, we see that the error due to the discretization in space is not significant.

6.3. Reduction of the number of systems. Let us now consider a signal that is smooth and of nearly limited time duration as follows:

$$(6.5) \quad g(\mathbf{r}, t) = \cos(5t - \mathbf{r} \cdot \alpha) \exp(-1.5(5t - \mathbf{r} \cdot \alpha - 5)^2), \quad \alpha = (1, 0).$$

TABLE 2

The results for scattering by the unit disk with $g(x, t) = \sin^5(t) \cos(3x)$ and the piecewise constant Galerkin basis $S = S_{-1,0}$. M is chosen so that $h^{3/2} \propto \Delta t^2$.

N	M	$\ \phi - \phi^{h,\lambda}\ _{-1/2,l^2}$	Rate
4	16	0.78	-/-
8	40	0.27	1.54
16	102	0.084	1.68
32	254	0.023	1.83
64	640	0.0062	1.93
128	1610	0.0016	1.98

TABLE 3

The results for scattering by the unit disk with $g(x, t) = \sin^5(t) \cos(3x)$ and the piecewise linear Galerkin basis $S = S_{0,1}$. M is chosen so that $h^{5/2} \propto \Delta t^2$.

N	M	$\ \phi - \phi^{h,\lambda}\ _{-1/2,l^2}$	Rate
4	22	0.66	-/-
8	40	0.26	1.34
16	68	0.082	1.67
32	116	0.023	1.84
64	204	0.0060	1.93
128	352	0.0015	1.99

TABLE 4

The results for scattering by the unit disk where the incoming wave is a Gaussian pulse and the piecewise linear Galerkin basis $S = S_{1,0}$ is used. The column $\#\mathbf{N}_z$ shows the number of Helmholtz problems actually solved.

N	$\#\mathbf{N}_z$	M	$\ \phi - \phi^{h,\lambda}\ _{-1/2,l^2}$	Rate
4	3	24	2.9	-/-
8	5	40	2.9	-0.03
16	9	68	1.4	1.09
32	17	116	0.42	1.70
64	24	204	0.11	1.92
128	24	352	0.028	1.98
256	24	612	0.0072	1.99

For such a Gaussian pulse our theory predicts that only $\mathcal{O}(N^\epsilon)$, for any fixed $\epsilon > 0$, Helmholtz systems need to be solved to obtain optimal convergence; see also Figure 2. The results for scattering by the unit disk and for piecewise-linear basis functions $S = S_{1,0}$ are given in Table 4. Since we approximate by zero only the solutions of those Helmholtz problems whose right-hand sides are zero almost to machine precision, the number of Helmholtz problems $\#\mathbf{N}_z$ is constant for large enough N . For this more complicated problem, for each N we have used as the reference solution the numerical solution using $2N$ steps in time and the corresponding number of nodes in the discretization in space.

7. Conclusion. We have described a method that requires the solution of a number of Helmholtz problems to obtain an approximate solution of the wave equation in an unbounded, homogeneous medium. We have proved stability and optimal convergence results for this approach. Further, we have indicated ways in which to efficiently solve the resulting system of Helmholtz problems. The stability and convergence results of the perturbations introduced by the efficient solvers have also been presented.

The fast methods we propose using are typically capable of computing a matrix-

vector product in almost linear time, i.e., $\mathcal{O}(M \log^a M)$, of a single dense $M \times M$ system arising from the discretization of the Helmholtz single layer potential. In order to solve efficiently the linear system by an iterative method requiring only matrix-vector multiplication, a good preconditioner is needed. The investigation of such a preconditioner is beyond the scope of this paper. With a preconditioned iterative solver we expect to obtain computational costs which scale linearly, up to logarithmic terms, with respect to the number of unknowns NM . An important observation is that in some cases only a few Helmholtz systems need to be solved. Although this does not change the overall complexity (the discrete Fourier transformation still requires $\mathcal{O}(MN \log N)$ operations), it can hugely reduce the absolute time for the computation. The storage costs will also scale linearly since at any one time only a single linear system representing the discretization of a Helmholtz problem needs to be stored. Since all the NM coefficients $\phi_{j,n}$ are stored, the storage costs are not better than linear. Crucially, since the Helmholtz problems to be solved are entirely decoupled, the proposed method is easily parallelizable.

These asymptotic estimates significantly improve both the storage and computational costs compared to the previously proposed approaches for the solution of the wave equation using the convolution quadrature discretization in time; see [24] and [22], [23], [28]. The asymptotic costs of the MOT method presented in [10], [18] are also almost linear in the number of degrees of freedom. Advantages of our method include the intrinsic parallel nature of the method, proven convergence and stability properties, and the relatively simple implementation details. In a forthcoming paper, algorithmic details for the data sparse approximations, a more in-depth asymptotic complexity analysis, and large-scale computational results will be presented.

Acknowledgments. We gratefully acknowledge the fruitful discussions with Ch. Lubich which led to significant improvements in the perturbation analysis.

REFERENCES

- [1] S. AMINI AND A. PROFIT, *Analysis of the truncation errors in the fast multipole method for scattering problems*, in Proceedings of the 8th International Congress on Computational and Applied Mathematics, ICCAM-98 (Leuven), J. Comput. Appl. Math., 115 (2000), pp. 23–33.
- [2] S. AMINI AND A. PROFIT, *Multi-level fast multipole solution of the scattering problem*, Eng. Anal. Boundary Elements, 27 (2003), pp. 547–564.
- [3] A. BAMBERGER AND T. HA DUONG, *Formulation variationnelle espace-temps pour le calcul par potentiel retardé de la diffraction d'une onde acoustique*. I, Math. Methods Appl. Sci., 8 (1986), pp. 405–435.
- [4] A. BAMBERGER AND T. HA DUONG, *Formulation variationnelle pour le calcul de la diffraction d'une onde acoustique par une surface rigide*, Math. Methods Appl. Sci., 8 (1986), pp. 598–608.
- [5] L. BANJAI AND W. HACKBUSCH, *Hierarchical matrix techniques for low- and high-frequency Helmholtz problems*, IMA J. Numer. Anal., 28 (2008), pp. 46–79.
- [6] L. BANJAI AND S. SAUTER, *Rapid solution of the wave equation in unbounded domains: Abridged version*, in Proceedings of Waves 2007, University of Reading, Reading, UK, 2007, pp. 38–40.
- [7] B. BIRGISSON, E. SIEBRITS, AND A. PIERCE, *Elastodynamic direct boundary element methods with enhanced numerical stability properties*, Internat. J. Numer. Methods Engrg., 46 (1999), pp. 871–888.
- [8] M. BLUCK AND S. WALKER, *Analysis of three-dimensional transient acoustic wave propagation using the boundary integral equation method*, Internat. J. Numer. Methods Engrg., 39 (1996), pp. 1419–1431.
- [9] H. CHENG, W. Y. CRUTCHFIELD, Z. GIMBUTAS, L. F. GREENGARD, J. F. ETHRIDGE, J. HUANG, V. ROKHLIN, N. YARVIN, AND J. ZHAO, *A wideband fast multipole method for the Helmholtz equation in three dimensions*, J. Comput. Phys., 216 (2006), pp. 300–325.

- [10] W. C. CHEW, J.-M. JIN, E. MICHELSEN, AND J. M. SONG, *Fast and Efficient Algorithms in Computational Electromagnetics*, Artech House, Boston, London, 2001.
- [11] P. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1987.
- [12] M. COSTABEL, *Developments in boundary element methods for time-dependent problems*, in Problems and Methods in Mathematical Physics, L. Jentsch and F. Tröltzsch, eds., Teubner, Leipzig, 1994, pp. 17–32.
- [13] E. DARVE AND P. HAVÉ, *Efficient fast multipole method for low-frequency scattering*, J. Comput. Phys., 197 (2004), pp. 341–363.
- [14] P. J. DAVIES, *Numerical stability and convergence of approximations of retarded potential integral equations*, SIAM J. Numer. Anal., 31 (1994), pp. 856–875.
- [15] P. DAVIES, *Averaging techniques for time marching schemes for retarded potential integral equations*, Appl. Numer. Math., 23 (1997), pp. 291–310.
- [16] P. J. DAVIES AND D. B. DUNCAN, *Stability and convergence of collocation schemes for retarded potential integral equations*, SIAM J. Numer. Anal., 42 (2004), pp. 1167–1188.
- [17] Y. DING, A. FORESTIER, AND T. HA-DUONG, *A Galerkin scheme for the time domain integral equation of acoustic scattering from a hard surface*, J. Acoust. Soc. Amer., 86 (1989), pp. 1566–1572.
- [18] A. ERGIN, B. SHANKER, AND E. MICHELSEN, *Fast analysis of transient acoustic wave scattering from rigid bodies using the multilevel plane wave time domain algorithm*, J. Acoust. Soc. Amer., 117 (2000), pp. 1168–1178.
- [19] M. FRIEDMAN AND R. SHAW, *Diffraction of pulses by cylindrical obstacles of arbitrary cross section*, J. Appl. Mech., 29 (1962), pp. 40–46.
- [20] T. HA-DUONG, *On retarded potential boundary integral equations and their discretization*, in Computational Methods in Wave Propagation, Vol. 31, M. Ainsworth, P. Davies, D. Duncan, P. Martin, and B. Rynne, eds., Springer, Heidelberg, 2003, pp. 301–336.
- [21] T. HA-DUONG, B. LUDWIG, AND I. TERRASSE, *A Galerkin BEM for transient acoustic scattering by an absorbing obstacle*, Internat. J. Numer. Methods Engrg., 57 (2003), pp. 1845–1882.
- [22] W. HACKBUSCH, W. KRESS, AND S. SAUTER, *Sparse convolution quadrature for time domain boundary integral formulations of the wave equation by cutoff and panel-clustering*, in Boundary Element Analysis: Mathematical Aspects and Applications, M. Schanz and O. Steinbach, eds., Lect. Notes Appl. Comput. Mech. 29, Springer, Berlin, 2006, pp. 113–134.
- [23] W. HACKBUSCH, W. KRESS, AND S. SAUTER, *Sparse convolution quadrature for time domain boundary integral formulations of the wave equation*, IMA J. Numer. Anal., to appear.
- [24] E. HAIBER, CH. LUBICH, AND M. SCHLICHTE, *Fast numerical solution of nonlinear Volterra convolution equations*, SIAM J. Sci. Stat. Comput., 6 (1985), pp. 532–541.
- [25] P. HENRICI, *Fast Fourier methods in computational complex analysis*, SIAM Rev., 21 (1979), pp. 481–527.
- [26] T. HOHAGE AND F.-J. SAYAS, *Numerical solution of a heat diffusion problem by boundary element methods using the Laplace transform*, Numer. Math., 102 (2005), pp. 67–92.
- [27] M. KÖHL AND S. RJASANOW, *Multifrequency analysis for the Helmholtz equation*, Comput. Mech., 32 (2003), pp. 234–239.
- [28] W. KRESS AND S. SAUTER, *Numerical treatment of retarded boundary integral equations by sparse panel clustering*, IMA J. Numer. Anal., 28 (2008), pp. 162–185.
- [29] C. LUBICH, *Convolution quadrature and discretized operational calculus I*, Numer. Math., 52 (1988), pp. 129–145.
- [30] C. LUBICH, *Convolution quadrature and discretized operational calculus II*, Numer. Math., 52 (1988), pp. 413–425.
- [31] C. LUBICH, *On the multistep time discretization of linear initial-boundary value problems and their boundary integral equations*, Numer. Math., 67 (1994), pp. 365–389.
- [32] C. LUBICH AND R. SCHNEIDER, *Time discretization of parabolic boundary integral equations*, Numer. Math., 63 (1992), pp. 455–481.
- [33] E. MILLER, *An overview of time-domain integral equations models in electromagnetics*, J. Electromagn. Waves Appl., 1 (1987), pp. 269–293.
- [34] S. OHNUKI AND W. C. CHEW, *Truncation error analysis of multipole expansion*, SIAM J. Sci. Comput., 25 (2003), pp. 1293–1306.
- [35] V. ROKHLIN, *Rapid solution of integral equations of scattering theory in two dimensions*, J. Comput. Phys., 86 (1990), pp. 414–439.
- [36] V. ROKHLIN, *Diagonal forms of translation operators for the Helmholtz equation in three dimensions*, Appl. Comput. Harmon. Anal., 1 (1993), pp. 82–93.

- [37] B. RYNNE AND P. SMITH, *Stability of time marching algorithms for the electric field integral equation*, J. Electromagn. Waves Appl., 4 (1990), pp. 1181–1205.
- [38] S. SAUTER AND C. SCHWAB, *Randelementmethoden*, Teubner, Leipzig, 2004.
- [39] A. SCHÄDLE, M. LÓPEZ-FERNÁNDEZ, AND C. LUBICH, *Fast and oblivious convolution quadrature*, SIAM J. Sci. Comput., 28 (2006), pp. 421–438.
- [40] M. SCHANZ, *Wave Propagation in Viscoelastic and Poroelastic Continua. A Boundary Element Approach*, Lecture Notes Appl. Comput. Mech. 2, Springer, Berlin, 2001.
- [41] M. SCHANZ AND H. ANTES, *Application of operational quadrature methods in time domain boundary element methods*, Meccanica, 32 (1997), pp. 179–186.
- [42] M. SCHANZ, H. ANTES, AND T. RÜBERG, *Convolution quadrature boundary element method for quasi-static visco- and poroelastic continua*, Comput. & Structures, 83 (2005), pp. 673–684.
- [43] D. SHEEN, I. H. SLOAN, AND V. THOMÉE, *A parallel method for time discretization of parabolic equations based on Laplace transformation and quadrature*, IMA J. Numer. Anal., 23 (2003), pp. 269–299.
- [44] O. VON ESTORFF, S. RJASANOW, M. STOLPER, AND O. ZALESKI, *Two efficient methods for a multifrequency solution of the Helmholtz equation*, Comput. Vis. Sci., 8 (2005), pp. 159–167.

A SEMI-IMPLICIT SCHEME FOR STATIONARY STATISTICAL PROPERTIES OF THE INFINITE PRANDTL NUMBER MODEL*

WENFANG (WENDY) CHENG[†] AND XIAOMING WANG[‡]

Abstract. We propose a semidiscrete in time semi-implicit numerical scheme for the infinite Prandtl model for convection. Besides the usual finite time convergence, this scheme enjoys the additional highly desirable feature that the stationary statistical properties of the scheme converge to those of the infinite Prandtl number model at vanishing time step. One of the key characteristics of the scheme is that it preserves the dissipativity of the infinite Prandtl number model uniformly in terms of the time step. So far as we know, this is the first rigorous result on convergence of stationary statistical properties of numerical schemes for infinite dimensional dissipative complex systems.

Key words. stationary statistical property, infinite Prandtl number model, uniformly dissipative scheme, Nusselt number

AMS subject classifications. 65M12, 65Z05, 65P99, 37M25, 76D06, 76M25, 76R10

DOI. 10.1137/080713501

1. Introduction. Many dynamical systems arising in applications are dissipative complex systems in the sense that they possess a compact global attractor and the dynamics on the global attractor are complex/chaotic [39]. Well-known examples include the simple Lorenz 63 model, Lorenz 96 model, the Navier–Stokes equations at large Reynolds number or Grashoff number, the Boussinesq system for convection at large Rayleigh number, the Kuramoto–Sivashinsky equation at large spatial size, and many models for the atmosphere, ocean, weather, and climate, etc. The dynamics of these systems are typically very complex/chaotic with generic sensitive dependence on data. Therefore, it is hardly meaningful to discuss long time behavior of a single trajectory for this kind of complex system. Instead, we should study statistical properties of the system since they are physically much more relevant than single trajectories [33, 15, 31, 29]. If the system reaches some kind of stationary state, then the objects that characterize the stationary statistical properties are the invariant measures or stationary statistical solutions of the system.

With a given complex system, analytical exact expressions for statistical properties are extremely rare, just as exact solution formulas are rare for single trajectory. Therefore we naturally turn to numerical methods, especially with today’s powerful computers and ever advancing computational technologies. The natural question then is what kind of numerical schemes would provide good approximations for the stationary statistical properties.

In terms of trajectory approximations, we are not aware of any effective long time integrator for dissipative complex/chaotic systems in general unless the long time dynamics is trivial or the trajectory under approximation is stable [21, 17, 28]. It is not

*Received by the editors January 17, 2008; accepted for publication (in revised form) July 23, 2008; published electronically October 31, 2008. This work was supported in part by grants from the NSF.

<http://www.siam.org/journals/sinum/47-1/71350.html>

[†]Department of Mathematics, Florida State University, Tallahassee, FL 32306 (wfcheng00@yahoo.com).

[‡]Corresponding author. Department of Mathematics, Florida State University, Tallahassee, FL 32306 (wxm@math.fsu.edu). Part of this author’s work was done while the author was a visitor at the Courant Institute at New York University, and at the Institute for Applied Mathematics at the University of Bonn.

at all clear whether those numerical methods that provide efficient and accurate approximations of the continuous complex dynamical system on a finite time interval are able to provide meaningful approximation for stationary statistical properties of the system since small errors (truncation and rounding) may accumulate and grow over a long time (think about the usual error estimates with a coefficient that grows exponentially in time due to the existence of chaotic behavior/positive Lyapunov exponent). Here we forego the idea of long time fidel approximation of any single trajectory, but ask if it is possible to approximate the mean or statistical properties faithfully. The numerical study of stationary statistical properties of complex system still is a very challenging task since it involves long time integration (so that the statistical averaging is computed utilizing time averaging under the assumption of ergodicity) and computation of a large number of trajectories (if no ergodicity is assumed).

We will demonstrate in this paper that a semidiscrete in time and semi-implicit scheme for the infinite Prandtl number model for convection is able to capture stationary statistical properties of the underlying infinite Prandtl number model. It seems that one of the key ingredients in the convergence of the stationary statistical properties is the *uniformly dissipativity* of the scheme; i.e., the scheme is dissipative uniformly with respect to the time step. Although this scheme may not approximate individual trajectory faithfully for a long time due to the accumulation of truncation and rounding errors and abundant instability/chaos, we will show that *stationary statistical properties characterized by the invariant measures (stationary statistical solutions) of the scheme converge to those of the continuous-in-time system*. This gives us strong evidence that these kinds of uniformly dissipative schemes are appropriate schemes in investigating statistics.

Although the idea of uniform dissipativity and convergence of stationary statistical properties is illustrated on the infinite Prandtl number model for convection and semidiscretization in time only, we believe that the methodology works for many more complex/chaotic dynamical systems [39] and fully discretized approximations. The key ingredients are uniform (in mesh size) dissipativity and finite time uniform convergence (see [48] for a somewhat general statement). The choice of the infinite Prandtl number model is both for its physical significance (see the next section) and for the sake of simplicity in exposition.

The idea of uniformly dissipative approximation for a dissipative dynamical system is a very natural one. Since the continuous-in-time dynamical system is dissipative (possess a global attractor), it is natural to consider numerical schemes that preserve the dissipativity in the sense that the solutions to the schemes should possess global attractors that are uniformly compact in some appropriate sense (say the union of the global attractors is precompact). These uniformly dissipative schemes are usually implicit in some way (to ensure long time stability) and therefore have not been very popular in practice so far. What we shall demonstrate below is that some of these uniformly dissipative schemes enjoy a highly desirable property in terms of approximating stationary statistical properties: the stationary statistical properties of the schemes converge to those of the continuous-in-time dynamical systems. We hope that our work will stimulate further study, both analytical and numerical, on approximating statistical properties of dissipative systems.

Earlier works on long time behavior of numerical schemes for dissipative systems mainly focused on the two-dimensional incompressible Navier–Stokes system and the Kuramoto–Sivashinsky equation (see [17, 22, 26, 36, 40, 13, 14, 25] among others) and the notion of long time stability or dissipativity. The uniform bound in the phase space and a finer/smaller space is called long time stability or dissipativity in

these works (some of the authors derived only uniform bound/long time stability in the phase space without other bounds that are necessary for ensuring the uniform dissipativity of the scheme). We prefer the term uniform dissipativity since long time stability could be misleading in the sense that it may imply the scheme is global in time, stable in one single phase space only, which is not sufficient to ensure the existence of the global attractor. Also, none of the authors discussed stationary statistical properties of their schemes. To the best of our knowledge, our work is the first in establishing the convergence of stationary statistical properties and therefore the usefulness of uniformly dissipative schemes in approximating stationary statistical properties. An announcement of the main results presented here can be found in [10].

The manuscript is organized as follows: we give an introduction in the first section; in section 2 we propose a semidiscrete (discrete in time) scheme for the infinite Prandtl number model and verify that it is uniformly dissipative and enjoys the property that the stationary statistical properties of the scheme converge to those of the continuous-in-time model; we then provide our conclusion and remarks in the third section.

2. A uniformly dissipative scheme for the infinite Prandtl number model.

2.1. The infinite Prandtl number model for convection. One of the fundamental systems in fluid dynamics is the *Boussinesq system for Rayleigh–Bénard convection*, which is a model for convection; i.e., fluid motion induced by differential heating under Boussinesq approximation [41, 16]. We assume that the fluids occupy the (nondimensionalized) region $\Omega = [0, L_x] \times [0, L_y] \times [0, 1]$ with periodicity imposed in the horizontal directions for simplicity.

The Boussinesq system exhibits extremely rich phenomena (see, for instance, [16, 41] and the recent reviews [4, 37]). In fact, the Boussinesq system is considered a fundamental paradigm for nonlinear dynamics including instabilities and bifurcations, pattern formation, chaotic dynamics, and fully developed turbulence [27]. On the other hand, we have very limited mathematical knowledge on the system. Therefore various physically relevant simplifications are highly desirable in order to make progress.

For fluids such as silicone oil or the earth’s mantle, the Prandtl number is large; therefore, we may formally set the Prandtl number to infinity in the nondimensional Boussinesq system, and we arrive at the following (see, for instance, [4, 6, 8, 18, 41] among others) *infinite Prandtl number model (nondimensional)*:

$$(2.1) \quad \nabla p = \Delta \mathbf{u} + Ra \mathbf{k}T, \quad \nabla \cdot \mathbf{u} = 0, \quad \mathbf{u}|_{z=0,1} = 0,$$

$$(2.2) \quad \frac{\partial T}{\partial t} + \mathbf{u} \cdot \nabla T = \Delta T, \quad T|_{z=0} = 1, T|_{z=1} = 0,$$

where \mathbf{u} is the Eulerian velocity of the fluid, p represents the kinematic pressure of the fluid, T is the temperature of the fluid, \mathbf{k} is a unit vector in the z direction, and Ra is the Rayleigh number measuring the ratio of differential heating over overall dissipation.

It is well known that for complex systems such as a convection system at large Rayleigh number where turbulent/chaotic behavior abounds (see, for instance, [6, 16, 27, 4, 37]), statistical properties for such systems are much more important and physically relevant than single trajectories [33, 15, 29, 31].

Although there have been extensive works on heat transport in Rayleigh–Bénard convection [1, 4, 6, 7, 18, 37, 23, 24], basic statistical properties of the system, such

as the heat transport in the vertical direction quantified by the Nusselt number and the mean velocity field, are not very well understood. On the other hand, the infinite Prandtl number model is much simpler than the Boussinesq system since the Navier–Stokes equations are replaced by the Stokes equations (and therefore the phase space is that of the temperature only). We also know that the statistics of the infinite Prandtl number model are close to those of the Boussinesq system at large Prandtl number [44, 45, 46]. Therefore it makes sense for us to study some fundamental statistical properties of convection utilizing the simple infinite Prandtl number model since we can generally expect to push to a physically more interesting higher Rayleigh number without sacrificing accuracy with the currently available computing resource.

In our case of infinite Prandtl number convection at large Rayleigh number, even the computation on order one time scale (diffusive time scale) is a challenge since it is in fact a long time integration in disguise. To see this, we can rewrite the infinite Prandtl number model as

$$(2.3) \quad \frac{\partial T}{\partial t} + Ra A^{-1}(\mathbf{k}T) \cdot \nabla T = \Delta T, \quad T|_{z=0} = 1, T|_{z=1} = 0,$$

where A denotes the Stokes operator with viscosity one and the associated boundary conditions. It is then apparent that this is an advection dominated problem (large Péclet number) for large Rayleigh number Ra . We divide both sides of the equation by Ra and introduce the fast time scale $\tau = Ra t$, for which we may rewrite the infinite Prandtl number model in the following alternative form with an order one advection term

$$(2.4) \quad \frac{\partial T}{\partial \tau} + A^{-1}(\mathbf{k}T) \cdot \nabla T = \frac{1}{Ra} \Delta T.$$

It appears that the leading order dynamics at large Rayleigh number is the nonlocal advection equation $\frac{\partial T}{\partial \tau} + A^{-1}(\mathbf{k}T) \cdot \nabla T = 0$. However, this is valid only on order one time scale for the fast time τ . What we are interested in is order one time scale for the diffusive time t , which means a long time for the fast time τ (of the order of Ra).

2.2. A semidiscrete in time scheme. In this subsection, we provide a specific semidiscrete in time convergent dissipative scheme for the infinite Prandtl number model. The scheme is semi-implicit and utilizes a background temperature profile. Indeed, consider a generic background temperature profile $\tau(z)$ which satisfies the nonhomogeneous Dirichlet boundary condition of T . We introduce the perturbative temperature field $\theta = T - \tau$. The exact form of the background profile τ to be used will be specified below. It is easy to see that θ satisfies the following equation:

$$(2.5) \quad \frac{\partial \theta}{\partial t} + Ra A^{-1}(\mathbf{k}\theta) \cdot \nabla \theta + Ra A^{-1}(\mathbf{k}\theta)_3 \tau'(z) = \Delta \theta + \tau''(z), \quad \theta|_{z=0,1} = 0,$$

and we are searching for a solution in the space $H_{0,per}^1$ (the subspace of H^1 with zero trace in the z direction and periodic in the horizontal directions). Here $A^{-1}(\mathbf{k}\theta)_3$ represents the third component (vertical velocity) of $A^{-1}(\mathbf{k}\theta)$.

The *semi-implicit semidiscrete in time scheme* that we propose is given by

$$(2.6) \quad \frac{\theta^{n+1} - \theta^n}{k} + Ra A^{-1}(\mathbf{k}\theta^n) \cdot \nabla \theta^{n+1} + Ra A^{-1}(\mathbf{k}\theta^{n+1})_3 \tau'(z) = \Delta \theta^{n+1} + \tau''(z),$$

where θ^n denotes the approximate solution at time kn where k is the time step. A more accurate notation would be θ_k^n to indicate the dependence on the time step k .

However, we will suppress the k dependence in the notion for simplicity except in the convergence proof.

Note that the scheme is linear although the PDE is nonlinear.

Also, we would arrive at a different scheme if we were to discretize in time first and then apply the translation/background profile (see (3.2) for the case of $\lambda = 0$).

Following the pioneering works of Constantin and Doering [7, 8], we set background temperature profile τ to be a locally smoothed (mollified) version of the following piecewise linear function:

$$(2.7) \quad \tau(z) = \begin{cases} 1 - \frac{z}{2\delta}, & 0 \leq z \leq \delta, \\ \frac{1}{2}, & \delta \leq z \leq 1 - \delta, \\ \frac{1-z}{2\delta}, & 1 - \delta \leq z \leq 1. \end{cases}$$

The choice of the parameter δ will be specified later.

We would like to remark here that the typical choice of τ being the conduction state $1 - z$ is not a good one. In fact, the linearized equation is unstable in this case [6] and the solutions (to the linearized problem) grow without bound for generic initial data. Therefore we have to utilize the nonlinear term (this is where the new background profile comes into the picture) to stabilize the whole system. It is also worthwhile to point out that boundary conditions play an important role in the stabilization process. For instance, if we choose $\tau = 1 - z$ (the pure conduction state) and utilize periodicity for the perturbative variables in all three directions (the so-called homogeneous Rayleigh–Bénard convection), then the nonlinear system is not stable [5] (look at solutions that are z independent).

2.3. Well-posedness. The well-posedness of the discrete scheme follows from the Lax–Milgram theorem [30].

The weak formulation of the scheme can be derived by multiplying the scheme (2.6) by a test function $\psi \in H_{0,per}^1$ and integrating by parts. The weak formulation of the discrete scheme can be rewritten into the form

$$(2.8) \quad B_n(\theta^{n+1}, \psi) = L_n(\psi),$$

where

$$(2.9) \quad B_n(\theta^{n+1}, \psi) = \left(\frac{1}{k} \theta^{n+1} + Ra A^{-1}(\mathbf{k}\theta^n) \cdot \nabla \theta^{n+1} + Ra A^{-1}(\mathbf{k}\theta^{n+1})_3 \tau' \right) + (\nabla \theta^{n+1}, \nabla \psi),$$

$$(2.10) \quad L_n(\psi) = - \left(\tau'(z), \frac{\partial \psi}{\partial z} \right) + \left(\frac{1}{k} \theta^n, \psi \right).$$

It is easy to see that B_n is a continuous bilinear form on $H_{0,per}^1 \times H_{0,per}^1$, and L_n is a continuous linear functional on $H_{0,per}^1$. We need only verify the coercivity for B_n in order to show the solvability thanks to the Lax–Milgram theorem.

For this purpose we notice that, thanks to the specific form of the background profile τ given in (2.7), the homogeneous boundary conditions for θ^{n+1} , $(A^{-1}(\mathbf{k}\theta^{n+1}))_3$, $\nabla(A^{-1}(\mathbf{k}\theta^{n+1}))_3$, elliptic regularity (for the Stokes operator), and Poincaré inequality

(three times), there exists a constant c_1 independent of k, n, Ra , such that

$$(2.11) \quad Ra \left| \int_{\Omega} \tau'(z)(A^{-1}(\mathbf{k}\theta^{n+1}))_3 \theta^{n+1} \right| \leq c_1 \delta^2 Ra \|\nabla \theta^{n+1}\|^2 \leq \frac{1}{4} \|\nabla \theta^{n+1}\|^2$$

provided that we choose¹

$$(2.12) \quad \delta = (4c_1 Ra)^{-\frac{1}{2}}.$$

Therefore

$$(2.13) \quad B_n(\theta^{n+1}, \theta^{n+1}) \geq \frac{1}{k} \|\theta^{n+1}\|^2 + \frac{3}{4} \|\nabla \theta^{n+1}\|^2$$

which proves the coercivity. Here and elsewhere $\|\theta\| = \sqrt{\int_{\Omega} |\theta|^2}$ denotes the spatial L^2 norm of θ , and $\|\theta\|_{\infty} = \text{esssup}_{\Omega} |\theta|$ denotes the spatial L^{∞} norm of θ .

This ends the proof of the well-posedness of the discrete scheme.

2.4. Uniform dissipativity. Next, we prove the uniform dissipativity. Here and below, the c_j s denote generic constants independent of k, n (but which may depend on the Rayleigh number).

We first derive a uniform bound in the L^2 space. For this purpose we take the inner product of the scheme with $\psi = \theta^{n+1}$ and utilize the identity $(a - b, a) = \frac{1}{2}(|a|^2 - |b|^2 + |a - b|^2)$ together with the estimate on the destabilizing term (2.11), and we have

$$\begin{aligned} & \frac{1}{2k} (\|\theta^{n+1}\|^2 - \|\theta^n\|^2 + \|\theta^{n+1} - \theta^n\|^2) + \|\nabla \theta^{n+1}\|^2 \\ & \leq \|\tau'\| \|\nabla \theta^{n+1}\| + Ra \left| \int_{\Omega} \tau'(z)(A^{-1}(\mathbf{k}\theta^{n+1}))_3 \theta^{n+1} \right| \\ & \leq \|\tau'\|^2 + \frac{1}{2} \|\nabla \theta^{n+1}\|^2. \end{aligned}$$

Therefore, there exists a constant c_2 such that

$$(2.14) \quad \frac{1}{k} (\|\theta^{n+1}\|^2 - \|\theta^n\|^2 + \|\theta^{n+1} - \theta^n\|^2) + \|\nabla \theta^{n+1}\|^2 \leq 2\|\tau'\|^2 \leq c_2 Ra^{\frac{1}{2}}$$

which further implies, thanks to the Poincaré inequality, that

$$(2.15) \quad (1 + k) \|\theta^{n+1}\|^2 \leq \|\theta^n\|^2 + c_2 k Ra^{\frac{1}{2}}.$$

This leads to, with the help of a simple iteration,

$$(2.16) \quad \|\theta^{n+1}\|^2 \leq (1 + k)^{-(n+1)} \|\theta_0\|^2 + c_2 Ra^{\frac{1}{2}}.$$

This is a uniform estimate in the L^2 space.

A byproduct of this estimate is that

$$(2.17) \quad \frac{1}{N} \sum_{n=0}^N \|\nabla \theta^{n+1}\|^2 \leq \frac{\|\theta_0\|^2}{kN} + c_2 Ra^{\frac{1}{2}},$$

which is a bound on the Nusselt number in this discretized case for large N (see the definition later for Nusselt number (2.58, 2.59)), and a bound in $L^2(H^1)$ for the scheme.

¹The choice of τ or δ given here is not optimal. A near optimal choice would be $\delta \sim Ra^{-\frac{1}{3}}$, but the control on the linear destabilizing term is much longer [8, 12, 46]. We use the simple one since the optimal bound is not our goal here.

We also have

$$(2.18) \quad \sum_{n=0}^N \|\theta^{n+1} - \theta^n\|^2 \leq \|\theta_0\|^2 + c_2 k N R a^{\frac{1}{2}}.$$

In order to obtain uniform estimates in H^1 , we take the inner product of the scheme with $\psi = -\Delta\theta^{n+1}$, and we have

$$\begin{aligned} & \frac{1}{2k} (\|\nabla\theta^{n+1}\|^2 - \|\nabla\theta^n\|^2 + \|\nabla(\theta^{n+1} - \theta^n)\|^2) + \|\Delta\theta^{n+1}\|^2 \\ & \leq \|\tau''\| \|\Delta\theta^{n+1}\| + Ra \|\tau'\| \|A^{-1}(\mathbf{k}\theta^{n+1})\|_{\infty} \|\Delta\theta^{n+1}\| \\ & \quad + Ra \|A^{-1}(\mathbf{k}\theta^n)\|_{\infty} \|\nabla\theta^{n+1}\| \|\Delta\theta^{n+1}\| \\ & \leq \|\tau''\| \|\Delta\theta^{n+1}\| + c_3 Ra \|\tau'\| \|\theta^{n+1}\| \|\Delta\theta^{n+1}\| + c_4 Ra \|\theta^n\| \|\theta^{n+1}\|^{\frac{1}{2}} \|\Delta\theta^{n+1}\|^{\frac{3}{2}} \\ & \leq \frac{1}{2} \|\Delta\theta^{n+1}\|^2 + c_5, \end{aligned}$$

where we have applied the regularity result for the Stokes operator, the Sobolev imbedding of H^2 into L^{∞} , interpolation inequality, the uniform L^2 estimate (2.16), and Hölder type inequality.

This implies that

$$(2.19) \quad (1+k)\|\nabla\theta^{n+1}\|^2 \leq \|\nabla\theta^n\|^2 + 2c_6 k,$$

which further implies, with the help of a simple iteration, that

$$(2.20) \quad \|\nabla\theta^{n+1}\|^2 \leq (1+k)^{-n} \|\nabla\theta^1\|^2 + 2c_6.$$

This is the desired uniform estimates in the H^1 space; i.e., there is a uniform in k bounded absorbing ball in H^1 which attracts all solutions with L^2 initial data.

Uniform estimates in Sobolev spaces with more derivatives can be derived just as in the case of a continuous-in-time system. Here we demonstrate that the H^2 norm of the solution is asymptotically uniformly bounded in time; i.e., there is an absorbing ball in H^2 which attracts all solutions with L^2 initial data uniformly for all k .

For this purpose we apply Δ to both sides of the scheme (2.6) and then multiply the scheme by $\Delta\theta^{n+1}$ and integrate over the domain. This leads to the following:

$$\begin{aligned} & \frac{1}{2k} (\|\Delta\theta^{n+1}\|^2 - \|\Delta\theta^n\|^2 + \|\Delta(\theta^{n+1} - \theta^n)\|^2) + \|\nabla\Delta\theta^{n+1}\|^2 \\ & \leq \|\tau^{(4)}\| \|\Delta\theta^{n+1}\| + Ra (\|\Delta(A^{-1}(\mathbf{k}\theta^n))\|_{L^6} \|\nabla\theta^{n+1}\|_{L^3} \\ & \quad + 2\|\nabla A^{-1}(\mathbf{k}\theta^n)\|_{L^{\infty}} \|\nabla^2\theta^{n+1}\|) \|\Delta\theta^{n+1}\| \\ & \quad + Ra (\|\Delta(A^{-1}(\mathbf{k}\theta^{n+1}))\| \|\tau'\|_{L^{\infty}} + 2\|\nabla(A^{-1}(\mathbf{k}\theta^{n+1}))\| \|\nabla\tau'\|_{L^{\infty}} \\ & \quad + \|A^{-1}(\mathbf{k}\theta^{n+1})\| \|\Delta\tau'\|_{L^{\infty}}) \|\Delta\theta^{n+1}\| \\ & \leq c_7 (\|\Delta\theta^{n+1}\| + \|\Delta\theta^{n+1}\|^2) \\ & \leq c_8 (\|\Delta\theta^{n+1}\| + \|\nabla\Delta\theta^{n+1}\| \|\nabla\theta^{n+1}\|) \\ & \leq \frac{1}{2} \|\nabla\Delta\theta^{n+1}\|^2 + c_9, \end{aligned}$$

where we have applied the identity $\int A^{-1}(\mathbf{k}\theta^n)\nabla\Delta\theta^{n+1}\Delta\theta^{n+1} = 0$, Hölder’s inequality, elliptic regularity, Sobolev imbedding, Cauchy–Schwarz, and interpolation inequality.

This leads to the inequality

$$(2.21) \quad (1+k)\|\Delta\theta^{n+1}\|^2 \leq \|\Delta\theta^n\|^2 + 2c_9k,$$

which further implies, with the help of a simple iteration,

$$(2.22) \quad \|\Delta\theta^{n+1}\|^2 \leq (1+k)^{-n+1}\|\Delta\theta^2\|^2 + 2c_9.$$

This is the desired uniform estimates in the H^2 space, i.e., there is a uniform in k bounded absorbing ball in H^2 which attracts all solutions with L^2 initial data.

To summarize, we have the following lemma.

LEMMA 1 (uniform bound/dissipativity). *There exists a constant c_9 independent of the time step k such that the scheme (2.6) possesses an absorbing ball in H^1 and H^2 with radius $2\sqrt{c_9}$ which attracts all bounded sets in L^2 .*

2.5. Consistency and convergence. Here we check the consistency first since this is what we need in the following.

Multiplying the scheme (2.6) by $k(\theta^{n+1} - \theta^n)$ and integrating over the domain we have

$$\begin{aligned} \|\theta^{n+1} - \theta^n\|^2 &\leq k \left\{ -\frac{1}{2}(\|\nabla\theta^{n+1}\|^2 - \|\nabla\theta^n\|^2 + \|\nabla(\theta^{n+1} - \theta^n)\|^2) \right. \\ &\quad + \|\tau'\| \|\nabla(\theta^{n+1} - \theta^n)\| + c_{10}\|\theta^n\|_\infty \|\nabla\theta^{n+1}\| \|\theta^{n+1} - \theta^n\| \\ &\quad \left. + c_{11}\|\theta^{n+1}\|_\infty \|\tau'\| \|\theta^{n+1} - \theta^n\| \right\} \\ &\leq k(c_{12} + c_{13}\|\theta^{n+1} - \theta^n\|), \end{aligned}$$

where we have applied the Cauchy–Schwarz inequality, Hölder’s inequality, elliptic regularity, and uniform bounds in H^1 (2.20).

This implies the following *consistency result*:

$$(2.23) \quad \|\theta^{n+1} - \theta^n\| \leq c_{14}k^{\frac{1}{2}},$$

provided that $\theta_0 \in H_{0,per}^1$.

If we assume $\theta_0 \in H_{0,per}^1 \cap H^2$, we may deduce from the scheme (2.6) and the uniform bound (2.22) the following stronger consistency result:

$$(2.24) \quad \|\theta^{n+1} - \theta^n\| \leq c_{15}k.$$

Next, we show that the solutions to the scheme converge to the solution of the infinite Prandtl number model in $L^2(0, T^*, L^2)$ for any given time $T^* > 0$ as $k \rightarrow 0$.

For this purpose, we rewrite the scheme (2.6) as

$$(2.25) \quad \begin{aligned} \frac{\partial \tilde{\theta}_k(t)}{\partial t} + Ra A^{-1}(\mathbf{k}\theta_k(t)) \cdot \nabla\theta_k(t+k) + Ra A^{-1}(\mathbf{k}\theta_k(t+k))_3\tau'(z) \\ = \Delta\theta_k(t+k) + \tau''(z), \end{aligned}$$

where

$$(2.26) \quad \theta_k(t) = \theta_k^n, \quad t \in [nk, (n+1)k),$$

$$(2.27) \quad \tilde{\theta}_k(t) = \theta_k^n + \frac{t - nk}{k}(\theta_k^{n+1} - \theta_k^n), \quad t \in [nk, (n+1)k).$$

The estimates (2.16, 2.17) imply that θ_k and $\tilde{\theta}_k$ are uniformly (in k) bounded in $L^\infty(0, T^*; L^2)$ and $L^2(0, T^*; H_{0,per}^1)$. Hence we have a subsequence, still denoted θ_k and $\tilde{\theta}_k$ and $\theta, \tilde{\theta} \in L^\infty(0, T^*; L^2) \cap L^2(0, T^*; H_{0,per}^1)$ such that

$$(2.28) \quad \theta_k \rightharpoonup \theta, \quad \text{weak } * \text{ in } L^\infty(0, T^*; L^2),$$

$$(2.29) \quad \theta_k \rightharpoonup \theta, \quad \text{weak in } L^2(0, T^*; H_{0,per}^1),$$

$$(2.30) \quad \tilde{\theta}_k \rightharpoonup \tilde{\theta}, \quad \text{weak } * \text{ in } L^\infty(0, T^*; L^2),$$

$$(2.31) \quad \tilde{\theta}_k \rightharpoonup \tilde{\theta}, \quad \text{weak in } L^2(0, T^*; H_{0,per}^1).$$

It is also easy to check, thanks to (2.18), for any $a < T^*$,

$$(2.32) \quad \int_0^{T^*-a} \|\theta_k(t+k) - \theta_k(t)\|^2 dt \leq c_{16}k,$$

$$(2.33) \quad \int_0^{T^*} \|\theta_k(t) - \tilde{\theta}_k(t)\|^2 dt \leq c_{17}k.$$

Therefore

$$(2.34) \quad \theta = \tilde{\theta},$$

$$(2.35) \quad \theta_k(\cdot + k) \rightharpoonup \theta, \quad \text{weak } * \text{ in } L^\infty(0, T^*; L^2),$$

$$(2.36) \quad \theta_k(\cdot + k) \rightharpoonup \theta, \quad \text{weak in } L^2(0, T^*; H_{0,per}^1).$$

Furthermore, thanks to a compactness theorem due to Témam ([38, Ch. 13, Theorem 13.3]), which states that a bounded set $\mathcal{G} \subset L^1(0, T^*; Y) \cap L^p(0, T^*; X)$, $p > 1$ with X, Y being two Banach spaces and the injection of Y into X being compact, and $\sup_{g \in \mathcal{G}} \int_0^{T^*-a} \|g(a+s) - g(s)\|_X^p ds \rightarrow 0$, as $a \rightarrow 0$, is necessarily precompact in $L^q(0, T^*; X) \forall q \in [1, p)$, there exists a sub-subsequence of $\tilde{\theta}_k$ which converges strongly in $L^q(0, T^*; L^2) \forall q \in [1, p)$. Indeed, testing the scheme (2.25) against a test function v and integrating from t to $t+a$, we have

$$(2.37) \quad \begin{aligned} & |(\tilde{\theta}_k(t+a) - \tilde{\theta}_k(t), v)| \\ & \leq \int_t^{t+a} \{ \|\nabla \theta_k(s+k)\| \|\nabla v\| \\ & \quad + \|\tau'\| \left\| \frac{\partial v}{\partial z} \right\| + Ra \|A^{-1}(\mathbf{k}\theta_k(s))\|_{L^\infty} \|\theta_k(s+k)\| \|\nabla v\| \\ & \quad + Ra \|A^{-1}(\mathbf{k}\theta_k(s+k))\|_3 \| \tau' \| \|v\| \} ds \\ & \leq c_{18} \|\nabla v\| a^{\frac{1}{2}}, \end{aligned}$$

where we have applied the regularity for the Stokes operator, Sobolev imbedding, Poincaré inequality, and the a priori estimates in $L^\infty(L^2)$ and $L^2(H^1)$ valid for L^2 initial data. Now set $v = \tilde{\theta}_k(t+a) - \tilde{\theta}_k(t)$ and utilize the $L^2(H^1)$ estimate on $\tilde{\theta}_k$. Then we have

$$(2.38) \quad \int_0^{T^*-a} \|\tilde{\theta}_k(t+a) - \tilde{\theta}_k(t)\|^2 \leq c_{19} a^{\frac{1}{2}}.$$

This implies the strong convergence by Témam’s compactness theorem.

Combining this strong convergence in $L^q(L^2)$, $q \in [1, 2)$ with the uniform $L^\infty(0, T^*; L^2)$ estimate, we conclude that the sub-subsequence in fact converges strongly in $L^q(0, T^*; L^2) \forall q \in [1, \infty)$. Hence we may summarize the a priori estimates as

$$(2.39) \quad \theta_k(\cdot), \theta_k(\cdot + k), \tilde{\theta}_k(\cdot), \tilde{\theta}_k(\cdot + k) \rightarrow \theta(\cdot), \text{ in } L^q(0, T^*; L^2) \forall q \in [1, \infty),$$

$$(2.40) \quad \theta_k(\cdot), \theta_k(\cdot + k), \tilde{\theta}_k(\cdot), \tilde{\theta}_k(\cdot + k) \rightharpoonup \theta(\cdot), \text{ weakly in } L^2(0, T^*; H_{0,per}^1).$$

Now for any $\phi \in H_{0,per}^1$ and $\psi \in C^1([0, T^*])$ with $\psi(T^*) = 0$, we can rewrite the scheme (2.25) in the following weak form:

$$(2.41) \quad \int_0^{T^*} \int_\Omega \left\{ -\tilde{\theta}_k(\mathbf{x}, t) \phi(\mathbf{x}) \psi'(t) + RaA^{-1}(\mathbf{k}\theta_k(\mathbf{x}, t)) \cdot \nabla \theta_k(\mathbf{x}, t+k) \phi(\mathbf{x}) \psi(t) \right. \\ \left. + RaA^{-1}(\mathbf{k}\theta_k(\mathbf{x}, t+k))_3 \tau'(z) \phi(\mathbf{x}) \psi(t) + \nabla \theta_k(\mathbf{x}, t+k) \cdot \nabla \phi(\mathbf{x}) \psi(t) \right. \\ \left. + \tau'(z) \frac{\partial}{\partial z} \phi(\mathbf{x}) \psi(t) \right\} d\mathbf{x} dt \\ = \int_\Omega \theta_0(\mathbf{x}) \phi(\mathbf{x}) \psi(0) d\mathbf{x}.$$

Utilizing the strong $L^q(L^2)$ convergence (2.39) and the weak $L^2(H_{0,per}^1)$ convergence (2.40) together with elliptic regularity, we can pass to the limit as $k \rightarrow 0$ and arrive at

$$(2.42) \quad \int_0^{T^*} \int_\Omega \left\{ -\theta \phi \psi' + RaA^{-1}(\mathbf{k}\theta) \cdot \nabla \theta \phi \psi + RaA^{-1}(\mathbf{k}\theta)_3 \tau' \phi \psi \right. \\ \left. + \nabla \theta \cdot \nabla \phi \psi + \tau' \frac{\partial}{\partial z} \phi \psi \right\} d\mathbf{x} dt = \int_\Omega \theta_0 \phi \psi(0),$$

which is exactly the weak form of the infinite Prandtl number model. Since the infinite Prandtl number model possesses a unique solution, θ must be the unique solution; hence, the whole sequence of θ_k and $\tilde{\theta}_k$ converges to θ as any subsequence has a sub-subsequence that converges to the same limit θ .

We summarize the result here as the following lemma.

LEMMA 2 (consistency and convergence). *For any given $T^* > 0$ and $\theta_0 \in L^2$, the solution to the numerical scheme (2.25) converges to the solution of the infinite Prandtl number model; i.e.,*

$$(2.43) \quad \theta_k(\cdot), \theta_k(\cdot + k), \tilde{\theta}_k(\cdot), \tilde{\theta}_k(\cdot + k) \rightarrow \theta(\cdot) \text{ in } L^q(0, T^*; L^2) \forall q \in [1, \infty),$$

$$(2.44) \quad \theta_k(\cdot), \theta_k(\cdot + k), \tilde{\theta}_k(\cdot), \tilde{\theta}_k(\cdot + k) \rightharpoonup \theta(\cdot) \text{ weakly in } L^2(0, T^*; H_{0,per}^1),$$

where θ is the unique solution to the infinite Prandtl number model.

Moreover, if $\theta_0 \in H_{0,per}^1 \cap H^2$, then there exists a generic constant c_{15} independent of k, n such that

$$(2.45) \quad \|\theta^{n+1} - \theta^n\| \leq c_{15}k.$$

We have established $L^q(L^2) \forall q < \infty$ convergence of the numerical scheme. Uniform in time convergence (on finite time interval, i.e., $L^\infty(0, T^*; L^2)$) of the scheme can be established as well if we assume all the compatibility conditions needed (so that the exact solution is smooth enough up to the initial time $t = 0$; see [38] for the case of Navier–Stokes equations). If no high order compatibility condition is assumed, then one can show the uniform in time convergence on any finite interval modulus an initial layer (see [20] for the case of Navier–Stokes equations). Uniform convergence without enough compatibility conditions is needed for the proof of convergence of the global attractors [48], but not required here and hence we skip the details.

2.6. Convergence of the stationary statistical properties. As we mentioned earlier, for complex systems with chaotic/turbulent behavior, statistical properties are much more important than individual trajectories. In fact it is essentially hopeless to try to find approximation schemes that possess the property that the approximate trajectory remain close to the “true” trajectory for all time due to abundant sensitive dependence on data and positive Lyapunov exponents.² Therefore, the natural question to ask is if stationary statistical properties are well approximated. These stationary statistical properties are characterized by stationary statistical solutions or invariant measures of the system. Hence the question that we ask here is if the invariant measures of the discrete time approximation approximate the invariant measures of the continuous-in-time infinite Prandtl number model.

We first observe that the numerical scheme (2.6) can be viewed as a *discrete time dynamical system* on the phase space L^2 with the notation

$$(2.46) \quad \theta^{n+1} = F_k(\theta^n).$$

Thanks to the well-posedness result, we see that F_k in fact maps L^2 into $H_{0,per}^1$, and F_k^2 maps L^2 into $H_{0,per}^1 \cap H^2$ by elliptic regularity. Moreover, the discrete dynamical system is uniformly (in k) dissipative thanks to the uniform H^1 estimate (2.20). Therefore, this dynamical system possesses a compact global attractor in H^1 which attracts all bounded sets in L^2 . This leads to the existence of invariant measures via a classical Krylov–Bogliubov argument [42, 15] for the numerical scheme (the discrete dynamical system).

We recall the definition of invariant measures.

DEFINITION 1 (invariant measures). *A Borel probability measure μ_k on L^2 is called an invariant measure for F_k if*

$$(2.47) \quad \int_{L^2} \Phi(F_k(\theta)) d\mu_k = \int_{L^2} \Phi(\theta) d\mu_k$$

for all bounded continuous test functional Φ .

The set of all invariant measures for F_k is denoted \mathcal{IM}_k .

²There is a notable exception when the system possesses an explicit hyperbolic structure for which numerical shadowing may be possible [35].

We also recall that a Borel probability measure μ on L^2 is an invariant measure, or stationary statistical solution, for the infinite Prandtl number model for convection if

1.

$$(2.48) \quad \int_{L^2} \|\nabla\theta\|^2 d\mu(\theta) < \infty,$$

2.

$$(2.49) \quad \int_{L^2} < -Ra A^{-1}(\mathbf{k}\theta) \cdot \nabla\theta - Ra A^{-1}(\mathbf{k}\theta)_3 \tau'(z) + \Delta\theta + \tau''(z), \Phi'(\theta) > d\mu(\theta) = 0$$

for any cylindrical test functional $\Phi(\theta) = \phi((\theta, w_1), \dots, (\theta, w_m))$, where ϕ is a C^1 function on R^m , $\{w_j, j \geq 1\}$ are the eigenfunctions of Δ which form an orthonormal basis for L^2 and $w_j \in H_{0,per}^1 \cap C^2 \forall j$, and $<, >$ denotes the $H^{-1}, H_{0,per}^1$ duality,

3.

$$(2.50) \quad \int_{L^2} \int_{\Omega} \{|\nabla\theta|^2 + Ra(A^{-1}(\mathbf{k}\theta))_3 \theta \tau' - \tau''\theta\} d\mathbf{x} d\mu(\theta) \leq 0.$$

The set of all stationary statistical solutions for the infinite Prandtl number model is denoted \mathcal{IM} .

Roughly speaking, the first condition says that the invariant measures are supported on the smaller and finer space of H^1 , the second condition is the differential form of the weak formulation of the invariance of the measure under the flow, and the third condition is a statistical version of the energy inequality.

Now let $\mu_k \in \mathcal{IM}_k$ be a sequence of invariant measures. Thanks to the uniform estimate in H^1 (2.20), we see that the support of μ_k is contained in a bounded ball in H^1 independent of k . Therefore, thanks to the Prokhorov compactness theorem and Rellich compactness theorem [3, 30], the sequence μ_k is weakly precompact in the set of all Borel probability measures on L^2 ; hence it must contain a weakly convergent subsequence (still denoted $\{\mu_k\}$) which converges to a Borel probability measure μ . Our goal is to show that μ must be an invariant measure of the infinite Prandtl number model.

The first condition in the definition is easily verified since the global attractors for the discrete dynamical systems are uniformly bounded in H^1 independent of the time step k , and the invariant measures are supported on the global attractor [15, 48].

In order to check the second condition, i.e., the differential form of the weak formulation of invariance, we let $\Phi(\theta) = \phi((\theta, w_1), \dots, (\theta, w_m)) = \phi(y_1, \dots, y_m)$ be a cylindrical test functional. Notice that

$$(2.51) \quad \Phi'(\theta) = \sum_{j=1}^m \frac{\partial}{\partial y_j} \phi((\theta, w_1), \dots, (\theta, w_m)) w_j.$$

Hence, denoting by $<, >$ the duality between H^{-1} and $H_{0,per}^1$, we have

$$\begin{aligned} & \int_{L^2} < -Ra A^{-1}(\mathbf{k}\theta) \cdot \nabla\theta - Ra A^{-1}(\mathbf{k}\theta)_3 \tau'(z) + \Delta\theta + \tau''(z), \Phi'(\theta) > d\mu(\theta) \\ &= \int_{L^2} < -Ra A^{-1}(\mathbf{k}\theta) \cdot \nabla\theta - Ra A^{-1}(\mathbf{k}\theta)_3 \tau'(z) \end{aligned}$$

$$\begin{aligned}
& +\Delta\theta + \tau''(z), \sum_{j=1}^m \frac{\partial\phi}{\partial y_j}((\theta, w_1), \dots, (\theta, w_m))w_j > d\mu(\theta) \\
& = \int_{L^2} \sum_{j=1}^m \frac{\partial\phi}{\partial y_j} \int_{\Omega} (Ra A^{-1}(\mathbf{k}\theta) \cdot \nabla w_j \theta - Ra A^{-1}(\mathbf{k}\theta)_3 \tau'(z)w_j \\
& \quad + \Delta w_j \theta + \tau''(z)w_j) d\mathbf{x} d\mu(\theta) \\
& = \lim_{k \rightarrow 0} \int_{L^2} \sum_{j=1}^m \frac{\partial\phi}{\partial y_j} \int_{\Omega} (Ra A^{-1}(\mathbf{k}\theta) \cdot \nabla w_j \theta - Ra A^{-1}(\mathbf{k}\theta)_3 \tau'(z)w_j \\
& \quad + \Delta w_j \theta + \tau''(z)w_j) d\mathbf{x} d\mu_k(\theta) \\
& = \lim_{k \rightarrow 0} \int_{L^2} \sum_{j=1}^m \frac{\partial\phi}{\partial y_j} \int_{\Omega} (Ra A^{-1}(\mathbf{k}\theta) \cdot \nabla w_j F_k(\theta) - Ra A^{-1}(\mathbf{k}F_k(\theta))_3 \tau'(z)w_j \\
& \quad + \Delta w_j F_k(\theta) + \tau''(z)w_j) d\mathbf{x} d\mu_k(\theta) \\
& = \lim_{k \rightarrow 0} \int_{L^2} \langle -Ra A^{-1}(\mathbf{k}\theta) \cdot \nabla F_k(\theta) - Ra A^{-1}(\mathbf{k}F_k(\theta))_3 \tau'(z) \\
& \quad + \Delta F_k(\theta) + \tau''(z), \Phi'(\theta) \rangle d\mu_k(\theta) \\
& = \lim_{k \rightarrow 0} \int_{L^2} \langle \frac{F_k(\theta) - \theta}{k}, \Phi'(\theta) \rangle d\mu_k(\theta) \\
& = \lim_{k \rightarrow 0} \int_{L^2} \frac{1}{k} (\Phi(F_k(\theta)) - \Phi(\theta)) d\mu_k(\theta) \\
& = 0,
\end{aligned}$$

where we have used the boundedness and continuity of $\frac{\partial\phi}{\partial y_j}$ on the union of the support of μ_k , the consistency estimate (2.24), the invariance of μ_k under F_k , and the following straightforward estimates valid on the union of the support of the μ_k s (uniformly bounded in $H_{0,per}^1 \cap H^2$):

$$\begin{aligned}
& \left| \int_{\Omega} A^{-1}(\mathbf{k}\theta) \cdot \nabla w_j (F_k(\theta) - \theta) d\mathbf{x} \right| \leq c \|F_k(\theta) - \theta\| = \mathcal{O}(k) \rightarrow 0, \\
& \left| \int_{\Omega} (A^{-1}(\mathbf{k}(F_k(\theta) - \theta)))_3 \tau'(z)w_j d\mathbf{x} \right| \leq c \|F_k(\theta) - \theta\| = \mathcal{O}(k) \rightarrow 0, \\
& \left| \int_{\Omega} \Delta w_j (F_k(\theta) - \theta) d\mathbf{x} \right| \leq c \|F_k(\theta) - \theta\| = \mathcal{O}(k) \rightarrow 0, \\
& \langle F_k(\theta) - \theta, \Phi'(\theta) \rangle = \Phi(F_k(\theta)) - \Phi(\theta) + \mathbf{o}(\|F_k(\theta) - \theta\|) \\
& \quad = \Phi(F_k(\theta)) - \Phi(\theta) + \mathbf{o}(k),
\end{aligned}$$

the weak convergence of μ_k to μ , the scheme, and the invariance of μ_k .

This proves the differential form of the weak invariance of μ under the infinite Prandtl number dynamics, i.e., 2 of Definition 1.

The energy type inequality 3 of Definition 1 can be verified easily as well. For this purpose, we first show that any invariant measure μ_k of the numerical scheme (2.6) must satisfy the same energy type estimate. The desired continuous one will be the limit as the time step approaches zero.

We first show that the invariant measures for F_k also satisfy the energy inequality. For this purpose we multiply the scheme (2.6) by θ^{n+1} and integrate over the domain, and we then have

$$\begin{aligned} & \frac{1}{2k}(\|\theta^{n+1}\|^2 - \|\theta^n\|^2 + \|\theta^{n+1} - \theta^n\|^2) + \|\nabla\theta^{n+1}\|^2 \\ & + \int_{\Omega} (-\tau''\theta^{n+1} + Ra \tau' A^{-1}(\mathbf{k}\theta^{n+1})_3\theta^{n+1}) = 0. \end{aligned}$$

This can be rewritten using the discrete dynamical system notation F_k as

$$(2.52) \quad \begin{aligned} & \frac{1}{2k}(\|F_k(\theta)\|^2 - \|\theta\|^2 + \|F_k(\theta) - \theta\|^2) + \|\nabla F_k(\theta)\|^2 \\ & + \int_{\Omega} (-\tau''F_k(\theta) + Ra \tau' A^{-1}(\mathbf{k}F_k(\theta))_3F_k(\theta)) = 0. \end{aligned}$$

Integrating this identity with respect to the invariant measure μ_k and utilizing the invariance of μ_k under F_k , we have

$$(2.53) \quad \begin{aligned} & \int_{L^2} \int_{\Omega} (|\nabla F_k(\theta)|^2 - \tau''F_k(\theta) + Ra \tau' A^{-1}(\mathbf{k}F_k(\theta))_3F_k(\theta)) d\mathbf{x} d\mu_k(\theta) \\ & = -\frac{1}{2k} \int_{L^2} \|F_k(\theta) - \theta\|^2 d\mu_k(\theta) \leq 0. \end{aligned}$$

Utilizing the invariance of μ_k under F_k again in the lower order terms we have

$$(2.54) \quad \int_{L^2} (\|\nabla F_k(\theta)\|^2 + \int_{\Omega} (-\tau''\theta + Ra \tau' (A^{-1}(\mathbf{k}\theta))_3\theta) d\mathbf{x}) d\mu_k(\theta) \leq 0.$$

Now we recall that the support of μ_k is uniformly bounded in H^1 and hence, since the w_j s are the eigenfunctions of Δ which form an orthonormal basis in L^2 , w_j also forms a complete orthogonal system in $H^1_{0,per}$ (with the inner product between f and g given by $\int_{\Omega} \nabla f \cdot \nabla g$),

$$\|\nabla F_k(\theta)\|^2 = \lim_{m \rightarrow \infty} \sum_{j=1}^m \frac{(\nabla F_k(\theta), \nabla w_j)^2}{\|\nabla w_j\|^2} = \lim_{m \rightarrow \infty} \sum_{j=1}^m \frac{(F_k(\theta), \Delta w_j)^2}{\|\nabla w_j\|^2}.$$

Therefore

$$(2.55) \quad \begin{aligned} \int_{L^2} \|\nabla F_k(\theta)\|^2 d\mu_k(\theta) &= \int_{L^2} \lim_{m \rightarrow \infty} \sum_{j=1}^m \frac{(F_k(\theta), \Delta w_j)^2}{\|\nabla w_j\|^2} d\mu_k(\theta) \\ &= \lim_{m \rightarrow \infty} \int_{L^2} \sum_{j=1}^m \frac{(F_k(\theta), \Delta w_j)^2}{\|\nabla w_j\|^2} d\mu_k(\theta) \\ &= \lim_{m \rightarrow \infty} \int_{L^2} \sum_{j=1}^m \frac{(\theta, \Delta w_j)^2}{\|\nabla w_j\|^2} d\mu_k(\theta) \\ &= \int_{L^2} \|\nabla\theta\|^2 d\mu_k(\theta), \end{aligned}$$

where we have used the uniform boundedness of $\|\nabla\theta\|$ and $\|\nabla F_k(\theta)\|$ on the support of μ_k , the Lebesgue dominated convergence theorem and the invariance of μ_k under F_k .

Hence we see that μ_k also satisfies the energy inequality, i.e.,

$$(2.56) \quad \int_{L^2} \int_{\Omega} (|\nabla\theta|^2 - \tau''\theta + Ra \tau'(A^{-1}(\mathbf{k}\theta))_3\theta) d\mu_k(\theta) \leq 0.$$

Next, we take the limit as k approaches zero. The last two terms in the discrete energy inequality above converge to the right limit by the very definition of weak convergence of μ_k to μ . As for the leading order quadratic term, we have

$$\begin{aligned} \int \|\nabla\theta\|^2 d\mu(\theta) &= \lim_{m \rightarrow \infty} \sum_{j=1}^m \int \frac{(\theta, \Delta w_j)^2}{\|\nabla w_j\|^2} d\mu(\theta) \\ &= \lim_{m \rightarrow \infty} \lim_{k \rightarrow 0} \sum_{j=1}^m \int \frac{(\theta, \Delta w_j)^2}{\|\nabla w_j\|^2} d\mu_k(\theta) \\ &\leq \liminf_{k \rightarrow 0} \sum_{j=1}^{\infty} \int \frac{(\theta, \Delta w_j)^2}{\|\nabla w_j\|^2} d\mu_k(\theta) \\ &= \liminf_{k \rightarrow 0} \int \|\nabla\theta\|^2 d\mu_k(\theta). \end{aligned}$$

This implies that

$$\begin{aligned} &\int_{L^2} \int_{\Omega} (|\nabla\theta|^2 - \tau''\theta + Ra \tau'(A^{-1}(\mathbf{k}\theta))_3\theta) d\mu(\theta) \\ &\leq \liminf_{k \rightarrow 0} \int \|\nabla\theta\|^2 d\mu_k(\theta) + \lim_{k \rightarrow 0} \int_{L^2} \int_{\Omega} (-\tau''\theta + Ra \tau'(A^{-1}(\mathbf{k}\theta))_3\theta) d\mu_k(\theta) \\ &\leq \liminf_{k \rightarrow 0} \int_{L^2} \int_{\Omega} (|\nabla\theta|^2 - \tau''\theta + Ra \tau'(A^{-1}(\mathbf{k}\theta))_3\theta) d\mu_k(\theta) \\ (2.57) &\leq 0. \end{aligned}$$

This completes the proof of the energy type inequality (3 in Definition 1) for the limit probability measure μ . Therefore we conclude that the limit μ must be an invariant measure of the infinite Prandtl number model.

Sometimes we impose a stronger version of the statistical energy inequality in the definition of stationary statistical solutions: we require that the statistical version of the energy inequality be true on any energy shells $e_1 \leq \|\theta\| \leq e_2$ instead of just one infinite shell from zero to infinity. Such a kind of energy inequalities is useful in some applications in the Navier–Stokes case (see, for instance, [15]). They can be verified with a little bit of extra work which involves approximating the finite difference by differentiation and utilizing the uniform in H^2 estimates (invariant measures are supported in a bounded ball in H^2). We shall supply details elsewhere. Likewise it is sometimes useful to have a stronger version of the invariance of μ under the continuous-in-time dynamics either as a straightforward pullback invariance or differential form of the weak invariance with a broader class of test functionals

that are bounded on bounded sets of L^2 , Fréchet differentiable for $\theta \in H^1_{0,per}$ with $\Phi'(\theta) \in H^1_{0,per}$, and the derivative is continuous and bounded as a function from $H^1_{0,per}$ to $H^1_{0,per}$. It can be shown that these variations yield the same definition just as in the case of two-dimensional Navier–Stokes equations [15].

Next, we turn our attention to one of the most important statistical quantities in convection: the heat transport in the vertical direction quantified by the Nusselt number. More specifically, we consider the limit of heat transport in the vertical direction; i.e., the Nusselt number, as the step size approaches zero. We first recall the definition of the *Nusselt number*.

DEFINITION 2 (Nusselt number). *For the infinite Prandtl number model, the nondimensional averaged heat transport in the vertical direction is defined as*

$$\begin{aligned}
 Nu &= \sup_{\theta_0 \in L^2} \limsup_{t \rightarrow \infty} \frac{1}{tL_xL_y} \int_0^t \int_{\Omega} |\nabla T(\mathbf{x}, s)|^2 \, d\mathbf{x}ds \\
 &= 1 + Ra \sup_{\theta_0 \in L^2} \limsup_{t \rightarrow \infty} \frac{1}{tL_xL_y} \int_0^t \int_{\Omega} A^{-1}(\mathbf{k}T(\mathbf{x}, s))_3 T(\mathbf{x}, s) \, d\mathbf{x}ds \\
 (2.58) \quad &= 1 + Ra \sup_{\theta_0 \in L^2} \limsup_{t \rightarrow \infty} \frac{1}{tL_xL_y} \int_0^t \int_{\Omega} A^{-1}(\mathbf{k}\theta(\mathbf{x}, s))_3 \theta(\mathbf{x}, s) \, d\mathbf{x}ds.
 \end{aligned}$$

Likewise, the nondimensional averaged heat transport in the vertical direction for the discrete in time scheme (2.6) is defined as

$$(2.59) \quad Nu_k = 1 + Ra \sup_{\theta_0 \in L^2} \limsup_{N \rightarrow \infty} \frac{1}{NL_xL_y} \sum_{n=1}^N \int_{\Omega} A^{-1}(\mathbf{k}\theta^n(\mathbf{x}))_3 \theta^n(\mathbf{x}) \, d\mathbf{x}.$$

It is well known that long time averages defined through Banach (generalized) limits are spatial averages with respect to appropriate invariant measures of the underlying dynamical system [2, 15, 42, 43, 47]. Moreover, for a given continuous test functional φ_0 (in the application here $\varphi_0(\theta) = 1 + \frac{Ra}{L_xL_y} \int_{\Omega} A^{-1}(\mathbf{k}\theta(\mathbf{x}))_3 \theta(\mathbf{x}) \, d\mathbf{x}$), and a particular trajectory (initial data), there exists a particular Banach limit, LIM_0 , so that $LIM_0 \frac{1}{T} \int_0^t \varphi_0(\theta(s)) \, ds = \limsup \frac{1}{T} \int_0^t \varphi_0(\theta(s)) \, ds$ [46, 47]. Therefore, when combined with the Prokhorov’s compactness theorem, we deduce the existence of an invariant measure $\mu_k \in \mathcal{IM}_k$ such that

$$(2.60) \quad Nu_k = 1 + \frac{Ra}{L_xL_y} \int_{L^2} \int_{\Omega} A^{-1}(\mathbf{k}\theta(\mathbf{x}))_3 \theta(\mathbf{x}) \, d\mathbf{x}d\mu_k.$$

Hence by the weak convergence result that we just proved, we see that for any sequence of Nu_k (and hence μ_k) there exists a subsequence (still denoted Nu_k and μ_k) and $\mu \in \mathcal{IM}$ such that

$$\begin{aligned}
 \lim_{k \rightarrow 0} Nu_k &= 1 + \frac{Ra}{L_xL_y} \lim_{k \rightarrow 0} \int_{L^2} \int_{\Omega} A^{-1}(\mathbf{k}\theta(\mathbf{x}))_3 \theta(\mathbf{x}) \, d\mathbf{x}d\mu_k \\
 &= 1 + \frac{Ra}{L_xL_y} \int_{L^2} \int_{\Omega} A^{-1}(\mathbf{k}\theta(\mathbf{x}))_3 \theta(\mathbf{x}) \, d\mathbf{x}d\mu \\
 &\leq 1 + \frac{Ra}{L_xL_y} \sup_{\nu \in \mathcal{IM}} \int_{L^2} \int_{\Omega} A^{-1}(\mathbf{k}\theta(\mathbf{x}))_3 \theta(\mathbf{x}) \, d\mathbf{x}d\nu \\
 (2.61) \quad &= 1 + \frac{Ra}{L_xL_y} \sup_{\theta_0 \in L^2} \lim_{t \rightarrow \infty} \frac{1}{tL_xL_y} \int_0^t \int_{\Omega} A^{-1}(\mathbf{k}\theta(\mathbf{x}, s))_3 \theta(\mathbf{x}, s) \, d\mathbf{x}ds,
 \end{aligned}$$

where we have used the weak convergence of μ_k to μ , the compactness of the set of all invariant measures due to Prokhorov's theorem and the a priori estimates, and the fact that extremal points of the set of invariant measures are ergodic (in the sense that phase space spatial average and time average are the same) [2, 43, 46].

To summarize, we have proved the following main result.

THEOREM 1 (convergence of stationary statistical properties). *Let μ_k be an arbitrary invariant measure of the numerical scheme (2.6) with time step k , i.e., $\mu_k \in \mathcal{IM}_k$, and let Nu_k be the Nusselt number characterizing the heat transport in the vertical direction for the scheme with time step k defined in (2.59). Then each subsequence of μ_k must contain a sub-subsequence (still denoted $\{\mu_k\}$) and an invariant measure μ of the infinite Prandtl number model so that μ_k weakly converges to μ ; i.e.,*

$$(2.62) \quad \mu_k \rightharpoonup \mu, k \rightarrow 0.$$

Moreover, the Nusselt number converges in an upper semicontinuous fashion in the sense that

$$(2.63) \quad \limsup_{k \rightarrow 0} Nu_k \leq Nu.$$

In particular, this implies that the convergent numerical schemes will not overestimate the Nusselt number asymptotically.

Notice that our asymptotic lower bound on the Nusselt number for the infinite Prandtl number model nicely complements the rigorous upper bound for the Nusselt number using a variational approach proposed by Constantin and Doering [7, 8].

3. Conclusions and remarks. Our main result clearly demonstrated the usefulness of the uniformly dissipative scheme that we proposed in terms of approximating stationary statistical properties of the infinite Prandtl number model for convection since the stationary statistical properties of the scheme converge to those of the continuous time model. To the best of our knowledge, this is the first rigorous result proving convergence of stationary statistical properties of numerical schemes to those of the continuous-in-time dynamical system under approximation. Our result may be viewed as a partial generalization of Lax's equivalence theorem in the sense that consistency and long time stability (uniform dissipativity) imply convergence of stationary statistical properties. We would like to emphasize that the methodology here can be applied to much more general dissipative systems (with chaotic behavior for relevance) although we have treated the infinite Prandtl number model only [48]. We hope that our work will stimulate further study on numerical schemes for approximating statistical properties of dissipative dynamical systems.

The convergence of the stationary statistical properties relies on the uniform bound in a space which is compactly embedded in the phase space (it is $H_{0,per}^1$ in the infinite Prandtl number case which is compactly imbedded in the phase space L^2 by Rellich's theorem). Simply having a uniform bound in the phase space may not imply the convergence of the statistical properties. (We could construct schemes that possess an absorbing ball in L^2 but not in H^1 .) Therefore, we would rather use the *uniformly dissipative* terminology instead of the *global in time stability* used by many other authors, which could mean uniform boundedness in the phase space only.

The convergence that we derived here is actually semiconvergence since different subsequences may converge to different invariant measures of the continuous-in-time dynamical system. There is no convergence rate either. This is perhaps the generic

picture in the sense that the result here is nearly optimal without additional assumption on the continuous dynamical system. One very useful physical assumption is the mixing of the continuous system. Indeed, if we assume that the continuous system is exponentially mixing with a rate of r [43], i.e., there is a physically relevant invariant measure μ so that

$$(3.1) \quad \left| \frac{1}{t} \int_0^t \Phi(\theta(\mathbf{x}, s)) ds - \int \Phi(\theta(\mathbf{x})) d\mu \right| \leq c \exp(-rt)$$

for all appropriate test functionals $\Phi(\theta)$ and almost all trajectories, then approximating a specific statistical quantity $\int \Phi(\theta(\mathbf{x})) d\mu$ becomes a finite time integration problem. Indeed, supposing the given tolerance level is 2ϵ , we first fix a time t so that $c \exp(-rt) \leq \epsilon$ (the time t is usually large for small mixing rate r). We then adjust our mesh size (small time step or mesh size) so that

$$\left| \frac{1}{t} \int_0^t \Phi(\theta(\mathbf{x}, s)) ds - \frac{1}{N} \sum_{n=1}^N \Phi(\theta^n) \right| \leq \epsilon,$$

where $Nk = t$ with k being the time step. Hence the infinite time approximation of a stationary statistical property becomes the problem of approximation on finite time interval $[0, t]$ for appropriate numerical schemes (say uniformly dissipative). This motivates us to work on higher order schemes so that the integration on $[0, t]$ can be calculated quickly.

Of course we do not have exponential mixing for generic dissipative complex/chaotic dynamical systems. One way to circumvent this difficulty is by considering noisy systems since our environment is intrinsically noisy. Exponential mixing can be verified for many dissipative systems with appropriate additive white noise [11, 49, 32, 29, 50]. Hence there is a strong incentive to generalize the notion of uniformly dissipative schemes to approximations of continuous-in-time stochastic dynamical systems (both SDE and SPDE; see [32] for the case of SDE and fully implicit approach). We will report results in this direction at another time.

The scheme that we presented here is not the only scheme that is able to capture stationary statistical properties of the underlying continuous system. The fully implicit backward Euler scheme is a uniformly dissipative scheme as one can readily verify. However, the backward Euler is nonlinear in the unknown and hence the computational cost at each time step is expected to be higher. There are other linear implicit uniformly dissipative schemes. For instance one may check that the following family of schemes is uniformly dissipative for $\lambda \in [0, 1]$:

$$(3.2) \quad \frac{\theta^{n+1} - \theta^n}{k} + Ra A^{-1}(\mathbf{k}\theta^n) \cdot \nabla \theta^{n+1} + Ra (A^{-1}(\mathbf{k}(\lambda\theta^{n+1} + (1 - \lambda)\theta^n)))_3 \tau'(z) = \Delta \theta^{n+1} + \tau''(z).$$

However, it seems that this family of schemes is uniformly dissipative under small time step restriction $k \leq \frac{1}{Ra^2}$. This kind of restriction may be expected since the linearly unstable modes grow as the Rayleigh number grows; hence the time step should reflect this through CFL condition. On the other hand, we observe that for the case of $\lambda = 0$, the scheme is the same as

$$\frac{T^{n+1} - T^n}{k} + Ra A^{-1}(\mathbf{k}T^n) \cdot \nabla T^{n+1} = \Delta T^{n+1},$$

whose inviscid part is stable (satisfies maximum principle). Hence if the viscous scheme is unstable, then the viscous term plays a destabilizing role. We also notice that the case with $\lambda = 0$ corresponds to discretization in time first followed by translation, as we mentioned in section 2.2.

It is also worthwhile to point out that at the time discretization only stage we should anticipate an implicit scheme due to the CFL condition.

An issue that we have not addressed here is spatial discretization. Since we have utilized the background temperature profile τ in our uniform dissipativity argument, it is expected that we need to resolve small scales within the background profile. A similar issue for the Navier–Stokes was investigated earlier [9]. We will report the details at another time.

Another issue that we have not addressed here is the behavior of the global attractors. We fully anticipate an upper semicontinuity result. The proof is a modification of the classical one [39, 34, 19] since we will not have uniform in time convergence of trajectories on finite time interval. In fact, we will have uniform in time convergence after a transitional period of time due to the fact that the points on the global attractors of the scheme may not satisfy high order compatibility condition for the infinite Prandtl number model and hence the solution may have an initial transitional layer (see [20] for the case of two-dimensional Navier–Stokes equations). We will report this at another time as well.

Acknowledgment. We acknowledge helpful conversations with Brian Ewald and Andy Majda.

REFERENCES

- [1] G. AMATI, K. KOAL, F. MASSAIOLI, K. R. SREENIVASAN, AND R. VERZICCO, *Turbulent thermal convection at large Rayleigh numbers for a Boussinesq fluid of constant Prandtl number*, Physics of Fluids, 17 (2005), p. 121701.
- [2] M. B. BEKKA AND M. MAYER, *Ergodic Theory and Topological Dynamics of Group Actions on Homogeneous Spaces*, London Mathematical Society Lecture Note Series 269, Cambridge University Press, Cambridge, UK, 2000.
- [3] P. BILLINGSLEY, *Weak Convergence of Measures: Applications in Probability*, SIAM, Philadelphia, 1971.
- [4] E. BODENSCHATZ, W. PESCH, AND G. AHLERS, *Recent developments in Rayleigh–Bénard convection*, Annu. Rev. Fluid Mech. 32, Palo Alto, CA, 2000, pp. 709–778.
- [5] E. CALZAVARINI, C. R. DOERING, J. D. GIBBON, D. LOHSE, A. TANABE, AND F. TOSCHI, *Exponentially growing solutions in homogeneous Rayleigh–Bénard convection*, Phys. Rev. E. (3), 73 (2006), p. 035301.
- [6] S. CHANDRASEKHAR, *Hydrodynamic and Hydromagnetic Stability*, Clarendon Press, Oxford, 1961.
- [7] P. CONSTANTIN AND C. R. DOERING, *Heat transfer in convective turbulence*, Nonlinearity, 9 (1996), pp. 1049–1060.
- [8] P. CONSTANTIN AND C. R. DOERING, *Infinite Prandtl number convection*, J. Stat. Phys., 94 (1999), pp. 159–172.
- [9] W. CHENG AND X. WANG, *A discrete Kato-type theorem on inviscid limit of Navier–Stokes flows*, J. Math. Phys., 48 (2007), p. 065303.
- [10] W. CHENG AND X. WANG, *A uniformly dissipative scheme for stationary statistical properties of the infinite Prandtl number model for convection*, Appl. Math. Lett., in press.
- [11] G. DA PRATO AND J. ZABCZYK, *Ergodicity for Infinite Dimensional Systems*, Cambridge University Press, Cambridge, UK, 1996.
- [12] C. R. DOERING, F. OTTO, AND M. G. REZNIKOFF, *Bounds on vertical heat transport for infinite-Prandtl-number Rayleigh–Bénard convection*, J. Fluid Mech., 560 (2006), pp. 229–241.
- [13] C. FOIAS, M. JOLLY, I. G. KEVREKIDIS, AND E. S. TITI, *Dissipativity of numerical schemes*, Nonlinearity, 4 (1991), pp. 591–613.

- [14] C. FOIAS, M. JOLLY, I. G. KEVREKIDIS, AND E. S. TITI, *On some dissipative fully discrete nonlinear Galerkin schemes for the Kuramoto–Sivashinsky equation*, Phys. Lett. A, 186 (1994), pp. 87–96.
- [15] C. FOIAS, O. MANLEY, R. ROSA, AND R. TEMAM, *Navier-Stokes equations and turbulence*, in Encyclopedia of Mathematics and its Applications 83, Cambridge University Press, Cambridge, UK, 2001.
- [16] A. V. GETLING, *Rayleigh–Bénard Convection*, Structures and Dynamics in Advanced Series in Nonlinear Dynamics 11, World Scientific Publishing Co., Inc., River Edge, NJ, 1998.
- [17] T. GEVECI, *On the convergence of a time discretization scheme for the Navier–Stokes equations*, Math. Comp., 53 (1989), pp. 43–53.
- [18] S. GROSSMANN AND D. LOHSE, *Scaling in thermal convection: A unifying theory*, J. Fluid Mech., 407 (2000), pp. 27–56.
- [19] J. K. HALE, *Asymptotic Behavior of Dissipative Systems*, American Mathematical Society, Providence, RI, 1988.
- [20] J. G. HEYWOOD AND R. RANNACHER, *Finite element approximation of the nonstationary Navier-Stokes problem. I. Regularity of solutions and second order error estimates for spatial discretization*, SIAM J. Numer. Anal., 19 (1982), pp. 275–311.
- [21] J. G. HEYWOOD AND R. RANNACHER, *Finite element approximation of the nonstationary Navier-Stokes problem. II. Stability of solutions and error estimates uniform in time*, SIAM J. Numer. Anal., 23 (1986), pp. 750–777.
- [22] A. T. HILL AND E. SÜLI, *Approximation of the global attractor for the incompressible Navier-Stokes equations*, IMA J. Numer. Anal., 20 (2000), pp. 663–667.
- [23] L. HOWARD, *Heat transport by turbulent convection*, J. Fluid Mech., 17 (1963), pp. 405–432.
- [24] G. R. IERLEY, R. R. KERSWELL, AND S. C. PLASTING, *Infinite-Prandtl-number convection. II. A singular limit of upper bound theory*, J. Fluid Mech., 560 (2006), pp. 159–227.
- [25] M. JOLLY, I. KEVREKIDIS, AND E. S. TITI, *Preserving dissipation in approximate inertial forms for the Kuramoto–Sivashinsky equation*, J. Dynam. Differential Equations, 3 (1991), pp. 179–197.
- [26] N. JU, *On the global stability of a temporal discretization scheme for the Navier-Stokes equations*, IMA J. Numer. Anal., 22 (2002), pp. 577–597.
- [27] L. P. KADANOFF, *Turbulent heat flow: Structures and scaling*, Physics Today, 54 (2001), pp. 34–39.
- [28] S. LARSSON, *The long-time behavior of finite-element approximations of solutions to semilinear parabolic problems*, SIAM J. Numer. Anal., 26 (1989), pp. 348–365.
- [29] A. LASOTA AND M. C. MACKEY, *Chaos, Fractals, and Noise*. Stochastic Aspects of Dynamics, 2nd ed., Springer-Verlag, New York, 1994.
- [30] P. D. LAX, *Functional Analysis*, John Wiley and Sons, New York, 2002.
- [31] A. J. MAJDA AND X. WANG, *Nonlinear Dynamics and Statistical Theory for Basic Geophysical Flows*, Cambridge University Press, Cambridge, UK, 2006.
- [32] J. C. MATTINGLY, A. M. STUART, AND D. J. HIGHAM, *Ergodicity for SDEs and approximations: Locally Lipschitz vector fields and degenerate noise*, Stochastic Process. Appl., 101 (2002), pp. 185–232.
- [33] A. S. MONIN AND A. M. YAGLOM, *Statistical Fluid Mechanics; Mechanics of Turbulence* (English ed. updated, augmented and rev. by the authors), MIT Press, Cambridge, MA, 1975.
- [34] G. RAUGEL, *Global attractors in partial differential equations*, in Handbook of Dynamical Systems, Vol. 2, North-Holland, Amsterdam, 2002, pp. 885–982.
- [35] T. SAUER AND J. A. YORKE, *Rigorous verification of trajectories for the computer simulation of dynamical systems*, Nonlinearity, 4 (1991), pp. 961–979.
- [36] J. SHEN, *Long time stability and convergence for the fully discrete nonlinear Galerkin methods*, Appl. Anal., 38 (1990), pp. 201–229.
- [37] E. D. SIGGIA, *High Rayleigh number convection*, Annual Review of Fluid Mechanics, Vol. 26, Annual Reviews, Palo Alto, CA, 1994, pp. 137–168.
- [38] R. M. TEMAM, *Navier-Stokes Equations and Nonlinear Functional Analysis*, 2nd ed., CBMS-NSF Regional Conference Series in Applied Mathematics 66, SIAM, 1995.
- [39] R. M. TEMAM, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, 2nd ed., Springer-Verlag, New York, 1997.
- [40] F. TONE AND D. WIROSOETISNO, *On the long-time stability of the implicit Euler scheme for the two-dimensional Navier-Stokes equations*, SIAM J. Numer. Anal., 44 (2006), pp. 29–40.
- [41] D. J. TRITTON, *Physical Fluid Dynamics*, 2nd ed., Oxford Science Publications, Clarendon Press, Oxford University Press, New York, 1988.
- [42] M. I. VISHIK AND A. V. FURSIKOV, *Mathematical Problems of Statistical Hydromechanics*, Kluwer Academic Publishers, Dordrecht/Boston/London, 1988.

- [43] P. WALTERS, *An Introduction to Ergodic Theory*, Springer-Verlag, New York-Berlin, 1982.
- [44] X. WANG, *Infinite Prandtl number limit of Rayleigh-Bénard convection*, *Comm. Pure Appl. Math.*, 57 (2004), pp. 1265–1282.
- [45] X. WANG, *Asymptotic behavior of global attractors to the Boussinesq system for Rayleigh-Bénard convection at large Prandtl number*, *Comm. Pure Appl. Math.*, 60 (2007), pp. 1293–1318.
- [46] X. WANG, *Stationary statistical properties of Rayleigh-Bénard convection at large Prandtl number*, *Comm. Pure Appl. Math.*, 61 (2008), pp. 789–815.
- [47] X. WANG, *Lecture Notes on Elementary Statistical Theories with Applications to Fluid Systems*, Shanghai Mathematics Summer School in Fudan University, 2007, High Education Press, in press.
- [48] X. WANG, *Temporal Approximations of Stationary Statistical Properties of Dissipative Systems*, submitted.
- [49] E. WEINAN, *Stochastic hydrodynamics*, in *Current Developments in Mathematics*, 2000, Intl. Press, Somerville, MA, 2001, pp. 109–147.
- [50] E. WEINAN AND D. LI, *The Andersen Thermostat in Molecular Dynamics*, preprint.

A LAGUERRE–LEGENDRE SPECTRAL METHOD FOR THE STOKES PROBLEM IN A SEMI-INFINITE CHANNEL*

MEJDI AZAIEZ[†], JIE SHEN[‡], CHUANJU XU[§], AND QINGQU ZHUANG[§]

Abstract. A mixed spectral method is proposed and analyzed for the Stokes problem in a semi-infinite channel. The method is based on a generalized Galerkin approximation with Laguerre functions in the x direction and Legendre polynomials in the y direction. The well-posedness of this method is established by deriving a lower bound on the inf-sup constant. Numerical results indicate that the derived lower bound is sharp. Rigorous error analysis is also carried out.

Key words. Laguerre functions, Legendre polynomials, Stokes equations, inf-sup condition, error analysis, spectral method

AMS subject classifications. Primary, 65N35; Secondary, 74S25, 76D07

DOI. 10.1137/070698269

1. Introduction. The Stokes problem plays an important role in fluid dynamics and in solid mechanics, and its numerical approximation has attracted much attention during the last three decades (see, for instance, [10, 5, 4] and the references therein). Most of these investigations have been concentrated on problems in bounded domains.

There is, however, a need to consider numerical approximations to the Stokes problem in unbounded domains. In particular, the flow in a channel and flow past a cylinder/sphere have important theoretical and practical applications. In most previous investigations for these problems, an artificial boundary is introduced, and an approximate boundary condition at the artificial boundary has to be used. The accuracy of these methods usually depends on how far downstream the artificial boundary is (cf. [17]). Therefore, even when high-order or spectral methods are applied for these problems, one can not achieve high-order or spectral accuracy for the original problem due to the approximation made to the “unknown” outflow boundary conditions.

We shall take a different approach in this paper. More precisely, we shall consider the problem directly in the unbounded domain without introducing an artificial boundary. In fact, many other problems in science and engineering are also set in unbounded domains, and there have been some investigations in using Laguerre polynomials/functions to approximate PDEs on semi-infinite intervals (see, among others, [8, 12, 18, 11, 15]). However, all these works are concerned with Poisson-type elliptic equations. To the best of our knowledge, no result is available for spectral methods to the Stokes problem in semi-infinite channels. Thus, results in this paper are the first of its kind and will play an important role for the numerical approximation of Stokes and Navier–Stokes equations in unbounded domains.

*Received by the editors July 25, 2007; accepted for publication (in revised form) May 27, 2008; published electronically November 21, 2008.

<http://www.siam.org/journals/sinum/47-1/69826.html>

[†]Laboratoire TREFLE (UMR CNRS 8508), ENSCPB, 33607 Pessac, France (azaiez@enscpb.fr).

[‡]Corresponding author. Department of Mathematics, Purdue University, West Lafayette, IN, 47907 (shen@math.purdue.edu). The work of this author was partially supported by NFS grant DMS-0610646.

[§]School of Mathematical Sciences, Xiamen University, 361005 Xiamen, China (cjxu@xmu.edu.cn, xmuyfd129@163.com). The research of the third author was partially supported by the NSF of China under grant 10531080, the Excellent Young Teachers Program by the Ministry of Education of China, and the 973 High Performance Scientific Computation Research Program. Part of this work was done when the third author was at the Université de Bordeaux-I as an invited professor.

More precisely, we consider the Stokes equations in a semi-infinite channel and introduce a mixed formulation based on Laguerre functions in the x direction and Legendre polynomials in the y direction. It is worthwhile to emphasize that we use Laguerre functions instead of Laguerre polynomials because the latter behaves wildly at infinity and is not suitable for approximation to flows which are well-behaved at infinity (cf. [18]). The well-posedness of this mixed formulation relies on the verification of the so-called inf-sup condition (cf. [1, 6]). The main contribution of this paper is the derivation of a lower bound on the inf-sup constant. We shall also present numerical results which indicate that the derived lower bound is sharp.

The rest of the paper is organized as follows. In the next section, we introduce some notations, derive some useful inverse inequalities for Laguerre functions and Legendre polynomials, and present the mixed Laguerre–Legendre formulation for the Stokes problem. Section 3 is devoted to deriving a lower bound for the inf-sup constant. In section 4, we carry out a complete error analysis for the mixed Laguerre–Legendre approximation. Finally, we present some implementation details and numerical results in section 5.

2. Mixed Laguerre–Legendre approximation. We start by introducing some notations. Let $R^+ = (0, +\infty)$, $\Lambda = (-1, 1)$, $\Omega = R^+ \times \Lambda$, and $\Gamma = \partial\Omega$. Let $\omega > 0$ be a weight function on Ω ; we denote by $(u, v)_{\Omega, \omega} := \int_{\Omega} uv\omega d\Omega$ the inner product of $L^2_{\omega}(\Omega)$, whose norm is denoted by $\|\cdot\|_{\omega, \Omega}$. We use $H^m_{\omega}(\Omega)$ and $H^m_{0, \omega}(\Omega)$ to denote the usual weighted Sobolev spaces, with norm $\|\cdot\|_{m, \omega, \Omega}$. In cases where no confusion would arise, ω (if $\omega \equiv 1$) and Ω may be dropped from the notations. Let M and N be the discretization parameters in x and in y . We denote by c a generic positive constant independent of the discretization parameters, and we use the expression $A \lesssim B$ to mean that $A \leq cB$. Throughout this paper we will use boldface letters to denote vectors and vector functions for ease of reading.

Let $\mathcal{L}_k(x)$ be the Laguerre polynomial of degree k ; we denote the Laguerre function by

$$\hat{\mathcal{L}}_i(x) = \mathcal{L}_i(x)e^{-x/2}$$

and set

$$\mathbb{P}_M = \text{span}\{\mathcal{L}_i(x), i = 0, 1, \dots, M\}$$

and

$$\hat{\mathbb{P}}_M = \text{span}\{\hat{\mathcal{L}}_i(x), i = 0, 1, \dots, M\}.$$

We now recall some definitions and related results which will be used in what follows.

Let $\omega(x) = e^{-x}$, and let

$$L^2_{\omega}(R^+) := \left\{ v; ve^{-x/2} \in L^2(R^+) \right\} = \left\{ v; \int_0^{\infty} v^2 \omega dx < \infty \right\}.$$

The space $L^2_{\omega}(R^+)$ is endowed with the norm $\|\cdot\|_{0, \omega, R^+}$ (also denoted by $\|\cdot\|_{0, \omega}$ or $\|\cdot\|_{\omega}$ when there is no confusion), defined by

$$\|v\|_{0, \omega, R^+} = \left(\int_0^{\infty} v^2 \omega dx \right)^{1/2}.$$

Let π_M^x be the L_ω^2 -orthogonal projector from $L_\omega^2(R^+)$ into $\mathbb{P}_M(R^+)$ defined by

$$\int_0^\infty (v - \pi_M^x v) \phi_M \omega dx = 0, \quad \forall v \in L_\omega^2(R^+), \quad \phi_M \in \mathbb{P}_M(R^+).$$

The projector π_M^x can be characterized by the following expression:

$$(2.1) \quad \pi_M^x v(x) = \sum_{m=0}^M \alpha_m \mathcal{L}_m(x) \quad \forall v(x) = \sum_{m=0}^\infty \alpha_m \mathcal{L}_m(x).$$

We define the operator $\hat{\pi}_M^x$ from $L^2(R^+)$ into $\hat{\mathbb{P}}_M(R^+)$ by (cf. [18])

$$\hat{\pi}_M^x v(x) = e^{-x/2} \pi_M^x(v(x)e^{x/2}) \quad \forall v \in L^2(R^+).$$

It can be easily verified that

$$(2.2) \quad \int_0^\infty (\hat{\pi}_M^x v - v) \phi_M dx = \int_0^\infty \left(\pi_M^x(v(x)e^{x/2}) - v(x)e^{x/2} \right) e^{-x/2} \phi_M dx = 0 \quad \forall \phi_M \in \hat{\mathbb{P}}_M(R^+).$$

Consequently, $\hat{\pi}_M^x$ is the orthogonal projector from $L^2(R^+)$ into $\hat{\mathbb{P}}_M(R^+)$.

We now present several useful results. We start with an inverse inequality for Laguerre functions.

LEMMA 2.1. *For all $\phi_M \in \hat{\mathbb{P}}_M(R^+)$, we have*

$$\|\partial_x \phi_M\|_{0,R^+} \lesssim M \|\phi_M\|_{0,R^+}.$$

Proof. Let $\phi_M(x) = \sum_{k=0}^M \tilde{\phi}_k \hat{\mathcal{L}}_k(x)$. Then, $\|\phi_M\|_{0,R^+}^2 = \sum_{k=0}^M \tilde{\phi}_k^2$ and

$$\partial_x \phi_M(x) = \sum_{k=0}^M \tilde{\phi}_k \hat{\mathcal{L}}'_k(x) = \sum_{k=0}^M \tilde{\phi}_k \left(\mathcal{L}'_k(x) - \frac{1}{2} \mathcal{L}_k(x) \right) e^{-\frac{x}{2}}.$$

Hence, the desired result is a direct consequence of the above and the inverse inequality for Laguerre polynomials (cf. [4]). \square

We now denote by $\mathbb{P}_N(\Lambda)$ the space of polynomials of degree less than or equal to N in Λ , and let π_N^y be the standard L^2 -orthogonal projector from $L^2(\Lambda)$ into $\mathbb{P}_N(\Lambda)$.

LEMMA 2.2. *For all $\phi_N \in \mathbb{P}_N(\Lambda) \cap H_0^1(\Lambda)$, we have*

$$\|\phi_N\|_{0,\Lambda} \leq N^{1/2} \|\pi_{N-2}^y \phi_N\|_{0,\Lambda}.$$

Remark 2.1. A proof of the above result, with a constant in front of $N^{1/2}$, can be found in [3]. In fact, a more precise computation as in [3] shows that the constant can be bounded by one.

A similar result with respect to $\hat{\mathbb{P}}_M(R^+) \cap H_0^1(R^+)$ is as follows.

LEMMA 2.3. *For all $\phi_M \in \hat{\mathbb{P}}_M(R^+) \cap H_0^1(R^+)$, we have*

$$\|\phi_M\|_{0,R^+} \leq (M+1)^{1/2} \|\hat{\pi}_{M-1}^x \phi_M\|_{0,R^+}.$$

Proof. Writing ϕ_M in the form

$$\phi_M(x) = \sum_{m=0}^M \alpha_m \mathcal{L}_m(x) e^{-x/2},$$

we derive from (2.1) that

$$\hat{\pi}_{M-1}^x \phi_M(x) = e^{-x/2} \pi_{M-1}^x \left(\sum_{m=0}^M \alpha_m \mathcal{L}_m(x) \right) = e^{-x/2} \sum_{m=0}^{M-1} \alpha_m \mathcal{L}_m(x).$$

Hence,

$$\phi_M(x) = \hat{\pi}_{M-1}^x \phi_M(x) + \alpha_M \mathcal{L}_M(x) e^{-x/2},$$

and by using the orthogonality relation,

$$(2.3) \quad \|\phi_M\|_{0,R^+}^2 = \|\hat{\pi}_{M-1}^x \phi_M(x)\|_{0,R^+}^2 + \alpha_M^2 \int_0^\infty (\mathcal{L}_M(x) e^{-x/2})^2 dx.$$

Note that $\phi_M(0) = 0$ implies that

$$\alpha_0 + \alpha_1 + \dots + \alpha_M = 0,$$

from which

$$(2.4) \quad |\alpha_M| = |\alpha_0 + \dots + \alpha_{M-1}| \leq \left[\sum_{m=0}^{M-1} \alpha_m^2 \right]^{1/2} M^{1/2}.$$

Combining (2.3) and (2.4) gives

$$\|\phi_M\|_{0,R^+}^2 \leq \|\hat{\pi}_{M-1}^x \phi_M(x)\|_{0,R^+}^2 + M \sum_{m=0}^{M-1} \alpha_m^2 = (M+1) \|\hat{\pi}_{M-1}^x \phi_M(x)\|_{0,R^+}^2. \quad \square$$

Now we consider the mixed Laguerre–Legendre approximation. Let $\mathbb{P}_{M,N}(\Omega)$ be the space of all polynomials in Ω of degree $\leq M$ in the x direction and $\leq N$ in the y direction, i.e.,

$$\mathbb{P}_{M,N}(\Omega) := \text{span}\{\mathcal{L}_i(x)L_j(y), i = 0, 1, \dots, M; j = 0, 1, \dots, N\},$$

where $\mathcal{L}_i(x)$ and $L_j(y)$ are, respectively, Laguerre and Legendre polynomials of degree i and j , satisfying

$$\int_{-1}^1 \int_0^\infty \mathcal{L}_i(x)L_j(y)\mathcal{L}_m(x)L_n(y)e^{-x} dx dy = \frac{2}{2n+1} \delta_{im} \delta_{jn}.$$

We also define

$$\hat{\mathbb{P}}_{M,N}(\Omega) := \text{span}\{\hat{\mathcal{L}}_i(x)L_j(y), i = 0, 1, \dots, M; j = 0, 1, \dots, N\}.$$

Let us denote by \mathcal{N} the pair of parameters (M, N) and set

$$X_{\mathcal{N}} = H_0^1(\Omega)^2 \cap \hat{\mathbb{P}}_{M,N}(\Omega)^2, \quad M_{\mathcal{N}} = L^2(\Omega) \cap \hat{\mathbb{P}}_{M-1,N-2}(\Omega).$$

LEMMA 2.4. *For all $\psi \in H^2(\Omega) \cap H_0^1(\Omega)$, we have*

$$(2.5) \quad \|\psi\|_{2,\Omega}^2 \lesssim \left\| \frac{\partial^2 \psi}{\partial y^2} \right\|_{0,\Omega}^2 + \left\| \frac{\partial^2 \psi}{\partial x^2} \right\|_{0,\Omega}^2.$$

Proof. For all $\psi \in H^2(\Omega)$, with $\psi(x, \pm 1) = 0 \forall x \in (0, +\infty)$, we have

$$\begin{aligned} |\psi(x, y)|^2 &= \left| \int_{-1}^y \frac{\partial \psi(x, s)}{\partial s} ds \right|^2 \leq \int_{-1}^y \left(\frac{\partial \psi(x, s)}{\partial s} \right)^2 ds \int_{-1}^y ds \\ &\leq (y + 1) \int_{-1}^1 \left(\frac{\partial \psi(x, s)}{\partial s} \right)^2 ds. \end{aligned}$$

Therefore,

$$(2.6) \quad \int_{-1}^1 \psi^2(x, y) dy \leq \int_{-1}^1 (y + 1) dy \int_{-1}^1 \left(\frac{\partial \psi(x, s)}{\partial s} \right)^2 ds = 2 \int_{-1}^1 \left(\frac{\partial \psi(x, y)}{\partial y} \right)^2 dy.$$

As a consequence,

$$(2.7) \quad \begin{aligned} \|\psi\|_{0,\Omega}^2 &= \int_0^\infty \int_{-1}^1 \psi^2(x, y) dy dx \leq 2 \int_0^\infty \int_{-1}^1 \left(\frac{\partial \psi(x, y)}{\partial y} \right)^2 dy dx \\ &= 2 \left\| \frac{\partial \psi}{\partial y} \right\|_{0,\Omega}^2. \end{aligned}$$

Using (2.7), we find

$$\begin{aligned} \left\| \frac{\partial \psi}{\partial y} \right\|_{0,\Omega}^2 &= \int_0^\infty \int_{-1}^1 \left(\frac{\partial \psi}{\partial y} \right)^2 dy dx = - \int_0^\infty \int_{-1}^1 \frac{\partial^2 \psi}{\partial y^2} \psi dy dx \\ &\leq \left\| \frac{\partial^2 \psi}{\partial y^2} \right\|_{0,\Omega} \|\psi\|_{0,\Omega} \leq \sqrt{2} \left\| \frac{\partial^2 \psi}{\partial y^2} \right\|_{0,\Omega} \left\| \frac{\partial \psi}{\partial y} \right\|_{0,\Omega}, \end{aligned}$$

from which we derive

$$\left\| \frac{\partial \psi}{\partial y} \right\|_{0,\Omega}^2 \leq 2 \left\| \frac{\partial^2 \psi}{\partial y^2} \right\|_{0,\Omega}^2.$$

We then derive from (2.7) and the above that

$$(2.8) \quad \|\psi\|_{0,\Omega}^2 \leq 4 \left\| \frac{\partial^2 \psi}{\partial y^2} \right\|_{0,\Omega}^2.$$

On the other hand, applying (2.7) to $\frac{\partial \psi}{\partial x}$ with $\psi \in H_0^1(\Omega)$, we obtain

$$(2.9) \quad \left\| \frac{\partial \psi}{\partial x} \right\|_{0,\Omega}^2 \leq 2 \left\| \frac{\partial^2 \psi}{\partial y \partial x} \right\|_{0,\Omega}^2.$$

Furthermore, we have, for all $\psi \in H_0^1(\Omega)$,

$$(2.10) \quad 2 \left\| \frac{\partial^2 \psi}{\partial y \partial x} \right\|_{0,\Omega}^2 = 2 \int_{-1}^1 \int_0^\infty \left(\frac{\partial^2 \psi}{\partial x^2} \right) \left(\frac{\partial^2 \psi}{\partial y^2} \right) dx dy \leq \left\| \frac{\partial^2 \psi}{\partial x^2} \right\|_{0,\Omega}^2 + \left\| \frac{\partial^2 \psi}{\partial y^2} \right\|_{0,\Omega}^2.$$

Finally, combining the above inequalities leads to

$$\begin{aligned}
 \|\nabla\psi\|_{1,\Omega}^2 &= \left\| \frac{\partial\psi}{\partial x} \right\|_{1,\Omega}^2 + \left\| \frac{\partial\psi}{\partial y} \right\|_{1,\Omega}^2 \\
 &= \left\| \frac{\partial\psi}{\partial x} \right\|_{0,\Omega}^2 + \left| \frac{\partial\psi}{\partial x} \right|_{1,\Omega}^2 + \left\| \frac{\partial\psi}{\partial y} \right\|_{0,\Omega}^2 + \left| \frac{\partial\psi}{\partial y} \right|_{1,\Omega}^2 \\
 &= \int_{-1}^1 \int_0^\infty \left[\left(\frac{\partial\psi}{\partial x} \right)^2 + \left(\frac{\partial\psi}{\partial y} \right)^2 + \left(\frac{\partial^2\psi}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2\psi}{\partial x\partial y} \right)^2 + \left(\frac{\partial^2\psi}{\partial y^2} \right)^2 \right] dx dy \\
 &\leq \int_{-1}^1 \int_0^\infty \left[4 \left(\frac{\partial^2\psi}{\partial x\partial y} \right)^2 + \left(\frac{\partial^2\psi}{\partial x^2} \right)^2 + 3 \left(\frac{\partial^2\psi}{\partial y^2} \right)^2 \right] dx dy \\
 &\leq \int_{-1}^1 \int_0^\infty \left[3 \left(\frac{\partial^2\psi}{\partial x^2} \right)^2 + 5 \left(\frac{\partial^2\psi}{\partial y^2} \right)^2 \right] dx dy \\
 &= 3 \left\| \frac{\partial^2\psi}{\partial x^2} \right\|_{0,\Omega}^2 + 5 \left\| \frac{\partial^2\psi}{\partial y^2} \right\|_{0,\Omega}^2.
 \end{aligned}$$

This estimate, together with (2.7), yields (2.5). \square

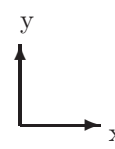
Now we set up the Stokes problem in a semi-infinite channel as depicted below:

(2.11)

Γ
 Ω
 Γ

Γ
 Γ

$\left\{ \begin{array}{l} -\Delta \mathbf{u} + \nabla p = \mathbf{f} \text{ in } \Omega, \\ \nabla \cdot \mathbf{u} = 0 \text{ in } \Omega, \\ \mathbf{u}|_\Gamma = 0, \\ \lim_{x \rightarrow \infty} \mathbf{u} = 0. \end{array} \right.$



Its weak formulation is as follows: Find $(\mathbf{u}, p) \in H_0^1(\Omega)^2 \times L^2(\Omega)$ such that

$$\begin{cases} (\nabla \mathbf{u}, \nabla \mathbf{v}) - (p, \nabla \cdot \mathbf{v}) = (\mathbf{f}, \mathbf{v}) \quad \forall \mathbf{v} \in H_0^1(\Omega)^2, \\ (q, \nabla \cdot \mathbf{u}) = 0 \quad \forall q \in L^2(\Omega). \end{cases}$$

Then, the mixed Laguerre–Legendre spectral approximation to (2.12) is as follows:

Find $\mathbf{u}_N \in X_N, p_N \in M_N$ such that

$$\begin{cases} (\nabla \mathbf{u}_N, \nabla \mathbf{v}_N)_N - (p_N, \nabla \cdot \mathbf{v}_N)_N = (\mathbf{f}, \mathbf{v}_N)_N \quad \forall \mathbf{v}_N \in X_N, \\ (q_N, \nabla \cdot \mathbf{u}_N)_N = 0 \quad \forall q_N \in M_N, \end{cases}$$

where the discrete inner product $(\cdot, \cdot)_N$ is defined by

$$(\phi, \psi)_N = \sum_{p=0}^M \sum_{q=0}^N \phi(\hat{\xi}_p, \xi_q) \psi(\hat{\xi}_p, \xi_q) \hat{\omega}_p \omega_q,$$

where $\{\hat{\xi}_p, \hat{\omega}_p\}_{p=0,1,\dots,N}$ are the Laguerre–Gauss–Radau points and the associated weights, such that the following quadrature rule holds:

$$\int_0^\infty \varphi(x) dx = \sum_{p=0}^M \varphi(\hat{\xi}_p) \hat{\omega}_p \quad \forall \varphi(x) \in \hat{\mathbb{P}}_{2M}(R^+);$$

$\{\xi_q, \omega_q\}_{q=0,1,\dots,N}$ are the Legendre–Gauss–Lobatto points and the associated weights, such that the following quadrature rule holds:

$$\int_{-1}^1 \varphi(y) dy = \sum_{q=0}^N \varphi(\xi_q) \omega_q, \quad \forall \varphi(y) \in \mathbb{P}_{2N-1}(\Lambda).$$

It is well known that, since the coercivity and continuity of the bilinear form $(\nabla \mathbf{w}_N, \nabla \mathbf{v}_N)_N$ and the continuity of the bilinear form $(\nabla \cdot \mathbf{v}_N, q_N)_N$ are evident, the well-posedness of the mixed formulation (2.13) relies on the so-called inf-sup condition [6]:

$$(2.14) \quad \inf_{q_N \in M_N} \sup_{\mathbf{v}_N \in X_N} \frac{-(\nabla \cdot \mathbf{v}_N, q_N)_N}{\|\mathbf{v}_N\|_{1,\Omega} \|q_N\|_{0,\Omega}} \geq \beta_N > 0,$$

where β_N is called the inf-sup constant. The next section is devoted to the estimation of this constant.

3. Estimation of the inf-sup constant. The main result in this section is what follows.

THEOREM 3.1.

$$(3.1) \quad \inf_{q_N \in M_N} \sup_{\mathbf{v}_N \in X_N} \frac{-(\nabla \cdot \mathbf{v}_N, q_N)_N}{\|\mathbf{v}_N\|_{1,\Omega} \|q_N\|_{0,\Omega}} \gtrsim \frac{1}{M}.$$

Remark 3.1. It is surprising that the inf-sup constant is independent of N , since it is well known that the inf-sup constant of the Legendre–Legendre $P_N^2 - P_{N-2}$ method in Λ^2 is of order $N^{-\frac{1}{2}}$ (see, for instance, [3]), and we have found numerically that in the Legendre–Legendre case in Λ^2 , the corresponding inf-sup constant behaves like $\max\{\frac{1}{\sqrt{M}}, \frac{1}{\sqrt{N}}\}$, where M, N are, respectively, the degrees of Legendre polynomials used in the x and y directions. However, our numerical results in section 5 indicate that the estimate (3.1) is sharp.

The proof of this result will be accomplished with a series of lemmas, which we present below. The confirmation of the result will be done by the numerical experiments carried out later.

LEMMA 3.1. *Given $q_N \in M_N$, the problem of finding $\psi_N \in H_0^1(\Omega) \cap \hat{\mathbb{P}}_{M,N}(\Omega)$ such that*

$$(3.2) \quad (\Delta \psi_N, r_N) = -(q_N, r_N) \quad \forall r_N \in M_N$$

admits a unique solution satisfying

$$(3.3) \quad \|\psi_N\|_{2,\Omega} \lesssim M \|q_N\|_{0,\Omega}.$$

Proof. Obviously, problem (3.2) defines a system with the number of unknowns equal to the number of equations, so the existence of such a function ψ_N is guaranteed by estimate (3.3), which we prove below.

By definition (3.2), we have

$$(\Delta \psi_N, r_1(x)r_2(y)) = -(q_N, r_1(x)r_2(y)), \quad \forall r_1 \in \hat{\mathbb{P}}_{M-1}(R^+), \quad \forall r_2 \in \mathbb{P}_{N-2}(\Lambda).$$

This implies

$$\begin{aligned} \int_0^\infty q_N(x, y) r_1(x) dx &= -\pi_{N-2}^y \int_0^\infty \Delta \psi_N(x, y) r_1(x) dx \\ &= -\int_0^\infty \pi_{N-2}^y (\Delta \psi_N(x, y)) r_1(x) dx \quad \forall r_1 \in \hat{\mathbb{P}}_{M-1}(R^+) \end{aligned}$$

and consequently

$$\begin{aligned} q_N(x, y) &= -\hat{\pi}_{M-1}^x \circ \pi_{N-2}^y (\Delta \psi_N(x, y)) = -\hat{\pi}_{M-1}^x \circ \pi_{N-2}^y \left(\frac{\partial^2 \psi_N}{\partial x^2} + \frac{\partial^2 \psi_N}{\partial y^2} \right) \\ &= -\hat{\pi}_{M-1}^x \circ \pi_{N-2}^y \frac{\partial^2 \psi_N}{\partial x^2} - \hat{\pi}_{M-1}^x \frac{\partial^2 \psi_N}{\partial y^2} \\ &= -\pi_{N-2}^y \circ \hat{\pi}_{M-1}^x \frac{\partial^2 \psi_N}{\partial x^2} - \hat{\pi}_{M-1}^x \frac{\partial^2 \psi_N}{\partial y^2}. \end{aligned}$$

Hence,

$$\begin{aligned} \|q_N\|_{0,\Omega}^2 &= \left\| \hat{\pi}_{M-1}^x \frac{\partial^2 \psi_N}{\partial y^2} \right\|_{0,\Omega}^2 + \left\| \pi_{N-2}^y \circ \hat{\pi}_{M-1}^x \frac{\partial^2 \psi_N}{\partial x^2} \right\|_{0,\Omega}^2 \\ &\quad + 2 \int_{\Omega} \hat{\pi}_{M-1}^x \frac{\partial^2 \psi_N}{\partial y^2} \pi_{N-2}^y \circ \hat{\pi}_{M-1}^x \frac{\partial^2 \psi_N}{\partial x^2} \\ (3.4) \quad &= \left\| \hat{\pi}_{M-1}^x \frac{\partial^2 \psi_N}{\partial y^2} \right\|_{0,\Omega}^2 + \left\| \pi_{N-2}^y \circ \hat{\pi}_{M-1}^x \frac{\partial^2 \psi_N}{\partial x^2} \right\|_{0,\Omega}^2 + 2 \int_{\Omega} \frac{\partial^2 \psi_N}{\partial y^2} \hat{\pi}_{M-1}^x \frac{\partial^2 \psi_N}{\partial x^2}. \end{aligned}$$

Observing that

$$(3.5) \quad \hat{\pi}_{M-1}^x \frac{\partial^2 \psi_N}{\partial x^2} = \frac{\partial^2 \psi_N}{\partial x^2} - \frac{1}{4} (I - \hat{\pi}_{M-1}^x) \psi_N,$$

where I denotes the identity operator, then the last term in (3.4) can be rewritten as

$$\int_{\Omega} \frac{\partial^2 \psi_N}{\partial y^2} \hat{\pi}_{M-1}^x \frac{\partial^2 \psi_N}{\partial x^2} = \int_{\Omega} \frac{\partial^2 \psi_N}{\partial y^2} \frac{\partial^2 \psi_N}{\partial x^2} - \frac{1}{4} \int_{\Omega} \frac{\partial^2 \psi_N}{\partial y^2} (I - \hat{\pi}_{M-1}^x) \psi_N.$$

For the first term on the right-hand side, we have, by integration by parts,

$$\int_{\Omega} \frac{\partial^2 \psi_N}{\partial y^2} \frac{\partial^2 \psi_N}{\partial x^2} = \int_{\Omega} \frac{\partial^2 \psi_N}{\partial x \partial y} \frac{\partial^2 \psi_N}{\partial x \partial y}.$$

To estimate the second term, we write

$$(3.6) \quad \psi_N = \sum_{m=0}^M \alpha_m(y) \mathcal{L}_m(x) e^{-x/2} \in H_0^1(\Omega) \cap \hat{\mathbb{P}}_{M,N}(\Omega).$$

Then, we have with $\alpha_m(y) \in \mathbb{P}_N^0(\Lambda)$, $m = 0, 1, \dots, M$, and by using the orthogonality of the Laguerre polynomials,

$$\begin{aligned} -\frac{1}{4} \int_{\Omega} \frac{\partial^2 \psi_N}{\partial y^2} (I - \hat{\pi}_{M-1}^x) \psi_N &= -\frac{1}{4} \int_{\Omega} \left[\sum_{m=0}^M \alpha_m''(y) \mathcal{L}_m(x) e^{-x/2} \right] \alpha_M(y) \mathcal{L}_M(x) e^{-x/2} \\ &= -\frac{1}{4} \sum_{m=0}^M \int_0^{\infty} \left(\int_{-1}^1 \alpha_m''(y) \alpha_M(y) dy \right) \mathcal{L}_m(x) \mathcal{L}_M(x) e^{-x} dx \\ &= -\frac{1}{4} \int_{-1}^1 \alpha_M''(y) \alpha_M(y) dy \\ &= \frac{1}{4} \int_{-1}^1 \alpha_M'(y) \alpha_M'(y) dy. \end{aligned}$$

Combining the above estimates leads to

$$(3.7) \quad \int_{\Omega} \frac{\partial^2 \psi_{\mathcal{N}}}{\partial y^2} \hat{\pi}_{M-1}^x \frac{\partial^2 \psi_{\mathcal{N}}}{\partial x^2} = \int_{\Omega} \frac{\partial^2 \psi_{\mathcal{N}}}{\partial x \partial y} \frac{\partial^2 \psi_{\mathcal{N}}}{\partial x \partial y} + \frac{1}{4} \int_{-1}^1 \alpha'_M(y) \alpha'_M(y) dy.$$

Hence, by using (3.7) and Lemma 2.3 in (3.4), we obtain

$$(3.8) \quad \begin{aligned} \|q_{\mathcal{N}}\|_{0,\Omega}^2 &= \left\| \hat{\pi}_{M-1}^x \frac{\partial^2 \psi_{\mathcal{N}}}{\partial y^2} \right\|_{0,\Omega}^2 + \left\| \pi_{N-2}^y \circ \hat{\pi}_{M-1}^x \frac{\partial^2 \psi_{\mathcal{N}}}{\partial x^2} \right\|_{0,\Omega}^2 \\ &\quad + 2 \int_{\Omega} \left(\frac{\partial^2 \psi_{\mathcal{N}}}{\partial x \partial y} \right)^2 dx dy + \frac{1}{2} \int_{-1}^1 (\alpha'_M(y))^2 dy \\ &\gtrsim M^{-1} \left\| \frac{\partial^2 \psi_{\mathcal{N}}}{\partial y^2} \right\|_{0,\Omega}^2 + \left\| \frac{\partial^2 \psi_{\mathcal{N}}}{\partial x \partial y} \right\|_{0,\Omega}^2. \end{aligned}$$

On the other hand, from the inverse inequality in the x direction (cf. Lemma 2.1) and the Poincaré inequality in the y direction, we have

$$\left\| \frac{\partial^2 \psi}{\partial x^2} \right\|_{0,\Omega} \lesssim M \left\| \frac{\partial \psi}{\partial x} \right\|_{0,\Omega} \lesssim M \left\| \frac{\partial^2 \psi}{\partial x \partial y} \right\|_{0,\Omega} \quad \forall \psi \in H_0^1(\Omega) \cap \hat{\mathbb{P}}_{M,N}(\Omega).$$

Using the above inequality and Lemma 2.4 in (3.8) gives

$$\begin{aligned} \|q_{\mathcal{N}}\|_{0,\Omega}^2 &\gtrsim M^{-1} \left\| \frac{\partial^2 \psi_{\mathcal{N}}}{\partial y^2} \right\|_{0,\Omega}^2 + M^{-2} \left\| \frac{\partial^2 \psi_{\mathcal{N}}}{\partial x^2} \right\|_{0,\Omega}^2 \\ &\gtrsim M^{-2} \|\psi_{\mathcal{N}}\|_{2,\Omega}^2. \end{aligned}$$

This leads to (3.3). \square

LEMMA 3.2. *For all $q_{\mathcal{N}} \in M_{\mathcal{N}}$, there exists $\mathbf{z}_{\mathcal{N}} \in (\hat{\mathbb{P}}_{M,N}(\Omega) \cap H_0^1(\Omega)) \times (\hat{\mathbb{P}}_{M,N+1}(\Omega) \cap H_0^1(\Omega))$ such that*

$$(\nabla \cdot \mathbf{z}_{\mathcal{N}}, r_{\mathcal{N}}) = -(q_{\mathcal{N}}, r_{\mathcal{N}}) \quad \forall r_{\mathcal{N}} \in M_{\mathcal{N}}$$

and

$$\|\mathbf{z}_{\mathcal{N}}\|_{1,\Omega} \lesssim M \|q_{\mathcal{N}}\|_{0,\Omega}.$$

Proof. For any $q_{\mathcal{N}} \in M_{\mathcal{N}}$, let $\psi_{\mathcal{N}}$ be defined by (3.2) and $\mathbf{w}_{\mathcal{N}} = \nabla \psi_{\mathcal{N}}$. Then, we have $\mathbf{w}_{\mathcal{N}} \in \hat{\mathbb{P}}_{M,N}(\Omega)^2$ satisfying

$$(3.9) \quad \begin{cases} (\nabla \cdot \mathbf{w}_{\mathcal{N}}, r_{\mathcal{N}}) = -(q_{\mathcal{N}}, r_{\mathcal{N}}) \quad \forall r_{\mathcal{N}} \in M_{\mathcal{N}}, \\ \mathbf{w}_{\mathcal{N}} \cdot \boldsymbol{\tau} = 0, \\ \|\mathbf{w}_{\mathcal{N}}\|_{1,\Omega} \lesssim M \|q_{\mathcal{N}}\|_{0,\Omega}, \end{cases}$$

where $\boldsymbol{\tau}$ is the unit tangent vector along $\partial\Omega$.

We now construct a lifting function $\phi_{\mathcal{N}} \in \hat{\mathbb{P}}_{M,N+1}(\Omega)$ such that

$$(3.10) \quad \begin{cases} \frac{\partial \phi_{\mathcal{N}}}{\partial \boldsymbol{\tau}} = -\mathbf{w}_{\mathcal{N}} \cdot \mathbf{n} \quad \text{on } \Gamma, \\ \frac{\partial \phi_{\mathcal{N}}}{\partial \mathbf{n}} = 0 \quad \text{on } \Gamma, \\ \|\phi_{\mathcal{N}}\|_{2,\Omega} \lesssim \|\mathbf{w}_{\mathcal{N}}\|_{1,\Omega}, \end{cases}$$

where \mathbf{n} is the outward normal to $\partial\Omega$. To this end, we define three functions on the boundaries $\Gamma_1 = \{(x, 1), 0 \leq x < \infty\}$, $\Gamma_2 = \{(0, y), -1 \leq y \leq 1\}$, $\Gamma_3 = \{(x, -1), 0 \leq x < \infty\}$, respectively, as follows:

$$\begin{aligned} b_{\mathcal{N}}^1(x, y) &= - \int_{\infty}^x (\mathbf{w}_{\mathcal{N}} \cdot \mathbf{n})(\sigma, y) d\sigma, \\ b_{\mathcal{N}}^2(x, y) &= b_{\mathcal{N}}^1(0, 1) - \int_1^y (\mathbf{w}_{\mathcal{N}} \cdot \mathbf{n})(x, \sigma) d\sigma, \\ b_{\mathcal{N}}^3(x, y) &= b_{\mathcal{N}}^2(0, -1) - \int_0^x (\mathbf{w}_{\mathcal{N}} \cdot \mathbf{n})(\sigma, y) d\sigma. \end{aligned}$$

Then, it can be easily verified that $b_{\mathcal{N}}^j$ ($j = 1, 2, 3$) satisfy the following continuity conditions:

$$\begin{aligned} b_{\mathcal{N}}^1(0, 1) &= b_{\mathcal{N}}^2(0, 1), \\ b_{\mathcal{N}}^2(0, -1) &= b_{\mathcal{N}}^3(0, -1), \\ \frac{\partial b_{\mathcal{N}}^1}{\partial \boldsymbol{\tau}}(0, 1) &= -\mathbf{w}_{\mathcal{N}} \cdot \mathbf{n}(0, 1) = 0 = \frac{\partial b_{\mathcal{N}}^2}{\partial \boldsymbol{\tau}}(0, 1), \\ \frac{\partial b_{\mathcal{N}}^2}{\partial \boldsymbol{\tau}}(0, -1) &= -\mathbf{w}_{\mathcal{N}} \cdot \mathbf{n}(0, -1) = 0 = \frac{\partial b_{\mathcal{N}}^3}{\partial \boldsymbol{\tau}}(0, -1), \\ \frac{\partial^2 b_{\mathcal{N}}^1}{\partial x \partial y}(0, 1) &= \frac{\partial^2 b_{\mathcal{N}}^2}{\partial x \partial y}(0, 1) = 0, \\ \frac{\partial^2 b_{\mathcal{N}}^2}{\partial x \partial y}(0, -1) &= \frac{\partial^2 b_{\mathcal{N}}^3}{\partial x \partial y}(0, -1) = 0. \end{aligned}$$

The above conditions, together with the fact that $b_{\mathcal{N}}^j \in \hat{\mathbb{P}}_{M, N+1}(\Omega)$ ($j = 1, 2, 3$), guarantee that there exists a $\phi_{\mathcal{N}} \in \hat{\mathbb{P}}_{M, N+1}(\Omega)$ satisfying (see [2])

$$(3.11) \quad \begin{cases} \phi_{\mathcal{N}} = b_{\mathcal{N}}^j & \text{on } \Gamma_j, \quad j = 1, 2, 3, \\ \frac{\partial \phi_{\mathcal{N}}}{\partial \mathbf{n}} = 0 & \text{on } \Gamma, \end{cases}$$

and

$$\|\phi_{\mathcal{N}}\|_{2, \Omega} \lesssim \sum_{j=1}^3 \|b_{\mathcal{N}}^j\|_{3/2, \Gamma} \lesssim \|\mathbf{w}_{\mathcal{N}} \cdot \mathbf{n}\|_{1/2, \Gamma} \lesssim \|\mathbf{w}_{\mathcal{N}}\|_{1, \Omega}.$$

Moreover, it is seen that the first equality of (3.11) implies

$$\frac{\partial \phi_{\mathcal{N}}}{\partial \boldsymbol{\tau}} = -\mathbf{w}_{\mathcal{N}} \cdot \mathbf{n} \quad \text{on } \Gamma.$$

This completes the construction of the lifting function $\phi_{\mathcal{N}}$. Now, let

$$\mathbf{z}_{\mathcal{N}} = \mathbf{w}_{\mathcal{N}} + \text{rot} \phi_{\mathcal{N}},$$

then $\mathbf{z}_{\mathcal{N}} \in \hat{\mathbb{P}}_{M, N}(\Omega) \times \hat{\mathbb{P}}_{M, N+1}(\Omega)$ and

$$\begin{aligned} \mathbf{z}_{\mathcal{N}} \cdot \mathbf{n}|_{\Gamma} &= \mathbf{w}_{\mathcal{N}} \cdot \mathbf{n} + \text{rot} \phi_{\mathcal{N}} \cdot \mathbf{n} = \mathbf{w}_{\mathcal{N}} \cdot \mathbf{n} + \frac{\partial \phi_{\mathcal{N}}}{\partial \boldsymbol{\tau}} = 0, \\ \mathbf{z}_{\mathcal{N}} \cdot \boldsymbol{\tau}|_{\Gamma} &= \mathbf{w}_{\mathcal{N}} \cdot \boldsymbol{\tau} + \text{rot} \phi_{\mathcal{N}} \cdot \boldsymbol{\tau} = \frac{\partial \phi_{\mathcal{N}}}{\partial \mathbf{n}} = 0, \end{aligned}$$

which means $\mathbf{z}_N \in (\hat{\mathbb{P}}_{M,N}(\Omega) \cap H_0^1(\Omega)) \times (\hat{\mathbb{P}}_{M,N+1}(\Omega) \cap H_0^1(\Omega))$. Moreover, we have

$$(\nabla \cdot \mathbf{z}_N, r_N) = (\nabla \cdot \mathbf{w}_N, r_N) + (\nabla \cdot \text{rot}\phi_N, r_N) = -(q_N, r_N) \quad \forall r_N \in M_N,$$

and, by (3.9) and the last inequality of (3.10),

$$(3.12) \quad \begin{aligned} \|\mathbf{z}_N\|_{1,\Omega} &= \|\mathbf{w}_N + \text{rot}\phi_N\|_{1,\Omega} \leq \|\mathbf{w}_N\|_{1,\Omega} + \|\text{rot}\phi_N\|_{1,\Omega} \\ &\lesssim \|\mathbf{w}_N\|_{1,\Omega} \lesssim M\|q_N\|_{0,\Omega}. \end{aligned}$$

The proof is complete. \square

LEMMA 3.3. *For all $q_N \in M_N$, there exists $\mathbf{v}_N \in X_N$ such that*

$$(\nabla \cdot \mathbf{v}_N, r_N) = -(q_N, r_N) \quad \forall r_N \in M_N$$

and

$$\|\mathbf{v}_N\|_{1,\Omega} \lesssim M\|q_N\|_{0,\Omega}.$$

Proof. For given q_N , let $\mathbf{z}_N := (z_N^{(1)}, z_N^{(2)}) \in (\hat{\mathbb{P}}_{M,N}(\Omega) \cap H_0^1(\Omega)) \times (\hat{\mathbb{P}}_{M,N+1}(\Omega) \cap H_0^1(\Omega))$ be a function associated to q_N in Lemma 3.2. Then, the second component of \mathbf{z}_N can be written under form

$$z_N^{(2)} = \sum_{i=2}^{N+1} \alpha_i(x) e^{-x/2} (L_i(y) - L_{i-2}(y)),$$

with $\alpha_i(x) \in \mathbb{P}_M(R^+) \cap H_0^1(R^+)$, $i = 2, \dots, M + 1$. We decompose $z_N^{(2)}$ into

$$z_N^{(2)} = \tilde{z}_N^{(2)} + \bar{z}_N^{(2)},$$

with

$$\begin{aligned} \tilde{z}_N^{(2)} &= \sum_{i=2}^N \alpha_i(x) e^{-x/2} (L_i(y) - L_{i-2}(y)), \\ \bar{z}_N^{(2)} &= \alpha_{N+1}(x) e^{-x/2} (L_{N+1}(y) - L_{N-1}(y)), \end{aligned}$$

and let

$$\mathbf{v}_N = \left(z_N^{(1)}, \tilde{z}_N^{(2)} \right).$$

Then, it is seen that $\mathbf{v}_N \in X_N$, moreover, for all $r_N \in M_N$, by using the orthogonality of the Legendre polynomials, we have

$$\begin{aligned} (\nabla \cdot \mathbf{v}_N, r_N) &= \left(\partial_x z_N^{(1)} + \partial_y \tilde{z}_N^{(2)}, r_N \right) \\ &= \left(\partial_x z_N^{(1)} + \partial_y \tilde{z}_N^{(2)}, r_N \right) - \left(\alpha_{N+1}(x) e^{-x/2} (L_{N+1}(y) - L_{N-1}(y)), \partial_y r_N \right) \\ &= \left(\partial_x z_N^{(1)} + \partial_y \tilde{z}_N^{(2)}, r_N \right) + \left(\alpha_{N+1}(x) e^{-x/2} \partial_y (L_{N+1}(y) - L_{N-1}(y)), r_N \right) \\ &= \left(\partial_x z_N^{(1)} + \partial_y z_N^{(2)}, r_N \right) \\ &= (\nabla \cdot \mathbf{z}_N, r_N) \\ &= -(q_N, r_N). \end{aligned}$$

It remains to prove $\|v_{\mathcal{N}}\|_{1,\Omega} \lesssim \|z_{\mathcal{N}}\|_{1,\Omega}$. Since $\tilde{z}_{\mathcal{N}}^{(2)} = z_{\mathcal{N}}^{(2)} - \bar{z}_{\mathcal{N}}^{(2)}$, we need only to prove $\|\bar{z}_{\mathcal{N}}^{(2)}\|_{1,\Omega} \lesssim \|z_{\mathcal{N}}^{(2)}\|_{1,\Omega}$. First, we have

$$\begin{aligned} \partial_x \bar{z}_{\mathcal{N}}^{(2)} &= \partial_x \left(\alpha_{N+1}(x)e^{-x/2} \right) (L_{N+1}(y) - L_{N-1}(y)), \\ \partial_x z_{\mathcal{N}}^{(2)} &= \sum_{i=2}^{N+1} \partial_x \left(\alpha_i(x)e^{-x/2} \right) (L_i(y) - L_{i-2}(y)) \\ &= \partial_x \left(\alpha_{N+1}(x)e^{-x/2} \right) L_{N+1}(y) + \partial_x \left(\alpha_N(x)e^{-x/2} \right) L_N(y) + \dots, \end{aligned}$$

thus

$$\begin{aligned} \left\| \partial_x z_{\mathcal{N}}^{(2)} \right\|_{0,\Omega}^2 &= \int_{-1}^1 \left[\int_0^\infty \left(\partial_x \left(\alpha_{N+1}(x)e^{-x/2} \right) \right)^2 dy \right] L_{N+1}(y)^2 dx + \dots \\ &\gtrsim \left\| \partial_x \left(\alpha_{N+1}(x)e^{-x/2} \right) \right\|_{0,R^+}^2 \frac{1}{N+1+1/2} \\ &\geq \frac{1}{6} \left| \alpha_{N+1}(x)e^{-x/2} \right|_{1,R^+}^2 \left(\frac{2}{2N+3} + \frac{2}{2N-1} \right) \\ &= \frac{1}{6} \left| \alpha_{N+1}(x)e^{-x/2} \right|_{1,R^+}^2 (\|L_{N+1}\|_{0,\Lambda}^2 + \|L_{N-1}\|_{0,\Lambda}^2) \\ &\gtrsim \left\| \partial_x \bar{z}_{\mathcal{N}}^{(2)} \right\|_{0,\Omega}^2. \end{aligned}$$

Similarly, we have

$$\left\| z_{\mathcal{N}}^{(2)} \right\|_{0,\Omega}^2 \gtrsim \left\| \bar{z}_{\mathcal{N}}^{(2)} \right\|_{0,\Omega}^2.$$

Second, from

$$\begin{aligned} \partial_y \bar{z}_{\mathcal{N}}^{(2)} &= \alpha_{N+1}(x)e^{-x/2} (L'_{N+1}(y) - L'_{N-1}(y)) = \alpha_{N+1}(x)e^{-x/2} (2N+1)L_N(y), \\ \partial_y z_{\mathcal{N}}^{(2)} &= \sum_{i=2}^{N+1} \alpha_i(x)e^{-x/2} (L'_i(y) - L'_{i-2}(y)) = \sum_{i=2}^{N+1} \alpha_i(x)e^{-x/2} (2i-1)L_{i-1}(y), \end{aligned}$$

we derive that

$$\begin{aligned} \left\| \partial_y z_{\mathcal{N}}^{(2)} \right\|_{0,\Omega}^2 &= \sum_{i=2}^{N+1} \left\| \alpha_i(x)e^{-x/2} \right\|_{0,R^+}^2 (2i-1)^2 \|L_{i-1}\|_{0,\Lambda}^2 \\ &\geq \left\| \alpha_{N+1}(x)e^{-x/2} \right\|_{0,R^+}^2 (2N+1)^2 \|L_N\|_{0,\Lambda}^2 \\ &= \left\| \partial_y \bar{z}_{\mathcal{N}}^{(2)} \right\|_{0,\Omega}^2. \end{aligned}$$

Combining all above estimations together gives

$$\left\| z_{\mathcal{N}}^{(2)} \right\|_{1,\Omega}^2 \gtrsim \left\| \bar{z}_{\mathcal{N}}^{(2)} \right\|_{1,\Omega}^2,$$

which yields

$$\left\| v_{\mathcal{N}}^{(2)} \right\|_{1,\Omega} = \left\| \tilde{z}_{\mathcal{N}}^{(2)} \right\|_{1,\Omega} = \left\| z_{\mathcal{N}}^{(2)} - \bar{z}_{\mathcal{N}}^{(2)} \right\|_{1,\Omega} \lesssim \left\| z_{\mathcal{N}}^{(2)} \right\|_{1,\Omega}.$$

This gives

$$\|\mathbf{v}_N\|_{1,\Omega} \lesssim \|\mathbf{z}_N\|_{1,\Omega} \lesssim M \|q_N\|_{0,\Omega}. \quad \square$$

Proof of Theorem 3.1. For all $q_N \in M_N$, let $\mathbf{v}_N \in X_N$ be the associated function given in Lemma 3.3, then

$$\begin{aligned} \frac{-(\nabla \cdot \mathbf{v}_N, q_N)_N}{\|\mathbf{v}_N\|_{1,\Omega} \|q_N\|_{0,\Omega}} &= \frac{-(\nabla \cdot \mathbf{v}_N, q_N)}{\|\mathbf{v}_N\|_{1,\Omega} \|q_N\|_{0,\Omega}} = \frac{(q_N, q_N)}{\|\mathbf{v}_N\|_{1,\Omega} \|q_N\|_{0,\Omega}} \\ &= \frac{\|q_N\|_{0,\Omega}}{\|\mathbf{v}_N\|_{1,\Omega}} \gtrsim \frac{1}{M}. \end{aligned}$$

This means (3.1) holds. \square

COROLLARY 3.1. For all $\mathbf{f} \in C^0(\Omega)^2$, problem (2.13) admits a unique solution (\mathbf{u}_N, p_N) satisfying

$$(3.13) \quad \|\mathbf{u}_N\|_{1,\Omega} + \frac{1}{M} \|p_N\|_{0,\Omega} \lesssim \|f\|_{L^\infty(\Omega)}.$$

Proof. First, it can be checked that, for each $\mathbf{u}_N \in X_N, \mathbf{v}_N \in X_N$,

$$\begin{aligned} |(\nabla \mathbf{u}_N, \nabla \mathbf{v}_N)_N| &\leq 3 \|\mathbf{u}_N\|_1 \|\mathbf{v}_N\|_1, \\ (\nabla \mathbf{v}_N, \nabla \mathbf{v}_N)_N &\geq |\mathbf{v}_N|_1^2 \geq \frac{1}{3} \|\mathbf{v}_N\|_1^2. \end{aligned}$$

Then, thanks to the above inequalities and (3.1), the well-posedness of problem (2.13) and stability estimate (3.13) are straightforward consequences of the abstract inf-sup theory (cf. [1, 6]). \square

4. Error estimation. We start with some notations and definitions which are needed in the following error analysis. Denote $\omega_r(x) = x^r e^{-x}$, $\hat{\omega}_r(x) = x^r$, and, in particular, we set $\omega(x) = \omega_0(x)$, $\hat{\omega}(x) = \hat{\omega}_0(x)$. Then, for any non-negative integer r , we define two Banach spaces

$$\begin{aligned} \hat{A}^r(R^+) &:= \{v; v \text{ is measurable on } R^+ \text{ and } \|v\|_{\hat{A}^r, R^+} < \infty\}, \\ A^r(R^+) &:= \{v; v \text{ is measurable on } R^+ \text{ and } \|v\|_{A^r, R^+} < \infty\}, \end{aligned}$$

equipped, respectively, with the following norms:

$$\begin{aligned} \|v\|_{\hat{A}^r, R^+} &= \left(\sum_{k=0}^r |v|_{\hat{A}^k, R^+}^2 \right)^{\frac{1}{2}}, \quad \text{with } |v|_{\hat{A}^k, R^+} = \|\partial_x^k v\|_{\omega_k, R^+} \quad \forall v \in \hat{A}^r(R^+), \\ \|v\|_{A^r, R^+} &= \left(\sum_{k=0}^r |v|_{A^k, R^+}^2 \right)^{\frac{1}{2}}, \quad \text{with } |v|_{A^k, R^+} = \|\partial_x^k v\|_{\hat{\omega}_k, R^+} \quad \forall v \in A^r(R^+). \end{aligned}$$

We now recall several approximation results. Let π_M^x and $\hat{\pi}_M^x$ be the projection operators defined in section 2.

LEMMA 4.1 (cf. [19]). For any $v \in \hat{A}^r(R^+)$ and integer $r \geq s \geq 0$,

$$(4.1) \quad \|v - \pi_M^x v\|_{\hat{A}^s, R^+} \lesssim M^{(s-r)/2} |v|_{\hat{A}^r, R^+}.$$

A direct consequence of Lemma 4.1 is that, for any integer $r \geq 0$, $v \in A^r(R^+)$,

$$(4.2) \quad \|v - \hat{\pi}_M^x v\|_{0, R^+} \lesssim M^{-r/2} |e^{x/2} v|_{\hat{A}^r, R^+}.$$

Denote $W_M = \{v \in \mathbb{P}_M(R^+); v(0) = 0\}$, and let $\pi_{1,M}^{x,0} : H_{0,\omega}^1(R^+) \rightarrow W_M$ be the $H_{0,\omega}^1(R^+)$ -orthogonal projection operator defined by

$$\int_0^\infty (\pi_{1,M}^{x,0} v)' \phi'_M \omega dx = \int_0^\infty v' \phi'_M \omega dx \quad \forall \phi_M \in W_M.$$

LEMMA 4.2. *If $v \in H_{0,\omega}^1(R^+)$, $\partial_x v \in \hat{A}^{r-1}(R^+)$, and integer $r \geq 1$, then*

$$\|v - \pi_{1,M}^{x,0} v\|_{1,\omega,R^+} \lesssim M^{\frac{1}{2}-\frac{r}{2}} |\partial_x v|_{\hat{A}^{r-1},R^+}.$$

Proof. Given $v \in H_{0,\omega}^1(R^+)$, let $v_M(z) = \int_0^z \pi_{M-1}^x \partial_x v(x) dx \quad \forall z \in R^+$, then $v_M \in W_M$ and $\partial_x v_M(x) = \pi_{M-1}^x(\partial_x v(x))$. Hence, by Lemma 2.2 of [11] and Lemma 4.1,

$$\begin{aligned} \|v - \pi_{1,M}^{x,0} v\|_{1,\omega,R^+} &\leq \|v - v_M\|_{1,\omega,R^+} \lesssim |v - v_M|_{1,\omega,R^+} \\ &= \|\partial_x v - \pi_{M-1}^x(\partial_x v)\|_{\omega,R^+} \lesssim M^{\frac{1}{2}-\frac{r}{2}} |\partial_x v|_{\hat{A}^{r-1},R^+}. \quad \square \end{aligned}$$

Now we set $\hat{W}_M = \{v e^{-x/2}; v \in W_M\}$ and define the projection operator $\hat{\pi}_{1,M}^{x,0}$ from $H_0^1(R^+)$ into \hat{W}_M by

$$\hat{\pi}_{1,M}^{x,0} v(x) := e^{-x/2} \pi_{1,M}^{x,0}(v(x) e^{x/2}) \quad \forall v \in H_0^1(R^+).$$

Then, it follows from Lemma 4.2 that, for $r \geq 1$,

$$(4.3) \quad \|v - \hat{\pi}_{1,M}^{x,0} v\|_{1,R^+} = \|v e^{x/2} - \pi_{1,M}^{x,0}(e^{x/2} v(x))\|_{1,\omega,R^+} \lesssim M^{\frac{1}{2}-\frac{r}{2}} |\partial_x(e^{x/2} v(x))|_{\hat{A}^{r-1},R^+}.$$

For $r \geq 1$, we introduce the space, suitable for analyzing the approximation properties of the Laguerre interpolation (cf. [19]),

$$B^r(R^+) := \{v; v \text{ is measurable on } R^+ \text{ and } \|v\|_{B^r,R^+} < \infty\},$$

with norm

$$\|v\|_{B^r,R^+} = \left(\sum_{k=1}^r \left\| x^{(r-1)/2} (x+1)^{1/2} \partial_x^k v \right\|_{0,R^+}^2 \right)^{1/2}.$$

Let I_M^x be the Laguerre–Gauss–Radau interpolation, and define $\hat{I}_M^x v(x) = e^{-x/2} I_M^x(e^{x/2} v(x))$; the following result is proved in [19].

LEMMA 4.3. *For any $v \in B^r(R^+)$, and $0 \leq \mu \leq 1 \leq r$,*

$$\begin{aligned} \|v - \hat{I}_M^x v\|_{\mu,R^+} &\lesssim (\ln M)^{1/2} M^{\mu+1/2-r/2} (|\partial_x v|_{\hat{A}^r,R^+} + |\partial_x(e^{x/2} v)|_{\hat{A}^{r-1},R^+}) \\ &\lesssim (\ln M)^{1/2} M^{\mu+1/2-r/2} \|v\|_{B^r,R^+}. \end{aligned}$$

Let $\pi_{1,N}^{y,0} : H_0^1(\Lambda) \rightarrow \mathbb{P}_N^0(\Lambda)$ be the $H_0^1(\Lambda)$ -orthogonal projector defined by

$$\int_{-1}^1 \partial_y(v - \pi_{1,N}^{y,0} v) \partial_y \phi dy = 0 \quad \forall \phi \in \mathbb{P}_N^0(\Lambda).$$

Then, it follows from [3] that, for all $s \geq 1$ and all $v \in H_0^1(\Lambda) \cap H^s(\Lambda)$, it holds that

$$(4.4) \quad \|v - \pi_{1,N}^{y,0} v\|_{k,\Lambda} \lesssim N^{k-s} \|v\|_{s,\Lambda}, \quad k = 0, 1.$$

We denote by $L^2(\Lambda, A^r(R^+))$ the space of the measurable functions $v : \Lambda \rightarrow A^r(R^+)$ such that

$$\|v\|_{A^r;0} := \left\{ \int_{\Lambda} \|v(\cdot, y)\|_{A^r, R^+}^2 dy \right\}^{\frac{1}{2}} < \infty.$$

Moreover, for any nonnegative integer s , we define

$$H^s(\Lambda, L^2(R^+)) := \left\{ v; \frac{\partial^j v}{\partial y^j} \in L^2(\Lambda, L^2(R^+)), 0 \leq j \leq s \right\}.$$

The norm of this space is given by

$$\|v\|_{0;s} = \left\{ \sum_{j=0}^s \left\| \frac{\partial^j v}{\partial y^j} \right\|_{0,\Omega}^2 \right\}^{\frac{1}{2}} = \left\{ \sum_{j=0}^s \left\| \frac{\partial^j v}{\partial y^j} \right\|_{0;0}^2 \right\}^{\frac{1}{2}}.$$

Now, for any nonnegative integer r and s , we define

$$A^{r;s}(\Omega) := H^s(\Lambda, L^2(R^+)) \cap L^2(\Lambda, A^r(R^+)),$$

with the following norm:

$$\|v\|_{A^{r;s}} = \left\{ \|v\|_{A^r;0}^2 + \|v\|_{0;s}^2 \right\}^{\frac{1}{2}} \quad \forall v \in A^{r;s}(\Omega).$$

We also define

$$\begin{aligned} \bar{B}^{r;s}(\Omega) &:= H^s(\Lambda, L^2(R^+)) \cap H^1(\Lambda, B^{r-1}(R^+)) \cap L^2(\Lambda, A^r(R^+)), \\ Y^{m;n}(\Omega) &:= H^n(\Lambda, L^2(R^+)) \cap H^1(\Lambda, A^{m-1}(R^+)) \cap H^{n-1}(\Lambda, H^1(R^+)) \\ &\quad \cap L^2(\Lambda, A^m(R^+)), \end{aligned}$$

equipped, respectively, with the following norms:

$$\begin{aligned} \|v\|_{\bar{B}^{r;s}} &= \left(\|v\|_{0;s}^2 + \|v\|_{B^{r-1};1}^2 + \|v\|_{A^r;0}^2 \right)^{\frac{1}{2}}, \\ \|v\|_{Y^{m;n}} &= \left(\|v\|_{0;n}^2 + \|v\|_{A^{m-1};1}^2 + \|v\|_{1;n-1}^2 + \|v\|_{A^m;0}^2 \right)^{\frac{1}{2}}. \end{aligned}$$

THEOREM 4.1. *If the solution (\mathbf{u}, p) of problem (2.12) satisfies $\mathbf{u} \in H_0^1(\Omega)^2 \cap Y^{m;n}(\Omega)^2 \cap C(\Omega)$, $p \in A^{m-1;n-1}(\Omega) \cap C(\Omega)$, $m \geq 1, n \geq 1$ and $\mathbf{f} \in \bar{B}^{r;s}(\Omega)^2 \cap C(\Omega)$, $r \geq 1, s \geq 1$, then the solution $(\mathbf{u}_{\mathcal{N}}, p_{\mathcal{N}})$ of (2.13) admits the following error estimates:*

$$\begin{aligned} \|\mathbf{u} - \mathbf{u}_{\mathcal{N}}\|_{1,\Omega} &\lesssim \left(M^{\frac{1}{2} - \frac{m}{2}} + N^{1-n} \right) (M \|\mathbf{u}\|_{Y^{m;n}} + \|p\|_{A^{m-1;n-1}}) \\ &\quad + \left((\ln M)^{\frac{1}{2}} M^{1 - \frac{r}{2}} + N^{-s} \right) \|\mathbf{f}\|_{\bar{B}^{r;s}}, \\ \|p - p_{\mathcal{N}}\|_{0,\Omega} &\lesssim M \left[\left(M^{\frac{1}{2} - \frac{m}{2}} + N^{1-n} \right) (M \|\mathbf{u}\|_{Y^{m;n}} + \|p\|_{A^{m-1;n-1}}) \right. \\ &\quad \left. + \left((\ln M)^{\frac{1}{2}} M^{1 - \frac{r}{2}} + N^{-s} \right) \|\mathbf{f}\|_{\bar{B}^{r;s}} \right]. \end{aligned}$$

Proof. Let

$$V_{\mathcal{N}} := \{\mathbf{v}_{\mathcal{N}} \in X_{\mathcal{N}}; (p_{\mathcal{N}}, \nabla \cdot \mathbf{v}_{\mathcal{N}})_{\mathcal{N}} = 0 \forall p_{\mathcal{N}} \in M_{\mathcal{N}}\}.$$

Then, for all $\mathbf{v}_{\mathcal{N}} \in X_{\mathcal{N}}, \mathbf{w}_{\mathcal{N}} \in V_{\mathcal{N}}$,

$$\begin{aligned} (\nabla \mathbf{w}_{\mathcal{N}}, \nabla \mathbf{v}_{\mathcal{N}}) - (\nabla \mathbf{w}_{\mathcal{N}}, \nabla \mathbf{v}_{\mathcal{N}})_{\mathcal{N}} &= (\nabla(\mathbf{w}_{\mathcal{N}} - \mathbf{u}), \nabla \mathbf{v}_{\mathcal{N}}) + (\nabla \mathbf{u}, \nabla \mathbf{v}_{\mathcal{N}}) \\ &\quad - (\nabla \mathbf{w}_{\mathcal{N}}, \nabla \mathbf{v}_{\mathcal{N}})_{\mathcal{N}} \\ &\lesssim (|\mathbf{u} - \mathbf{w}_{\mathcal{N}}|_{1,\Omega} + |\mathbf{u} - \hat{\pi}_{1,M}^{x,0} \pi_{1,N-1}^{y,0} \mathbf{u}|_{1,\Omega}) |\mathbf{v}_{\mathcal{N}}|_{1,\Omega} \\ &\quad + |\mathbf{w}_{\mathcal{N}} - \hat{\pi}_{1,M}^{x,0} \pi_{1,N-1}^{y,0} \mathbf{u}|_{1,\Omega} |\mathbf{v}_{\mathcal{N}}|_{1,\Omega} \\ &\lesssim (|\mathbf{u} - \mathbf{w}_{\mathcal{N}}|_{1,\Omega} + |\mathbf{u} - \hat{\pi}_{1,M}^{x,0} \pi_{1,N-1}^{y,0} \mathbf{u}|_{1,\Omega}) |\mathbf{v}_{\mathcal{N}}|_{1,\Omega}. \end{aligned}$$

This result, together with Theorem IV.2.5 and Remark IV.2.7 of [3], leads to

$$\begin{aligned} |\mathbf{u} - \mathbf{u}_{\mathcal{N}}|_{1,\Omega} &\lesssim \inf_{\mathbf{w}_{\mathcal{N}} \in V_{\mathcal{N}}} \left(|\mathbf{u} - \mathbf{w}_{\mathcal{N}}|_{1,\Omega} + \sup_{\mathbf{v}_{\mathcal{N}} \in X_{\mathcal{N}}} \frac{(\nabla \mathbf{w}_{\mathcal{N}}, \nabla \mathbf{v}_{\mathcal{N}}) - (\nabla \mathbf{w}_{\mathcal{N}}, \nabla \mathbf{v}_{\mathcal{N}})_{\mathcal{N}}}{|\mathbf{v}_{\mathcal{N}}|_{1,\Omega}} \right) \\ &\quad + \inf_{q_{\mathcal{N}} \in M_{\mathcal{N}}} \|p - q_{\mathcal{N}}\|_{0,\Omega} + \sup_{\mathbf{v}_{\mathcal{N}} \in X_{\mathcal{N}}} \frac{(\mathbf{f}, \mathbf{v}_{\mathcal{N}})_{\mathcal{N}} - (\mathbf{f}, \mathbf{v}_{\mathcal{N}})}{|\mathbf{v}_{\mathcal{N}}|_{1,\Omega}} \\ &\lesssim \inf_{\mathbf{w}_{\mathcal{N}} \in V_{\mathcal{N}}} |\mathbf{u} - \mathbf{w}_{\mathcal{N}}|_{1,\Omega} + |\mathbf{u} - \hat{\pi}_{1,M}^{x,0} \pi_{1,N-1}^{y,0} \mathbf{u}|_{1,\Omega} + \inf_{q_{\mathcal{N}} \in M_{\mathcal{N}}} \|p - q_{\mathcal{N}}\|_{0,\Omega} \\ &\quad + \sup_{\mathbf{v}_{\mathcal{N}} \in X_{\mathcal{N}}} \frac{(\mathbf{f}, \mathbf{v}_{\mathcal{N}})_{\mathcal{N}} - (\mathbf{f}, \mathbf{v}_{\mathcal{N}})}{|\mathbf{v}_{\mathcal{N}}|_{1,\Omega}} \\ &\lesssim M \inf_{\mathbf{v}_{\mathcal{N}} \in X_{\mathcal{N}}} |\mathbf{u} - \mathbf{v}_{\mathcal{N}}|_{1,\Omega} + |\mathbf{u} - \hat{\pi}_{1,M}^{x,0} \pi_{1,N-1}^{y,0} \mathbf{u}|_{1,\Omega} \\ &\quad + \inf_{q_{\mathcal{N}} \in M_{\mathcal{N}}} \|p - q_{\mathcal{N}}\|_{0,\Omega} + \sup_{\mathbf{v}_{\mathcal{N}} \in X_{\mathcal{N}}} \frac{(\mathbf{f}, \mathbf{v}_{\mathcal{N}})_{\mathcal{N}} - (\mathbf{f}, \mathbf{v}_{\mathcal{N}})}{|\mathbf{v}_{\mathcal{N}}|_{1,\Omega}}. \end{aligned}$$

And,

$$\begin{aligned} \|p - p_{\mathcal{N}}\|_{0,\Omega} &\lesssim M \left[|\mathbf{u} - \hat{\pi}_{1,M}^{x,0} \pi_{1,N-1}^{y,0} \mathbf{u}|_{1,\Omega} + \inf_{q_{\mathcal{N}} \in M_{\mathcal{N}}} \|p - q_{\mathcal{N}}\|_{0,\Omega} \right. \\ &\quad \left. + M \inf_{\mathbf{v}_{\mathcal{N}} \in X_{\mathcal{N}}} |\mathbf{u} - \mathbf{v}_{\mathcal{N}}|_{1,\Omega} + \sup_{\mathbf{v}_{\mathcal{N}} \in X_{\mathcal{N}}} \frac{(\mathbf{f}, \mathbf{v}_{\mathcal{N}})_{\mathcal{N}} - (\mathbf{f}, \mathbf{v}_{\mathcal{N}})}{|\mathbf{v}_{\mathcal{N}}|_{1,\Omega}} \right]. \end{aligned}$$

Now, for all $\mathbf{f}_{M,N-1} \in \hat{\mathbb{P}}_{M,N-1}(\Omega)^2$, we have

$$(\mathbf{f}, \mathbf{v}_{\mathcal{N}})_{\mathcal{N}} - (\mathbf{f}, \mathbf{v}_{\mathcal{N}}) \lesssim (\|\mathbf{f} - I_N^y \hat{I}_M^x \mathbf{f}\|_{0,\Omega} + \|\mathbf{f} - \mathbf{f}_{M,N-1}\|_{0,\Omega}) |\mathbf{v}_{\mathcal{N}}|_{1,\Omega}.$$

We know from the interpolation results of I_N^y and \hat{I}_M^x that

$$\begin{aligned} \|\mathbf{f} - I_N^y \hat{I}_M^x \mathbf{f}\|_{0,\Omega} &\leq \|\mathbf{f} - I_N^y \mathbf{f}\|_{0;0} + \|I_N^y (\mathbf{f} - \hat{I}_M^x \mathbf{f})\|_{0;0} \\ &\lesssim N^{-s} \|\mathbf{f}\|_{0;s} + \|\mathbf{f} - \hat{I}_M^x \mathbf{f}\|_{0;1} \\ &\lesssim N^{-s} \|\mathbf{f}\|_{0;s} + (\ln M)^{\frac{1}{2}} M^{1-\frac{s}{2}} \|\mathbf{f}\|_{B^{r-1;1}}. \end{aligned}$$

Furthermore,

$$\|\mathbf{f} - \mathbf{f}_{M,N-1}\|_{0,\Omega} \leq \|\mathbf{f} - \pi_{N-1}^y \circ \hat{\pi}_M^x \mathbf{f}\|_{0;0} \lesssim N^{-s} \|\mathbf{f}\|_{0;s} + M^{-\frac{s}{2}} \|\mathbf{f}\|_{A^r;0}.$$

Combining the above two inequalities, we get

$$(\mathbf{f}, \mathbf{v}_{\mathcal{N}})_{\mathcal{N}} - (\mathbf{f}, \mathbf{v}_{\mathcal{N}}) \lesssim \left((\ln M)^{\frac{1}{2}} M^{1-\frac{r}{2}} + N^{-s} \right) \|\mathbf{f}\|_{\bar{B}^{r;s}} |\mathbf{v}_{\mathcal{N}}|_{1,\Omega}.$$

Now we estimate $\inf_{\mathbf{v}_{\mathcal{N}} \in X_{\mathcal{N}}} |\mathbf{u} - \mathbf{v}_{\mathcal{N}}|_{1,\Omega}$. Since

$$|\mathbf{u} - \mathbf{v}_{\mathcal{N}}|_{1,\Omega} = \left\| \frac{\partial}{\partial x} (\mathbf{u} - \mathbf{v}_{\mathcal{N}}) \right\|_{0,\Omega} + \left\| \frac{\partial}{\partial y} (\mathbf{u} - \mathbf{v}_{\mathcal{N}}) \right\|_{0,\Omega},$$

by choosing $\mathbf{v}_{\mathcal{N}} = \hat{\pi}_{1,M}^{x,0} \pi_{1,N}^{y,0} \mathbf{u}$, we know from the approximation results of $\hat{\pi}_{1,M}^{x,0}$ and $\pi_{1,N}^{y,0}$ that

$$\begin{aligned} \left\| \frac{\partial}{\partial x} (\mathbf{u} - \mathbf{v}_{\mathcal{N}}) \right\|_{0,\Omega} &\leq \left\| \frac{\partial}{\partial x} (\mathbf{u} - \hat{\pi}_{1,M}^{x,0} \mathbf{u}) \right\|_{0,\Omega} + \left\| \frac{\partial}{\partial x} \hat{\pi}_{1,M}^{x,0} (\mathbf{u} - \pi_{1,N}^{y,0} \mathbf{u}) \right\|_{0,\Omega} \\ &\lesssim M^{\frac{1}{2}-\frac{m}{2}} |\partial_x (e^{x/2} \mathbf{u})|_{\hat{A}^{m-1;0}} + N^{1-n} \left\| \frac{\partial}{\partial x} \hat{\pi}_{1,M}^{x,0} \mathbf{u} \right\|_{0;n-1} \\ &\lesssim M^{\frac{1}{2}-\frac{m}{2}} \|\mathbf{u}\|_{A^{m;0}} + N^{1-n} \left\| \frac{\partial}{\partial x} (e^{x/2} \mathbf{u}) \right\|_{\omega_0;n-1} \\ &\lesssim M^{\frac{1}{2}-\frac{m}{2}} \|\mathbf{u}\|_{A^{m;0}} + N^{1-n} \|\mathbf{u}\|_{1;n-1}; \\ \left\| \frac{\partial}{\partial y} (\mathbf{u} - \mathbf{v}_{\mathcal{N}}) \right\|_{0,\Omega} &\leq \left\| \frac{\partial}{\partial y} (\mathbf{u} - \pi_{1,N}^{y,0} \mathbf{u}) \right\|_{0,\Omega} + \left\| \frac{\partial}{\partial y} \pi_{1,N}^{y,0} (\mathbf{u} - \hat{\pi}_{1,M}^{x,0} \mathbf{u}) \right\|_{0,\Omega} \\ &\lesssim N^{1-n} \|\mathbf{u}\|_{0;n} + M^{\frac{1}{2}-\frac{m}{2}} \left| e^{x/2} \frac{\partial}{\partial y} \pi_{1,N-1}^{y,0} \mathbf{u} \right|_{\hat{A}^{m-1;0}} \\ &\lesssim N^{1-n} \|\mathbf{u}\|_{0;n} + M^{\frac{1}{2}-\frac{m}{2}} \|\mathbf{u}\|_{A^{m-1;1}}. \end{aligned}$$

Combining the above two results leads to

$$\inf_{\mathbf{v}_{\mathcal{N}} \in X_{\mathcal{N}}} |\mathbf{u} - \mathbf{v}_{\mathcal{N}}|_{1,\Omega} \leq |\mathbf{u} - \hat{\pi}_{1,M}^{x,0} \pi_{1,N}^{y,0} \mathbf{u}|_{1,\Omega} \lesssim (M^{\frac{1}{2}-\frac{m}{2}} + N^{1-n}) \|\mathbf{u}\|_{Y^{m;n}}.$$

Similarly, we have

$$|\mathbf{u} - \hat{\pi}_{1,M}^{x,0} \pi_{1,N-1}^{y,0} \mathbf{u}|_{1,\Omega} \lesssim (M^{\frac{1}{2}-\frac{m}{2}} + N^{1-n}) \|\mathbf{u}\|_{Y^{m;n}}.$$

Now it remains to estimate $\inf_{q_{\mathcal{N}} \in M_{\mathcal{N}}} \|p - q_{\mathcal{N}}\|_{0,\Omega}$. By using the known properties of the projectors $\hat{\pi}_M^x$ and π_N^y in [16], it follows that

$$\begin{aligned} \inf_{q_{\mathcal{N}} \in M_{\mathcal{N}}} \|p - q_{\mathcal{N}}\|_{0,\Omega} &\leq \|p - \pi_{N-2}^y \circ \hat{\pi}_{M-1}^x p\|_{0;0} \\ &\lesssim (N-2)^{1-n} \|p\|_{0;n-1} + \|p - \hat{\pi}_{M-1}^x p\|_{0;0} \\ &\lesssim N^{1-n} \|p\|_{0;n-1} + M^{\frac{1}{2}-\frac{m}{2}} \|p\|_{A^{m-1;0}} \\ &\lesssim (M^{\frac{1}{2}-\frac{m}{2}} + N^{1-n}) \|p\|_{A^{m-1;n-1}}. \end{aligned}$$

As a direct consequence of the above estimates, we finally obtain

$$\begin{aligned} \|\mathbf{u} - \mathbf{u}_{\mathcal{N}}\|_{1,\Omega} &\lesssim (M^{\frac{1}{2}-\frac{m}{2}} + N^{1-n}) (M \|\mathbf{u}\|_{Y^{m;n}} + \|p\|_{A^{m-1;n-1}}) \\ &\quad + ((\ln M)^{\frac{1}{2}} M^{1-\frac{r}{2}} + N^{-s}) \|\mathbf{f}\|_{\bar{B}^{r;s}}, \\ \|p - p_{\mathcal{N}}\|_{0,\Omega} &\lesssim M \left[(M^{\frac{1}{2}-\frac{m}{2}} + N^{1-n}) (M \|\mathbf{u}\|_{Y^{m;n}} + \|p\|_{A^{m-1;n-1}}) \right. \\ &\quad \left. + ((\ln M)^{\frac{1}{2}} M^{1-\frac{r}{2}} + N^{-s}) \|\mathbf{f}\|_{\bar{B}^{r;s}} \right]. \quad \square \end{aligned}$$

5. Numerical results and discussions. We start with some implementation details. Let $\mathbf{u}_N = (u_N^1, u_N^2)^t$, and we write

$$u_N^r(x, y) = \sum_{j=1}^{N-1} \sum_{i=1}^M u_N^r(\hat{\xi}_i, \xi_j) \hat{h}_i(x) h_j(y), \quad r = 1, 2,$$

where $h_j \in \mathbb{P}_N(\Lambda)$ ($0 \leq j \leq N$) are the Legendre–Gauss–Lobatto interpolants satisfying $h_j(\xi_q) = \delta_{qj}$, while $\hat{h}_i \in \hat{\mathbb{P}}_M(R^+)$ ($0 \leq i \leq M$) are the Laguerre–Gauss–Radau interpolants satisfying $\hat{h}_i(\hat{\xi}_q) = \delta_{qi}$. We use $\underline{\mathbf{u}}_N$ to denote the vector consisting of the values of \mathbf{u}_N at the nodes $(\hat{\xi}_i, \xi_j)_{1 \leq i \leq M, 1 \leq j \leq N-1}$.

Similarly, we write

$$p_N(x, y) = \sum_{j=1}^{N-1} \sum_{i=1}^M p_N(\hat{\zeta}_i, \zeta_j) \hat{\ell}_i(x) \ell_j(y),$$

where $(\hat{\zeta}_i)_{1 \leq i \leq M}$ and $(\zeta_j)_{1 \leq j \leq N-1}$ are, respectively, the Laguerre–Gauss and Legendre–Gauss points, and $\ell_j \in \mathbb{P}_{N-2}(\Lambda)$ ($1 \leq j \leq N-1$) are the Legendre–Gauss interpolants satisfying $\ell_j(\xi_q) = \delta_{qj}$, while $\hat{\ell}_i \in \hat{\mathbb{P}}_{M-1}(R^+)$ ($1 \leq i \leq M$) are the Laguerre–Gauss interpolants satisfying $\hat{\ell}_i(\hat{\zeta}_q) = \delta_{qi}$. We use $\underline{\mathbf{p}}_N$ to denote the vector consisting of the values of p_N at the nodes $(\hat{\zeta}_i, \zeta_j)_{1 \leq i \leq M, 1 \leq j \leq N-1}$.

Inserting the expansions of \mathbf{u}_N and p_N into (2.13), the resulting set of algebraic equations can be written under a matrix form:

$$(5.1) \quad \mathbf{A}_N \underline{\mathbf{u}}_N + \mathbf{D}_N \underline{\mathbf{p}}_N = \mathbf{B}_N \underline{\mathbf{f}}_N,$$

$$(5.2) \quad \mathbf{D}_N^T \underline{\mathbf{u}}_N = 0,$$

where $\underline{\mathbf{f}}_N$ is a vector representation of the \mathbf{f} at the nodes $(\hat{\xi}_i, \xi_j)$. The matrices \mathbf{A}_N , \mathbf{D}_N , and \mathbf{B}_N are block-diagonal matrices with 2 blocks each. The blocks of \mathbf{A}_N are the discrete Laplace operators, and those of \mathbf{D}_N are associated to the different components of the discrete gradient operators, while blocks of \mathbf{B}_N are the mass matrices with respect to each component of \mathbf{f} .

Eliminating $\underline{\mathbf{u}}_N$ from (5.1)–(5.2), we obtain

$$(5.3) \quad \underbrace{\mathbf{D}_N^T \mathbf{A}_N^{-1} \mathbf{D}_N}_{\mathbf{S}_N} \underline{\mathbf{p}}_N = \mathbf{D}_N^T \mathbf{A}_N^{-1} \mathbf{B}_N \underline{\mathbf{f}}_N.$$

The matrix $\mathbf{S}_N := \mathbf{D}_N^T \mathbf{A}_N^{-1} \mathbf{D}_N$ is usually referred as the Uzawa matrix. A typical procedure for solving (5.1)–(5.2) is to first solve $\underline{\mathbf{p}}_N$ from (5.3) and then solve $\underline{\mathbf{u}}_N$ from the Poisson equation (5.1) with known $\underline{\mathbf{p}}_N$.

The Uzawa matrix is of dimension $M \times (N-1)$, full, symmetric, and semidefinite. A usual procedure is to use a preconditioned conjugate gradient procedure with the Gauss mass matrix $\tilde{\mathbf{B}}_N$ as a preconditioner [3, 7, 9, 14]. Each outer iteration requires the inversion of two Laplace operators (\mathbf{A}_N matrix), which can be carried out by the fast diagonalization method (see [13]). Hence, the efficiency of the method is dictated by the condition number κ_N of $\mathbf{B}_N^{-1} \mathbf{S}_N$. Another important consequence of the inf-sup constant is that $\kappa_N = \frac{1}{\beta_N^2}$ [14].

The first computational investigation is concerned with the sharpness of the lower bound on the inf-sup constant derived in section 3. In the left of Figure 1, we plot

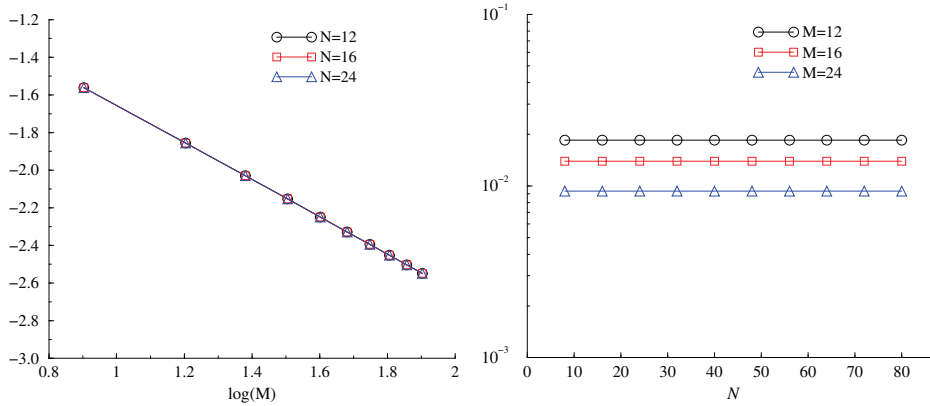


FIG. 1. Left: inf-sup constant β_N vs. M in log-log scale; right: inf-sup constant β_N vs. N .

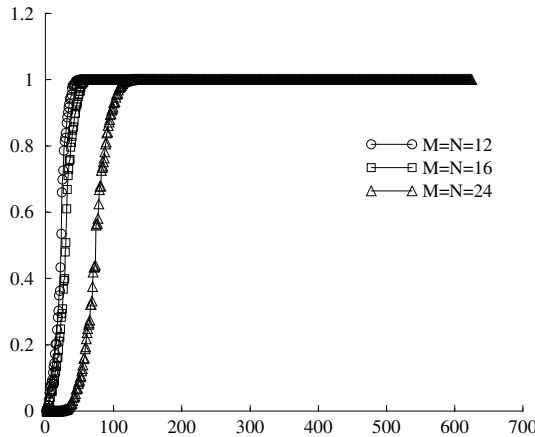


FIG. 2. Spectra of the Uzawa operator for three different values of N with $M = N$.

the variations of β_N versus (vs.) M in log-log scale for several N . We observe that β_N is independent of N while it decays as $\frac{1}{M}$. In the right of Figure 1, we plot the variations of β_N vs. N for several M . We observe that β_N remains to be constant as we vary N with M fixed. These results are fully consistent with Theorem 3.1, indicating that our estimate for the inf-sup constant is sharp.

In view of inverting the Uzawa operator, the knowledge of the eigenvalues' distribution of the matrix $\mathbf{B}_N^{-1}\mathbf{S}_N$ may help to design adapted preconditioners for (5.3). The efficiency of the iterating methods depends on how the preconditioners affect the eigenvalues of \mathbf{S}_N . In Figure 2 we plot all of the eigenvalues of $\tilde{\mathbf{B}}_N^{-1}\mathbf{S}_N$ for some values of $M = N \in \{12, 16, 24\}$.

The first feature of the spectra is the similarity of their distribution for different values of N with $M = N$. Another interesting aspect is a strong concentration of the eigenvalues around the largest value 1. It is known that this type of clustering is very advantageous for the conjugate gradient iteration since the contribution of the eigenspaces associated with a given multiple eigenvalue is resolved in only one iteration.

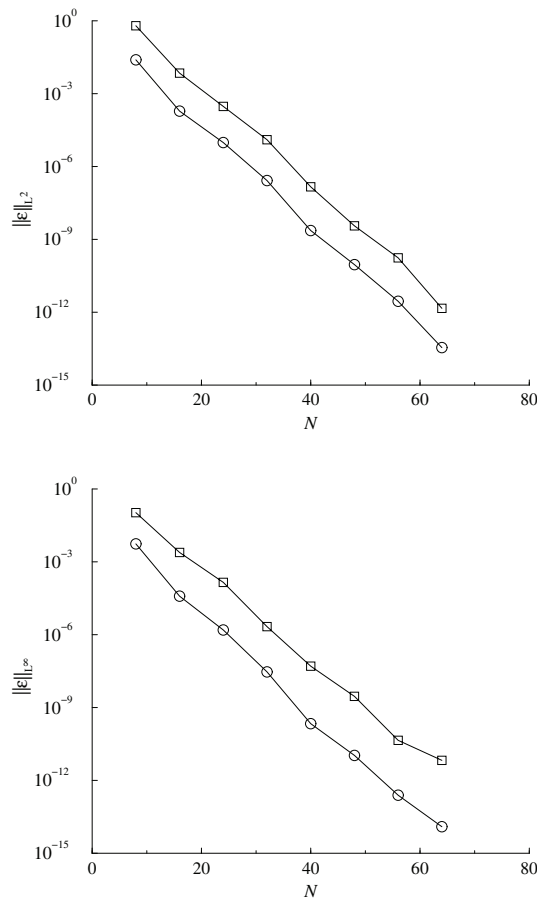


FIG. 3. The velocity (\circ) and pressure (\square) errors as a function of N with $M = N$: left, in L^2 norm; right, in L^∞ norm.

We now present some numerical tests to validate the error estimates. We consider the Stokes problem with the following analytical solution:

$$\mathbf{u} = \begin{pmatrix} \sin(x) \cos(y) e^{-x} \\ (\sin(x) - \cos(x)) \sin(y) e^{-x} \end{pmatrix}, \quad p = \cos(x) \cos(y) e^{-x}.$$

In Figure 3, we plot, in a semilogarithmic scale, the L^2 -velocity and the L^2 -pressure errors (top figure), and the L^∞ -velocity and the L^∞ -pressure errors (bottom figure) with respect to N with $M = N$. We observe that the errors converge exponentially, which is a typical behavior for spectral methods with analytical solutions.

Finally, in order to justify the use of compatible discrete velocity and pressure spaces, we show via a simple test that the equal-order velocity-pressure approximation $\mathbb{P}_{M,N}(\Omega)^2 \times \mathbb{P}_{M,N}(\Omega)$ is ill-posed. In Figure 4, we present the velocity and the pressure errors in the L^2 -norm as a function of N with $M = N$. Obviously the pressure fails to converge when the polynomial degree increases. The reason for this failure is that there are spurious pressure modes in the pressure space, similar to the well-known case of the Legendre–Legendre $P_N^2 - P_N$ method for the Stokes problem in a rectangular domain.

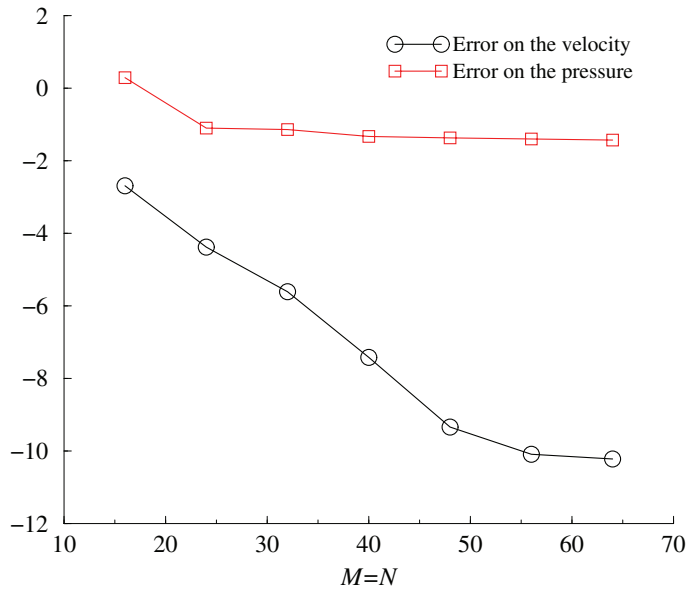


FIG. 4. The velocity and pressure errors as a function of $N(M = N)$ by the incompatible $\mathbb{P}_{M,N}(\Omega)^2 \times \mathbb{P}_{M,N}(\Omega)$ method.

In summary, we have presented a mixed Laguerre–Legendre spectral method for the Stokes problem on a semi-infinite channel. We established the well-posedness of this method by deriving a lower bound on the inf-sup constant and presented numerical results which indicated that the derived lower bound is sharp. We have also derived error estimates by using the inf-sup condition and the Laguerre and Legendre approximation properties.

REFERENCES

- [1] I. BABUŠKA, *The finite element method with Lagrangian multipliers*, Numer. Math., 20 (1972/73), pp. 179–192.
- [2] C. BERNARDI, M. DAUGE, AND Y. MADAY, *Polynomial in the Sobolev World*, preprint, Laboratoire Jacques-Louis Lions, Paris, 2003, <http://www.ann.jussieu.fr/publications/2003/R03038.html>.
- [3] C. BERNARDI AND Y. MADAY, *Approximations Spectrales de Problèmes aux Limites Elliptiques*, Springer-Verlag, Paris, 1992.
- [4] C. BERNARDI AND Y. MADAY, *Spectral method*, in Handb. Numer. Anal. 5 (Part 2), P. G. Ciarlet and L. L. Lions, eds., North-Holland, Amsterdam, 1997.
- [5] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
- [6] F. BREZZI, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers*, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge, 8 (1974), pp. 129–151.
- [7] C. CANUTO, M. Y. HUSSAINI, A. QUARTERONI, AND T. A. ZANG, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, Berlin, 1987.
- [8] O. COULAUD, D. FUNARO, AND O. KAVIAN, *Laguerre spectral approximation of elliptic problems in exterior domains*, Comput. Methods Appl. Mech. Engrg., 80 (1990), pp. 451–458.
- [9] M. O. DEVILLE, P. F. FISCHER, AND E. H. MUND, *High-order Methods for Incompressible Fluid Flow*, Camb. Monogr. Appl. Comput. Math. 9, Cambridge University Press, Cambridge, 2002.
- [10] V. GIRAULT AND P. A. RAVIART, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.

- [11] B. GUO AND J. SHEN, *Laguerre-Galerkin method for nonlinear partial differential equations on a semi-infinite interval*, Numer. Math., 86 (2000), pp. 635–654.
- [12] V. IRANZO AND A. FALQUÉS, *Some spectral approximations for differential equations in unbounded domains*, Comput. Methods Appl. Mech. Engrg., 98 (1992), pp. 105–126.
- [13] R. E. LYNCH, J. R. RICE, AND D. H. THOMAS, *Direct solution of partial differential equations by tensor product methods*, Numer. Math., 6 (1964), pp. 185–199.
- [14] Y. MADAY, D. MEIRON, A. T. PATERA, AND E. M. RÖNQUIST, *Analysis of iterative methods for the steady and unsteady Stokes problem: Application to spectral element discretizations*, SIAM J. Sci. Comput., 14 (1993), pp. 310–337.
- [15] H.-P. MA AND B.-Y. GUO, *Composite Legendre-Laguerre pseudospectral approximation in unbounded domains*, IMA J. Numer. Anal., 21 (2001), pp. 587–602.
- [16] A. QUARTERONI AND A. VALLI, *Numerical Approximation of Partial Differential Equations*, Springer Ser. Comput. Math. 23, Springer-Verlag, Berlin, 1994.
- [17] R. L. SANI AND P. M. GRESHO, *Résumé and remarks on the open boundary condition minisymposium*, Internat. J. Numer. Methods Fluids, 18 (1994), pp. 983–1008.
- [18] J. SHEN, *A new fast Chebyshev-Fourier algorithm for the Poisson-type equations in polar geometries*, Appl. Numer. Math., 33 (2000), pp. 183–190.
- [19] L. WANG AND B. GUO, *Modified Laguerre pseudospectral method refined by multidomain Legendre pseudospectral approximation*, J. Comput. Appl. Math., 190 (2006), pp. 304–324.

OPTIMAL IMPORTANCE SAMPLING PARAMETER SEARCH FOR LÉVY PROCESSES VIA STOCHASTIC APPROXIMATION*

REIICHIRO KAWAI†

Abstract. The author proposes stochastic approximation methods of finding the optimal measure change by the exponential tilting for Lévy processes in Monte Carlo importance sampling variance reduction. In accordance with the structure of the underlying Lévy measure, either a constrained or unconstrained algorithm of the stochastic approximation is chosen. For both cases, the almost sure convergence to a unique stationary point is proved. Numerical examples are presented to illustrate the effectiveness of our method.

Key words. Esscher transform, Girsanov theorem, Monte Carlo simulation, infinitely divisible distribution, stochastic approximation algorithm, variance reduction

AMS subject classifications. 65C05, 62L20, 60E07, 60G51

DOI. 10.1137/070680564

1. Introduction. The importance sampling method is aimed at reducing the variance of independently and identically distributed (i.i.d.) Monte Carlo summands by appropriately transforming the underlying probability measure, from which interested random variables or stochastic processes are generated, so as to put more weight on important events and less on undesirable ones. Due to its practical effectiveness, it has long been thought of as one of the most important variance reduction methods in the Monte Carlo simulation and has been intensively studied with a view towards a wide range of applications, such as mathematical finance, queueing theory, and sequential analysis, to mention just a few. For its principle with some numerical examples, see, for instance, section 4.6 of Glasserman [7].

In the importance sampling “variance” reduction, the optimal measure change means nothing but the one attaining the minimal variance of i.i.d. Monte Carlo summands. In the Gaussian framework, the Girsanov measure change is often indexed by a single parameter, that is, the drift parameter, and several attempts have been made to find its optimum. In financial applications, for example, Glasserman, Heidelberger, and Shahabuddin [6] propose an optimization procedure to find a nearly optimal measure change in pricing financial derivatives, while Su and Fu [14] and Arouna [1] apply the stochastic approximation so as to search for the root of the gradient of the Monte Carlo variance with respect to the measure change parameter.

The aim of the present work is to apply the idea of [1, 14] to Lévy processes without the Brownian motion, or equivalently after discretization, infinitely divisible laws without Gaussian component. In general, the measure change for Lévy processes involves every single jump, which forms the sample paths. (See section 33 of Sato [12] for details. For an importance sampling method with such intricate measure changes, see Kawai [8].) In this paper, we, however, restrict our attention to the simplest measure change, often called the Esscher transform, which has only to look at the terminal marginals. The Esscher transform is nothing but the well-known exponential

*Received by the editors January 22, 2007; accepted for publication (in revised form) August 27, 2008; published electronically November 21, 2008.

<http://www.siam.org/journals/sinum/47-1/68056.html>

†Center for the Study of Finance and Insurance, Osaka University, Toyonaka, 560-8531, Japan (reiichiro_kawai@ybb.ne.jp).

tilting of laws and is thus indexed by a single (multidimensional) parameter. As we will investigate later, a crucial difficulty in the case of Lévy processes without Gaussian component is that, depending on the structure of the underlying Lévy measure, the exponential tilting parameter might have to stay in a suitable compact set, while the drift parameter of the Gaussian distribution may be arbitrarily taken.

The rest of the paper is organized as follows. Section 2 recalls the Esscher transform and the principle of the importance sampling variance reduction, and constructs the basis of our analysis. In section 3, the almost sure convergence of the stochastic approximation is proved separately for the constrained and unconstrained algorithms, depending on the structure of the underlying Lévy measure. Section 4 illustrates the effectiveness of our method via numerical examples for both constrained and unconstrained stochastic approximation algorithms. Finally, section 5 concludes.

2. Preliminaries. Let us begin with some notations which will be used throughout the text. \mathbb{N} is the collection of all positive numbers, with $\mathbb{N}_0 := \{0\} \cup \mathbb{N}$. \mathbb{R}^d is the d -dimensional Euclidean space with the norm $\|\cdot\|$ and the inner product $\langle \cdot, \cdot \rangle$, $\mathbb{R}_0^d := \mathbb{R}^d \setminus \{0\}$ and $\mathcal{B}(\mathbb{R}_0^d)$ is the Borel σ -field of \mathbb{R}_0^d . $(\Omega, \mathcal{F}, \mathbb{P})$ is our underlying probability space. $\text{Leb}(\cdot)$ denotes the Lebesgue measure, while $\mathbb{P}|_{\mathcal{F}_t}$ is the restriction of a probability measure \mathbb{P} to the σ -field \mathcal{F}_t . Denote by ∇ the gradient, and by $\text{Hess}[\cdot]$ the Hessian matrix. The interval $(0, -1]$ is understood to be $[-1, 0)$. The expression $f(x) \sim g(x)$ means $f(x)/g(x)$ tends to 1. The identity in law is denoted by $\stackrel{L}{\sim}$. We say that a stochastic process $\{X_t : t \geq 0\}$ in \mathbb{R}^d is a Lévy process if it has independent and stationary increments, if it is continuous in probability, and if $X_0 = 0$, *a.s.* By the Lévy–Khinchine representation theorem, the characteristic function of its marginal law is uniquely given by

$$\mathbb{E} \left[e^{i\langle y, X_t \rangle} \right] = \exp \left[t \left(i\langle y, \gamma \rangle - \frac{1}{2} \langle y, Ay \rangle + \int_{\mathbb{R}_0^d} \left(e^{i\langle y, z \rangle} - 1 - i\langle y, z \rangle \mathbb{1}_{(0,1]}(\|z\|) \right) \nu(dz) \right) \right],$$

where $\gamma \in \mathbb{R}^d$, A is a symmetric nonnegative-definite $d \times d$ matrix, and ν is a Lévy measure on \mathbb{R}_0^d , that is, $\int_{\mathbb{R}_0^d} (\|z\|^2 \wedge 1) \nu(dz) < +\infty$. If the above holds, then we say that the Lévy process $\{X_t : t \geq 0\}$ is generated by the triplet (γ, A, ν) . *In this paper, we restrict our attention to pure-jump Lévy processes, that is, we set $A \equiv 0$ throughout. Moreover, we also assume that all components are nondegenerate.* A function $f : \mathbb{R}^d \mapsto [0, \infty)$ is said to be submultiplicative if there exists a positive constant a such that $f(x + y) \leq af(x)f(y)$ for $x, y \in \mathbb{R}^d$. Letting $c \in \mathbb{R}$, $\gamma \in \mathbb{R}^d$, and $b > 0$, if $f(x)$ is submultiplicative on \mathbb{R}^d , then $f(cx + \gamma)^b$ is submultiplicative, and the functions $\|x\| \vee 1$, $e^{\langle c, x \rangle}$ are submultiplicative, and a product of two submultiplicative functions is submultiplicative. We recall an important moment property of Lévy processes, which will be used often in what follows.

THEOREM 2.1 (Sato [12], Theorem 25.3). *Let f be a submultiplicative, locally bounded, measurable function on \mathbb{R}^d , and let $\{X_t : t \geq 0\}$ be a Lévy process in \mathbb{R}^d with Lévy measure ν . Then, $\mathbb{E}[f(X_t)]$ is finite for every $t > 0$ if and only if $\int_{\|z\|>1} f(z)\nu(dz) < +\infty$.*

2.1. Esscher transform. Among the density transformations of Lévy processes, there is a simple class ending up with looking only at the marginals, which is built via the exponential tilting. The class is often called the Esscher transform in mathematical finance and actuarial science. Let $\{X_t : t \geq 0\}$ be a Lévy process in \mathbb{R}^d generated by

$(\gamma, 0, \nu)$, and let $(\mathcal{F}_t)_{t \geq 0}$ be the natural filtration of $\{X_t : t \geq 0\}$. Define

$$\Lambda_1 := \left\{ \lambda \in \mathbb{R}^d : \mathbb{E}_{\mathbb{P}} \left[e^{\langle \lambda, X_1 \rangle} \right] < +\infty \right\} = \left\{ \lambda \in \mathbb{R}^d : \int_{\|z\| > 1} e^{\langle \lambda, z \rangle} \nu(dz) < +\infty \right\},$$

where the second equality holds by Theorem 2.1. We impose the condition $\text{Leb}(\Lambda_1) > 0$ throughout. Clearly, the set Λ_1 contains the origin and is convex. For $\lambda \in \Lambda_1$, we denote by φ the cumulant generating function of the marginal law at unit time of $\{X_t : t \geq 0\}$ under the probability measure \mathbb{P} ; that is, $\varphi(\lambda) := \ln \mathbb{E}_{\mathbb{P}}[e^{\langle \lambda, X_1 \rangle}]$. For ease in notation, we also write $\varphi_t(\lambda) := \ln \mathbb{E}_{\mathbb{P}}[e^{\langle \lambda, X_t \rangle}]$, $t > 0$, in view of

$$\varphi_t(\lambda) = \ln \mathbb{E}_{\mathbb{P}} \left[e^{\langle \lambda, X_t \rangle} \right] = t \ln \mathbb{E}_{\mathbb{P}} \left[e^{\langle \lambda, X_1 \rangle} \right] = t\varphi(\lambda),$$

where the second equality holds by the infinite divisibility of the marginal laws of Lévy processes. Note that $\varphi(\lambda)$ is continuous and $\nabla\varphi(\lambda)$ is well defined in $\lambda \in \Lambda_1$. Under the probability measure \mathbb{Q}_λ , where $\lambda \in \Lambda_1$ and which is defined via the Radon–Nikodym derivative, for every $t \in (0, +\infty)$,

$$\left. \frac{d\mathbb{Q}_\lambda}{d\mathbb{P}} \right|_{\mathcal{F}_t} = \frac{e^{\langle \lambda, X_t \rangle}}{\mathbb{E}_{\mathbb{P}} \left[e^{\langle \lambda, X_t \rangle} \right]} = e^{\langle \lambda, X_t \rangle - \varphi_t(\lambda)}, \quad \mathbb{P}\text{-a.s.},$$

the stochastic process $\{X_t : t \geq 0\}$ is again a Lévy process generated by $(\gamma_\lambda, 0, \nu_\lambda)$, where $\gamma_\lambda = \gamma + \int_{\|z\| \leq 1} z(\nu_\lambda - \nu)(dz)$, and

$$(2.1) \quad \nu_\lambda(dz) = e^{\langle \lambda, z \rangle} \nu(dz).$$

Then, the probability measures $\mathbb{P}|_{\mathcal{F}_t}$ and $\mathbb{Q}_\lambda|_{\mathcal{F}_t}$ are mutually absolutely continuous for every $t \in (0, +\infty)$. We also have $\mathbb{E}_{\mathbb{Q}_\lambda} [e^{-\langle \lambda, X_1 \rangle}] < +\infty$, and

$$\left. \frac{d\mathbb{P}}{d\mathbb{Q}_\lambda} \right|_{\mathcal{F}_t} = \left(\left. \frac{d\mathbb{Q}_\lambda}{d\mathbb{P}} \right|_{\mathcal{F}_t} \right)^{-1} = e^{-\langle \lambda, X_t \rangle + \varphi_t(\lambda)}, \quad \mathbb{Q}_\lambda\text{-a.s.}$$

For $t > 0$, let p be a probability density function on \mathbb{R}^d of the random vector X_t under \mathbb{P} , provided that it is well defined. Then, a density function p_λ of X_t under \mathbb{Q}_λ is given by

$$(2.2) \quad p_\lambda(x) = e^{\langle \lambda, x \rangle - \varphi_t(\lambda)} p(x), \quad x \in \mathbb{R}^d.$$

2.2. Importance sampling variance reduction. Suppose we are interested in evaluating

$$C := \mathbb{E}_{\mathbb{P}}[F(X)]$$

by Monte Carlo simulation, where $F(X) := F(\{X_t : t \in [0, T]\}) \in L^2(\Omega, \mathcal{F}_T, \mathbb{P})$, and assume $\mathbb{P}(F(X) \neq 0) > 0$. In view of the equality

$$\begin{aligned} \mathbb{E}_{\mathbb{P}}[F(X)] &= \mathbb{E}_{\mathbb{Q}_\lambda} \left[\left. \frac{d\mathbb{P}}{d\mathbb{Q}_\lambda} \right|_{\mathcal{F}_T} F(X) \right] \\ &= \mathbb{E}_{\mathbb{Q}_\lambda} \left[\left(\left. \frac{d\mathbb{Q}_\lambda}{d\mathbb{P}} \right|_{\mathcal{F}_T} \right)^{-1} F(X) \right] \\ &= \mathbb{E}_{\mathbb{Q}_\lambda} \left[e^{-\langle \lambda, X_T \rangle + \varphi_T(\lambda)} F(X) \right], \end{aligned}$$

define a set

$$\Lambda_2 := \Lambda_1 \cap \left\{ \lambda \in \mathbb{R}^d : \mathbb{E}_{\mathbb{P}} \left[e^{-\langle \lambda, X_T \rangle} F(X)^2 \right] < +\infty \right\},$$

and suppose that $\text{Leb}(\Lambda_2) > 0$. Let us now give a lemma, whose proof will be often adapted in what follows.

LEMMA 2.2. *The set Λ_2 is convex.*

Proof. For any $\lambda_1, \lambda_2 \in \Lambda_2$, and for any $m \in (0, 1)$ and $n = 1 - m$, the Hölder inequality gives

$$\mathbb{E}_{\mathbb{P}} \left[e^{-\langle m\lambda_1 + n\lambda_2, X_T \rangle} F(X)^2 \right] \leq \mathbb{E}_{\mathbb{P}} \left[e^{-\langle \lambda_1, X_T \rangle} F(X)^2 \right]^m \mathbb{E}_{\mathbb{P}} \left[e^{-\langle \lambda_2, X_T \rangle} F(X)^2 \right]^n < +\infty.$$

The claim then follows from the convexity of Λ_1 . \square

For $\lambda \in \Lambda_2$, the variance under the probability measure \mathbb{Q}_λ is given by

$$\begin{aligned} V(\lambda) &:= \mathbb{E}_{\mathbb{Q}_\lambda} \left[\left(\frac{d\mathbb{P}}{d\mathbb{Q}_\lambda} \Big|_{\mathcal{F}_T} \right)^2 F(X)^2 \right] - C^2 \\ &= \mathbb{E}_{\mathbb{P}} \left[\left(\frac{d\mathbb{Q}_\lambda}{d\mathbb{P}} \Big|_{\mathcal{F}_T} \right)^{-1} F(X)^2 \right] - C^2 \\ &= \mathbb{E}_{\mathbb{P}} \left[e^{-\langle \lambda, X_T \rangle + \varphi_T(\lambda)} F(X)^2 \right] - C^2. \end{aligned}$$

Define also a set

$$\Lambda_3 := \Lambda_2 \cap \left\{ \lambda \in \mathbb{R}^d : \mathbb{E}_{\mathbb{P}} \left[\|X_T\|^2 e^{-\langle \lambda, X_T \rangle} F(X)^2 \right] < +\infty \right\},$$

and assume that $\text{Leb}(\Lambda_3) > 0$.

PROPOSITION 2.3. *The set Λ_3 is convex and $V(\lambda)$ is strictly convex in $\lambda \in \Lambda_3$.*

Proof. The convexity of Λ_3 can be proved in a similar manner to the proof of Lemma 2.2.

Since $\lambda \in \Lambda_3$, by the Hölder inequality, we have

$$\mathbb{E}_{\mathbb{P}} \left[\|X_T\| e^{-\langle \lambda, X_T \rangle} F(X)^2 \right]^2 \leq \mathbb{E}_{\mathbb{P}} \left[e^{-\langle \lambda, X_T \rangle} F(X)^2 \right] \mathbb{E}_{\mathbb{P}} \left[\|X_T\|^2 e^{-\langle \lambda, X_T \rangle} F(X)^2 \right] < +\infty,$$

and thus with the help of the dominated convergence theorem, we obtain the gradient

$$\nabla V(\lambda) = \mathbb{E}_{\mathbb{P}} \left[(\nabla \varphi_T(\lambda) - X_T) e^{-\langle \lambda, X_T \rangle + \varphi_T(\lambda)} F(X)^2 \right],$$

and also the Hessian

$$\begin{aligned} \text{Hess}[V(\lambda)] &= \mathbb{E}_{\mathbb{P}} \left[\left(\text{Hess}[\varphi_T(\lambda)] + (\nabla \varphi_T(\lambda) - X_T)(\nabla \varphi_T(\lambda) - X_T)' \right) e^{-\langle \lambda, X_T \rangle + \varphi_T(\lambda)} F(X)^2 \right]. \end{aligned}$$

Then, we have for $y \in \mathbb{R}_0^d$,

$$\begin{aligned} y' \text{Hess}[V(\lambda)] y &= \mathbb{E}_{\mathbb{P}} \left[\left(y' \text{Hess}[\varphi_T(\lambda)] y + \langle y, \nabla \varphi_T(\lambda) - X_T \rangle^2 \right) e^{-\langle \lambda, X_T \rangle + \varphi_T(\lambda)} F(X)^2 \right] > 0, \end{aligned}$$

since $\text{Hess}[\varphi_T(\lambda)]$ reduces to the variance-covariance matrix of the random vector X_T under the probability measure \mathbb{Q}_λ , which is clearly positive definite. \square

Remark 2.4. The definition of the sets Λ_2 and Λ_3 is less intuitive and is of less practical use. We may instead give more intuitive definition in connection with the Lévy measure by giving up some part of its domain as

$$\Lambda'_2 = \left\{ \lambda \in \mathbb{R}^d : \int_{\|z\|>1} e^{-q\langle\lambda,z\rangle} \nu(dz) < +\infty, \mathbb{E}_{\mathbb{P}} [|F(X)|^{2p}] < +\infty, \frac{1}{p} + \frac{1}{q} = 1 \text{ for some } p > 1 \right\},$$

and

$$\Lambda'_3 = \left\{ \lambda \in \mathbb{R}^d : \int_{\|z\|>1} \|z\|^{2q} e^{-q\langle\lambda,z\rangle} \nu(dz) < +\infty, \mathbb{E}_{\mathbb{P}} [|F(X)|^{2p}] < +\infty, \frac{1}{p} + \frac{1}{q} = 1 \text{ for some } p > 1 \right\}.$$

It is easy to check that both Λ'_2 and Λ'_3 are convex, and that $\Lambda'_2 \subseteq \Lambda_2$, $\Lambda'_3 \subseteq \Lambda_3$, and $\Lambda'_3 \subseteq \Lambda'_2$. They are derived as follows. By the Hölder inequality, with $1/p + 1/q = 1$ for some $p > 1$ and for $k = 0, 2$,

$$\mathbb{E}_{\mathbb{P}} \left[\|X_T\|^k e^{-\langle\lambda,X_T\rangle} F(X)^2 \right] \leq \mathbb{E}_{\mathbb{P}} [|F(X)|^{2p}]^{1/p} \mathbb{E}_{\mathbb{P}} \left[\|X_T\|^{kq} e^{-q\langle\lambda,X_T\rangle} \right]^{1/q}.$$

By Theorem 2.1, the finiteness of the second expectation of the above right-hand side for each $k = 0, 2$ is equivalent to $\int_{\|z\|>1} \|z\|^{kq} e^{-q\langle\lambda,z\rangle} \nu(dz) < +\infty$ for corresponding k . This, with $k = 0$, asserts the definition of Λ_2 , while the definition of Λ_3 is verified with $k = 2$.

Meanwhile, as soon as $F(X)$ reduces to $f(X_T)$ with f being submultiplicative, the set Λ_3 is identical to

$$\left\{ \lambda \in \mathbb{R}^d : \int_{\|z\|>1} \left[e^{\langle\lambda,z\rangle} \vee \left(\|z\|^2 e^{-\langle\lambda,z\rangle} f(z)^2 \right) \right] \nu(dz) < +\infty \right\},$$

by Theorem 2.1.

3. Convergence of stochastic approximation algorithms. We begin with recalling the stochastic approximation algorithms. Let $\{X_{n,t} : t \in [0, T]\}_{n \in \mathbb{N}}$ be i.i.d. copies of the stochastic process $\{X_t : t \in [0, T]\}$. For ease in notation, we will write $X_n := X_{n,T}$ for $n \in \mathbb{N}$, and $F(X)_n := F(\{X_{n,t} : t \in [0, T]\})$. Let H be a connected set in \mathbb{R}^d with $\{0\} \in H$, and define a sequence $\{Y_n\}_{n \in \mathbb{N}}$ of random vectors in \mathbb{R}^d by

$$Y_{n+1} = (\nabla\varphi(\lambda_n) - X_{n+1}) e^{-\langle\lambda_n, X_{n+1}\rangle + \varphi(\lambda_n)} F(X)_{n+1}^2,$$

where $\lambda_0 \in H$, $\{\lambda_n\}_{n \in \mathbb{N}}$ is a sequence of random vectors in \mathbb{R}^d iteratively generated by

$$(3.1) \quad \lambda_{n+1} = \Pi_H [\lambda_n - \epsilon_n Y_{n+1}],$$

where Π_H is the projection onto the constraint set H and where $\{\epsilon_n\}_{n \in \mathbb{N}_0}$ is a sequence of positive constants satisfying

$$(3.2) \quad \sum_{n \in \mathbb{N}_0} \epsilon_n = +\infty, \quad \sum_{n \in \mathbb{N}_0} \epsilon_n^2 < +\infty.$$

Moreover, define the filtration $(\mathcal{G}_n)_{n \in \mathbb{N}_0}$ by $\mathcal{G}_n := \sigma(\{\lambda_k\}_{k \leq n}, \{X_k\}_{k \leq n})$.

In what follows, the term “the constrained algorithms” means the algorithms where the constraint set H in (3.1) is not \mathbb{R}^d and the sequence $\{\lambda_n\}_{n \in \mathbb{N}_0}$ is required to stay in the set, while by “the unconstrained algorithms,” we mean the ones whose constraint set H is extended to \mathbb{R}^d ; that is, the elements of $\{\lambda_n\}_{n \in \mathbb{N}_0}$ may be arbitrarily taken in \mathbb{R}^d .

Define a set

$$\Lambda_4 := \Lambda_1 \cap \left\{ \lambda \in \mathbb{R}^d : \mathbb{E}_{\mathbb{P}} \left[\|X_T\|^k e^{-2\langle \lambda, X_T \rangle} F(X)^4 \right] < +\infty, k = 0, 2 \right\}.$$

We will below see that the algorithm is unconstrained if $\Lambda_4 = \mathbb{R}^d$. It is, however, difficult to check whether or not that is the case, since the the operator F is involved. Meanwhile, to have an unconstrained algorithm, we need at least $\Lambda_4 \subseteq \Lambda_1 = \mathbb{R}^d$. In this sense, let us give a rough illustration of the situation in the following.

LEMMA 3.1. *If the Lévy measure ν has a compact support, then $\Lambda_1 = \mathbb{R}^d$. If $\int_{\|z\| > 1} e^{\|z\|^{1+\delta}} \nu(dz) < +\infty$ for some $\delta > 0$, then $\Lambda_1 = \mathbb{R}^d$.*

3.1. Constrained algorithms. The following proves the almost sure convergence of the constrained algorithms. Their gradient-based structure simplifies the argument.

THEOREM 3.2. *Assume that $\text{Leb}(\Lambda_4) \in (0, +\infty)$, and $\lambda_0 \in \Lambda_4$. Let H be a compact set such that $H \subseteq \Lambda_4$. Then, there exists $\lambda^* \in H$ such that the sequence $\{\lambda_n\}_{n \in \mathbb{N}_0}$ in (3.1) converges \mathbb{P} -a.s. to λ^* . Moreover, $V(\lambda^*) \leq V(0)$.*

Proof. First, note that $\mathbb{E}_{\mathbb{P}}[e^{-\langle \lambda, X_T \rangle} F(X)^2] < +\infty$ since $\mathbb{E}_{\mathbb{P}}[e^{-2\langle \lambda, X_T \rangle} F(X)^4] < +\infty$, and that by the Cauchy–Schwarz inequality,

$$\mathbb{E}_{\mathbb{P}} \left[\|X_T\|^2 e^{-\langle \lambda, X_T \rangle} F(X)^2 \right]^2 \leq \mathbb{E}_{\mathbb{P}} \left[\|X_T\|^2 \right] \mathbb{E}_{\mathbb{P}} \left[\|X_T\|^2 e^{-2\langle \lambda, X_T \rangle} F(X)^4 \right] < +\infty.$$

Hence, $\Lambda_4 \subseteq \Lambda_3$. The convexity of Λ_4 can be proved in a similar manner to the proof of Lemma 2.2. Moreover, we have

$$\begin{aligned} \mathbb{E}_{\mathbb{P}} \left[\|X_T\| e^{-2\langle \lambda, X_T \rangle} F(X)^4 \right]^2 \\ \leq \mathbb{E}_{\mathbb{P}} \left[e^{-2\langle \lambda, X_T \rangle} F(X)^4 \right] \mathbb{E}_{\mathbb{P}} \left[\|X_T\|^2 e^{-2\langle \lambda, X_T \rangle} F(X)^4 \right] < +\infty. \end{aligned}$$

Now, since

$$\sup_{n \in \mathbb{N}} \mathbb{E}_{\mathbb{P}} \left[\|Y_n\|^2 \right] \leq \sup_{\lambda \in H} \mathbb{E}_{\mathbb{P}} \left[\|\nabla \varphi_T(\lambda) - X_T\|^2 e^{-2\langle \lambda, X_T \rangle + 2\varphi_T(\lambda)} F(X)^4 \right],$$

and since $\|\nabla \varphi(\lambda)\| < +\infty$ and $\varphi(\lambda) < +\infty$, for $\lambda \in H$, the expectation of the above right-hand side is finite if and only if $\mathbb{E}_{\mathbb{P}}[\|X_T\|^k e^{-2\langle \lambda, X_T \rangle} F(X)^4] < +\infty$ for each $k = 0, 1, 2$. This proves $\sup_{n \in \mathbb{N}} \mathbb{E}_{\mathbb{P}}[\|Y_n\|^2] < +\infty$. Since Λ_4 is convex, it follows from Theorem 2.1 (p. 127) of Kushner and Yin [11] that the sequence $\{\lambda_n\}_{n \in \mathbb{N}_0}$ converges

\mathbb{P} -a.s. to a unique stationary point in H . The last claim holds by the strict convexity of V on H . \square

Remark 3.3. It is not clear whether or not there exists $\lambda \in H$ such that $\nabla V(\lambda) = 0$, and thus the above stationary point $\lambda^* \in H$ does not necessarily attain $\nabla V(\lambda^*) = 0$. If, however, there happens to exist $\lambda \in H$ such that $\nabla V(\lambda) = 0$, then $\nabla V(\lambda^*) = 0$ is guaranteed by the strict convexity of V on Λ_4 .

Remark 3.4. We may give some modifications of the set Λ_4 so that it looks more intuitive, as in Remark 3.3. If $F(X) = f(X_T)$ with f being submultiplicative, then Λ_4 can be rewritten as

$$\Lambda_4 = \left\{ \lambda \in \mathbb{R}^d : \int_{\|z\|>1} \left[e^{\langle \lambda, z \rangle} \vee \|z\|^2 e^{-2\langle \lambda, z \rangle} f(z)^4 \right] \nu(dz) < +\infty \right\}.$$

Otherwise, by the Hölder inequality,

$$\Lambda'_4 = \left\{ \lambda \in \mathbb{R}^d : \int_{\|z\|>1} \|z\|^{2q} e^{-2q\langle \lambda, z \rangle} \nu(dz) < +\infty, \mathbb{E}_{\mathbb{P}} [|F(X)|^{4p}] < +\infty, \frac{1}{p} + \frac{1}{q} = 1 \text{ for some } p > 1 \right\},$$

which is a convex subset of Λ_4 .

3.2. Unconstrained algorithms. We begin with the main result.

PROPOSITION 3.5. *If $\Lambda_4 = \mathbb{R}^d$ and if there exists $c > 0$ such that*

$$(3.3) \quad M := \inf_{\|y\|=1} \int_{\|z\|\leq c} \langle y, z \rangle^2 \mathbb{1}_{[0,+\infty)}(\langle y, z \rangle) \nu(dz) > 0,$$

then there exists a unique $\lambda^ \in \mathbb{R}^d$ such that $\nabla V(\lambda^*) = 0$.*

Remark 3.6. In most applications, Lévy processes are chosen to have independent components, each of which possesses small jumps in both positive and negative directions. Then, their Lévy measures are supported on all the axes of \mathbb{R}^d , that is,

$$\nu(dz_1, \dots, dz_d) = \sum_{k=1}^d \delta_0(dz_1) \cdots \delta_0(dz_{k-1}) \nu_k(dz_k) \delta_0(dz_{k+1}) \cdots \delta_0(dz_d),$$

for some Lévy measures $\{\nu_k\}_{k=1, \dots, d}$ on \mathbb{R}_0 . We then get

$$M = \inf_{\|y\|=1} \sum_{k=1}^d y_k^2 \int_{z_k \in (0, \text{sgn}(y_k)c]} z_k^2 \nu_k(dz_k) > 0.$$

In Example 4.2 below, we discretize the sample paths of Lévy processes with both positive and negative jumps on a finite time horizon into a few independent increments, and thus the condition (3.3) holds true.

Proof. By Proposition 2.3 with $\Lambda_3 \supseteq \Lambda_4 = \mathbb{R}^d$, it suffices to show that $\lim_{\|\lambda\| \uparrow +\infty} V(\lambda) = +\infty$. First, note that with a suitable $\gamma_c \in \mathbb{R}^d$,

$$\varphi(\lambda) - \langle \lambda, \gamma_c \rangle = \int_{\|z\|>c} \left(e^{\langle \lambda, z \rangle} - 1 \right) \nu(dz) + \int_{\|z\|\leq c} \left(e^{\langle \lambda, z \rangle} - 1 - \langle \lambda, z \rangle \right) \nu(dz).$$

The first component of the right-hand side above is bounded from below by $-\nu(\{z \in \mathbb{R}_0^d : \|z\| > c\})$ since $\Lambda_1 = \mathbb{R}^d$. For the second component, since $e^x - 1 - x \geq 0$, $x \in \mathbb{R}$, we have

$$\begin{aligned} & \int_{\|z\| \leq c} \left(e^{\langle \lambda, z \rangle} - 1 - \langle \lambda, z \rangle \right) \nu(dz) \\ & \geq \inf_{\|y\|=1} \int_{\|z\| \leq c} \left(e^{\|\lambda\| \langle y, z \rangle} - 1 - \|\lambda\| \langle y, z \rangle \right) \nu(dz) \\ & \geq \inf_{\|y\|=1} \int_{\|z\| \leq c} \left(e^{\|\lambda\| \langle y, z \rangle} - 1 - \|\lambda\| \langle y, z \rangle \right) \mathbb{1}_{[0, +\infty)}(\langle y, z \rangle) \nu(dz) \\ & \geq M \|\lambda\|^2. \end{aligned}$$

Therefore, we get

$$\begin{aligned} & \mathbb{E}_{\mathbb{P}} \left[e^{-\langle \lambda, X_T \rangle + \varphi_T(\lambda)} F(X)^2 \right] \\ & = e^{\varphi_T(\lambda) - T \langle \lambda, \gamma_c \rangle} \mathbb{E}_{\mathbb{P}} \left[e^{-\langle \lambda, X_T - T \gamma_c \rangle} F(X)^2 \right] \\ & \geq e^{\varphi_T(\lambda) - T \langle \lambda, \gamma_c \rangle} \mathbb{E}_{\mathbb{P}} \left[e^{-\|\lambda\| \|X_T - T \gamma_c\|} F(X)^2 \mathbb{1}(\|X_T - T \gamma_c\| \leq M \|\lambda\| / 2) \right] \\ & \geq e^{T(M \|\lambda\|^2 / 2 - \nu(\{z \in \mathbb{R}_0^d : \|z\| > c\}))} \mathbb{E}_{\mathbb{P}} \left[F(X)^2 \mathbb{1}(\|X_T - T \gamma_c\| \leq M \|\lambda\| / 2) \right], \end{aligned}$$

which explodes as $\|\lambda\| \uparrow +\infty$. This proves the claim. \square

The unconstrained algorithms often show a rough numerical behavior. This phenomenon is mainly due to the extremely fast growth of $\mathbb{E}_{\mathbb{P}}[\|\nabla \varphi_T(\lambda) - X_T\|^2 e^{-2\langle \lambda, X_T \rangle + 2\varphi_T(\lambda)} F(X)^4]$ with respect to $\|\lambda\|$. Alternatively, Chen, Guo and Gao [4] proposes a projection procedure. In essence, by forcing the iterates to stay in an increasing sequence of compact sets, the procedure avoids the explosion of the algorithm during the early stage. Meanwhile, we adapt the results of Chen and Zhu [3] and Delyon [5]. Let $\{H_n\}_{n \in \mathbb{N}_0}$ be an increasing sequence of compact sets such that $\cup_{n \in \mathbb{N}_0} H_n = \mathbb{R}^d$, and modify the algorithm (3.1) as

$$(3.4) \quad \lambda_{n+1} = \Pi_{H_{\sigma(n)}} [\lambda_n - \epsilon_n Y_{n+1}],$$

where $\sigma(n)$ counts the number of projections up to the n th step.

THEOREM 3.7. *Assume that $\Lambda_4 = \mathbb{R}^d$ and that there exists a unique λ^* such that $\nabla V(\lambda^*) = 0$. Then, the sequence $\{\lambda_n\}_{n \in \mathbb{N}_0}$ in (3.4) converges \mathbb{P} -a.s. to λ^* . Moreover, $\lim_{n \uparrow +\infty} \sigma(n) < +\infty$, \mathbb{P} -a.s.*

Proof. Let $m \in \mathbb{N}$ and define for $n \in \mathbb{N}_0$,

$$M_n := \sum_{k=0}^n \epsilon_k (Y_{k+1} - \mathbb{E}_{\mathbb{P}} [Y_{k+1} | \mathcal{G}_k]) \mathbb{1}(\|\lambda_k\| < m).$$

By Proposition 3.5 and the results in [3, 5], we are only to show that for each $m \in \mathbb{N}$, $\{M_n\}_{n \in \mathbb{N}_0}$ converges \mathbb{P} -a.s. Since the sequence $\{M_n\}_{n \in \mathbb{N}_0}$ is a martingale with respect to the filtration $(\mathcal{G}_n)_{n \in \mathbb{N}_0}$, it suffices to show that $\{M_n\}_{n \in \mathbb{N}_0}$ is a L^2 -martingale. To this end, for each $m \in \mathbb{N}$, we will show that, \mathbb{P} -a.s.,

$$\begin{aligned} & \sum_{n \in \mathbb{N}_0} \epsilon_n^2 \mathbb{E}_{\mathbb{P}} [\|Y_{n+1}\|^2 \mathbb{1}(\|\lambda_n\| \leq m) | \mathcal{G}_n] \\ & = \sum_{n \in \mathbb{N}_0} \epsilon_n^2 \mathbb{E}_{\mathbb{P}} \left[\|\nabla \varphi_T(\lambda_n) - X_T\|^2 e^{-2\langle \lambda_n, X_T \rangle + 2\varphi_T(\lambda_n)} F(X)^4 \mathbb{1}(\|\lambda_n\| \leq m) | \mathcal{G}_n \right] < +\infty. \end{aligned}$$

We begin with proving that for each $m \in \mathbb{N}$, the following four quantities are well defined:

$$\begin{aligned} C_1(m) &:= \sup_{\|\lambda\| \leq m} \left| \int_{\|z\| > 1} \left(e^{\langle \lambda, z \rangle} - 1 \right) \nu(dz) \right|, \\ C_2(m) &:= \sup_{\|\lambda\| \leq m} \left| \int_{\|z\| \leq 1} \left(e^{\langle \lambda, z \rangle} - 1 - \langle \lambda, z \rangle \right) \nu(dz) \right|, \\ C_3(m) &:= \sup_{\|\lambda\| \leq m} \int_{\|z\| > 1} \|z\| e^{\langle \lambda, z \rangle} \nu(dz), \\ C_4(m) &:= \sup_{\|\lambda\| \leq m} \int_{\|z\| \leq 1} \|z\| \left| e^{\langle \lambda, z \rangle} - 1 \right| \nu(dz). \end{aligned}$$

Clearly, $C_1(m)$ is finite since $\Lambda_1 = \mathbb{R}^d$ and $\nu(\{z \in \mathbb{R}_0^d : \|z\| > 1\}) < +\infty$, while the finiteness of $C_2(m)$ follows from $e^{\langle \lambda, z \rangle} - 1 - \langle \lambda, z \rangle \sim \langle \lambda, z \rangle^2 \leq \|\lambda\|^2 \|z\|^2$ as $\|z\| \downarrow 0$. For $C_3(m)$, the Hölder inequality gives the assertion, that is, with $1/p + 1/q = 1$ for some $p > 1$,

$$\int_{\|z\| > 1} \|z\| e^{\langle \lambda, z \rangle} \nu(dz) \leq \left[\int_{\|z\| > 1} \|z\|^p \nu(dz) \right]^{1/p} \left[\int_{\|z\| > 1} e^{q\langle \lambda, z \rangle} \nu(dz) \right]^{1/q} < +\infty,$$

again with the help of $\Lambda_1 = \mathbb{R}^d$. Finally, the finiteness of $C_4(m)$ holds by $\|z\| |e^{\langle \lambda, z \rangle} - 1| \sim \|z\| |\langle \lambda, z \rangle| \leq \|\lambda\| \|z\|^2$ as $\|z\| \downarrow 0$.

Let us now proceed to the main part of the proof. First, as previously, note that

$$\varphi(\lambda) - \langle \lambda, \gamma \rangle = \int_{\|z\| > 1} \left(e^{\langle \lambda, z \rangle} - 1 \right) \nu(dz) + \int_{\|z\| \leq 1} \left(e^{\langle \lambda, z \rangle} - 1 - \langle \lambda, z \rangle \right) \nu(dz).$$

Both the first and the second integrals of the right-hand side above are well defined due to the finiteness of $C_1(m)$ and $C_2(m)$, respectively. Hence, we get

$$|\varphi(\lambda) - \langle \lambda, \gamma \rangle| \leq C_1(m) + C_2(m) =: C_5(m).$$

Next, note that

$$\nabla (\varphi(\lambda) - \langle \lambda, \gamma \rangle) = \int_{\|z\| > 1} z e^{\langle \lambda, z \rangle} \nu(dz) + \int_{\|z\| \leq 1} z \left(e^{\langle \lambda, z \rangle} - 1 \right) \nu(dz),$$

where the passages to the gradient operator are verified by the finiteness of $C_3(m)$ and $C_4(m)$, and thus

$$\|\nabla (\varphi(\lambda) - \langle \lambda, \gamma \rangle)\| \leq C_3(m) + C_4(m) =: C_6(m).$$

In total, we get for each $m \in \mathbb{N}$,

$$\begin{aligned} &\mathbb{E}_{\mathbb{P}} \left[\|\nabla \varphi_T(\lambda) - X_T\|^2 e^{-2\langle \lambda, X_T \rangle + 2\varphi_T(\lambda)} F(X)^4 \mathbf{1}(\|\lambda\| \leq m) \mid \mathcal{G}_n \right] \\ &\leq \mathbb{E}_{\mathbb{P}} \left[(\|X_T - \gamma T\| + C_6(m)T)^2 e^{-2\langle \lambda, X_T - \gamma T \rangle + 2C_5(m)T} F(X)^4 \mathbf{1}(\|\lambda\| \leq m) \mid \mathcal{G}_n \right], \end{aligned}$$

which is bounded \mathbb{P} -a.s., since $\Lambda_4 = \mathbb{R}^d$. The proof is complete. \square

4. Numerical illustrations. In this section, we give two numerical examples, corresponding to the constrained algorithm and the unconstrained one. We will evaluate the efficiency of the importance sampling variance reduction by the ratio of variances (vratio), defined by

$$(\text{vratio}) := \frac{\text{Var}_{\mathbb{P}}(F(X))}{\text{Var}_{\mathbb{Q}_{\lambda_N}}((d\mathbb{P}/d\mathbb{Q}_{\lambda_N})F(X))}.$$

Example 4.1 (constrained algorithm). Let $X := (X^1, \dots, X^5)'$ be an infinitely divisible random vector with independent and identically distributed components under the probability measure \mathbb{P} , where the common Lévy measure ν on \mathbb{R}_0 for each component is of the Meixner type of Schoutens and Teugels [13]. It is characterized by three parameters (a, b, d) in the form

$$\nu(dz) = d \frac{\exp(bz/a)}{z \sinh(\pi z/a)} dz, \quad z \in \mathbb{R}_0,$$

where $a > 0$, $b \in (-\pi, \pi)$, and $d > 0$, while the probability density function p of X_1 is given in closed form by

$$(4.1) \quad p(x) = \frac{(2 \cos(b/2))^{2d}}{2a\pi\Gamma(2d)} e^{bx/a} \left| \Gamma\left(d + \frac{ix}{a}\right) \right|^2.$$

We can derive that

$$\Lambda_1 = \left\{ \lambda \in \mathbb{R} : \int_{|z|>1} e^{\lambda z} \nu(dz) < +\infty \right\}^5 = \left(\frac{-\pi - b}{a}, \frac{\pi - b}{a} \right)^5,$$

and that for $\lambda \in \Lambda_1$,

$$\varphi(\lambda) = \sum_{k=1}^5 2d \left[\ln \left(\cos \frac{b}{2} \right) - \ln \left(\cos \frac{b + a\lambda_k}{2} \right) \right],$$

and

$$\nabla \varphi(\lambda) = \left(2ad \tan \frac{b + a\lambda_1}{2}, \dots, 2ad \tan \frac{b + a\lambda_5}{2} \right)'.$$

Consider an Asian payoff

$$F(X) = \max \left[0, \frac{1}{5} \sum_{k=1}^5 S_0 e^{\sum_{i=1}^k X_i - k\varphi((1,0,0,0,0)')} - K \right].$$

For the condition $\mathbb{E}_{\mathbb{P}}[|F(X)|^{2p}] < +\infty$, it is sufficient to have $\int_{|z|>1} e^{2pz} \nu(dz) < +\infty$. With $p > 1$, we get $p \in (1, +\infty) \cap ((-\pi - b)/(2a), (\pi - b)/(2a))$, provided that $\pi - b > 2a$. Next, the condition $\int_{|z|>1} |z|^{2q} e^{-q\lambda z} \nu(dz) < +\infty$ yields $b/a - q\lambda \in (-\pi/a, \pi/a)$, and in view of the interval of q , we get

$$\Lambda'_2 = \Lambda_1 \cap \left(\frac{b - \pi}{a} \left(\left[1 - \frac{2a}{\pi - b} \right] \wedge 1 \right), \frac{b + \pi}{a} \left(\left[1 - \frac{2a}{\pi - b} \right] \wedge 1 \right) \right)^5,$$

provided that $\pi - b > 2a$. In a similar manner, we can prove $\Lambda'_3 = \Lambda'_2$ and

$$\Lambda'_4 = \Lambda_1 \cap \left(\frac{b - \pi}{2a} \left(\left[1 - \frac{4a}{\pi - b} \right] \wedge 1 \right), \frac{b + \pi}{2a} \left(\left[1 - \frac{4a}{\pi - b} \right] \wedge 1 \right) \right)^5,$$

provided that $\pi - b > 4a$.

We set the parameters of the Meixner distribution $(a, b, d) = (0.1, 0.0, 1.0)$, and thus an effective domain is approximately $\Lambda'_4 = (-13.707963, 13.707963)^5$. The constraint set H must be compact, so it is safe to set $H = [-13.70796, 13.70796]^5 \subset \Lambda'_4$. We generate $N = 1e + 5$ Monte Carlo runs with the full help of the closed form density function (4.1). With those runs, we perform the constrained algorithm (3.1) with $\epsilon_n = \alpha/(n + 1)$ and $\lambda_0 = \{0\}$. We examine three cases: the ATM case ($K = 100$), an OTM case ($K = 125$), and a deep OTM case ($K = 150$). The left figures in Figure 1 draw a sequence $\{\|\nabla V(\lambda_n)\|\}_{n \in \mathbb{N}_0}$ of the absolute gradient levels, which is “desired” to achieve $\lim_{n \uparrow +\infty} \|\nabla V(\lambda_n)\| = 0$, \mathbb{P} -*a.s.* (As pointed out in Remark 2.4, it is not clear whether or not the constraint set H contains λ^* such that $\nabla V(\lambda^*) = 0$.) The figures on the right present the convergence of the Monte Carlo estimate $\mathbb{E}_{\mathbb{P}}[F(X)]$ (MC) and that of the importance sampling Monte Carlo estimate $\mathbb{E}_{\mathbb{Q}_{\lambda_N}}[(d\mathbb{P}/d\mathbb{Q}_{\lambda_N})F(X)]$ (IS MC), of which λ_N is the exponential tilting parameter obtained after $N = 1e + 5$ of the stochastic approximation iterations, while the three vertical lines indicate $\tilde{C} := \mathbb{E}_{\mathbb{Q}_{\lambda_N}}[(d\mathbb{P}/d\mathbb{Q}_{\lambda_N})F(X)]$, $0.99\tilde{C}$ and $1.01\tilde{C}$.

The absolute gradient level tends to decrease as desired, and the resulting importance sampling succeeds in reducing the Monte Carlo variance. The absolute gradient level seems to have already converged to zero, while we have observed that a component of $\{\lambda_n\}_{n \in \mathbb{N}_0}$ seems to stay at the upper boundary ($= +13.70796$) in the ATM ($K = 100$) and in the OTM ($K = 125$). Those are delicate issues in the constrained algorithms.

Example 4.2 (unconstrained algorithm). Let $X := (X_1, \dots, X_5)'$ be an infinitely divisible random vector in \mathbb{R}^5 with independent and identically distributed components, whose common Lévy measure ν on \mathbb{R}_0 of each component is given in the form of the standard Gaussian density function,

$$\nu(dz) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz.$$

Evidently, for each $\lambda \in \mathbb{R}$, $\int_{|z|>1} e^{\lambda z} \nu(dz) < +\infty$. Letting $\lambda := (\lambda_1, \dots, \lambda_5)'$, with the help of the independence of the components, we get $\Lambda_1 = \mathbb{R}^5$,

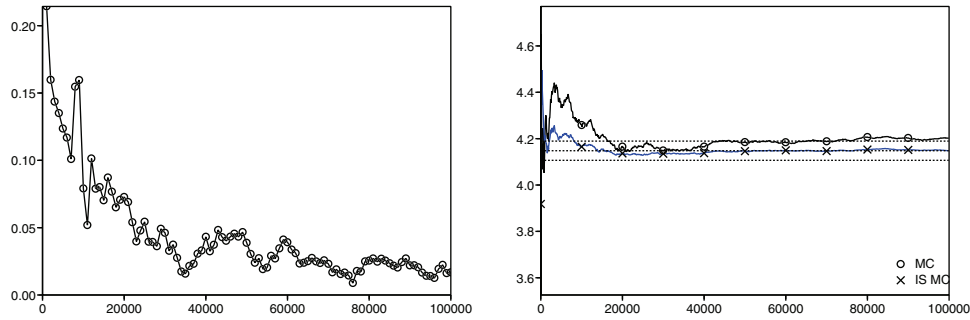
$$\varphi(\lambda) = \sum_{k=1}^5 \left(e^{\frac{1}{2}\lambda_k^2} - 1 \right),$$

and

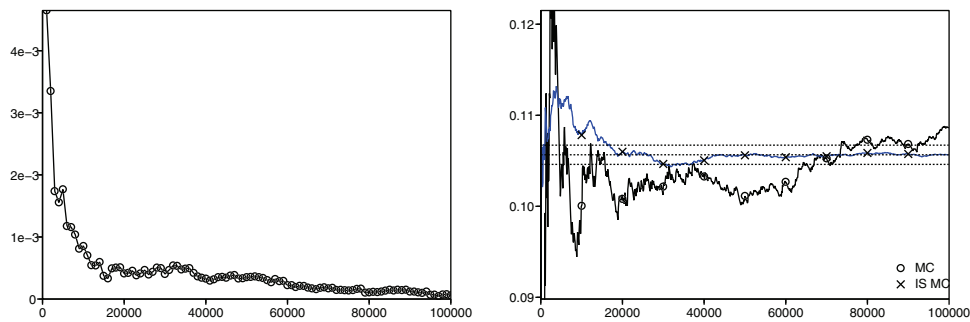
$$\nabla \varphi(\lambda) = \left(\lambda_1 e^{\frac{1}{2}\lambda_1^2}, \dots, \lambda_5 e^{\frac{1}{2}\lambda_5^2} \right)'.$$

Due to the compound Poisson structure, the random vector under the probability measure \mathbb{P} can be generated via

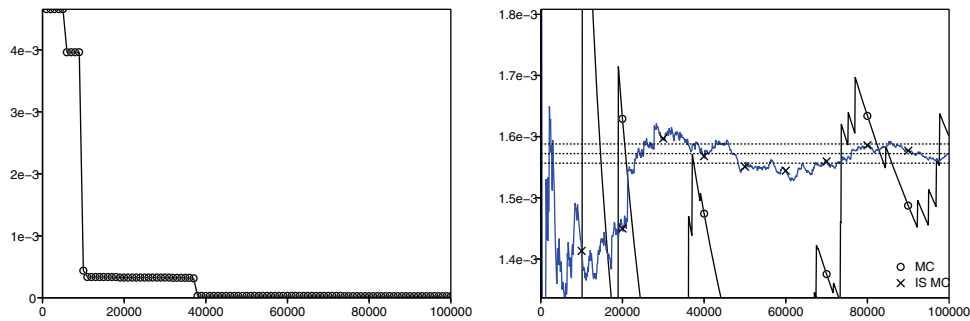
$$X \stackrel{\mathcal{L}}{=} \left(\sum_{n=1}^{N_1} W_{1,n}, \dots, \sum_{n=1}^{N_5} W_{5,n} \right)',$$



The ATM (K=100) case with $\alpha = 5e+0$. The ratio is 5.063.



An OTM (K=125) case with $\alpha = 1e+3$. The ratio is 35.26.



A deep OTM (K=150) case with $\alpha = 1e+4$. The ratio is 60.67.

FIG. 1. Results for the constrained algorithm; $\{\|\nabla V(\lambda_n)\|\}_{n \in \mathbb{N}_0}$ (left figures), while $\mathbb{E}_{\mathbb{P}}[F(X)]$ (MC) and $\mathbb{E}_{\mathbb{Q}_{\lambda_N}}[(d\mathbb{P}/d\mathbb{Q}_{\lambda_N})F(X)]$ (IS MC) (right figures).

where $\{N_n\}_{n \leq 5}$ is a sequence of i.i.d. Poisson random variables with unit parameter and $\{W_{k,n}\}_{k \leq 5, n \in \mathbb{N}}$ is an i.i.d. standard Gaussian random array. In view of (2.1), the Lévy measure under the probability measure \mathbb{Q}_λ is given by

$$\nu_\lambda(dz) = e^{\lambda z} \nu(dz) = \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}\lambda^2} e^{-\frac{1}{2}(z-\lambda)^2} dz,$$

which is just like a drift shift of the Gaussian density by λ (up to the constant $e^{\frac{1}{2}\lambda^2}$). Hence, the random vector under the new probability measure \mathbb{Q}_λ can be generated

via the identity

$$(4.2) \quad \begin{aligned} X &\stackrel{\mathcal{L}}{=} \left(\sum_{n=1}^{N_1} (W_{1,n} + \lambda_1), \dots, \sum_{n=1}^{N_5} (W_{5,n} + \lambda_5) \right)' \\ &= \left(\sum_{n=1}^{N_1} W_{1,n} + \lambda_1 N_1, \dots, \sum_{n=1}^{N_5} W_{5,n} + \lambda_5 N_5 \right)', \end{aligned}$$

where $\{N_n\}_{n \leq 5}$ is now a sequence of i.i.d. Poisson random variables with parameter $e^{\frac{1}{2}\lambda^2}$ (≥ 1) and where $\{W_{k,n}\}_{k \leq 5, n \in \mathbb{N}}$ remains to be an i.i.d. standard Gaussian random array. For any $\lambda \in \mathbb{R}_0^d$, the componentwise variance tends to increase by factor $e^{\frac{1}{2}\lambda_k^2}$, since $\mathbb{E}_{\mathbb{Q}_\lambda}[\sum_{n=1}^{N_k} W_{k,n}] = \mathbb{E}_{\mathbb{Q}_\lambda}[N_k] \mathbb{E}_{\mathbb{Q}_\lambda}[W_{k,1}]$, while the drift shift λ_k is further accelerated by factor $e^{\frac{1}{2}\lambda_k^2}$ on average.

Consider a digital payoff

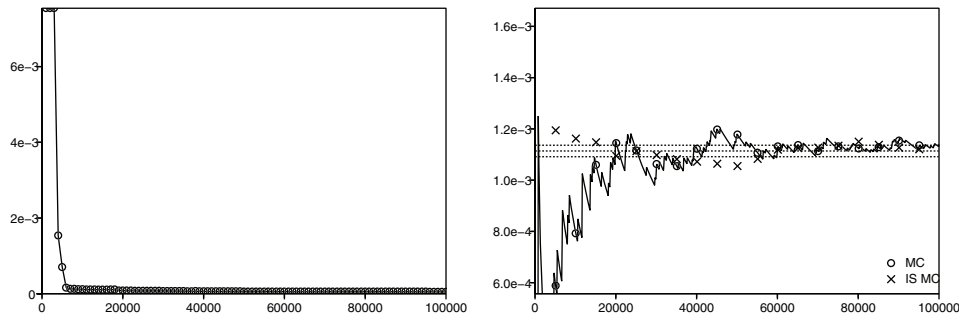
$$F(X) = \mathbf{1}(S_1 < 100 - K, S_2 > 100 + K, S_3 < 100 - K, S_4 > 100 + K, S_5 < 100 - K),$$

for a suitable K and where $S_n = 100 \exp[\sum_{k=1}^n X_k - n\varphi((1, 0, 0, 0, 0)')]$, $n \leq 5$. Since $|F(X)| \leq 1$, \mathbb{P} -a.s., we get $\Lambda_4 = \mathbb{R}^5$. We generate $N = 1e + 5$ Monte Carlo runs and perform the unconstrained algorithm (3.4) with $\epsilon_n := \alpha/(n + 1)$, $H_n := \{\lambda \in \mathbb{R}^d : \|\lambda\| \leq 10 \ln(100(n + 1))\}$, and $\lambda_0 := \{0\}$. We examine the three cases; $K = 5, 20$, and 40 . We present the results in Figure 2, where the three vertical lines in the right figures here indicate $\tilde{C} := \mathbb{E}_{\mathbb{Q}_{\lambda_N}}[(d\mathbb{P}/d\mathbb{Q}_{\lambda_N})F(X)]$, $0.98\tilde{C}$ and $1.02\tilde{C}$.

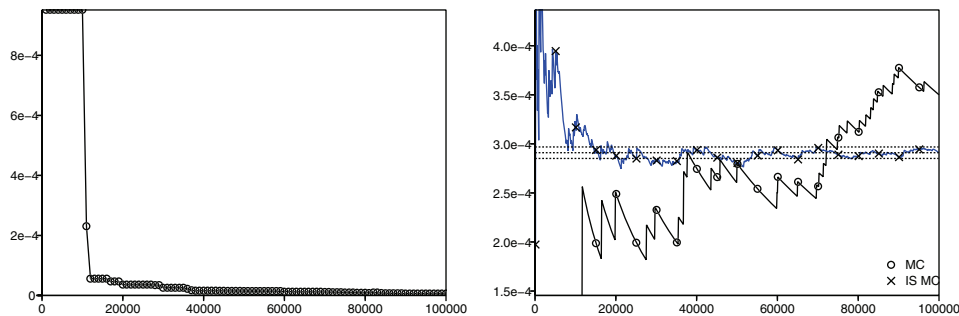
In similar to the results in Example 4.1, the algorithm reduces the absolute gradient level, and the resulting importance sampling succeeds in reducing the Monte Carlo variance. Unlike in the constrained algorithm case, we know that there exists a unique optimum λ^* which makes the absolute gradient level zero. It seems that the zero is fairly attained.

Remark 4.3. In the above numerical illustrations, we have chosen the Lévy measures of the Meixner type and of the Gaussian density. It is a clear reason of the choice that they are somewhat invariant with respect to the Esscher transform owing to the exponential component of their Lévy measure and thus remain very easy to generate in simulation even after the measure change. It should be mentioned here that from a computational point of view, Lévy measures without such an invariance property may not be a good candidate for simulation in our framework. However, this should not be a crucial drawback since most recent popular Lévy measures possess an exponential component, for instance, the Lévy measure of gamma processes.

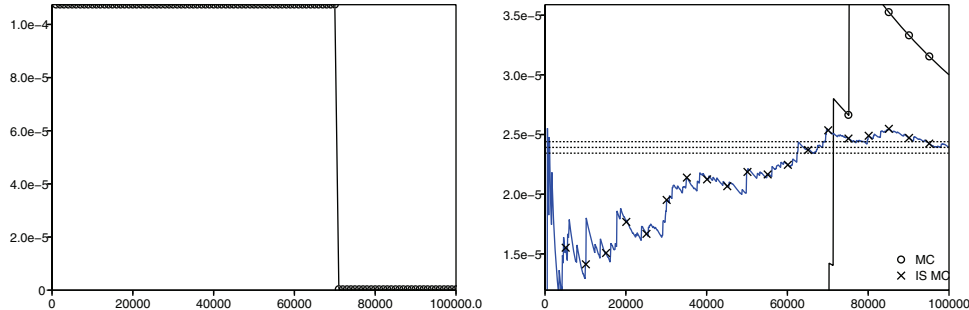
5. Concluding remarks. In this paper, we have developed stochastic approximation methods of finding the optimal measure change for Lévy processes in Monte Carlo importance sampling variance reduction. Our analysis is valid on the basis of the restriction to the exponential tilting measure change, that is, limiting the density to a function the terminal value X_T . Nevertheless, our method should be applicable to a variety of applications since its principle is not specific to the structure of the Monte Carlo estimator itself. It may be of interest to extend to the intricate series representation setting of [8] by using characterizing parameters of the Lévy measure in the stochastic approximation procedure. Extensions to an optimal parameter search for the combined importance sampling and control variates variance



K=5 with $\alpha = 8e+2$. The ratio is 6.605.



K=20 with $\alpha = 1e+4$. The ratio is 11.75.



K=40 with $\alpha = 3e+4$. The ratio is 57.27.

FIG. 2. Results for the unconstrained algorithm; $\{\|\nabla V(\lambda_n)\|\}_{n \in \mathbb{N}_0}$ (left figures), while $\mathbb{E}_{\mathbb{P}}[F(X)]$ (MC) and $\mathbb{E}_{\mathbb{Q}_{\lambda_N}}[(d\mathbb{P}/d\mathbb{Q}_{\lambda_N})F(X)]$ (IS MC) (right figures).

reduction are studied using a two-time-scale version of the stochastic approximation algorithm in subsequent papers [9, 10].

Acknowledgments. The author is grateful to anonymous referees for a careful reading and various useful suggestions, and in particular for remarks which helped improve Proposition 3.5. Part of this work was done at Daiwa Securities SMBC Co. Ltd. The author would like to thank Tatsuya Toshida and the Financial Engineering team for their support and encouragement. This work is supported in part by JSPS Core-to-Core program. At last but not least, it is a great pleasure to express my appreciation to Arturo Kohatsu-Higa for stimulating discussions.

REFERENCES

- [1] B. AROUNA, *Robbins-Monro algorithms and variance reduction in finance*, J. Comput. Finance, 7 (2004), pp. 35–61.
- [2] B. AROUNA, *Adaptive Monte Carlo method, a variance reduction technique*, Monte Carlo Methods Appl., 10 (2004), pp. 1–24.
- [3] K.-F. CHEN AND Y. ZHU, *Stochastic Approximation Procedure with randomly varying truncations*, Scientia Sinica Series, 1986.
- [4] H.-F. CHEN, L. GUO, AND A.-J. GAO, *Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds*, Stochastic Process. Appl., 27 (1987), pp. 217–231.
- [5] B. DELYON, *General results on the convergence of stochastic algorithms*, IEEE Trans. Automat. Control, 41 (1996), pp. 1245–1255.
- [6] P. GLASSERMAN, P. HEIDELBERGER, AND P. SHAHABUDDIN, *Asymptotic optimal importance sampling and stratification for pricing path-dependent options*, Math. Finance, 9 (1999), pp. 117–152.
- [7] P. GLASSERMAN, *Monte Carlo Methods in Financial Engineering*, Springer-Verlag, New York, 2004.
- [8] R. KAWAI, *An importance sampling method based on the density transformation of Lévy processes*, Monte Carlo Methods Appl., 12 (2006), pp. 171–186.
- [9] R. KAWAI, *Adaptive Monte Carlo variance reduction with two-time-scale stochastic approximation*, Monte Carlo Methods Appl., 13 (2007), pp. 197–217.
- [10] R. KAWAI, *Adaptive Monte Carlo variance reduction for Lévy processes with two-time-scale stochastic approximation*, Methodol. Comput. Appl. Probab., 10 (2008), pp. 199–223.
- [11] H.J. KUSHNER AND G. YIN, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed., Springer-Verlag, New York, 2003.
- [12] K. SATO, *Lévy Processes and Infinitely Divisible Distributions*, Cambridge University Press, Cambridge, 1999.
- [13] W. SCHOUTENS AND J.L. TEUGELS, *Lévy processes, polynomials and martingales*, Commun. Statistics: Stochastic Models, 14 (1998), pp. 335–349.
- [14] Y. SU AND M.-C. FU, *Optimal importance sampling in securities pricing*, J. Comput. Finance, 5 (2002), pp. 27–50.

CONVERGENCE OF THE FORWARD EULER METHOD FOR NONCONVEX DIFFERENTIAL INCLUSIONS*

MATTIAS SANDBERG†

Abstract. The convergence of reachable sets for nonconvex differential inclusions is considered. When the right-hand side in the differential inclusion is a compact-valued, Lipschitz continuous set-valued function it is shown that the convergence in Hausdorff distance of reachable sets for a forward Euler discretization is linear in the time step.

Key words. differential inclusion, forward Euler, convergence order, convexification

AMS subject classifications. 34A60, 65L20, 49M25

DOI. 10.1137/070686093

1. Introduction. A natural method to approximate solutions to a differential inclusion is the forward Euler method. Before this method is described, some notation is introduced. We denote by B the closed unit ball in \mathbb{R}^d . The Minkowski sum of two nonempty sets $C, D \subset \mathbb{R}^d$, is defined by

$$C + D = \{c + d \mid c \in C \text{ and } d \in D\},$$

the multiplication by a scalar, $\lambda > 0$, by

$$\lambda C = \{\lambda c \mid c \in C\},$$

and the sum of an element $c \in \mathbb{R}^d$ and a set C by

$$c + C = \{c\} + C.$$

The Hausdorff distance is given by

$$\mathcal{H}(C, D) = \inf \{\lambda \geq 0 \mid C \subset D + \lambda B \text{ and } D \subset C + \lambda B\},$$

and the convex hull of a set C is denoted $\text{co}(C)$. The Euclidean norm is denoted $|\cdot|$. The differential inclusion to be studied is

$$(1.1) \quad \begin{aligned} x'(t) &\in F(x(t)), \\ x(0) &= x_0, \end{aligned}$$

where $x_0 \in \mathbb{R}^d$ is the starting position. The function F is *set-valued* $\mathbb{R}^d \rightsquigarrow \mathbb{R}^d$, i.e., for each $x \in \mathbb{R}^d$, $F(x)$ is a subset of \mathbb{R}^d . It is assumed that the images $F(x)$ are compact sets, uniformly bounded in the sense that

$$(1.2) \quad |y| \leq K \quad \text{for all } y \in \bigcup_{x \in \mathbb{R}^d} F(x),$$

*Received by the editors March 23, 2007; accepted for publication (in revised form) July 7, 2008; published electronically November 26, 2008.

<http://www.siam.org/journals/sinum/47-1/68609.html>

†Centre of Mathematics for Applications, University of Oslo, P.O. Box 1053 Blindern, NO-0316 Oslo, Norway (mattias.sandberg@cma.uio.no).

and are Lipschitz continuous with respect to the Hausdorff distance:

$$(1.3) \quad \mathcal{H}(F(x), F(y)) \leq L|x - y| \quad \text{for all } x, y \in \mathbb{R}^d.$$

These assumptions are made for the sake of convenient notation, and could clearly be relaxed. We could, e.g., change the global Lipschitz continuity in (1.3) to local Lipschitz continuity, since, by the boundedness of F , the solutions to the differential inclusion remain in a bounded set in the finite time interval we will consider. It is also possible to treat the nonautonomous inclusion $x'(t) \in F(t, x(t))$ with the theory presented for the autonomous inclusion (1.1), if there is Lipschitz continuity in both space and time, i.e.,

$$(1.4) \quad \mathcal{H}(F(t, x), F(s, y)) \leq L(|t - s| + |x - y|).$$

This is done by introducing the variable $z = (t, x) \in \mathbb{R}^{d+1}$, which solves the inclusion

$$(1.5) \quad \begin{aligned} z' &\in \tilde{F}(z), \\ z(0) &= (0, x_0), \end{aligned}$$

where $\tilde{F}(z) = (1, F(z_1, z_s))$, and where z_1 is the first coordinate of the vector z and z_s is the vector containing the last d coordinates in z . Note that the inclusion (1.5) is autonomous in the variable z , with a right-hand side which by (1.4) is Lipschitz continuous with constant $\sqrt{2}L$.

By a solution to the differential inclusion in some interval, say $[0, T]$, we mean an absolutely continuous function $x : [0, T] \rightarrow \mathbb{R}^d$ such that (1.1) holds almost everywhere in $[0, T]$. General theory on differential inclusions can be found in [2, 3, 8]. This paper is about how well reachable sets of the differential inclusion can be approximated by reachable sets of a forward Euler discretization of it. The interval $[0, T]$ is split into N equal parts so that the step length is $\Delta t = T/N$. We shall consider solutions $\{\xi_n\}_{n=0}^N$ to the forward Euler discretized inclusion

$$(1.6) \quad \begin{aligned} \xi_{n+1} &\in \xi_n + \Delta t F(\xi_n), \quad n = 0, 1, \dots, N-1, \\ \xi_0 &= x_0. \end{aligned}$$

For convex differential inclusions, i.e., those where the function F is convex-valued, it has been shown in [9] that the forward Euler method converges with rate one. If the differential inclusion (1.1) is not convex, the relaxation theorem, restated in Theorem 2.2 (see also [2]), gives that its set of solutions is dense in the set of solutions to its convexified version, where $F(x)$ has been changed to the convex hull $\text{co}(F(x))$. Therefore, a straightforward way to approximate solutions of a nonconvex differential inclusion is to first convexify, and then to use the convergence result for the Euler method applied to convex differential inclusions, mentioned above. The new result presented here, in Theorem 2.5, is that the forward Euler method is convergent with rate one, even *without* the step of convexification.

The convergence result in Theorem 2.5 is weaker than the result for convex problems in [9] in two ways. First, the convergence is in the sense of the Hausdorff distance between reachable sets, to be explained in section 2, whereas the result for convex problems concerns convergence of entire solution paths. Second, the reachable sets to the forward Euler scheme are shown to be of the order $d^2\Delta t$ from the reachable sets to the original differential inclusion (1.1), measured in Hausdorff distance; i.e., there is a dependence on the dimension which is not present in the result for convex differential inclusions.

A benefit of using the forward Euler method directly, without convexifying, is that when applied to optimal control problems, the minimization in the approximating optimization problem might be over a finite set. The connection with optimal control is discussed further in section 2.1.

It will be shown that the Hausdorff distance between reachable sets for (1.1) and (1.6) is of the order Δt . When this is proved we use the fact that the forward Euler method for convex differential inclusions converges with rate one. The convexified variant of (1.6) is therefore also introduced:

$$(1.7) \quad \begin{aligned} \eta_{n+1} &\in \eta_n + \Delta t \operatorname{co}(F(\eta_n)), \quad n = 0, 1, \dots, N-1, \\ \eta_0 &= x_0. \end{aligned}$$

In this paper the forward Euler method for approximating differential inclusions is treated. This method is studied for convex problems in [1, 9, 13, 14]. For convex differential inclusions there are also other methods available; for surveys on such approximation methods, see [10, 15]. Under a condition of strong convexity, a Runge–Kutta-type scheme is shown to be convergent of order two in [18]. In [19] the second order convergence is established for linear differential inclusions that are convex, but not necessarily strongly convex.

2. The convergence result. Introduce the *reachable sets*

$$(2.1) \quad \begin{aligned} C_n &= \{x(n\Delta t) \mid x : [0, T] \rightarrow \mathbb{R}^d \text{ solution to (1.1)}\}, \\ D_n &= \{\xi_n \mid \{\xi_i\}_{i=0}^N \text{ solution to (1.6)}\}, \\ E_n &= \{\eta_n \mid \{\eta_i\}_{i=0}^N \text{ solution to (1.7)}\}. \end{aligned}$$

The main result is that the Hausdorff distance between the sets C_n and D_n is of the order Δt . The proof uses two previously known results, and a new one. Before stating these, a lemma about the Lipschitz continuity of $\operatorname{co}(F(x))$ is formulated.

LEMMA 2.1. *Let F be a function from \mathbb{R}^d into the compact subsets of \mathbb{R}^d which satisfies (1.3). Then*

$$\mathcal{H}(\operatorname{co}(F(x)), \operatorname{co}(F(y))) \leq L|x - y|.$$

Proof. This is a direct consequence of the fact that for any nonempty compact sets A and B in \mathbb{R}^d ,

$$\mathcal{H}(\operatorname{co}(A), \operatorname{co}(B)) \leq \mathcal{H}(A, B),$$

a fact which may be found in, e.g., [20]. \square

We now formulate the known results used in the proof of the convergence result for nonconvex differential inclusions in Theorem 2.5. The first is the relaxation theorem (see, e.g., [2]), which states that the set of solutions to the differential inclusion (1.1) is dense in the set of solutions to its convexified version, where $F(x)$ has been changed to $\operatorname{co}(F(x))$.

THEOREM 2.2. *Let F be a Lipschitz continuous function from \mathbb{R}^d into the nonempty compact subsets of \mathbb{R}^d . Let $y : [0, T] \rightarrow \mathbb{R}^d$ be a solution to*

$$y'(t) \in \operatorname{co}(F(y(t))), \quad y(0) = x_0.$$

Then for every positive ε , there exists a solution $x : [0, T] \rightarrow \mathbb{R}^d$ to (1.1) such that for all $t \in [0, T]$, $|x(t) - y(t)| \leq \varepsilon$.

The relaxation theorem makes it possible to move from the setting of nonconvex differential inclusions to the convex one. For convex differential inclusions it has been proved in [9] that the forward Euler method converges with rate one. The result in [9] is formulated in a slightly more general setting than here, but we give a version adapted to the present assumptions.

THEOREM 2.3. *Let F be a function from \mathbb{R}^d into the nonempty compact convex subsets of \mathbb{R}^d , which satisfies (1.2) and (1.3). For any solution $x : [0, T] \rightarrow \mathbb{R}^d$ to (1.1) there exists a solution $\{\xi_n\}_{n=0}^N$ to (1.6) such that*

$$(2.2) \quad \max_{0 \leq n \leq N} |x(n\Delta t) - \xi_n| \leq KLT e^{LT} \Delta t.$$

Moreover, for any solution $\{\xi_n\}_{n=0}^N$ to (1.6) there exists a solution $x : [0, T] \rightarrow \mathbb{R}^d$ to (1.1) such that (2.2) holds. Hence the sets C_n and D_n , defined in (2.1), satisfy

$$(2.3) \quad \max_{0 \leq n \leq N} \mathcal{H}(C_n, D_n) \leq KLT e^{LT} \Delta t.$$

Remark 2.4. Even though the constant in front of Δt in (2.2) and (2.3) does not occur explicitly in [9], it can be found by working through the proofs.

One method for approximation of (1.1) is to use a forward Euler scheme with a convexified right-hand side. Theorems 2.2 and 2.3 show that this method converges with order one. This paper, however, concerns the forward Euler scheme with the right-hand side unchanged. What remains to show in order to prove the desired convergence for this method is that the Hausdorff distance between the reachable sets to (1.6) and to (1.7) is of the order Δt . This is shown in section 3, but first we state the result for the forward Euler method for nonconvex differential inclusions.

THEOREM 2.5. *Let F be a function from \mathbb{R}^d into the nonempty compact subsets of \mathbb{R}^d which satisfies (1.2) and (1.3). Then the sets C_n and D_n , defined in (2.1), satisfy*

$$(2.4) \quad \max_{0 \leq n \leq N} \mathcal{H}(C_n, D_n) \leq (Ke^{LT}(Kd(d+1) + LT) + 2Kd)\Delta t.$$

Proof. Theorem 2.2 implies that

$$\bar{C}_n = \{y(n\Delta t) \mid y : [0, T] \rightarrow \mathbb{R}^d \text{ solves } y'(t) \in \text{co}(F(y(t))), y(0) = x_0\},$$

where the bar denotes set closure. Lemma 2.1 and Theorem 2.3 imply that

$$(2.5) \quad \max_{0 \leq n \leq N} \mathcal{H}(\bar{C}_n, E_n) \leq KLT e^{LT} \Delta t,$$

while Theorem 3.6 says that

$$(2.6) \quad \max_{0 \leq n \leq N} \mathcal{H}(D_n, E_n) \leq (Ke^{LT}d(d+1) + 2Kd)\Delta t.$$

Since $\mathcal{H}(C_n, D_n) = \mathcal{H}(\bar{C}_n, D_n)$, (2.4) follows by (2.5) and (2.6). \square

Let us compare Theorem 2.5 with other results regarding the convergence of the forward Euler method for differential inclusions. First, we note that the constant in front of Δt in (2.4) depends on the dimension d , while such explicit dependence on the dimension is not present in the estimate for convex differential inclusions in Theorem 2.3. Although this dependence on d might be weaker, it is not possible to avoid it completely. Consider, e.g., the case when F is constant and consists of the

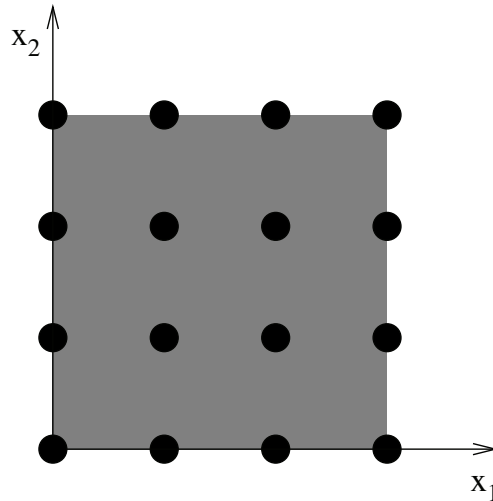


FIG. 2.1. Part of the sets C_n (shaded) and D_n (dots), for some n , in two dimensions when $F = \{(0, 0), (1, 0), (0, 1)\}$. The closest distance between the dots is Δt , so the distance “center of square”—“corner of square” is $\frac{1}{2}\sqrt{2}\Delta t$, and hence $\mathcal{H}(C_n, D_n) = \frac{1}{2}\sqrt{2}\Delta t$. The analogous situation in dimension d gives $\mathcal{H}(C_n, D_n) = \frac{1}{2}\sqrt{d}\Delta t$.

unit vectors in the coordinate directions in \mathbb{R}^d , together with the zero vector. In this situation the Hausdorff distance between the sets C_n and D_n in Theorem 2.5 will be of the order $\sqrt{d}\Delta t$; see Figure 2.1. In [11] the same square root dependence on the dimension is shown to hold when the integral of a nonconvex set-valued function is approximated by Riemann sums.

In [12], equation (1.1) is considered with F a function from \mathbb{R}^d into the compact convex subsets of \mathbb{R}^d . The forward Euler approximation is however performed with the scheme

$$\begin{aligned}\xi_{n+1} &\in \xi_n + \Delta t G(\xi_n), \quad n = 0, 1, \dots, N-1, \\ \xi_0 &= x_0,\end{aligned}$$

where $G(x)$ is either $\partial F(x)$, the boundary points of $F(x)$, or $\text{ext}F(x)$, the set of extreme points of F . When $G(x) = \partial F(x)$, linear convergence is proved, while the result for $G(x) = \text{ext}F(x)$, assuming Lipschitz continuity of G , only is convergence of order $\sqrt{\Delta t}$. Theorem 2.5 covers the same situation if we let $\text{ext}F(x)$ in [12] be the F we are using in this paper. It improves the half-order convergence in [12] partially, since it proves linear convergence. The convergence in Theorem 2.5 is however of a weaker form as it gives convergence of the sets D_n to the sets C_n , for all n , while the result in [12] is of the same kind as in Theorem 2.3, i.e., convergence of approximating paths. The result in Theorem 2.5 is however what is needed in many applications. The relevance for optimal control is discussed in the following section.

2.1. Convergence of approximations in optimal control. We will consider the optimal control problem to minimize the functional

$$(2.7) \quad \int_0^T h(x(t), \alpha(t)) dt + g(x(T))$$

over all solutions to the equation

$$(2.8) \quad x'(t) = f(x(t), \alpha(t)), \quad x(0) = x_s,$$

where $\alpha : [0, T] \rightarrow D$ is the control variable. In order to be able to use the result in Theorem 2.5 to prove a convergence result for approximations of optimal control, the following result on equivalence between control problems and differential inclusions is needed. The theorem is taken from [2, Corollary 1, p. 91].

THEOREM 2.6. *Let $f : \mathbb{R}^d \times D \rightarrow \mathbb{R}^d$ be continuous where D is a compact separable metric space and assume that there exist a $T > 0$ and an absolutely continuous $x : [0, T] \rightarrow \mathbb{R}^d$ such that*

$$(2.9) \quad x'(t) \in \bigcup_{a \in D} f(x(t), a) \quad \text{for almost every } t \in [0, T].$$

Then there exists a Lebesgue measurable $\alpha : [0, T] \rightarrow D$ such that for almost every $t \in [0, T]$, (2.8) holds.

Remark 2.7. The theorem shows that every solution to the differential inclusion (2.9) is also a solution to the control equation (2.8) with a measurable control α . That every solution to (2.8) is a solution to (2.9) is obvious. Hence the solution sets to the control equation and the corresponding differential inclusion coincide.

With this duality between optimal control and differential inclusions, it is possible to use Theorem 2.5 to prove the following convergence result for the optimal value of the functional in (2.7).

THEOREM 2.8. *Let D and f satisfy the conditions in Theorem 2.6, and let $h : \mathbb{R}^d \times D \rightarrow \mathbb{R}$ be continuous. Let also f and h be such that the set-valued function $J : \mathbb{R}^{d+1} \rightsquigarrow \mathbb{R}^{d+1}$ defined by $J(y) = \bigcup_{a \in D} (f(x, a), h(x, a))$, where x is the vector consisting of the d first components in y , has compact values, and is uniformly bounded and uniformly Lipschitz continuous in \mathbb{R}^d . Let the function $g : \mathbb{R}^d \rightarrow \mathbb{R}$ be uniformly Lipschitz continuous. Furthermore, let x_s be an element in \mathbb{R}^d , let T be a positive real number, and split the interval $[0, T]$ into N equal parts of length $\Delta t = T/N$. Then*

$$|u - \bar{u}| = \mathcal{O}(\Delta t),$$

where the values u and \bar{u} are defined by

$$(2.10) \quad \begin{aligned} u = \inf \left\{ \int_0^T h(x(t), \alpha(t)) dt + g(x(T)) \mid \alpha : [0, T] \rightarrow D \text{ measurable,} \right. \\ \left. x : [0, T] \rightarrow \mathbb{R}^d \text{ measurable and satisfies (2.8) a.e. in } [0, T] \right\}, \\ \bar{u} = \inf \left\{ \Delta t \sum_0^{N-1} h(x_n, \alpha_n) + g(x_N) \mid \right. \\ \left. x_{n+1} = x_n + \Delta t f(x_n, \alpha_n), x_0 = x_s, \alpha_n \in D \text{ for all } 0 \leq n \leq N-1 \right\}. \end{aligned}$$

Remark 2.9. The conditions on the set-valued function J , appearing in the theorem, are satisfied, e.g., if the set D is a compact subset of a Euclidean space \mathbb{R}^n

(any n), and the functions f and h are uniformly bounded and uniformly Lipschitz continuous in $\mathbb{R}^d \times D$.

Proof. The value u can be written as

$$u = \inf \{ \tilde{g}(y(T)) \mid y : [0, T] \rightarrow \mathbb{R}^{d+1} \text{ measurable and} \\ y'(t) = p(y(t), \alpha(t)) \text{ a.e. } t \in [0, T], y(0) = (x_s, 0), \alpha : [0, T] \rightarrow D \text{ measurable} \},$$

where $p(y, a) = (f(x, a), h(x, a))$, $\tilde{g}(y) = g(x) + y_{d+1}$, and x and y_{d+1} are the d first components and the last component in y , respectively. The function p inherits the fact that it satisfies the conditions in Theorem 2.6 from f and h . Hence it is possible to express the value u in yet another way,

$$u = \inf \{ \tilde{g}(y(T)) \mid y : [0, T] \rightarrow \mathbb{R}^{d+1} \text{ measurable,} \\ y'(t) \in J(y(t)), \text{ a.e. } t \in [0, T], y(0) = (x_s, 0) \}.$$

Similarly, the value \bar{u} can be written as

$$\bar{u} = \inf \{ \tilde{g}(y_N) \mid y_{n+1} \in y_n + \Delta t J(y_n) \text{ for all } 0 \leq n \leq N - 1, y_0 = (x_s, 0) \}.$$

Pick an $\varepsilon > 0$ and let $z : [0, T] \rightarrow \mathbb{R}^{d+1}$ be an absolutely continuous function such that $z'(t) \in J(z(t))$ for a.e. $t \in [0, T]$, $z(0) = (x_s, 0)$, and such that $\tilde{g}(z(T)) \leq u + \varepsilon$.

As J satisfies the conditions in Theorem 2.5, it follows that there exists $\{z_n\}_0^N$ with $|z_N - z(T)| = \mathcal{O}(\Delta t)$ such that $z_{n+1} \in z_n + \Delta t J(z_n)$ for $0 \leq n \leq N$ and $z_0 = (x_s, 0)$. The terminal cost function \tilde{g} inherits Lipschitz continuity from g , and therefore $|\tilde{g}(z_N) - \tilde{g}(z(T))| = \mathcal{O}(\Delta t)$. Since $\bar{u} \leq \tilde{g}(z_N)$, it follows that there exists a constant C which is independent of Δt such that $\bar{u} - u \leq C\Delta t + \varepsilon$. As ε was arbitrary it may be removed from this expression. The relation $u - \bar{u} \leq C\Delta t$ is proved similarly. \square

In cases where the control set D is finite, the minimization in (2.10) is over a finite set, while convexification generally results in minimization over infinite sets.

The result in the theorem uses a minimization directly in the control variable. Another method for approximation of optimal control problems based on analysis of the underlying Hamilton–Jacobi equation can be found in [17]. General theory for optimal control problems can be found in [5, 6, 7, 8, 16].

3. Convexification of the forward Euler scheme. In order to be able to show that the distance between the solution sets to (1.6) and (1.7) is of the order Δt we introduce two set-valued maps. Let φ and ψ be functions from \mathbb{R}^d into the nonempty compact subsets of \mathbb{R}^d , defined by

$$\varphi(x) = x + \Delta t F(x), \\ \psi(x) = x + \Delta t \text{co}(F(x)).$$

If A is a subset of \mathbb{R}^d we define

$$\varphi(A) = \bigcup_{x \in A} \varphi(x),$$

and similarly for the set-valued maps ψ and F . The solution set to the forward Euler equation (1.6) is given by iterates of the function φ , while the solution set to (1.7) is given by iterates of ψ . We therefore introduce the notation

$$\varphi^n(x_0) = \underbrace{\varphi \circ \varphi \circ \dots \circ \varphi}_{n}(x_0),$$

and similarly for ψ .

It will be shown that

$$(3.1) \quad \max_{0 \leq n \leq N} \mathcal{H}(\varphi^n(x_0), \psi^n(x_0)) = \mathcal{O}(\Delta t).$$

Since we have the immediate inclusion

$$\varphi^n(x_0) \subset \psi^n(x_0),$$

this amounts to proving that

$$(3.2) \quad \psi^n(x_0) \subset \varphi^n(x_0) + \mathcal{O}(\Delta t)B \quad \text{for all } 0 \leq n \leq N.$$

When proving this, the Carathéodory theorem will play an important role.

THEOREM 3.1. *The convex hull of an arbitrary subset A of \mathbb{R}^d is given by*

$$\text{co}(A) = \left\{ \sum_{i=1}^{d+1} \lambda_i a_i \mid a_i \in A, \lambda_i \geq 0, \sum_{i=1}^{d+1} \lambda_i = 1 \right\}.$$

For a proof, see, e.g., [4].

The first step in establishing (3.2) is to show that

$$(3.3) \quad \psi^{d+1}(x_0) \approx \psi^d(\varphi(x_0)).$$

Theorem 3.2 shows that equality holds in (3.3) when F is constant.

THEOREM 3.2. *Let P be any nonempty subset of \mathbb{R}^d . For any integer $s \geq d$ it holds that*

$$(3.4) \quad (s + 1) \text{co}(P) = P + s \text{co}(P).$$

Proof. We begin by considering the case when P is a set containing $d + 1$ points,

$$P = \{p_i\}_{i=1}^{d+1},$$

which do not lie in a $(d - 1)$ -dimensional hyperplane; i.e., the dimension of $\text{co}(P)$ is d . To start with, (3.4) is shown with $s = d$. It holds that

$$P + d \text{co}(P) \subset (d + 1) \text{co}(P),$$

since

$$\text{co}(P + d \text{co}(P)) = \text{co}(P) + d \text{co}(P) = (d + 1) \text{co}(P).$$

It therefore remains to show that

$$(3.5) \quad P + d \text{co}(P) \supset (d + 1) \text{co}(P).$$

The polytope spanned by P is translated so that it is centered at the origin of \mathbb{R}^d ; i.e., it is assumed that

$$\sum_{i=1}^{d+1} p_i = 0.$$

Denote by Q a polytope spanned by $d + 1$ unit vectors, $\{q_i\}_{i=1}^{d+1}$, directed so that the angle between any two unit vectors is the same as the angle between any other two vectors and such that

$$\sum_{i=1}^{d+1} q_i = 0.$$

There exists a linear transformation $A : \mathbb{R}^d \rightarrow \mathbb{R}^d$ such that $A(\text{co}(P)) = Q$. It can be constructed by demanding that $Ap_i = q_i$, $i = 1, \dots, d$. Then also

$$Ap_{d+1} = A(-p_1 - p_2 - \dots - p_d) = -q_1 - q_2 - \dots - q_d = q_{d+1}.$$

All the sets $p_i + d \text{co}(P)$, $i = 1, \dots, d+1$ contain the zero vector. Furthermore, all these sets have d faces lying in the same planes as d faces of $(d+1) \text{co}(P)$. The remaining face of $p_i + d \text{co}(P)$ is parallel to the remaining face of $(d+1) \text{co}(P)$. Assume there exists a point $p \in (d+1) \text{co}(P)$ such that $p \notin P + d \text{co}(P)$. Denote by e_i the outward normal of the face of $A(p_i + d \text{co}(P))$ which contains the origin. It must then hold that

$$(3.6) \quad e_i \cdot Ap > 0, \quad i = 1, \dots, d+1.$$

Because of the symmetry, $\sum_{i=1}^{d+1} e_i = 0$, which contradicts (3.6). Hence (3.5) holds.

By this it easily follows that (3.4) holds also for $s > d$ in the situation where P contains $d + 1$ points and has a d -dimensional convex hull:

$$\begin{aligned} P + s \text{co}(P) &= P + d \text{co}(P) + (s - d) \text{co}(P) \\ &= (d + 1) \text{co}(P) + (s - d) \text{co}(P) = (s + 1) \text{co}(P). \end{aligned}$$

Now consider the case when P is any subset of \mathbb{R}^d such that $\text{co}(P)$ has dimension d . Denote by $\{P_d\}$ the set of all subsets of P with $d + 1$ elements, such that the dimension of $\text{co}(P_d)$ is d . By Carathéodory's theorem (Theorem 3.1) it then follows that

$$\text{co}(P) = \bigcup_{\{P_d\}} \text{co}(P_d).$$

With the result for sets containing $d + 1$ points already proved, we have that

$$(s + 1) \text{co}(P) = \bigcup_{\{P_d\}} (s + 1) \text{co}(P_d) = \bigcup_{\{P_d\}} (P_d + s \text{co}(P_d)) \subset P + s \text{co}(P).$$

As before, $P + s \text{co}(P) \subset (s + 1) \text{co}(P)$, since their convex hulls coincide. It therefore holds that $P + s \text{co}(P) = (s + 1) \text{co}(P)$.

It remains to show that (3.4) also holds when the dimension of $\text{co}(P)$ is lower than d . In this case there exists a vector $v \in \mathbb{R}^d$ such that $R \equiv v + P$ is contained in a linear subspace which also contains the origin. Consider the linear subspace of the dimension equal to the dimension of $\text{co}(P)$. By what has already been proved, (3.4) holds for the set R . This together with the fact that

$$\text{co}(R) = v + \text{co}(P)$$

implies that

$$v + P + s(v + \text{co}(P)) = (s + 1)(v + \text{co}(P)),$$

so that (3.4) holds also for P . \square

When F is not constant, equality does not hold in (3.3), but instead we have the following inclusion.

THEOREM 3.3.

$$(3.7) \quad \psi^{d+1}(x_0) \subset \psi^d(\varphi(x_0)) + KLd(d+1)\Delta t^2 B \quad \text{for any } x_0 \in \mathbb{R}^d.$$

Proof. By the definition of ψ , and by the simple fact that $F(\psi^i(x_0)) \supset F(x)$ for $x \in \psi^i(x_0)$, it follows that

$$(3.8) \quad \psi^{d+1}(x_0) \subset x_0 + \Delta t \operatorname{co}(F(x_0)) + \Delta t \operatorname{co}(F(\psi(x_0))) + \cdots + \Delta t \operatorname{co}(F(\psi^d(x_0))).$$

The boundedness of F implies

$$\mathcal{H}(\psi^i(x_0), x_0) \leq iKL\Delta t,$$

and by the Lipschitz continuity of $\operatorname{co}(F)$ (Lemma 2.1) it therefore holds that

$$\operatorname{co}(F(\psi^i(x_0))) \subset \operatorname{co}(F(x_0)) + iKL\Delta t B.$$

This fact in (3.8) yields

$$(3.9) \quad \begin{aligned} \psi^{d+1}(x_0) &\subset x_0 + (d+1)\Delta t \operatorname{co}(F(x_0)) + \sum_{i=1}^d iKL\Delta t^2 B \\ &= x_0 + (d+1)\Delta t \operatorname{co}(F(x_0)) + KL \frac{d(d+1)}{2} \Delta t^2 B \\ &= x_0 + \Delta t F(x_0) + d\Delta t \operatorname{co}(F(x_0)) + KL \frac{d(d+1)}{2} \Delta t^2 B, \end{aligned}$$

where the last equality follows by Theorem 3.2. Once again, the Lipschitz continuity of $\operatorname{co}(F)$ is used, so that

$$\operatorname{co}(F(x_0)) \subset \operatorname{co}(F(\psi^i(\varphi(x_0)))) + (i+1)KL\Delta t B.$$

This fact will be used to prove that

$$(3.10) \quad x_0 + \Delta t F(x_0) + d\Delta t \operatorname{co}(F(x_0)) \subset \psi^d(\varphi(x_0)) + KL \frac{d(d+1)}{2} \Delta t^2 B.$$

Let z be any element in $x_0 + \Delta t F(x_0) + d\Delta t \operatorname{co}(F(x_0))$. It then holds that

$$z = x_0 + \Delta t x_1 + \Delta t x_2 + \cdots + \Delta t x_{d+1},$$

where $x_1 \in F(x_0)$, and $x_2, \dots, x_{d+1} \in \operatorname{co}(F(x_0))$. Now let $\tilde{x}_1 = x_1$, let \tilde{x}_2 be the projection of x_2 on $\operatorname{co}(F(x_0 + \Delta t \tilde{x}_1))$ (i.e., the element in $\operatorname{co}(F(x_0 + \Delta t \tilde{x}_1))$ such that $d(x_2, \operatorname{co}(F(x_0 + \Delta t \tilde{x}_1))) = |\tilde{x}_2 - x_2|$), let \tilde{x}_3 be the projection of x_3 on $\operatorname{co}(F(x_0 + \Delta t \tilde{x}_1 + \Delta t \tilde{x}_2))$, and so on. The Lipschitz continuity of $\operatorname{co}(F)$ (Lemma 2.1) implies that

$$|\tilde{x}_i - x_i| \leq (i-1)KL\Delta t \quad \text{for } i = 2, \dots, d+1.$$

From this (3.10) follows, which together with (3.9) gives (3.7). \square

The next step to establish (3.2) is to show that

$$\psi^{d+n}(x_0) \subset \psi^d(\varphi^n(x_0)) + \mathcal{O}(\Delta t)B.$$

This is done in Theorems 3.4 and 3.5. The desired result, relation (3.2), follows from this and is given in Theorem 3.6.

THEOREM 3.4. *Assume that*

$$(3.11) \quad \psi^{d+n}(x_0) \subset \psi^d(\varphi^n(x_0)) + \varepsilon B.$$

Then

$$(3.12) \quad \psi^{d+n+1}(x_0) \subset \psi^d(\varphi^{n+1}(x_0)) + (KLd(d+1)\Delta t^2 + \varepsilon(1+L\Delta t))B.$$

Proof. By (3.11),

$$(3.13) \quad \psi^{d+n+1}(x_0) \subset \psi(\psi^d(\varphi^n(x_0)) + \varepsilon B).$$

Consider elements

$$\begin{aligned} z &\in \psi(\psi^d(\varphi^n(x_0)) + \varepsilon B), \\ x &\in \psi^d(\varphi^n(x_0)) + \varepsilon B \end{aligned}$$

such that $z \in x + \Delta t \operatorname{co}(F(x))$. Let x' be an element in $\psi^d(\varphi^n(x_0))$ such that $|x' - x| \leq \varepsilon$. The Lipschitz continuity of $\operatorname{co}(F)$ (Lemma 2.1) gives the inclusion

$$\operatorname{co}(F(x)) \subset \operatorname{co}(F(x')) + L\varepsilon B.$$

It therefore holds that

$$z \in x' + \Delta t \operatorname{co}(F(x')) + \varepsilon(1+L\Delta t)B = \psi(x') + \varepsilon(1+L\Delta t)B.$$

Hence,

$$\begin{aligned} \psi(\psi^d(\varphi^n(x_0)) + \varepsilon B) &\subset \psi^{d+1}(\varphi^n(x_0)) + \varepsilon(1+L\Delta t)B \\ &\subset \psi^d(\varphi^{n+1}(x_0)) + (KLd(d+1)\Delta t^2 + \varepsilon(1+L\Delta t))B, \end{aligned}$$

where the last inclusion follows from Theorem 3.3. Together with (3.13) this shows (3.12). \square

THEOREM 3.5. *For all $1 \leq n \leq N - d$ the following inclusion holds:*

$$(3.14) \quad \begin{aligned} \psi^{d+n}(x_0) &\subset \psi^d(\varphi^n(x_0)) + Ke^{LTn/N}d(d+1)\Delta tB \\ &\subset \psi^d(\varphi^n(x_0)) + Ke^{LT}d(d+1)\Delta tB. \end{aligned}$$

Proof. By Theorems 3.3 and 3.4 there are constants ε^n satisfying

$$(3.15) \quad \begin{aligned} \varepsilon^1 &= KLd(d+1)\Delta t^2, \\ \varepsilon^{n+1} &= KLd(d+1)\Delta t^2 + \varepsilon^n(1+L\Delta t) \quad \text{for } n \geq 1 \end{aligned}$$

such that

$$\psi^{d+n}(x_0) \subset \psi^d(\varphi^n(x_0)) + \varepsilon^n B.$$

By (3.15) the constants ε^n satisfy

$$\begin{aligned}\varepsilon^n &= K L d(d+1) \Delta t^2 \sum_{i=0}^{n-1} (1+L \Delta t)^i = K L d(d+1) \Delta t^2 \frac{(1+L \Delta t)^n - 1}{L \Delta t} \\ &= K((1+L \Delta t)^n - 1) d(d+1) \Delta t \leq K e^{L T n / N} d(d+1) \Delta t \\ &\leq K e^{L T} d(d+1) \Delta t,\end{aligned}$$

which proves (3.14). \square

THEOREM 3.6. *For all $0 \leq n \leq N$ the following inclusion holds:*

$$(3.16) \quad \psi^n(x_0) \subset \varphi^n(x_0) + (K e^{L T} d(d+1) + 2Kd) \Delta t B.$$

Hence the sets D_n and E_n , defined in (2.1), satisfy

$$\max_{0 \leq n \leq N} \mathcal{H}(D_n, E_n) \leq (K e^{L T} d(d+1) + 2Kd) \Delta t.$$

Proof. The bound (1.2) on the function F implies that for any $m \geq 1$

$$\varphi^d(\varphi^m(x_0)) \subset \varphi^m(x_0) + Kd \Delta t B \subset \varphi^{d+m}(x_0) + 2Kd \Delta t B.$$

Together with Theorem 3.5 this implies (3.16) when $n \geq d+1$. When $n \leq d$ it holds that

$$(3.17) \quad \psi^n(x_0) \subset \varphi^n(x_0) + 2Kn \Delta t B \subset \varphi^n(x_0) + 2Kd \Delta t B,$$

so (3.16) holds for all $0 \leq n \leq N$. \square

Acknowledgments. I would like to thank Anders Szepessy for proofreading this article and suggesting improvements. I would also like to thank the referees for their constructive comments and suggestions which have improved the paper.

REFERENCES

- [1] Z. ARTSTEIN, *First-order approximations for differential inclusions*, Set-Valued Anal., 2 (1994), pp. 7–17.
- [2] J.-P. AUBIN AND A. CELLINA, *Differential Inclusions*, Grundlehren Math. Wiss. 264, Springer-Verlag, Berlin, 1984.
- [3] J.-P. AUBIN AND H. FRANKOWSKA, *Set-Valued Analysis*, Systems Control Found. Appl. 2, Birkhäuser Boston, Boston, MA, 1990.
- [4] M. BERGER, *Geometry*. I, Universitext, Springer-Verlag, Berlin, 1977 (in French); Universitext, Springer-Verlag, Berlin, 1987 (in English).
- [5] D. P. BERTSEKAS, *Dynamic Programming and Optimal Control*. Vol. II, 2nd ed., Athena Scientific, Belmont, MA, 2001.
- [6] D. P. BERTSEKAS, *Dynamic Programming and Optimal Control*. Vol. I, 3rd ed., Athena Scientific, Belmont, MA, 2005.
- [7] P. CANNARSA AND C. SINISTRARI, *Semiconcave Functions, Hamilton–Jacobi Equations, and Optimal Control*, Prog. Nonlinear Differential Equations Appl. 58, Birkhäuser Boston, Boston, MA, 2004.
- [8] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Canad. Math. Soc. Ser. Monogr. Adv. Texts, John Wiley & Sons, New York, 1983.
- [9] A. L. DONTCHEV AND E. M. FARKHI, *Error estimates for discretized differential inclusion*, Computing, 41 (1989), pp. 349–358.
- [10] A. DONTCHEV AND F. LEMPIO, *Difference methods for differential inclusions: A survey*, SIAM Rev., 34 (1992), pp. 263–294.

- [11] N. DYN AND E. FARKHI, *Set-valued approximations with Minkowski averages—convergence and convexification rates*, Numer. Funct. Anal. Optim., 25 (2004), pp. 363–377.
- [12] G. GRAMMEL, *Towards fully discretized differential inclusions*, Set-Valued Anal., 11 (2003), pp. 1–8.
- [13] F. LEMPPIO, *Modified Euler methods for differential inclusions*, in Set-Valued Analysis and Differential Inclusions (Pamporovo, 1990), Progr. Systems Control Theory 16, Birkhäuser Boston, Boston, MA, 1993, pp. 131–148.
- [14] F. LEMPPIO, *Euler’s method revisited*, Trudy Mat. Inst. Steklov., 211 (1995), pp. 473–494.
- [15] F. LEMPPIO AND V. VELIOV, *Discrete approximations of differential inclusions*, Bayreuth. Math. Schr., 54 (1998), pp. 149–232.
- [16] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Macmillan, New York, 1964.
- [17] M. SANDBERG AND A. SZEPESSY, *Convergence rates of symplectic Pontryagin approximations in optimal control theory*, M2AN Math. Model. Numer. Anal., 40 (2006), pp. 149–173.
- [18] V. VELIOV, *Second order discrete approximations to strongly convex differential inclusions*, Systems Control Lett., 13 (1989), pp. 263–269.
- [19] V. VELIOV, *Second-order discrete approximation to linear differential inclusions*, SIAM J. Numer. Anal., 29 (1992), pp. 439–451.
- [20] R. WEBSTER, *Convexity*, Oxford Sci. Publ., Clarendon Press, Oxford University Press, New York, 1994.

STABILITY PRESERVATION ANALYSIS FOR FREQUENCY-BASED METHODS IN NUMERICAL SIMULATION OF FRACTIONAL ORDER SYSTEMS*

MOHAMMAD SALEH TAVAZOEI[†], MOHAMMAD HAERI[†], SADEGH BOLOUKI[†], AND
MILAD SIAMI[†]

Abstract. In this paper, the frequency domain-based numerical methods for simulation of fractional order systems are studied in the sense of stability preservation. First, the stability boundary curve is exactly determined for these methods. Then, this boundary is analyzed and compared with an accurate (ideal) boundary in different frequency ranges. Also, the critical regions in which the stability does not preserve are determined. Finally, the analytical achievements are confirmed via some numerical illustrations.

Key words. fractional order system, frequency domain-based numerical method, fractional operator approximation, stability, stability boundary

AMS subject classifications. 26A33, 68U20, 34D23

DOI. 10.1137/080715949

1. Introduction. Fractional calculus as an extension of ordinary calculus is a mathematical topic with more than 300 years of history. Even though fractional calculus has a long history, its application to physics and engineering has attracted lots of attention only in the last few decades. It has been found that many real-world physical systems can be described by fractional differential equations. For instance, the fractional derivatives have been widely used in the mathematical modeling of viscoelastic materials [1, 2]. Some electromagnetic problems are described using fractional differentiation operators [3]. The anomalous diffusion phenomena in inhomogeneous media can be explained by noninteger derivative-based equations of diffusion [4, 5]. The RLC interconnect model of a transmission line is a fractional order model [6]. Heat conduction as a dynamical process can be more adequately modeled by fractional order models rather than their integer order counterparts [7]. In biology, it has been deduced that the membranes of the cells of a biological organism have fractional order electrical conductance [8] and then are classified in a group of noninteger order systems. In economy, it is known that some finance systems can display fractional order dynamics [9]. More examples from fractional order dynamics can be found in [10, 11] and references therein. Also, in recent years fractional order dynamic systems have been widely studied in the design and practice of control systems (for example, [12, 13, 14, 15, 16, 17]).

Although the integer order models can be considered as a special form of the more general fractional order models, there are basic differences between fractional order and integer order models. The main difference between them arises from an inherent attribute of fractional derivatives. In fact, the fractional derivatives are not local operators in opposition with integer derivatives that are local operators [11]. In other words, the fractional derivative of a function depends on its whole

*Received by the editors February 19, 2008; accepted for publication (in revised form) June 10, 2008; published electronically November 26, 2008. This work was supported financially by the Iranian National Science Foundation grant 86094/45.

<http://www.siam.org/journals/sinum/47-1/71594.html>

[†]Advanced Control System Lab, Electrical Engineering Department, Sharif University of Technology, Tehran 11155-9363, Iran (m.tavazoei@ee.sharif.edu, haeri@sina.sharif.edu, bolouki@ee.sharif.edu, siami@ee.sharif.edu).

past values. This property makes a fractional order model behave like a system with an “infinite memory” or “long memory.” Due to this property, simulation of these systems is more complicated [18]. Up to now, some analytic methods have been proposed to find numerical solution of the fractional differential equations (for example, [19, 20, 21, 22, 23, 24]). Since convergence, stability, and existence of bound for the estimation error have been proved for analytic methods, these direct methods are reliable and can be properly used for simulating fractional order systems. However, since the long memory behavior of these systems directly appears in the direct methods, simulation of fractional order systems via these methods sometimes requires a very long simulation time. Applying some ideas such as the short memory (fixed memory) principle and penalizing some forms of inaccuracy [11] may partly reduce the computational cost of time domain methods [25].

There is another popular way to simulate fractional order systems which is based on frequency domain approximation of fractional operators [26, 27, 28, 29, 30, 31, 32]. Simulation of a fractional order system by using rational approximation of the fractional operators consists of two steps as follows. First, the fractional order equation of the system is converted to the frequency domain, and the Laplace transform of the fractional integral operator is replaced by its integer order approximation. Then, the approximated equation in frequency domain is transformed back into the time domain. The resulted ordinary differential equation can now be numerically solved by applying the well-known numerical methods such as Runge–Kutta or the Adams–Bashforth–Moulton algorithm. Contrary to using direct methods, simulation of fractional order systems via using fractional operator approximation is simple, because, in this case, an ordinary differential system is simulated instead of the original fractional order system. But, unfortunately it has been shown that the results of simulations using the fractional operator approximation are not always reliable and the frequency-based numerical methods have some limitations in special cases [33, 34]. In this paper, a rigorous stability analysis is done to clear the problems arising from using frequency domain approximation in numerical simulations of fractional order systems. We show that this approximation can cause undesired changes in the stability. More precisely, stable systems may be converted to unstable approximated systems and vice versa.

This paper is organized as follows. Section 2 summarizes the basic concepts in the fractional calculus. Section 3 contains the main results of stability investigation for a frequency domain-based approximated system. Results of section 3 are numerically verified in section 4, and, finally, conclusions in section 5 close the paper.

2. Basic concepts. By extending the concept of integer order integral and derivative, the fractional integral and derivative have been defined. The definition of fractional integral is an outgrowth of the Cauchy formula for evaluating the integration. The q th order fractional integral of function $f(t)$ with respect to t is defined by [11]:

$$(1) \quad J^q f(t) = \frac{1}{\Gamma(q)} \int_0^t (t - \tau)^{q-1} f(\tau) d\tau,$$

where Γ is the Gamma function. Also, there are some definitions for fractional derivatives [11]. The Riemann–Liouville definition is the simplest and easiest definition to use. Based on this definition, the q th order fractional derivative of function $f(t)$ with respect to t and the terminal value 0 is given by

$$(2) \quad \frac{d^q f(t)}{dt^q} = \frac{1}{\Gamma(m - q)} \frac{d^m}{dt^m} \int_0^t (t - \tau)^{m-q-1} f(\tau) d\tau,$$

where m is the first integer larger than q , i.e., $m - 1 \leq q < m$. The Laplace transform of the Riemann–Liouville derivative is given as follows:

$$(3) \quad L \left\{ \frac{d^q f(t)}{dt^q} \right\} = s^q L\{f(t)\} - \sum_{k=0}^{m-1} s^k \frac{d^{q-k-1} f(0)}{dt^{q-k-1}}, \quad m - 1 < q \leq m.$$

Unfortunately, the Riemann–Liouville fractional derivative appears unsuitable to be treated by the Laplace transform technique in that it requires knowledge of the non-integer order derivatives of the function at $t = 0$. The mentioned problem does not exist in the Caputo definition of the fractional derivative. This definition of derivative, which is sometimes called a smooth fractional derivative, is described as

$$(4) \quad \frac{d^q f(t)}{dt^q} = \begin{cases} \frac{1}{\Gamma(m-q)} \int_0^t \frac{f^{(m)}(\tau)}{(t-\tau)^{q+1-m}} d\tau, & m - 1 < q \leq m, \\ \frac{d^m}{dt^m} f(t), & q = m, \end{cases}$$

where m is the first integer larger than q . The Laplace transform of the Caputo fractional derivative is

$$(5) \quad L \left\{ \frac{d^q f(t)}{dt^q} \right\} = s^q L\{f(t)\} - \sum_{k=0}^{m-1} s^{q-1-k} f^{(k)}(0), \quad m - 1 < q \leq m.$$

Contrary to the Riemann–Liouville fractional derivative, only integer order derivatives of function f appear in the Laplace transform of the Caputo fractional derivative. For zero initial conditions, (5) reduces to

$$(6) \quad L \left\{ \frac{d^q f(t)}{dt^q} \right\} = s^q L\{f(t)\}.$$

A fractional order linear time invariant system can be represented in the following state space form:

$$(7) \quad \begin{cases} \frac{d^q x}{dt^q} = Ax + Bu, \\ y = Cx, \end{cases}$$

where $x \in R^n$, $u \in R^m$, and $y \in R^p$ are states, inputs, and outputs vectors of the system and $A \in R^{n \times n}$, $B \in R^{n \times m}$, $C \in R^{p \times n}$, and q is the fractional commensurate order, respectively. Now, we state two stability theorems from the fractional calculus.

THEOREM 1 (see [35]). *The following autonomous system*

$$(8) \quad \frac{d^q x}{dt^q} = Ax, \quad x(0) = x_0,$$

where $0 < q \leq 1$, $x \in R^n$, and A is an $n \times n$ matrix is asymptotically stable if and only if $|\arg(\lambda)| > q\pi/2$ is satisfied for all eigenvalues (λ) of matrix A . Also, this system is stable if and only if $|\arg(\lambda)| \geq q\pi/2$ is satisfied for all eigenvalues (λ) of matrix A with those critical eigenvalues satisfying $|\arg(\lambda)| = q\pi/2$ have geometric multiplicity equal to algebraic multiplicity.

Theorem 1 can be proved by finding the solution of system (8) based on the eigenfunction of the smooth derivation operator for an eigenvalue and checking the

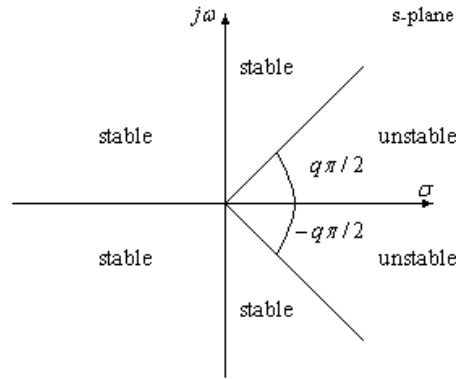


FIG. 1. Stability region of fractional order linear time invariant system with order $0 < q \leq 1$.

asymptotic behavior of this eigenfunction [35, 36]. The stable and unstable regions for $0 < q \leq 1$ is shown in Figure 1. In this paper, we will denote the stable and unstable regions for $0 < q \leq 1$ by C^{q-} and C^{q+} , respectively. Now, consider the following commensurate fractional order system:

$$(9) \quad \frac{d^q x}{dt^q} = f(x),$$

where $0 < q \leq 1$ and $x \in R^n$. The equilibrium points of system (9) are calculated by solving the following equation:

$$(10) \quad f(x) = 0.$$

THEOREM 2 (see [37]). *The equilibrium points of system (9) are locally asymptotically stable if all eigenvalues (λ) of the Jacobian matrix $J = \partial f / \partial x$ evaluated at the equilibrium points satisfy*

$$(11) \quad |\arg(\lambda)| > q\pi/2.$$

3. Analysis of frequency-based approximated model. There are many different methods to find frequency domain approximation of fractional operators (for example, see [38, 39, 40, 41, 42, 43]). In most of these methods, first a frequency range $[\omega_L, \omega_H]$ is chosen and then an integer order transfer function is determined to approximate the fractional operator in the selected frequency range. Suppose that the fractional integral operator $1/s^q$ is approximated by transfer function $G(s)$ in the given frequency range $[\omega_L, \omega_H]$:

$$(12) \quad \frac{1}{s^q} \stackrel{[\omega_L, \omega_H]}{\approx} G(s).$$

The main criticism that can be raised about using approximation in (12) in simulation of the fractional order system (9) is eliminating the inherent attribute of a fractional order system, i.e., omitting its long memory characteristics. In fact, the fractional integral operator $1/s^q$ is a nonlocal operator, whereas its approximation ($G(s)$) does not have this property. Hence, replacing $1/s^q$ by transfer function $G(s)$ causes the role of history to be ignored in simulation of the fractional order system (9).

Consequently, none of the simulations done via using frequency domain approximation of the fractional operators can preserve the long memory attribute of a fractional order system. Moreover, using approximation in (12) in simulation of the fractional order system (9) may cause some inaccuracies in the number and the location of fixed points. Fixed points of the original system are achieved by solving (10), whereas the fixed points of its frequency-based approximated model are solutions of the following equation:

$$(13) \quad f(x) = gx,$$

where g is the inverse of the steady state gain of the approximating filter $G(s)$ [34]. If the steady state gain of the approximating filter is infinite, fixed points of the approximated system and the fixed points of the original system are the same. Otherwise, the number and the location of fixed points of the original system and the approximated one may be different. In this section, our aim is to analyze stability of the approximated integer order system and compare it with that of the original fractional order system. First, the fractional order linear systems are considered and then the results of linear case are extended to a field of fractional order nonlinear systems.

3.1. Linear case. Suppose that the original system is a fractional order linear time invariant system described by

$$(14) \quad \frac{d^q x}{dt^q} = Ax,$$

where $x \in R^n$, $0 < q < 1$, and $A \in R^{n \times n}$. Also, without loss of generality, let the approximating filter $G(s)$ be strictly proper as described below:

$$(15) \quad G(s) = \frac{b_{m-1}s^{m-1} + \dots + b_1s + b_0}{s^m + a_{m-1}s^{m-1} + \dots + a_1s + a_0}.$$

The approximated model using the filter (15) is

$$(16) \quad \frac{d^m}{dt^m}x + (a_{m-1}I_n - b_{m-1}A)\frac{d^{m-1}}{dt^{m-1}}x + \dots + (a_1I_n - b_1A)\frac{d}{dt}x + (a_0I_n - b_0A)x = 0.$$

According to (13), the original system (14) and the approximated system (16) have the same fixed points if the inverse of the steady state gain of the approximating filter $G(s)$, i.e., a_0/b_0 , is not an eigenvalue of the matrix A . The high order descriptor system (16) can be realized by a first order state space model as

$$(17) \quad \dot{\tilde{x}} = M\tilde{x},$$

where

$$(18) \quad M = \begin{bmatrix} 0 & I_n & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & I_n \\ b_0A - a_0I_n & b_1A - a_1I_n & \dots & b_{m-1}A - a_{m-1}I_n \end{bmatrix}.$$

M is an $mn \times mn$ matrix and $\tilde{x} \in R^{mn}$ [34, 44]. The original system is asymptotically stable if and only if all eigenvalues of the matrix A are settled in stable region C^q (Figure 1), whereas the approximated system is asymptotically stable if and only if

$$(19) \quad \text{Re}(\text{eig}(M)) < 0.$$

In the following theorem, we state a property for eigenvalues of the matrix M .

THEOREM 3. *If the matrix A is diagonalizable in the complex field C , then eigenvalues of the matrix M defined by (18) depend only on the coefficients $a_0, \dots, a_{m-1}, b_0, \dots, b_{m-1}$ and eigenvalues of A .*

Proof. Since the matrix A is diagonalizable in the complex field C , there is a complex matrix Q such that $Q^{-1}AQ$ is a diagonal matrix with eigenvalues of A in its diagonal. Let us define

$$(20) \quad P = \begin{bmatrix} I_n & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & I_n & 0 \\ 0 & \cdots & 0 & Q \end{bmatrix},$$

where P is an $mn \times mn$ matrix. We have

$$(21) \quad P^{-1} = \begin{bmatrix} I_n & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & I_n & 0 \\ 0 & \cdots & 0 & Q^{-1} \end{bmatrix}.$$

Therefore,

$$(22) \quad P^{-1}MP = \begin{bmatrix} 0 & I_n & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & I_n \\ b_0(Q^{-1}AQ) - a_0I_n & b_1(Q^{-1}AQ) - a_1I_n & \cdots & b_{m-1}(Q^{-1}AQ) - a_{m-1}I_n \end{bmatrix}.$$

Since $Q^{-1}AQ$ depends only on the eigenvalues of A , eigenvalues of $P^{-1}MP$ depend only on the coefficients $a_0, \dots, a_{m-1}, b_0, \dots, b_{m-1}$ and eigenvalues of A . Also, M and $P^{-1}MP$ are similar matrices, and we know that similar matrices have the same set of eigenvalues. This completes the proof. \square

In the following theorem, we provide a more general property of eigenvalues of M . By this theorem, eigenvalues of M can be determined by solving an m -degree equation.

THEOREM 4. *μ is an eigenvalue of M if and only if there exists an eigenvalue of A , like λ , such that*

$$(23) \quad \mu^m - (b_{m-1}\lambda - a_{m-1})\mu^{m-1} - \cdots - (b_1\lambda - a_1)\mu - (b_0\lambda - a_0) = 0.$$

Proof. First, we prove the sufficiency of the condition. Let λ be an eigenvalue of A with corresponding eigenvector v , i.e.,

$$(24) \quad Av = \lambda v.$$

Suppose that $\mu = \mu_0$ is a solution of (23), i.e.,

$$(25) \quad \mu_0^m - (b_{m-1}\lambda - a_{m-1})\mu_0^{m-1} - \cdots - (b_1\lambda - a_1)\mu_0 - (b_0\lambda - a_0) = 0.$$

Define

$$(26) \quad w = \begin{bmatrix} v \\ \mu_0 v \\ \vdots \\ \mu_0^{m-1} v \end{bmatrix}.$$

We claim that $Mw = \mu_0 w$. First, note that

$$(27) \quad \begin{aligned} & [b_0 A - a_0 I_n \quad b_1 A - a_1 I_n \quad \cdots \quad b_{m-1} A - a_{m-1} I_n] w \\ &= (b_0 A - a_0 I_n) v + (b_1 A - a_1 I_n) \mu_0 v + \cdots + (b_{m-1} A - a_{m-1} I_n) \mu_0^{m-1} v \\ &= (b_0 + b_1 \mu_0 + \cdots + b_{m-1} \mu_0^{m-1}) A v - (a_0 + a_1 \mu_0 + \cdots + a_{m-1} \mu_0^{m-1}) v. \end{aligned}$$

Using (24) and (25), results in the above expression equal to

$$(28) \quad \begin{aligned} & (b_0 + b_1 \mu_0 + \cdots + b_{m-1} \mu_0^{m-1}) \lambda v - (a_0 + a_1 \mu_0 + \cdots + a_{m-1} \mu_0^{m-1}) v \\ &= [(b_0 \lambda - a_0) + (b_1 \lambda - a_1) \mu_0 + \cdots + (b_{m-1} \lambda - a_{m-1}) \mu_0^{m-1}] v \\ &= \mu_0^m v. \end{aligned}$$

Now, it is straightforward to verify that

$$(29) \quad \begin{aligned} Mw &= \begin{bmatrix} 0 & I_n & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & I_n \\ b_0 A - a_0 I_n & b_1 A - a_1 I_n & \cdots & b_{m-1} A - a_{m-1} I_n \end{bmatrix} \begin{bmatrix} \nu \\ \mu_0 \nu \\ \vdots \\ \mu_0^{m-1} \nu \end{bmatrix} \\ &= \begin{bmatrix} \mu_0 \nu \\ \mu_0^2 \nu \\ \vdots \\ \mu_0^m \nu \end{bmatrix} = \mu_0 w. \end{aligned}$$

Thus, the claim is proved. Therefore, μ_0 is an eigenvalue of M . Now, we prove the necessity of the condition given in Theorem 4. Let μ be an eigenvalue of M and w be the eigenvector of M corresponding to μ , i.e.,

$$(30) \quad Mw = \mu w.$$

We assume that

$$(31) \quad w = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix},$$

where v_i 's ($i = 1, 2, \dots, m$) are $n \times 1$ vectors. Hence,

$$(32) \quad Mw = \begin{bmatrix} 0 & I_n & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & I_n \\ b_0 A - a_0 I_n & b_1 A - a_1 I_n & \cdots & b_{m-1} A - a_{m-1} I_n \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{m-1} \\ v_m \end{bmatrix} = \begin{bmatrix} v_2 \\ v_3 \\ \vdots \\ v_m \\ \nu \end{bmatrix},$$

where ν is an $n \times 1$ vector. According to (30) and (32), one finds $v_{i+1} = \mu v_i$ for $i = 1, 2, \dots, m-1$. Thus, $v_i = \mu^{i-1} v_1$ for $i = 1, 2, \dots, m$. Therefore,

$$(33) \quad w = \begin{bmatrix} v_1 \\ \mu v_1 \\ \vdots \\ \mu^{m-1} v_1 \end{bmatrix}.$$

From (32), we know that $\nu = \mu^m v_1$. On the other hand,

$$(34) \quad \begin{aligned} \nu &= [b_0 A - a_0 I_n \quad b_1 A - a_1 I_n \quad \cdots \quad b_{m-1} A - a_{m-1} I_n] w \\ &= (b_0 + b_1 \mu + \cdots + b_{m-1} \mu^{m-1}) Av_1 - (a_0 + a_1 \mu + \cdots + a_{m-1} \mu^{m-1}) v_1. \end{aligned}$$

Therefore,

$$(35) \quad (b_0 + b_1 \mu + \cdots + b_{m-1} \mu^{m-1}) Av_1 - (a_0 + a_1 \mu + \cdots + a_{m-1} \mu^{m-1}) v_1 = \mu^m v_1.$$

So,

$$(36) \quad Av_1 = \frac{\mu^m + a_{m-1} \mu^{m-1} + \cdots + a_1 \mu + a_0}{b_{m-1} \mu^{m-1} + \cdots + b_1 \mu + b_0} v_1.$$

Thus, the expression $(\mu^m + a_{m-1} \mu^{m-1} + \cdots + a_1 \mu + a_0)/(b_{m-1} \mu^{m-1} + \cdots + b_1 \mu + b_0)$ is an eigenvalue of A . In other words, the following relation holds for any eigenvalue λ_0 of A :

$$(37) \quad \lambda_0 = \frac{\mu^m + a_{m-1} \mu^{m-1} + \cdots + a_1 \mu + a_0}{b_{m-1} \mu^{m-1} + \cdots + b_1 \mu + b_0}.$$

Therefore,

$$(38) \quad \mu^m - (b_{m-1} \lambda_0 - a_{m-1}) \mu^{m-1} - \cdots - (b_1 \lambda_0 - a_1) \mu - (b_0 \lambda_0 - a_0) = 0,$$

and this closes the proof of Theorem 4. \square

According to (23), each eigenvalue of the matrix A converts to m eigenvalues in the approximated model, where m is the order of the approximating filter. Now, we want to find stable and unstable regions for the approximated model concerning the eigenvalues of A . The stable region for the approximated model is the region in which all m converted eigenvalues corresponding to each eigenvalue of A are stable. The rest of the complex plane forms an unstable region for the approximated model. Let $C_{G(s)}^{q-}$ and $C_{G(s)}^{q+}$, respectively, denote stable and unstable regions for the approximated model constructed by the approximating filter $G(s)$ as an approximation for operator $1/s^q$. Also, the phrase ‘‘stability boundary’’ indicates the boundary between $C_{G(s)}^{q-}$ and $C_{G(s)}^{q+}$. The ideal stability boundary is defined by $|\arg(z)| = q\pi/2$ as shown in Figure 1.

By Theorem 4, we know that if μ is an eigenvalue of M , then $(\mu^m + a_{m-1} \mu^{m-1} + \cdots + a_1 \mu + a_0)/(b_{m-1} \mu^{m-1} + \cdots + b_1 \mu + b_0)$ is an eigenvalue of A . Suppose that

$$(39) \quad H(s) = (G(s))^{-1} = \frac{s^m + a_{m-1} s^{m-1} + \cdots + a_1 s + a_0}{b_{m-1} s^{m-1} + \cdots + b_1 s + b_0}.$$

Now, we define the region D as follows:

$$(40) \quad D = \{H(s) | \operatorname{Re}(s) \geq 0\}.$$

Thus, the approximated system is stable if and only if A has no eigenvalue in D . Hence, the stability boundary of the approximated system is the following curve in the complex plane:

$$(41) \quad \gamma = \{H(j\omega) | -\infty < \omega < \infty\}.$$

Since γ is the stability boundary of the approximated system, it can precisely determine the stable region for the approximated system. In fact, the stability boundary is mapping of the vertical imaginary axis by relation $H(s) = 1/G(s)$. It is clear that the curve γ is symmetrical with respect to the horizontal real axis of the complex plane. Theorem 5 investigates the behavior of the curve γ when $\omega \rightarrow 0$.

THEOREM 5. *When ω is close enough to zero, the curve γ in the complex plane can be approximately determined by a horizontal parabola with the following equation:*

$$(42) \quad \operatorname{Re}(z) = A + B(\operatorname{Im}(z))^2,$$

where

$$(43) \quad A = a_0/b_0$$

and

$$(44) \quad B = \frac{(a_1b_1 + a_0b_2 - a_2b_0 - a_0b_1^2/b_0)}{(a_1 - a_0b_1/b_0)^2}.$$

Proof. We have

$$(45) \quad H(j\omega) = \frac{[a_0 - a_2\omega^2 + a_4\omega^4 - \dots] + j[a_1\omega - a_3\omega^3 + a_5\omega^5 - \dots]}{[b_0 - b_2\omega^2 + b_4\omega^4 - \dots] + j[b_1\omega - b_3\omega^3 + b_5\omega^5 - \dots]}.$$

Therefore,

$$(46) \quad \begin{aligned} &\operatorname{Re}(H(j\omega)) \\ &= \frac{[a_0 - a_2\omega^2 + \dots][b_0 - b_2\omega^2 + \dots] + [a_1\omega - a_3\omega^3 + \dots][b_1\omega - b_3\omega^3 + \dots]}{[b_0 - b_2\omega^2 + \dots]^2 + [b_1\omega - b_3\omega^3 + \dots]^2} \end{aligned}$$

and

$$(47) \quad \begin{aligned} &\operatorname{Im}(H(j\omega)) \\ &= \frac{[a_1\omega - a_3\omega^3 + \dots][b_0 - b_2\omega^2 + \dots] - [a_0 - a_2\omega^2 + \dots][b_1\omega - b_3\omega^3 + \dots]}{[b_0 - b_2\omega^2 + \dots]^2 + [b_1\omega - b_3\omega^3 + \dots]^2}. \end{aligned}$$

If $\omega \rightarrow 0$,

$$(48) \quad \begin{aligned} \operatorname{Re}(H(j\omega)) &\approx \frac{a_0b_0 + (a_1b_1 - a_0b_2 - a_2b_0)\omega^2}{b_0^2 + (b_1^2 - 2b_0b_2)\omega^2} \\ &\approx \frac{a_0}{b_0} + \frac{(a_1b_1 + a_0b_2 - a_2b_0 - a_0b_1^2/b_0)\omega^2}{b_0^2 + (b_1^2 - 2b_0b_2)\omega^2} \\ &\approx \frac{a_0}{b_0} + \frac{(a_1b_1 + a_0b_2 - a_2b_0 - a_0b_1^2/b_0)\omega^2}{b_0^2} \end{aligned}$$

and

$$(49) \quad \operatorname{Im}(H(j\omega)) \approx \frac{(a_1b_0 - a_0b_1)\omega}{b_0^2}.$$

Use of (48) and (49) results in (42). \square

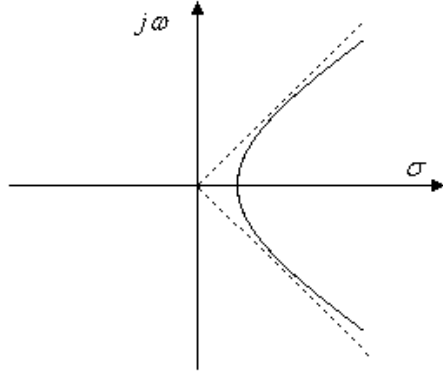


FIG. 2. The stability boundary for low frequencies. The dashed lines indicate the actual stability boundary.

Figure 2 schematically shows the behavior of curve γ for low frequencies. According to this figure, it is clear that there is an area between sector $|\arg(z)| < q\pi/2$ and the parabola-like shape of boundary, which can be problematic in the sense of stability. In other words, if A has an unstable eigenvalue settled in this area, it converts to m stable eigenvalues in the approximated model. Thus, the original system may be unstable, whereas its approximation is stable. Now, we state another theorem to clear the behavior of curve γ when $\omega \rightarrow \infty$.

THEOREM 6. *If $b_{m-1} \neq 0$, when $\omega \rightarrow \infty$, the curve γ in the complex plane tends to the vertical line*

$$(50) \quad \operatorname{Re}(z) = \frac{\sum_{i=1}^{m-1} z_i - \sum_{i=1}^m p_i}{b_{m-1}},$$

where z_i ($i = 1, 2, \dots, m-1$) and p_i ($i = 1, 2, \dots, m$), respectively, denote zeros and poles of the approximating filter (15).

Proof. Since $G(s) = 1/H(s)$ is proper, $\lim_{\omega \rightarrow \infty} |H(j\omega)| = \infty$. Hence, it suffices to show that

$$(51) \quad \lim_{\omega \rightarrow \infty} \operatorname{Re}(H(j\omega)) = \frac{\sum_{i=1}^{m-1} z_i - \sum_{i=1}^m p_i}{b_{m-1}}.$$

If m is even ($m = 2k$, $k \in \mathbb{N}$),

$$(52) \quad H(j\omega) = \frac{[a_0 - a_2\omega^2 + \dots + (-1)^k a_{m-1}\omega^{m-1}] + j[a_1\omega - a_3\omega^3 + \dots + (-1)^{k-1} a_{m-1}\omega^{m-1}]}{[b_0 - b_2\omega^2 + \dots + (-1)^{k-1} b_{m-2}\omega^{m-2}] + j[b_1\omega - b_3\omega^3 + \dots + (-1)^{k-1} b_{m-1}\omega^{m-1}]},$$

and if m is odd ($m = 2k + 1$, $k \in \mathbb{N}$),

$$(53) \quad H(j\omega) = \frac{[a_0 - a_2\omega^2 + \dots + (-1)^k a_{m-1}\omega^{m-1}] + j[a_1\omega - a_3\omega^3 + \dots + (-1)^k \omega^m]}{[b_0 - b_2\omega^2 + \dots + (-1)^k b_{m-1}\omega^{m-1}] + j[b_1\omega - b_3\omega^3 + \dots + (-1)^{k-1} b_{m-2}\omega^{m-2}]}.$$

From (52) and (53),

$$(54) \quad \operatorname{Re}(H(j\omega)) = \frac{(a_{m-1}b_{m-1} - b_{m-2})\omega^{2m-2} + P_1(\omega)}{b_{m-1}^2\omega^{2m-2} + P_2(\omega)},$$

where $P_1(\omega)$ and $P_2(\omega)$ are polynomials with degrees less than $2m - 2$. So,

$$(55) \quad \lim_{\omega \rightarrow \infty} \operatorname{Re}(H(j\omega)) = \frac{a_{m-1}b_{m-1} - b_{m-2}}{b_{m-1}^2} = \frac{a_{m-1} - b_{m-2}/b_{m-1}}{b_{m-1}}.$$

We know that $a_{m-1} = -\sum_{i=1}^m p_i$ and $b_{m-2}/b_{m-1} = -\sum_{i=1}^{m-1} z_i$. This completes the proof. \square

The behavior of curve γ in low and high frequencies has been discussed in Theorems 5 and 6. Now, we want to analyze the behavior of curve γ in the intermediate frequencies. Suppose

$$(56) \quad \tilde{H}(s) = s^q,$$

where s belongs to the principal Riemann surface [45] and define curve $\tilde{\gamma}$ in the complex plane as follows:

$$(57) \quad \tilde{\gamma} = \{\tilde{H}(j\omega) \mid -\infty < \omega < \infty\}.$$

It is clear that the curve $\tilde{\gamma}$ is the boundary between C^{q-} and C^{q+} regions (ideal stability boundary). This means if we could use $1/s^q$ instead of $G(s)$ in numerical simulations, there is no problem in the sense of stability. We know $G(s)$ is an approximation for the fractional operator $1/s^q$ in the frequency range $[\omega_L, \omega_H]$. Suppose

$$(58) \quad \begin{aligned} \exists \varepsilon_0 > 0 : |G(j\omega)| &= \omega^{-q} + \varepsilon, & |\varepsilon| < \varepsilon_0, \\ \exists \delta_0 > 0 : \arg(G(j\omega)) &= -q\frac{\pi}{2} + \delta, & |\delta| < \delta_0 \end{aligned}$$

for $\omega_L < \omega < \omega_H$. From (58),

$$(59) \quad \frac{\operatorname{Im}(H(j\omega))}{\operatorname{Re}(H(j\omega))} = \tan\left(q\frac{\pi}{2} - \delta\right), \quad \omega_L < \omega < \omega_H.$$

If $\delta_0 < q\pi/2$, (59) results in

$$(60) \quad \tan\left(q\frac{\pi}{2} - \delta_0\right) < \frac{\operatorname{Im}(H(j\omega))}{\operatorname{Re}(H(j\omega))} < \tan\left(q\frac{\pi}{2} + \delta_0\right), \quad \omega_L < \omega < \omega_H.$$

The inequality (60) guarantees that the curve γ settle in the sector $|\arg(z) - q\pi/2| < \delta_0$. Therefore, the allowable phase error of approximating filter has a very effective role in the accuracy of the stability boundary, whereas in most approximation methods, the allowable magnitude error is considered as one of the determinable parameters [38, 39, 40]. Figure 3 schematically shows the behavior of curve γ for intermediate and high frequencies. According to this figure, it is obvious that if the matrix A has stable eigenvalues settled in the right side of vertical asymptote, the approximated system has at least one unstable eigenvalue. Thus, the original system may be stable, whereas its approximation is unstable.

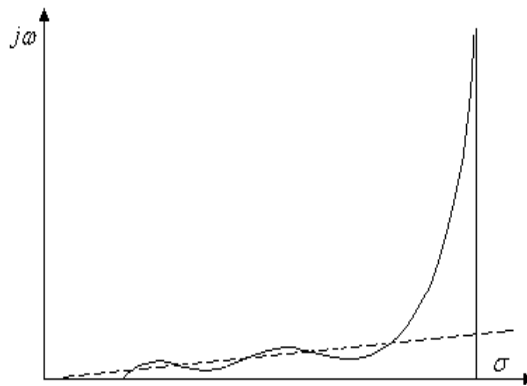


FIG. 3. The stability boundary for intermediate and high frequencies. The dashed line indicates the actual stability boundary.

3.2. Nonlinear case. Suppose that the original system is a fractional order nonlinear time invariant system described by (9). According to Theorem 2, the fixed point of this system x_e (a solution of (10)) is asymptotically stable if

$$(61) \quad |\arg(\text{eig}(J))| > q\pi/2,$$

where

$$(62) \quad J = \frac{\partial f}{\partial x} \Big|_{x=x_e}.$$

Using the approximating filter (15), the approximated system is described by the following relation:

$$(63) \quad \frac{d^m}{dt^m}x + \dots + a_1 \frac{d}{dt}x + a_0x = b_{m-1} \frac{d^{m-1}}{dt^{m-1}}f(x) + \dots + b_1 \frac{d}{dt}f(x) + b_0f(x).$$

In [34], by using a generalization of Faa di Bruno’s formula [46], it has been shown that the fixed point of system (63) x_e^* (a solution of (13)) is asymptotically stable if

$$(64) \quad \text{Re}(\text{eig}(M^*)) < 0,$$

where

$$(65) \quad M^* = \begin{bmatrix} 0 & I_n & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & I_n \\ b_0J^* - a_0I_n & b_1J^* - a_1I_n & \dots & b_{m-1}J^* - a_{m-1}I_n \end{bmatrix}$$

and

$$(66) \quad J^* = \frac{\partial f}{\partial x} \Big|_{x=x_e^*}.$$

Results obtained in the previous part can be applied here for stability analysis of the equilibrium x_e^* . As in the previous case, eigenvalues of matrix M^* depend on eigenvalues of matrix J^* as well as the coefficients $a_0, \dots, a_{m-1}, b_0, \dots, b_{m-1}$ (Theorem 3). It should be noted that since the fixed point of original system x_e and the fixed point of the approximated system x_e^* are not equal, J^* is not necessarily equal to the original matrix J . However, if the steady state gain of the approximating filter (15) is

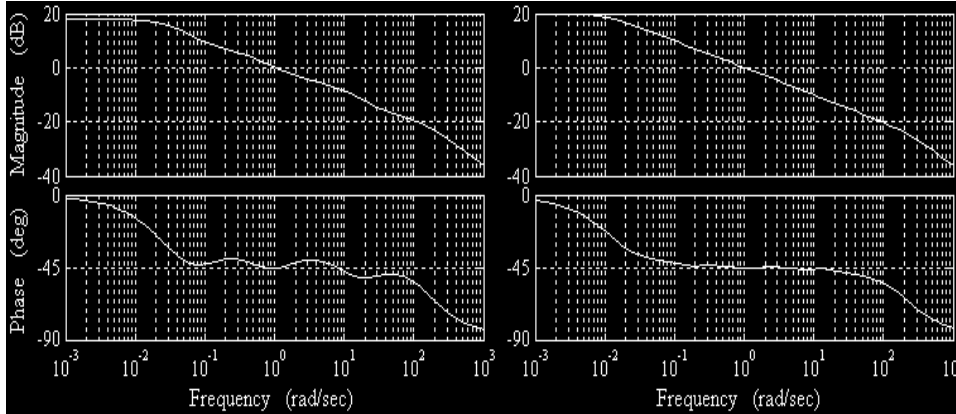


FIG. 4. Bode diagrams for transfer functions $G_1(s)$ (left) and $G_2(s)$ (right).

large enough, the fixed points of approximated system (63) are very close to the fixed points of the original system (9) [34].

4. Some numerical illustrations. In this section, results of the previous section are examined via numerical examples. Also, we analyze some wrong results reported in the literature and show that the negligence of authors in the limitations of the frequency domain methods has caused these mistakes. Table I of [26] has given approximations for $1/s^q$, with $q = 0.1 - 0.9$ in steps of 0.1. These approximations, obtained by trial and error, have maximum discrepancy of 2 dB from $\omega = 10^{-2}$ to 10^2 rad./sec. In this table, the integer order approximation of operator $1/s^{0.5}$ is given as

$$(67) \quad G_1(s) = \frac{15.97s^4 + 593.2s^3 + 1080s^2 + 135.4s + 1}{s^5 + 134.3s^4 + 1072s^3 + 543.4s^2 + 20.10s + 0.1259}.$$

Another integer order approximation for operator $1/s^{0.5}$ with maximum discrepancy of 2 dB from $\omega = 10^{-2}$ to 10^2 rad./sec. is given in Table 1 of [28]. This approximation, found by Charef’s method [39], is given as follows:

$$(68) \quad G_2(s) = \frac{15.8489(s + 0.03981)(s + 0.2512)(s + 1.585)(s + 10)(s + 63.1)}{(s + 0.01585)(s + 0.1)(s + 0.631)(s + 3.981)(s + 25.12)(s + 158.5)}.$$

Figure 4 shows the Bode diagrams for transfer functions $G_1(s)$ and $G_2(s)$. According to (41), curves $\gamma_1 = \{1/G_1(j\omega) | -\infty < \omega < \infty\}$ and $\gamma_2 = \{1/G_2(j\omega) | -\infty < \omega < \infty\}$ determine stability boundaries for the approximated systems constructed based on filters $G_1(s)$ and $G_2(s)$, respectively. The boundaries of low frequency range are illustrated in Figure 5(a). It is seen that the left side of the boundaries have a horizontal parabola-like shape which verifies Theorem 5. If the original system has an unstable eigenvalue settled between sector $|\arg(z)| < \pi/4$ and a parabola-like shape of boundaries, this eigenvalue is interpreted to stable eigenvalues in the approximated system. Consequently, results of the numerical simulation would not be reliable in this case. Figure 5(b) shows stability boundaries for intermediate frequencies. In these frequencies, the stability boundary of the approximating filter $G_2(s)$ is more consistent with the ideal boundary than the stability boundary of the approximating filter $G_1(s)$. The reason arises from accuracy of the approximation in the sense of phase. The phase of the approximating filter $G_2(s)$ is more accurate than the phase of the approximating

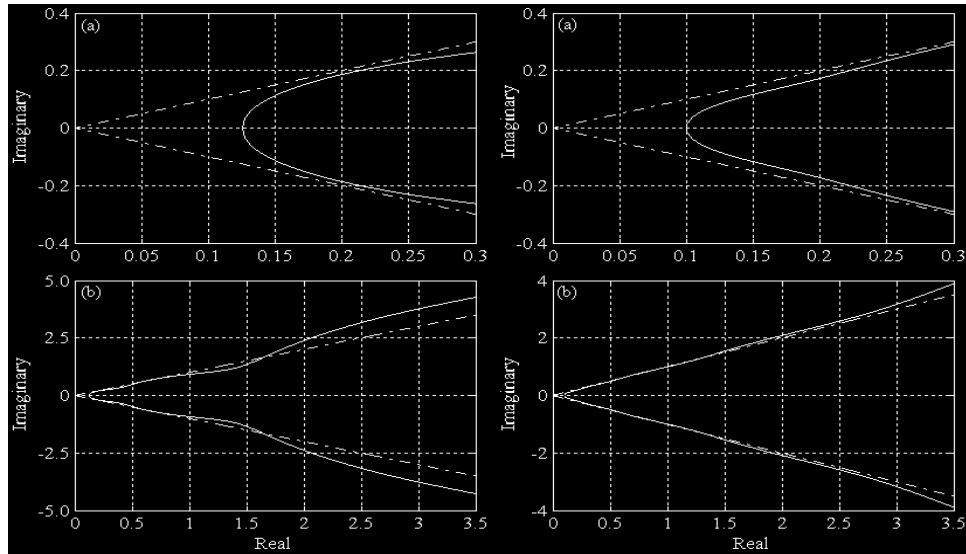


FIG. 5. Stability boundaries resulted from using approximating filters $G_1(s)$ (left) and $G_2(s)$ (right). (a): low frequencies, (b): intermediate frequencies. The dashed lines indicate the actual stability boundary.

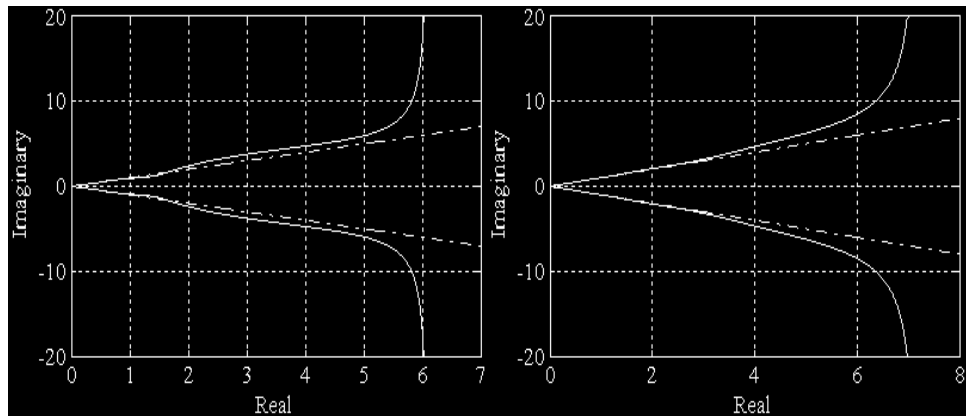


FIG. 6. Stability boundaries resulted from using approximating filters $G_1(s)$ (left) and $G_2(s)$ (right). The dashed lines indicate the actual stability boundary.

filter $G_1(s)$ (Figure 4). Therefore, according to (60), this approximation preserves the stability boundary more accurately. Figure 6 shows the stability boundaries for a wide range of frequencies. The stability boundaries tend to vertical asymptotes obtained by (50). It is clear that if the original fractional order system has a stable eigenvalue in the right side of asymptotes, the approximated system has at least one unstable eigenvalue. Therefore, results of the numerical simulation in this case are not reliable. This fallacy has occurred in some papers. In the following, we debate about one of these mistakes. Existence of chaotic behavior has been reported for a fractional order Lu system based on numerical simulations in [29]. The fractional order Lu system is

described by

$$(69) \quad \begin{cases} \frac{d^q x}{dt^q} = a(y - x), \\ \frac{d^q y}{dt^q} = -xz + cy, \\ \frac{d^q z}{dt^q} = xy - bz. \end{cases}$$

For $bc > 0$, the Lu system has three equilibriums at

$$(70) \quad O = (0, 0, 0), \quad C^\pm = (\pm\sqrt{bc}, \pm\sqrt{bc}, c).$$

The Jacobian matrix of the system (69), evaluated at (x^*, y^*, z^*) , is

$$(71) \quad \begin{bmatrix} -a & a & 0 \\ -z^* & c & -x^* \\ y^* & x^* & -b \end{bmatrix}.$$

For instance, a chaotic attractor of the fractional order Lu system has been shown for parameter set $(a, b, c) = (26, 3, 28)$ and order $q = 0.5$ in [29]. For this parameter set, the fixed points and their corresponding eigenvalues are

$$\begin{aligned} O = (0, 0, 0) : & \quad \lambda_1 = -26, \quad \lambda_2 = 28, \quad \lambda_3 = -3, \\ O_{2,3} = (\pm 9.1652, \pm 9.1652, 28) : & \quad \lambda_1 = -15.0661, \quad \lambda_{2,3} = 7.0331 \pm j15.5067. \end{aligned}$$

The maximum fractional order q , for which the fixed points C^\pm remain unstable and consequently, the fractional order Lu system is susceptible to be chaotic, is about 0.73 [33]. Therefore, for $q = 0.5$ and parameter set $(a, b, c) = (26, 3, 28)$, the fixed points C^\pm are locally stable. This means the fractional order Lu system cannot be chaotic for parameter set $(a, b, c) = (26, 3, 28)$ and order $q = 0.5$. But why has the chaotic behavior been demonstrated in the numerical simulations of [29]? Numerical simulations in [29] have been performed based on the frequency domain approximations. The reference of this paper for frequency domain approximations is Table I of [26]. Similar to [29], we also used the approximating filter (67) to simulate the system (69) with the parameter set $(a, b, c) = (26, 3, 28)$ and order $q = 0.5$. The chaotic behavior was demonstrated in this case as well (Figure 7). By using approximating filter (67), the approximated system has three fixed points. From (13) and (71), these fixed points and their corresponding eigenvalues are

$$\begin{aligned} O = (0, 0, 0) : & \quad \lambda_1 = -26, \quad \lambda_2 = 28, \quad \lambda_3 = -3, \\ \tilde{O}_{2,3} = (\pm 9.3344, \pm 9.3796, 28.0091) : & \quad \lambda_1 = -15.2335, \quad \lambda_{2,3} = 7.1167 \pm j15.7336. \end{aligned}$$

All eigenvalues of fixed points $O_{2,3}$ are settled in region $C^{0.5-}$, but eigenvalues $\lambda_{2,3}$ of fixed points $\tilde{O}_{2,3}$ are not settled in $C_{G_1(s)}^-$ (Figure 8). Therefore, the fixed points $\tilde{O}_{2,3}$ of the approximated system are not stable and consequently, this system is capable to generate chaos. This inconsistency can be solved if one uses another approximating filter whose stability boundary is more coincident with the original stability boundary. Figure 8 shows a stability boundary of another approximating filter $G_3(s)$ for operator $1/s^{0.5}$ constructed by Charef's method [39] and has maximum discrepancy of 2 dB from $\omega = 10^{-3}$ to 10^3 rad./sec.

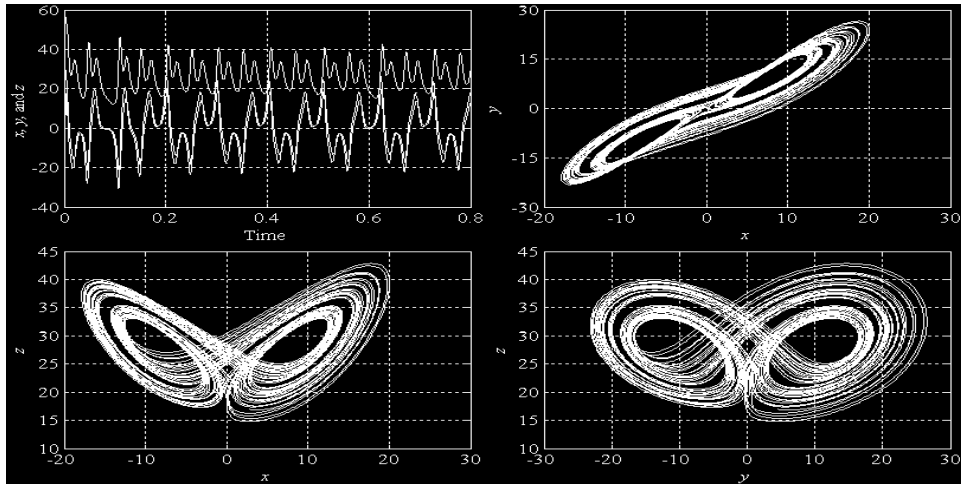


FIG. 7. Simulation results for Lu system with parameter set $(a, b, c) = (26, 3, 28)$ and order $q = 0.5$ (Simulation has been done by using approximating filter $G_1(s)$).

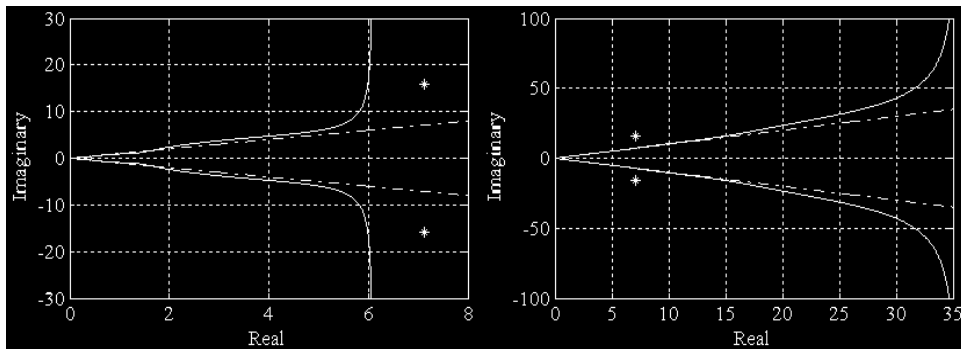


FIG. 8. Stability boundaries for $G_1(s)$ (left) and $G_3(s)$ (right). Locations of critical eigenvalues are highlighted by *. The dashed lines indicate the actual stability boundary.

(72)

$$G_3(s) = \frac{80.18s^8 + 1.51 \times 10^5 s^7 + 3.891 \times 10^7 s^6 + 1.555 \times 10^9 s^5 + 9.818 \times 10^9 s^4 + 9.813 \times 10^9 s^3 + 1.549 \times 10^9 s^2 + 3.793 \times 10^7 s + 1.271 \times 10^5}{s^9 + 4731s^8 + 3.062 \times 10^6 s^7 + 3.074 \times 10^8 s^6 + 4.875 \times 10^9 s^5 + 1.225 \times 10^{10} s^4 + 4.872 \times 10^9 s^3 + 3.062 \times 10^8 s^2 + 2.985 \times 10^6 s + 3981}$$

The above approximation does not change the stability of the fixed points of the original system. Thus, using it in the simulation of system (69) with the parameter set $(a, b, c) = (26, 3, 28)$ and order $q = 0.5$ does not lead to wrong consequences. The similar wrong results can be noticed in the numerical simulations of [30] and [31].

5. Conclusions. Before any conclusion can be made about the results of a performed numerical simulation, the reliability of the numerical method used in that simulation must be considered and checked. Hence, the reliability verification of numerical methods is of great importance. One of the basic points which should be considered in reliability verification of a numerical simulation is stability preserva-

tion. In fact, the original system and its simulated model must be similar in the sense of stability. In this paper, the stability preservation problem was investigated in one of the popular methods used in the simulation of fractional order systems. The method discussed in this paper is based on using integer order approximation of fractional operators. We found the “stability boundary” for this method and showed that the found boundary is not the same as the original boundary, and there are critical regions in which the stability can not preserve. Two types of inaccuracies can occur:

- The approximated model, obtained by frequency domain methods, is stable, whereas the original system is not actually stable.
- The original system is stable, but its frequency domain-based approximation is not stable.

Also, we showed that the second type of inaccuracies can lead to wrong results in detecting chaotic behaviors which are reported in some recent papers. Moreover, we found that the accuracy in the phase of integer order approximations has an undeniable role in the stability preservation of the frequency domain-based numerical methods. It should be noted that the methods available for finding integer order approximation of fractional order operators do not directly try to rectify the phase of the approximation.

REFERENCES

- [1] R. L. BAGLEY AND R. A. CALICO, *Fractional order state equations for the control of viscoelastically damped structures*, J. Guidance Control Dyn., 14 (1991), pp. 304–311.
- [2] Y. A. ROSSIKHIN AND M. V. SHITIKOVA, *Application of fractional derivatives to the analysis of damped vibrations of viscoelastic single mass system*, Acta Mech., 120 (1997), pp. 109–125.
- [3] N. ENGHETA, *On fractional calculus and fractional multipoles in electromagnetism*, IEEE Trans. Antennas and Propagation, 44 (1996), pp. 554–566.
- [4] V. E. ARKHINCHEEV, *Anomalous diffusion in inhomogeneous media: Some exact results*, Model. Meas. Control A, 26 (1993), pp. 11–29.
- [5] A. M. A. EL-SAYED, *Fractional order diffusion wave equation*, Internat. J. Theoret. Phys., 35 (1996), pp. 311–322.
- [6] G. CHEN AND G. FRIEDMAN, *An RLC interconnect model based on Fourier analysis*, IEEE Trans. Computer Aided Des. Integr. Circ. Syst., 24 (2005), pp. 170–183.
- [7] V. G. JENSON AND G. V. JEFFREYS, *Mathematical Methods in Chemical Engineering*, 2nd ed., Academic Press, New York, 1977.
- [8] K. S. COLE, *Electric conductance of biological systems*, in Proceedings of the Cold Spring Harbor Symposium on Quantitative Biology, Cold Spring Harbor, New York, 1993, pp. 107–126.
- [9] N. LASKIN, *Fractional market dynamics*, Phys. A, 287 (2000), pp. 482–492.
- [10] R. HILFER (ED.), *Applications of Fractional Calculus in Physics*, World Scientific, Singapore, 2000.
- [11] I. PODLUBNY, *Fractional Differential Equations*, Academic Press, San Diego, 1999.
- [12] A. OUSTALOUP, J. SABATIER, AND P. LANUSSE, *From fractal robustness to CRONE control*, Fract. Calc. Appl. Anal., 2 (1999), pp. 1–30.
- [13] H. F. RAYNAUD AND A. Z. INOH, *State-space representation for fractional order controllers*, Automatica, 36 (2000), pp. 1017–1021.
- [14] I. PODLUBNY, *Fractional order systems and $PI^\lambda D^\mu$ -controllers*, IEEE Trans. Automat. Control, 44 (1999), pp. 208–214.
- [15] G. MAIONE AND P. LINO, *New tuning rules for fractional PI^α controllers*, Nonlinear Dynam., 49 (2007), pp. 251–257.
- [16] M. S. TAVAZOEI AND M. HAERI, *Chaos control via a simple fractional order controller*, Phys. Lett. A, 372 (2008), pp. 798–807.
- [17] V. FELIU-BATLLE, R. RIVAS PEREZ, AND L. SANCHEZ RODRIGUEZ, *Fractional robust control of main irrigation canals with variable dynamic parameters*, Control Engr. Pract., 15 (2007), pp. 673–686.
- [18] A. OUSTALOUP, *La Dérivation Non Entière: Théorie, Synthèse et Applications*, Hermès, Paris, 1995.
- [19] K. DIETHELM, N. J. FORD, AND A. D. FREED, *A predictor-corrector approach for the numerical solution of fractional differential equations*, Nonlinear Dynam., 29 (2002), pp. 3–22.

- [20] K. DIETHELM, N. J. FORD, A. D. FREED, AND Y. LUCHKO, *Algorithms for the fractional calculus: A selection of numerical methods*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 743–773.
- [21] N. J. FORD AND A. C. SIMPSON, *The numerical solution of fractional differential equations: Speed versus accuracy*, Numer. Algorithms, 26 (2001), pp. 333–346.
- [22] L. YUAN AND O. P. AGRAWAL, *A numerical scheme for dynamic systems containing fractional derivatives*, Trans. ASME. J. Vibration Acoust., 124 (2002), pp. 321–324.
- [23] P. KUMAR AND O. P. AGRAWAL, *An approximate method for numerical solution of fractional differential equations*, Signal Process., 86 (2006), pp. 2602–2610.
- [24] K. DIETHELM AND N. J. FORD, *Multi-order fractional differential equations and their numerical solution*, Appl. Math. Comput., 154 (2004), pp. 621–640.
- [25] W. DENG, *Short memory principle and a predictor-corrector approach for fractional differential equations*, J. Comput. Appl. Math., 206 (2007), pp. 174–188.
- [26] T. T. HARTLEY, C. F. LORENZO, AND H. K. QAMMER, *Chaos in a fractional order Chua's system*, IEEE Trans. Circuits Syst. I, 42 (1995), pp. 485–490.
- [27] C. LI AND G. CHEN, *Chaos and hyperchaos in the fractional order Rössler equations*, Physica A., 341 (2004), pp. 55–61.
- [28] W. M. AHMAD AND J. C. SPOTT, *Chaos in fractional order autonomous nonlinear systems*, Chaos Solitons Fractals, 16 (2003), pp. 339–351.
- [29] J. G. LU, *Chaotic dynamics of the fractional order Lü system and its synchronization*, Phys. Lett. A, 354 (2006), pp. 305–311.
- [30] J. G. LU AND G. CHEN, *A note on the fractional order Chen system*, Chaos Solitons Fractals, 27 (2006), pp. 685–688.
- [31] J. S. H. TSAI, T. H. CHIEN, S. M. GUO, Y. P. CHANG, AND L. S. SHIEH, *State space self tuning control for stochastic chaotic fractional order systems*, IEEE Trans. Circuits Syst. I, 54 (2007), pp. 632–642.
- [32] J. L. ADAMS, T. T. HARTLEY, AND C. F. LORENZO, *Chaos in a Chua system with order less than two*, The IASTED International Workshop on Modern Nonlinear Theory, Montreal, Canada, May 30, 2007.
- [33] M. S. TAVAZOEI AND M. HAERI, *A necessary condition for double scroll attractor existence in fractional order systems*, Phys. Lett. A, 367 (2007), pp. 102–113.
- [34] M. S. TAVAZOEI AND M. HAERI, *Unreliability of frequency domain approximation in recognizing chaos in fractional order systems*, IET Signal Process., 1 (2007), pp. 171–181.
- [35] D. MATIGNON, *Stability results for fractional differential equations with applications to control processing*, Computational Engineering in Systems and Application Multiconference, IMACS, IEEE-SMC Proceedings, Lille, France, Vol. 2, 1996, pp. 963–968.
- [36] D. MATIGNON, *Stability properties for generalized fractional differential systems*, ESAIM: Proceedings, 5 (1998), pp. 145–158.
- [37] E. AHMED, A. M. A. EL-SAYED, AND H. A. A. EL-SAKA, *Equilibrium points, stability and numerical solutions of fractional order predator-prey and rabies models*, J. Math. Anal. Appl., 325 (2007), pp. 542–553.
- [38] B. M. VINAGRE, I. PODLUBNY, A. HERNANDEZ, AND V. FELIU, *Some approximations of fractional order operators used in control theory and applications*, Fract. Calc. Appl. Anal., 3 (2000), pp. 945–950.
- [39] A. CHAREF, H. H. SUN, Y. Y. TSAO, AND B. ONARAL, *Fractal system as represented by singularity function*, IEEE Trans. Automat. Control, 37 (1992), pp. 1465–1470.
- [40] A. CHAREF, *Analogue realization of fractional order integrator, differentiator and fractional $PI^{\lambda}D^{\mu}$ controller*, IEE Proc. Control Theory Appl., 153 (2006), pp. 714–720.
- [41] M. AOUN, R. MALTI, F. LEVRON, AND A. OUSTALOUP, *Numerical simulations of fractional systems: An overview of existing methods and improvements*, Nonlinear Dynam., 38 (2004), pp. 117–131.
- [42] G. E. CARLSON AND C. A. HALIYAK, *Approximation of fractional capacitors $(1/s)^{1/n}$ by a regular Newton process*, IRE Trans. Circuit Theory, 11 (1964), pp. 210–213.
- [43] A. OUSTALOUP, F. LEVRON, B. MATHIEU, AND F. M. NANOT, *Frequency band complex non integer differentiator: Characterization and synthesis*, IEEE Trans. Circuits Syst. I. Fundam. Theory Appl., 47 (2000), pp. 25–39.
- [44] G. R. DUAN, *Parametric control systems design: Theory and applications*, SICE-ICASE International Joint Conference, Korea, October 18–21, 2006, pp. I-15–I-26.
- [45] M. D. ORTIGUEIRA, *Introduction to fractional linear systems. Part 1: Continuous-time systems*, IEE Proc. Vision, Image Signal Process., 147 (2000), pp. 62–70.
- [46] R. L. MISHKOV, *Generalization of the formula of Faa di Bruno for a composite function with a vector argument*, Internat. J. Math. Math. Sci., 24 (2000), pp. 481–491.

L^1 -APPROXIMATION OF STATIONARY HAMILTON–JACOBI EQUATIONS*

JEAN-LUC GUERMOND[†] AND BOJAN POPOV[‡]

Abstract. We describe a nonlinear finite element technique to approximate the solutions of stationary Hamilton–Jacobi equations in two space dimensions using continuous finite elements of arbitrary degree. The method consists of minimizing a functional containing the L^1 -norm of the Hamiltonian plus a discrete entropy. It is shown that the approximate sequence converges to the unique viscosity solution under appropriate hypotheses on the Hamiltonian and the mesh family.

Key words. finite elements, best L^1 -approximation, viscosity solution, Hamilton–Jacobi equations

AMS subject classifications. 35F30, 35F99, 49L25, 65N30

DOI. 10.1137/070681922

1. Introduction.

1.1. Formulation of the problem. Let Ω be an open, bounded, Lipschitz, and connected domain in \mathbb{R}^2 . We consider the stationary Hamilton–Jacobi equation

$$(1.1) \quad H(x, u, Du) = 0, \quad \text{a.e. } x \text{ in } \Omega, \quad u|_{\Gamma} = 0,$$

where Du denotes the gradient of u . We restrict ourselves to homogeneous boundary conditions to simplify the analysis. Nonhomogeneous boundary conditions can be accounted for by introducing appropriate continuous liftings provided the boundary data are compatible with our solution class; see (1.4)–(1.6).

The problem (1.1) has been extensively studied and is known to be particularly challenging in regard to the question of uniqueness. It turns out that adding a vanishing viscosity to the equation and passing to the limit usually leads to unique solutions under appropriate assumptions on the structure of the Hamiltonian H . We refer to Evans [10] for an introduction to this topic. Crandall and Lions [8] thoroughly characterized limit solutions by using the maximum principle and introducing the notions of subsolution and supersolution. They showed that the solution obtained by the vanishing viscosity limit is a subsolution and a supersolution. We refer to Barles [4] for additional details on this technique. When H is convex with respect to the gradient, Kruřkov [16] characterized the limit solution by proving that second finite differences satisfy a one-sided bound. This criteria has been significantly weakened by Lions and Souganidis [18]. It is the one-sided bound characterization of Lions and Souganidis that will be used in the present paper; see hypothesis (5.5).

The literature on the approximation of Hamilton–Jacobi equations is abundant; we refer to Sethian [19] for a thorough review. Most successful algorithms are based

*Received by the editors February 6, 2007; accepted for publication (in revised form) June 24, 2008; published electronically November 26, 2008. This material is based upon work supported by National Science Foundation grant DMS-0510650.

<http://www.siam.org/journals/sinum/47-1/68192.html>

[†]Department of Mathematics, Texas A&M University, 3368 TAMU, College Station, TX 77843, and LMSI, UPRR 3251 CNRS, BP 133, 91403 Orsay cedex, France (guermond@math.tamu.edu).

[‡]Department of Mathematics, Texas A&M University, 3368 TAMU, College Station, TX 77843 (popov@math.tamu.edu).

on monotonicity and Lax–Friedrichs approximate Hamiltonians; see, e.g., Kao, Osher, and Tsai [15]. Monotonicity is at the core of most convergence proofs for low-order approximations; see, e.g., Crandall and Lions [9], Barles and Souganidis [5], and Abgrall [1, 2]. For higher-order approximations, limiters are typically used and monotonicity cannot be preserved. Convergence results are difficult to obtain. For instance, it is shown in [18] that MUSCL-like finite difference approximations converge to viscosity solutions. In the present paper we take a radically different point of view by formulating the discrete problem as a minimization in $L^1(\Omega)$. The motivation behind this approach is based on observations made in [11] that L^1 -minimization is capable of selecting viscosity solutions of transport equations equipped with ill-posed boundary conditions. This fact has indeed been proved in [13] in one space dimension. Numerical computations in [12] confirm that this is also the case for stationary one-dimensional Hamilton–Jacobi equations. Moreover, results from Lin and Tadmor [17] show that the L^1 -metric is appropriate for deriving error estimates for time-dependent Hamilton–Jacobi equations. This encouraged us to develop a research program in this direction, and the purpose of this paper is to report that indeed L^1 -minimization is a viable technique.

In the present paper we describe a nonlinear finite element technique for approximating viscosity solutions to (1.1) in two space dimensions using continuous finite elements of arbitrary degree. The method is based on the minimization over the finite element space of a functional containing the L^1 -norm of the Hamiltonian plus a discrete entropy. Under appropriate hypotheses on the Hamiltonian, it is shown that the algorithm converges to the unique viscosity solution. The main results of this paper are Theorems 4.5, 5.3, and 6.3.

The paper is organized as follows. In the rest of this section we introduce notation and structural hypotheses for (1.1). The discrete finite element setting, along with the minimization problem, is introduced in section 2. The existence of minimizers for the discrete problem is proved in section 3. The passage to the limit is done in section 4; i.e., it is shown in this section that the limit solution solves (1.1). The proof that the limit solution is indeed a viscosity solution is reported in section 5. Since the proof reported in section 5 is based on a hypothesis which we do not know how to verify on arbitrary grids (see (3.2)), we give an alternative proof in section 6 using a vertex-based entropy on Cartesian grids. The main argument in sections 5 and 6 consists of proving a one-sided bound.

1.2. Structure hypotheses. We make the following assumptions on the Hamiltonian:

$$(1.2) \quad \|p\| \leq c_s (|H(x, v, p)| + |v| + 1) \quad \forall (x, v, p) \in \Omega \times \mathbb{R} \times \mathbb{R}^2,$$

$$(1.3) \quad H(x, \cdot, \cdot) \in C^{0,1}(\overline{B_{\mathbb{R}}(0, R)} \times \overline{B_{\mathbb{R}^2}(0, R)}; \mathbb{R}) \quad \forall R > 0 \text{ uniformly in } x \in \Omega,$$

where $\|\cdot\|$ denotes the Euclidean norm in \mathbb{R}^2 . A typical example is the eikonal equation, $H(x, v, p) = |p| - 1$, or modified versions of this equation, say, $H(x, v, p) = v + F(|p|) - f(x)$, where F is a convex and f a bounded positive function.

DEFINITION 1.1. *A function u in $W^{1,\infty}(\Omega)$ is said to be q -semiconcave if there is a concave function $v_c \in W^{1,\infty}(\Omega)$ and a function $w \in W^{2,q}(\Omega)$ so that $u = v_c + w$.*

We assume that (1.1) has a unique viscosity solution u such that

$$(1.4) \quad u \in W^{1,\infty}(\overline{\Omega}),$$

$$(1.5) \quad u \text{ is } q\text{-semiconcave for some } q > 2,$$

$$(1.6) \quad Du \in \text{BV}(\Omega),$$

where we have set $W^{1,\infty}(\overline{\Omega}) := W^{1,\infty}(\Omega) \cap C^0(\overline{\Omega})$ and D is the gradient operator.

Remark 1.1. The class of problems we are working on is not empty. In particular, it is known that the unique viscosity solution to (1.1) satisfies the above hypotheses when the Hamiltonian is convex; see [16, 18].

Remark 1.2. Hypothesis (1.2) seems nonstandard. Typical hypotheses in the literature (see [4, p. 189]) consist of assuming H to be convex with respect to p and $H(x, v, p) \rightarrow +\infty$ when $|p| \rightarrow +\infty$. It can be shown that these two conditions (convexity plus growth at infinity) imply (1.2).

Remark 1.3. Recall that a function v in $W^{1,\infty}(\Omega)$ is usually called uniformly semiconcave in textbooks if and only if it can be decomposed into $v(x) = v_c(x) + c_v x^2$, where c_v is a nonnegative constant and v_c is concave and in $W^{1,\infty}(\Omega)$; see [10, p. 130]. Definition 1.1 is a slight generalization of semiconcavity.

Remark 1.4. When Ω can be finitely covered by open convex subsets, it can be shown that (1.4) and (1.5) imply (1.6). Actually, Definition 1.1 implies $u = v_c + w$, where $v_c \in W^{1,\infty}(\Omega)$, $w \in W^{2,q}(\Omega) \subset W^{2,1}(\Omega)$. Clearly $Dw \in W^{1,1}(\Omega) \subset \text{BV}(\Omega)$. It is also known that convex functions in $W^{1,\infty}(\Omega)$ have gradients in $\text{BV}(\Omega)$; see [3, Prop. 5.1, 7.11].

To be able to collectively refer to (1.4)–(1.6), we define

$$(1.7) \quad X = \{v \in W^{1,\infty}(\Omega); Dv \in \text{BV}(\Omega); v \text{ is } q\text{-semiconcave}\}$$

with the norm

$$(1.8) \quad \|v\|_X := \|v\|_{W^{1,\infty}(\Omega)} + \|Dv\|_{\text{BV}(\Omega)} + \inf_{v=v_c+w} \|w\|_{W^{2,q}(\Omega)}.$$

In the remainder of the paper c is a generic constant that does not depend on the mesh size and whose value may change at each occurrence. For any real number $r \geq 1$, we denote by r' the conjugate of r , i.e., $\frac{1}{r} + \frac{1}{r'} = 1$.

2. The discrete problem.

2.1. The meshes. Let $\{\mathcal{T}_h\}_{h>0}$ be a family of shape regular finite element meshes. For the sake of simplicity we assume that the mesh elements are triangles and the mesh family is quasi uniform. For each mesh \mathcal{T}_h , the subscript h refers to the maximum mesh size in the mesh. We denote by \mathcal{F}_h^i the set of mesh interfaces: F is a member of \mathcal{F}_h^i if and only if there are two elements $K_1(F)$, $K_2(F)$ in \mathcal{T}_h such that $F = K_1(F) \cap K_2(F)$. The intersection of two cells is either empty, a vertex, or an entire edge. For every function $v \in C^0(K_1(F)) \cup C^0(K_2(F))$, we denote

$$(2.1) \quad \forall x \in F, \quad \{v\}(x) = \frac{1}{2}(v|_{K_1(F)}(x) + v|_{K_2(F)}(x)).$$

DEFINITION 2.1 (see Figure 1). (1) *We call a chain a numbered collection of triangles $\Lambda = \{K_j\}_{1 \leq j \leq J}$ such that K_1 has one edge on Γ , K_j shares one edge with K_{j-1} and one edge with K_{j+1} for $1 < j < J$, and $K_i \neq K_j$ if $i \neq j$.*

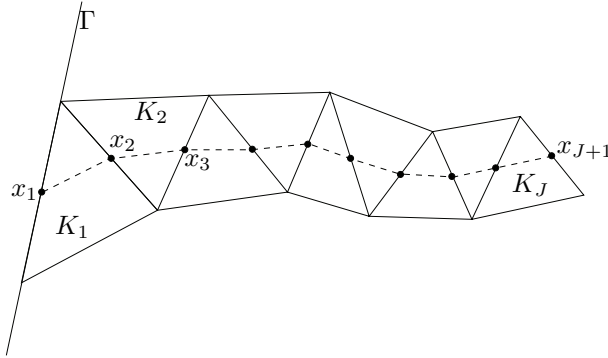


FIG. 1. Illustration of the notation for the chain and path in Definition 2.1.

(2) The path associated with a chain $\Lambda = \{K_j\}_{1 \leq j \leq J}$ is the broken line traversing the chain in such a way that (i) it crosses K_j , $1 < j < J$, by connecting the two midpoints of the two interfaces that connect K_j to the chain; (ii) it connects the midpoint of the interface connecting K_1 to the chain and the midpoint of the face of K_1 that lies on Γ ; and (iii) it connects the midpoint of the interface connecting K_J to the chain and the midpoint of another arbitrary face of K_J .

Since the mesh family $\{\mathcal{T}_h\}_{h>0}$ is quasi uniform, the following property holds: For each mesh \mathcal{T}_h , there exists a collection of chains $\{\Lambda_l\}_{1 \leq l \leq L_h}$ such that every triangle in \mathcal{T}_h belongs to at least one chain (there are $L_h \in \mathbb{N}$ of those chains). Moreover, there is c independent of h such that

$$(2.2) \quad \text{card } \Lambda_l \leq ch^{-1}, \quad 1 \leq l \leq L_h,$$

$$(2.3) \quad L_h \leq ch^{-1}.$$

Let $k \geq 1$ be an integer and denote by \mathbb{P}_k the set of real-valued polynomials in \mathbb{R}^2 of total degree at most k . We introduce

$$(2.4) \quad X_h = \{v_h \in C^0(\bar{\Omega}); v_h|_K \in \mathbb{P}_k \ \forall K \in \mathcal{T}_h; v_h|_\Gamma = 0\},$$

$$(2.5) \quad X_{(h)} = X + X_h.$$

For every function v in $X_{(h)}$ we denote by $\lambda_+(v) : \mathcal{T}_h \rightarrow \mathbb{R}^+$ the mapping such that for every $x \in K \in \mathcal{T}_h$, $\lambda_+(v)(x)$ is the largest positive eigenvalue of the Hessian of v at x . Observe that Remark 1.3 implies that λ_+ is well defined on X ; that is, λ_+ is well defined on the space $X_{(h)}$.

Similarly, for every function v in $X_{(h)}$ we denote by $\{-\partial_n v\}_+ : \mathcal{F}_h^i \rightarrow \mathbb{R}^+$ the mapping such that for all $x \in F = K_1 \cap K_2 \in \mathcal{F}_h^i$,

$$\{-\partial_n v\}_+(x) = \left(-\frac{1}{2}(Dv|_{K_1}(x) \cdot n_1 + Dv|_{K_2}(x) \cdot n_2) \right)_+,$$

where n_1 and n_2 are the unit outward normals to K_1 and K_2 at x , respectively, and $(t)_+ := \frac{1}{2}(t + |t|)$ denotes the positive part of t for all $t \in \mathbb{R}$.

2.2. The discrete minimization problem. Let p_1 and p_2 be two fixed real numbers such that

$$(2.6) \quad 1 \leq p_1 \leq q \quad \text{and} \quad 1 \leq p_2 \leq q.$$

We now define the following functional $J_h : X_{(h)} \ni v \mapsto J_h(v) \in \mathbb{R}^+$ by

$$(2.7) \quad J_h(v) = \int_{\Omega} |H(x, v, Dv)| dx + h \sum_{K \in \mathcal{T}_h} \int_K [\lambda_+(v)]^{p_1} dx + h^{2-p_2} \sum_{F \in \mathcal{F}_h^i} \int_F \{-\partial_n v\}_+^{p_2} d\sigma.$$

For every function v in $X_{(h)}$ we refer to $H(x, v, Dv)$ as the residual. The two extra terms in the right-hand side above are referred to as the volume entropy and the interface entropy, respectively.

Remark 2.1. Whenever $v \in W^{1,\infty}(\Omega)$ is concave, the two entropy terms are zero; i.e., these two terms do not add extra viscosity. They act to prevent the occurrence of large positive second derivatives.

The discrete problem on which we shall henceforth focus our attention consists of the following minimization problem: Seek u_h in X_h such that

$$(2.8) \quad J_h(u_h) = \inf_{v_h \in X_h} J_h(v_h).$$

The goal of the rest of the paper is to show that minimizers exist for each mesh and every sequence of minimizers converges to the unique viscosity solution to (1.1).

3. Existence of minimizers. The goal of this section is to show the existence of (at least) one minimizer to problem (2.8). This is done by deriving a priori bounds and using a simple compactness argument.

3.1. Consistency. We start by deriving a consistency property. Since the mesh is quasi uniform, one can always construct a family of linear approximation operators on piecewise linear polynomials $\mathcal{I}_h : X \rightarrow X_h$ that are stable on $W^{1,\infty}(\Omega)$ and such that the following property holds:

$$(3.1) \quad \|v - \mathcal{I}_h v\|_{W^{1,1}(\Omega)} \leq ch \|Dv\|_{\text{BV}(\Omega)} \quad \forall v \in X.$$

This is a standard approximation property; for instance, the linear Clément operator [7, 6] satisfies this property (see also section 6.2 for a precise definition of the Clément approximation). It is also clear that for $p_2 = 1$ the Clément approximation (or any other reasonable approximation operator) satisfies

$$(3.2) \quad h^{2-p_2} \sum_{F \in \mathcal{F}_h^i} \int_F \{-\partial_n \mathcal{I}_h(u)\}_+^{p_2} d\sigma \leq c(\|u\|_X) h.$$

When $p_2 > 1$, (3.2) becomes a nontrivial property. In our previous paper [14], which deals with the one-dimensional case, we showed that the piecewise linear Lagrange interpolant of the exact solution satisfies (3.2) with $p_2 > 1$. Unfortunately, this argument does not hold in two space dimensions. To see this, assume that the gradient of u is discontinuous across a line L and the mesh family is such that $\mathcal{O}(h^{-1})$ cell interfaces cross L . Then the left-hand side of (3.2) is bounded from below and above by ch^{2-p_2} which is larger than ch unless $p_2 = 1$. In other words, (3.2) is hard to verify when $p_2 > 1$. Of course (3.2) can be shown to hold with $p_2 > 1$ if we are allowed to optimize or control the mesh. For instance, the Lagrange interpolant of u satisfies (3.2) with $p_2 > 1$ if the mesh family is such that no cell interface crosses the

lines across which the normal derivative of u is discontinuous; i.e., the mesh is aligned with the discontinuity lines of the gradient of u .

Henceforth we make the following assumption:

$$(3.3) \quad \left\{ \begin{array}{l} \text{If } p_2 > 1, \text{ there exists an approximation operator } \mathcal{I}_h \text{ such} \\ \text{that (3.1) and (3.2) hold simultaneously.} \end{array} \right.$$

The existence of such an operator (for $p_2 > 1$) is an open question when the mesh is not aligned with the discontinuities of the gradient of u . Note that the assumption is empty when $p_2 = 1$.

The following lemma is the first key step of the theory.

LEMMA 3.1. *Let u solve (1.1) and assume (3.3); then there is $c(u)$ independent of h such that*

$$(3.4) \quad J_h(\mathcal{I}_h u) \leq c(u) h.$$

Proof. (1) Since $\mathcal{I}_h u$ is piecewise linear, the restriction of the Hessian of $\mathcal{I}_h u$ to every mesh element is zero, i.e., $\lambda_+(\mathcal{I}_h u)|_K = 0$ for all $K \in \mathcal{T}_h$.

(2) Since \mathcal{I}_h is uniformly stable in $W^{1,\infty}(\Omega)$, there is $c \geq 0$, independent of h , such that $\|\mathcal{I}_h u\|_{W^{1,\infty}} \leq c \|u\|_{W^{1,\infty}}$. Let us set $R = c \|u\|_{W^{1,\infty}}$; then owing to (1.3), there is $c_R \geq 0$ such that for all $x \in \Omega$,

$$\begin{aligned} |H(x, \mathcal{I}_h u, D(\mathcal{I}_h u))| &= |H(x, \mathcal{I}_h u, D(\mathcal{I}_h u)) - H(x, u, Du)| \\ &\leq c_R (|\mathcal{I}_h u - u| + \|D(\mathcal{I}_h u - u)\|). \end{aligned}$$

Then together with (3.1), this implies

$$\int_{\Omega} |H(x, \mathcal{I}_h u, D(\mathcal{I}_h u))| \leq c_R \|\mathcal{I}_h u - u\|_{W^{1,1}} \leq c c_R h \|Du\|_{\text{BV}(\Omega)}.$$

(3) We now conclude by using (3.2) and collecting the above results:

$$J_h(\mathcal{I}_h u) \leq c(\|u\|_X) h \leq c' h.$$

This concludes the proof. \square

Remark 3.1. Note that it was critical to use the L^1 -norm of the residual to obtain (3.4). This is compatible with the fact that Du is in $\text{BV}(\Omega)$ only. Using any other L^p -norm would yield a suboptimal exponent on h .

3.2. The $W^{1,1}(\Omega) \cap L^\infty(\Omega)$ bound. Let $\alpha > 0$ be a positive real number. Define the set $S_{h,\alpha} = \{v_h \in X_h; \int_{\Omega} |H(x, v_h, Dv_h)| dx \leq \alpha h\}$. Using (3.4), we infer that $S_{h,c(u)}$ is not empty, i.e., $\mathcal{I}_h(u) \in S_{h,c(u)}$.

LEMMA 3.2. *Let $\alpha > 0$ and assume that $S_{h,\alpha}$ is not empty; then there is $c_0(\alpha) > 0$, independent of h , and $h_0 > 0$ such that*

$$(3.5) \quad \forall h < h_0, \quad \forall v_h \in S_{h,\alpha}, \quad \|v_h\|_{W^{1,1}} + \|v_h\|_{L^\infty} \leq c_0(\alpha).$$

Proof. Let Λ_l be a chain in the collection $\{\Lambda_l\}_{1 \leq l \leq L_h}$ (see Figure 1). Set $N_l = \text{card}(\Lambda_l)$, let v_h be a member of the nonempty set $S_{h,\alpha}$, and define

$$F_j^l = \sum_{i=1}^j \int_{K_i} \|Dv_h\| dx, \quad 1 \leq j \leq N_l.$$

Owing to (1.2), we infer

$$F_j^l \leq \sum_{i=1}^j \int_{K_i} c_s (|H(x, v_h, Dv_h)| dx + |v_h| + 1) dx.$$

Then using the fact that v_h is a member of $S_{h,\alpha}$ implies

$$F_j^l \leq \alpha c_s h + c_s N_l \max_{K \in \mathcal{T}_h} \text{meas}(K) + \sum_{i=1}^j c_s \int_{K_i} |v_h| dx.$$

In other words, using (2.2) together with $\max_{K \in \mathcal{T}_h} \text{meas}(K) \leq ch^2$, we infer

$$(3.6) \quad F_j^l \leq ch + c_s \sum_{i=1}^j \int_{K_i} |v_h| dx.$$

Let x_1, \dots, x_i be the points of the path traversing Λ_l such that $(x_m, x_{m+1}) = \Lambda_l \cap K_m$, $1 \leq m < N_l$ (see Figure 1). Denote $\tau_m = (x_{m+1} - x_m) / \|x_{m+1} - x_m\|$. Let us now consider a cell K_i , $1 \leq i \leq j$, and let y be an arbitrary point in K_i ; then the fundamental theorem of calculus implies

$$v_h(y) = v_h(x_1) + \sum_{m=1}^{i-1} \int_{x_m}^{x_{m+1}} \tau_m \cdot Dv_h(x) d\sigma + \int_{x_i}^y \frac{y - x_i}{\|y - x_i\|} \cdot Dv_h(x) d\sigma,$$

with the obvious convention if $i = 1$ or $y = x_i$. This in turn implies

$$|v_h(y)| \leq ch \sum_{m=1}^i \|Dv_h\|_{L^\infty(K_i)} \quad \forall y \in K_i.$$

Since $v_h|_{K_m}$ is a polynomial, the equivalence of norms on finite-dimensional normed spaces gives in two space dimensions

$$(3.7) \quad |v_h(y)| \leq ch^{-1} \sum_{m=1}^i \|Dv_h\|_{L^1(K_i)} = ch^{-1} F_i^l \quad \forall y \in K_i.$$

By integrating over K_i , we obtain

$$(3.8) \quad \int_{K_i} |v_h| dx \leq ch F_i^l.$$

By combining (3.6) and (3.8), we infer $F_j^l \leq ch + c'h \sum_{i=1}^j F_i^l$, which (provided h is small enough) immediately implies

$$F_j^l \leq \frac{ch}{1 - c'h} \left(1 + \frac{c'h}{1 - c'h}\right)^j.$$

Then, owing to (2.2) we have $j \leq N_l \leq ch^{-1}$, which in turn yields

$$F_j^l \leq ch,$$

which owing to (3.7) implies the desired L^∞ -bound on v_h .

We obtain the $W^{1,1}$ -bound by using the property saying that Ω can be covered with chains, i.e.,

$$\|Dv_h\|_{L^1(\Omega)} \leq \sum_{l=1}^{L_h} \sum_{K \in \Lambda_l} \|Dv_h\|_{L^1(K)} \leq c \sum_{l=1}^{L_h} F_l^{N_l} \leq chL_h \leq c,$$

which concludes the proof. \square

We are now in a position to prove the existence of a minimizer solving problem (2.8).

COROLLARY 3.3. *The discrete problem (2.8) has at least one minimizer u_h , and there is $c > 0$ independent of h such that*

$$(3.9) \quad J_h(u_h) \leq ch,$$

$$(3.10) \quad \|u_h\|_{W^{1,1}} + \|u_h\|_{L^\infty} \leq c.$$

Proof. Observe first that Lemma 3.1 implies that $S_{h,c(u)}$ is not empty, since $\mathcal{I}_h(u) \in S_{h,c(u)}$. Second, define $\mathcal{K}_h = \{v_h \in X_h; J_h(v_h) \leq J_h(\mathcal{I}_h u)\}$. Clearly $\mathcal{I}_h u$ is a member of \mathcal{K}_h . Moreover, owing to Lemma 3.1, for every v_h in \mathcal{K}_h ,

$$\int_{\Omega} |H(x, v_h, Dv_h)| dx \leq J_h(v_h) \leq J_h(\mathcal{I}_h u) \leq c(u) h.$$

That is, $\mathcal{K}_h \subset S_{h,c(u)}$. Lemma 3.2 implies that there is $c'(u)$ independent of h such that for all $v_h \in \mathcal{K}_h$, $\|v_h\|_{L^\infty} + \|v_h\|_{W^{1,1}} \leq c'(u)$. In other words, \mathcal{K}_h is uniformly bounded in $W^{1,1}(\Omega) \cap L^\infty(\Omega)$. Finite-dimensionality then implies that \mathcal{K}_h is compact. It is also clear that $J_h : \mathcal{K}_h \rightarrow \mathbb{R}$ is continuous in every norm (possibly not uniformly with respect to h). Then, there exists $u_h \in \mathcal{K}_h$ that minimizes J_h on \mathcal{K}_h . Since for every function v_h in $X_h \setminus \mathcal{K}_h$, $J_h(v_h)$ is larger than $J_h(\mathcal{I}_h u)$, we conclude that

$$\inf_{v_h \in X_h} J_h(v_h) = \inf_{v_h \in \mathcal{K}_h} J_h(v_h) = \min_{v_h \in \mathcal{K}_h} J_h(v_h) = J_h(u_h),$$

which concludes the proof. \square

Since in practice u_h might not be computed exactly or might be approximated to some extent by using some iterative process (the details of the process in question are irrelevant for our discussion), we now define the notion of an *almost minimizer*.

DEFINITION 3.1. *We say that a family of functions $\{v_h \in X_h\}_{h>0}$ is a sequence of almost minimizers for (1.1) if there is $c > 0$ such that for all $h > 0$,*

$$(3.11) \quad J_h(v_h) \leq ch.$$

It is clear that minimizers are almost minimizers, thus showing that the class of almost minimizers is not empty. Almost minimizers also satisfy the following uniform bound owing to Lemma 3.2:

$$(3.12) \quad \|v_h\|_{W^{1,1}} + \|v_h\|_{L^\infty} \leq c.$$

The rest of the paper consists of proving that sequences of almost minimizers for (1.1) converge to the viscosity solution of (1.1).

4. Passage to the limit. Henceforth $\{u_h \in X_h\}_{h>0}$ denotes a sequence of almost minimizers for (1.1) as defined above. We show in this section that sequences of almost minimizers for (1.1) converge to weak solutions of (1.1). The main result of this section is Theorem 4.5. That the limit solution is indeed the viscosity limit will be shown in sections 5 and 6 by proving a one-sided bound.

4.1. The $W^{1,\infty}(\overline{\Omega})$ -bound. We prove a $W^{1,\infty}(\overline{\Omega})$ -bound by using a regularization technique. Let $\rho : \mathbb{R}^2 \rightarrow \mathbb{R}^+$ be a positive kernel, i.e., ρ is compactly supported in $B_{\mathbb{R}^2}(0, 1)$ and $\int_{\mathbb{R}^2} \rho dx = 1$. We then define the sequence of mollifiers $\rho_\epsilon(x) = \rho(x/\epsilon)$ with

$$(4.1) \quad \epsilon = h^{1/2}.$$

This scaling is justified by the fact that $h\|\rho_\epsilon\|_{L^\infty} \leq c$, and this property is used in the proof of Lemma 4.1; see also Remark 4.1. Let us denote by \tilde{u}_h the extension of u_h to \mathbb{R}^2 by setting $u_h|_{\mathbb{R}^2 \setminus \Omega} = 0$. Since $u_h|_\Gamma = 0$, u_h is continuous and piecewise polynomial in $\overline{\Omega}$, this extension is $W^{1,\infty}$ -stable, i.e., $\|\tilde{u}_h\|_{W^{1,\infty}(\mathbb{R}^2)} \leq \|u_h\|_{W^{1,\infty}(\Omega)}$. We now define

$$(4.2) \quad u_{\epsilon,h} = \rho_\epsilon * \tilde{u}_h.$$

The main result of this section is the following lemma.

LEMMA 4.1. *There is a constant c , independent of h , such that*

$$\|u_{\epsilon,h}\|_{W^{1,\infty}(\overline{\Omega})} \leq c.$$

Proof. Let x be any point in \mathbb{R}^2 . We have

$$\begin{aligned} \|Du_{\epsilon,h}(x)\| &= \left| \int_{\mathbb{R}^2} \rho_\epsilon(x-y) D\tilde{u}_h(y) dy \right| \\ &\leq \int_{\Omega} \rho_\epsilon(x-y) \|D\tilde{u}_h(y)\| dy = \int_{\Omega} \rho_\epsilon(x-y) \|Du_h(y)\| dy \\ &\leq c_s \int_{\Omega} \rho_\epsilon(x-y) (|H(y, u_h(y), Du_h(y))| + |u_h(y)| + 1) dy \\ &\leq c_s \|\rho_\epsilon\|_{L^\infty(\Omega)} \int_{\Omega} |H(y, u_h(y), Du_h(y))| dy + (\|u_h\|_{L^\infty(\Omega)} + 1) \|\rho_\epsilon\|_{L^1(\Omega)}. \end{aligned}$$

The estimates (3.11) and (3.12) imply

$$\|Du_{\epsilon,h}(x)\| \leq c(h\|\rho_\epsilon\|_{L^\infty(\Omega)} + \|\rho_\epsilon\|_{L^1(\Omega)}).$$

Then the estimates $\|\rho_\epsilon\|_{L^\infty(\Omega)} \leq c\epsilon^{-2}$ and $\|\rho_\epsilon\|_{L^1(\Omega)} = 1$ along with the definition of ϵ (4.1) imply the desired result. \square

Remark 4.1. The above result generalizes to any space dimension d if estimates (3.11) and (3.12) hold and provided we take the scaling $\epsilon = h^{1/d}$.

4.2. The BV-bound on Du_h . We prove in this section an a priori bound on the BV-norm of Du_h . We start with a technical result concerning interface averages of the gradient of functions in X_h . Let (e_1, e_2) be the canonical basis of \mathbb{R}^2 .

LEMMA 4.2. *For all $v_h \in X_h$ and all $F \in \mathcal{F}_h^i$, the following holds:*

$$(4.3) \quad \{(n \cdot e_j) \partial_i v_h\}|_F = \{\partial_n v_h\}|_F (n_1 \cdot e_i)(n_1 \cdot e_j) = \{\partial_n v_h\}|_F (n_2 \cdot e_i)(n_2 \cdot e_j),$$

where n_1 (resp., n_2) is the unit outer normal of $K_1(F)$ on F (resp., $K_2(F)$ on F).

Proof. Let $\tau_1 := -\tau_2$ be one of the two unit vectors that are parallel to F (see Figure 2). Upon observing that $e_i = (n_1 \cdot e_i)n_1 + (\tau_1 \cdot e_i)\tau_1 = (n_2 \cdot e_i)n_2 + (\tau_2 \cdot e_i)\tau_2$,

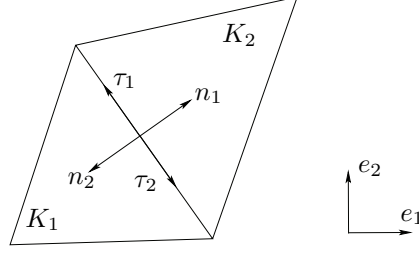


FIG. 2. Illustration of the notation for Lemma 4.2.

$(n_1 \cdot e_i)(n_1 \cdot e_j) = (n_2 \cdot e_i)(n_2 \cdot e_j)$, and $(n_1 \cdot e_i)(\tau_1 \cdot e_j) = (n_2 \cdot e_i)(\tau_2 \cdot e_j)$, we infer

$$\begin{aligned}
 \{(n \cdot e_j) \partial_i v_h\} &= \frac{1}{2} ((n_1 \cdot e_j) \partial_i v_h|_{K_1} + (n_2 \cdot e_i) \partial_i v_h|_{K_2}) \\
 &= \frac{1}{2} ((n_1 \cdot e_j) (Dv_h|_{K_1}) \cdot ((n_1 \cdot e_i) n_1 + (\tau_1 \cdot e_i) \tau_1) \\
 &\quad + (n_2 \cdot e_j) (Dv_h|_{K_2}) \cdot ((n_2 \cdot e_i) n_2 + (\tau_2 \cdot e_i) \tau_2)) \\
 &= \{\partial_n v_h\} (n_1 \cdot e_i) (n_1 \cdot e_j) + \{\partial_\tau v_h\} (n_1 \cdot e_i) (\tau_1 \cdot e_j).
 \end{aligned}$$

Then, we conclude by observing that functions in X_h are continuous across interfaces, which implies $\{\partial_\tau v_h\} = 0$. \square

Another preliminary result consists of bounding the normal derivative of u_h at the boundary of the domain. This is achieved by means of the following lemma.

LEMMA 4.3. *There is c , independent of h , such that*

$$(4.4) \quad \int_{\Gamma} |\partial_n u_h| dx \leq c.$$

Proof. Let us denote by \mathcal{L}_h the layer of triangles that have at least one edge on Γ . Using an inverse inequality and (1.2), we deduce

$$\begin{aligned}
 \int_{\Gamma} |\partial_n u_h| dx &\leq c h^{-1} \sum_{K \in \mathcal{L}_h} \int_K \|Du_h\| dx \leq c h^{-1} \sum_{K \in \mathcal{L}_h} \int_K (|H(x, u_h, Du_h)| + |u_h| + 1) dx \\
 &\leq c h^{-1} \int_{\Omega} |H(x, u_h, Du_h)| dx + c h^{-1} \sum_{K \in \mathcal{L}_h} \int_K (|u_h| + 1) dx.
 \end{aligned}$$

Then we conclude using the estimates (3.11)–(3.12) together with the fact that $\sum_{K \in \mathcal{L}_h} \text{meas}(K) \leq ch$. \square

LEMMA 4.4. *There is c , independent of h , such that*

$$(4.5) \quad \|Du_h\|_{BV(\Omega)} \leq c.$$

Proof. Using (4.3) and the definition of the BV-seminorm implies

$$\begin{aligned} |Du_h|_{\text{BV}(\Omega)} &= \sum_{i,j=1}^2 \sup_{\substack{\phi \in C_0^\infty(\Omega) \\ \|\phi\|_{L^\infty} \leq 1}} \int_{\Omega} \partial_i u_h \partial_j \phi dx \\ &= \sum_{i,j=1}^2 \left(\sup_{\substack{\phi \in C_0^\infty(\Omega) \\ \|\phi\|_{L^\infty} \leq 1}} \sum_{K \in \mathcal{T}_h} \int_K -\partial_{ij} u_h \phi dx + \sum_{F \in \mathcal{F}_h^i} \int_F 2\{\partial_i u_h(n \cdot e_j)\} \phi ds \right) \\ &\leq \sum_{i,j=1}^2 \left(\sum_{K \in \mathcal{T}_h} \int_K |\partial_{ij} u_h| dx + \sum_{F \in \mathcal{F}_h^i} \int_F 2\{|\partial_n u_h|\} ds \right) \\ &\leq \sum_{K \in \mathcal{T}_h} \int_K (|\partial_{11} u_h| + |\partial_{22} u_h| + 2|\partial_{12} u_h|) dx + \sum_{F \in \mathcal{F}_h^i} \int_F 8\{|\partial_n u_h|\} ds. \end{aligned}$$

Now we use the relation $|x| = 2x_+ - x$ as follows:

$$\begin{aligned} |Du_h|_{\text{BV}(\Omega)} &\leq \sum_{K \in \mathcal{T}_h} \int_K (2((\partial_{11} u_h)_+ + (\partial_{22} u_h)_+) - \Delta u_h + 2|\partial_{12} u_h|) dx \\ &\quad + \sum_{F \in \mathcal{F}_h^i} \int_F (16\{|\partial_n u_h|\}_+ - 8\{|\partial_n u_h|\}) ds. \end{aligned}$$

Moreover, the definition of λ_+ implies that for all $x \in K$ and all $K \in \mathcal{T}_h$,

$$\begin{aligned} \max((\partial_{11} u_h)_+(x), (\partial_{22} u_h)_+(x)) &\leq \lambda_+(x), \\ |\partial_{12} u_h(x)| &\leq \lambda_+(x) - \frac{1}{2} \Delta u_h(x). \end{aligned}$$

Then,

$$\begin{aligned} |Du_h|_{\text{BV}(\Omega)} &\leq \sum_{K \in \mathcal{T}_h} \int_K (6\lambda_+ - 2\Delta u_h) dx + \sum_{K \in \mathcal{T}_h} \int_K 4\Delta u_h dx \\ &\quad + \sum_{F \in \mathcal{F}_h^i} \int_F (16\{|\partial_n u_h|\}_+) ds - \sum_{F \in \mathcal{F}_h^0} \int_F 4\partial_n u_h ds. \end{aligned}$$

Now we use $\Delta u_h \leq 2\lambda_+$ to derive

$$|Du_h|_{\text{BV}(\Omega)} \leq \sum_{K \in \mathcal{T}_h} \int_K 10\lambda_+ dx + \sum_{F \in \mathcal{F}_h^i} \int_F (16\{|\partial_n u_h|\}_+) ds + \int_{\Gamma} 4|\partial_n u_h| ds.$$

Let R_1 , R_2 , and R_3 be the three terms in the right-hand side of the above inequality.

We bound $R_1 + R_2$ as follows:

$$\begin{aligned} R_1 + R_2 &\leq c_1 \left(\sum_{K \in \mathcal{T}_h} \text{meas}(K) \right)^{1/p_1'} \left(\sum_{K \in \mathcal{T}_h} \int_K \lambda_+^{p_1} \right)^{1/p_1} \\ &\quad + c_2 \left(h^{\frac{p_2-1}{p_2} p_2'} \sum_{F \in \mathcal{F}_h^i} \text{meas}(F) \right)^{1/p_2'} \left(h^{1-p_2} \sum_{F \in \mathcal{F}_h^i} \int_F \{|\partial_n u_h|\}_+^{p_2} ds \right)^{1/p_2}. \end{aligned}$$

Then using the estimate on $J_h(u_h)$ in (3.11), we derive

$$R_1 + R_2 \leq c.$$

To conclude that R_3 is also bounded, we use the estimate (4.4). We conclude that $\|Du_h\|_{\text{BV}(\Omega)}$ is uniformly bounded by using the fact that Du_h is also uniformly bounded in $L^1(\Omega)$. \square

4.3. Convergence to a weak solution. We say that $v \in W^{1,\infty}(\overline{\Omega})$ is a weak solution to (1.1) if v solves (1.1) almost everywhere.

THEOREM 4.5. *Assume that (1.1) has a solution u in X and that the mesh family satisfies (3.3). Then the sequence of almost minimizers $\{u_h\}_{h>0}$ converges, up to subsequences, to a weak solution to (1.1).*

Proof. Owing to Corollary 3.3 and Lemma 4.4, the sequence $\{u_h\}_{h>0}$ is precompact in $W^{1,1}(\Omega)$. Let u be the limit, up to subsequences, of $\{u_h\}_{h>0}$ in $W^{1,1}(\Omega)$. We need to show that u is also in $W^{1,\infty}(\overline{\Omega})$. To see this, we observe that, up to subsequences again, $\{u_h\}_{h>0}$ and $\{u_{\epsilon,h}\}_{h>0}$ have the same limit in $W^{1,1}(\Omega)$ since

$$\|u_h - u_{\epsilon,h}\|_{W^{1,1}(\Omega)} \leq \|u_h - u\|_{W^{1,1}(\Omega)} + \|u - \rho_\epsilon * u\|_{W^{1,1}(\Omega)} + \|\rho_\epsilon * (u - u_h)\|_{W^{1,1}(\Omega)},$$

and the right-hand side goes to zero as $h \rightarrow 0$, owing to well-known properties of mollifiers, recalling that $\epsilon = h^{\frac{1}{2}}$. Moreover, the sequence $\{u_{\epsilon,h}\}_{h>0}$, being uniformly bounded in $W^{1,\infty}(\overline{\Omega}) \subset W^{1,\infty}(\Omega)$, converges in $W^{1,\infty}(\Omega)$ in the weak-* topology, up to subsequences. The uniqueness of limits implies that u is in $W^{1,\infty}(\Omega)$. The sequence $\{u_{\epsilon,h}\}_{h>0}$ being uniformly bounded in $W^{1,\infty}(\overline{\Omega})$ means that it is equicontinuous on $\overline{\Omega}$; as a result, the limit is continuous, i.e., $u \in C^0(\overline{\Omega})$. Combining the two above results implies $u \in W^{1,\infty}(\overline{\Omega}) = C^0(\overline{\Omega}) \cap W^{1,\infty}(\Omega)$.

We now prove that u is a weak solution to (1.1) by showing that $\|H(\cdot, u, Du)\|_{L^1(\Omega)} = 0$. Using that $u_h \rightarrow u$ in $W^{1,1}(\Omega)$, we conclude that, up to subsequences, $u_h \rightarrow u$ and $Du_h \rightarrow Du$ a.e. in Ω . Then, we can apply Egorov's theorem. Given $\epsilon' > 0$, there exists a set E with $\text{meas}(E) < \epsilon'$, such that the convergence of $u_h \rightarrow u$ on $\Omega \setminus E$ is uniform. Therefore, for every $\epsilon'', 1 \geq \epsilon'' > 0$, we can find $h(\epsilon'') > 0$ such that for every $h < h(\epsilon'')$,

$$|u_h(x) - u(x)| < \epsilon'' \quad \text{and} \quad \|Du_h(x) - Du(x)\| < \epsilon'' \quad \forall x \in \Omega \setminus E.$$

Note also that for every $x \in \Omega \setminus E$ and every $h < h(1)$, we have

$$\max(|u_h(x)|, |u(x)|, \|Du_h(x)\|, \|Du(x)\|) \leq R,$$

where $R := \|u\|_{W^{1,\infty}(\Omega)} + 1$. Hence, we can use the Lipschitz continuity of H to derive that there exists a value of $\epsilon'' > 0$ such that

$$(4.6) \quad |H(x, u, Du) - H(x, u_h, Du_h)| < \epsilon'$$

for every $x \in \Omega \setminus E$ and every $h < h(\epsilon'')$. Note that at this point the value of ϵ'' solely depends on ϵ' . We now split $\|H(\cdot, u, Du)\|_{L^1(\Omega)}$ in the following way:

$$(4.7) \quad \|H(\cdot, u, Du)\|_{L^1(\Omega)} = \|H(\cdot, u, Du)\|_{L^1(\Omega \setminus E)} + \|H(\cdot, u, Du)\|_{L^1(E)}.$$

We use that for every $R > 0$, $H(x, \cdot, \cdot)$ is Lipschitz continuous on $\overline{B_{\mathbb{R}}(0, R)} \times \overline{B_{\mathbb{R}^2}(0, R)}$ uniformly with respect to x to estimate

$$\|H(\cdot, u, Du)\|_{L^1(E)} \leq c \text{meas}(E) = c\epsilon'.$$

The other term in the right-hand side of (4.7) is estimated as follows:

$$\begin{aligned} & \|H(\cdot, u, Du)\|_{L^1(\Omega \setminus E)} \\ & \leq \|H(\cdot, u, Du) - H(\cdot, u_h, Du_h)\|_{L^1(\Omega \setminus E)} + \|H(\cdot, u_h, Du_h)\|_{L^1(\Omega \setminus E)} \\ & \leq \epsilon' \operatorname{meas}(\Omega \setminus E) + \|H(\cdot, u_h, Du_h)\|_{L^1(\Omega)} \\ & \leq c\epsilon' + ch, \end{aligned}$$

where we used (4.6) and (3.11) to derive the above inequality. As a result, for every $\epsilon' > 0$ and every $h < h(\epsilon')$,

$$\|H(\cdot, u, Du)\|_{L^1(\Omega)} \leq c(\epsilon' + h),$$

which means $\|H(\cdot, u, Du)\|_{L^1(\Omega)} = 0$. □

Remark 4.2. Recall that when $p_2 = 1$, hypothesis (3.3) is empty, i.e., Theorem 4.5 holds without any assumption on the mesh family other than that it is quasi uniform.

Now we have to address the question of whether this weak solution is indeed the viscosity solution.

5. One-sided bound. The goal of this section is to show that the algorithm described in this paper using the functional defined in (2.7) converges to the viscosity solution to (1.1) under the assumptions

$$(5.1) \quad p_1 > 2 \quad \text{and} \quad p_2 > 2.$$

Throughout section 5 we conjecture (3.3). That is, there exists an approximation operator \mathcal{I}_h satisfying (3.1) and (3.2) simultaneously for every mesh family. We have not been able to prove this statement for arbitrary meshes (unless the discontinuity lines of the gradient of u are aligned with the mesh). An alternative proof of convergence is reported in section 6 using a vertex-based entropy assuming that the mesh family is Cartesian and $p_1 = p_2 = 1$ so that the Clément interpolant always satisfies (3.1)–(3.2); i.e., the assumption (3.3) is empty.

Observe that if a function v is q -semiconcave, then there is $c > 0$ such that for all $\delta > 0$, all $\omega \subset \Omega$ so that $\omega + \delta e \subset \Omega$, and for every unit vector $e \in \mathbb{R}^2$, the following hold:

$$(5.2) \quad u(x + \delta e) - 2u(x) + u(x - \delta e) \leq c\delta^{2-\frac{2}{q}} \quad \forall x \in \omega,$$

$$(5.3) \quad \|(u(\cdot + \delta e) - 2u(\cdot) + u(\cdot - \delta e))_+\|_{L^q(\omega)} \leq c\delta^2.$$

Note that (5.2) implies that for every orthonormal basis of \mathbb{R}^2 , say (f_1, f_2) , every $\delta > 0$, every $\gamma \leq 1 - \frac{2}{q}$, and every $x \in \omega$, the following holds:

$$(5.4) \quad \Delta_\delta u(x) := \sum_{i=1}^2 u(x + \delta f_i) - 2u(x) + u(x - \delta f_i) \leq c\delta^{1+\gamma}.$$

To stay general in the remainder of the paper we make the following assumption:

$$(5.5) \quad \left\{ \begin{array}{l} \text{A weak solution } u \text{ to (1.1) is the unique viscosity solution if } u \in \\ W^{1,\infty}(\Omega) \text{ and there exist an orthonormal basis } (f_1, f_2) \text{ of } \mathbb{R}^2 \text{ and} \\ \gamma > 0 \text{ such that (5.4) is satisfied.} \end{array} \right.$$

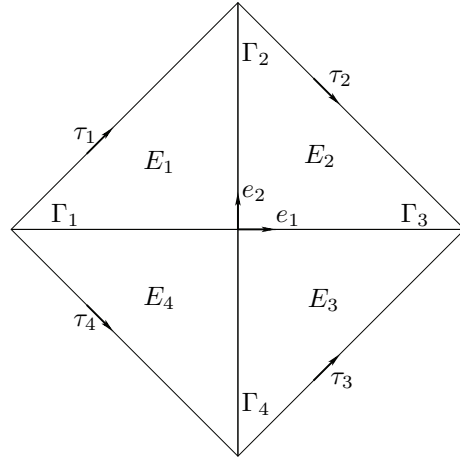


FIG. 3. Illustration of the notation for Lemma 5.1.

This property is known to characterize viscosity solutions to stationary Hamilton–Jacobi equations with $H(x, u, Du) = u + F(Du)$, where $F : \mathbb{R}^2 \rightarrow \mathbb{R}$ is convex as shown by Lions and Souganidis [18, Thm. 2.6].

Throughout section 5 the orthonormal basis that we use is the canonical one (e_1, e_2) and the discrete Laplacian $\Delta_\delta u(x)$ is defined using this basis.

Let $x \in \Omega$ and $\delta > 0$ such that $B_{\mathbb{R}^2}(x, \delta) \subset \Omega$ and let us consider the square whose four vertices are $x - \delta e_1, x + \delta e_2, x + \delta e_1,$ and $x - \delta e_2$. This square is the union of the following four triangles:

$$\begin{aligned}
 E_1 &= x + \{(x_1, x_2) \in \mathbb{R}^2; x_1 \leq 0; x_2 \geq 0; x_1 - x_2 + \delta \geq 0\}, \\
 E_2 &= x + \{(x_1, x_2) \in \mathbb{R}^2; x_1 \geq 0; x_2 \geq 0; x_1 + x_2 - \delta \leq 0\}, \\
 E_3 &= x + \{(x_1, x_2) \in \mathbb{R}^2; x_1 \geq 0; x_2 \leq 0; x_1 - x_2 - \delta \leq 0\}, \\
 E_4 &= x + \{(x_1, x_2) \in \mathbb{R}^2; x_1 \leq 0; x_2 \leq 0; x_1 + x_2 + \delta \geq 0\}.
 \end{aligned}
 \tag{5.6}$$

The interior of E_i is henceforth denoted by $\dot{E}_i, i \in \{1, 2, 3, 4\}$. We also set

$$\begin{aligned}
 \Gamma_1 &= x + \{(x_1, 0) \in \mathbb{R}^2; -\delta \leq x_1 \leq 0\}, \\
 \Gamma_2 &= x + \{(0, x_2) \in \mathbb{R}^2; 0 \leq x_2 \leq \delta\}, \\
 \Gamma_3 &= x + \{(x_1, 0) \in \mathbb{R}^2; 0 \leq x_1 \leq \delta\}, \\
 \Gamma_4 &= x + \{(0, x_2) \in \mathbb{R}^2; -\delta \leq x_2 \leq 0\}.
 \end{aligned}
 \tag{5.7}$$

We now define the unit vectors $\tau_1 = 2^{-\frac{1}{2}}(e_1 + e_2), \tau_2 = 2^{-\frac{1}{2}}(e_1 - e_2), \tau_3 = \tau_1,$ and $\tau_4 = \tau_2$. See Figure 3.

We are now in a position to derive an integral representation of $\Delta_\delta u_h(x)$ over the square $E_1 \cup E_2 \cup E_3 \cup E_4$.

LEMMA 5.1. *The following holds for all $v_h \in X_h$ and all $x \in \Omega$ and $\delta > 0$ such that $B_{\mathbb{R}^2}(x, \delta) \subset \Omega$:*

$$(5.8) \quad \Delta_\delta v_h(x) = \sum_{l=1}^4 \sum_{\substack{K \in \mathcal{T}_h \\ K \cap E_l \neq \emptyset}} \int_{K \cap E_l} \partial_{\tau_l \tau_l} v_h + 2 \sum_{l=1}^4 \sum_{\substack{F \in \mathcal{F}_h^i \\ F \cap E_l \neq \emptyset}} \int_{F \cap E_l} \{-\partial_n v_h\}(\tau_l \cdot n)^2.$$

Proof. Consider first triangle E_1 . Upon integrating by parts two times and using Lemma 4.2 and the fact that $v_h \in C^0(\Omega)$, we infer the following:

$$\begin{aligned} 0 &= \int_{E_1} v_h \partial_{\tau_1 \tau_1}(1) dx = \sum_{\substack{K \in \mathcal{T}_h \\ K \cap E_1 \neq \emptyset}} \int_{K \cap E_1} v_h \partial_{\tau_1 \tau_1}(1) dx = \sum_{\substack{K \in \mathcal{T}_h \\ K \cap E_1 \neq \emptyset}} - \int_{K \cap E_1} \partial_{\tau_1} v_h \partial_{\tau_1}(1) dx \\ &= \sum_{\substack{K \in \mathcal{T}_h \\ K \cap E_1 \neq \emptyset}} \int_{K \cap E_1} \partial_{\tau_1 \tau_1} v_h dx - \int_{\partial(K \cap E_1)} (\tau_1 \cdot n) \partial_{\tau_1} v_h ds \\ &= \sum_{\substack{K \in \mathcal{T}_h \\ K \cap E_1 \neq \emptyset}} \int_{K \cap E_1} \partial_{\tau_1 \tau_1} v_h dx - \sum_{\substack{F \in \mathcal{F}_h^i \\ F \cap E_1 \neq \emptyset}} \int_{F \cap E_1} 2\{\partial_n v_h\}(\tau_1 \cdot n)^2 ds \\ &\quad + \frac{1}{2} \int_{\Gamma_1} \partial_1 v_h ds - \frac{1}{2} \int_{\Gamma_2} \partial_2 v_h ds - \int_{\Gamma_1} \partial_n v_h (\tau_1 \cdot n)^2 ds - \int_{\Gamma_2} \partial_n v_h (\tau_1 \cdot n)^2 ds. \end{aligned}$$

By proceeding similarly with the other triangles $E_2, E_3,$ and $E_4,$ and adding the four results, we obtain

$$\begin{aligned} & - \int_{\Gamma_1} \partial_1 v_h ds + \int_{\Gamma_2} \partial_2 v_h ds + \int_{\Gamma_3} \partial_1 v_h ds - \int_{\Gamma_4} \partial_2 v_h ds \\ &= \sum_{l=1}^4 \sum_{\substack{K \in \mathcal{T}_h \\ K \cap E_l \neq \emptyset}} \int_{K \cap E_l} \partial_{\tau_l \tau_l} v_h dx + 2 \sum_{l=1}^4 \sum_{\substack{F \in \mathcal{F}_h^i \\ F \cap E_l \neq \emptyset}} \int_{F \cap E_l} \{-\partial_n v_h\}(\tau_l \cdot n)^2. \end{aligned}$$

We conclude by observing that

$$\Delta_\delta v_h(x) = - \int_{\Gamma_1} \partial_1 v_h ds + \int_{\Gamma_2} \partial_2 v_h ds + \int_{\Gamma_3} \partial_1 v_h ds - \int_{\Gamma_4} \partial_2 v_h ds.$$

This concludes the proof. \square

We are now in a position to prove a one-sided bound similar to that in (5.5).

LEMMA 5.2. *For all sequences of almost minimizers for (2.8), say $\{u_h\}_{h>0},$ there exist $c > 0$ and $\gamma := \min(\frac{p_1-2}{p_1}, \frac{p_2-2}{p_2})$ such that for all $x \in \Omega$ and $\delta > h$ such that $B_{\mathbb{R}^2}(x, \delta) \subset \Omega,$ the following one-sided bound holds:*

$$(5.9) \quad \Delta_\delta u_h(x) \leq c\delta^{1+\gamma}.$$

Proof. Let us set $E := E_1 \cup E_2 \cup E_3 \cup E_4$. Using Lemma 5.1 together with the estimate (3.11), we infer

$$\begin{aligned} \Delta_\delta u_h(x) &\leq \sum_{K \in \mathcal{T}_h \cap E} \int_K \lambda_+(u_h) dx + 2 \sum_{F \in \mathcal{F}_h^i \cap E} \int_F \{-\partial_n u_h\}_+ ds \\ &\leq \left(\sum_{K \in \mathcal{T}_h \cap E} \text{meas } K \right)^{\frac{1}{p_1}} \left(\sum_{K \in \mathcal{T}_h} \int_K \lambda_+(u_h)^{p_1} dx \right)^{\frac{1}{p_1}} \\ &\quad + 2 \left(h^{\frac{p_2-1}{p_2} p_2'} \sum_{F \in \mathcal{F}_h^i \cap E} \text{meas } F \right)^{\frac{1}{p_2}} \left(h^{1-p_2} \sum_{F \in \mathcal{F}_h^i} \int_F \{-\partial_n u_h\}_+^{p_2} ds \right)^{\frac{1}{p_2}} \\ &\leq c \left(\delta^{\frac{2}{p_1}} + \delta^{\frac{2}{p_2}} \right), \end{aligned}$$

where we used $h \leq \delta$ in the last inequality. (We have bounded from above the number of cells in $\mathcal{T}_h \cap E$ and the number of interfaces in $\mathcal{F}_h^i \cap E$ by δ/h and this number cannot be less than 1.) We conclude by observing that $\frac{2}{p_i} = 1 + \frac{p_i-2}{p_i}$ and $p_i > 2$ for $i = 1, 2$. \square

THEOREM 5.3. *Let $u \in X$ be the unique solution to (1.1). Under the mesh assumption (3.3), the uniqueness assumption (5.5), and the restriction $p_1 > 2, p_2 > 2$, every sequence of almost minimizers converges to the unique viscosity solution to (1.1).*

Proof. Let $\{u_h\}_{h>0}$ be a sequence of almost minimizers. Let $\delta > 0$ and let $\Omega_\delta = \{x \in \Omega; B_{\mathbb{R}^2}(x, \delta) \subset \Omega\}$. Then owing to Lemma 5.9 the following holds for every $x \in \Omega_\delta$:

$$\Delta_\delta u_h(x) \leq c\delta^{1+\gamma}.$$

Since u_h converges strongly to u in L^1 , we infer that $u_h \rightarrow u$ a.e. in Ω_δ , that is,

$$\Delta_\delta u(x) \leq c\delta^{1+\gamma}, \quad \text{a.e. } x \text{ in } \Omega_\delta.$$

We then conclude that the above inequality holds for every $x \in \Omega_\delta$ since u is continuous. \square

Remark 5.1. Recall that whether hypothesis (3.3) holds for every quasi-uniform mesh family is an open question. It definitely holds on aligned meshes. We remove this assumption in the next section for uniform meshes by taking $p_2 = 1$ and adding an extra term in the entropy.

6. One-sided bound on uniform meshes. The goal of this section is to prove an analogue of Theorem 5.3 in the case $p_2 = 1$, i.e., the mesh assumption (3.3) is empty. For this purpose we assume that the mesh is uniform and add a vertex-centered entropy to the functional J_h . We prove the one-sided bound (5.5) using the orthonormal basis $\frac{1}{\sqrt{2}}(e_1 + e_2, e_1 - e_2)$.

6.1. The vertex-centered entropy. We henceforth assume that the mesh is uniform in the sense that the set of vertices is

$$(6.1) \quad \Omega_h := \{(sh, kh) \in \mathbb{R}^2; (s, k) \in I_h\} \subset \overline{\Omega},$$

where $\{I_h\}_{h>0}$ is a family of subsets of \mathbb{N}^2 . The set of interior vertices is denoted by $\dot{\Omega}_h$. The mesh cells are triangles whose edges are parallel to e_1, e_2 , or $e_1 + e_2$.

In order to understand how a vertex-centered entropy can be constructed, let us consider a point $x := (ih, jh) \in \Omega_h$ and $\delta > 0$ such that $x + B_h(0, \sqrt{2}\delta) \subset \Omega_h$, where $B_h(0, \mu) := \{(sh, kh) = z; (k, l) \in \mathbb{N}^2; \|z\| < \mu\}$. Assume for the time being that $\delta = nh$ with $n \geq 1$. We now define the following index sets

$$\begin{aligned}
 \Lambda_0 &= \{(s, k); |k - j| \leq n \text{ and } |s - i| \leq n\}, \\
 \Lambda_1 &= \{(s, k); 1 \leq k - j \leq n \text{ and } |s - i| \leq k - j - 1\}, \\
 \Lambda_2 &= \{(s, k); 1 \leq s - i \leq n \text{ and } |k - j| \leq s - i - 1\}, \\
 \Lambda_3 &= \{(s, k); 1 \leq j - k \leq n \text{ and } |s - i| \leq j - k - 1\}, \\
 \Lambda_4 &= \{(s, k); 1 \leq i - s \leq n \text{ and } |k - j| \leq i - s - 1\}, \\
 \Lambda_5 &= \{(s, k); 0 < |s - i| = |k - j| < n\}.
 \end{aligned}
 \tag{6.2}$$

Up to a $\pi/4$ rotation and an appropriate rescaling, the sets $\Lambda_m, m \in \{1, 2, 3, 4\}$ correspond to the triangles E_m defined in (5.6). The set Λ_0 corresponds to the square $E_1 \cup E_2 \cup E_3 \cup E_4$. The set Λ_5 corresponds to $\Gamma_1 \cup \Gamma_2 \cup \Gamma_3 \cup \Gamma_4$ minus the center and the four corners of the square (the Γ_m 's have been defined in (5.7)); see also Figure 3.

Let v be a member of X_h . To simplify notation we set $v_{s,k} := v(sh, kh)$. Now our goal is to find a discrete analogue of the integral representation (5.8) of the discrete Laplacian

$$\Delta_{\sqrt{2}\delta} u(x) = u_{i-n,j-n} + u_{i-n,j+n} + u_{i+n,j-n} + u_{i+n,j+n} - 4u_{i,j}.
 \tag{6.3}$$

Let $z = (sh, kh) \in \dot{\Omega}_h$ be an interior vertex. We introduce the following additional notation for the second-order directional finite differences at z :

$$D_1^2 v_{s,k} = u_{s-1,k} - 2u_{s,k} + u_{s+1,k}, \quad D_2^2 v_{s,k} = u_{s,k-1} - 2u_{s,k} + u_{s,k+1}.
 \tag{6.4}$$

Then the following discrete representation of $\Delta_{\sqrt{2}\delta} v(x)$ holds:

$$\begin{aligned}
 \Delta_{\sqrt{2}\delta} v(x) &= R_1(v, x, \delta) + D_1^2 v_{i,j} + D_2^2 v_{i,j} + \sum_{(s,k) \in \Lambda_1 \cup \Lambda_3} D_1^2 v_{s,k} + \sum_{(s,k) \in \Lambda_2 \cup \Lambda_4} D_2^2 v_{s,k} \\
 &+ \sum_{(s,k) \in \Lambda_5} \left(\frac{1}{2} D_1^2 v_{s,k} + \frac{1}{2} D_2^2 v_{s,k} \right),
 \end{aligned}
 \tag{6.5}$$

where the remainder $R_1(v, x, \delta)$ is defined by

$$\begin{aligned}
 2R_1(v, x, \delta) &= (v_{i-n+1,j+n} - v_{i-n,j+n}) + (v_{i-n,j+n-1} - v_{i-n,j+n}) \\
 &+ (v_{i+n-1,j+n} - v_{i+n,j+n}) + (v_{i+n,j+n-1} - v_{i+n,j+n}) \\
 &+ (v_{i+n-1,j-n} - v_{i+n,j-n}) + (v_{i+n,j-n+1} - v_{i+n,j-n}) \\
 &+ (v_{i-n+1,j-n} - v_{i-n,j-n}) + (v_{i-n,j-n+1} - v_{i-n,j-n}).
 \end{aligned}
 \tag{6.6}$$

We now define a vertex-centered entropy as follows:

$$E_h(v) := \sum_{(ih,jh) \in \dot{\Omega}_h} (D_1^2 v_{i,j})_+^{p_3} + (D_2^2 v_{i,j})_+^{p_3},
 \tag{6.7}$$

where

$$(6.8) \quad p_3 > 2.$$

We modify the functional J_h by setting $p_1 = p_2 = 1$ and by adding the vertex-centered entropy

$$(6.9) \quad J_h(v) = \int_{\Omega} |H(x, v, Dv)| dx + h \sum_{K \in \mathcal{T}_h} \int_K \lambda_+(v) dx + h \sum_{F \in \mathcal{F}_h^i} \int_F \{-\partial_n v\}_+ d\sigma + h^{3-2p_3} E_h(v),$$

and we henceforth denote by $\{u_h\}_{h>0}$ a sequence of almost minimizers for (2.8) using the above modified functional.

6.2. Consistency. The goal of this section is to show that, with the above choice of entropy, it is possible to construct an approximation of u , say $\mathcal{I}_h u$, that satisfies the estimate

$$(6.10) \quad J_h(\mathcal{I}_h u) \leq ch.$$

We make use of the Clément interpolation operator [7, 6] to this purpose. Let v be an arbitrary function in $L^1(\Omega)$. We define $\mathcal{I}_h(v)$ to be a piecewise linear function on the mesh \mathcal{T}_h as follows. Let a be any vertex of \mathcal{T}_h . If a is on Γ , we set $\mathcal{I}_h(v)(a) = 0$. If a is an interior vertex, we define Δ_a to be the set of all those triangles that have a as a vertex. We define $r(v) \in \mathbb{P}_1$ to the linear polynomial such that

$$(6.11) \quad \int_{\Delta_a} (r(v)(x) - v(x))q(x) dx = 0 \quad \forall q \in \mathbb{P}_1.$$

Then we set $\mathcal{I}_h(v)(a) = r(v)(a)$. The interpolant \mathcal{I}_h thus defined is $W^{1,\infty}$ -stable and there is $c > 0$ such that for all $m \in \{0, 1\}$, $k \in \{0, 1\}$, any number $p \geq 1$, and all $v \in W^{k+1,p}(\Omega) \cap W_0^{1,p}(\Omega)$ the following holds (see [6, 7]):

$$(6.12) \quad \|\mathcal{I}_h(v) - v\|_{W^{m,p}(\Omega)} \leq ch^{k+1-m} \|v\|_{W^{k+1,p}(\Omega)},$$

$$(6.13) \quad \left(\sum_{F \in \mathcal{F}_h^i} \|\mathcal{I}_h v - v\|_{W^{m,p}(F)}^p \right)^{1/p} \leq ch^{k+1-m-\frac{1}{p}} \|v\|_{W^{k+1,p}(\Omega)}.$$

LEMMA 6.1. *Under the above hypotheses and with the definition (6.9) of the functional J_h , there is c , uniform in h , such that the linear Clément interpolant of u , say $\mathcal{I}_h(u)$, satisfies*

$$(6.14) \quad J_h(\mathcal{I}_h(u)) \leq ch.$$

Proof. The $W^{1,\infty}$ -stability and the approximation property (6.12) with $m = p = k = 1$ of the Clément interpolant together with the assumption (1.3) yields $\|H(\cdot, \mathcal{I}_h(u), D\mathcal{I}_h(u))\|_{L^1(\Omega)} \leq ch$. Since $\mathcal{I}_h(u)$ is piecewise linear, the volume entropy in J_h involving $\lambda_+(\mathcal{I}_h(u))$ is zero. The q -semiconcavity of u implies $\{-\partial_n \mathcal{I}_h u\}_+ = \{-\partial_n(\mathcal{I}_h u - u)\}_+$ across every $F \in \mathcal{F}_h^i$, since the $W^{2,q}$ -component of u has a continuous gradient by embedding (recall that $q > 2$). This together with (6.13) (using $k = 1$,

$m = 0, p = 1$) yields that the interface entropy involving $\{-\partial_n \mathcal{I}_h u\}_+$ is bounded from above by ch .

Now we have to make sure that the vertex-centered entropy is appropriately controlled. Let $a \in \hat{\Omega}_h$ be an interior mesh vertex. Let us evaluate $r(u)$ at a as defined in (6.11). To do so, we expand $r(u)$ in the following manner:

$$r(u)(x) = \alpha + \beta \cdot (x - a),$$

where $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^2$ are yet to be determined. Since the mesh is structured, the following holds:

$$\int_{\Delta_a} (x - a) dx = 0.$$

This immediately implies

$$0 = \int_{\Delta_a} (r(u) - u) dx = |\Delta_a| \left(\alpha - \Delta_a^{-1} \int_{\Delta_a} u dx \right),$$

i.e., $r(u)(a) = \alpha = \frac{1}{|\Delta_a|} \int_{\Delta_a} u dx$.

Let i be an index in $\{1, 2\}$. Let us set $\Delta_a^+ = \Delta_a + he_i$ and $\Delta_a^- = \Delta_a - he_i$. Then, with obvious notation,

$$\begin{aligned} |\Delta_a| D_i^2 r(u)(a) &= \int_{\Delta_a^+} u(x) dx - 2 \int_{\Delta_a} u(x) dx + \int_{\Delta_a^-} u(x) dx \\ &= \int_{\Delta_a} (u(x + he_i) - 2u(x) + u(x - he_i)) dx \\ &= \int_{\Delta_a} D_i^2 u(x) dx. \end{aligned}$$

Now, using the q -semiconcavity hypothesis (1.5), we have $u = v_c + w$, where $v_c \in W^{1,\infty}(\Omega)$ is concave and $w \in W^{2,q}(\Omega)$. Then, using the concavity of v_c and the $W^{2,q}$ -regularity of w , infer

$$E(\mathcal{I}_h(u)) = E(\mathcal{I}_h(w)) \leq ch^{2p_3-2}.$$

This implies $h^{3-2p_3} E(\mathcal{I}_h(u)) \leq ch$. This completes the proof. \square

Lemma 6.1 means that the set of almost minimizers is not empty when using definition (6.9). No extra assumptions need to be made.

Remark 6.1. Note that the fact that the mesh is structured is a key argument in the proof of Lemma 6.1.

6.3. Convergence to the viscosity solution. Let $\delta \geq h$ be a real number that we assume for the time being to be a multiple of h , i.e., $\delta = nh$ with $n \geq 1$. Consider a point $x := (ih, jh) \in \Omega_h$ such that $x + B_h(0, \sqrt{2}\delta) \subset \Omega_h$.

LEMMA 6.2. *Under the above hypotheses, for all sequences of almost minimizers of (2.8), say $\{u_h\}_{h>0}$, there is c , independent of x, h , and δ , and there is $\gamma := 1 - \frac{2}{p_3} > 0$ so that*

$$(6.15) \quad \Delta_{\sqrt{2}\delta} u_h(x) \leq c\delta^{1+\gamma} + R_1(u_h, x, \delta).$$

Proof. From (6.5) we infer

$$\Delta_{\sqrt{2}\delta}u_h(x) \leq R_1(u_h, x, \delta) + \frac{5}{2} \sum_{(s,k) \in \Lambda_0} (D_1^2(u_h)_{s,k})_+ + \frac{5}{2} \sum_{(s,k) \in \Lambda_0} (D_2^2(u_h)_{s,k})_+.$$

Using Hölder’s inequality, this implies

$$\Delta_{\sqrt{2}\delta}u_h(x) \leq R_1(u_h, x, \delta) + \text{card}(\Lambda_0)^{\frac{1}{p_3}} (E_h(u_h))^{\frac{1}{p_3}},$$

where $\text{card}(\Lambda_0)$ is the cardinal number of Λ_0 . Clearly $\text{card}(\Lambda_0) \leq c(\delta/h)^2$. Moreover, since $\{u_h\}_{h>0}$ is a sequence of almost minimizers, it comes that $E_h(u_h) \leq ch^{2(p_3-1)}$. That is to say,

$$\Delta_{\sqrt{2}\delta}u_h(x) \leq R_1(u_h, x, \delta) + c\delta^{2/p_3}h^{-2/p_3}h^{2(p_3-1)/p_3} \leq R_1(u_h, x, \delta) + c\delta^{1+\gamma},$$

where $\gamma = 1 - 2/p_3 > 0$ since $p_3 > 2$. □

We now conclude the following theorem.

THEOREM 6.3. *Let $u \in X$ be the unique solution to (1.1). Consider the uniform mesh family defined by (6.1). Under the uniqueness assumption (5.5), and the restriction $p_3 > 2$, every sequence of almost minimizers for the functional (6.9) converges to the unique viscosity solution to (1.1).*

Proof. Let $\{u_h\}_{h>0}$ be a sequence of almost minimizers. Let x be a point in Ω . There exists δ_0 such that $B_{\mathbb{R}^2}(x, \delta_0) \subset \Omega$. Let δ be a fixed number in $(0, \delta_0/2\sqrt{2}]$ and let h be an arbitrary mesh size such that $h \leq \delta$.

Since the ratio δ/h may not be an integer and/or x is almost surely not a member of Ω_h , we define $n := \lfloor \delta/h \rfloor$, $i := \lfloor (x \cdot e_1)/h \rfloor$, and $j := \lfloor (x \cdot e_2)/h \rfloor$, where $\lfloor \cdot \rfloor$ is the floor function. Then we set $\bar{\delta} = nh$ and $\bar{x} = (ih, jh)$. Note that with our choice of parameters, \bar{z} is in Ω_h , where \bar{z} is either \bar{x} or $\bar{x} \pm \bar{\delta}e_1 \pm \bar{\delta}e_2$. These definitions imply

$$(6.16) \quad \Delta_{\sqrt{2}\delta}u_h(x) = \Delta_{\sqrt{2}\bar{\delta}}u_h(\bar{x}) + R_2(u_h, x, \delta),$$

where the remainder is defined by $R_2(u_h, x, \delta) := \Delta_{\sqrt{2}\delta}u_h(x) - \Delta_{\sqrt{2}\bar{\delta}}u_h(\bar{x})$.

Now we use the one-sided bound (6.15) from Lemma 6.2 to obtain

$$\Delta_{\sqrt{2}\delta}u_h(x) \leq c\bar{\delta}^{1+\gamma} + R_1(u_h, \bar{x}, \bar{\delta}) + R_2(u_h, x, \delta).$$

We conclude by passing to the limit on h . Clearly $\bar{\delta} \rightarrow \delta$. Since $u_h \rightarrow u$ a.e. in Ω , we infer $\Delta_{\sqrt{2}\delta}u_h(x) \rightarrow \Delta_{\sqrt{2}\delta}u(x)$ for a.e. x in Ω . Moreover, owing to Lemma 6.4, $R_1(u_h, \bar{x}, \bar{\delta}) \rightarrow 0$ and $R_2(u_h, x, \delta) \rightarrow 0$ for a.e. x in Ω . As a result,

$$\Delta_{\sqrt{2}\delta}u(x) \leq c\delta^{1+\gamma} \text{ for a.e. } x \in \Omega,$$

and the constant c does not depend on x . Since u is continuous, this implies that the inequality holds for every x in Ω and every δ such that $B_{\mathbb{R}^2}(x, \delta_0) \subset \Omega$. □

Remark 6.2. Recall that the class of stationary Hamilton–Jacobi equations defined by $H(x, u, Du) = u + F(Du)$, where $F : \mathbb{R}^2 \rightarrow \mathbb{R}$ is convex, has a unique viscosity solution characterized by (5.5); see [18, Thm. 2.6]. Moreover, it is known that the solution is ∞ -semiconcave under appropriate restrictions on the domain. In other words, Theorem 6.3 holds at least for the above class of Hamilton–Jacobi equations.

LEMMA 6.4. *Under the above hypotheses, $R_1(u_h, \bar{x}(x), \bar{\delta}) \rightarrow 0$ for a.e. $x \in \Omega$ and $R_2(u_h, x, \delta) \rightarrow 0$ for a.e. $x \in \Omega$.*

Proof. R_1 is composed of eight terms which have the generic form

$$r_h(\bar{x}, \bar{\delta}, h) := u_h(\bar{x} + s_1 \bar{\delta} e_{\pm} + s_2 h e_i) - u_h(\bar{x} + s_1 \bar{\delta} e_{\pm}),$$

where $e_{\pm} = e_1 \pm e_2$, $i \in \{1, 2\}$ and $s_1, s_2 \in \{-1, +1\}$. To avoid boundary issues, we extend r_h to \mathbb{R}^2 by replacing u_h by \tilde{u}_h ; the extension in question is denoted by \tilde{r}_h . We now evaluate the L^1 -norm of $\tilde{r}_h(\bar{x}(x), \bar{\delta}, h)$ as follows:

$$\begin{aligned} \|\tilde{r}_h(\bar{x}(\cdot), \bar{\delta}, h)\|_{L^1(\mathbb{R}^2)} &= \int_{\mathbb{R}^2} |\tilde{u}_h(\bar{x}(x) + s_1 \bar{\delta} e_{\pm} + s_2 h e_i) - \tilde{u}_h(\bar{x}(x) + s_1 \bar{\delta} e_{\pm})| dx \\ &= \int_{\mathbb{R}^2} |\tilde{u}_h(\bar{x}(x) + s_2 h e_i) - \tilde{u}_h(\bar{x}(x))| dx \\ &= h^2 \sum_{k,l=-\infty}^{+\infty} |\tilde{u}_h(x_{k,l} + s_2 h e_i) - \tilde{u}_h(x_{k,l})| \\ &= h^2 \sum_{k,l=-\infty}^{+\infty} \left| \int_{F_{k,l}} \partial_{e_i} \tilde{u}_h(y) dy \right|, \end{aligned}$$

where we have denoted $x_{k,l} = (kh, lh)$ and $F_{k,l}$ is the segment $(x_{k,l}, x_{k,l} + s_2 h e_i)$. Note that $F_{k,l}$ is equal to $(x_{k,l}, x_{(k+s_2),l})$ if $i = 1$ and $(x_{k,l}, x_{k,(l+s_2)})$ if $i = 2$. Let us denote by $\Delta_{x_{k,l}}$ the set of all those triangles that have $x_{k,l}$ as a vertex (there are six of those). Then a trace and an inverse inequality yield

$$\left| \int_{F_{k,l}} \partial_{e_i} \tilde{u}_h(y) dy \right| \leq ch^{-1} \sum_{K \in \Delta_{x_{k,l}}} \int_K \|D\tilde{u}_h\|_{L^1(K)}.$$

This then yields

$$\begin{aligned} \|\tilde{r}_h(\bar{x}(\cdot), \bar{\delta}, h)\|_{L^1(\mathbb{R}^2)} &\leq ch \sum_{k,l=-\infty}^{+\infty} \sum_{K \in \Delta_{x_{k,l}}} \int_K \|D\tilde{u}_h\|_{L^1(K)} \\ &\leq ch \|\tilde{u}_h\|_{W^{1,1}(\mathbb{R}^2)} \leq c' h \|u_h\|_{W^{1,1}(\Omega)}. \end{aligned}$$

This means $\tilde{r}_h(\bar{x}(\cdot), \bar{\delta}, h) \rightarrow 0$ in $L^1(\mathbb{R}^2)$, which immediately implies $r_h(\bar{x}(x), \bar{\delta}, h) \rightarrow 0$ for a.e. x in Ω .

For R_2 , we observe that R_2 is composed of five terms which have the generic form

$$r_h(x, \delta, h) := u_h(x + s\delta e_{\pm}) - u_h(\bar{x}(x) + s\bar{\delta} e_{\pm}),$$

where $s \in \{-1, 0, +1\}$. To avoid boundary issues, we again extend r_h to \mathbb{R}^2 by replacing u_h by \tilde{u}_h ; the extension in question is denoted by \tilde{r}_h . We now evaluate the

L^1 -norm of $\tilde{r}_h(x, \delta, h)$ as follows:

$$\begin{aligned} \|\tilde{r}_h(\cdot, \bar{\delta}, h)\|_{L^1(\mathbb{R}^2)} &= \int_{\mathbb{R}^2} |\tilde{u}_h(x + s\delta e_{\pm}) - \tilde{u}_h(\bar{x}(x) + s\bar{\delta}e_{\pm})| dx \\ &\leq \int_{\mathbb{R}^2} |\tilde{u}_h(x + s\delta e_{\pm}) - \tilde{u}_h(x + s\bar{\delta}e_{\pm})| dx \\ &\quad + \int_{\mathbb{R}^2} |\tilde{u}_h(x + s\bar{\delta}e_{\pm}) - \tilde{u}_h(\bar{x}(x) + s\bar{\delta}e_{\pm})| dx \\ &\leq \int_{\mathbb{R}^2} |\tilde{u}_h(x) - \tilde{u}_h(x + s(\bar{\delta} - \delta)e_{\pm})| dx \\ &\quad + \int_{\mathbb{R}^2} |\tilde{u}_h(x) - \tilde{u}_h(\bar{x}(x))| dx. \end{aligned}$$

Let r_1, r_2 be the two integrals in the right-hand side, respectively. For r_1 we have

$$\begin{aligned} r_1 &= \int_{\mathbb{R}^2} \left| \int_0^1 D\tilde{u}_h(x + \theta s(\bar{\delta} - \delta)e_{\pm}) \cdot (\bar{\delta} - \delta)e_{\pm} d\theta \right| dx \\ &\leq |\bar{\delta} - \delta| \int_0^1 \int_{\mathbb{R}^2} \|D\tilde{u}_h(x + \theta s(\bar{\delta} - \delta)e_{\pm})\| dx d\theta \leq h \|D\tilde{u}_h\|_{L^1(\mathbb{R}^2)} \\ &\leq ch \|u_h\|_{W^{1,1}(\Omega)}. \end{aligned}$$

For the second residual we have

$$r_2 = \sum_{k,l=-\infty}^{+\infty} \int_{S_{k,l}} |\tilde{u}_h(x) - \tilde{u}_h(x_{k,l})| dx,$$

where $x_{k,l} = (kh, lh)$ and $S_{k,l}$ is the square $(x_{k,l}, x_{k+1,l}) \times (x_{k,l}, x_{k,l+1})$. Then a trace inequality and an inverse inequality yields

$$\begin{aligned} r_2 &= \sum_{k,l=-\infty}^{+\infty} \int_{S_{k,l}} \left| \int_0^1 D\tilde{u}_h(x + \theta(x_{k,l} - x)) \cdot (x_{k,l} - x) d\theta \right| dx \\ &\leq ch^{-1} \sum_{k,l=-\infty}^{+\infty} \int_{S_{k,l}} \|D\tilde{u}_h\|_{L^1(S_{k,l})} \leq ch \sum_{k,l=-\infty}^{+\infty} \|D\tilde{u}_h\|_{L^1(S_{k,l})} \\ &\leq ch \|u_h\|_{W^{1,1}(\Omega)}. \end{aligned}$$

We then conclude as above. \square

7. Numerical experiments. A one-dimensional theory for the L^1 -approximation of stationary Hamilton–Jacobi equations is developed in [14], and efficient numerical algorithms are proposed and analyzed in [12].

The purpose of this section is to support our theory by reporting two-dimensional numerical experiments. Our goal is not to analyze or discuss the optimality of any given numerical strategy to solve (2.8) but to show that L^1 -minimizers are computable

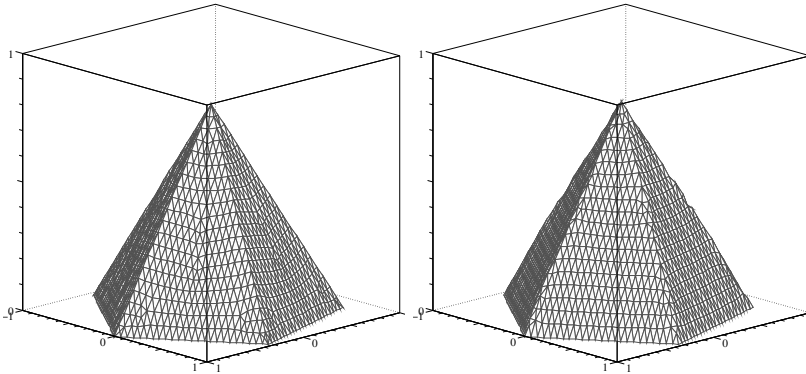


FIG. 4. *Pentagon: Aligned unstructured mesh (left); nonaligned unstructured mesh (right).*

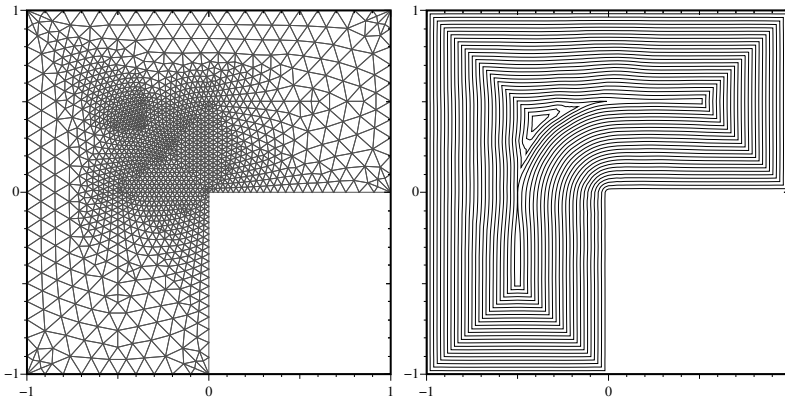


FIG. 5. *L-shaped domain: Unstructured mesh (left); iso-lines of approximate minimizer (right).*

and are very accurate nonoscillatory approximations to viscosity solutions of stationary two-dimensional Hamilton–Jacobi equations. We henceforth focus our attention on the eikonal equation, $\|Du\| = 1$, equipped with homogeneous Dirichlet boundary conditions. The computations are done using piecewise linear continuous finite elements. The entropy is defined using $p = 2$. The discrete problem (2.8) is solved by using an iterative regularization method described in [11]. In a few words, the algorithm consists of computing $\lim_{h \rightarrow 0} \lim_{\epsilon \rightarrow 0} \operatorname{argmin}_{v_h \in X_h} J_{h,\epsilon}(v_h)$. The functional $J_{h,\epsilon}(v_h)$ is a regularized version of $J_h(v_h)$, where the absolute value defining the L^1 -norm and the $(\cdot)_+$ function are replaced by $x \mapsto x^2/(|x| + \epsilon)$. The minimization problem is solved by using a Newton method. The number ϵ is used as a continuation parameter. The computation stops when $\epsilon = 1.10^{-5}$. The mesh size h is also used as a continuation parameter in the sense that the computation is done on three grids successively refined. The result on a coarse grid is used to initialize the solution on the next grid.

In the first example the domain Ω is a pentagon. The computation is done on two types of meshes. The first type is composed of meshes that are aligned with the discontinuities of the gradient and the second type consists of unstructured meshes. Typical results are reported in Figure 4. For both mesh types, we observe that the approximate L^1 -minimizer is similar to the Lagrange interpolant of the exact solution

on the same mesh. This is what we should expect intuitively. The L^1 -minimization process solves the equation in the region where the solution is smooth and simply ignores the PDE in the regions where the gradient of the exact solution is discontinuous. For more details we refer to [13, 14].

The second example is the eikonal equation on an L -shaped domain. The viscosity solution to this problem is in $W^{1,\infty}(\overline{\Omega})$ and is q semiconcave for every $q < 2$. This is a borderline case not covered by our theory (we a priori need $q > 2$). We nevertheless do the computations using $p = 2$ for the entropy. We show a mesh and the corresponding approximate minimizer in Figure 5. Once again, we observe that the solution is accurate. The iso-lines are not oscillating and are very sharp.

REFERENCES

- [1] R. ABGRALL, *Numerical discretization of the first-order Hamilton–Jacobi equation on triangular meshes*, Comm. Pure Appl. Math., 49 (1996), pp. 1339–1373.
- [2] R. ABGRALL, *Numerical discretization of boundary conditions for first order Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 41 (2003), pp. 2233–2261.
- [3] G. ALBERTI AND L. AMBROSIO, *A geometrical approach to monotone functions in \mathbf{R}^n* , Math. Z., 230 (1999), pp. 259–316.
- [4] G. BARLES, *Solutions de viscosité des équations de Hamilton–Jacobi*, Math. Appl. (Berlin) 17, Springer-Verlag, Paris, 1994.
- [5] G. BARLES AND P. E. SOUGANIDIS, *Convergence of approximation schemes for fully nonlinear second order equations*, Asymptot. Anal., 4 (1991), pp. 271–283.
- [6] C. BERNARDI AND V. GIRAULT, *A local regularization operator for triangular and quadrilateral finite elements*, SIAM J. Numer. Anal., 35 (1998), pp. 1893–1916.
- [7] P. CLÉMENT, *Approximation by finite element functions using local regularization*, RAIRO Anal. Numer., 9 (1975), pp. 77–84.
- [8] M. G. CRANDALL AND P.-L. LIONS, *Viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.
- [9] M. G. CRANDALL AND P.-L. LIONS, *Two approximations of solutions of Hamilton–Jacobi equations*, Math. Comp., 43 (1984), pp. 1–19.
- [10] L. C. EVANS, *Partial Differential Equations*, Grad. Stud. Math. 19, American Mathematical Society, Providence, RI, 1998.
- [11] J. L. GUERMOND, *A finite element technique for solving first-order PDEs in L^p* , SIAM J. Numer. Anal., 42 (2004), pp. 714–737.
- [12] J.-L. GUERMOND, F. MARPEAU, AND B. POPOV, *A fast algorithm for solving first-order PDEs by L^1 -minimization*, Commun. Math. Sci., 6 (2008), pp. 199–216.
- [13] J.-L. GUERMOND AND B. POPOV, *Linear advection with ill-posed boundary conditions via L^1 -minimization*, Int. J. Numer. Anal. Model., 4 (2007), pp. 39–47.
- [14] J.-L. GUERMOND AND B. POPOV, *L^1 -minimization methods for Hamilton–Jacobi equations: The one-dimensional case*, Numer. Math., 109 (2008), pp. 269–284.
- [15] C.-Y. KAO, S. OSHER, AND Y.-H. TSAI, *Fast sweeping methods for static Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 42 (2005), pp. 2612–2632.
- [16] S. N. KRUKOV, *Generalized solutions of Hamilton–Jacobi equations of eikonal type. I. Statement of the problems; existence, uniqueness and stability theorems; certain properties of the solutions*, Mat. Sb., 98 (1975), pp. 450–493, 496.
- [17] C.-T. LIN AND E. TADMOR, *L^1 -stability and error estimates for approximate Hamilton–Jacobi solutions*, Numer. Math., 87 (2001), pp. 701–735.
- [18] P.-L. LIONS AND P. E. SOUGANIDIS, *Convergence of MUSCL and filtered schemes for scalar conservation laws and Hamilton–Jacobi equations*, Numer. Math., 69 (1995), pp. 441–470.
- [19] J. A. SETHIAN, *Fast marching methods*, SIAM Rev., 41 (1999), pp. 199–235.

FAST MARCHING METHODS FOR STATIONARY HAMILTON–JACOBI EQUATIONS WITH AXIS-ALIGNED ANISOTROPY*

KEN ALTON[†] AND IAN M. MITCHELL[†]

Abstract. The fast marching method (FMM) has proved to be a very efficient algorithm for solving the isotropic Eikonal equation. Because it is a minor modification of Dijkstra’s algorithm for finding the shortest path through a discrete graph, FMM is also easy to implement. In this paper we describe a new class of Hamilton–Jacobi (HJ) PDEs with axis-aligned anisotropy which satisfy a causality condition for standard finite-difference schemes on orthogonal grids and can hence be solved using the FMM; the only modification required to the algorithm is in the local update equation for a node. This class of HJ PDEs has applications in anelliptic wave propagation and robotic path planning, and brief examples are included. Since our class of HJ PDEs and grids permit asymmetries, we also examine some methods of improving the efficiency of the local update that do not require symmetric grids and PDEs. Finally, we include explicit update formulas for variations of the Eikonal equation that use the Manhattan, Euclidean, and infinity norms on orthogonal grids of arbitrary dimension and with variable node spacing.

Key words. fast marching method, anisotropic optimal control, Hamilton–Jacobi equation, viscosity solution

AMS subject classifications. 35F30, 49L20, 49L25, 49N90, 65N06, 65N12

DOI. 10.1137/070680357

1. Introduction. The fast marching method (FMM) [29, 23] has become a popular algorithm to use when solving the Dirichlet problem for an isotropic static Hamilton–Jacobi partial differential equation (HJ PDE), also known as the Eikonal equation $\|Du(x)\|_2 = c(x)$. FMM has proven to be particularly efficient in practice because it can approximately solve this problem in a single pass through the nodes of a grid. It is also straightforward to implement, requiring only a small modification of Dijkstra’s algorithm [9], which is a popular method for finding the shortest path through a graph.

While the isotropic case is the most common, there are applications which require the solution of anisotropic HJ PDEs. Unfortunately, FMM produces a correct approximation only under certain causality conditions on the values of nodes and their neighbors. This limitation has motivated the development of a more generally applicable version of FMM called ordered upwind methods (OUMs) [21] and also several recent works such as [31, 13, 19] on sweeping methods. However, OUMs are much more complex to implement than FMM, and sweeping methods can be much less efficient for problems with curved characteristics and practical grid sizes [12, 11].

Consequently, we have motivation to seek classes of anisotropic problems to which FMM might still be applied. One such class of problems was identified in [20] and includes the Eikonal equation where an energy norm replaces the standard Euclidean norm. In [3] we identified another such class of problems. Because its characteristics are minimum time paths to the boundary, the Eikonal equation has often been

*Received by the editors January 18, 2007; accepted for publication (in revised form) July 7, 2008; published electronically November 26, 2008. This work was supported by a grant from the National Science and Engineering Research Council of Canada.

<http://www.siam.org/journals/sinum/47-1/68035.html>

[†]Department of Computer Science, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (kalton@cs.ubc.ca, mitchell@cs.ubc.ca, <http://www.cs.ubc.ca/~mitchell>).

proposed for robotic path planning; for example, see [15]. However, for some robots, using the Euclidean norm in this equation is inappropriate. Consider a robot arm, where each joint has its own motor. If each motor can rotate at some maximum speed independent of the action of the other motors, then the action of the whole arm is best bounded in an appropriately-scaled infinity norm. The corresponding Eikonal equation should use the dual Manhattan norm and is thus anisotropic. Other scenarios where such problems arise were considered in [3]—such as planning collision-free optimal paths for multiple robots—and experimental evidence suggested that FMM would be successful on these problems.

As a group, the anisotropy in these problems is axis-aligned. In this paper we describe a broader class of such axis-aligned problems (section 2) and demonstrate that FMM can be applied to approximate their solution on axis-aligned orthogonal grids without modification of the algorithm beyond the local update function for a single node (section 3). The examples (section 4) include an anelliptic wave propagation problem and a new multirobot scenario. In Appendix A, we propose some methods by which the local update's efficiency might be improved even if the grid and/or PDE lack symmetry. Lastly, in Appendix B, we provide analytic update formulas for the Eikonal equation with the $p = 1, 2,$ and ∞ norms on variably spaced orthogonal grids in any dimension.

Some proofs of theorems and experimental details have been omitted from this paper and may be found in [2].

1.1. The problem. The Dirichlet problem of a static HJ PDE is to find a function u such that

$$(1.1a) \quad H(x, Du(x)) = 0, \quad x \in \Omega,$$

$$(1.1b) \quad u(x) = g(x), \quad x \in \partial\Omega,$$

where $Du(x)$ is the gradient of u at x , $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain, and $\partial\Omega$ is the domain's boundary. In general, it is not possible to find a classical solution to the Dirichlet problem (1.1) where u is differentiable for all x , so we seek instead the viscosity solution [7], a unique weak solution which is continuous and almost everywhere differentiable.

To appreciate the difference between isotropic and anisotropic problems, it is useful to consider a control-theoretic formulation of the Hamiltonian

$$(1.2) \quad H(x, q) = \max_{a \in \mathcal{A}(x)} (-q \cdot a) - 1,$$

where a is an action and $\mathcal{A}(x) \subset \mathbb{R}^d$ is a compact, convex action set containing the origin in its interior. In an isotropic problem $\mathcal{A}(x)$ is a hypersphere centered on the origin for all x , although its radius may depend on x . In such a problem (1.2) reduces to

$$(1.3) \quad H(x, q) = \|q\|_2 - c(x),$$

where $c(x) = 1/r(x)$ and $r(x)$ is the radius of the hyperspherical $\mathcal{A}(x)$. In this case (1.1a) becomes the Eikonal equation. For an anisotropic problem, $\mathcal{A}(x)$ is not always an origin-centered hypersphere. Since not all Hamiltonians H fit the control-theoretic formulation, more generally, for an isotropic problem, the set of q solving $H(x, q) = 0$ is the surface of an origin-centered hypersphere. Several examples of anisotropic problems that do not fit this criterion are included in sections 2.2 and 4.

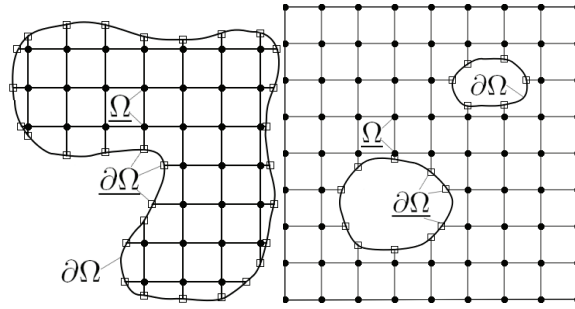


FIG. 1.1. Orthogonal grids combining discretizations $\underline{\Omega}$ and $\partial\Omega$. (a) boundary conditions are given around the outside of Ω . (b) boundary conditions are given on the inside of Ω .

1.2. The FMM. Since we typically cannot solve for the viscosity solution exactly, we compute an approximate solution \underline{u} on an axis-aligned orthogonal grid with nodes forming both a discretization $\underline{\Omega}$ of Ω , and a discretization $\partial\Omega$ of $\partial\Omega$; for example, see Figure 1.1. We take $\underline{\Omega}$ and $\partial\Omega$ to be disjoint sets. We allow any axis-aligned orthogonal grid, including those with node spacing that varies between dimensions and within a single dimension; the latter capability makes it easier to more accurately manage an irregular boundary [11]. It is important that the orthogonal grid and the Hamiltonian H are aligned to the same axis. What it means for H to be aligned to an axis is explained in section 2.

Let $\mathcal{N}(\underline{x})$ be the set of neighbors of node $\underline{x} \in \underline{\Omega}$. Whenever we refer to a simplex of a node \underline{x} , we mean a simplex specified by the node \underline{x} and d of its neighbors, each in a distinct dimension. Since we are restricted to orthogonal grids, each simplex of \underline{x} corresponds to a particular orthant.

Informally, we refer to $\underline{u}(\underline{x})$ as the value of node \underline{x} . In what follows, we may use \underline{u} to refer to either the values of the nodes in the operation or the output of FMM or to the solution of the discretized PDE (3.3). This ambiguity becomes less bothersome when we point out in Proposition 3.2 that the output of FMM is, in fact, the solution to (3.3).

Algorithm 1 outlines a simple dynamic programming algorithm. The algorithm can become either Dijkstra’s algorithm or FMM depending on the choice of the **Update** function. Consider, for example, the **Update** function in the context of optimal control, where we are computing the minimal cost over all possible paths. For Dijkstra’s algorithm, **Update** computes $\underline{u}(\underline{x}_0)$ as a simple minimization over the neighboring nodes of \underline{x}_0 of the path costs to \underline{x}_0 via each neighbor. For FMM, the **Update** function computes $\underline{u}(\underline{x}_0)$ as a minimization over the neighboring simplices of \underline{x}_0 of the minimum path costs to \underline{x}_0 via each simplex.

The **Update** function must satisfy a *causality* property in order for Algorithm 1 to terminate with a correct solution: **Update** must compute a node value $\underline{u}(\underline{x})$ based only on information from neighboring nodes with smaller values, so that \underline{u} is computed in increasing order of $\underline{u}(\underline{x})$ [28, 24]. In Dijkstra’s algorithm and FMM for a standard Euclidean norm Eikonal equation on an orthogonal grid, this property is automatic. A major contribution of this paper is to demonstrate that, for a class of static HJ PDEs with axis-aligned anisotropy, an **Update** function that is consistent with the PDE and satisfies the causality property can be defined, and thus FMM can be used.

While the **Update** function in Algorithm 1 is determined by the underlying equation which we seek to solve, it is assumed that its execution time is independent of

```

1 foreach  $\underline{x} \in \underline{\Omega}$  do  $\underline{u}(\underline{x}) \leftarrow \infty$ 
2 foreach  $\underline{x} \in \partial\underline{\Omega}$  do  $\underline{u}(\underline{x}) \leftarrow g(\underline{x})$ 
3  $\mathcal{Q} \leftarrow \underline{\Omega} \cup \partial\underline{\Omega}$ 
4 while  $\mathcal{Q} \neq \emptyset$  do
5    $\underline{y} \leftarrow \text{ExtractMin}(\mathcal{Q})$ 
6   foreach  $\underline{x}_0 \in (\mathcal{N}(\underline{y}) \cap \mathcal{Q}) \setminus \partial\underline{\Omega}$  do  $\underline{u}(\underline{x}_0) \leftarrow \text{Update}(\underline{x}_0, \underline{u})$ 
7 end

```

Algorithm 1: Dynamic Programming Algorithm.

grid resolution, and hence it does not affect the algorithm’s asymptotic complexity. The `Update` functions in this paper maintain this property. FMM is usually described as being $\mathcal{O}(n \log n)$, where $n = |\underline{\Omega}|$ is the number of grid points in the discretized domain. This complexity is derived by noting that each node is removed from \mathcal{Q} once by `ExtractMin` and, in the usual binary heap implementation of \mathcal{Q} , extraction of the minimum value node costs $\mathcal{O}(\log |\mathcal{Q}|) \leq \mathcal{O}(\log n)$. Note that the heap need only sort nodes with finite values. Because we restrict our modifications of Algorithm 1 to the `Update` function, all of the results here can be used with other versions of FMM; for example, the $\mathcal{O}(n)$ algorithm described in [30], which uses an untidy priority queue for \mathcal{Q} to reduce the cost of `ExtractMin` and hence the whole algorithm. However, for implementation simplicity, we have used the standard binary heap version of \mathcal{Q} in our experiments.

1.3. Related work. The first Dijkstra-like method for a first-order semi-Lagrangian discretization of the isotropic Eikonal PDE on an orthogonal grid was developed in [28]. The Dijkstra-like FMM was later independently developed in [23] for the first-order upwind Eulerian finite-difference discretization of the same Eikonal PDE. FMM was then extended to handle higher-order upwind discretizations on grids and unstructured meshes in \mathbb{R}^n and on manifolds [14, 25, 20]. In [24] it was shown that Dijkstra’s method on a uniform orthogonal grid produces the solution for the anisotropic maximum norm Eikonal equation. By solving an isotropic problem on a manifold and then projecting the solution into a subspace, FMM can solve certain anisotropic problems [20]; for example, (1.2) with a constant elliptic $\mathcal{A}(x) = \mathcal{A}$ can be solved by running isotropic FMM on an appropriately tilted planar manifold and then projecting away one dimension. Some anisotropic etching problems have also been solved using FMM [17].

The fact that correct operation of Dijkstra-like algorithms for approximating the Eikonal PDE requires the causality property that $\underline{u}(\underline{x})$ can be written only in terms of smaller values \underline{u} at neighboring nodes was stated in [28], but a reader might incorrectly infer from further comments in that paper that such algorithms would not work for any unstructured grid or anisotropic problem. That FMM is applicable for any consistent, orthogonal, causality satisfying, finite-difference discretization of a general static convex HJ PDE is stated in [24]; however, it is now understood that this criterion applies even more generally, since a Dijkstra-like method can be used to efficiently solve on a graph any nonlinear system of equations for which $\underline{u}(\underline{x})$ is dependent only on smaller values \underline{u} at neighboring nodes. A sufficient criterion (see section 2.1) under which FMM can be used for orthogonal, finite-difference discretizations of static HJ PDEs—now commonly referred to as “Osher’s criterion”—is widely attributed to an unpublished work by Osher and Helmsen, but the earliest published description seems

to be [17]. While it is stronger than the causality conditions described earlier, it is useful because it is stated as a condition on the analytic Hamiltonian instead of the equations created by the discretization. In this paper we likewise seek conditions under which FMM is applicable that are closer to the problem’s definition than the algorithm’s implementation.

OUMs [21, 22] can solve general convex anisotropic problems on unstructured grids with an asymptotic complexity only a constant factor (related to the degree of anisotropy) worse than FMM. FMM fails for these general problems because the neighboring simplex from which the characteristic approaches a node \underline{x}_0 may contain another node \underline{x} such that causality does not hold: $\underline{u}(\underline{x}_0) < \underline{u}(\underline{x})$. OUM avoids this difficulty by searching along the active front to find a set of neighboring nodes (which may not be direct neighbors of \underline{x}_0) whose values have been accepted, and then constructing a virtual simplex with these nodes from which to update $\underline{u}(\underline{x}_0)$. Although this search along the active front does not degrade the asymptotic complexity, it does significantly increase the computational cost in practice. This effect can be partially mitigated by using nontrivial data structures such as 2^d -trees to speed up the search.

An alternative to these single-pass (or label-setting) algorithms are the sweeping (or label-correcting) algorithms, which are often even simpler to implement than FMM. Sweeping algorithms are also capable of handling anisotropic and even non-convex problems. The simplest sweeping algorithm is to just iterate through the grid updating each node in a Gauss–Seidel (GS) fashion (so a new value for a node is used immediately in subsequent updates) until \underline{u} converges. GS converges quickly if the node update order is aligned with the characteristics of the solution, so better sweeping algorithms [8, 6, 31, 13, 19] alternate among a collection of static node orderings so that all possible characteristic directions will align with at least one ordering. It is argued in [31] that these methods achieve $\mathcal{O}(n)$ asymptotic complexity (assuming that the node orderings are already determined); however, unlike FMM and OUM, the constant depends on the problem. For practical grid resolutions on problems with curved characteristics, FMM does better despite the difference in asymptotic complexity [12, 11].

There are also a number of sweeping algorithms which use dynamic node orderings; for example [18, 5]. These algorithms attempt to approximate the optimal ordering generated by single-pass methods such as FMM without the overhead associated with managing an accurate queue. These methods have been demonstrated to be comparable to or better than single-pass methods for certain problems and grid resolutions [18, 5]. However, in general, these methods may need to revisit nodes multiple times.

Accurate robotic path planning is only required in cluttered environments where optimal paths—and hence the characteristics of the HJ PDE—are not straight. No alternative algorithm proposed approaches the simple implementation and guaranteed speed of FMM for these types of problems. Consequently, we set out in this paper to characterize another class of anisotropic HJ PDEs for which FMM will work and also to explore their efficient implementation. It should be noted that the update procedures discussed in this paper can be applied to any of the sweeping algorithms without modification.

2. Class of Hamiltonians. FMM can be extended to handle a class of axis-aligned anisotropic problems, defined by a restriction of the Hamiltonian H to that satisfying Properties 1 to 4. We let $q, \tilde{q} \in \mathbb{R}^d$ and make the following definitions.

DEFINITION 2.1. *Write $q \succeq \tilde{q}$ if $q_j \tilde{q}_j \geq 0$ and $|q_j| \geq |\tilde{q}_j|$, for $1 \leq j \leq d$.*

DEFINITION 2.2. Write $q \triangleright \tilde{q}$ if (i) $q \neq 0$ and (ii) $q_j \tilde{q}_j \geq 0$ and $|q_j| > |\tilde{q}_j|$ or $q_j = \tilde{q}_j = 0$, for $1 \leq j \leq d$.

The following properties are satisfied by H .

PROPERTY 1. H is continuous: $H \in C(\Omega \times \mathbb{R}^d)$.

PROPERTY 2. H is coercive: $H(x, q) \rightarrow \infty$ as $\|q\| \rightarrow \infty$ for all $x \in \Omega$.

PROPERTY 3. H is strictly compatible: $H(x, 0) < 0$ for all $x \in \Omega$.

PROPERTY 4. H is strictly one-sided monotone: If $q \triangleright \tilde{q}$, then $H(x, q) > H(x, \tilde{q})$.

We note that Properties 1, 2, and 3 are similar to some properties on the Hamiltonian in [5]. In this paper, we typically deal only with the `Update` function. For this reason, we usually consider a fixed $x \in \Omega$ and may write $H(q) = H(x, q)$ wherever no ambiguity results. When discussing properties of H , these are in reference to the q parameter. The source of the *axis-aligned* description of the problem class is the strict one-sided monotonicity property of H .

2.1. Connection to Osher’s criterion. Although there are earlier statements of the conditions on node values under which a Dijkstra-like algorithm can or cannot be used to solve the problem [28, 23], in this section we outline the connection between the properties described above and Osher’s criterion [17] because the latter directly provides a condition on the Hamiltonian rather than on the solution values. In section 3.3, we make the connection between Properties 1 to 4 and the earlier conditions.

Osher’s fast marching criterion is defined in [17, 27] as

$$q_j \frac{\partial H(x, q)}{\partial q_j} \geq 0$$

for $1 \leq j \leq d$. The authors state there that as long as this criterion is satisfied, a simple fast marching algorithm based on a one-sided upwind finite-difference discretization can be applied to solve the problem. However, we use Properties 1 to 4 instead of Osher’s criterion because Osher’s criterion requires H to be differentiable so that $D_q H(x, q)$ exists, but we are interested in potentially nondifferentiable H (e.g., see section 2.2). Note that strict one-sided monotonicity is applicable even when $D_q H(x, q)$ does not exist for all x .

Propositions 2.3, 2.4, and 2.5 explain the relationship between strict one-sided monotonicity of H (Property 4) and Osher’s criterion. Proposition 2.3 shows that Property 4 implies one-sided monotonicity (Property 5). Then, Proposition 2.4 shows that Property 5 is the same as Osher’s criterion as long as H is differentiable. Finally, Proposition 2.5 demonstrates that Property 5 with the addition of one-sided homogeneity (Property 6) implies Property 4.

PROPERTY 5. H is one-sided monotone: If $q \succeq \tilde{q}$, then $H(x, q) \geq H(x, \tilde{q})$.

PROPOSITION 2.3. Let H be continuous (Property 1). Then strict one-sided monotonicity of H (Property 4) implies one-sided monotonicity of H (Property 5).

Proof. Let H be strictly one-sided monotone. Let $q, \tilde{q} \in \mathbb{R}^d$ be such that $q \succeq \tilde{q}$. Let $r \in \{-1, 1\}^d$ be such that

$$r_j = \begin{cases} +1, & \text{if } q_j \geq 0, \\ -1, & \text{otherwise,} \end{cases}$$

and let $\epsilon > 0$. Note that $q + \epsilon r \triangleright \tilde{q}$ and thus we have $H(q + \epsilon r) > H(\tilde{q})$. By the continuity of H , we have

$$\lim_{\epsilon \rightarrow 0^+} H(q + \epsilon r) \geq H(\tilde{q})$$

and also

$$\lim_{\epsilon \rightarrow 0^+} H(q + \epsilon r) = H(q).$$

Therefore, $H(q) \geq H(\tilde{q})$. \square

PROPOSITION 2.4. *Let H be continuous (Property 1), and let $D_q H(q)$ exist for all $q \in \mathbb{R}^d$. Then the following conditions on H are equivalent:*

- (a) $q_j \frac{\partial H(q)}{\partial q_j} \geq 0$ for all j and $q \in \mathbb{R}^d$.
- (b) H is one-sided monotone (Property 5).

Proof. We begin by proving that (a) implies (b). Let $q, \tilde{q} \in \mathbb{R}^d$ be such that $q \succeq \tilde{q}$. If $q = \tilde{q}$, then $H(q) = H(\tilde{q})$. Otherwise, define the function $\bar{q} : [0, 1] \rightarrow \mathbb{R}^d$ such that $\bar{q}(t) = \tilde{q} + t(q - \tilde{q})$ to represent the line segment between \tilde{q} and q parameterized by $t \in [0, 1]$. Because $q \succeq \tilde{q}$ we have

$$(q - \tilde{q})_j \bar{q}_j(t) \geq 0$$

for $1 \leq j \leq d$ and for $t \in [0, 1]$. Thus, by condition (a), we have

$$(2.1) \quad (q - \tilde{q})_j \frac{\partial H(\bar{q}_j(t))}{\partial \bar{q}_j(t)} \geq 0$$

for $1 \leq j \leq d$ and for $t \in [0, 1]$. We know that

$$\begin{aligned} H(q) &= H(\tilde{q}) + \int_0^1 \frac{d\bar{q}(t)}{dt} \cdot D_q H(\bar{q}(t)) dt \\ &= H(\tilde{q}) + \int_0^1 (q - \tilde{q}) \cdot D_q H(\bar{q}(t)) dt \\ &= H(\tilde{q}) + \int_0^1 \sum_{i=1}^n (q - \tilde{q})_j \frac{\partial H(\bar{q}_j(t))}{\partial \bar{q}_j(t)} dt \\ &\geq H(\tilde{q}). \end{aligned}$$

The first equality follows from integrating the change in H along the line segment connecting \tilde{q} and q . The second equality is because the derivative $\frac{d\bar{q}(t)}{dt}$ is simply the vector $q - \tilde{q}$. The third equality breaks up the vector dot product into a sum of scalar products. The inequality results from (2.1) and the fact that an integral of a nonnegative function is nonnegative. Thus, for all q, \tilde{q} such that $q \succeq \tilde{q}$, including $q = \tilde{q}$, we have $H(q) \geq H(\tilde{q})$.

We now prove that (b) implies (a). Let $q \in \mathbb{R}^d$ and $1 \leq j \leq d$. Define the function $s : \mathbb{R} \rightarrow \{-1, +1\}$ such that

$$s(y) = \begin{cases} +1, & \text{if } y \geq 0, \\ -1, & \text{otherwise,} \end{cases}$$

let $\epsilon > 0$, and let e_j be the j th vector in the standard basis. Note that $q + \epsilon s(q_j)e_j \succeq q$ and thus by (b) we have $H(q + \epsilon s(q_j)e_j) - H(q) \geq 0$. Consequently, by the existence of $D_q H(q)$ for all $q \in \mathbb{R}^d$, we have

$$q_j \frac{\partial H(q)}{\partial q_j} = \lim_{\epsilon \rightarrow 0^+} q_j \frac{H(q + \epsilon s(q_j)e_j) - H(q)}{\epsilon s(q_j)} \geq 0. \quad \square$$

The following property is used to state Proposition 2.5.

PROPERTY 6. H is one-sided homogeneous: $H(tq) - H(0) = t(H(q) - H(0))$ for all $t \geq 0$ and $q \in \mathbb{R}^d$.

PROPOSITION 2.5. Let H satisfy Properties 1, 2, and 3, and let H be one-sided monotone (Property 5) and one-sided homogeneous (Property 6). Then H is strictly one-sided monotone (Property 4).

Proof. Let $q \triangleright \tilde{q}$. Then $q \supseteq \tilde{q}$ and $H(q) \geq H(\tilde{q})$ by one-sided monotonicity.

First consider the case $\tilde{q} = 0$. Assume $H(q) = H(\tilde{q}) = H(0)$. By the one-sided homogeneity of H ,

$$\lim_{t \rightarrow \infty} [H(tq) - H(0)] = \lim_{t \rightarrow \infty} [t(H(q) - H(0))] = 0.$$

But by the coercivity of H ,

$$\lim_{t \rightarrow \infty} [H(tq) - H(0)] = \infty,$$

since $\lim_{t \rightarrow \infty} \|tq\| = \infty$ and by compatibility $H(0) < 0$. Thus, we have a contradiction, and it must be that $H(q) > H(\tilde{q})$.

Second, consider the case where $\tilde{q} \neq 0$. Let $\mathcal{J} = \{j \mid |q_j| > |\tilde{q}_j|\}$. Note that by Definition 2.2 since $\tilde{q} \neq 0$, we have $\mathcal{J} \neq \emptyset$ and there exist $j \in \mathcal{J}$ such that $\tilde{q}_j \neq 0$. Define a scalar multiple of q :

$$\check{q} = tq = \left(\max_{j \in \mathcal{J}} \frac{|\tilde{q}_j|}{|q_j|} \right) q.$$

Since $|q_j| > |\tilde{q}_j|$, for all $j \in \mathcal{J}$, we have $0 < t < 1$. Furthermore, for $j \in \mathcal{J}$,

$$|\check{q}_j| = \left(\max_{j \in \mathcal{J}} \frac{|\tilde{q}_j|}{|q_j|} \right) |q_j| \geq |\tilde{q}_j|,$$

while for $j \notin \mathcal{J}$,

$$\check{q}_j = tq_j = 0 = \tilde{q}_j.$$

Consequently, $|\check{q}_j| \geq |\tilde{q}_j|$ for $1 \leq j \leq d$. Also, since $t > 0$, we have $\check{q}_j \tilde{q}_j = tq_j \tilde{q}_j \geq 0$ for $1 \leq j \leq d$. This implies, by one-sided monotonicity of H , that $H(\check{q}) \geq H(\tilde{q})$. Moreover, by one-sided homogeneity of H , $H(\check{q}) - H(0) = H(tq) - H(0) = t(H(q) - H(0))$. It follows that $H(q) - H(0) = (H(\check{q}) - H(0))/t > H(\tilde{q}) - H(0)$, since $0 < t < 1$ and $H(\check{q}) \geq H(0)$ by one-sided monotonicity. Therefore, $H(q) > H(\tilde{q}) \geq H(\tilde{q})$. \square

We impose strict one-sided monotonicity on H because it guarantees a unique solution to a first-order upwind finite-difference discretization of (1.1a), as shown in section 3.1. Simply imposing one-sided monotonicity on H or Osher's condition on differentiable H is not sufficient for a unique solution. However, Proposition 2.5 states that when H satisfies one-sided homogeneity in addition to one-sided monotonicity, then it also satisfies strict one-sided monotonicity, and there is a unique solution to the discretization. Moreover, by Propositions 2.4 and 2.5, when differentiable H satisfies one-sided homogeneity in addition to Osher's criterion, then H also satisfies strict one-sided monotonicity, and there is a unique solution to the discretization. Note that there exist conditions other than one-sided homogeneity, such as strict convexity, that in combination with Osher's criterion, result in strict one-sided monotonicity of H .

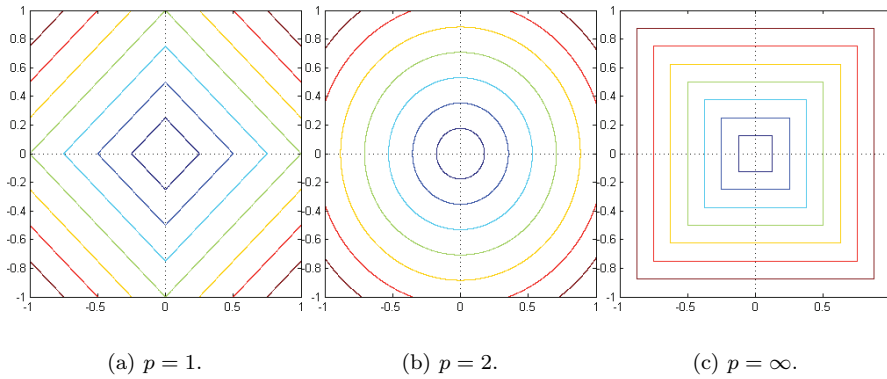


FIG. 2.1. Contour plots of $\|q\|_p$.

2.2. Example H functions. A Hamiltonian H that satisfies Properties 1 to 4 encompasses a fairly broad range of anisotropic problems. We consider examples of H that satisfy Properties 1 to 4. In particular, we look at the case

$$(2.2) \quad H(x, q) = G(x, q) - c(x),$$

where G is a p -norm or some variant and c is a positive cost. We must ensure that G is strictly one-sided monotone, which is not true of all norms.

The p -norm is a useful category of strictly one-sided monotone norms. Let a p -norm, $\|\cdot\|_p$, be defined by

$$\|q\|_p = \left(\sum_{j=1}^d |q_j|^p \right)^{1/p},$$

where $p \geq 1$. Commonly used p -norms, illustrated in Figure 2.1, are the Manhattan norm ($p = 1$), the Euclidean norm ($p = 2$), and the maximum norm ($p = \infty$). The following proposition is proved in [2].

PROPOSITION 2.6. $\|\cdot\|_p$ is strictly one-sided monotone.

Define a linearly-transformed p -norm $\|\cdot\|_{B,p}$ to be

$$\|q\|_{B,p} = \|Bq\|_p,$$

where $p \geq 1$ and B is a nonsingular $d \times d$ matrix. Note that B must be nonsingular so that $\|\cdot\|_{B,p}$ satisfies the properties of a norm such as definiteness and homogeneity. Such a norm is not strictly one-sided monotone in general. Figure 2.2(a) shows a simple example where a vector is rotated by $-\pi/4$ and scaled by 3 in the q_2 -axis before the Euclidean norm is taken; i.e.,

$$(2.3) \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} \cos(-\pi/4) & -\sin(-\pi/4) \\ \sin(-\pi/4) & \cos(-\pi/4) \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -3/\sqrt{2} & 3/\sqrt{2} \end{bmatrix}.$$

Let $q = (2, 2)^T$ and $\tilde{q} = (\sqrt{2}, 0)^T$. We have $q \succeq \tilde{q}$, but

$$\|Bq\|_2 = \left\| \begin{pmatrix} 2\sqrt{2} \\ 0 \end{pmatrix} \right\|_2 = \sqrt{8} < \sqrt{10} = \|(1, -3)^T\|_2 = \|B\tilde{q}\|_2.$$

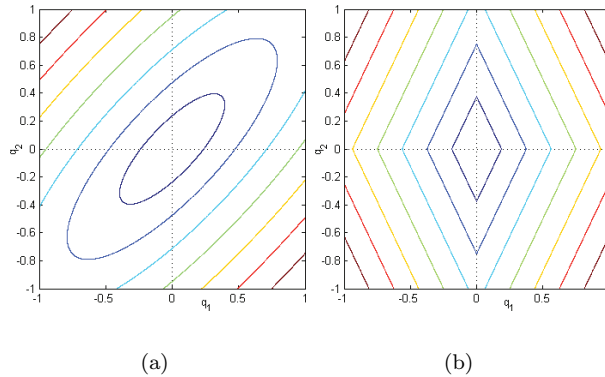


FIG. 2.2. Contour plots of $\|Bq\|_p$. (a) is not strictly one-sided monotone: $p = 2$ and B is defined by (2.3). (b) is strictly one-sided monotone: $p = 1$ and B scales by 2 in the q_1 -axis.

Consequently, this particular linearly transformed p -norm is not strictly one-sided monotone. However, in this case an inverse transformation B^{-1} of the grid coordinates will result in a strictly one-sided monotone p -norm, while maintaining the grid's orthogonality. More generally, we conjecture that if the Hamiltonian is of the form $H(q) = \tilde{H}(Bq)$, where B is a rotation (which may be followed by scaling) and \tilde{H} satisfies Properties 1 to 4, a transformation of the grid coordinates by B^{-1} will result in a transformed H that also satisfies Properties 1 to 4, while maintaining the grid's orthogonality. More complex coordinate modifications might be possible, but we have not yet adequately investigated conditions or procedures.

A scaled p -norm (Figure 2.2(b)) is a special case of a linearly transformed p -norm. Such a norm scales the components of its argument before applying a p -norm by restricting B to be a nonsingular diagonal matrix. It is simple to show that a scaled p -norm is strictly one-sided monotone, considering Proposition 2.6.

A mixed p -norm is a recursive composition of p -norms, and it is strictly one-sided monotone. The following is an example (Figure 2.3(a)) of a mixed p -norm that takes the Euclidean norm of the first two components and then takes the Manhattan norm of the result and the last component:

$$(2.4) \quad \begin{aligned} \|q\| &= \|(\|(q_1, q_2)\|_2, q_3)\|_1 \\ &= \sqrt{(q_1)^2 + (q_2)^2} + |q_3|, \end{aligned}$$

where $q = (q_1, q_2, q_3)$. This particular norm was used as a G function in [3] for a simple two-robot coordinated optimal control problem.

Finally, the one-sidedness of Property 4 allows G to be asymmetric, which is not permitted for a norm. An example of such an asymmetric norm-like function is shown in Figure 2.3(b) and is given by

$$(2.5) \quad G(q) = \begin{cases} \|B_a q\|_\infty, & \text{if } q_1 \leq 0 \text{ and } q_2 \leq 0, \\ \|B_b q\|_1, & \text{if } q_1 \leq 0 \text{ and } q_2 > 0, \\ \|B_c q\|_2, & \text{if } q_1 > 0 \text{ and } q_2 \leq 0, \\ \|B_d q\|_2, & \text{if } q_1 > 0 \text{ and } q_2 > 0, \end{cases}$$

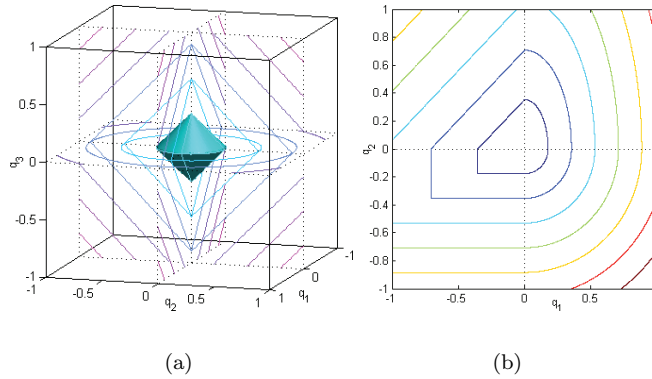


FIG. 2.3. Contour plots of $G(q)$. (a) mixed p -norm: G is defined by (2.4). (b) asymmetric norm-like function: G is defined by (2.5).

where

$$B_a = \begin{bmatrix} 1/2 & 0 \\ 0 & 1 \end{bmatrix} \quad B_b = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix} \quad B_c = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad B_d = \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix}.$$

We solve the anisotropic problem characterized by (2.5) as well as an anelliptic wave propagation problem and a multirobot optimal path planning problem in section 4. Other examples of G functions which satisfy strict one-sided monotonicity are some polygonal norms such as axis aligned hexagonal or octagonal norms; however, we do not further investigate these options here.

3. FMM and the discretized problem. We define a discretized analogue of the Dirichlet problem (1.1). By describing the `Update` function in Algorithm 1, we also formalize the FMM algorithm. Finally, we examine important properties of the `Update` function.

We recall that the nodes in $\underline{\Omega}$ lie on an axis-aligned orthogonal grid. Let $\underline{x}_0 \in \underline{\Omega}$. The neighborhood of \underline{x}_0 is shown in Figure 3.1. Let \underline{x}_j^\pm be the neighbors of \underline{x}_0 in the $\pm e_j$ directions, e_j being the j th vector in the standard basis. The set of neighbors is

$$\mathcal{N}(\underline{x}_0) = \{\underline{x}_1^\pm, \underline{x}_2^\pm, \dots, \underline{x}_d^\pm\},$$

and the neighborhood vector is

$$N(\underline{x}_0) = (\underline{x}_0, \underline{x}_1^\pm, \underline{x}_2^\pm, \dots, \underline{x}_d^\pm).$$

Let $h_j^\pm = \pm \|\underline{x}_0 - \underline{x}_j^\pm\|$ be signed distances to the neighbors in the $\pm e_j$ directions. Let

$$\mathcal{S} = \{(s_1, s_2, \dots, s_d) \mid s_j \in \{-1, +1\}, 1 \leq j \leq d\}$$

such that $s \in \mathcal{S}$ represents one of the 2^d neighboring simplices of \underline{x}_0 . Note that we abuse notation by using $s_j \in \{-1, +1\}$ as a superscript indexing \underline{x}_j^\pm or h_j^\pm .

Let $B(\Omega)$ be the set of bounded functions on domain Ω . We define the numerical Hamiltonian $\underline{H} : \Omega^{1+2d} \times B(\Omega) \times \mathbb{R} \rightarrow \mathbb{R}$ as follows:

$$(3.1) \quad \underline{H}(N, \phi, \mu) = \max_{s \in \mathcal{S}} [H(\underline{x}_0, \underline{D}^s(N, \phi, \mu))],$$

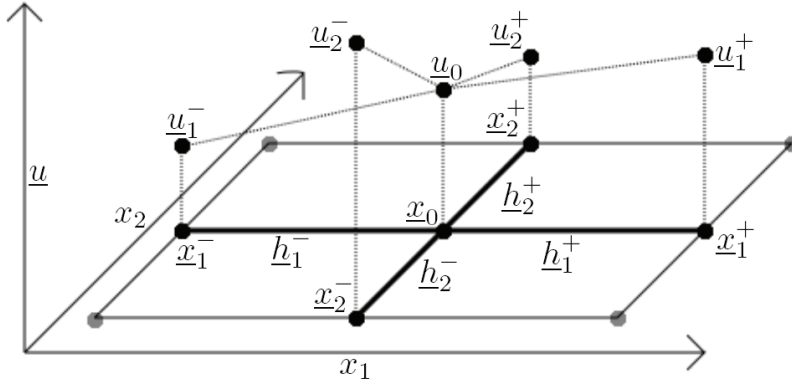


FIG. 3.1. Neighborhood of \underline{x}_0 with $d = 2$.

where H is as defined in section 2 and

$$\underline{D}^s(N, \phi, \mu) = (\underline{D}_1^s(N, \phi, \mu), \underline{D}_2^s(N, \phi, \mu), \dots, \underline{D}_d^s(N, \phi, \mu))$$

is a first-order, upwind, finite-difference gradient approximation from the simplex represented by s ; that is,

$$(3.2) \quad \underline{D}_j^s(N, \phi, \mu) = \frac{\max(0, \mu - \phi(\underline{x}_j^{s_j}))}{-h_j^{s_j}}$$

for $1 \leq j \leq d$. Although \underline{H} is defined on domain $\Omega^{1+2d} \times B(\Omega) \times \mathbb{R}$, for FMM it will only be used on domain $\underline{\Omega}^{1+2d} \times B(\underline{\Omega}) \times \mathbb{R}$. The broader definition of domain is important for consistency [4]. The restriction of Ω^{1+2d} to $\underline{\Omega}^{1+2d}$ poses no problems to the definition of \underline{H} . Furthermore, to evaluate \underline{H} , ϕ need only be defined on \mathcal{N} , which is true of any function in $B(\underline{\Omega})$.

The discretized Dirichlet problem is to find a function $\underline{u} : (\underline{\Omega} \cup \partial\underline{\Omega}) \rightarrow \mathbb{R}$ such that

$$(3.3a) \quad \underline{H}(N(\underline{x}), \underline{u}, \underline{u}(\underline{x})) = 0, \quad \underline{x} \in \underline{\Omega},$$

$$(3.3b) \quad \underline{u}(\underline{x}) = g(\underline{x}), \quad \underline{x} \in \partial\underline{\Omega}.$$

DEFINITION 3.1. Let FMM be Algorithm 1 with the Update function defined as follows. A call to $\text{Update}(\underline{x}_0, \underline{u})$ returns the solution $\mu = \tilde{\mu}$ to

$$(3.4) \quad \underline{H}(N(\underline{x}_0), \underline{u}, \mu) = 0.$$

In this way it determines a node’s value $\underline{u}(\underline{x}_0) \leftarrow \tilde{\mu}$ given the values of its neighbors, $\underline{u}_j^\pm = \underline{u}(\underline{x}_j^\pm)$. When we are varying only μ , it will be convenient to write $\underline{H}(\mu) = \underline{H}(N, \phi, \mu)$ and $\underline{D}^s(\mu) = \underline{D}^s(N, \phi, \mu)$. For the lemmas and theorems stated below, we assume H satisfies Properties 1 to 4.

PROPOSITION 3.2. Let $\underline{u} : (\underline{\Omega} \cup \partial\underline{\Omega}) \rightarrow \mathbb{R}$ be the grid function after FMM terminates. Then \underline{u} is the unique solution of (3.3).

This proposition states that the grid function \underline{u} that results from running FMM solves the discretized problem (3.3). We use a method similar to those for isotropic FMM in [29, 23] to prove Proposition 3.2 in [2]. The causality of the Update function is essential so that FMM can be used to solve (3.3).

A method for proving the convergence of \underline{u} to the solution of (1.1) as the grid spacing goes to zero is presented in [4]. It is shown there that the consistency, monotonicity, and stability of the numerical scheme are sufficient for convergence. We closely follow the technique described in [4] to prove convergence in [2]. Also, uniqueness and monotonicity of the solution to (3.4) are useful for using numerical root finders to implement `Update`. We include proofs of uniqueness, monotonicity, and causality of the `Update` function below. For more details regarding convergence, including the proofs of consistency and stability, see [2].

3.1. Unique update. Let the minimum value of all neighbors of \underline{x}_0 be

$$(3.5) \quad \check{u} = \min_{\underline{x} \in \mathcal{N}(\underline{x}_0)} (\underline{u}(\underline{x})).$$

We show there is a unique solution $\mu = \tilde{\mu}$ to (3.4) such that $\tilde{\mu} > \check{u}$. First, we prove two useful lemmas.

LEMMA 3.3. *$\underline{H}(\mu)$ is strictly increasing on $\mu \geq \check{u}$.*

Proof. Let $\mu_a > \mu_b \geq \check{u}$. Let $s \in \mathcal{S}$ and $1 \leq j \leq d$. If $\mu_a > \underline{u}_j^{s_j}$, then $\underline{D}_j^s(\mu_a)\underline{D}_j^s(\mu_b) \geq 0$ and $|\underline{D}_j^s(\mu_a)| > |\underline{D}_j^s(\mu_b)|$. On the other hand, if $\mu_a \leq \underline{u}_j^{s_j}$, then $\underline{D}_j^s(\mu_a) = \underline{D}_j^s(\mu_b) = 0$. Also, there exists at least one $s \in \mathcal{S}$ and $1 \leq j \leq d$ such that $\underline{D}_j^s(\mu_a) \neq 0$, since $\mu_a > \check{u}$. For such s , $H(\underline{D}^s(\mu_a)) > H(\underline{D}^s(\mu_b))$, by strict one-sided monotonicity (Property 4). For all other s , $H(\underline{D}^s(\mu_a)) = H(\underline{D}^s(\mu_b)) = H(0)$. Therefore, by (3.1) $\underline{H}(\mu_a) > \underline{H}(\mu_b)$, so $\underline{H}(\mu)$ is strictly increasing on $\mu \geq \check{u}$. \square

LEMMA 3.4. *The numerical Hamiltonian $\underline{H}(\mu)$ satisfies the following:*

- (a) $\underline{H}(\mu) = H(0) < 0$ for $\mu \leq \check{u}$.
- (b) $\underline{H}(\mu) \rightarrow \infty$ as $\mu \rightarrow \infty$.
- (c) $\underline{H}(\mu)$ is nondecreasing on all μ .

Proof. If $\mu \leq \check{u}$, then by (3.2) and (3.5), we have $\underline{D}_j^s(\mu) = 0$ for all $s \in \mathcal{S}$, $1 \leq j \leq d$. By the strict compatibility of H , $H(\underline{D}^s(v_j)) = H(0) < 0$ for all s . By (3.1), we have $\underline{H}(\mu) = H(0) < 0$, for $\mu \leq \check{u}$, proving (a).

Let $s \in \mathcal{S}$ and $1 \leq j \leq d$. As $\mu \rightarrow \infty$, we have $\underline{D}_j^s(\mu) \rightarrow \infty$ and $\|\underline{D}^s(\mu)\| \rightarrow \infty$ for all $s \in \mathcal{S}$, $1 \leq j \leq d$. By the coercivity of H , as $\mu \rightarrow \infty$, we have $H(\underline{D}^s(\mu)) \rightarrow \infty$ for all $s \in \mathcal{S}$. By (3.1), we have $\underline{H}(\mu) \rightarrow \infty$ as $\mu \rightarrow \infty$, proving (b).

Because $\underline{H}(\mu)$ is constant on $\mu \leq \check{u}$ and by Lemma 3.3 increasing on $\mu \geq \check{u}$, $\underline{H}(\mu)$ is nondecreasing on all μ , proving (c). \square

THEOREM 3.5. *There exists a unique solution $\mu = \tilde{\mu}$ to $\underline{H}(\mu) = 0$ such that $\tilde{\mu} > \check{u}$.*

Proof. Each $\underline{D}_j^s(\mu)$ is continuous on μ . Furthermore, by the continuity of H , $H(\underline{D}^s(\mu))$ is continuous on μ for all s . Since \max is continuous, $\underline{H}(\mu)$ is continuous. By Lemma 3.4(a/b), $\underline{H}(\mu) < 0$ for $\mu \leq \check{u}$ and $\underline{H}(\mu) \rightarrow \infty$ as $\mu \rightarrow \infty$. Therefore, by the intermediate value theorem, there exists a solution $\mu = \tilde{\mu}$ to $\underline{H}(\mu) = 0$ such that $\check{u} < \tilde{\mu} < \infty$. Moreover, since \underline{H} is strictly increasing on $\mu \geq \check{u}$ by Lemma 3.3, the solution is unique. \square

Remark 1. We note that strict one-sided monotonicity (Property 4) of H is used to prove Lemma 3.3, and Lemma 3.3 is then used to show that the solution to $\underline{H}(\mu) = 0$ is unique. We might consider whether or not one-sided monotonicity (Property 5) of H is sufficient for a unique solution. However, Property 5 would not be sufficient to prove Lemma 3.3, and we would find that $\underline{H}(\mu)$ is only nondecreasing on $\mu \geq \check{u}$. A solution to $\underline{H}(\mu) = 0$ would still be guaranteed but not unique in this case. Analogously, for differentiable H , Osher’s criterion on H implies a solution that

may not be unique unless H satisfies some additional property, such as one-sided homogeneity (Property 6) or convexity.

3.2. Monotonicity. We show that \underline{H} and the `Update` function are monotone in the neighbor’s values. Monotonicity of \underline{H} requires that if none of the neighbor’s values decreases, the numerical Hamiltonian \underline{H} should not increase. Additionally, monotonicity of the `Update` function requires that if none of the neighbor’s values decreases, the solution to (3.4) should not decrease. Monotonicity is useful both for showing that FMM finds a unique solution and for proving convergence. We note that monotonicity does not require strict one-sided monotonicity of H , but rather one-sided monotonicity of H is sufficient.

THEOREM 3.6. *Let \underline{v} and \underline{u} be grid functions. Let $\underline{v}_j^\pm \geq \underline{u}_j^\pm$ for $1 \leq j \leq d$. Then for $\mu \in \mathbb{R}$, we have $\underline{H}(N, \underline{v}, \mu) \leq \underline{H}(N, \underline{u}, \mu)$. Furthermore, if $\mu = \mu_v$ is the unique solution to $\underline{H}(N, \underline{v}, \mu) = 0$ and $\mu = \mu_u$ is the unique solution to $\underline{H}(N, \underline{u}, \mu) = 0$, then $\mu_v \geq \mu_u$.*

Proof. Let $\mu \in \mathbb{R}$. We have $\underline{D}^s(N, \underline{u}, \mu) \supseteq \underline{D}^s(N, \underline{v}, \mu)$ for all $s \in \mathcal{S}$. Also, by Proposition 2.3, H satisfies one-sided monotonicity (Property 5). Thus,

$$H(\underline{D}^s(N, \underline{u}, \mu)) \geq H(\underline{D}^s(N, \underline{v}, \mu)) = 0$$

for all $s \in \mathcal{S}$. Consequently, $\underline{H}(N, \underline{u}, \mu) \geq \underline{H}(N, \underline{v}, \mu)$, proving the first claim.

To prove the second claim, we let μ_v and μ_u be as defined above. We note that $\underline{H}(N, \underline{u}, \mu_u) = 0 \geq \underline{H}(N, \underline{v}, \mu_u)$. By Lemma 3.4(c), $\underline{H}(N, \underline{v}, \mu)$ is nondecreasing on all μ , so in order that $\underline{H}(N, \underline{v}, \mu_v) = 0$, it must be that $\mu_v \geq \mu_u$. \square

3.3. Causality. We note that (3.3) defines a very large system of nonlinear equations, one equation for each node $\underline{x} \in \underline{\Omega}$. FMM can be used to solve this system very efficiently, if the solution $\mu = \tilde{\mu}$ to (3.4) is dependent only on neighbors with smaller values. This property represents a causal relationship between node values. There is an information flow from nodes with smaller values to those with larger values. The causal relationship is meant to mimic that of the PDE (1.1). The solution u of (1.1) is completely defined at x using only values of u from states that are backwards along the characteristic line that passes through x .

FMM exploits the causal property of \underline{H} by computing $\underline{u}(\underline{x})$ in increasing order in a single pass through the nodes. This causal property has been discussed as a requirement for Dijkstra-like single-pass methods in several works [28, 24, 26, 17, 22]. The following theorem states that \underline{H} and the `Update` function are causal. The `Update` function is considered causal if any change to the value of a neighboring node, such that both the new and old values are no smaller than the solution $\mu = \tilde{\mu}$ to $\underline{H}(N, \underline{u}, \mu) = 0$, has no effect on the solution.

THEOREM 3.7. *Let \underline{v} and \underline{u} be grid functions. Let*

$$\tilde{\mathcal{N}}(\underline{x}_0) = \{\underline{x} \in \mathcal{N}(\underline{x}_0) \mid \underline{v}(\underline{x}) \neq \underline{u}(\underline{x})\}.$$

Let

$$\tilde{w} = \begin{cases} \min_{\underline{x} \in \tilde{\mathcal{N}}(\underline{x}_0)} \min(\underline{v}(\underline{x}), \underline{u}(\underline{x})), & \text{if } \tilde{\mathcal{N}}(\underline{x}_0) \neq \emptyset, \\ +\infty, & \text{otherwise.} \end{cases}$$

Then $\underline{H}(N, \underline{v}, \mu) = \underline{H}(N, \underline{u}, \mu)$ for $\mu \leq \tilde{w}$.

Furthermore, let $\mu = \tilde{\mu}_u$ be the unique solution to $\underline{H}(N, \underline{u}, \mu) = 0$, and let $\mu = \tilde{\mu}_v$ be the unique solution to $\underline{H}(N, \underline{v}, \mu) = 0$. If $\tilde{\mu}_u \leq \tilde{w}$ or $\tilde{\mu}_v \leq \tilde{w}$, then $\tilde{\mu}_u = \tilde{\mu}_v$.

TABLE 4.1

Errors of approximate solution computed by FMM compared to exact solution of (1.1), where H is as in (2.2) and $G(Du(x)) = \|Du(x)\|_p$. The variables d , m , and n are the dimension, the number of nodes in each dimension, and the total number of nodes, respectively. Other variables are the spacing h between grid nodes, the \mathcal{L}_∞ -error e_∞ , the \mathcal{L}_∞ convergence rate r_∞ , the \mathcal{L}_1 -error e_1 , and the \mathcal{L}_1 convergence rate r_1 .

d	m	n	h	$p = 1$				$p = 2$			
				e_∞	r_∞	e_1	r_1	e_∞	r_∞	e_1	r_1
2	11	1.2e2	2.0e-1	2.2e-1		6.3e-2		1.2e-1		6.2e-2	
	21	4.4e2	1.0e-1	1.7e-1	.41	3.7e-2	.77	7.8e-2	.56	4.3e-2	.53
	41	1.7e3	5.0e-2	1.2e-1	.46	2.0e-2	.85	5.0e-2	.65	2.8e-2	.63
	81	6.6e3	2.5e-2	8.8e-2	.48	1.1e-2	.90	3.1e-2	.70	1.7e-2	.69
	161	2.6e4	1.3e-2	6.3e-2	.49	5.7e-3	.94	1.8e-2	.75	1.0e-2	.73
	321	1.0e5	6.3e-3	4.4e-2	.49	2.9e-3	.96	1.1e-2	.78	6.1e-3	.77
	641	4.1e5	3.1e-3	3.1e-2	.50	1.5e-3	.97	6.1e-3	.81	3.5e-3	.79
	1281	1.6e6	1.6e-3	2.2e-2	.50	7.6e-4	.98	3.4e-3	.83	2.0e-3	.82
3	11	1.3e3	2.0e-1	3.5e-1		1.2e-1		2.1e-1		1.2e-1	
	21	9.3e3	1.0e-1	2.6e-1	.43	6.9e-2	.78	1.4e-1	.58	8.4e-2	.57
	41	6.9e4	5.0e-2	1.9e-1	.47	3.9e-2	.85	8.7e-2	.66	5.4e-2	.65
	81	5.3e5	2.5e-2	1.3e-1	.49	2.1e-2	.89	5.3e-2	.72	3.3e-2	.70
	161	4.2e6	1.3e-2	9.5e-2	.50	1.1e-2	.92	3.1e-2	.76	2.0e-2	.74
4	11	1.5e4	2.0e-1	4.4e-1		1.7e-1		2.9e-1		1.8e-1	
	21	1.9e5	1.0e-1	3.2e-1	.45	9.8e-2	.78	1.9e-1	.60	1.2e-1	.58
	41	2.8e6	5.0e-2	2.3e-1	.48	5.5e-2	.83	1.2e-1	.67	7.7e-2	.66

Proof. Let $\mu \leq \check{w}$. By (3.2) and the definition of \check{w} , we have $D_j^s(N, \underline{v}, \mu) = D_j^s(N, \underline{u}, \mu)$ for all $s \in \mathcal{S}$, $1 \leq j \leq d$. This implies that $\underline{H}(N, \underline{v}, \mu) = \underline{H}(N, \underline{u}, \mu)$, proving the first claim.

For the second claim, let $\tilde{\mu}_u$ and $\tilde{\mu}_v$ be as defined above. Let $\tilde{\mu}_u \leq \check{w}$. Then $\underline{H}(N, \underline{v}, \tilde{\mu}_u) = \underline{H}(N, \underline{u}, \tilde{\mu}_u) = 0$, so $\mu = \tilde{\mu}_u$ is a solution to $\underline{H}(N, \underline{v}, \mu) = 0$. By Theorem 3.5, this solution is unique. By a symmetric argument, if $\tilde{\mu}_v \leq \check{w}$, then $\mu = \tilde{\mu}_v$ is the unique solution to $\underline{H}(N, \underline{u}, \mu) = 0$. \square

4. Experiments. We conduct experiments to show numerical evidence that the result of FMM converges to the viscosity solution of (1.1), to demonstrate types of anisotropic problems that can be solved, and to determine the effectiveness of the node and simplex elimination techniques described in Appendix A. Throughout this section, the boundary conditions are $g(x) = 0$ for $x \in \partial\Omega$. For all experiments below, excluding that in section 4.4, we discretize $[-1, 1]^d$ such that there are m uniformly spaced nodes in each dimension, and we ensure that there is a node at the origin O .

4.1. Convergence study. We examine the difference between the solution to (3.3) and the solution to (1.1) for two simple Dirichlet problems. In particular, we look at how the absolute error changes as the grid spacing decreases toward zero. For the problems considered, $\Omega = [-1, 1]^d \setminus \{O\}$. We take H to have the form in (2.2), where $G(Du(x)) = \|Du(x)\|_p$ and $p = 1$ or $p = 2$. The boundary conditions are $g(O) = 0$. We use the analytic node value update equations provided in Appendix B.

Since there is a node at O , any error introduced is from the discretization of H and not from the discretization of the boundary condition. The approximation errors are summarized in Table 4.1.

4.2. Asymmetric anisotropic problem. For this anisotropic problem, H is as in (2.2), where G is defined by (2.5) (see Figure 2.3(b)). The domain is given by $\Omega = [-1, 1]^2 \setminus \{O\}$ and $\partial\Omega = \{O\}$. The cost is $c(x) = 1$, except in four rectangular regions shown in black in Figure 4.1, where $c(x) \gg 1$. In the `Update` function, we

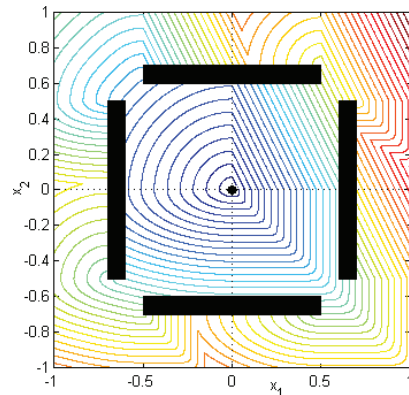


FIG. 4.1. Contours of \underline{u} computed for the anisotropic problem where Hamiltonian H is as in (2.2) and G is as in (2.5). The black circle at $O = (0, 0)$ indicates $\partial\Omega$, and in the black rectangles $c(x) \gg 1$. In these regions, \underline{u} has purposefully not been computed.

analytically computed the solution to (3.4) using the equations for updating from a single simplex given in Appendix B. The number of nodes in each dimension is $m = 1281$. We plot the contours of \underline{u} computed by FMM in Figure 4.1. Note the asymmetric contours where the characteristics bend through gaps. The relationship between the shape of the contours of G in Figure 2.3(b) and those of \underline{u} is explained by the duality articulated in Proposition 2.7 of [1].

4.3. Anelliptic elastic wave propagation. As is done in [10], we consider elastic wave propagation in VTI media, which are transversely isotropic media with a vertical axis of symmetry. In particular, we wish to find the arrival times of quasi-longitudinal (quasi-P or qP) waves propagating in two dimensions from a point source at the origin O . We solve the anisotropic HJ PDE given by defining the Hamiltonian

$$(4.1) \quad H(q) = \frac{1}{2}(q_1^2 + q_2^2) \left\{ (a+l)q_1^2 + (c+l)q_2^2 + \sqrt{[(a-l)q_1^2 - (c-l)q_2^2]^2 + 4(f+l)^2 q_1^2 q_2^2} \right\} - 1.$$

This Hamiltonian is derived from the anisotropic Eikonal equation and the exact qP-wave phase velocity equation in [10]. The parameters $a = 14.47$, $l = 2.28$, $c = 9.57$, and $f = 4.51$ are taken from [10].

The Hamiltonian H and the approximate solution \underline{u} resulting from FMM are shown in Figure 4.2. We have not shown analytically that (4.1) satisfies strict one-sided monotonicity for some range of parameters. However, the level sets of H as shown in Figure 4.2(a) indicate that H is strictly one-sided monotone for the given parameters. Furthermore, the level sets of H indicate that H is convex and a computation of the derivative of H using the symbolic mathematics program Maple shows that H satisfies Osher's criterion for the given parameters. As a result, the analysis in this paper can be applied to the problem, and FMM can be used to compute the solution.

We used a grid of size 201×201 . In the `Update` function, we used the interval method to solve (3.4) numerically. We computed the maximum relative error of \underline{u} to be 0.0076 when compared to the travel-time computed with the group-velocity

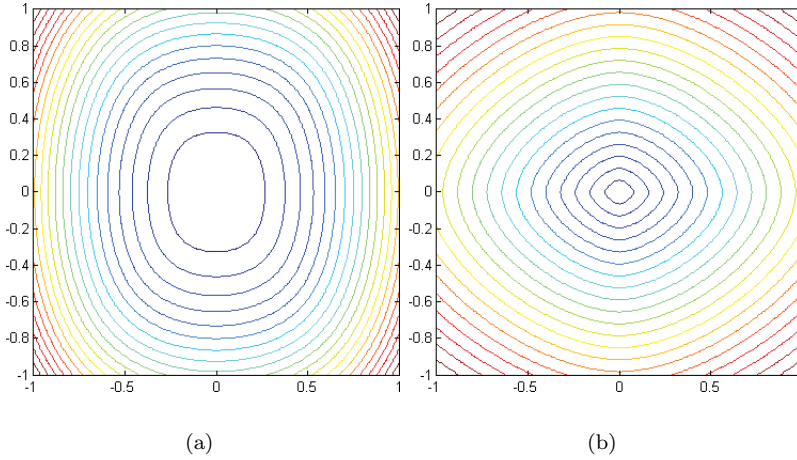


FIG. 4.2. Using FMM for computation of travel times of qP-waves in two-dimensional VTI media. (a) Contours of Hamiltonian H as given in (4.1). (b) Contours of approximate solution \underline{u} computed by FMM.

approximation for qP-waves presented in [10]. In turn, the group-velocity approximation is claimed to have a maximum error of about 0.003 when compared to the true solution.

4.4. Two robots. We consider the two-robot coordinated navigation problem illustrated in Figure 4.3. The circular robots are free to move independently in a two-dimensional plane but may not collide with each other or the obstacles (black region). Each may travel at a maximum speed of $1/c(x)$ in any direction. The robots attempt to achieve a joint goal state. This goal should be achieved in minimal time from any initial state in the domain without incurring collisions.

Let the state of the dark-colored robot be $(x_1, x_2) \in \mathbb{R}^2$ and the state of the light-colored robot be $(x_3, x_4) \in \mathbb{R}^2$ so that the combined state of the two robots is $(x_1, x_2, x_3, x_4) \in \mathbb{R}^4$. We define the control-theoretic action set

$$\mathcal{A}(x) = \{a \mid F(a) = \|(\|a_1, a_2\|_2, \|a_3, a_4\|_2)\|_\infty \leq 1/c(x)\}.$$

Proposition 2.7 of [1] states that we can use the dual of F to obtain

$$(4.2) \quad G(x, Du(x)) = \|(\|(\partial_1 u(x), \partial_2 u(x))\|_2, \|(\partial_3 u(x), \partial_4 u(x))\|_2)\|_1,$$

where $Du(x) = (\partial_1 u(x), \partial_2 u(x), \partial_3 u(x), \partial_4 u(x))$. Where x is a collision state, we set $c(x) \gg 1$. For all other states x , $c(x) = 1$.

We can compute \underline{u} using FMM since G is a mixed p -norm, and thus H satisfies Properties 1 to 4 (see section 2.2). The domain Ω is discretized using a uniform orthogonal grid of $(81 \times 21)^2$ nodes. The discretization of (4.2) is quartic in \underline{u}_0 , so it is difficult to solve analytically. However, Theorem 3.5 tells us that we can determine the solution to (3.4) uniquely. As a result, numerical root-finders can easily be used to compute this solution in the `Update` function. Once an approximation of u is generated by FMM, a gradient descent algorithm is used to find optimal paths [3, 2]. The optimal trajectories from a single starting condition are shown in Figure 4.3.

4.5. Efficient implementation. Appendix A describes three different methods for improving the efficiency of the `Update` function: symmetry, causality, and solution

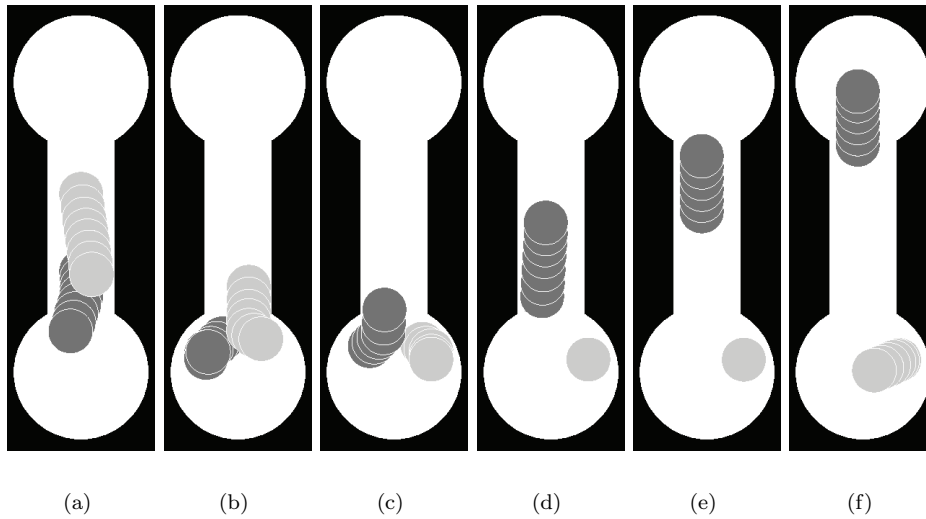


FIG. 4.3. *Two-robot coordinated optimal navigation problem. The joint goal is for the dark-colored robot to reach the center of the upper bulb and the light-colored robot to reach the center of the lower bulb. Black indicates an obstacle region. The sequence shows the robots achieving their joint goal without collision from a particular initial state. The solution of (1.1), where H is given by (2.2) and G is given by (4.2) allows quick determination of the optimal collision-free trajectories for both robots from any initial condition [3].*

elimination. Some of these methods are related to those found in [15, 31]. However, experimental results indicate that the efficiency gains from using these methods are not substantial for an already efficient implementation of FMM. In such an implementation the `Update` function computes only updates from those nodes that have already been extracted from \mathcal{Q} using the `ExtractMin` function. Also, only simplices that include the most-recently extracted node \underline{y} are considered in `Update`. Our experiments show that in many calls to `Update`, only a single simplex fits these criteria, and the fraction of updates for which only a single simplex fits the criteria grows as the grid is refined. For this reason, further techniques for eliminating nodes and simplices, such as those described in Appendix A, are largely ineffective.

However, for coarse grid resolutions and problems where characteristics intersect often, multiple simplices are considered by `Update` frequently enough that symmetry elimination, which is very cheap, significantly improves efficiency. In some cases, a node value update can be ignored altogether if the most-recently extracted node is eliminated by symmetry.

Despite the fact that the node and simplex elimination techniques described in Appendix A are useful only in limited circumstances, we include them for theoretical interest and because they may be applied in other algorithms, such as sweeping methods, that also require the `Update` function.

5. Conclusion. We have described a new class of static HJ PDEs with axis-aligned but potentially asymmetric anisotropy. Assuming Properties 1 to 4 of the Hamiltonian, we showed that uniqueness, monotonicity, and causality hold for a standard finite-difference discretization of these PDEs on an orthogonal grid, and so the FMM can be used to approximate their solution. In the appendix, we also demonstrate several methods for reducing the number of neighboring simplices which must

be considered when computing node updates, including novel methods which work when the PDE and/or grid are asymmetric. In future work, these results might be generalized to unstructured grids.

Appendix A. Efficient Implementation of Update. We discuss ways to improve the efficiency of the `Update` function, which calculates the unique solution $\mu = \tilde{\mu}$ to (3.4). We note that these improvements may be used for any type of solution method, including FMM and sweeping methods, as long as (3.4) is being solved. Some efficiency improvements are related to similar ideas specific to the isotropic Eikonal equation found in [15, 31].

Efficiency can be gained by determining which neighbors $\underline{x} \in \mathcal{N}(\underline{x}_0)$ have no influence on the solution and eliminating them from consideration. Let

$$\sigma = (\sigma_1, \sigma_2, \dots, \sigma_d),$$

where $\sigma_j \subseteq \{\pm 1\}$, indicate which $\underline{x} \in \mathcal{N}$ are considered in determining the solution $\mu = \tilde{\mu}$. Let \mathcal{N}_σ be the reduced set of neighbor nodes defined by σ . Let \mathcal{S}_σ be the set of neighboring simplices that can be formed by the neighbors in \mathcal{N}_σ . For example, in $d = 4$ dimensions, take

$$\sigma = (\emptyset, \{\pm 1\}, \{-1\}, \{\pm 1\}).$$

We have

$$\begin{aligned} \mathcal{N}_\sigma &= \{\underline{x}_2^\pm, \underline{x}_3^-, \underline{x}_4^\pm\} \quad \text{and} \\ \mathcal{S}_\sigma &= \{(0, -1, -1, -1), (0, +1, -1, -1), (0, -1, -1, +1), (0, +1, -1, +1)\}. \end{aligned}$$

Let $\underline{H}_\sigma(N, \phi, \mu) = \max_{s \in \mathcal{S}_\sigma} [H(\underline{x}_0, \underline{D}^s(N, \phi, \mu))]$ be the *reduced-neighbor numerical Hamiltonian*, a modification of (3.1) that considers only the neighbors and simplices indicated by σ . For $s \in \mathcal{S}_\sigma$ and $1 \leq j \leq d$, $s_j = 0$ indicates that $\underline{x}_j^{s_j}$ is not considered in computing the gradient approximation $D^s \underline{u}(\mu)$; that is, $\underline{D}_j^s(N, \phi, \mu) = 0$ if $s_j = 0$, and \underline{D}_j^s satisfies (3.2) otherwise.

To implement `Update`, we first reduce the set of considered neighbors and then solve

$$(A.1) \quad \underline{H}_\sigma(N(\underline{x}_0), \underline{u}, \mu) = 0$$

for $\mu = \tilde{\mu}$ to determine a node’s value $\underline{u}(\underline{x}_0)$. As in section 3, we may write $\underline{H}_\sigma(\mu) = \underline{H}_\sigma(N, \phi, \mu)$ and $\underline{D}^s(\mu) = \underline{D}^s(N, \phi, \mu)$, where no ambiguity results. Note that some properties of (A.1) are retained from (3.4) as long as at least one considered neighbor remains in σ . Let

$$\check{\underline{u}}_\sigma = \min_{\underline{x} \in \mathcal{N}_\sigma} (\underline{u}(\underline{x})).$$

PROPOSITION A.1 (analogue of Lemma 3.3). $\underline{H}_\sigma(\mu)$ is strictly increasing on $\mu \geq \check{\underline{u}}_\sigma$.

PROPOSITION A.2 (analogue of Lemma 3.4). The numerical Hamiltonian $\underline{H}_\sigma(\mu)$ satisfies the following:

- (a) $\underline{H}_\sigma(\mu) = H(0) < 0$ for $\mu \leq \check{\underline{u}}_\sigma$.
- (b) $\underline{H}_\sigma(\mu) \rightarrow \infty$ as $\mu \rightarrow \infty$.
- (c) $\underline{H}_\sigma(\mu)$ is nondecreasing on all μ .

PROPOSITION A.3 (analogue of Theorem 3.5). There exists a unique solution $\mu = \tilde{\mu}$ to $\underline{H}_\sigma(\mu) = 0$ such that $\tilde{\mu} > \check{\underline{u}}_\sigma$.

A.1. Symmetry. We show how the considered neighbors σ can be reduced by keeping only the neighbor with the smaller value of a pair of opposite neighbors in the j th dimension when (3.1) is symmetric in that dimension. This procedure is a generalization of those in [15, 31] to all axis-aligned anisotropic problems on unequally spaced grids. First, we introduce useful notation.

Let $q \in \mathbb{R}^d$. Let $T^i(q)$ be a reflection of q in the hyperplane orthogonal to the i th axis, such that

$$T_j^i(q) = \begin{cases} -q_j, & \text{if } j = i, \\ q_j, & \text{otherwise,} \end{cases}$$

for $1 \leq j \leq d$. Let Ψ_j indicate symmetry of (3.1) in the j th dimension, as follows:

$$\Psi_j = \begin{cases} 1, & \text{if } |h_j^-| = |h_j^+| \text{ and for all } q \in \mathbb{R}^d, H(q) = H(T^j(q)), \\ 0, & \text{otherwise.} \end{cases}$$

In other words, $\Psi_j = 1$ if and only if the grid spacing and H are symmetric in the j th dimension. The following theorem is proved in [2].

THEOREM A.4. *Let σ be such that $\sigma_j \subseteq \{\pm 1\}$ for $1 \leq j \leq d$. Let $\tilde{\sigma}$ be defined by*

$$\tilde{\sigma}_j = \begin{cases} \{-1\}, & \text{if } \sigma_j = \{\pm 1\}, \Psi_j = 1, \text{ and } \underline{u}_j^- \leq \underline{u}_j^+, \\ \{+1\}, & \text{if } \sigma_j = \{\pm 1\}, \Psi_j = 1, \text{ and } \underline{u}_j^- > \underline{u}_j^+, \\ \sigma_j, & \text{otherwise,} \end{cases}$$

for $1 \leq j \leq d$. Let $\mu = \mu_\sigma$ be the unique solution to $\underline{H}_\sigma(\mu) = 0$. Let $\mu = \mu_{\tilde{\sigma}}$ be the unique solution to $\underline{H}_{\tilde{\sigma}}(\mu) = 0$. Then $\mu_{\tilde{\sigma}} = \mu_\sigma$.

An implementation of the **Update** function can use the result obtained in Theorem A.4 to eliminate $\underline{x} \in \mathcal{N}$ from consideration in solving (A.1) by exploiting symmetries in (3.1). We call this *symmetry elimination*.

Remark 2. Theorem A.4 can be generalized to an asymmetric version. We let $1 \leq j \leq d$, and let $s_j, \tilde{s}_j \in \{\pm 1\}$ such that $s_j \neq \tilde{s}_j$. Node $\underline{x}_j^{s_j} \in \mathcal{N}$ may be eliminated from consideration if

- $|h_j^{\tilde{s}_j}| \leq |h_j^{s_j}|$;
- for all $q \in \mathbb{R}^d$ such that $s_j q_j \geq 0$, $H(q) \leq H(T^j(q))$;
- and $\underline{u}_j^{\tilde{s}_j} \leq \underline{u}_j^{s_j}$.

A.2. Causality. The causality of (3.1) can also be exploited to eliminate $\underline{x} \in \mathcal{N}_\sigma$ from consideration. This observation was used in two distinct but equivalent methods for analytically computing the **Update** from a single simplex to solve an isotropic Eikonal equation [15, 31]. We show with the following theorem that the condition $\underline{H}_\sigma(\underline{u}(\underline{x})) \geq 0$ can be checked to determine that a node \underline{x} is noncausal, i.e., that the solution $\mu = \mu_\sigma$ to (A.1) is not dependent on the node \underline{x} and its value $\underline{u}(\underline{x})$.

THEOREM A.5. *Let σ be such that $\sigma_j \subseteq \{\pm 1\}$ for $1 \leq j \leq d$. Pick any $s \in \mathcal{S}_\sigma$ and $i \in \{1, 2, \dots, d\}$ such that $s_i \neq 0$ and $\underline{H}_\sigma(\underline{u}_i^{s_i}) \geq 0$. Let $\tilde{\sigma}$ be defined by*

$$\tilde{\sigma}_j = \begin{cases} \sigma_j \setminus \{s_j\}, & \text{if } j = i, \\ \sigma_j, & \text{otherwise.} \end{cases}$$

Let $\mu = \mu_\sigma$ be the unique solution to $\underline{H}_\sigma(\mu) = 0$. Let $\mu = \mu_{\tilde{\sigma}}$ be the unique solution to $\underline{H}_{\tilde{\sigma}}(\mu) = 0$. Then $\mu_{\tilde{\sigma}} = \mu_\sigma$.

Proof. Let $\sigma, s, i, \tilde{\sigma}, \mu_\sigma,$ and $\mu_{\tilde{\sigma}}$ be as defined above. By Proposition A.2(c), $\underline{H}_\sigma(\mu)$ is nondecreasing. Since $\underline{H}_\sigma(\underline{u}_i^{s_i}) \geq 0 = \underline{H}_\sigma(\mu_\sigma)$, it must be that $\mu_\sigma \leq \underline{u}_i^{s_i}$. Note that $\underline{H}_{\tilde{\sigma}}(\mu)$ is identical to $\underline{H}_\sigma(\mu)$ except for $D_i^s \underline{u}(\mu)$, which is set to zero in $\underline{H}_{\tilde{\sigma}}(\mu)$. But for $\mu \leq \underline{u}_i^{s_i}$, we also have $D_i^s \underline{u}(\mu) = 0$ in $\underline{H}_\sigma(\mu)$. Consequently, $\underline{H}_{\tilde{\sigma}}(\mu) = \underline{H}_\sigma(\mu)$ for $\mu \leq \underline{u}_i^{s_i}$. In particular, $\underline{H}_{\tilde{\sigma}}(\mu_\sigma) = \underline{H}_\sigma(\mu_\sigma) = 0$. Therefore, $\mu_{\tilde{\sigma}} = \mu_\sigma$. \square

Theorem A.5 states that the unique solution μ to (A.1) does not change when a noncausal node is removed from σ . This node removal can be repeated until all noncausal nodes have been removed, and the solution $\mu = \mu_\sigma$ will remain unchanged. We call this *causality elimination*. A binary or linear search through sorted neighbors' values can be used to determine the largest node value that might be causal. Note that causality elimination does not require symmetry in (3.1). However, the test for noncausality requires an evaluation of \underline{H}_σ , which is more expensive than the comparison of two neighbors' values used for symmetry elimination.

A.3. Solution. After eliminating from consideration nodes in σ using symmetry and causality elimination, we can determine the solution $\mu = \tilde{\mu}$ to (A.1). Let

$$(A.2) \quad \tilde{\mu} = \min_{s \in \mathcal{S}_\sigma} (\mu_s),$$

where $\mu = \mu_s$ is the unique solution to

$$(A.3) \quad H(D^s \underline{u}(\mu)) = 0.$$

We show with the following proposition that, instead of solving (A.1) directly, we can solve (A.3) for each $s \in \mathcal{S}_\sigma$ and take the minimum such solution $\tilde{\mu}$. It can be shown that $H(D^s \underline{u}(\mu))$ is continuous and nondecreasing on μ and that (A.3) has a unique solution in an analogous but simpler manner as the proof of Theorem 3.5.

PROPOSITION A.6. *Let $\hat{\mu}$ be the unique solution to (A.1). Then $\hat{\mu} = \tilde{\mu}$.*

Proof. Let $\mu_s, \tilde{\mu},$ and $\hat{\mu}$ be as defined above. For any $s \in \mathcal{S}_\sigma$, we know $\mu_s \geq \tilde{\mu}$. Since $H(D^s \underline{u}(\mu))$ is nondecreasing on μ , it must be that $H(D^s \underline{u}(\mu)) \leq H(D^s \underline{u}(\mu_s)) = 0$ for all $\mu \leq \mu_s$. In particular, $H(D^s \underline{u}(\tilde{\mu})) \leq 0$. Furthermore, by the definition of $\tilde{\mu}$, there exists an $\tilde{s} \in \mathcal{S}_\sigma$ such that $H(D^{\tilde{s}} \underline{u}(\tilde{\mu})) = 0$. Consequently,

$$(A.4) \quad \underline{H}_\sigma(\tilde{\mu}) = \max_{s \in \mathcal{S}_\sigma} H(D^s \underline{u}(\tilde{\mu})) = 0.$$

Therefore, $\hat{\mu} = \tilde{\mu}$ solves (A.1), and it is a unique solution by Proposition A.3. \square

We further show that we may be able to determine $\tilde{\mu}$ without solving (A.3) for each $s \in \mathcal{S}_\sigma$. We demonstrate using the following proposition that if we have computed a solution $\mu = \mu_s$ of (A.3) for some $s \in \mathcal{S}_\sigma$, we can easily determine if $\mu_{\tilde{s}} \geq \mu_s$, where $\mu = \mu_{\tilde{s}}$ is the solution to $H(D^{\tilde{s}} \underline{u}(\mu)) = 0$ for some other $\tilde{s} \in \mathcal{S}_\sigma$. Note we do not necessarily need to compute $\mu_{\tilde{s}}$ to rule it out as a minimal solution.

PROPOSITION A.7. *Let $s \in \mathcal{S}_\sigma$ and $\tilde{s} \in \mathcal{S}_\sigma$. Let $\mu = \mu_s$ be the unique solution to $H(D^s \underline{u}(\mu)) = 0$ and $\mu = \mu_{\tilde{s}}$ be the unique solution to $H(D^{\tilde{s}} \underline{u}(\mu)) = 0$. Then $\mu_{\tilde{s}} < \mu_s$ if and only if $H(D^{\tilde{s}} \underline{u}(\mu_s)) > H(D^s \underline{u}(\mu_s))$.*

Proof. Let μ_s and $\mu_{\tilde{s}}$ be as defined above. If $H(D^{\tilde{s}} \underline{u}(\mu_s)) > H(D^s \underline{u}(\mu_s)) = 0$, then the unique solution $\mu = \mu_{\tilde{s}}$ to $H(D^{\tilde{s}} \underline{u}(\mu)) = 0$ must be such that $\mu_{\tilde{s}} < \mu_s$, since $H(D^{\tilde{s}} \underline{u}(\mu))$ is nondecreasing on μ . Similarly, if $H(D^{\tilde{s}} \underline{u}(\mu_s)) \leq H(D^s \underline{u}(\mu_s))$, then the unique solution $\mu = \mu_{\tilde{s}}$ to $H(D^{\tilde{s}} \underline{u}(\mu)) = 0$ must be such that $\mu_{\tilde{s}} \geq \mu_s$. \square

The result of Proposition A.7 can be used to eliminate simplices $s \in \mathcal{S}_\sigma$ for which solutions to (A.3) are irrelevant to the computation. We call this process *solution elimination*.

Appendix B. Analytic solutions. We provide analytic node value update equations for the cases where H is given by (2.2), where $G(Du(x)) = \|Du(x)\|_p$ and $p = 1$, $p = 2$, or $p = \infty$. In these cases, there is an exact solution to (3.4). For derivations of these equations, see [2]. The equation for $p = 2$ fixes some errors in the appendix of [16]. In [3] we demonstrated that these cases could be treated by FMM and are useful for robotic applications. However, here we generalize the update equations to any dimension and grid spacing.

Let (v_1, v_2, \dots, v_m) be the values of the neighboring nodes in the simplex $s \in \mathcal{S}_\sigma$ and (h_1, h_2, \dots, h_m) be the corresponding grid spacings. We are solving for μ . In order to use the analytic updates below, noncausal node values must already have been eliminated using causality elimination, so $\mu > \max_{1 \leq j \leq m} v_j$. However, in the case of the efficient implementation of FMM discussed in section 4.5, any nodes that would be removed from consideration by causality elimination could not already have been extracted from \mathcal{Q} , and so the analytic updates below can be applied directly.

The update formula for $p = 1$ is

$$\mu = \frac{\sum_j \left(\prod_{l \neq j} h_l \right) v_j + \prod_l h_l c}{\sum_j \prod_{l \neq j} h_l}.$$

The update formula for $p = 2$ is

$$\mu = \frac{\sum_j \left(\prod_{l \neq j} h_l^2 \right) v_j + \prod_l h_l \sqrt{\left(\sum_j \prod_{l \neq j} h_l^2 \right) c^2 - \sum_{j_1} \sum_{j_2 > j_1} \left(\prod_{l \neq j_1, j_2} h_l^2 \right) (v_{j_1} - v_{j_2})^2}}{\sum_j \prod_{l \neq j} h_l^2}.$$

The update formula for $p = \infty$ is

$$\mu = \min_j (v_j + h_j c).$$

The $p = \infty$ case is identical to the update formula for Dijkstra's algorithm for shortest path on a discrete graph.

Acknowledgments. We would like to thank Alexander Vladimirsky for enlightening discussions about FMM and OUM and Adam Oberman for discussions and insights regarding [4]. Furthermore, we wish to thank the reviewers for constructive and insightful comments.

REFERENCES

- [1] K. ALTON AND I. M. MITCHELL, *Fast Marching Methods for a Class of Anisotropic Stationary Hamilton-Jacobi Equations*, Technical report TR-2006-27, Department of Computer Science, University of British Columbia, Vancouver, 2007.
- [2] K. ALTON AND I. M. MITCHELL, *Fast Marching Methods for Hamilton-Jacobi Equations with Axis-aligned Anisotropy*, Technical report TR-2008-02, Department of Computer Science, University of British Columbia, Vancouver, 2008.
- [3] K. ALTON AND I. MITCHELL, *Optimal path planning under different norms in continuous state spaces*, in Proceedings of the International Conference on Robotics and Automation, IEEE, 2006, pp. 866–872.
- [4] G. BARLES AND P. E. SOUGANIDIS, *Convergence of approximation schemes for fully nonlinear second order equations*, *Asymptot. Anal.*, 4 (1991), pp. 271–283.
- [5] F. BORNEMANN AND C. RASCH, *Finite-element discretization of static Hamilton-Jacobi equations based on a local variational principle*, *Comput. Vis. Sci.*, 9 (2006), pp. 57–69.

- [6] M. BOUE AND P. DUPUIS, *Markov chain approximations for deterministic control problems with affine dynamics and quadratic cost in the control*, SIAM J. Numer. Anal., 36 (1999), pp. 667–695.
- [7] M. G. CRANDALL, H. ISHII, AND P. LIONS, *User’s guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.
- [8] P.-E. DANIELSSON, *Euclidean distance mapping*, Comput. Graph. Image Process., 14 (1980), pp. 227–248.
- [9] E. W. DIJKSTRA, *A note on two problems in connection with graphs*, Numer. Math., 1 (1959), pp. 269–271.
- [10] S. FOMEL, *On anelliptic approximations for qp velocities in vti media*, Geophys. Prospecting, 52 (2004), pp. 247–259.
- [11] P. A. GREMAUD AND C. M. KUSTER, *Computational study of fast methods for the Eikonal equation*, SIAM J. Sci. Comput., 27 (2006), pp. 1803–1816.
- [12] S.-R. HYSING AND S. TUREK, *The Eikonal equation: Numerical efficiency vs. algorithmic complexity on quadrilateral grids*, in Proceedings of Algoritmy 2005, Vsake Tatry, Pobanske, Slovakia, 2005, pp. 22–31.
- [13] C. Y. KAO, S. OSHER, AND Y. H. TSAI, *Fast sweeping methods for static Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 42 (2005), pp. 2612–2632.
- [14] R. KIMMEL AND J. A. SETHIAN, *Fast marching methods on triangulated domains*, Proc. Natl. Acad. Sci. USA, 95 (1998), pp. 8341–8435.
- [15] R. KIMMEL AND J. A. SETHIAN, *Optimal algorithm for shape from shading and path planning*, J. Math. Imaging Vision, 14 (2001), pp. 237–244.
- [16] I. M. MITCHELL AND S. SASTRY, *Continuous Path Planning with Multiple Constraints*, Technical report UCB/ERL M03/34, UC Berkeley Engineering Research Laboratory, Berkeley, CA, 2003.
- [17] S. OSHER AND R. P. FEDKIW, *Level set methods: An overview and some recent results*, J. Comput. Phys., 169 (2001), pp. 463–502.
- [18] L. C. POLYMENAKOS, D. P. BERTSEKAS, AND J. N. TSITSIKLIS, *Implementation of efficient algorithms for globally optimal trajectories*, IEEE Trans. Automat. Control, 43 (1998), pp. 278–283.
- [19] J. QIAN, Y. ZHANG, AND H. ZHAO, *A fast sweeping method for static convex Hamilton–Jacobi equations*, J. Sci. Comput., 31 (2007), pp. 237–271.
- [20] J. A. SETHIAN AND A. VLADIMIRSKY, *Fast methods for Eikonal and related Hamilton–Jacobi equations on unstructured meshes*, Proc. Natl. Acad. Sci. USA, 97 (2000), pp. 5699–5703.
- [21] J. A. SETHIAN AND A. VLADIMIRSKY, *Ordered upwind methods for static Hamilton–Jacobi equations*, Proc. Natl. Acad. Sci. USA, 98 (2001), pp. 11069–11074.
- [22] J. A. SETHIAN AND A. VLADIMIRSKY, *Ordered upwind methods for static Hamilton–Jacobi equations: Theory and algorithms*, SIAM J. Numer. Anal., 41 (2003), pp. 325–363.
- [23] J. A. SETHIAN, *A fast marching level set method for monotonically advancing fronts*, Proc. Natl. Acad. Sci. USA, 93 (1996), pp. 1591–1595.
- [24] J. A. SETHIAN, *Level Set Methods*, Cambridge University Press, Cambridge, UK, 1996.
- [25] J. A. SETHIAN, *Fast marching methods*, SIAM Rev., 41 (1999), pp. 199–235.
- [26] J. A. SETHIAN, *Level Set Methods and Fast Marching Methods*, Cambridge University Press, Cambridge, UK, 1999.
- [27] Y.-H. R. TSAI, L.-T. CHENG, S. OSHER, AND H.-K. ZHAO, *Fast sweeping algorithms for a class of Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 41 (2003), pp. 673–694.
- [28] J. N. TSITSIKLIS, *Efficient algorithms for globally optimal trajectories*, in Proceedings of the 33rd IEEE Conference on Decision and Control, Lake Buena Vista, FL, 1994, pp. 1368–1373.
- [29] J. N. TSITSIKLIS, *Efficient algorithms for globally optimal trajectories*, IEEE Trans. Automat. Control, 40 (1995), pp. 1528–1538.
- [30] L. YATZIV, A. BARTESAGHI, AND G. SAPIRO, *O(n) implementation of the fast marching method*, J. Comput. Phys., 212 (2006), pp. 393–399.
- [31] H. ZHAO, *A fast sweeping method for Eikonal equations*, Math. Comp., 74 (2004), pp. 603–627.

NEW INTERIOR PENALTY DISCONTINUOUS GALERKIN METHODS FOR THE KELLER–SEGEL CHEMOTAXIS MODEL*

YEKATERINA EPSHTEYN[†] AND ALEXANDER KURGANOV[‡]

Abstract. We develop a family of new interior penalty discontinuous Galerkin methods for the Keller–Segel chemotaxis model. This model is described by a system of two nonlinear PDEs: a convection-diffusion equation for the cell density coupled with a reaction-diffusion equation for the chemoattractant concentration. It has been recently shown that the convective part of this system is of a mixed hyperbolic–elliptic-type, which may cause severe instabilities when the studied system is solved by straightforward numerical methods. Therefore, the first step in the derivation of our new methods is made by introducing the new variable for the gradient of the chemoattractant concentration and by reformulating the original Keller–Segel model in the form of a convection-diffusion-reaction system with a hyperbolic convective part. We then design interior penalty discontinuous Galerkin methods for the rewritten Keller–Segel system. Our methods employ the central-upwind numerical fluxes, originally developed in the context of finite-volume methods for hyperbolic systems of conservation laws. In this paper, we consider Cartesian grids and prove error estimates for the proposed high-order discontinuous Galerkin methods. Our proof is valid for pre-blow-up times since we assume boundedness of the exact solution. We also show that the blow-up time of the exact solution is bounded from above by the blow-up time of our numerical solution. In the numerical tests presented below, we demonstrate that the obtained numerical solutions have no negative values and are oscillation-free, even though no slope-limiting technique has been implemented.

Key words. Keller–Segel chemotaxis model, convection-diffusion-reaction systems, discontinuous Galerkin methods, nonsymmetric interior penalty Galerkin, incomplete interior penalty Galerkin, and symmetric interior penalty Galerkin methods, Cartesian meshes

AMS subject classifications. 65M60, 65M12, 65M15, 92C17, 35K57

DOI. 10.1137/07070423X

1. Introduction. The goal of this work is to design new discontinuous Galerkin (DG) methods for the two-dimensional Keller–Segel chemotaxis model [13, 29, 30, 31, 36, 38]. The DG methods have recently become increasingly popular thanks to their attractive features such as local, elementwise mass conservation; flexibility to use high-order polynomial and nonpolynomial basis functions; ability to easily increase the order of approximation on each mesh element independently; ability to achieve almost an exponential convergence rate when smooth solutions are captured on appropriate meshes; block diagonal mass matrices, which are of great computational advantage if an explicit time integration is used; suitability for parallel computations due to (relatively) local data communications; applicability to problems with discontinuous coefficients and/or solutions. The DG methods have been successfully applied to a wide variety of problems ranging from solid mechanics to fluid mechanics (see, e.g., [3, 7, 14, 15, 17, 21, 23, 41] and references therein).

*Received by the editors October 1, 2007; accepted for publication (in revised form) August 26, 2008; published electronically November 26, 2008.

<http://www.siam.org/journals/sinum/47-1/70423.html>

[†]Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 (rina10@andrew.cmu.edu). The research of this author is based upon work supported by the Center for Nonlinear Analysis (CNA) under NSF grant DMS-0635983.

[‡]Mathematics Department, Tulane University, New Orleans, LA 70118 (kurganov@math.tulane.edu). The research of this author was supported in part by NSF grant DMS-0610430.

In this paper, we consider the most common formulation of the Keller–Segel system [13], which can be written in the dimensionless form as

$$(1.1) \quad \begin{cases} \rho_t + \nabla \cdot (\chi \rho \nabla c) = \Delta \rho, \\ c_t = \Delta c - c + \rho, \end{cases} \quad (x, y) \in \Omega, \quad t > 0,$$

subject to the Neumann boundary conditions:

$$\nabla \rho \cdot \mathbf{n} = \nabla c \cdot \mathbf{n} = 0, \quad (x, y) \in \partial\Omega.$$

Here, $\rho(x, y, t)$ is the cell density, $c(x, y, t)$ is the chemoattractant concentration, χ is a chemotactic sensitivity constant, Ω is a bounded domain in \mathbb{R}^2 , $\partial\Omega$ is its boundary, and \mathbf{n} is a unit normal vector.

It is well-known that solutions of this system may blow up in finite time; see, e.g., [27, 28] and references therein. This blow up represents a mathematical description of a cell concentration phenomenon that occurs in real biological systems; see, e.g., [1, 8, 10, 11, 16, 39].

Capturing blowing up solutions numerically is a challenging problem. Finite-volume [22] and finite element [35] methods have been proposed for a simpler version of the Keller–Segel model,

$$\begin{cases} \rho_t + \nabla \cdot (\chi \rho \nabla c) = \Delta \rho, \\ \Delta c - c + \rho = 0, \end{cases}$$

in which the equation for concentration c has been replaced by an elliptic equation using an assumption that the chemoattractant concentration c changes over much smaller time scales than the density ρ . A fractional step numerical method for a fully time-dependent chemotaxis system from [42] has been proposed in [43]. However, the operator-splitting approach may not be applicable when a convective part of the chemotaxis system is not hyperbolic, which is a generic situation for the original Keller–Segel model as it was shown in [12], where the finite-volume Godunov-type central-upwind scheme was derived for (1.1) and extended to some other chemotaxis and haptotaxis models.

The starting point in the derivation of the central-upwind scheme in [12] was rewriting the original system (1.1) in an equivalent form, in which the concentration equation is replaced with the corresponding equation for the gradient of c :

$$\begin{cases} \rho_t + \nabla \cdot (\chi \rho \mathbf{w}) = \Delta \rho, \\ \mathbf{w}_t - \nabla \rho = \Delta \mathbf{w} - \mathbf{w}, \end{cases} \quad \mathbf{w} \equiv (u, v) := \nabla c.$$

This form can be considered as a convection-diffusion-reaction system

$$(1.2) \quad \mathbf{U}_t + \mathbf{f}(\mathbf{U})_x + \mathbf{g}(\mathbf{U})_y = \Delta \mathbf{U} + \mathbf{r}(\mathbf{U}),$$

where $\mathbf{U} := (\rho, u, v)^T$, $\mathbf{f}(\mathbf{U}) := (\chi \rho u, -\rho, 0)^T$, $\mathbf{g}(\mathbf{U}) := (\chi \rho v, 0, -\rho)^T$, and $\mathbf{r}(\mathbf{U}) := (0, -u, -v)^T$. The system (1.2) is an appropriate form of the chemotaxis system if one wants to solve it numerically by a finite-volume method. Even though the convective part of the system (1.2) is not hyperbolic, some stability of the resulting central-upwind scheme was ensured by proving its positivity-preserving property; see [12].

A major disadvantage of the system (1.2) is a mixed type of its convective part. When a high-order numerical method is applied to (1.2), a switch from a hyperbolic

region to an elliptic one may cause severe instabilities in the numerical solution since the propagation speeds in the elliptic region are infinite. Therefore, in order to develop high-order DG methods for (1.1), we rewrite it in a different form, which is suitable for DG settings:

$$(1.3) \quad \rho_t + (\chi\rho u)_x + (\chi\rho v)_y = \Delta\rho,$$

$$(1.4) \quad c_t = \Delta c - c + \rho,$$

$$(1.5) \quad u = c_x,$$

$$(1.6) \quad v = c_y,$$

where the new unknowns $\rho, c, u,$ and v satisfy the following boundary conditions:

$$(1.7) \quad \nabla\rho \cdot \mathbf{n} = \nabla c \cdot \mathbf{n} = (u, v)^T \cdot \mathbf{n} = 0, \quad (x, y) \in \partial\Omega.$$

The new system (1.3)–(1.6) may also be considered as a system of convection-diffusion-reaction equations

$$(1.8) \quad k\mathbf{Q}_t + \mathbf{F}(\mathbf{Q})_x + \mathbf{G}(\mathbf{Q})_y = k\Delta\mathbf{Q} + \mathbf{R}(\mathbf{Q}),$$

where $\mathbf{Q} := (\rho, c, u, v)^T$, the fluxes are $\mathbf{F}(\mathbf{Q}) := (\chi\rho u, 0, -c, 0)^T$ and $\mathbf{G}(\mathbf{Q}) := (\chi\rho v, 0, 0, -c)^T$, the reaction term is $\mathbf{R}(\mathbf{Q}) := (0, \rho - c, -u, -v)$, the constant $k = 1$ in the first two equations in (1.8), and $k = 0$ in the third and the fourth equations there. As we show in section 3, the convective part of the system (1.8) is hyperbolic.

In this paper, we develop a family of high-order DG methods for the system (1.8). The proposed methods are based on three primal DG methods: the nonsymmetric interior penalty Galerkin (NIPG), the symmetric interior penalty Galerkin (SIPG), and the incomplete interior penalty Galerkin (IIPG) methods [4, 18, 19, 40]. The numerical fluxes in the proposed DG methods are the fluxes developed for the semidiscrete finite-volume central-upwind schemes in [33] (see also [32, 34] and references therein). These schemes belong to the family of nonoscillatory central schemes, which are highly accurate and efficient methods applicable to general multidimensional systems of conservation laws and related problems. Like other central fluxes, the central-upwind ones are obtained without using the (approximate) Riemann problem solver, which is unavailable for the system under consideration. At the same time, certain upwinding information—one-sided speeds of propagation—is incorporated into the central-upwind fluxes.

We consider Cartesian grids and prove the error estimates for the proposed high-order DG methods under the assumption of boundedness of the exact solution. We also show that the blow-up time of the exact solution is bounded from above by the blow-up time of the solution of our DG methods. In numerical tests presented in section 6, we demonstrate that the obtained numerical solutions have no negative values and are oscillation-free, even though no slope-limiting technique has been implemented. We also demonstrate a high order of numerical convergence, achieved even when the final computational time gets close to the blow-up time and the spiky structure of the solution is well-developed.

The paper is organized as follows. In section 2, we introduce our notations and assumptions and state some standard results. The new DG methods are presented in section 3. The consistency and error analysis of the proposed methods are established in sections 4 and 5; some technical details which are omitted from the proof can be found in [20]. Finally, in section 6, we perform several numerical experiments.

2. Assumptions, notations, and standard results. We denote by \mathcal{E}_h a nondegenerate quasi-uniform rectangular subdivision of the domain Ω (the quasi-uniformity requirement will be used only in section 5 for establishing the rate of convergence with respect to the polynomial degree). The maximum diameter over all mesh elements is denoted by h , and the set of the interior edges is denoted by Γ_h . To each edge e in Γ_h , we associate a unit normal vector $\mathbf{n}_e = (n_x, n_y)$. We assume that \mathbf{n}_e is directed from the element E^1 to E^2 , where E^1 denotes a certain element and E^2 denotes an element that has a common edge with the element E^1 and a larger index (this simplified element notation will be used throughout the paper). For a boundary edge, \mathbf{n}_e is chosen so that it coincides with the outward normal.

The discrete space of discontinuous piecewise polynomials of degree r is denoted by

$$\mathcal{W}_{r,h}(\mathcal{E}_h) = \{w \in L^2(\Omega) : \forall E \in \mathcal{E}_h, w|_E \in P_r(E)\},$$

where $P_r(E)$ is a space of polynomials of degree r over the element E . For any function $w \in \mathcal{W}_{r,h}$, we denote the jump and average operators over a given edge e by $[w]$ and $\{w\}$, respectively:

$$\begin{aligned} \text{for an interior edge } e = \partial E^1 \cap \partial E^2, \quad [w] &:= w_e^{E^1} - w_e^{E^2}, \quad \{w\} := 0.5w_e^{E^1} + 0.5w_e^{E^2}, \\ \text{for a boundary edge } e = \partial E^1 \cap \partial \Omega, \quad [w] &:= w_e^{E^1}, \quad \{w\} := w_e^{E^1}, \end{aligned}$$

where $w_e^{E^1}$ and $w_e^{E^2}$ are the corresponding polynomial approximations from the elements E^1 and E^2 , respectively. We also recall that the following identity between the jump and the average operators is satisfied:

$$(2.1) \quad [w_1 w_2] = \{w_1\}[w_2] + \{w_2\}[w_1].$$

For the finite-element subdivision \mathcal{E}_h , we define the broken Sobolev space

$$H^s(\mathcal{E}_h) = \{w \in L^2(\Omega) : w|_{E^j} \in H^s(E^j), j = 1, \dots, N_h\}$$

with the norms

$$\|w\|_{0,\Omega} = \left(\sum_{E \in \mathcal{E}_h} \|w\|_{0,E}^2 \right)^{\frac{1}{2}} \quad \text{and} \quad \|w\|_{s,\Omega} = \left(\sum_{E \in \mathcal{E}_h} \|w\|_{s,E}^2 \right)^{\frac{1}{2}}, \quad s > 0,$$

where $\|\cdot\|_{s,E}$ denotes the Sobolev s -norm over the element E .

We now recall some well-known facts that will be used in the error analysis in section 5. First, let us state some approximation properties and inequalities for the finite-element space.

LEMMA 2.1 (*hp approximation* [5, 6]). *Let $E \in \mathcal{E}_h$ and $\psi \in H^s(E)$, $s \geq 0$. Then there exist a positive constant C , independent of ψ, r , and h , and a sequence $\tilde{\psi}_r^h \in P_r(E)$, $r = 1, 2, \dots$, such that for any $q \in [0, s]$*

$$(2.2) \quad \left\| \psi - \tilde{\psi}_r^h \right\|_{q,E} \leq C \frac{h^{\mu-q}}{r^{s-q}} \|\psi\|_{s,E}, \quad \mu := \min(r+1, s).$$

LEMMA 2.2 (*trace inequalities* [2]). *Let $E \in \mathcal{E}_h$. Then for the trace operators γ_0 and γ_1 , there exists a constant C_t , independent of h , such that*

$$(2.3) \quad \forall w \in H^s(E), \quad s \geq 1, \quad \|\gamma_0 w\|_{0,e} \leq C_t h^{-\frac{1}{2}} \left(\|w\|_{0,E} + h \|\nabla w\|_{0,E} \right),$$

$$(2.4) \quad \forall w \in H^s(E), \quad s \geq 2, \quad \|\gamma_1 w\|_{0,e} \leq C_t h^{-\frac{1}{2}} \left(\|\nabla w\|_{0,E} + h \|\nabla^2 w\|_{0,E} \right),$$

where e is an edge of the element E .

LEMMA 2.3 (see [40]). *Let E be a mesh element with an edge e . Then there is a constant C_t , independent of h and r , such that*

$$(2.5) \quad \forall w \in P_r(E), \quad \|\gamma_0 w\|_{0,e} \leq C_t h^{-\frac{1}{2}} r \|w\|_{0,E}.$$

LEMMA 2.4 (see [4, 9]). *There exists a constant C , independent of h and r , such that*

$$\forall w \in \mathcal{W}_{r,h}(\mathcal{E}_h), \quad \|w\|_{0,\Omega}^2 \leq C \left(\sum_{E \in \mathcal{E}_h} \|\nabla w\|_{0,E}^2 + \sum_{e \in \Gamma_h} \frac{1}{|e|} \|[w]\|_{0,e}^2 \right)^{\frac{1}{2}},$$

where $|e|$ denotes the measure of e .

LEMMA 2.5 (inverse inequalities). *Let $E \in \mathcal{E}_h$ and $w \in P_r(E)$. Then there exists a constant C , independent of h and r , such that*

$$(2.6) \quad \|w\|_{L^\infty(E)} \leq Ch^{-1} r \|w\|_{0,E},$$

$$(2.7) \quad \|w\|_{1,E} \leq Ch^{-1} r \|w\|_{0,E}.$$

We also recall the following form of Gronwall’s lemma.

LEMMA 2.6 (Gronwall). *Let φ, ψ , and ϕ be continuous nonnegative functions defined on the interval $a \leq t \leq b$, and the function ϕ is nondecreasing. If $\varphi(t) + \psi(t) \leq \phi(t) + \int_a^t \varphi(s) ds$ for all $t \in [a, b]$, then $\varphi(t) + \psi(t) \leq e^{t-a} \phi(t)$.*

In the analysis below we also make the following assumptions:

- Ω is a rectangular domain with the boundary $\partial\Omega = \partial\Omega_{\text{ver}} \cup \partial\Omega_{\text{hor}}$, where $\partial\Omega_{\text{ver}}$ and $\partial\Omega_{\text{hor}}$ denote the vertical and horizontal pieces of the boundary $\partial\Omega$, respectively. We also split the set of interior edges Γ_h into two sets of vertical Γ_h^{ver} and horizontal Γ_h^{hor} edges, respectively.

- The degree of basis polynomials is $r \geq 2$, and the maximum diameter of the elements is $h < 1$ (the latter assumption is needed only for simplification of the error analysis).

3. Description of the numerical scheme. We consider the Keller–Segel system (1.8). First, notice that the Jacobians of \mathbf{F} and \mathbf{G} are

$$\frac{\partial \mathbf{F}}{\partial \mathbf{Q}} = \begin{pmatrix} \chi u & 0 & \chi \rho & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \frac{\partial \mathbf{G}}{\partial \mathbf{Q}} = \begin{pmatrix} \chi v & 0 & 0 & \chi \rho \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}$$

and their eigenvalues are

$$(3.1) \quad \lambda_1^{\mathbf{F}} = \chi u, \quad \lambda_2^{\mathbf{F}} = \lambda_3^{\mathbf{F}} = \lambda_4^{\mathbf{F}} = 0 \quad \text{and} \quad \lambda_1^{\mathbf{G}} = \chi v, \quad \lambda_2^{\mathbf{G}} = \lambda_3^{\mathbf{G}} = \lambda_4^{\mathbf{G}} = 0,$$

respectively. Hence, the convective part of (1.8) is hyperbolic. We now design semidiscrete interior penalty Galerkin methods for this system.

We assume that at any time level $t \in [0, T]$ the solution $(\rho, c, u, v)^T$ is approximated by (discontinuous) piecewise polynomials of the corresponding degrees r_ρ, r_c, r_u , and r_v , which satisfy the following relation:

$$(3.2) \quad \frac{r_{\max}}{r_{\min}} \leq a, \quad r_{\max} := \max\{r_\rho, r_c, r_u, r_v\}, \quad r_{\min} := \min\{r_\rho, r_c, r_u, r_v\},$$

where a is a constant independent of r_ρ, r_c, r_p , and r_q .

Our new DG methods are formulated as follows: find a continuous-in-time solution

$$(\rho^{\text{DG}}(\cdot, t), c^{\text{DG}}(\cdot, t), u^{\text{DG}}(\cdot, t), v^{\text{DG}}(\cdot, t)) \in \mathcal{W}_{r_\rho, h}^\rho \times \mathcal{W}_{r_c, h}^c \times \mathcal{W}_{r_u, h}^u \times \mathcal{W}_{r_v, h}^v$$

which satisfies the following weak formulation of the chemotaxis system (1.3)–(1.6):

$$\begin{aligned} & \int_{\Omega} \rho_t^{\text{DG}} w^\rho + \sum_{E \in \mathcal{E}_h} \int_E \nabla \rho^{\text{DG}} \nabla w^\rho - \sum_{e \in \Gamma_h} \int_e \{\nabla \rho^{\text{DG}} \cdot \mathbf{n}_e\} [w^\rho] \\ & + \varepsilon \sum_{e \in \Gamma_h} \int_e \{\nabla w^\rho \cdot \mathbf{n}_e\} [\rho^{\text{DG}}] \\ & + \sigma_\rho \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \int_e [\rho^{\text{DG}}] [w^\rho] - \sum_{E \in \mathcal{E}_h} \int_E \chi \rho^{\text{DG}} u^{\text{DG}} (w^\rho)_x \\ & + \sum_{e \in \Gamma_h^{\text{ver}}} \int_e (\chi \rho^{\text{DG}} u^{\text{DG}})^* n_x [w^\rho] \\ (3.3) \quad & - \sum_{E \in \mathcal{E}_h} \int_E \chi \rho^{\text{DG}} v^{\text{DG}} (w^\rho)_y + \sum_{e \in \Gamma_h^{\text{hor}}} \int_e (\chi \rho^{\text{DG}} v^{\text{DG}})^* n_y [w^\rho] = 0, \end{aligned}$$

$$\begin{aligned} & \int_{\Omega} c_t^{\text{DG}} w^c + \sum_{E \in \mathcal{E}_h} \int_E \nabla c^{\text{DG}} \nabla w^c - \sum_{e \in \Gamma_h} \int_e \{\nabla c^{\text{DG}} \cdot \mathbf{n}_e\} [w^c] + \varepsilon \sum_{e \in \Gamma_h} \int_e \{\nabla w^c \cdot \mathbf{n}_e\} [c^{\text{DG}}] \\ (3.4) \quad & + \sigma_c \sum_{e \in \Gamma_h} \frac{r_c^2}{|e|} \int_e [c^{\text{DG}}] [w^c] + \int_{\Omega} c^{\text{DG}} w^c - \int_{\Omega} \rho^{\text{DG}} w^c = 0, \end{aligned}$$

$$\begin{aligned} & \int_{\Omega} u^{\text{DG}} w^u + \sum_{E \in \mathcal{E}_h} \int_E c^{\text{DG}} (w^u)_x + \sum_{e \in \Gamma_h^{\text{ver}}} \int_e (-c^{\text{DG}})_u^* n_x [w^u] \\ (3.5) \quad & - \sum_{e \in \partial\Omega_{\text{ver}}} \int_e c^{\text{DG}} n_x w^u + \sigma_u \sum_{e \in \Gamma_h \cup \partial\Omega_{\text{ver}}} \frac{r_u^2}{|e|} \int_e [u^{\text{DG}}] [w^u] = 0, \end{aligned}$$

$$\begin{aligned} & \int_{\Omega} v^{\text{DG}} w^v + \sum_{E \in \mathcal{E}_h} \int_E c^{\text{DG}} (w^v)_y + \sum_{e \in \Gamma_h^{\text{hor}}} \int_e (-c^{\text{DG}})_v^* n_y [w^v] \\ (3.6) \quad & - \sum_{e \in \partial\Omega_{\text{hor}}} \int_e c^{\text{DG}} n_y w^v + \sigma_v \sum_{e \in \Gamma_h \cup \partial\Omega_{\text{hor}}} \frac{r_v^2}{|e|} \int_e [v^{\text{DG}}] [w^v] = 0, \end{aligned}$$

with the initial conditions

$$\begin{aligned} (3.7) \quad & \int_{\Omega} \rho^{\text{DG}}(\cdot, 0) w^\rho = \int_{\Omega} \rho(\cdot, 0) w^\rho, \quad \int_{\Omega} c^{\text{DG}}(\cdot, 0) w^c = \int_{\Omega} c(\cdot, 0) w^c, \\ & \int_{\Omega} u^{\text{DG}}(\cdot, 0) w^u = \int_{\Omega} u(\cdot, 0) w^u, \quad \int_{\Omega} v^{\text{DG}}(\cdot, 0) w^v = \int_{\Omega} v(\cdot, 0) w^v. \end{aligned}$$

Here, $(w^\rho, w^c, w^u, w^v) \in \mathcal{W}_{r_\rho, h}^\rho \times \mathcal{W}_{r_c, h}^c \times \mathcal{W}_{r_u, h}^u \times \mathcal{W}_{r_v, h}^v$ are the test functions, and $\sigma_\rho, \sigma_c, \sigma_u$, and σ_v are real positive penalty parameters. The parameter ε is equal to either $-1, 0$, or 1 ; these values of ε correspond to the SIPG, IIPG, or NIPG method, respectively.

To approximate the convective terms in (3.3) and (3.5)–(3.6), we use the central-upwind fluxes from [33]:

$$\begin{aligned}
 (\chi\rho^{\text{DG}}u^{\text{DG}})^* &= \frac{a^{\text{out}}(\chi\rho^{\text{DG}}u^{\text{DG}})_e^{E^1} - a^{\text{in}}(\chi\rho^{\text{DG}}u^{\text{DG}})_e^{E^2}}{a^{\text{out}} - a^{\text{in}}} - \frac{a^{\text{out}}a^{\text{in}}}{a^{\text{out}} - a^{\text{in}}} [\rho^{\text{DG}}], \\
 (\chi\rho^{\text{DG}}v^{\text{DG}})^* &= \frac{b^{\text{out}}(\chi\rho^{\text{DG}}v^{\text{DG}})_e^{E^1} - b^{\text{in}}(\chi\rho^{\text{DG}}v^{\text{DG}})_e^{E^2}}{b^{\text{out}} - b^{\text{in}}} - \frac{b^{\text{out}}b^{\text{in}}}{b^{\text{out}} - b^{\text{in}}} [\rho^{\text{DG}}], \\
 (-c^{\text{DG}})_u^* &= -\frac{a^{\text{out}}(c^{\text{DG}})_e^{E^1} - a^{\text{in}}(c^{\text{DG}})_e^{E^2}}{a^{\text{out}} - a^{\text{in}}} - \frac{a^{\text{out}}a^{\text{in}}}{a^{\text{out}} - a^{\text{in}}} [u^{\text{DG}}], \\
 (-c^{\text{DG}})_v^* &= -\frac{b^{\text{out}}(c^{\text{DG}})_e^{E^1} - b^{\text{in}}(c^{\text{DG}})_e^{E^2}}{b^{\text{out}} - b^{\text{in}}} - \frac{b^{\text{out}}b^{\text{in}}}{b^{\text{out}} - b^{\text{in}}} [v^{\text{DG}}].
 \end{aligned}
 \tag{3.8}$$

Here, a^{out} , a^{in} , b^{out} , and b^{in} are the one-sided local speeds in the x - and y -directions. Since the convective part of the system (1.3)–(1.6) is hyperbolic, these speeds can be estimated using the largest and the smallest eigenvalues of the Jacobian $\frac{\partial \mathbf{F}}{\partial \mathbf{Q}}$ and $\frac{\partial \mathbf{G}}{\partial \mathbf{Q}}$ (see (3.1)):

$$\begin{aligned}
 a^{\text{out}} &= \max\left((\chi u^{\text{DG}})_e^{E^1}, (\chi u^{\text{DG}})_e^{E^2}, 0\right), & a^{\text{in}} &= \min\left((\chi u^{\text{DG}})_e^{E^1}, (\chi u^{\text{DG}})_e^{E^2}, 0\right), \\
 b^{\text{out}} &= \max\left((\chi v^{\text{DG}})_e^{E^1}, (\chi v^{\text{DG}})_e^{E^2}, 0\right), & b^{\text{in}} &= \min\left((\chi v^{\text{DG}})_e^{E^1}, (\chi v^{\text{DG}})_e^{E^2}, 0\right).
 \end{aligned}
 \tag{3.9}$$

Remark. If $a^{\text{out}} - a^{\text{in}} = 0$ at a certain element edge e , we set

$$\begin{aligned}
 (\chi\rho^{\text{DG}}u^{\text{DG}})^* &= \frac{(\chi\rho^{\text{DG}}u^{\text{DG}})_e^{E^1} + (\chi\rho^{\text{DG}}u^{\text{DG}})_e^{E^2}}{2}, & (-c^{\text{DG}})_u^* &= -\frac{(c^{\text{DG}})_e^{E^1} + (c^{\text{DG}})_e^{E^2}}{2}, \\
 (\chi\rho^{\text{DG}}v^{\text{DG}})^* &= \frac{(\chi\rho^{\text{DG}}v^{\text{DG}})_e^{E^1} + (\chi\rho^{\text{DG}}v^{\text{DG}})_e^{E^2}}{2}, & (-c^{\text{DG}})_v^* &= -\frac{(c^{\text{DG}})_e^{E^1} + (c^{\text{DG}})_e^{E^2}}{2}
 \end{aligned}$$

there. Notice that in any case the following inequalities are satisfied:

$$\frac{a^{\text{out}}}{a^{\text{out}} - a^{\text{in}}} \leq 1, \quad \frac{-a^{\text{in}}}{a^{\text{out}} - a^{\text{in}}} \leq 1, \quad \frac{b^{\text{out}}}{b^{\text{out}} - b^{\text{in}}} \leq 1, \quad \text{and} \quad \frac{-b^{\text{in}}}{b^{\text{out}} - b^{\text{in}}} \leq 1.
 \tag{3.10}$$

From now on we will assume that $a^{\text{out}} - a^{\text{in}} > 0$ and $b^{\text{out}} - b^{\text{in}} > 0$ throughout the computational domain.

4. Consistency of the numerical scheme. In this section, we show that the proposed DG methods (3.3)–(3.6) are strongly consistent with the Keller–Segel system (1.3)–(1.6).

LEMMA 4.1. *If the solution of (1.3)–(1.6) is sufficiently regular, namely, if $(\rho, c) \in H^1([0, T]) \cap H^2(\mathcal{E}_h)$ and $(u, v) \in L^2([0, T]) \cap H^2(\mathcal{E}_h)$, then it satisfies the formulation (3.3)–(3.6).*

Proof. We first multiply (1.3) by $w^\rho \in \mathcal{W}_{r_\rho, h}^\rho$ and integrate by parts on one element E to obtain

$$\begin{aligned}
 \int_E \rho_t w^\rho + \int_E \nabla \rho \nabla w^\rho - \int_{\partial E} \nabla \rho \cdot \mathbf{n}_e w^\rho - \int_E \chi \rho u (w^\rho)_x + \int_{\partial E} \chi \rho u n_x w^\rho \\
 - \int_E \chi \rho v (w^\rho)_y + \int_{\partial E} \chi \rho v n_y w^\rho = 0.
 \end{aligned}
 \tag{4.1}$$

Notice that the continuity of ρ and u implies that at the edge e , $\rho_e^{E^1} = \rho_e^{E^2}$ and $(\chi\rho u)_e^{E^1} = (\chi\rho u)_e^{E^2}$. Therefore, $[\rho] = 0$ and

$$\begin{aligned} \{\chi\rho u\} &= \frac{1}{2}(\chi\rho u)_e^{E^1} + \frac{1}{2}(\chi\rho u)_e^{E^2} = (\chi\rho u)_e^{E^1} = \frac{a^{\text{out}} - a^{\text{in}}}{a^{\text{out}} - a^{\text{in}}}(\chi\rho u)_e^{E^1} \\ &= \frac{a^{\text{out}}}{a^{\text{out}} - a^{\text{in}}}(\chi\rho u)_e^{E^1} - \frac{a^{\text{in}}}{a^{\text{out}} - a^{\text{in}}}(\chi\rho u)_e^{E^2} \\ &= \frac{a^{\text{out}}(\chi\rho u)_e^{E^1} - a^{\text{in}}(\chi\rho u)_e^{E^2}}{a^{\text{out}} - a^{\text{in}}} = (\chi\rho u)^*. \end{aligned}$$

Summing now (4.1) over all elements $E \in \mathcal{E}_h$, using the jump-average identity (2.1), adding the penalty terms $\varepsilon \sum_{e \in \Gamma_h} \int_e \{\nabla w^\rho \cdot \mathbf{n}_e\}[\rho]$ and $\sigma_\rho \sum_{e \in \Gamma_h} \frac{r_e^2}{|e|} \int_e [\rho][w^\rho]$, and using the Neumann boundary conditions (1.7), we obtain that the solution of the system (1.3)–(1.6) satisfies (3.3). A similar procedure can be applied to show that the solution of (1.3)–(1.6) satisfies (3.4)–(3.6) as well. This concludes the consistency proof. \square

5. Error analysis. In this section, we prove the existence and show the convergence of the numerical solution using Schauder’s fixed point theorem [25].

In the analysis below, we will assume that the exact solution of the system (1.3)–(1.6) is sufficiently regular for $t \leq T$, where T is a pre-blow-up time. In particular we will assume that

$$(5.1) \quad (\rho, c, u, v) \in H^{s_1}([0, T]) \cap H^{s_2}(\Omega), \quad s_1 > 3/2, \quad s_2 \geq 3,$$

which is needed for the h -analysis (convergence rate with respect to the mesh size), or

$$(5.2) \quad (\rho, c, u, v) \in H^{s_1}([0, T]) \cap H^{s_2}(\Omega), \quad s_1 > 3/2, \quad s_2 \geq 5,$$

which is needed for the r -analysis (convergence rate with respect to the polynomial degree). Notice that these assumptions are reasonable since classical solutions of the Keller–Segel system (1.1) are regular (before the blow-up time), provided the initial data are sufficiently smooth; see [27] and references therein.

We denote by $\tilde{\rho}, \tilde{c}, \tilde{u}$, and \tilde{v} the piecewise polynomial interpolants of the exact solution components ρ, c, u , and v of the Keller–Segel system (1.3)–(1.6) and assume that these interpolants satisfy the approximation property (2.2). We then use the idea similar to [37] and define the following subset of the broken Sobolev space:

$$\begin{aligned} S = \left\{ (\phi^\rho, \phi^c, \phi^u, \phi^v) \in L^2([0, T]) \cap L^\infty([0, T]) \cap \mathcal{W}_{r_\rho, h}^\rho \times \mathcal{W}_{r_c, h}^c \times \mathcal{W}_{r_u, h}^u \times \mathcal{W}_{r_v, h}^v : \right. \\ \sup_{t \in [0, T]} \|\phi^\rho - \tilde{\rho}\|_{0, \Omega}^2 + \int_0^T \left(\|\nabla(\phi^\rho - \tilde{\rho})\|_{0, \Omega}^2 + \sum_{e \in \Gamma_h} \frac{r_e^2}{|e|} \|[(\phi^\rho - \tilde{\rho})]\|_{0, e}^2 \right) \leq C_\rho \mathcal{E}_1, \\ \sup_{t \in [0, T]} \|\phi^c - \tilde{c}\|_{0, \Omega}^2 + \int_0^T \left(\|\nabla(\phi^c - \tilde{c})\|_{0, \Omega}^2 + \sum_{e \in \Gamma_h} \frac{r_e^2}{|e|} \|[(\phi^c - \tilde{c})]\|_{0, e}^2 \right) \leq C_c \mathcal{E}_1, \\ \sup_{[0, T]} \|\phi^u - \tilde{u}\|_{0, \Omega} \leq \mathcal{E}_2, \quad \int_0^T \left(\|\phi^u - \tilde{u}\|_{0, \Omega}^2 + \sum_{e \in \Gamma_h} \frac{r_e^2}{|e|} \|[(\phi^u - \tilde{u})]\|_{0, e}^2 \right) \leq C_u \mathcal{E}_1, \\ \left. \sup_{[0, T]} \|\phi^v - \tilde{v}\|_{0, \Omega} \leq \mathcal{E}_2, \quad \int_0^T \left(\|\phi^v - \tilde{v}\|_{0, \Omega}^2 + \sum_{e \in \Gamma_h} \frac{r_e^2}{|e|} \|[(\phi^v - \tilde{v})]\|_{0, e}^2 \right) \leq C_v \mathcal{E}_1 \right\}, \end{aligned}$$

where

$$(5.3) \quad \mathcal{E}_1 := \sum_{\alpha \in \{\rho, c, u, v\}} \frac{h^{2 \min(r_\alpha + 1, s_\alpha) - 2}}{r_\alpha^{2s_\alpha - 4}}, \quad \mathcal{E}_2 := Ch \left(\frac{1}{r_\rho} + \frac{1}{r_c} + \frac{1}{r_u} + \frac{1}{r_v} \right),$$

$C, C_\rho, C_c, C_u,$ and C_v are positive constants (which will be defined later) independent of h and the polynomial degrees (r_ρ, r_c, r_u, r_v) , and the parameters $s_\rho, s_c, s_u,$ and s_v denote the regularity of the corresponding components of the exact solution. Clearly the subset S is a closed convex subset of the broken Sobolev space, and it is not empty since it contains the element $(\tilde{\rho}, \tilde{c}, \tilde{u}, \tilde{v})$. We first show that the functions in S are bounded.

LEMMA 5.1. *For any $(\phi^\rho, \phi^c, \phi^u, \phi^v) \in S$, there exist positive constants $M_\rho, M_c, M_u,$ and M_v , independent of $h, r_\rho, r_c, r_u,$ and r_v , such that*

$$(5.4) \quad \sup_{t \in [0, T]} \|\phi^\alpha\|_{\infty, \Omega} \leq M_\alpha, \quad \alpha \in \{\rho, c, u, v\}.$$

Proof. From the definition of the subset S , we have $\sup_{t \in [0, T]} \|\phi^\rho - \tilde{\rho}\|_{0, \Omega}^2 \leq C_\rho \mathcal{E}_1$ and hence $\sup_{t \in [0, T]} \|\phi^\rho - \tilde{\rho}\|_{0, \Omega} \leq M \frac{h}{r_{\min}}$. Using the inverse inequality (2.6), we obtain

$$\sup_{t \in [0, T]} \|\phi^\rho - \tilde{\rho}\|_{\infty, \Omega} \leq M_1 r_\rho h^{-1} \sup_{t \in [0, T]} \|\phi^\rho - \tilde{\rho}\|_{0, \Omega} \leq \frac{r_{\max}}{r_{\min}} M^* \leq M.$$

This estimate implies that $\sup_{t \in [0, T]} \|\phi^\rho\|_{\infty, \Omega} \leq M + \sup_{[0, T]} \|\tilde{\rho}\|_{\infty, \Omega}$, which, together with the hp approximation property (see Lemma 2.1), yields the bound (5.4) for $\alpha = \rho$. The estimates for $\alpha = c, u,$ and v are obtained in a similar manner. \square

We now define the solution operator A on S as follows:

$$\forall (\phi^\rho, \phi^c, \phi^u, \phi^v) \in S, \quad A(\phi^\rho, \phi^c, \phi^u, \phi^v) = (\phi_L^\rho, \phi_L^c, \phi_L^u, \phi_L^v),$$

where the initial conditions are $(\phi_L^{\rho, 0}, \phi_L^{c, 0}, \phi_L^{u, 0}, \phi_L^{v, 0}) = (\tilde{\rho}^0, \tilde{c}^0, \tilde{u}^0, \tilde{v}^0)$ and the functions

$$\begin{aligned} \phi_L^\rho &\in \mathcal{W}_{r_\rho, h, t}^\rho := H^s([0, T]) \cap \mathcal{W}_{r_\rho, h}^\rho, & \phi_L^c &\in \mathcal{W}_{r_c, h, t}^c := H^s([0, T]) \cap \mathcal{W}_{r_c, h}^c, & s > 3/2, \\ \phi_L^\alpha &\in \mathcal{W}_{r_\alpha, h, t}^\alpha := L^2([0, T]) \cap L^\infty([0, T]) \cap \mathcal{W}_{r_\alpha, h}^\alpha, & \alpha &\in \{u, v\}, \end{aligned}$$

are such that

$$(5.5) \quad \begin{aligned} &\int_\Omega (\phi_L^\rho)_t w^\rho + \sum_{E \in \mathcal{E}_h} \int_E \nabla(\phi_L^\rho) \nabla w^\rho - \sum_{e \in \Gamma_h} \int_e \{\nabla \phi_L^\rho \cdot \mathbf{n}_e\} [w^\rho] + \varepsilon \sum_{e \in \Gamma_h} \int_e \{\nabla w^\rho \cdot \mathbf{n}_e\} [\phi_L^\rho] \\ &+ \sigma_\rho \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \int_e [\phi_L^\rho] [w^\rho] - \sum_{E \in \mathcal{E}_h} \int_E \chi \phi_L^\rho \phi^u (w^\rho)_x + \sum_{e \in \Gamma_h^{\text{ver}}} \int_e (\chi \phi_L^\rho \phi^u)^* n_x [w^\rho] \\ &- \sum_{E \in \mathcal{E}_h} \int_E \chi \phi_L^\rho \phi^v (w^\rho)_y + \sum_{e \in \Gamma_h^{\text{hor}}} \int_e (\chi \phi_L^\rho \phi^v)^* n_y [w^\rho] = 0 \quad \forall w^\rho \in \mathcal{W}_{r_\rho, h}^\rho, \end{aligned}$$

$$(5.6) \quad \begin{aligned} &\int_\Omega (\phi_L^c)_t w^c + \sum_{E \in \mathcal{E}_h} \int_E \nabla(\phi_L^c) \nabla w^c - \sum_{e \in \Gamma_h} \int_e \{\nabla \phi_L^c \cdot \mathbf{n}_e\} [w^c] + \varepsilon \sum_{e \in \Gamma_h} \int_e \{\nabla w^c \cdot \mathbf{n}_e\} [\phi_L^c] \\ &+ \sigma_c \sum_{e \in \Gamma_h} \frac{r_c^2}{|e|} \int_e [\phi_L^c] [w^c] + \int_\Omega \phi_L^c w^c - \int_\Omega \phi_L^\rho w^c = 0 \quad \forall w^c \in \mathcal{W}_{r_c, h}^c, \end{aligned}$$

$$\begin{aligned}
 & \int_{\Omega} \phi_L^u w^u + \sum_{E \in \mathcal{E}_h} \int_E \phi_L^c (w^u)_x + \sum_{e \in \Gamma_h^{\text{ver}}} \int_e (-\phi_L^c)_u^* n_x [w^u] - \sum_{e \in \partial\Omega_{\text{ver}}} \int_e \phi_L^c n_x w^u \\
 (5.7) \quad & + \sigma_u \sum_{e \in \Gamma_h \cup \partial\Omega_{\text{ver}}} \frac{r_u^2}{|e|} \int_e [\phi_L^u] [w^u] = 0 \quad \forall w^u \in \mathcal{W}_{r_u, h}^u,
 \end{aligned}$$

$$\begin{aligned}
 & \int_{\Omega} \phi_L^v w^v + \sum_{E \in \mathcal{E}_h} \int_E \phi_L^c (w^v)_y + \sum_{e \in \Gamma_h^{\text{hor}}} \int_e (-\phi_L^c)_v^* n_y [w^v] - \sum_{e \in \partial\Omega_{\text{hor}}} \int_e \phi_L^c n_y w^v \\
 (5.8) \quad & + \sigma_v \sum_{e \in \Gamma_h \cup \partial\Omega_{\text{hor}}} \frac{r_v^2}{|e|} \int_e [\phi_L^v] [w^v] = 0 \quad \forall w^v \in \mathcal{W}_{r_v, h}^v.
 \end{aligned}$$

As before, the central-upwind numerical fluxes are utilized in (5.5)–(5.8):

$$\begin{aligned}
 (\chi \phi_L^\rho \phi^u)^* &= \frac{a_L^{\text{out}} (\chi \phi_L^\rho \phi^u)_e^{E^1} - a_L^{\text{in}} (\chi \phi_L^\rho \phi^u)_e^{E^2}}{a_L^{\text{out}} - a_L^{\text{in}}} - \frac{a_L^{\text{out}} a_L^{\text{in}}}{a_L^{\text{out}} - a_L^{\text{in}}} [\phi_L^\rho], \\
 (\chi \phi_L^\rho \phi^v)^* &= \frac{b_L^{\text{out}} (\chi \phi_L^\rho \phi^v)_e^{E^1} - b_L^{\text{in}} (\chi \phi_L^\rho \phi^v)_e^{E^2}}{b_L^{\text{out}} - b_L^{\text{in}}} - \frac{b_L^{\text{out}} b_L^{\text{in}}}{b_L^{\text{out}} - b_L^{\text{in}}} [\phi_L^\rho], \\
 (-\phi_L^c)_u^* &= -\frac{a_L^{\text{out}} (\phi_L^c)_e^{E^1} - a_L^{\text{in}} (\phi_L^c)_e^{E^2}}{a_L^{\text{out}} - a_L^{\text{in}}} - \frac{a_L^{\text{out}} a_L^{\text{in}}}{a_L^{\text{out}} - a_L^{\text{in}}} [\phi_L^u], \\
 (-\phi_L^c)_v^* &= -\frac{b_L^{\text{out}} (\phi_L^c)_e^{E^1} - b_L^{\text{in}} (\phi_L^c)_e^{E^2}}{b_L^{\text{out}} - b_L^{\text{in}}} - \frac{b_L^{\text{out}} b_L^{\text{in}}}{b_L^{\text{out}} - b_L^{\text{in}}} [\phi_L^v],
 \end{aligned}
 \tag{5.9}$$

where the one-sided local speeds are

$$\begin{aligned}
 a_L^{\text{out}} &:= \max \left((\chi \phi^u)_e^{E^1}, (\chi \phi^u)_e^{E^2}, 0 \right), \quad a_L^{\text{in}} := \min \left((\chi \phi^u)_e^{E^1}, (\chi \phi^u)_e^{E^2}, 0 \right), \\
 b_L^{\text{out}} &:= \max \left((\chi \phi^v)_e^{E^1}, (\chi \phi^v)_e^{E^2}, 0 \right), \quad b_L^{\text{in}} := \min \left((\chi \phi^v)_e^{E^1}, (\chi \phi^v)_e^{E^2}, 0 \right).
 \end{aligned}
 \tag{5.10}$$

Notice that the inequalities similar to (3.10),

$$(5.11) \quad \frac{a_L^{\text{out}}}{a_L^{\text{out}} - a_L^{\text{in}}} \leq 1, \quad \frac{-a_L^{\text{in}}}{a_L^{\text{out}} - a_L^{\text{in}}} \leq 1, \quad \frac{b_L^{\text{out}}}{b_L^{\text{out}} - b_L^{\text{in}}} \leq 1, \quad \text{and} \quad \frac{-b_L^{\text{in}}}{b_L^{\text{out}} - b_L^{\text{in}}} \leq 1,$$

which are needed in our convergence proof, are satisfied for the local speeds defined in (5.10) as well (for simplicity, we assume that $a^{\text{out}} - a^{\text{in}} \neq 0$ and $b^{\text{out}} - b^{\text{in}} \neq 0$ throughout the computational domain).

We now show that the operator A is well-defined by proving the existence and uniqueness of $(\phi_L^\rho, \phi_L^c, \phi_L^u, \phi_L^v)$.

LEMMA 5.2. *There exists a unique solution $(\phi_L^\rho, \phi_L^c, \phi_L^u, \phi_L^v) \in \mathcal{W}_{r_\rho, h, t}^\rho \times \mathcal{W}_{r_c, h, t}^c \times \mathcal{W}_{r_u, h, t}^u \times \mathcal{W}_{r_v, h, t}^v$ of (5.5)–(5.8).*

Proof. First, notice that (5.5)–(5.6) can be rewritten as the explicit linear differential equations for ϕ_L^ρ and ϕ_L^c . Hence, there exists a unique solution $(\phi_L^\rho, \phi_L^c) \in \mathcal{W}_{r_\rho, h, t}^\rho \times \mathcal{W}_{r_c, h, t}^c$.

Equations (5.7)–(5.8) can be rewritten as

$$\begin{aligned}
 \int_{\Omega} \phi_L^u w^u + \sigma_u \sum_{e \in \Gamma_h \cup \partial\Omega_{\text{ver}}} \frac{r_u^2}{|e|} \int_e [\phi_L^u][w^u] &= - \sum_{E \in \mathcal{E}_h} \int_E \phi_L^c (w^u)_x - \sum_{e \in \Gamma_h^{\text{ver}}} \int_e (-\phi_L^c)_u^* n_x [w^u] \\
 (5.12) \qquad \qquad \qquad &+ \sum_{e \in \partial\Omega_{\text{ver}}} \int_e \phi_L^c n_x w^u \quad \forall w^u \in \mathcal{W}_{r_u, h}^u,
 \end{aligned}$$

$$\begin{aligned}
 \int_{\Omega} \phi_L^v w^v + \sigma_v \sum_{e \in \Gamma_h \cup \partial\Omega_{\text{hor}}} \frac{r_v^2}{|e|} \int_e [\phi_L^v][w^v] &= - \sum_{E \in \mathcal{E}_h} \int_E \phi_L^c (w^v)_y - \sum_{e \in \Gamma_h^{\text{hor}}} \int_e (-\phi_L^c)_v^* n_y [w^v] \\
 (5.13) \qquad \qquad \qquad &+ \sum_{e \in \partial\Omega_{\text{hor}}} \int_e \phi_L^c n_y w^v \quad \forall w^v \in \mathcal{W}_{r_v, h}^v.
 \end{aligned}$$

The bilinear form on the left-hand side (LHS) of (5.12) is coercive since for all $\varphi \in \mathcal{W}_{r_u, h}^u$, $\int_{\Omega} \varphi \varphi + \sigma_u \sum_{e \in \Gamma_h \cup \partial\Omega_{\text{ver}}} \frac{r_u^2}{|e|} \int_e [\varphi][\varphi] \geq \|\varphi\|_{0, \Omega}^2$. It is also continuous on $\mathcal{W}_{r_u, h}^u \times \mathcal{W}_{r_u, h}^u$, while the linear form on the right-hand side (RHS) of (5.12) is continuous on $\mathcal{W}_{r_u, h}^u$. Hence, there exists a unique solution of (5.12). The same argument is true for (5.13). This concludes the proof of the lemma. \square

Our next goal is to show that the operator A maps S into itself and that A is compact. By the second Shauder fixed-point theorem [25], this will imply that the nonlinear mapping $(\phi^\rho, \phi^c, \phi^u, \phi^v) \in S \rightarrow A(\phi^\rho, \phi^c, \phi^u, \phi^v)$ has a fixed point denoted by $(\rho^{\text{DG}}, c^{\text{DG}}, u^{\text{DG}}, v^{\text{DG}})$.

THEOREM 5.3. *Let the solution of (1.3)–(1.6) satisfy the assumption (5.1). Then for any $(\phi^\rho, \phi^c, \phi^u, \phi^v) \in S$, $A(\phi^\rho, \phi^c, \phi^u, \phi^v) \in S$.*

Proof. Let $(\phi^\rho, \phi^c, \phi^u, \phi^v) \in S$ and $(\phi_L^\rho, \phi_L^c, \phi_L^u, \phi_L^v) = A(\phi^\rho, \phi^c, \phi^u, \phi^v)$. We introduce the following notation:

$$\begin{aligned}
 (5.14) \qquad \tau^\rho &:= \phi_L^\rho - \tilde{\rho}, & \xi^\rho &:= \rho - \tilde{\rho}, & \tau^c &:= \phi_L^c - \tilde{c}, & \xi^c &:= c - \tilde{c}, \\
 \tau^u &:= \phi_L^u - \tilde{u}, & \xi^u &:= u - \tilde{u}, & \tau^v &:= \phi_L^v - \tilde{v}, & \xi^v &:= v - \tilde{v}.
 \end{aligned}$$

It follows from the consistency Lemma 4.1 that the exact solution of (1.3)–(1.6) satisfies not only (3.3) but also the similar equation

$$\begin{aligned}
 \int_{\Omega} \rho_t w^\rho + \sum_{E \in \mathcal{E}_h} \int_E \nabla \rho \nabla w^\rho - \sum_{e \in \Gamma_h} \int_e \{\nabla \rho \cdot \mathbf{n}_e\} [w^\rho] + \varepsilon \sum_{e \in \Gamma_h} \int_e \{\nabla w^\rho \cdot \mathbf{n}_e\} [\rho] \\
 + \sigma_\rho \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \int_e [\rho][w^\rho] - \sum_{E \in \mathcal{E}_h} \int_E \chi \rho u (w^\rho)_x + \sum_{e \in \Gamma_h^{\text{ver}}} \int_e (\chi \rho u)^{**} n_x [w^\rho] \\
 (5.15) \qquad \qquad \qquad - \sum_{E \in \mathcal{E}_h} \int_E \chi \rho v (w^\rho)_y + \sum_{e \in \Gamma_h^{\text{hor}}} \int_e (\chi \rho v)^{**} n_y [w^\rho] = 0,
 \end{aligned}$$

where

$$\begin{aligned}
 (\chi \rho u)^{**} &:= \frac{a_L^{\text{out}} (\chi \rho u)_e^{E^1} - a_L^{\text{in}} (\chi \rho u)_e^{E^2}}{a_L^{\text{out}} - a_L^{\text{in}}} - \frac{a_L^{\text{out}} a_L^{\text{in}}}{a_L^{\text{out}} - a_L^{\text{in}}} [\rho], \\
 (\chi \rho v)^{**} &:= \frac{b_L^{\text{out}} (\chi \rho v)_e^{E^1} - b_L^{\text{in}} (\chi \rho v)_e^{E^2}}{b_L^{\text{out}} - b_L^{\text{in}}} - \frac{b_L^{\text{out}} b_L^{\text{in}}}{b_L^{\text{out}} - b_L^{\text{in}}} [\rho]
 \end{aligned}$$

and the local speeds a_L^{out} , a_L^{in} , b_L^{out} , and b_L^{in} are given by (5.10). Using (5.14), (5.15) can be rewritten as

$$\begin{aligned}
 & \int_{\Omega} \tilde{\rho}_t w^\rho + \sum_{E \in \mathcal{E}_h} \int_E \nabla \tilde{\rho} \nabla w^\rho - \sum_{e \in \Gamma_h} \int_e \{\nabla \tilde{\rho} \cdot \mathbf{n}_e\} [w^\rho] + \varepsilon \sum_{e \in \Gamma_h} \int_e \{\nabla w^\rho \cdot \mathbf{n}_e\} [\tilde{\rho}] \\
 & + \sigma_\rho \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \int_e [\tilde{\rho}] [w^\rho] - \sum_{E \in \mathcal{E}_h} \int_E \chi \tilde{\rho} \phi^u (w^\rho)_x + \sum_{e \in \Gamma_h^{\text{ver}}} \int_e (\chi \rho u)^{**} n_x [w^\rho] \\
 & - \sum_{E \in \mathcal{E}_h} \int_E \chi \tilde{\rho} \phi^v (w^\rho)_y + \sum_{e \in \Gamma_h^{\text{hor}}} \int_e (\chi \rho v)^{**} n_y [w^\rho] \\
 & = - \int_{\Omega} \xi_t^\rho w^\rho - \sum_{E \in \mathcal{E}_h} \int_E \nabla \xi^\rho \nabla w^\rho + \sum_{e \in \Gamma_h} \int_e \{\nabla \xi^\rho \cdot \mathbf{n}_e\} [w^\rho] \\
 & \quad - \varepsilon \sum_{e \in \Gamma_h} \int_e \{\nabla w^\rho \cdot \mathbf{n}_e\} [\xi^\rho] \\
 & \quad - \sigma_\rho \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \int_e [\xi^\rho] [w^\rho] + \sum_{E \in \mathcal{E}_h} \int_E \chi \xi^\rho u (w^\rho)_x - \sum_{E \in \mathcal{E}_h} \int_E \chi \tilde{\rho} (\phi^u - u) (w^\rho)_x \\
 (5.16) \quad & + \sum_{E \in \mathcal{E}_h} \int_E \chi \xi^\rho v (w^\rho)_y - \sum_{E \in \mathcal{E}_h} \int_E \chi \tilde{\rho} (\phi^v - v) (w^\rho)_y.
 \end{aligned}$$

Subtracting (5.16) from (5.5) and choosing $w^\rho = \tau^\rho$, we obtain

$$\begin{aligned}
 & \frac{1}{2} \frac{d}{dt} \left(\|\tau^\rho\|_{0,\Omega}^2 \right) + \|\nabla \tau^\rho\|_{0,\Omega}^2 + \sigma_\rho \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \|\tau^\rho\|_{0,e}^2 \\
 & = (1 - \varepsilon) \sum_{e \in \Gamma_h} \int_e \{\nabla \tau^\rho \cdot \mathbf{n}_e\} [\tau^\rho] + \sum_{E \in \mathcal{E}_h} \int_E \chi \tau^\rho \phi^u (\tau^\rho)_x \\
 & \quad - \sum_{e \in \Gamma_h^{\text{ver}}} \int_e ((\chi \phi_L^\rho \phi^u)^* - (\chi \rho u)^{**}) n_x [\tau^\rho] + \sum_{E \in \mathcal{E}_h} \int_E \chi \tau^\rho \phi^v (\tau^\rho)_y \\
 & \quad - \sum_{e \in \Gamma_h^{\text{hor}}} \int_e ((\chi \phi_L^\rho \phi^v)^* - (\chi \rho v)^{**}) n_y [\tau^\rho] + \int_{\Omega} \xi_t^\rho \tau^\rho \\
 & \quad + \sum_{E \in \mathcal{E}_h} \int_E \nabla \xi^\rho \nabla \tau^\rho - \sum_{e \in \Gamma_h} \int_e \{\nabla \xi^\rho \cdot \mathbf{n}_e\} [\tau^\rho] + \varepsilon \sum_{e \in \Gamma_h} \int_e \{\nabla \tau^\rho \cdot \mathbf{n}_e\} [\xi^\rho] \\
 & \quad + \sigma_\rho \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \int_e [\xi^\rho] [\tau^\rho] - \sum_{E \in \mathcal{E}_h} \int_E \chi \xi^\rho u (\tau^\rho)_x - \sum_{E \in \mathcal{E}_h} \int_E \chi \xi^\rho v (\tau^\rho)_y \\
 (5.17) \quad & + \sum_{E \in \mathcal{E}_h} \int_E \chi \tilde{\rho} (\phi^u - u) (\tau^\rho)_x + \sum_{E \in \mathcal{E}_h} \int_E \chi \tilde{\rho} (\phi^v - v) (\tau^\rho)_y =: T_1^\rho + T_2^\rho + \dots + T_{14}^\rho.
 \end{aligned}$$

Next, we bound each term on the RHS of (5.17) using standard DG techniques. The quantities ε_i in the estimates below are positive real numbers, which will be defined later.

We begin with the first term on the RHS of (5.17). The Cauchy–Schwarz inequality yields

$$|T_1^\rho| \leq (1 - \varepsilon) \sum_{e \in \Gamma_h} \|\{\nabla \tau^\rho\}\|_{0,e} \|\tau^\rho\|_{0,e}.$$

As before, we denote by E^1 and E^2 the two elements sharing the edge e . Then, using the inequality (2.5), we obtain

$$\begin{aligned} \sum_{e \in \Gamma_h} \|\{\nabla \tau^\rho\}\|_{0,e} \|\tau^\rho\|_{0,e} &\leq \sum_{e \in \Gamma_h} \frac{1}{2} \left(\left\| (\nabla \tau^\rho)_e^{E^1} \right\|_{0,e} + \left\| (\nabla \tau^\rho)_e^{E^2} \right\|_{0,e} \right) \|\tau^\rho\|_{0,e} \\ &\leq \frac{C_t r_\rho}{2\sqrt{h}} \sum_{e \in \Gamma_h} \left(\|\nabla \tau^\rho\|_{0,E^1} + \|\nabla \tau^\rho\|_{0,E^2} \right) \|\tau^\rho\|_{0,e}, \end{aligned}$$

and hence, using the fact that $|e| \leq \sqrt{h}$, we end up with the following bound on T_1^ρ :

$$|T_1^\rho| \leq \varepsilon_1^\rho \sum_{E \in \mathcal{E}_h} \|\nabla \tau^\rho\|_{0,E}^2 + C_1^\rho \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \|\tau^\rho\|_{0,e}^2 = \varepsilon_1^\rho \|\nabla \tau^\rho\|_{0,\Omega}^2 + C_1^\rho \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \|\tau^\rho\|_{0,e}^2. \quad (5.18)$$

Consider now the second term on the RHS of (5.17). From Lemma 5.1 we know that ϕ^u is a bounded function, and hence T_2^ρ can be bounded as follows:

$$(5.19) \quad |T_2^\rho| \leq \varepsilon_2^\rho \sum_{E \in \mathcal{E}_h} \|(\tau^\rho)_x\|_{0,E}^2 + C_2^\rho \|\tau^\rho\|_{0,\Omega}^2 \leq \varepsilon_2^\rho \|(\tau^\rho)_x\|_{0,\Omega}^2 + C_2^\rho \|\tau^\rho\|_{0,\Omega}^2.$$

Next, we bound the third term on the RHS of (5.17) as

$$\begin{aligned} |T_3^\rho| &\leq \sum_{e \in \Gamma_h^{\text{ver}}} \left(\left| \int_e \frac{a_L^{\text{out}}}{a_L^{\text{out}} - a_L^{\text{in}}} \left((\chi \phi_L^\rho \phi^u)_e^{E^1} - (\chi \rho u)_e^{E^1} \right) n_x[\tau^\rho] \right| \right. \\ &\quad + \left| \int_e \frac{-a_L^{\text{in}}}{a_L^{\text{out}} - a_L^{\text{in}}} \left((\chi \phi_L^\rho \phi^u)_e^{E^2} - (\chi \rho u)_e^{E^2} \right) n_x[\tau^\rho] \right| \\ (5.20) \quad &\quad \left. + \left| \int_e \frac{-a_L^{\text{in}} a_L^{\text{out}}}{a_L^{\text{out}} - a_L^{\text{in}}} [\phi_L^\rho - \rho] n_x[\tau^\rho] \right| \right) =: \text{I} + \text{II} + \text{III}. \end{aligned}$$

Using (5.11) and (5.14), the first term on the RHS of (5.20) can be estimated by

$$\begin{aligned} \text{I} &\leq \chi \sum_{e \in \Gamma_h^{\text{ver}}} \left| \int_e \left((\phi_L^\rho \phi^u)_e^{E^1} - (\rho u)_e^{E^1} \right) n_x[\tau^\rho] \right| \\ &\leq \chi \sum_{e \in \Gamma_h^{\text{ver}}} \left(\left| \int_e (\tau^\rho \phi^u)_e^{E^1} n_x[\tau^\rho] \right| + \left| \int_e (\xi^\rho \phi^u)_e^{E^1} n_x[\tau^\rho] \right| \right. \\ &\quad \left. + \left| \int_e ((\phi^u - \tilde{u})\rho)_e^{E^1} n_x[\tau^\rho] \right| + \left| \int_e (\xi^u \rho)_e^{E^1} n_x[\tau^\rho] \right| \right) =: \tilde{\text{I}}. \end{aligned}$$

We now use the Cauchy–Schwarz inequality, the trace inequality (2.3), the inequality (2.5), the assumption (3.2), the approximation inequality (2.2), and the bound on ϕ^u from Lemma 5.1 to obtain the bound on $\tilde{\text{I}}$:

$$\tilde{\text{I}} \leq \frac{1}{2} \|\tau^\rho\|_{0,\Omega}^2 + K \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \|\tau^\rho\|_{0,e}^2 + C^* \sum_{\alpha=\rho,u} \frac{h^{2 \min(r_\alpha+1, s_\alpha)}}{r_\alpha^{2s_\alpha}} + C^{**} \|\phi^u - \tilde{u}\|_{0,\Omega}^2.$$

A similar bound can be derived for the second term II on the RHS of (5.20). To estimate the last term on the RHS of (5.20), we first use (5.14) and the definition of

the one-sided local speeds (5.10) to obtain

$$\text{III} \leq C \sum_{e \in \Gamma_h^{\text{ver}}} \left(\|[\tau^\rho]\|_{0,e}^2 + \left| \int_e [\xi^\rho][\tau^\rho] \right| \right) := \widetilde{\text{III}}.$$

Then, using the Cauchy–Schwarz inequality, the trace inequality (2.3), and the approximation inequality (2.2), we bound $\widetilde{\text{III}}$ as follows:

$$\widetilde{\text{III}} \leq \left(\frac{K_1 h}{r_\rho^2} + K_2 \right) \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \|[\tau^\rho]\|_{0,e}^2 + C \frac{h^{2 \min(r_\rho+1, s_\rho)}}{r_\rho^{2s_\rho}}.$$

Combining the above bounds on I, II, and III, we arrive at

$$|T_3^\rho| \leq \|\tau^\rho\|_{0,\Omega}^2 + C_3^\rho \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \|[\tau^\rho]\|_{0,e}^2 + C^* \sum_{\alpha \in \{\rho, u\}} \frac{h^{2 \min(r_\alpha+1, s_\alpha)}}{r_\alpha^{2s_\alpha}} + C^{**} \|\phi^u - \tilde{u}\|_{0,\Omega}^2. \tag{5.21}$$

The terms T_4^ρ and T_5^ρ are bounded in the same way as the terms T_2^ρ and T_3^ρ , respectively, and the bounds are

$$|T_4^\rho| \leq \varepsilon_2^\rho \|(\tau^\rho)_y\|_{0,\Omega}^2 + C_4^\rho \|\tau^\rho\|_{0,\Omega}^2 \tag{5.22}$$

and

$$|T_5^\rho| \leq \|\tau^\rho\|_{0,\Omega}^2 + C_5^\rho \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \|[\tau^\rho]\|_{0,e}^2 + C^* \sum_{\alpha \in \{\rho, v\}} \frac{h^{2 \min(r_\alpha+1, s_\alpha)}}{r_\alpha^{2s_\alpha}} + C^{**} \|\phi^v - \tilde{v}\|_{0,\Omega}^2. \tag{5.23}$$

The term T_6^ρ is bounded using the Cauchy–Schwarz inequality and the approximation inequality (2.2):

$$|T_6^\rho| \leq \|\tau^\rho\|_{0,\Omega}^2 + C^* \frac{h^{2 \min(r_\rho+1, s_\rho)}}{r_\rho^{2s_\rho}}. \tag{5.24}$$

Using the Cauchy–Schwarz inequality, Young’s inequality, and the approximation inequality (2.2) for ρ , we obtain the following bound for the term T_7^ρ :

$$|T_7^\rho| \leq \varepsilon_7^\rho \|\nabla \tau^\rho\|_{0,\Omega}^2 + C^* \frac{h^{2 \min(r_\rho+1, s_\rho)-2}}{r_\rho^{2s_\rho-2}}. \tag{5.25}$$

The term T_8^ρ is bounded using the Cauchy–Schwarz inequality, the trace inequality (2.4), and the approximation inequality (2.2):

$$|T_8^\rho| \leq C_8^\rho \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \|[\tau^\rho]\|_{0,e}^2 + C^* \frac{h^{2 \min(r_\rho+1, s_\rho)-2}}{r_\rho^{2s_\rho-2}}. \tag{5.26}$$

To bound the term T_9^ρ , we use the trace inequality (2.5), inequality (2.3), the Cauchy–Schwarz inequality, and Young’s inequality:

$$|T_9^\rho| \leq \varepsilon_9^\rho \|\nabla \tau^\rho\|_{0,\Omega}^2 + C^* \frac{h^{2 \min(r_\rho+1, s_\rho)-2}}{r_\rho^{2s_\rho-4}}. \tag{5.27}$$

Similarly, we bound the term T_{10}^ρ by

$$(5.28) \quad |T_{10}^\rho| \leq C_{10}^\rho \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \|[\tau^\rho]\|_{0,e}^2 + C^* \frac{h^{2 \min(r_\rho+1, s_\rho)-2}}{r_\rho^{2s_\rho-4}}.$$

For the terms T_{11}^ρ and T_{12}^ρ , we use our assumption on the smoothness of the exact solution together with the Cauchy–Schwarz inequality and the approximation inequality (2.2) to obtain the following bounds:

$$(5.29) \quad |T_{11}^\rho| \leq \varepsilon_{11}^\rho \|(\tau^\rho)_x\|_{0,\Omega}^2 + C^* \frac{h^{2 \min(r_\rho+1, s_\rho)}}{r_\rho^{2s_\rho}}, \quad |T_{12}^\rho| \leq \varepsilon_{11}^\rho \|(\tau^\rho)_y\|_{0,\Omega}^2 + C^* \frac{h^{2 \min(r_\rho+1, s_\rho)}}{r_\rho^{2s_\rho}}.$$

Consider now the term T_{13}^ρ . We first use (5.14) to obtain

$$|T_{13}^\rho| \leq C \sum_{E \in \mathcal{E}_h} \left(\left| \int_E (\phi^u - \tilde{u})(\tau^\rho)_x \right| + \left| \int_E \xi^u(\tau^\rho)_x \right| \right).$$

Then we apply the Cauchy–Schwarz inequality and the approximation inequality (2.2), which result in

$$(5.30) \quad |T_{13}^\rho| \leq \varepsilon_{13}^\rho \|(\tau^\rho)_x\|_{0,\Omega}^2 + C^* \frac{h^{2 \min(r_u+1, s_u)}}{r_u^{2s_u}} + C^{**} \|\phi^u - \tilde{u}\|_{0,\Omega}^2.$$

The bound on the term T_{14}^ρ is obtained in the same way as the bound on T_{13}^ρ :

$$(5.31) \quad |T_{14}^\rho| \leq \varepsilon_{13}^\rho \|(\tau^\rho)_y\|_{0,\Omega}^2 + C^* \frac{h^{2 \min(r_v+1, s_v)}}{r_v^{2s_v}} + C^{**} \|\phi^v - \tilde{v}\|_{0,\Omega}^2.$$

Finally, we plug the estimates (5.18)–(5.19) and (5.21)–(5.31) into (5.17) and use the assumption that $h < 1$ to obtain

$$(5.32) \quad \begin{aligned} & \frac{1}{2} \frac{d}{dt} \|\tau^\rho\|_{0,\Omega}^2 + (1 - \varepsilon_1^\rho - \varepsilon_2^\rho - \varepsilon_7^\rho - \varepsilon_9^\rho - \varepsilon_{11}^\rho - \varepsilon_{13}^\rho) \|\nabla \tau^\rho\|_{0,\Omega}^2 \\ & + (\sigma_\rho - C_1^\rho - C_3^\rho - C_5^\rho - C_8^\rho - C_{10}^\rho) \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \|[\tau^\rho]\|_{0,e}^2 \leq (3 + C_2^\rho + C_4^\rho) \|\tau^\rho\|_{0,\Omega}^2 \\ & + C_\rho^* \left(\frac{h^{2 \min(r_\rho+1, s_\rho)-2}}{r_\rho^{2s_\rho-4}} + \sum_{\alpha \in \{u,v\}} \frac{h^{2 \min(r_\alpha+1, s_\alpha)}}{r_\alpha^{2s_\alpha}} \right) + C^{**} \sum_{\alpha \in \{u,v\}} (\|\phi^\alpha - \tilde{\alpha}\|_{0,\Omega}^2). \end{aligned}$$

We now choose ε_i^ρ and the penalty parameter σ_ρ so that the coefficients of $\|\nabla \tau^\rho\|_{0,\Omega}^2$ and $\sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \|[\tau^\rho]\|_{0,e}^2$ on the LHS of (5.32) are equal to 1/2. We then multiply (5.32) by 2 and integrate it in time from 0 to t . Taking into account that $(\phi^u, \phi^v) \in S$ and using the fact that $\tau^0 = 0$, we obtain

$$(5.33) \quad \|\tau^\rho\|_{0,\Omega}^2 + \int_0^t \left(\|\nabla \tau^\rho\|_{0,\Omega}^2 + \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \|[\tau^\rho]\|_{0,e}^2 \right) \leq \tilde{C}^\rho \int_0^t \|\tau^\rho\|_{0,\Omega}^2 + C^{uv} \mathcal{E}_1.$$

Next, we apply Gronwall's Lemma 2.6 and take the supremum with respect to t of both sides of (5.33):

$$(5.34) \quad \sup_{[0,T]} \|\tau^\rho\|_{0,\Omega}^2 + \int_0^T \left(\|\nabla \tau^\rho\|_{0,\Omega}^2 + \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \|[\tau^\rho]\|_{0,e}^2 \right) \leq C^I \mathcal{E}_1,$$

where \mathcal{E}_1 is given in (5.3) and C^I is a constant that depends on $\|\rho\|_{(L^\infty([0,T]);H^2(\Omega))}$, $\|\rho_t\|_{(L^\infty([0,T]);L^2(\Omega))}$, $\|u\|_{(L^\infty([0,T]);L^2(\Omega))}$, $\|v\|_{(L^\infty([0,T]);L^2(\Omega))}$, and T only.

According to the definition of the broken Sobolev space S given on p. 393, the estimate (5.34) implies that $\phi_L^\rho \in S$. Using similar techniques, it can be shown that $(\phi_L^c, \phi_L^u, \phi_L^v) \in S$ as well (see [20] for the detailed proof). Therefore, we have shown that $A(S) \subset S$, and the proof of Theorem 5.3 is now complete. \square

Let us recall that our goal is to show that the operator A has a fixed point. Equipped with Theorem 5.3, it remained to prove that A is compact. To this end, we need to show that A is continuous and equicontinuous.

LEMMA 5.4. *The operator A is continuous and equicontinuous.*

Proof. We consider the sequence $\{(\phi_n^\rho, \phi_n^c, \phi_n^u, \phi_n^v)\}$ and assume that

$$\sup_{t \in [0, T]} (\|(\phi_n^\rho, \phi_n^c, \phi_n^u, \phi_n^v) - (\phi^\rho, \phi^c, \phi^u, \phi^v)\|_S) \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Let

$$(5.35) \quad (\phi_{L,n}^\rho, \phi_{L,n}^c, \phi_{L,n}^u, \phi_{L,n}^v) = A(\phi_n^\rho, \phi_n^c, \phi_n^u, \phi_n^v)$$

and

$$(5.36) \quad (\phi_L^\rho, \phi_L^c, \phi_L^u, \phi_L^v) = A(\phi^\rho, \phi^c, \phi^u, \phi^v)$$

be two solutions of (5.5)–(5.8). We denote by $(\widehat{\phi}_L^\rho, \widehat{\phi}_L^c, \widehat{\phi}_L^u, \widehat{\phi}_L^v)$ the difference between these two solutions (note that $(\widehat{\phi}_L^{\rho,0}, \widehat{\phi}_L^{c,0}, \widehat{\phi}_L^{u,0}, \widehat{\phi}_L^{v,0}) = (0, 0, 0, 0)$), subtract (5.36) from (5.35), and choose the test function in the resulting equation for ρ to be $w^\rho = \widehat{\phi}_L^\rho$. This yields

$$\begin{aligned} & \frac{1}{2} \frac{d}{dt} \|\widehat{\phi}_L^\rho\|_{0,\Omega}^2 + \|\widehat{\nabla} \widehat{\phi}_L^\rho\|_{0,\Omega}^2 + \sigma_\rho \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \|\llbracket \widehat{\phi}_L^\rho \rrbracket\|_{0,e}^2 \\ &= (1 - \varepsilon) \sum_{e \in \Gamma_h} \int_e \{ \widehat{\nabla} \widehat{\phi}_L^\rho \cdot \mathbf{n}_e \} \llbracket \widehat{\phi}_L^\rho \rrbracket + \sum_{E \in \mathcal{E}_h} \int_E \chi \widehat{\phi}_L^\rho \phi^u (\widehat{\phi}_L^\rho)_x \\ & \quad + \sum_{E \in \mathcal{E}_h} \int_E \chi \phi_{L,n}^\rho (\phi_n^u - \phi^u) (\widehat{\phi}_L^\rho)_x \\ & \quad - \sum_{e \in \Gamma_h^{\text{ver}}} \int_e (\chi \widehat{\phi}_L^\rho \phi^u)^* n_x \llbracket \widehat{\phi}_L^\rho \rrbracket + \sum_{e \in \Gamma_h^{\text{ver}}} \int_e ((\chi \phi_{L,n}^\rho \phi^u)^* - (\chi \phi_{L,n}^\rho \phi_n^u)^*) n_x \llbracket \widehat{\phi}_L^\rho \rrbracket \\ & \quad + \sum_{E \in \mathcal{E}_h} \int_E \chi \widehat{\phi}_L^\rho \phi^v (\widehat{\phi}_L^\rho)_y + \sum_{E \in \mathcal{E}_h} \int_E \chi \phi_{L,n}^\rho (\phi_n^v - \phi^v) (\widehat{\phi}_L^\rho)_y \\ & \quad - \sum_{e \in \Gamma_h^{\text{hor}}} \int_e (\chi \widehat{\phi}_L^\rho \phi^v)^* n_y \llbracket \widehat{\phi}_L^\rho \rrbracket \\ (5.37) \quad & + \sum_{e \in \Gamma_h^{\text{hor}}} \int_e ((\chi \phi_{L,n}^\rho \phi^v)^* - (\chi \phi_{L,n}^\rho \phi_n^v)^*) n_y \llbracket \widehat{\phi}_L^\rho \rrbracket =: R_1 + R_2 + \dots + R_9. \end{aligned}$$

We now bound each term on the RHS of (5.37).

The term R_1 can be bounded using the Cauchy–Schwarz inequality, Young’s inequality, and the inequality (2.5):

$$(5.38) \quad |R_1| \leq \frac{1}{6} \|\widehat{\nabla} \widehat{\phi}_L^\rho\|_{0,\Omega}^2 + C_1 \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \|\llbracket \widehat{\phi}_L^\rho \rrbracket\|_{0,e}^2.$$

Next, applying the Cauchy–Schwarz and Young’s inequalities and using the boundness of $\|\phi^u\|_{\infty,\Omega}$, established in Lemma 5.1, we obtain the following bound on R_2 :

$$(5.39) \quad |R_2| \leq \frac{1}{6} \left\| \left(\widehat{\phi}_L^\rho \right)_x \right\|_{0,\Omega}^2 + C_2 \left\| \widehat{\phi}_L^\rho \right\|_{0,\Omega}^2.$$

Using the Cauchy–Schwarz and Young’s inequalities and the fact that $\phi_{L,n}^\rho \in S$, we bound the term R_3 by

$$(5.40) \quad |R_3| \leq \frac{1}{6} \left\| \left(\widehat{\phi}_L^\rho \right)_x \right\|_{0,\Omega}^2 + C_3 \|\phi_n^u - \phi^u\|_{0,\Omega}^2.$$

We then use the Cauchy–Schwarz inequality, the inequality (2.5), and the first numerical flux formula in (5.9) to estimate R_4 :

$$(5.41) \quad |R_4| \leq \left\| \widehat{\phi}_L^\rho \right\|_{0,\Omega}^2 + C_4 \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \left\| \left[\widehat{\phi}_L^\rho \right] \right\|_{0,e}^2.$$

We now consider the term R_5 . It follows from formulas (5.9)–(5.10) that the numerical fluxes $(\chi \phi_{L,n}^\rho \phi^u)^*$ are the composition of the continuous functions with respect to the variables $(\phi^u)_e^{E^1}$ and $(\phi^u)_e^{E^2}$. Hence, we can apply the Cauchy–Schwarz inequality and the inequality (2.5) to R_5 so that it is bounded by

$$(5.42) \quad |R_5| \leq \|(\chi \phi_{L,n}^\rho \phi^u)^* - (\chi \phi_{L,n}^\rho \phi_n^u)^*\|_{0,\Omega}^2 + C_5 \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \left\| \left[\widehat{\phi}_L^\rho \right] \right\|_{0,e}^2.$$

The terms R_6 , R_7 , R_8 , and R_9 are similar to the terms R_2 , R_3 , R_4 , and R_5 estimated in (5.39), (5.40), (5.41), and (5.42), respectively. Therefore, we obtain

$$(5.43) \quad |R_6| \leq \frac{1}{6} \left\| \left(\widehat{\phi}_L^\rho \right)_y \right\|_{0,\Omega}^2 + C_6 \left\| \widehat{\phi}_L^\rho \right\|_{0,\Omega}^2,$$

$$(5.44) \quad |R_7| \leq \frac{1}{6} \left\| \left(\widehat{\phi}_L^\rho \right)_y \right\|_{0,\Omega}^2 + C_7 \|\phi_n^v - \phi^v\|_{0,\Omega}^2,$$

$$(5.45) \quad |R_8| \leq \left\| \widehat{\phi}_L^\rho \right\|_{0,\Omega}^2 + C_8 \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \left\| \left[\widehat{\phi}_L^\rho \right] \right\|_{0,e}^2,$$

$$(5.46) \quad |R_9| \leq \|(\chi \phi_{L,n}^\rho \phi^v)^* - (\chi \phi_{L,n}^\rho \phi_n^v)^*\|_{0,\Omega}^2 + C_9 \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \left\| \left[\widehat{\phi}_L^\rho \right] \right\|_{0,e}^2.$$

Substituting the estimates (5.38)–(5.46) into (5.37) yields

$$\begin{aligned} & \frac{1}{2} \frac{d}{dt} \left\| \widehat{\phi}_L^\rho \right\|_{0,\Omega}^2 + \frac{1}{2} \left\| \nabla \widehat{\phi}_L^\rho \right\|_{0,\Omega}^2 + (\sigma_\rho - C) \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \left\| \left[\widehat{\phi}_L^\rho \right] \right\|_{0,e}^2 \leq C^* \left\| \widehat{\phi}_L^\rho \right\|_{0,\Omega}^2 \\ & + C^{**} \sum_{\alpha \in \{u,v\}} \left(\|\phi_n^\alpha - \phi^\alpha\|_{0,\Omega}^2 + \|(\chi \phi_{L,n}^\rho \phi^\alpha)^* - (\chi \phi_{L,n}^\rho \phi_n^\alpha)^*\|_{0,\Omega}^2 \right), \end{aligned}$$

where the penalty parameter σ_ρ is chosen sufficiently large so that the coefficient $(\sigma_\rho - C)$ is nonnegative.

We now integrate the latter inequality with respect to time from 0 to t and apply Gronwall's Lemma 2.6 to obtain

$$\begin{aligned} & \left\| \widehat{\phi}_L^\rho \right\|_{0,\Omega}^2 + \int_0^t \left(\left\| \nabla \widehat{\phi}_L^\rho \right\|_{0,\Omega}^2 + (\sigma_\rho - C) \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \left\| [\widehat{\phi}_L^\rho] \right\|_{0,e}^2 \right) dt \leq M \left(\left\| \phi_L^{\rho,0} \right\|_{0,\Omega}^2 \right. \\ & \left. + \sum_{\alpha \in \{u,v\}} \int_0^t \left(\left\| \phi_n^\alpha - \phi^\alpha \right\|_{0,\Omega}^2 + \left\| (\chi \phi_{L,n}^\rho \phi^\alpha)^* - (\chi \phi_{L,n}^\rho \phi_n^\alpha)^* \right\|_{0,\Omega}^2 \right) \right). \end{aligned}$$

Finally, taking the supremum over t and since $\widehat{\phi}_L^{\rho,0} = 0$, we arrive at

$$\begin{aligned} & \sup_{t \in [0,T]} \left\| \widehat{\phi}_L^\rho \right\|_{0,\Omega}^2 + \int_0^T \left(\left\| \nabla \widehat{\phi}_L^\rho \right\|_{0,\Omega}^2 + \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \left\| [\widehat{\phi}_L^\rho] \right\|_{0,e}^2 \right) dt \\ & \leq M^* \sum_{\alpha \in \{u,v\}} \int_0^T \left(\left\| \phi_n^\alpha - \phi^\alpha \right\|_{0,\Omega}^2 + \left\| (\chi \phi_{L,n}^\rho \phi^\alpha)^* - (\chi \phi_{L,n}^\rho \phi_n^\alpha)^* \right\|_{0,\Omega}^2 \right) dt. \end{aligned}$$

This inequality together with the similar inequalities for $\widehat{\phi}^c$, $\widehat{\phi}^u$, and $\widehat{\phi}^v$, which can be obtained in an analogous way, imply continuity of the operator A .

Applying similar techniques to the difference $(\overline{\phi}_L^\rho, \overline{\phi}_L^c, \overline{\phi}_L^u, \overline{\phi}_L^v) := (\phi_L^\rho, \phi_L^c, \phi_L^u, \phi_L^v)(t_1, x_1, y_1) - (\phi_L^\rho, \phi_L^c, \phi_L^u, \phi_L^v)(t_2, x_2, y_2)$ and using the fact that $(\phi^u, \phi^v) \in S$, one can show that the operator A is equicontinuous. \square

Equipped with Lemma 5.4, we conclude that the operator A is compact. Hence, by the second Schauder fixed-point theorem [25], it has at least one fixed point $(\rho^{\text{DG}}, c^{\text{DG}}, u^{\text{DG}}, v^{\text{DG}})$, which is the DG solution of (3.3)–(3.6). For this solution, we establish the convergence rate results, stated in the following theorem.

THEOREM 5.5 ($L^2(H^1)$ - and $L^\infty(L^2)$ -error estimates). *Let the solution of the Keller–Segel system (1.3)–(1.6) satisfy the smoothness assumption (5.2). If the penalty parameters σ_ρ , σ_c , σ_u , and σ_v in the DG method (3.3)–(3.9) are sufficiently large and $r_{\min} \geq 2$, then there exist constants C_ρ and C_c , independent of h , r_ρ , r_c , r_u , and r_v , such that the following two error estimates hold:*

$$\begin{aligned} & \left\| \rho^{\text{DG}} - \rho \right\|_{L^\infty([0,T];L^2(\Omega))} + \left\| \nabla (\rho^{\text{DG}} - \rho) \right\|_{L^2([0,T];L^2(\Omega))} \\ & \quad + \left(\int_0^T \sum_{e \in \Gamma_h} \frac{r_\rho^2}{|e|} \left\| [\rho^{\text{DG}} - \rho] \right\|_{0,e}^2 \right)^{\frac{1}{2}} \leq C_\rho E, \\ & \left\| c^{\text{DG}} - c \right\|_{L^\infty([0,T];L^2(\Omega))} + \left\| \nabla (c^{\text{DG}} - c) \right\|_{L^2([0,T];L^2(\Omega))} \\ & \quad + \left(\int_0^T \sum_{e \in \Gamma_h} \frac{r_c^2}{|e|} \left\| [c^{\text{DG}} - c] \right\|_{0,e}^2 \right)^{\frac{1}{2}} \leq C_c E, \end{aligned}$$

where

$$E := \sum_{\alpha \in \{\rho, c, u, v\}} \frac{h^{\min(r_\alpha+1, s_\alpha)-1}}{r_\alpha^{s_\alpha-2}}.$$

Proof. The result follows from the definition of space S , the fact that the DG solution is a fixed point of the compact operator A (defined above), the hp approximation Lemma 2.1, and the triangle inequality. \square

Remark. The obtained error estimates are h -optimal but only suboptimal for r .

Finally, equipped with the results established in Theorem 5.5, we obtain the following bound for the blow-up time of the exact solution of the Keller–Segel system.

THEOREM 5.6. *Let us denote by t_b the blow-up time of the exact solution of the Keller–Segel system (1.1) and by t_b^{DG} the blow-up time of the DG solution of (3.3)–(3.9). Then $t_b \leq t_b^{\text{DG}}$.*

Proof. The solution ρ of the Keller–Segel model blows up if $\|\rho\|_{L^\infty(\Omega)}$ becomes unbounded in either finite or infinite time (see, e.g., [27, 28]). Therefore, in order to prove the theorem we need to establish an L^∞ -error bound.

From Theorem 5.5 we have the following L^2 -error bound: $\|\rho^{\text{DG}} - \rho\|_{L^2(\Omega)} \leq C_\rho E$, and hence from (2.6) we obtain $\|\rho^{\text{DG}} - \rho\|_{L^\infty(\Omega)} \leq C_\rho E_1$, which, in turn, implies that $\|\rho^{\text{DG}}\|_{L^\infty(\Omega)} \leq \|\rho\|_{L^\infty(\Omega)} + C_\rho E_1$, where $E_1 := \sum_{\alpha \in \{\rho, c, u, v\}} \frac{h^{\min(r_\alpha + 1, s_\alpha) - 2}}{r_\alpha^{s_\alpha - 3}}$. From the last estimate the statement of the theorem follows. \square

6. Numerical example. In this section, we demonstrate the performance of the proposed DG method. In all of our numerical experiments, we have used the third-order strong stability-preserving Runge–Kutta method for the time discretization [24]. No slope-limiting technique has been implemented. The values of the penalty parameters used are $\sigma_\rho = \sigma_c = 1$ and $\sigma_u = \sigma_v = 0.01$. We note that no instabilities have been observed when the latter two parameters were taken as zero; however, since our convergence proof requires σ_u and σ_v to be positive, we show only the results obtained with positive σ_u and σ_v , which are almost identical to the ones obtained with $\sigma_u = \sigma_v = 0$.

We consider the initial boundary value problem for the Keller–Segel system in the square domain $[-\frac{1}{2}, \frac{1}{2}] \times [-\frac{1}{2}, \frac{1}{2}]$. We take the chemotactic sensitivity $\chi = 1$ and the bell-shaped initial data

$$\rho(x, y, 0) = 1200e^{-120(x^2+y^2)}, \quad c(x, y, 0) = 600e^{-60(x^2+y^2)}.$$

According to the results in [26], both components ρ and c of the solution are expected to blow up at the origin in finite time. This situation is especially challenging since capturing the blowing up solution with shrinking support is extremely hard.

In Figures 6.1–6.4, we plot the logarithmically scaled density $\ln(1 + \rho^{\text{DG}})$ computed at different times on two different uniform grids with $h = 1/51$ (Figures 6.1 and 6.3) and $h = 1/101$ (Figures 6.2 and 6.4). The results shown in Figures 6.1–6.2 have been obtained with quadratic polynomials (i.e., $r_\rho = r_c = r_u = r_v = r = 2$), while the solution shown in Figures 6.3–6.4 have been computed with the help of cubic polynomials (i.e., $r_\rho = r_c = r_u = r_v = r = 3$).

Numerical convergence of the scheme is verified by refining the mesh and by increasing the polynomial degree. As one can see, the computed solutions are in very good agreement at the smaller times ($t = 1.46 \cdot 10^{-5}$, $2.99 \cdot 10^{-5}$, and $6.03 \cdot 10^{-5}$). However, at time close to the blow-up time ($t = 1.21 \cdot 10^{-4}$) the maximum value of ρ^{DG} grows, while its support shrinks, and no mesh-refinement convergence is observed: the numerical solution keeps increasing when the mesh is refined. Using Theorem 5.6, we can conclude that in this example the blow-up time of the exact solution is less than or equal to the blow-up time of the DG solution, which is approximately $t_b^{\text{DG}} \approx 1.21 \cdot 10^{-4}$.

We note that even though no slope-limiting or any other positivity-preserving techniques have been implemented, the computed solutions have never developed negative values and are oscillation-free.

Finally, we check the numerical order of the convergence of the proposed DG method. We first consider the smooth solution at a very small time $t = 1.0 \cdot 10^{-7}$

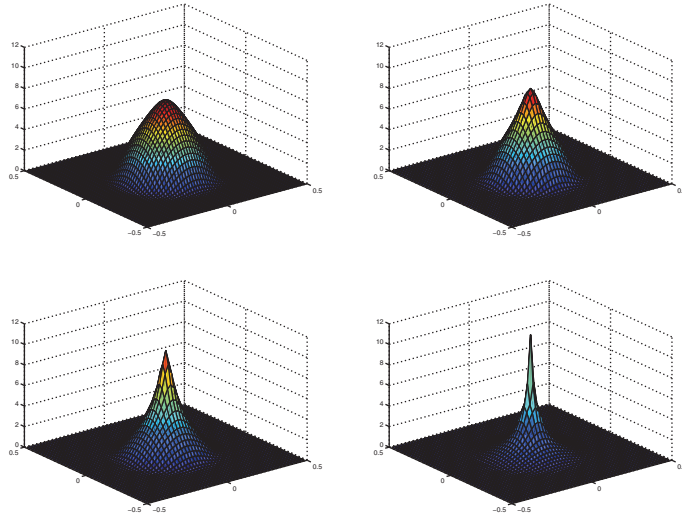


FIG. 6.1. $h = 1/51, r = 2$. Logarithmically scaled density computed at $t = 1.46 \cdot 10^{-5}$ (top left), $t = 2.99 \cdot 10^{-5}$ (top right), $t = 6.03 \cdot 10^{-5}$ (bottom left), and $t = 1.21 \cdot 10^{-4} \approx t_b^{\text{DG}}$ (bottom right).

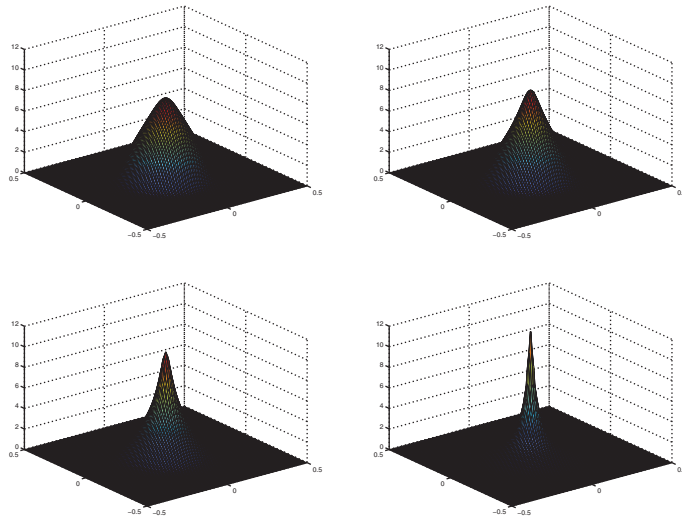


FIG. 6.2. The same as in Figure 6.1 but with $h = 1/101, r = 2$.

and test the convergence with respect to the mesh size h for the fixed $r = 2$ (piecewise quadratic polynomials). Since the exact solution for the Keller–Segel system is unavailable, we compute the reference solution by the proposed DG method on a fine mesh with $h = 1/128$ and using the fifth-order ($r = 5$) piecewise polynomials. We then use the obtained reference solution to compute the relative L^2 - and relative H^1 -errors. These errors are presented in Table 6.1. From this table, one can see that the solution numerically converges to the reference solution with the (optimal) second order in the H^1 -norm which confirms the theoretical results predicted by our convergence analysis. Moreover, the achieved third order of convergence in the L^2 -norm is optimal for quadratic piecewise polynomials.

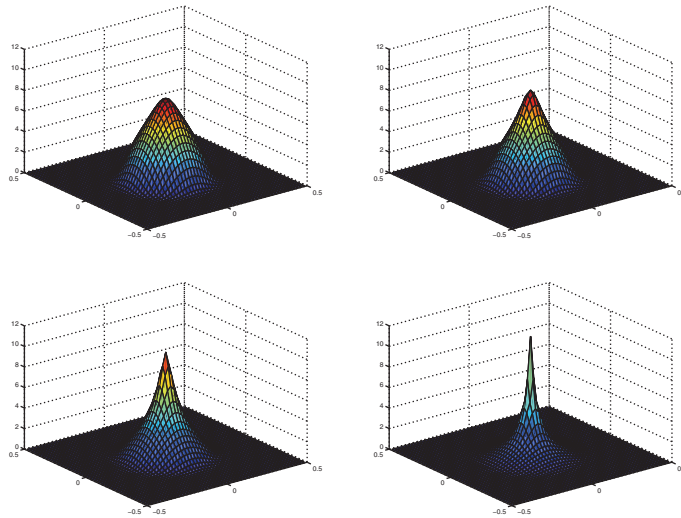


FIG. 6.3. *The same as in Figures 6.1–6.2 but with $h = 1/51, r = 3$.*

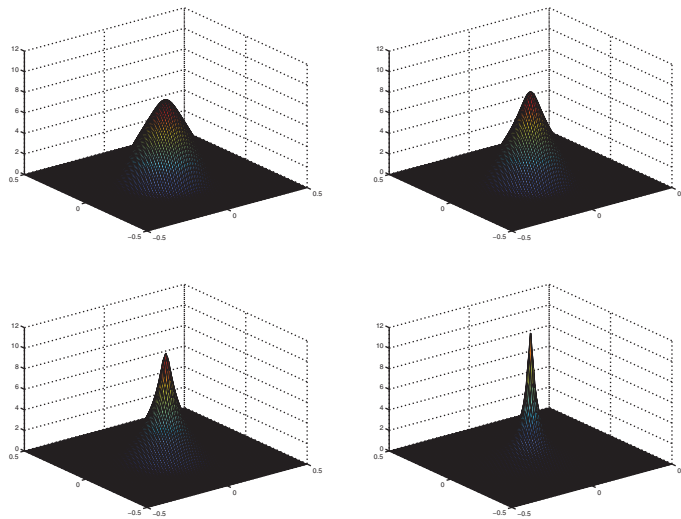


FIG. 6.4. *The same as in Figures 6.1–6.3 but with $h = 1/101, r = 3$.*

We then test the convergence of the proposed DG method with respect to the degree r of piecewise polynomials for the fixed $h = 1/32$. The obtained results, reported in Table 6.2, show that the error decreases almost exponentially when the polynomial degree increases (this is a typical situation when DG methods capture smooth solutions).

We also compute the L^2 -errors with respect to the reference solution for the solutions plotted on Figures 6.1 and 6.2 at times $t = 2.99 \cdot 10^{-5}$ and $t = 6.03 \cdot 10^{-5}$. These times are close to the blow-up time, and the solutions develop a pick at the origin. The obtained errors are reported in Table 6.3. As one can see, even for the spiky solutions, the convergence rate is very high though it, as expected, deteriorates as t approaches t_b^{DG} .

TABLE 6.1
Relative errors as functions of the mesh size h ; $r = 2$ is fixed.

h	L^2 -error	Rate	H^1 -error	Rate
1/4	3.0578	–	1.5591	–
1/8	1.0290	1.6	1.2348	0.35
1/16	0.0796	3.7	0.5206	1.3
1/32	0.0075	3.4	0.0937	2.5
1/64	0.0006	3.6	0.0157	2.6

TABLE 6.2
Relative errors as functions of the piecewise polynomial degree r ; $h = 1/32$ is fixed.

r	L^2 -error	Rate	H^1 -error	Rate
2	7.5e-03	–	9.4e-02	–
3	9.0e-04	5.2	2.2e-02	3.6
4	8.0e-05	8.4	2.6e-03	7.4
5	6.9e-06	11.0	2.9e-04	9.8

TABLE 6.3
Relative L^2 -errors at two different times; $r = 2$ is fixed.

h	$t = 2.99 \cdot 10^{-5}$		$t = 6.03 \cdot 10^{-5}$	
	L^2 -error	Rate	L^2 -error	Rate
1/51	5.5e-02	–	5.0e-02	–
1/101	5.2e-03	3.4	1.1e-02	2.2

REFERENCES

- [1] J. ADLER, *Chemotaxis in bacteria*, Annu. Rev. Biochem., 44 (1975), pp. 341–356.
- [2] S. AGMON, *Lectures on Elliptic Boundary Value Problems*, Van Nostrand, Princeton, NJ, 1965.
- [3] V. AZINGER, C. DAWSON, B. COCKBURN, AND P. CASTILLO, *Local discontinuous Galerkin methods for contaminant transport*, Adv. Water Res., 24 (2000), pp. 73–87.
- [4] D. N. ARNOLD, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.
- [5] I. BABUŠKA AND M. SURI, *The h - p version of the finite element method with quasiuniform meshes*, RAIRO Modél. Math. Numér., 21 (1987), pp. 199–238.
- [6] I. BABUŠKA AND M. SURI, *The optimal convergence rate of the p -version of the finite element method*, SIAM J. Numer. Anal., 24 (1987), pp. 750–776.
- [7] G. A. BAKER, W. N. JUREIDINI, AND O. A. KARAKASHIAN, *Piecewise solenoidal vector fields and the Stokes problem*, SIAM J. Numer. Anal., 27 (1990), pp. 1466–1485.
- [8] J. BONNER, *The Cellular Slime Molds*, 2nd ed., Princeton University Press, Princeton, NJ, 1967.
- [9] S. C. BRENNER, *Poincaré–Friedrichs inequalities for piecewise H^1 functions*, SIAM J. Numer. Anal., 41 (2003), pp. 306–324.
- [10] E. BUDRENE AND H. BERG, *Complex patterns formed by motile cells of escherichia coli*, Nat., 349 (1991), pp. 630–633.
- [11] E. BUDRENE AND H. BERG, *Dynamics of formation of symmetrical patterns by chemotactic bacteria*, Nat., 376 (1995), pp. 49–53.
- [12] A. CHERTOCK AND A. KURGANOV, *A positivity preserving central-upwind scheme for chemotaxis and haptotaxis models*, Numer. Math., DOI 10.1007/s00211-008-0188-0.
- [13] S. CHILDRESS AND J. PERCUS, *Nonlinear aspects of chemotaxis*, Math. Biosci., 56 (1981), pp. 217–237.
- [14] B. COCKBURN, G. KARNIADAKIS, AND C.-W. SHU, EDS., *First International Symposium on Discontinuous Galerkin Methods*, Lecture Notes in Comput. Sci. Engrg. 11, Springer-Verlag, Berlin, 2000.
- [15] B. COCKBURN AND C.-W. SHU, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.

- [16] M. COHEN AND A. ROBERTSON, *Wave propagation in the early stages of aggregation of cellular slime molds*, J. Theoret. Biol., 31 (1971), pp. 101–118.
- [17] C. DAWSON, E. KUBATKO, AND J. WESTERINK, *hp discontinuous Galerkin methods for advection-dominated problems in shallow water*, Comput. Methods Appl. Mech. Engrg., to appear.
- [18] C. DAWSON, S. SUN, AND M. WHEELER, *Compatible algorithms for coupled flow and transport*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 2565–2580.
- [19] J. DOUGLAS AND T. DUPONT, *Interior penalty procedures for elliptic and parabolic Galerkin methods*, Lecture Notes in Physics, Vol. 58, Springer-Verlag, Berlin, 1976.
- [20] Y. EPSHTEYN AND A. KURGANOV, *New Discontinuous Galerkin Methods for the Keller-Segel Chemotaxis Model*, CNA report 07-CNA-006, Carnegie Mellon University, Pittsburgh, PA, 2007; available online from <http://www.math.cmu.edu/cna/pub2007.html>.
- [21] Y. EPSHTEYN AND B. RIVIÈRE, *On the solution of incompressible two-phase flow by a p-version discontinuous Galerkin method*, Comm. Numer. Methods Engrg., 22 (2006), pp. 741–751.
- [22] F. FILBET, *A finite volume scheme for the Patlak-Keller-Segel chemotaxis model*, Numer. Math., 104 (2006), pp. 457–488.
- [23] V. GIRAULT, B. RIVIÈRE, AND M. WHEELER, *A discontinuous Galerkin method with non-overlapping domain decomposition for the Stokes and Navier-Stokes problems*, Math. Comp., 74 (2005), pp. 53–84.
- [24] S. GOTTLIEB, C.-W. SHU, AND E. TADMOR, *Strong stability-preserving high-order time discretization methods*, SIAM Rev., 43 (2001), pp. 89–112.
- [25] D. GRIFFEL, *Applied Functional Analysis*, Dover, New York, 2002.
- [26] M. HERRERO AND J. VELÁZQUEZ, *A blow-up mechanism for a chemotaxis model*, Ann. Sc. Norm. Super. Pisa Cl. Sci., 24 (1997), pp. 633–683.
- [27] D. HORSTMANN, *From 1970 until now: The Keller-Segel model in chemotaxis and its consequences I*, Jahresber. Deutsch. Math.-Verein., 105 (2003), pp. 103–165.
- [28] D. HORSTMANN, *From 1970 until now: The Keller-Segel model in chemotaxis and its consequences II*, Jahresber. Deutsch. Math.-Verein., 106 (2004), pp. 51–69.
- [29] E. KELLER AND L. SEGEL, *Initiation of slime mold aggregation viewed as an instability*, J. Theoret. Biol., 26 (1970), pp. 399–415.
- [30] E. KELLER AND L. SEGEL, *Model for chemotaxis*, J. Theoret. Biol., 30 (1971), pp. 225–234.
- [31] E. KELLER AND L. SEGEL, *Traveling bands of chemotactic bacteria: A theoretical analysis*, J. Theoret. Biol., 30 (1971), pp. 235–248.
- [32] A. KURGANOV AND C.-T. LIN, *On the reduction of numerical dissipation in central-upwind schemes*, Commun. Comput. Phys., 2 (2007), pp. 141–163.
- [33] A. KURGANOV, S. NOELLE, AND G. PETROVA, *Semidiscrete central-upwind schemes for hyperbolic conservation laws and Hamilton–Jacobi equations*, SIAM J. Sci. Comput., 23 (2001), pp. 707–740.
- [34] A. KURGANOV AND G. PETROVA, *Central-upwind schemes on triangular grids for hyperbolic systems of conservation laws*, Numer. Methods Partial Differential Equations, 21 (2005), pp. 536–552.
- [35] A. MARROCCO, *2d simulation of chemotaxis bacteria aggregation*, M2AN Math. Model. Numer. Anal., 37 (2003), pp. 617–630.
- [36] V. NANJUNDIAH, *Chemotaxis, signal relaying and aggregation morphology*, J. Theoret. Biol., 42 (1973), pp. 63–105.
- [37] C. ORTNER AND E. SÜLI, *Discontinuous Galerkin finite element approximation of nonlinear second-order elliptic and hyperbolic systems*, SIAM J. Numer. Anal., 45 (2007), pp. 1370–1397.
- [38] C. PATLAK, *Random walk with persistence and external bias*, Bull. Math. Biophys., 15 (1953), pp. 311–338.
- [39] L. PRESCOTT, J. HARLEY, AND D. KLEIN, *Microbiology*, 3rd ed., W. C. Brown, Chicago, London, 1996.
- [40] B. RIVIÈRE, M. F. WHEELER, AND V. GIRAULT, *A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems*, SIAM J. Numer. Anal., 39 (2001), pp. 902–931.
- [41] S. SUN AND M. F. WHEELER, *Symmetric and nonsymmetric discontinuous Galerkin methods for reactive transport in porous media*, SIAM J. Numer. Anal., 43 (2005), pp. 195–219.
- [42] R. TYSON, S. LUBKIN, AND J. MURRAY, *A minimal mechanism for bacterial pattern formation*, Proc. R. Soc. Lond. Ser. B Biol. Sci., 266 (1999), pp. 299–304.
- [43] R. TYSON, L. STERN, AND R. LEVEQUE, *Fractional step methods applied to a chemotaxis model*, J. Math. Biol., 41 (2000), pp. 455–475.

GALERKIN FINITE ELEMENT METHODS WITH SYMMETRIC PRESSURE STABILIZATION FOR THE TRANSIENT STOKES EQUATIONS: STABILITY AND CONVERGENCE ANALYSIS*

ERIK BURMAN[†] AND MIGUEL A. FERNÁNDEZ[‡]

Abstract. We consider the stability and convergence analysis of pressure stabilized finite element approximations of the transient Stokes equation. The analysis is valid for a class of symmetric pressure stabilization operators, but also for standard, inf-sup stable, velocity/pressure spaces without stabilization. Provided the initial data are chosen as a specific (method-dependent) Ritz-projection, we get unconditional stability and optimal convergence for both pressure and velocity approximations, in natural norms. For arbitrary interpolations of the initial data, a condition between the space and time discretization parameters has to be verified in order to guarantee pressure stability.

Key words. transient Stokes equations, finite element methods, symmetric pressure stabilization, time discretization, Ritz-projection

AMS subject classifications. 65N12, 65N30, 76M10, 76D07

DOI. 10.1137/070707403

1. Introduction. In this paper we consider stabilized finite element methods for the transient Stokes problem. For methods of standard pressure stabilized Petrov–Galerkin (PSPG) or Galerkin least squares (GLS) type, the analysis of time-discretization schemes is a difficult issue, unless a space-time approach is applied with a discontinuous Galerkin discretization in time. Indeed, for standard finite difference type time discretizations, the finite difference term must be included in the stabilization operator to ensure consistency (see, e.g., [11, 23]). It has been shown in [3] that even for first order backward difference (BDF1) schemes this perturbs the stability of the numerical scheme when the time step is small, unless the following condition between the space mesh size and the time step is verified:

$$(1.1) \quad \delta t \geq Ch^2,$$

where δt denotes the time step and h the space discretization parameter. For higher order schemes, such as Crank–Nicholson or second order backward differencing, the strongly consistent scheme appears to be unstable (see, e.g., [1]). Similar initial time-step instabilities were observed in [19] for the algebraic (static) subscale stabilization scheme applied to the Navier–Stokes equations, and they were cured by including time dependent subscales.

Our goal in this work is to consider a fairly large class of pressure stabilization methods and show that convergence of velocities and pressures, for the transient Stokes problem, can be obtained without conditions on the space- and time-discretization parameters (like (1.1)), provided the initial data are chosen as a specific (method-dependent) Ritz-projection (see, e.g., [33, 34]) onto a space of discretely *divergence-free* functions. *Discretely divergence-free* should here be interpreted

*Received by the editors November 6, 2007; accepted for publication (in revised form) June 23, 2008; published electronically December 5, 2008.

<http://www.siam.org/journals/sinum/47-1/70740.html>

[†]Department of Mathematics, University of Sussex, Brighton, BN1 9RF, UK (E.N.Burman@sussex.ac.uk).

[‡]INRIA, Rocquencourt, B.P. 105, F–78153 Le Chesnay Cedex, France (miguel.fernandez@inria.fr).

in the sense of the stabilized method. If, on the other hand, the initial data are chosen as some interpolant that does not conserve the discrete divergence-free character, the condition

$$(1.2) \quad \delta t \geq \tilde{C}h^{2k},$$

with k the polynomial degree of the velocity approximation space, has to be respected in order to avoid pressure oscillations in the transient solution for small times.

Although the stability conditions (1.1) and (1.2) are similar, their natures are different. As mentioned above, if (1.2) fails to be satisfied, pressure instabilities appear when dealing with nondiscrete divergence-free initial velocity approximation, but they are not related to the structure of the pressure stabilization. For residual-based stabilization methods (PSPG, GLS, etc.) on the other hand, the finite difference/pressure coupling of the stabilization perturbs the coercivity of the discrete pressure operator (see [3]) unless condition (1.1) is satisfied (irrespective of the divergence-free character of the initial velocity approximation).

The analysis carried out in this paper is valid not only for pressure stabilization operators that are symmetric and weakly consistent but also for standard methods using inf-sup stable velocity/pressure pairs, but it does not apply to residual-based pressure stabilizations (PSPG, GLS, etc.). In particular, space and time discretizations commute (i.e., lead to the same fully discrete scheme) for the methods we analyze.

We prove unconditional stability of velocities and pressures and optimal convergence (in natural norms) when the initial data are chosen as a certain Ritz-type projection. In the case when a standard interpolation of the initial data is applied, an *inverse parabolic Courant–Friedrich–Lewy (CFL)-type* condition must be respected in order to maintain pressure stability for small time steps. We give the full analysis only for the backward difference formula of order one, and we indicate how the analysis changes in the case of second order approximations in time. Indeed, any \mathcal{A} -stable implicit scheme is expected to yield optimal performance.

The remainder of the paper is organized as follows. In the next section we introduce the problem under consideration and some useful notation. The space- and time-discretized formulations are introduced in section 3. In subsection 3.1, the space discretization is formulated using a general framework; we also discuss how some known pressure stabilized finite element methods enter this setting. The time discretization is performed in subsection 3.2 using the first order backward difference (BDF1), Crank–Nicholson, and second order backward difference (BDF2) schemes. Section 4 is devoted to the stability analysis of the resulting fully discrete formulations. The convergence analysis for the BDF1 scheme is carried out in section 5. We illustrate the theoretical results with some numerical experiments in section 6, using interior penalty stabilization of the gradient jumps. Finally, some conclusions are given in section 7.

2. Problem setting. Let Ω be a domain in \mathbb{R}^d ($d = 2$ or 3) with a polyhedral boundary $\partial\Omega$. For $T > 0$ we consider the problem of solving, for $\mathbf{u} : \Omega \times (0, T) \rightarrow \mathbb{R}^d$ and $p : \Omega \times (0, T) \rightarrow \mathbb{R}$, the following time-dependent Stokes problem:

$$(2.1) \quad \begin{cases} \partial_t \mathbf{u} - \nu \Delta \mathbf{u} + \nabla p = \mathbf{f} & \text{in } \Omega \times (0, T), \\ \nabla \cdot \mathbf{u} = 0 & \text{in } \Omega \times (0, T), \\ \mathbf{u} = \mathbf{0} & \text{on } \partial\Omega \times (0, T), \\ \mathbf{u}(\cdot, 0) = \mathbf{u}_0 & \text{in } \Omega. \end{cases}$$

Here, $\mathbf{f} : \Omega \times (0, T) \rightarrow \mathbb{R}$ stands for the source term, $\mathbf{u}_0 : \Omega \rightarrow \mathbb{R}^d$ for the initial velocity, and $\nu > 0$ for a given constant viscosity. In order to introduce a variational setting for (2.1) we consider the following standard velocity and pressure spaces:

$$V \stackrel{\text{def}}{=} [H_0^1(\Omega)]^d, \quad H \stackrel{\text{def}}{=} [L^2(\Omega)]^d, \quad Q \stackrel{\text{def}}{=} L_0^2(\Omega),$$

normed with

$$\|\mathbf{v}\|_H \stackrel{\text{def}}{=} (\mathbf{v}, \mathbf{v})^{\frac{1}{2}}, \quad \|\mathbf{v}\|_V \stackrel{\text{def}}{=} \|\nu^{\frac{1}{2}} \nabla \mathbf{v}\|_H, \quad \|q\|_Q \stackrel{\text{def}}{=} \|\nu^{-\frac{1}{2}} q\|_H,$$

where (\cdot, \cdot) denotes the standard L^2 -inner product in Ω .

Problem (2.1) can be formulated in weak form as follows: For all $t > 0$, find $\mathbf{u}(t) \in V$ and $p(t) \in Q$ such that

$$(2.2) \quad \begin{cases} (\partial_t \mathbf{u}, \mathbf{v}) + a(\mathbf{u}, \mathbf{v}) + b(p, \mathbf{v}) = (\mathbf{f}, \mathbf{v}) & \text{a.e. in } (0, T), \\ b(q, \mathbf{u}) = 0 & \text{a.e. in } (0, T), \\ \mathbf{u}(\cdot, 0) = \mathbf{u}_0 & \text{a.e. in } \Omega \end{cases}$$

for all $\mathbf{v} \in V$, $q \in Q$ and with

$$a(\mathbf{u}, \mathbf{v}) \stackrel{\text{def}}{=} (\nu \nabla \mathbf{u}, \nabla \mathbf{v}), \quad b(p, \mathbf{v}) \stackrel{\text{def}}{=} -(p, \nabla \cdot \mathbf{v}).$$

From these definitions, the following classical coercivity and continuity estimates hold:

$$(2.3) \quad a(\mathbf{v}, \mathbf{v}) \geq \|\mathbf{v}\|_V^2, \quad a(\mathbf{u}, \mathbf{v}) \leq \|\mathbf{u}\|_V \|\mathbf{v}\|_V, \quad b(\mathbf{v}, q) \leq \|\mathbf{v}\|_V \|q\|_Q$$

for all $\mathbf{u}, \mathbf{v} \in V$ and $q \in Q$. It is known (see, e.g., [22]) that if $\mathbf{f} \in C^0([0, T]; H)$ and that $\mathbf{u}_0 \in V \cap H_0(\text{div}; \Omega)$, problem (2.2) admits a unique solution (\mathbf{u}, p) in $L^2(0, T; V) \times L^2(0, T; Q)$ with $\partial_t \mathbf{u} \in L^2(0, T; V')$.

Throughout this paper, C stands for a generic positive constant independent of the physical and discretization parameters.

3. Space and time discretization. In this section we discretize problem (2.2) with respect to the space and time variables. Symmetric pressure stabilized finite elements are used for the space discretization (subsection 3.1), and some known \mathcal{A} -stable schemes are used for the time discretization (subsection 3.2).

3.1. Space semidiscretization: Symmetric pressure stabilized formulations. Let $\{\mathcal{T}_h\}_{0 < h \leq 1}$ denote a shape-regular family of triangulations of the domain Ω . For each triangulation \mathcal{T}_h , the subscript $h \in (0, 1]$ refers to the level of refinement of the triangulation, which is defined by

$$h \stackrel{\text{def}}{=} \max_{K \in \mathcal{T}_h} h_K,$$

with h_K the diameter of K . In order to simplify the analysis, we assume that the family of triangulations $\{\mathcal{T}_h\}_{0 < h \leq 1}$ is quasi uniform. For more precise information on the constraint on the mesh, we refer the reader to the analysis of the various finite element methods in the steady case; see subsection 3.1.1.

In this paper, we let X_h^k and M_h^l denote, respectively, the standard spaces of continuous and (possibly) discontinuous piecewise polynomial functions of degree $k \geq 1$ and $l \geq 0$ ($k - 1 \leq l \leq k$),

$$X_h^k \stackrel{\text{def}}{=} \{v_h \in C^0(\bar{\Omega}) : v_h|_K \in \mathbb{P}_k(K) \quad \forall K \in \mathcal{T}_h\},$$

$$M_h^l \stackrel{\text{def}}{=} \{q_h \in L^2(\Omega) : q_h|_K \in \mathbb{P}_l(K) \quad \forall K \in \mathcal{T}_h\}.$$

For the approximated velocities, we will use the space $[V_h^k]^d \stackrel{\text{def}}{=} [X_h^k \cap H_0^1(\Omega)]^d$, and for the pressure, we will use either $Q_h^l \stackrel{\text{def}}{=} M_h^l \cap L_0^2(\Omega)$ or $Q_h^l \stackrel{\text{def}}{=} M_h^l \cap L_0^2(\Omega) \cap C^0(\overline{\Omega})$. In order to stabilize the pressure we introduce a bilinear form $j : Q_h \times Q_h \rightarrow \mathbb{R}$ satisfying the following properties:

- Symmetry:

$$(3.1) \quad j(p_h, q_h) = j(q_h, p_h) \quad \forall p_h, q_h \in Q_h^l;$$

- continuity:

$$(3.2) \quad |j(p_h, q_h)| \leq j(p_h, p_h)^{\frac{1}{2}} j(q_h, q_h)^{\frac{1}{2}} \leq C \|p_h\|_Q \|q_h\|_Q \quad \forall p_h, q_h \in Q_h^l;$$

- weak consistency:

$$(3.3) \quad j(\Pi_h^l q, \Pi_h^l q)^{\frac{1}{2}} \leq C \frac{h^{s_p}}{\nu} \|q\|_{s_p, \Omega} \quad \forall q \in H^s(\Omega),$$

with $s_p \stackrel{\text{def}}{=} \min\{s, \tilde{l}, l + 1\}$, $\tilde{l} \geq 1$, denoting the order of weak consistency of the stabilization operator, and $\Pi_h^l : Q \rightarrow Q_h^l$ a given projection operator such that

$$(3.4) \quad \|q - \Pi_h^l q\|_Q \leq \frac{C}{\nu^{\frac{1}{2}}} h^{l+1} \|q\|_{l+1, \Omega}$$

for all $q \in H^{l+1}(\Omega)$.

Finally, we assume that there exists a projection operator $\mathcal{I}_h^k : V \rightarrow V_h^k$ satisfying the following approximation properties:

$$(3.5) \quad \|\mathbf{v} - \mathcal{I}_h^k \mathbf{v}\|_H + h\nu^{-\frac{1}{2}} \|\mathbf{v} - \mathcal{I}_h^k \mathbf{v}\|_V \leq C_{\mathcal{I}} h^{r_u} \|\mathbf{v}\|_{r_u, \Omega},$$

$$(3.6) \quad |b(q_h, \mathbf{v} - \mathcal{I}_h^k \mathbf{v})| \leq C j(q_h, q_h)^{\frac{1}{2}} \left(\nu^{\frac{1}{2}} \|h^{-1}(\mathbf{v} - \mathcal{I}_h^k \mathbf{v})\|_H + \|\mathbf{v} - \mathcal{I}_h^k \mathbf{v}\|_V \right)$$

for all $\mathbf{v} \in [H^r(\Omega)]^d$, $r_u = \min\{r, k + 1\}$, and $(q_h, \mathbf{v}_h) \in Q_h^l \times [V_h^k]^d$.

Our space semidiscretized scheme reads as follows: For all $t \in (0, T)$, find $(\mathbf{u}_h(t), p_h(t)) \in [V_h^k]^d \times Q_h^l$ such that

$$(3.7) \quad \begin{aligned} (\partial_t \mathbf{u}_h, \mathbf{v}_h) + a(\mathbf{u}_h, \mathbf{v}_h) + b(p_h, \mathbf{v}_h) - b(q_h, \mathbf{u}_h) + j(p_h, q_h) &= (\mathbf{f}, \mathbf{v}_h), \\ \mathbf{u}_h(0) &= \mathbf{u}_h^0, \end{aligned}$$

for all $(\mathbf{v}_h, q_h) \in [V_h^k]^d \times Q_h^l$ and with \mathbf{u}_h^0 a suitable approximation of \mathbf{u}_0 in $[V_h^k]^d$.

The following modified inf-sup condition states the stability of the discrete pressures in (3.7).

LEMMA 3.1. *There exists two constants $C, \beta > 0$, independent of h and ν , such that*

$$(3.8) \quad \sup_{\mathbf{v}_h \in [V_h^k]^d} \frac{|b(q_h, \mathbf{v}_h)|}{\|\mathbf{v}_h\|_V} + C j(q_h, q_h)^{\frac{1}{2}} \geq \beta \|q_h\|_Q$$

for all $q_h \in Q_h^l$.

Proof. Let $q_h \in Q_h^l$; from [25, Corollary 2.4] and (3.5) there exists $\mathbf{v}_q \in H_0^1(\Omega)$ such that $\nabla \cdot \mathbf{v}_q = \nu^{-1} q_h$ and

$$(3.9) \quad \|\mathcal{I}_h^k \mathbf{v}_q\|_V \leq C \|\mathbf{v}_q\|_V \leq C \|q_h\|_Q.$$

On the other hand, using (3.6), we have

$$\begin{aligned} \|q_h\|_Q^2 &= b(q_h, \mathbf{v}_q) \\ &= b(q_h, \mathbf{v}_q - \mathcal{I}_h^k \mathbf{v}_q) + b(q_h, \mathcal{I}_h^k \mathbf{v}_q) \\ &\leq Cj(q_h, q_h)^{\frac{1}{2}} \left(\|\nu^{\frac{1}{2}} h^{-1}(\mathbf{v} - \mathcal{I}_h^k \mathbf{v})\|_H + \|\mathbf{v} - \mathcal{I}_h^k \mathbf{v}\|_V \right) + b(q_h, \mathcal{I}_h^k \mathbf{v}_q) \\ &\leq Cj(q_h, q_h)^{\frac{1}{2}} \|q_h\|_Q + b(q_h, \mathcal{I}_h^k \mathbf{v}_q). \end{aligned}$$

We conclude the proof by dividing this last inequality by $\|\mathcal{I}_h^k \mathbf{v}_q\|_V$ and using (3.9). \square

The above lemma ensures the well-posedness of problem (3.7). This is stated in the following theorem.

THEOREM 3.2. *The discrete problem (3.7) with $\mathbf{u}_h^0 \in V_{h,k}^{\text{div}} \stackrel{\text{def}}{=} \{\mathbf{v}_h \in V_j^k : b(q_h, \mathbf{v}_h) = 0 \text{ for all } q_h \in Q_h \cap \text{Ker } j\}$ has a unique solution $(\mathbf{u}_h, p_h) \in C^1((0, T]; [V_h^k]^d) \times C^0((0, T]; Q_h^k)$.*

To facilitate the analysis we introduce the following (mesh-dependent) seminorm, which is a norm for the velocity and a seminorm for the pressure:

$$(3.10) \quad \|(\mathbf{v}_h, q_h)\|_h^2 \stackrel{\text{def}}{=} \|\mathbf{v}_h\|_V^2 + j(q_h, q_h).$$

Remark 3.3. If the velocity/pressure finite element pair V_h^k/Q_h^k is inf-sup stable, we can take $j(\cdot, \cdot) \stackrel{\text{def}}{=} 0$ in (3.7), as usual. Obviously, this choice is compatible with hypothesis (3.1)–(3.3) so that the results of this paper still apply. In particular, the relation (3.8) becomes the standard inf-sup condition between V_h^k and Q_h^k .

3.1.1. Examples. In this section we will review some of the most well-known pressure projection stabilization methods and discuss how they enter the abstract framework of the previous subsection. For detailed results on analysis for the respective methods, we refer the reader to the references considering the stationary case.

Recently, several different weakly consistent symmetric pressure stabilized finite element methods have been proposed. These methods take their origin from the works of Silvester [32] and Codina and Blasco [17]. Further developments include the work by Becker and Braack [2] on local projection schemes; the extension of the interior penalty method, using penalization of gradient jumps, to the case of pressure stabilization by Burman and Hansbo [14]; and the interpretation of these methods as minimal stabilization procedures by Brezzi and Fortin [9]. Similar approaches have been advocated in Dohrmann and Bochev in [21], and a review of the analysis (with special focus on discontinuous pressure spaces and the Darcy problem) is given in [12].

The main idea underpinning all these methods is that, when using a velocity-pressure space pair $V_h \times Q_h$, the inf-sup stability constraint on the spaces may be relaxed by the addition of an operator penalizing the difference between the discrete pressure variable and its projection onto a subspace $\tilde{Q}_h \subset Q_h$, such that $V_h \times \tilde{Q}_h$ is inf-sup stable. The penalization may either act directly on the pressure, as in [21, 12], or on the gradient of the pressure, as in [2, 18, 14]. Generally speaking, the pressure approximation properties of the numerical scheme will be given by \tilde{Q}_h , expressed in the weak consistency satisfied by the penalty operator. For the Oseen’s problem, some of these methods may be extended to include high Reynolds number effects (see, e.g., [13, 6, 16]). The advantages and disadvantages of symmetric weakly consistent pressure stabilization methods compared to GLS or PSPG approaches is discussed in a recent review paper [7].

The methods of Brezzi and Pitkäranta, Silvester, and Dohrmann and Bochev. The original pressure stabilized finite element method was proposed by Brezzi and Pitkäranta in [10]. Here, the velocity and pressure discrete spaces are chosen as the standard finite element space of piecewise affine continuous functions, $[V_h^1]^d \times Q_h^1$. The operator $j(\cdot, \cdot)$ is given by

$$(3.11) \quad j(p_h, q_h) = \left(\frac{h^2}{\nu} \nabla p_h, \nabla q_h \right).$$

A variant of this method was recently proposed by Dohrmann and Bochev in [21], using an equivalent stabilization operator, namely,

$$(3.12) \quad j(p_h, q_h) = \left(\frac{1}{\nu} (I - \pi_0) p_h, (I - \pi_0) q_h \right),$$

where $\pi_0 : Q \rightarrow Q_h^0$ denotes the (elementwise) projection onto piecewise constants. Property (3.6) is verified after an integration by parts, with \mathcal{I}_h^1 simply the Scott–Zhang interpolant onto $[V_h^1]^d$ (see, e.g., [31, 22]),

$$\begin{aligned} b(q_h, \mathbf{v} - \mathcal{I}_h^1 \mathbf{v}) &= (\nabla q_h, \mathbf{v} - \mathcal{I}_h^1 \mathbf{v}) \\ &\leq j(q_h, q_h)^{\frac{1}{2}} \left(h^{-1} \nu^{\frac{1}{2}} \|\mathbf{v} - \mathcal{I}_h^1 \mathbf{v}\|_H + \|\mathbf{v} - \mathcal{I}_h^1 \mathbf{v}\|_V \right). \end{aligned}$$

One readily verifies that (3.2) and (3.3) hold. Moreover, in both cases (3.11) and (3.12), the weak consistency property holds (with $\tilde{l} = 1$),

$$j(\Pi_h^1 p, \Pi_h^1 p)^{\frac{1}{2}} \leq \frac{C}{\nu^{\frac{1}{2}}} h \|p\|_{1,\Omega},$$

with Π_h^1 being, for instance, the L^2 -projection onto Q_h^1 (we could use instead the Clément [15] or Scott–Zhang interpolants). Indeed, for (3.11) we apply the H^1 -stability of the L^2 -projection (see, e.g., [22, 20, 8, 5]), whereas for (3.12) we add and subtract suitable terms (p and $\pi_0 p$) and use the approximation properties of π_0 and Π_h^1 (see, e.g., [22]). As a result, our analysis for the time discretization is valid.

Another low order scheme, covered by the analysis, is the method which consists of using piecewise affine continuous velocities and elementwise constants pressures, $[V_h^1]^d \times Q_h^0$; see, e.g., [27]. Stability is obtained by the addition of the jump over element faces of the discontinuous pressure, namely,

$$j(p_h, q_h) = \sum_{K \in \mathcal{T}_h} \int_{\partial K \setminus \partial \Omega} \frac{h}{\nu} \llbracket p_h \rrbracket \llbracket q_h \rrbracket.$$

Here, $\llbracket q_h \rrbracket$ denotes the jump of q_h over the interelement boundary, defined by

$$\llbracket q_h \rrbracket(\mathbf{x}) \stackrel{\text{def}}{=} \lim_{\epsilon \rightarrow 0} (q_h(\mathbf{x} + \epsilon \mathbf{n}) - q_h(\mathbf{x} - \epsilon \mathbf{n})) \quad \forall \mathbf{x} \in F,$$

with \mathbf{n} standing for a fixed, but arbitrary, normal to the internal face F . In this case, (3.6) is obtained after an integration by parts in the term $b(q_h, \mathbf{v} - \mathcal{I}_h^1 \mathbf{v})$ and an elementwise trace inequality (see, e.g., [22]),

$$\begin{aligned} b(q_h, \mathbf{v} - \mathcal{I}_h^1 \mathbf{v}) &= - \sum_K \int_{\partial K \setminus \partial \Omega} \llbracket q_h \rrbracket (\mathbf{v} - \mathcal{I}_h^1 \mathbf{v}) \cdot \mathbf{n} \\ &\leq j(q_h, q_h)^{\frac{1}{2}} \left(h^{-1} \nu^{\frac{1}{2}} \|\mathbf{v} - \mathcal{I}_h^1 \mathbf{v}\|_H + \|\mathbf{v} - \mathcal{I}_h^1 \mathbf{v}\|_V \right). \end{aligned}$$

In addition, by taking, for instance, Π_h^0 as the L^2 -projection onto Q_h^0 , using an elementwise trace inequality and the approximation properties of Π_h^0 (see, e.g., [22]), one also easily shows that the weak consistency property holds,

$$j(\Pi_h^0 p, \Pi_h^0 p)^{\frac{1}{2}} = j((I - \Pi_h^0)p, (I - \Pi_h^0)p)^{\frac{1}{2}} \leq \frac{C}{\nu^{\frac{1}{2}}} h \|p\|_{1,\Omega},$$

and hence $\tilde{l} = 1$.

For details on the cases of stabilization of the pressure jumps only in macroelements, or the generalization to higher order finite element spaces of the Taylor–Hood family with discontinuous pressures, we refer the reader to [12].

Orthogonal subscale stabilization. The orthogonal subscale stabilization was proposed by Codina and Blasco in [17]. Equal order ($k = l \geq 1$) continuous approximation spaces are used for the velocities and the pressures.

Here the main idea is to penalize the difference between the pressure gradient and its projection onto the finite element space. This imposes the introduction of an auxiliary variable for the projection since it may not be localized and is given only implicitly. Hence, the stabilization operator is given by

$$j(p_h, q_h) = \left(\frac{h^2}{\nu} (\nabla p_h - \pi_h^k \nabla p_h), \nabla q_h \right),$$

where $\pi_h^k : [L^2(\Omega)]^d \rightarrow [V_h^k]^d$ stands for the L^2 -projection onto $[V_h^k]^d$, which is given as the solution of the (global) problem

$$(\pi_h^k \nabla p_h, \boldsymbol{\xi}_h) = (\nabla p_h, \boldsymbol{\xi}_h) \quad \forall \boldsymbol{\xi}_h \in [V_h^k]^d.$$

One may readily show that (3.2) and (3.3) hold. Disregarding for simplicity the boundary conditions, the projection operator $\mathcal{I}_h^k = \pi_h^k$ of (3.6) is here chosen also as the L^2 -projection onto $[V_h^k]^d$. This can be justified if boundary conditions are imposed weakly, for instance, using Nitsche’s method (see [29, 24]), and V_h^k includes the degrees of freedom on the boundary. Indeed, then we have

$$\begin{aligned} b(q_h, \mathbf{v} - \mathcal{I}_h^k \mathbf{v}) &= (\nabla q_h - \pi_h^k \nabla q_h, \mathbf{v} - \mathcal{I}_h^k \mathbf{v}) \\ (3.13) \quad &\leq j(q_h, q_h)^{\frac{1}{2}} \left(h^{-1} \nu^{\frac{1}{2}} \|\mathbf{v} - \mathcal{I}_h^k \mathbf{v}\|_H + \|\mathbf{v} - \mathcal{I}_h^k \mathbf{v}\|_V \right). \end{aligned}$$

Finally, by taking $\Pi_h^k : Q \rightarrow Q_h^k$ as the L^2 -projection onto Q_h^k , adding and subtracting suitable terms (∇p and $\pi_h^k \nabla p$), and using the approximation properties of π_h^k and Π_h^k (see, e.g., [22]), one readily verifies the weak consistency

$$j(\Pi_h^k p, \Pi_h^k p)^{\frac{1}{2}} = \frac{h}{\nu^{\frac{1}{2}}} \|(I - \pi_h^k) \nabla \Pi_h^k p\|_{0,\Omega} \leq \frac{C}{\nu^{\frac{1}{2}}} h^{s_p} \|p\|_{s_p,\Omega},$$

for all $p \in H^s(\Omega)$ and with $s_p = \min\{k + 1, s\}$. In particular, $\tilde{l} = l = k$. The above analysis is hence valid also in this case (with some modifications of a technical nature due to the weakly imposed boundary conditions).

Local projection stabilization. In the local projection stabilization proposed in [2], stability is obtained by penalizing the projection of the gradient onto piecewise discontinuous functions defined on patches consisting of several elements, obtained by

using hierarchic meshes, or by penalizing the gradient of the difference of the pressure and its projection on polynomials of lower polynomial order. The construction relies on the inf-sup stability of a velocity/pressure pair typically of mini-element character or of the Taylor–Hood family. Similar ideas were advocated in [21]. The stabilization operator is written as

$$j(p_h, q_h) = \sum_{\tilde{K}} \left(\frac{h^2}{\nu} \kappa \nabla p_h, \kappa \nabla q_h \right),$$

where κ is the so-called *fluctuation operator* defined as $\kappa \stackrel{\text{def}}{=} I - \tilde{\pi}_h$, where $\tilde{\pi}_h$ denotes a *local* projection operator onto either a polynomial of order k on a macropatch consisting of three triangles (or four quadrilaterals) or a polynomial of order $k - 1$ on the element. One may show that (3.6), (3.2), and (3.3) hold (for details on the construction of \mathcal{I}_h^k , see [2, 6], and for general conditions on the finite element spaces and stabilization operators, see [28]). In the case when we consider the projection $\tilde{\pi}_h$ onto polynomials of order $k - 1$, the stabilization operator may be written as

$$(3.14) \quad j(p_h, q_h) = \sum_{\tilde{K}} \left(\frac{h^2}{\nu} \nabla(\kappa p_h), \nabla(\kappa q_h) \right),$$

or, equivalently, following [21], as

$$j(p_h, q_h) = \sum_{\tilde{K}} \left(\frac{1}{\nu} \kappa p_h, \kappa q_h \right).$$

In these latter cases, condition (3.6) is obtained by choosing \mathcal{I}_h^k as the Fortin interpolation operator associated with $[V_h^k]^d \times \tilde{Q}_h$, where \tilde{Q}_h is the space of continuous piecewise polynomial functions of order $k - 1$. Clearly, we then have

$$\begin{aligned} b(q_h, \mathbf{v} - \mathcal{I}_h^k \mathbf{v}) &= b(\kappa q_h, \mathbf{v} - \mathcal{I}_h^k \mathbf{v}) \\ &\leq j(q_h, q_h)^{\frac{1}{2}} \left(h^{-1} \nu^{\frac{1}{2}} \|\mathbf{v} - \mathcal{I}_h^k \mathbf{v}\|_H + \|\mathbf{v} - \mathcal{I}_h^k \mathbf{v}\|_V \right), \end{aligned}$$

since $b(\tilde{q}_h, \mathbf{v} - \mathcal{I}_h^k \mathbf{v}) = 0$ for all $\tilde{q}_h \in \tilde{Q}_h$. The form (3.14) is treated in a similar fashion after an integration by parts. On the other hand, by taking $\Pi_h^l : Q \rightarrow Q_h^l$ as the L^2 -projection operator onto Q_h^l and using approximation properties of Π_h^l and $\tilde{\pi}_h$, we have

$$\begin{aligned} j(\Pi_h^l p, \Pi_h^l p)^{\frac{1}{2}} &\leq j((I - \Pi_h^l)p, (I - \Pi_h^l)p)^{\frac{1}{2}} + j(p, p)^{\frac{1}{2}} \\ &\leq \frac{C}{\nu^{\frac{1}{2}}} h^{s_p} \|p\|_{s_p, \Omega} \quad \forall p \in H^s(\Omega), \end{aligned}$$

where $s_p = \min\{\tilde{l}, s, l + 1\}$ and $\tilde{l} - 1$ denotes the polynomial order of the space on which the local projection is taken. Clearly, if we project on polynomials of order $k - 1$, the stabilization operator loses one order in the weak consistency; however, the estimates remain optimal since we expect the velocities to be one order more regular than the pressure.

Continuous interior penalty (CIP) stabilization. The CIP stabilization for the stationary Stokes problem was proposed in [14] and generalized to Oseen’s problem in [13]. It uses equal order continuous approximation spaces for velocities and pressures ($k = l \geq 1$) and relies on the fact that the component of the pressure gradient orthogonal to the finite element space may be controlled by the gradient jumps using an interpolation estimate between discrete spaces. Indeed, it was shown in [13] that the following inequality holds:

$$(3.15) \quad \|h(\nabla p_h - \tilde{i}\nabla p_h)\|_H^2 \leq \sum_{K \in \mathcal{T}_h} \int_{\partial K \setminus \partial \Omega} h_K^3 \llbracket \nabla p_h \cdot \mathbf{n} \rrbracket^2$$

for a certain Clément-type quasi-interpolation operator \tilde{i} . This motivates the use of the pressure stabilization operator

$$j(p_h, q_h) = \sum_{K \in \mathcal{T}_h} \int_{\partial K \setminus \partial \Omega} \frac{h^3}{\nu} \llbracket \nabla p_h \cdot \mathbf{n} \rrbracket \llbracket \nabla q_h \cdot \mathbf{n} \rrbracket.$$

Clearly (3.2) and (3.3) are verified in this case. Moreover, (3.6) may be shown to hold if \mathcal{I}_h^k is chosen to be the L^2 -projection onto $[V_h^k]^d$ and boundary conditions are imposed weakly [13]. To show the inequality we combine (3.13) with (3.15). Finally, by taking Π_h^k as the L^2 -projection onto Q_h^k , since $\llbracket \mathcal{C}_h^k \nabla p \rrbracket = \mathbf{0}$ (with \mathcal{C}_h^k the Clément interpolant onto $[X_h^k]^d$), using an elementwise trace inequality, adding and subtracting ∇p , and using the approximation properties of \mathcal{C}_h^k and Π_h^k , one readily verifies (see [13, Lemma 4.7]) the weak consistency

$$\begin{aligned} j(\Pi_h^k p, \Pi_h^k p)^{\frac{1}{2}} &\leq C \frac{h}{\nu^{\frac{1}{2}}} \|\nabla \Pi_h^k p - \mathcal{C}_h^k \nabla p\|_{0,\Omega} \\ &\leq \frac{C}{\nu^{\frac{1}{2}}} h^{s_p} \|p\|_{s_p,\Omega} \quad \forall p \in H^s(\Omega), \end{aligned}$$

with $s_p = \min\{k + 1, s\}$, so that $\tilde{l} = l = k$.

We refer the reader to [13] for the details on the technical issue related to the weak imposition of the boundary conditions using Nitsche’s method.

3.1.2. The Ritz-projection operator. For the purpose of the stability and convergence analysis below we introduce the Ritz-projection operator

$$S_h^{k,l} : [H^1(\Omega)]^d \times L^2(\Omega) \longrightarrow V_h^k \times Q_h^l.$$

For each $(\mathbf{u}, p) \in [H^1(\Omega)]^d \times L^2(\Omega)$, the projection $S_h^{k,l}(\mathbf{u}, p) \stackrel{\text{def}}{=} (P_h^k(\mathbf{u}, p), R_h^l(\mathbf{u}, p)) \in [V_h^k]^d \times Q_h^l$ is defined as the unique solution of

$$(3.16) \quad \begin{cases} a(P_h^k(\mathbf{u}, p), \mathbf{v}_h) + b(R_h^l(\mathbf{u}, p), \mathbf{v}_h) = a(\mathbf{u}, \mathbf{v}_h) + b(p, \mathbf{v}_h), \\ -b(q_h, P_h^k(\mathbf{u}, p)) + j(R_h^l(\mathbf{u}, p), q_h) = 0 \end{cases}$$

for all $(\mathbf{v}_h, q_h) \in [V_h^k]^d \times Q_h^l$.

Problem (3.16) is well-posed thanks to the inf-sup condition (3.8); in particular, we have the following a priori stability estimate:

$$(3.17) \quad \|(P_h^k(\mathbf{u}, p), R_h^l(\mathbf{u}, p))\|_h^2 \leq C (\|\mathbf{u}\|_V^2 + \|p\|_Q^2),$$

with $C > 0$ a constant independent of h and ν .

Finally, we have the following approximation result.

LEMMA 3.4. *Let $(\mathbf{u}, p) \in C^1([0, T], [H^r(\Omega) \cap H_0^1(\Omega)]^d \cap H_0(\text{div}; \Omega) \times H^s(\Omega))$ with $r \geq 2$ and $s \geq 1$. The following error estimate for the projection $S_h^{k,l}$ holds with $\alpha = 0, 1$:*

$$\begin{aligned} \|(\partial_t^\alpha(\mathbf{u} - P_h^k(\mathbf{u}, p)), \partial_t^\alpha R_h^l(\mathbf{u}, p))\|_h &\leq C \left(\nu^{\frac{1}{2}} h^{r_u-1} \|\partial_t^\alpha \mathbf{u}\|_{r_u, \Omega} + \nu^{-\frac{1}{2}} h^{s_p} \|\partial_t^\alpha p\|_{s_p, \Omega} \right), \\ \|p - R_h^l(\mathbf{u}, p)\|_Q &\leq C \left(\nu^{\frac{1}{2}} h^{r_u-1} \|\mathbf{u}\|_{r_u, \Omega} + \nu^{-\frac{1}{2}} h^{s_p} \|p\|_{s_p, \Omega} \right) \end{aligned}$$

for all $t \in [0, T]$ and with $r_u \stackrel{\text{def}}{=} \min\{r, k+1\}$ and $s_p \stackrel{\text{def}}{=} \min\{s, \tilde{l}, l+1\}$, and $C > 0$ independent of ν and h . Moreover, provided the domain Ω is sufficiently smooth and, if $\tilde{l} \geq 1$, there also holds

$$(3.18) \quad \|\partial_t^\alpha(\mathbf{u} - P_h^k(\mathbf{u}, p))\|_H \leq Ch \|(\partial_t^\alpha(\mathbf{u} - P_h^k(\mathbf{u}, p)), \partial_t^\alpha R_h^l p)\|_h.$$

Proof. For simplicity we here use the notation $\mathbf{u}_h \stackrel{\text{def}}{=} P_h^k(\mathbf{u}, p)$ and $p_h \stackrel{\text{def}}{=} R_h^l(\mathbf{u}, p)$. From (3.10), the V -coercivity of $a(\cdot, \cdot)$ (see (2.3)), and the orthogonality provided by (3.16), we have

$$\begin{aligned} \|(\mathbf{u}_h - \mathcal{I}_h^k \mathbf{u}, p_h - \Pi_h^l p)\|_h^2 &= a(\mathbf{u} - \mathcal{I}_h^k \mathbf{u}, \mathbf{u}_h - \mathcal{I}_h^k \mathbf{u}) + b(p - \Pi_h^l p, \mathbf{u}_h - \mathcal{I}_h^k \mathbf{u}) \\ &\quad + b(p_h - \Pi_h^l p, \mathbf{u} - \mathcal{I}_h^k \mathbf{u}) + j(\Pi_h^l p, p_h - \Pi_h^l p). \end{aligned}$$

Finally, using (2.3) and (3.6), we have that

$$\begin{aligned} \|(\mathbf{u}_h - \mathcal{I}_h^k \mathbf{u}, p_h - \Pi_h^l p)\|_h^2 &\leq (\|\mathbf{u} - \mathcal{I}_h^k \mathbf{u}\|_V + \|p - \Pi_h^l p\|_Q) \|\mathbf{u}_h - \mathcal{I}_h^k \mathbf{u}\|_V \\ &\quad + C \left(\nu^{\frac{1}{2}} \|h^{-1}(\mathbf{u} - \mathcal{I}_h^k \mathbf{u})\|_H + \|\mathbf{u} - \mathcal{I}_h^k \mathbf{u}\|_V + j(\Pi_h^l p, \Pi_h^l p)^{\frac{1}{2}} \right) j(p_h - \Pi_h^l p, p_h - \Pi_h^l p)^{\frac{1}{2}}. \end{aligned}$$

We obtain the estimation for the velocity ($\alpha = 0$) using the approximation properties of \mathcal{I}_h^k and Π_h^l (see (3.5) and (3.4)) and the weak consistency (3.3) of the stabilizing term $j(\cdot, \cdot)$. The convergence for the time derivative ($\alpha = 1$) is obtained in a similar fashion after the time derivation of (3.16).

For the pressure estimate, we use the generalized inf-sup condition (3.8) and the orthogonality provided by (3.16). We then have

$$\begin{aligned} &\beta \|\Pi_h^l p - p_h\|_Q \\ &\leq \sup_{\mathbf{v}_h \in [V_h^k]^d} \frac{b(\Pi_h^l p - p_h, \mathbf{v}_h)}{\|\mathbf{v}_h\|_V} + C j(\Pi_h^l p - p_h, \Pi_h^l p - p_h)^{\frac{1}{2}} \\ &\leq \sup_{\mathbf{v}_h \in [V_h^k]^d} \frac{b(\Pi_h^k p - p, \mathbf{v}_h) - a(\mathbf{u} - \mathbf{u}_h, \mathbf{v}_h)}{\|\mathbf{v}_h\|_V} + C j(\Pi_h^l p - p_h, \Pi_h^l p - p_h)^{\frac{1}{2}}. \end{aligned}$$

We conclude by using the continuity of $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$, approximability, the weak consistency of $j(\cdot, \cdot)$, and the previous error estimate. For a proof of the optimality in the H -norm, see, e.g., [13, Theorem 4.14]. \square

3.2. Fully discrete formulation: Time discretization. In this subsection we discretize (3.7) with respect to the time variable. To this end, we will use some known \mathcal{A} -stable time discretization schemes for ODEs.

Let $N \in \mathbb{N}^*$ be given. We consider a uniform partition $\{[t_n, t_{n+1}]\}_{0 \leq n \leq N-1}$, with $t_n \stackrel{\text{def}}{=} n\delta t$, of the time interval of interest $[0, T]$ with time-step size $\delta t \stackrel{\text{def}}{=} T/N$. The discrete pair (\mathbf{u}_h^n, p_h^n) stands for an approximation of $(\mathbf{u}(t_n), p(t_n))$ in $[V_h^k]^d \times Q_h^l$.

First order backward difference formula (BDF1). By introducing the first order backward difference quotient

$$\bar{D}u_h^{n+1} \stackrel{\text{def}}{=} \frac{u_h^{n+1} - u_h^n}{\delta t},$$

our first fully discrete scheme reads as follows: For $0 \leq n \leq N-1$, find $(\mathbf{u}_h^{n+1}, p_h^{n+1}) \in [V_h^k]^d \times Q_h^l$ such that

$$(3.19) \quad (\bar{D}\mathbf{u}_h^{n+1}, \mathbf{v}_h) + a(\mathbf{u}_h^{n+1}, \mathbf{v}_h) + b(p_h^{n+1}, \mathbf{v}_h) - b(q_h, \mathbf{u}_h^{n+1}) \\ + j(p_h^{n+1}, q_h) = (\mathbf{f}(t_{n+1}), \mathbf{v}_h)$$

for all $(\mathbf{v}_h, q_h) \in V_h^k \times Q_h^l$ and with \mathbf{u}_h^0 a suitable approximation of \mathbf{u}_0 in $[V_h^k]^d$.

Crank–Nicholson scheme. Let us consider now the scheme given by the following: For $0 \leq n \leq N-1$, find $(\mathbf{u}_h^{n+1}, p_h^{n+\frac{1}{2}}) \in [V_h^k]^d \times Q_h^l$ such that

$$(3.20) \quad (\bar{D}\mathbf{u}_h^{n+1}, \mathbf{v}_h) + a(\mathbf{u}_h^{n+\frac{1}{2}}, \mathbf{v}_h) + b(p_h^{n+\frac{1}{2}}, \mathbf{v}_h) - b(q_h, \mathbf{u}_h^{n+\frac{1}{2}}) \\ + j(p_h^{n+\frac{1}{2}}, q_h) = (\mathbf{f}^{n+\frac{1}{2}}, \mathbf{v}_h)$$

for all $(\mathbf{v}_h, q_h) \in [V_h^k]^d \times Q_h^l$, where $\mathbf{u}_h^{n+\frac{1}{2}} \stackrel{\text{def}}{=} \frac{1}{2}(\mathbf{u}_h^{n+1} + \mathbf{u}_h^n)$ and \mathbf{u}_h^0 is a suitable approximation of \mathbf{u}_0 in $[V_h^k]^d$.

Remark 3.5. Note that (3.20) uniquely determines \mathbf{u}_h^{n+1} , since \mathbf{u}_h^0 is given. For the pressure, however, neither p_h^{n+1} nor p_h^n is used in (3.20). Therefore, by working with $p_h^{n+\frac{1}{2}}$ as the pressure variable, we do not need to provide an initial condition for the pressure. On the other hand, we do not have an approximation of p_h^{n+1} unless one is constructed by extrapolation.

Second order backward difference (BDF2). Finally, by considering the second order backward difference quotient

$$\tilde{D}\mathbf{u}^{n+1} \stackrel{\text{def}}{=} \frac{1}{2\delta t}(3\mathbf{u}_h^{n+1} - 4\mathbf{u}_h^n + \mathbf{u}_h^{n-1}),$$

we obtain the following BDF2 scheme: For $1 \leq n \leq N-1$, find $(\mathbf{u}_h^{n+1}, p_h^{n+1}) \in [V_h^k]^d \times Q_h^l$ such that

$$(3.21) \quad (\tilde{D}\mathbf{u}_h^{n+1}, \mathbf{v}_h) + a(\mathbf{u}_h^{n+1}, \mathbf{v}_h) + b(p_h^{n+1}, \mathbf{v}_h) - b(q_h, \mathbf{u}_h^{n+1}) \\ + j(p_h^{n+1}, q_h) = (\mathbf{f}^{n+1}, \mathbf{v}_h)$$

for all $(\mathbf{v}_h, q_h) \in [V_h^k]^d \times Q_h^l$ and $(\mathbf{u}_h^1, p_h^1) \in [V_h^k]^d \times Q_h^l$ given by the first step of backward Euler scheme (3.19).

4. Stability. In this section we analyze the stability properties of the fully discrete schemes introduced in subsection 3.2. For the sake of simplicity, full details will be given only for the backward scheme (3.19). Nevertheless, in subsection 4.2, we will discuss how the results extend to the second order time-stepping schemes Crank–Nicholson and BDF2.

4.1. First order \mathcal{A} -stable scheme. The next result provides the unconditional stability of the velocity. It also provides a uniform estimate for the pressure, in terms of the discrete velocity time derivative. Theorem 4.2 points out the role of the initial velocity approximation on the stability of the velocity time derivative approximations. Finally, Corollary 4.3 states the (conditional or unconditional) stability of the pressure, depending on the choice of the initial velocity approximation.

THEOREM 4.1. *Let \mathbf{u}_h^0 be a given H -stable approximation of \mathbf{u}_0 in $[V_h^k]^d$, and let $\{(\mathbf{u}_h^n, p_h^n)\}_{n=1}^N$ be the solution of the fully discrete problem (3.19). The following estimate holds for $1 \leq n \leq N$:*

$$(4.1) \quad \begin{aligned} & \|\mathbf{u}_h^n\|_H^2 + \sum_{m=0}^{n-1} \delta t \|(\mathbf{u}_h^{m+1}, p_h^{m+1})\|_h^2 \leq C \|\mathbf{u}_0\|_H^2 + \frac{C_P^2}{\nu} \sum_{m=0}^{n-1} \delta t \|\mathbf{f}(t_{m+1})\|_H^2, \\ & \sum_{m=0}^{n-1} \delta t \|p_h^{m+1}\|_Q^2 \\ & \leq \frac{C}{\beta^2} \sum_{m=0}^{n-1} \delta t \left(\|(\mathbf{u}_h^{m+1}, p_h^{m+1})\|_h^2 + \nu^{-1} \|\bar{D}\mathbf{u}_h^{m+1}\|_H^2 + \nu^{-1} \|\mathbf{f}(t_{m+1})\|_H^2 \right), \end{aligned}$$

with $C_P > 0$ the Poincaré constant.

Proof. Taking $\mathbf{v}_h = \mathbf{u}_h^{n+1}$ and $q_h = p_h^{n+1}$ in (3.19), using the coercivity of the bilinear form, the Cauchy–Schwarz inequality, and the Poincaré inequality, we have

$$(4.2) \quad (\bar{D}\mathbf{u}_h^{n+1}, \mathbf{u}_h^{n+1}) + \frac{1}{2} \|(\mathbf{u}_h^{n+1}, p_h^{n+1})\|_h^2 \leq \frac{C_P^2}{2\nu} \|\mathbf{f}(t_{n+1})\|_H^2.$$

Now, recalling that

$$(4.3) \quad (\bar{D}\mathbf{u}_h^{n+1}, \mathbf{u}_h^{n+1}) = \frac{1}{2} \bar{D} \|\mathbf{u}_h^{n+1}\|_H^2 + \frac{1}{2\delta t} \|\mathbf{u}_h^{n+1} - \mathbf{u}_h^n\|_H^2,$$

we have

$$\bar{D} \|\mathbf{u}_h^{n+1}\|_H^2 + \|(\mathbf{u}_h^{n+1}, p_h^{n+1})\|_h^2 \leq \frac{C_P^2}{\nu} \|\mathbf{f}(t_{n+1})\|_H^2,$$

leading to, after summation over $0 \leq m \leq n-1$,

$$\|\mathbf{u}_h^n\|_H^2 + \sum_{m=0}^{n-1} \delta t \|(\mathbf{u}_h^{m+1}, p_h^{m+1})\|_h^2 \leq \|\mathbf{u}_h^0\|_H^2 + \frac{C_P^2}{\nu} \sum_{m=0}^{n-1} \delta t \|\mathbf{f}(t_{m+1})\|_H^2.$$

For the pressure estimate, from (3.8), (3.19) (with $q_h = 0$) and the Poincaré inequality, we have

$$\beta \|p_h^{n+1}\|_Q \leq C \left(\|(\mathbf{u}_h^{n+1}, p_h^{n+1})\|_h + \nu^{-\frac{1}{2}} \|\bar{D}\mathbf{u}_h^{n+1}\|_H + \nu^{-\frac{1}{2}} \|\mathbf{f}(t_{n+1})\|_H \right),$$

which completes the proof. \square

The next theorem states some a priori estimates of the approximations of the velocity time derivative.

THEOREM 4.2. *Let $\{(\mathbf{u}_h^n, p_h^n)\}_{n=1}^N$ be the solution of the fully discrete problem (3.19).*

- *If $\mathbf{u}_0 \in [H^1(\Omega)]^d$ and $\mathbf{u}_h^0 = P_h^k(\mathbf{u}_0, 0)$, the following estimate holds for $1 \leq n \leq N$:*

$$(4.4) \quad \sum_{m=0}^{n-1} \delta t \|\bar{D}\mathbf{u}_h^{m+1}\|_H^2 + \|(\mathbf{u}_h^n, p_h^n)\|_h^2 \leq C \left(\|\mathbf{u}_0\|_V^2 + \sum_{m=0}^{n-1} \delta t \|\mathbf{f}(t_{m+1})\|_H^2 \right).$$

- *If $\mathbf{u}_0 \in [H^r(\Omega) \cap H_0^1(\Omega)]^d \cap H_0(\text{div}; \Omega)$, $r \geq 2$, and $\mathbf{u}_h^0 = \mathcal{I}_h^k \mathbf{u}_0$, the following estimate holds for $1 \leq n \leq N$:*

$$(4.5) \quad \sum_{m=0}^{n-1} \delta t \|\bar{D}\mathbf{u}_h^{m+1}\|_H^2 + \|(\mathbf{u}_h^n, p_h^n)\|_h^2 \leq C \left(\|\mathbf{u}_0\|_V^2 + \nu h^{2(r_u-1)} \|p_h^1\|_Q^2 + \|\mathbf{u}_0\|_{r_u, \Omega}^2 + \sum_{m=0}^{n-1} \delta t \|\mathbf{f}(t_{m+1})\|_H^2 \right),$$

with $r_u \stackrel{\text{def}}{=} \min\{k+1, r\}$.

Proof. For $0 \leq n \leq N-1$, by taking $\mathbf{v}_h = \bar{D}\mathbf{u}_h^{n+1}$ and $q_h = 0$ in (3.19) and using the Cauchy–Schwarz inequality, we have

$$(4.6) \quad \frac{1}{2} \|\bar{D}\mathbf{u}_h^{n+1}\|_H^2 + a(\mathbf{u}_h^{n+1}, \bar{D}\mathbf{u}_h^{n+1}) + b(p_h^{n+1}, \bar{D}\mathbf{u}_h^{n+1}) = \frac{1}{2} \|\mathbf{f}(t_{n+1})\|_H^2.$$

On the other hand, for $1 \leq n \leq N-1$, testing (3.19) at the time levels n and $n+1$ with $\mathbf{v}_h = \mathbf{0}$ and $q_h = p_h^{n+1}$, we have

$$(4.7) \quad \begin{aligned} b(p_h^{n+1}, \mathbf{u}_h^{n+1}) &= j(p_h^{n+1}, p_h^{n+1}), \\ b(p_h^{n+1}, \mathbf{u}_h^n) &= j(p_h^n, p_h^{n+1}). \end{aligned}$$

Therefore, by subtracting these equalities and using the bilinearity of $j(\cdot, \cdot)$, we obtain

$$(4.8) \quad b(p_h^{n+1}, \bar{D}\mathbf{u}_h^{n+1}) = j(\bar{D}p_h^{n+1}, p_h^{n+1})$$

for $1 \leq n \leq N-1$. It then follows from (4.6) that

$$(4.9) \quad \frac{1}{2} \|\bar{D}\mathbf{u}_h^{n+1}\|_H^2 + a(\mathbf{u}_h^{n+1}, \bar{D}\mathbf{u}_h^{n+1}) + j(p_h^{n+1}, \bar{D}p_h^{n+1}) \leq \frac{1}{2} \|\mathbf{f}(t_{n+1})\|_H^2.$$

On the other hand, using the symmetry and bilinearity of $a(\cdot, \cdot)$ and $j(\cdot, \cdot)$, we have

$$\begin{aligned} a(\mathbf{u}_h^{n+1}, \bar{D}\mathbf{u}_h^{n+1}) &= \frac{1}{2} \bar{D}a(\mathbf{u}_h^{n+1}, \mathbf{u}_h^{n+1}) + \frac{\delta t}{2} a(\bar{D}\mathbf{u}_h^{n+1}, \bar{D}\mathbf{u}_h^{n+1}), \\ j(p_h^{n+1}, \bar{D}p_h^{n+1}) &= \frac{1}{2} \bar{D}j(p_h^{n+1}, p_h^{n+1}) + \frac{\delta t}{2} j(\bar{D}p_h^{n+1}, \bar{D}p_h^{n+1}). \end{aligned}$$

Hence,

$$\|\bar{D}\mathbf{u}_h^{n+1}\|_H^2 + \bar{D}(a(\mathbf{u}_h^{n+1}, \mathbf{u}_h^{n+1}) + j(p_h^{n+1}, p_h^{n+1})) \leq \|\mathbf{f}(t_{n+1})\|_H^2$$

for $1 \leq n \leq N - 1$. After multiplication by δt and summation over $1 \leq m \leq n - 1$, it follows that

$$(4.10) \quad \sum_{m=1}^{n-1} \delta t \|\bar{D}\mathbf{u}_h^{m+1}\|_H^2 + \|(\mathbf{u}_h^n, p_h^n)\|_h^2 \leq \|(\mathbf{u}_h^1, p_h^1)\|_h^2 + \sum_{m=1}^{n-1} \delta t \|\mathbf{f}(t_{m+1})\|_H^2.$$

In order to highlight the impact of the initial velocity approximation on the stability of the time derivative, we consider now the first time level ($n = 0$) of (3.19). By testing with $\mathbf{v}_h = \bar{D}\mathbf{u}_h^1$, $q_h = 0$, after multiplication by $2\delta t$ and using the symmetry and bilinearity of $a(\cdot, \cdot)$, we get

$$(4.11) \quad \delta t \|\bar{D}\mathbf{u}_h^1\|_H^2 + a(\mathbf{u}_h^1, \mathbf{u}_h^1) - a(\mathbf{u}_h^0, \mathbf{u}_h^0) + 2\delta t b(p_h^1, \bar{D}\mathbf{u}_h^1) \leq \delta t \|\mathbf{f}(t_1)\|_H^2.$$

If the initial velocity approximation is given in terms of the Ritz-projection, $\mathbf{u}_h^0 = P_h^k(\mathbf{u}_0, 0)$ with $\mathbf{u}_0 \in [H^1(\Omega)]^d$, by setting $p_h^0 \stackrel{\text{def}}{=} R_h^l(\mathbf{u}_0, 0)$ it follows that (4.7) also holds for $n = 0$. Therefore,

$$(4.12) \quad b(p_h^1, \bar{D}\mathbf{u}_h^1) = j(\bar{D}p_h^1, p_h^1).$$

Thus, from the symmetry and bilinearity of $j(\cdot, \cdot)$ and (4.11), we have

$$(4.13) \quad \delta t \|\bar{D}\mathbf{u}_h^1\|_H^2 + \|(\mathbf{u}_h^1, p_h^1)\|_h^2 \leq \|(\mathbf{u}_h^0, p_h^0)\|_h^2 + \delta t \|\mathbf{f}(t_1)\|_H^2.$$

Estimate (4.4) is obtained by adding this last inequality to (4.10) and using the stability of the Ritz-projection (3.17), $\|(\mathbf{u}_h^0, p_h^0)\|_h^2 \leq C\|\mathbf{u}_0\|_V^2$.

If the initial velocity approximation is given in terms of a general interpolant, $\mathbf{u}_h^0 = \mathcal{I}_h^k \mathbf{u}_0$ with $\mathbf{u}_0 \in [H^r(\Omega) \cap H_0^1(\Omega)]^d \cap H_0(\text{div}; \Omega)$, equality (4.12) does not hold in general. Instead, we can use an approximation argument to obtain

$$(4.14) \quad \begin{aligned} b(p_h^1, \bar{D}\mathbf{u}_h^1) &= \frac{1}{\delta t} (j(p_h^1, p_h^1) - (p_h^1, \nabla \cdot (\mathcal{I}_h^k \mathbf{u}_0 - \mathbf{u}_0))) \\ &\geq \frac{1}{\delta t} j(p_h^1, p_h^1) - \frac{C_{\mathcal{I}}}{\delta t} (\nu h^{2(r_u-1)} \|p_h^1\|_Q^2 + \|\mathbf{u}_0\|_{r_u, \Omega}^2), \end{aligned}$$

with $r_u \stackrel{\text{def}}{=} \min\{k + 1, r\}$. As a result, from (4.11) it follows that

$$\delta t \|\bar{D}\mathbf{u}_h^1\|_H^2 + \|(\mathbf{u}_h^1, p_h^1)\|_h^2 \leq a(\mathbf{u}_h^0, \mathbf{u}_h^0) + C_{\mathcal{I}} (\nu h^{2(r_u-1)} \|p_h^1\|_Q^2 + \|\mathbf{u}_0\|_{r_u, \Omega}^2) + \delta t \|\mathbf{f}(t_1)\|_H^2.$$

We conclude the proof by adding this equality to (4.10) and using the stability of the Ritz-projection. \square

The next corollary solves the problem of the stability of the pressures by combining the results of Theorems 4.1 and 4.2.

COROLLARY 4.3. *Let $\{(\mathbf{u}_h^n, p_h^n)\}_{n=1}^N$ be the solution of the fully discrete problem (3.19). Then*

- if $\mathbf{u}_0 \in [H^1(\Omega)]^d$ and $\mathbf{u}_h^0 = P_h^k(\mathbf{u}_0, 0)$, the following estimate holds for $1 \leq n \leq N$:

$$(4.15) \quad \begin{aligned} \sum_{m=0}^{n-1} \delta t \|p_h^{m+1}\|_Q^2 &\leq \frac{C}{\beta^2 \nu} \|\mathbf{u}_0\|_V^2 \\ &+ \frac{C}{\beta^2} \sum_{m=0}^{n-1} \delta t \left(\|(\mathbf{u}_h^{m+1}, p_h^{m+1})\|_h^2 + \nu^{-1} \|\mathbf{f}(t_{m+1})\|_H^2 \right). \end{aligned}$$

- if $\mathbf{u}_0 \in [H^r(\Omega) \cap H_0^1(\Omega)]^d \cap H_0(\text{div}; \Omega)$, $r \geq 2$, $\mathbf{u}_h^0 = \mathcal{I}_h^k \mathbf{u}_0$, and

$$(4.16) \quad \frac{2C_{\mathcal{I}}}{\beta^2} h^{2(r_{\mathbf{u}}-1)} \leq \delta t,$$

the following estimate holds for $1 \leq n \leq N$:

$$(4.17) \quad \sum_{m=0}^{n-1} \delta t \|p_h^{m+1}\|_Q^2 \leq \frac{C}{\beta^2 \nu} (\|\mathbf{u}_0\|_V^2 + \|\mathbf{u}_0\|_{r_{\mathbf{u}}, \Omega}^2) \\ + \frac{C}{\beta^2} \sum_{m=0}^{n-1} \delta t \left(\|(\mathbf{u}_h^{m+1}, p_h^{m+1})\|_h^2 + \nu^{-1} \|\mathbf{f}(t_{m+1})\|_H^2 \right),$$

with $r_{\mathbf{u}} \stackrel{\text{def}}{=} \min\{k+1, r\}$.

Proof. Estimate (4.17) is a direct consequence of Theorem 4.1 and estimate (4.4). On the other hand, from Theorem 4.1 and estimate (4.5), we have

$$\left(\beta^2 \delta t - C_{\mathcal{I}} h^{2(r_{\mathbf{u}}-1)} \right) \|p_h^1\|_Q^2 + \beta^2 \sum_{m=1}^{n-1} \delta t \|p_h^{m+1}\|_Q^2 \\ \leq \frac{C}{\nu} (\|\mathbf{u}_0\|_V^2 + \|\mathbf{u}_0\|_{r_{\mathbf{u}}, \Omega}^2) + C \sum_{m=0}^{n-1} \delta t \left(\|(\mathbf{u}_h^{m+1}, p_h^{m+1})\|_h^2 + \nu^{-1} \|\mathbf{f}(t_{m+1})\|_H^2 \right),$$

which combined with the stability condition (4.16) leads to (4.17). \square

A few observations are now in order. Corollary 4.3 states the unconditional stability of the pressure provided the initial velocity approximation \mathbf{u}_h^0 is given in terms of the Ritz-projection operator (3.16). In the general case, i.e., whenever \mathbf{u}_h^0 does not satisfy a discrete divergence-free condition (as \mathbf{u}_h^1 does), only conditional stability can be guaranteed. As a matter of fact, from the stability condition (4.16), pressure instabilities are expected for very small time steps. This issue will be illustrated by numerical experiments in section 6.

Finally, let us mention that residual-based stabilization methods, such as PSPG and GLS, combined with finite difference time discretization schemes, are known to give rise to pressure instabilities in the small time-step limit; see [3, 19]. Indeed, it has been shown in [3] that the finite difference/pressure coupling of the stabilization perturbs the coercivity of the discrete pressure operator unless a condition of the type

$$(4.18) \quad Ch^2 \leq \delta t$$

is satisfied. It is worth emphasizing that, although the stability conditions (4.18) and (4.16) are somehow similar, their natures are different. Actually, the instabilities anticipated by Corollary 4.3 are related to the discrete divergence-free character of the initial velocity approximation, but not to the structure of the pressure stabilization $j(\cdot, \cdot)$.

4.2. Second order \mathcal{A} -stable schemes. In this subsection we discuss how the results of Theorems 4.1 and 4.2 and Corollary 4.3 extend to the second order time-stepping schemes Crank–Nicholson and BDF2.

Crank–Nicholson. The following theorem summarizes the resulting stability estimates.

THEOREM 4.4. *Let \mathbf{u}_h^0 be a given H -stable approximation of \mathbf{u}_0 in $[V_h^k]^d$, and let $\{(\mathbf{u}_h^n, p_h^n)\}_{n=1}^N$ be the solution of the discrete scheme (3.20). Then the following estimate holds for $1 \leq n \leq N$:*

$$\|\mathbf{u}_h^n\|_H^2 + \sum_{m=0}^{n-1} \delta t \|(\mathbf{u}_h^{m+\frac{1}{2}}, p_h^{m+\frac{1}{2}})\|_h^2 \leq C \|\mathbf{u}_0\|_H^2 + \frac{C_P^2}{\nu} \sum_{m=0}^{n-1} \delta t \|\mathbf{f}(t_{m+\frac{1}{2}})\|_H^2.$$

Moreover, if $\mathbf{u}_0 \in [H^1(\Omega)]^d$ and $\mathbf{u}_h^0 = P_h^k(\mathbf{u}_0, 0)$, the following estimate holds for $1 \leq n \leq N$:

$$\sum_{m=0}^{n-1} \delta t \|p_h^{m+\frac{1}{2}}\|_Q^2 \leq \frac{C}{\beta^2 \nu} \|\mathbf{u}_0\|_V^2 + \frac{C}{\beta^2} \sum_{m=0}^{n-1} \delta t \left(\|(\mathbf{u}_h^{m+\frac{1}{2}}, p_h^{m+\frac{1}{2}})\|_h^2 + \nu^{-1} \|\mathbf{f}(t_{m+\frac{1}{2}})\|_H^2 \right).$$

On the other hand, if $\mathbf{u}_0 \in [H^r(\Omega) \cap H_0^1(\Omega)]^d \cap H_0(\operatorname{div}; \Omega)$, $r \geq 2$, $\mathbf{u}_h^0 = \mathcal{I}_h^k \mathbf{u}_0$, and the stability condition (4.16) is satisfied, the following estimate holds for $1 \leq n \leq N$:

$$\begin{aligned} \sum_{m=0}^{n-1} \delta t \|p_h^{m+\frac{1}{2}}\|_Q^2 &\leq \frac{C}{\beta^2 \nu} (\|\mathbf{u}_0\|_V^2 + \|\mathbf{u}_0\|_{r_u, \Omega}^2) \\ &\quad + \frac{C}{\beta^2} \sum_{m=0}^{n-1} \delta t \left(\|(\mathbf{u}_h^{m+\frac{1}{2}}, p_h^{m+\frac{1}{2}})\|_h^2 + \nu^{-1} \|\mathbf{f}(t_{m+\frac{1}{2}})\|_H^2 \right). \end{aligned}$$

Proof. The first estimate, corresponding to Theorem 4.1, holds by taking $\mathbf{v}_h = \mathbf{u}_h^{n+\frac{1}{2}}$ and $q_h = p_h^{n+\frac{1}{2}}$ in (3.20).

The pressure estimate requires an a priori bound of the discrete velocity time derivative. As in Theorem 4.2, such an estimate can be obtained by taking $\mathbf{v}_h = \bar{D}\mathbf{u}_h^{n+1}$ and $q_h = 0$ in (3.20) for $0 \leq n \leq N-1$. The main difference, with respect to the proof of Theorem 4.2, arises in the treatment of the coupling term $b(p_h^{n+\frac{1}{2}}, \bar{D}\mathbf{u}_h^{n+1})$. Indeed, in the Crank–Nicholson scheme incompressibility is enforced on $\mathbf{u}_h^{n+\frac{1}{2}}$ instead of \mathbf{u}_h^{n+1} . We first note that, since

$$\mathbf{u}_h^{n+1} - \mathbf{u}_h^n = 2 \left(\mathbf{u}_h^{n+\frac{1}{2}} - \mathbf{u}_h^n \right), \quad \mathbf{u}_h^n = 2\mathbf{u}_h^{n-1+\frac{1}{2}} - \mathbf{u}_h^{n-1},$$

we have

$$\mathbf{u}_h^{n+1} - \mathbf{u}_h^n = 2\mathbf{u}_h^{n+\frac{1}{2}} + 4 \sum_{l=1}^n (-1)^l \mathbf{u}_h^{n-l+\frac{1}{2}} - (-1)^n 2\mathbf{u}_h^0$$

for $0 \leq n \leq N-1$. Therefore, from (3.20) and using the bilinearity of $j(\cdot, \cdot)$, we get

$$\begin{aligned} (4.19) \quad b(p_h^{n+\frac{1}{2}}, \mathbf{u}_h^{n+1} - \mathbf{u}_h^n) &= j \left(2p_h^{n+\frac{1}{2}} + 4 \sum_{l=1}^n (-1)^l p_h^{n-l+\frac{1}{2}}, p_h^{n+\frac{1}{2}} \right) \\ &\quad - 2(-1)^n b(p_h^{n+\frac{1}{2}}, \mathbf{u}_h^0). \end{aligned}$$

On the other hand, for $0 \leq n \leq N - 1$, we introduce the following change of variables (or extrapolation):

$$\frac{1}{2} (p_h^{n+1} + p_h^n) \stackrel{\text{def}}{=} p_h^{n+\frac{1}{2}},$$

with $p_h^0 \in Q_h^l$ to be specified later on. By inserting this expression into (4.19), we obtain

$$(4.20) \quad \begin{aligned} b(p_h^{n+\frac{1}{2}}, \mathbf{u}_h^{n+1} - \mathbf{u}_h^n) &= j(p_h^{n+1} - p_h^n, p_h^{n+\frac{1}{2}}) \\ &+ 2(-1)^n \left[j(p_h^0, p_h^{n+\frac{1}{2}}) - b(p_h^{n+\frac{1}{2}}, \mathbf{u}_h^0) \right]. \end{aligned}$$

If $\mathbf{u}_h^0 = P_h^k(\mathbf{u}_0, 0)$ and we choose $p_h^0 \stackrel{\text{def}}{=} R_h^l(\mathbf{u}_0, 0)$, from (3.16)₂ it follows that the last term in (4.20) cancels. Thus, we have

$$\begin{aligned} b(p_h^{n+\frac{1}{2}}, \bar{D}\mathbf{u}_h^{n+1}) &= j(\bar{D}p_h^{n+1}, p_h^{n+\frac{1}{2}}) \\ &= \frac{1}{2} \bar{D}j(p_h^{n+1}, p_h^{n+1}), \end{aligned}$$

which corresponds to the Crank–Nicholson counterpart of (4.8).

Finally, when the initial velocity approximation is given in terms of a general interpolant, $\mathbf{u}_h^0 = \mathcal{I}_h^k \mathbf{u}_0$ with $\mathbf{u}_0 \in [H^r(\Omega) \cap H_0^1(\Omega)]^d \cap H_0(\text{div}; \Omega)$, we take $p_h^0 \stackrel{\text{def}}{=} 0$. Therefore, from (4.20) and using an approximation argument (as in (4.14)), we get

$$(4.21) \quad \begin{aligned} b(p_h^{n+\frac{1}{2}}, \bar{D}\mathbf{u}_h^{n+1}) &= \frac{1}{2} \bar{D}j(p_h^{n+1}, p_h^{n+1}) - \frac{2}{\delta t} (-1)^n b(p_h^{n+\frac{1}{2}}, \mathbf{u}_h^0) \\ &\geq \frac{1}{2} \bar{D}j(p_h^{n+1}, p_h^{n+1}) \\ &\quad - \frac{2C_{\mathcal{I}}}{\delta t} \left(\nu h^{2(r_u-1)} \|p_h^{n+\frac{1}{2}}\|_Q^2 + \|\mathbf{u}_0\|_{r_u, \Omega}^2 \right), \end{aligned}$$

which leads to the stability condition (4.16). The rest of the proof follows with minor modifications. \square

Remark 4.5. By comparing the proofs of Corollary 4.3 and the previous theorem, we can notice that, if the initial velocity approximation is not discretely divergence free, the stability condition (4.16) has to be satisfied at each time level when using the Crank–Nicholson scheme (due to (4.21)), whereas for the backward Euler scheme that condition is needed only at the first time step (thanks to (4.8) and (4.14)).

BDF2. The following theorem summarizes the resulting stability estimates.

THEOREM 4.6. *Let \mathbf{u}_h^0 be a given H -stable approximation of \mathbf{u}_0 in $[V_h^k]^d$, let (u_h^1, p_h^1) be the corresponding first time step of the backward Euler scheme (3.19), and let $\{(\mathbf{u}_h^n, p_h^n)\}_{n=2}^N$ be the solution of the discrete scheme (3.21). Then, the following estimate holds for $2 \leq n \leq N$:*

$$\|\mathbf{u}_h^n\|_H^2 + 2 \sum_{m=1}^{n-1} \delta t \|(\mathbf{u}_h^{m+1}, p_h^{m+1})\|_h^2 \leq C (\|\mathbf{u}_0\|_H^2 + \|\mathbf{u}_h^1\|_H^2) + \frac{2C_{\text{P}}^2}{\nu} \sum_{m=1}^{n-1} \delta t \|\mathbf{f}(t_{m+1})\|_H^2.$$

Moreover, if $\mathbf{u}_0 \in [H^1(\Omega)]^d$ and $\mathbf{u}_h^0 = P_h^k(\mathbf{u}_0, 0)$, the following estimate holds for $2 \leq n \leq N$:

$$\begin{aligned} \sum_{m=1}^{n-1} \delta t \|p_h^{m+1}\|_Q^2 &\leq \frac{C}{\beta^2 \nu} \left(\|\mathbf{u}_0\|_V^2 + \|(\mathbf{u}_h^1, p_h^1)\|_h^2 \right) \\ &\quad + \frac{C}{\beta^2} \sum_{m=1}^{n-1} \delta t \left(\|(\mathbf{u}_h^{m+1}, p_h^{m+1})\|_h^2 + \nu^{-1} \|\mathbf{f}(t_{m+1})\|_H^2 \right). \end{aligned}$$

On the other hand, if $\mathbf{u}_0 \in [H^r(\Omega) \cap H_0^1(\Omega)]^d \cap H_0(\operatorname{div}; \Omega)$, $r \geq 2$, $\mathbf{u}_h^0 = \mathcal{I}_h^k \mathbf{u}_0$, and the stability condition (4.16) is satisfied, the following estimate holds for $2 \leq n \leq N$:

$$\begin{aligned} \sum_{m=1}^{n-1} \delta t \|p_h^{m+1}\|_Q^2 &\leq \frac{C}{\beta^2 \nu} \left(\|\mathbf{u}_0\|_V^2 + \|\mathbf{u}_0\|_{r, \Omega}^2 + \|(\mathbf{u}_h^1, p_h^1)\|_h^2 \right) \\ &\quad + \frac{C}{\beta^2} \sum_{m=1}^{n-1} \delta t \left(\|(\mathbf{u}_h^{m+1}, p_h^{m+1})\|_h^2 + \nu^{-1} \|\mathbf{f}(t_{m+1})\|_H^2 \right). \end{aligned}$$

Proof. The first estimate, corresponding to Theorem 4.1, holds by taking $\mathbf{v}_h = \mathbf{u}_h^{n+1}$ and $q_h = p_h^{n+1}$ in (3.21) and applying the standard identity

$$(4.22) \quad (3a - 4b + c)a = \frac{1}{2} [a^2 - b^2 + (2a - b)^2 - (2b - c)^2 + (a - 2b + c)^2],$$

which provides the numerical dissipation of the BDF2 scheme.

Since the pressure estimate is here based on the control of the time derivative, $\tilde{D}\mathbf{u}_h^{n+1}$, we take $\mathbf{v}_h = \tilde{D}\mathbf{u}_h^{n+1}$ and $q_h = 0$ in (3.21). In particular, for the coupling term $b(p_h^{n+1}, \tilde{D}\mathbf{u}_h^{n+1})$, using (3.21) and (4.22), we have

$$(4.23) \quad \begin{aligned} b(p_h^{n+1}, \tilde{D}\mathbf{u}_h^{n+1}) &= j(\tilde{D}p_h^{n+1}, p_h^{n+1}) \\ &\geq \frac{1}{4} \bar{D} (j(p_h^{n+1}, p_h^{n+1}) + j(2p_h^{n+1} - p_h^n, 2p_h^{n+1} - p_h^n)) \end{aligned}$$

for $2 \leq n \leq N - 1$, which corresponds to the BDF2 counterpart of (4.8). On the other hand, for $n = 1$, from (3.21) and (3.19), we obtain

$$(4.24) \quad b(p_h^2, \tilde{D}\mathbf{u}_h^2) = \frac{1}{2\delta t} (3j(p_h^2, p_h^2) - 4j(p_h^1, p_h^2) + b(p_h^2, \mathbf{u}_h^0)).$$

If the initial velocity approximation is given in terms of the Ritz-projection, $\mathbf{u}_h^0 = P_h^k(\mathbf{u}_0, 0)$, it follows that $b(p_h^2, \mathbf{u}_h^0) = j(p_h^0, p_h^2)$, with $p_h^0 \stackrel{\text{def}}{=} R_h^l(\mathbf{u}_0, 0)$. Thus, (4.24) reduces to

$$b(p_h^2, \tilde{D}\mathbf{u}_h^2) = j(\tilde{D}p_h^2, p_h^2),$$

so that (4.23) holds true also for $n = 1$.

Finally, if the initial velocity approximation is given in terms of a general interpolant, $\mathbf{u}_h^0 = \mathcal{I}_h^k \mathbf{u}_0$, we apply an approximation argument (as in (4.14)). Hence,

from (4.24)

(4.25)

$$b\left(p_h^2, \tilde{D}\mathbf{u}_h^2\right) \geq \frac{1}{2\delta t} \left[3j(p_h^2, p_h^2) - 4j(p_h^1, p_h^1) - C_{\mathcal{I}} \left(\nu h^{2(r_{\mathbf{u}}-1)} \|p_h^2\|_Q^2 + \|\mathbf{u}_0\|_{r_{\mathbf{u}}, \Omega}^2 \right) \right],$$

which is the BDF2 counterpart of (4.14) and leads to the stability condition (4.16). The rest of the proof follows with minor modifications. \square

Remark 4.7. A bound for the backward Euler initialization terms $\|\mathbf{u}_h^1\|_H$ and $\|(\mathbf{u}_h^1, p_h^1)\|_h$, appearing in the above estimates, is provided by Theorems 4.1 and 4.2 with $n = 1$.

Remark 4.8. When the initial velocity approximation is not discretely divergence free, the stability condition (4.16) has to be satisfied twice when using BDF2, at the first time step (according to (4.25)) and at the backward Euler initialization (see Theorem 4.2).

5. Convergence. In this section we provide optimal convergence error estimates for the discrete formulation (3.19), the backward Euler scheme.

Theorem 5.2 concerns the convergence for the velocity and gives an estimate for the pressure in terms of the error in the velocity time derivative. Theorem 5.3 answers the question of optimal convergence of the pressure by providing an optimal error estimate for the time derivative, provided the exact pressure is smooth. Finally, Theorem 5.4 provides an improved $L^\infty((0, T), H)$ estimate that justifies the initialization of the BDF2 scheme with a backward Euler step.

The following result expresses the modified Galerkin orthogonality in terms of the consistency error in space and time.

LEMMA 5.1 (consistency error). *Let (\mathbf{u}, p) be the solution of (2.1) and let $\{(\mathbf{u}_h^n, p_h^n)\}_{0 \leq n \leq N}$ be the solution of (3.19). Assume that $\mathbf{u} \in C^0([0, T]; V) \cap C^1((0, T]; H)$, and let $p \in C^0((0, T]; Q)$. Then, for $0 \leq n \leq N - 1$, there holds*

$$\begin{aligned} & (\bar{D}\mathbf{u}(t_{n+1}) - \bar{D}\mathbf{u}_h^{n+1}, \mathbf{v}_h) + a(\mathbf{u}(t_{n+1}) - \mathbf{u}_h^{n+1}, \mathbf{v}_h) + b(p(t_{n+1}) - p_h^{n+1}, \mathbf{v}_h) \\ & - b(q_h, \mathbf{u}(t_{n+1}) - \mathbf{u}_h^{n+1}) = j(p_h^{n+1}, q_h) + (\bar{D}\mathbf{u}(t_{n+1}) - \partial_t \mathbf{u}(t_{n+1}), \mathbf{v}_h) \end{aligned}$$

for all $(\mathbf{v}_h, q_h) \in [V_h^k]^d \times Q_h^l$.

THEOREM 5.2. *Assume that $\mathbf{u} \in H^1(0, T; [H^r(\Omega)]^d) \cap H^2(0, T; [L^2(\Omega)]^d)$ and $p \in C^0((0, T]; H^s(\Omega))$ with $r \geq 2$ and $s \geq 1$, and set $\mathbf{u}_h^0 \in [V_h^k]^d$ as a given approximation of \mathbf{u}_0 . Then the following estimate holds for $1 \leq n \leq N$:*

$$\begin{aligned} \|\mathbf{u}_h^n - \mathbf{u}(t_n)\|_H^2 &+ \sum_{m=0}^{n-1} \delta t \|(\mathbf{u}_h^{m+1} - \mathbf{u}(t_{m+1}), p_h^{m+1})\|_h^2 \leq \|\mathcal{I}_h^k \mathbf{u}_0 - \mathbf{u}_h^0\|_H^2 \\ &+ Ch^{2r_{\mathbf{u}}} \left(\|\mathbf{u}\|_{C^0([t_1, t_n]; H^{r_{\mathbf{u}}}(\Omega))}^2 + \nu^{-1} \|\partial_t \mathbf{u}\|_{L^2(0, t_n; H^{r_{\mathbf{u}}}(\Omega))}^2 \right) \\ &+ C \left(\frac{\delta t^2}{\nu} \|\partial_{tt} \mathbf{u}\|_{L^2(0, t_n; H)}^2 + \frac{h^{2s_p}}{\nu} t_n \|p\|_{C^0([t_1, t_n]; H^{s_p}(\Omega))}^2 \right. \\ &\quad \left. + \nu h^{2(r_{\mathbf{u}}-1)} t_n \|\mathbf{u}\|_{C^0([t_1, t_n]; H^{r_{\mathbf{u}}}(\Omega))}^2 \right), \end{aligned}$$

$$\begin{aligned} \sum_{m=0}^{n-1} \delta t \|p_h^{m+1} - p(t_{m+1})\|_Q^2 &\leq C \left(1 + \frac{1}{\beta^2}\right) \frac{h^{2s_p}}{\nu} t_n \|p\|_{C^0([t_1, t_n]; H^{s_p}(\Omega))}^2 \\ &+ \frac{C}{\beta^2} \sum_{m=0}^{n-1} \delta t \left(\|(\mathbf{u}_h^{m+1} - \mathbf{u}(t_{m+1}), p_h^{m+1})\|_h^2 + \nu^{-1} \|\partial_t \mathbf{u}(t_{m+1}) - \bar{D}\mathbf{u}_h^{m+1}\|_H^2 \right), \end{aligned}$$

with $C > 0$ a positive constant independent of h , δt , and ν .

Proof. The error estimate for the velocity follows standard energy arguments, and for the pressure we use the modified inf-sup condition (3.8). We start by decomposing the velocity and pressure error using, respectively, the projections \mathcal{I}_h^k and Π_h^l . This yields

$$\begin{aligned} \mathbf{u}(t_{n+1}) - \mathbf{u}_h^{n+1} &= \underbrace{\mathbf{u}(t_{n+1}) - \mathcal{I}_h^k \mathbf{u}(t_{n+1})}_{\boldsymbol{\theta}_\pi^{n+1}} + \underbrace{\mathcal{I}_h^k \mathbf{u}(t_{n+1}) - \mathbf{u}_h^{n+1}}_{\boldsymbol{\theta}_h^{n+1}} = \boldsymbol{\theta}_\pi^{n+1} + \boldsymbol{\theta}_h^{n+1}, \\ (5.1) \quad p(t_{n+1}) - p_h^{n+1} &= \underbrace{p(t_{n+1}) - \Pi_h^l p(t_{n+1})}_{y_\pi^{n+1}} + \underbrace{\Pi_h^l p(t_{n+1}) - p_h^{n+1}}_{y_h^{n+1}} = y_\pi^{n+1} + y_h^{n+1}. \end{aligned}$$

The first term $\boldsymbol{\theta}_\pi^{n+1}$ can be bounded using approximation (3.5). In order to estimate $\boldsymbol{\theta}_h^{n+1}$ we first note, using (4.3) and the coercivity of the bilinear form $a(\cdot, \cdot) + j(\cdot, \cdot)$,

$$\begin{aligned} (5.2) \quad \frac{1}{2} \bar{D} \|\boldsymbol{\theta}_h^{n+1}\|_H^2 + \|(\boldsymbol{\theta}_h^{n+1}, y_h^{n+1})\|_h^2 &\leq (\bar{D}\boldsymbol{\theta}_h^{n+1}, \boldsymbol{\theta}_h^{n+1}) + \|(\boldsymbol{\theta}_h^{n+1}, y_h^{n+1})\|_h^2 \\ &\leq \underbrace{(\bar{D}\boldsymbol{\theta}_h^{n+1}, \boldsymbol{\theta}_h^{n+1}) + a(\boldsymbol{\theta}_h^{n+1}, \boldsymbol{\theta}_h^{n+1}) + b(y_h^{n+1}, \boldsymbol{\theta}_h^{n+1}) - b(y_h^{n+1}, \boldsymbol{\theta}_h^{n+1}) + j(y_h^{n+1}, y_h^{n+1})}_{T_1^{n+1}}. \end{aligned}$$

In addition, using (5.1) we have

$$\begin{aligned} T_1^{n+1} &= -(\bar{D}\boldsymbol{\theta}_\pi^{n+1}, \boldsymbol{\theta}_h^{n+1}) - a(\boldsymbol{\theta}_\pi^{n+1}, \boldsymbol{\theta}_h^{n+1}) + j(\Pi_h^l p(t_{n+1}), y_h^{n+1}) - b(y_\pi^{n+1}, \boldsymbol{\theta}_h^{n+1}) \\ &\quad + b(y_h^{n+1}, \boldsymbol{\theta}_\pi^{n+1}) + (\bar{D}\mathbf{u}(t_{n+1}) - \bar{D}\mathbf{u}_h^{n+1}, \boldsymbol{\theta}_h^{n+1}) + a(\mathbf{u}(t_{n+1}) - \mathbf{u}_h^{n+1}, \boldsymbol{\theta}_h^{n+1}) \\ &\quad + b(p(t_{n+1}) - p_h^{n+1}, \boldsymbol{\theta}_h^{n+1}) - b(y_h^{n+1}, \mathbf{u}(t_{n+1}) - \mathbf{u}_h^{n+1}) - j(p_h^{n+1}, y_h^{n+1}). \end{aligned}$$

By the modified Galerkin orthogonality (Lemma 5.1), this expression reduces to

$$\begin{aligned} (5.3) \quad T_1^{n+1} &= -(\bar{D}\boldsymbol{\theta}_\pi^{n+1}, \boldsymbol{\theta}_h^{n+1}) + (\bar{D}\mathbf{u}(t_{n+1}) - \partial_t \mathbf{u}(t_{n+1}), \boldsymbol{\theta}_h^{n+1}) \\ &\quad - a(\boldsymbol{\theta}_\pi^{n+1}, \boldsymbol{\theta}_h^{n+1}) + j(\Pi_h^l p(t_{n+1}), y_h^{n+1}) - b(y_\pi^{n+1}, \boldsymbol{\theta}_h^{n+1}) + b(y_h^{n+1}, \boldsymbol{\theta}_\pi^{n+1}). \end{aligned}$$

Now, using the Cauchy–Schwarz and the Poincaré inequalities and (3.6), we have

$$\begin{aligned} (5.4) \quad T_1^{n+1} &\leq \underbrace{\left(\|\bar{D}\mathbf{u}(t_{n+1}) - \partial_t \mathbf{u}(t_{n+1})\|_H + \|\bar{D}\boldsymbol{\theta}_\pi^{n+1}\|_H \right)}_{T_2^{n+1}} \frac{C_P}{\nu^{\frac{1}{2}}} \|(\boldsymbol{\theta}_h^{n+1}, y_h^{n+1})\| \\ &\quad + \left(\|\boldsymbol{\theta}_\pi^{n+1}\|_V + \|y_\pi^{n+1}\|_Q + \nu^{\frac{1}{2}} \|h^{-1} \boldsymbol{\theta}_\pi^{n+1}\|_H \right. \\ &\quad \left. + j(\Pi_h^l p(t_{n+1}), \Pi_h^l p(t_{n+1}))^{\frac{1}{2}} \right) \|(\boldsymbol{\theta}_h^{n+1}, y_h^{n+1})\|. \end{aligned}$$

The term T_2^{n+1} can be treated, in a standard way (see, e.g., [30]), using a Taylor expansion and the Cauchy–Schwarz inequality, which yields

$$(5.5) \quad \begin{aligned} T_2^{n+1} &\leq \frac{1}{\delta t} \int_{t_n}^{t_{n+1}} (\delta t \|\partial_{tt} \mathbf{u}(s)\|_H + \|\partial_t \boldsymbol{\theta}_\pi(s)\|_H) ds \\ &\leq \delta t^{\frac{1}{2}} \|\partial_{tt} \mathbf{u}(s)\|_{L^2((t_n, t_{n+1}); H)} + \delta t^{-\frac{1}{2}} \|\partial_t \boldsymbol{\theta}_\pi\|_{L^2((t_n, t_{n+1}); H)}. \end{aligned}$$

Thus, from (5.4), using Young’s inequality, it follows that

$$\begin{aligned} T_1^{n+1} &\leq \frac{1}{2} \|(\boldsymbol{\theta}_h^{n+1}, y_h^{n+1})\|_h^2 \\ &\quad + C \left[\frac{C_P^2}{\nu} \left(\delta t \|\partial_{tt} \mathbf{u}\|_{L^2((t_n, t_{n+1}); H)}^2 + \delta t^{-1} \|\partial_t \boldsymbol{\theta}_\pi\|_{L^2((t_n, t_{n+1}); H)}^2 \right) \right. \\ &\quad \left. + \|\boldsymbol{\theta}_\pi^{n+1}\|_V^2 + \|y_\pi^{n+1}\|_Q^2 + \nu \|h^{-1} \boldsymbol{\theta}_\pi^{n+1}\|_H^2 + j(\Pi_h^l p(t_{n+1}), \Pi_h^l p(t_{n+1})) \right]. \end{aligned}$$

By inserting this expression into (5.2), multiplying the resulting expression by $2\delta t$, and summing over $0 \leq m \leq n-1$, we obtain

$$\begin{aligned} &\|\boldsymbol{\theta}_h^n\|_H^2 + \sum_{m=0}^{n-1} \delta t \|(\boldsymbol{\theta}_h^{m+1}, y_h^{m+1})\|_h^2 \\ &\leq \|\boldsymbol{\theta}_h^0\|_H^2 + C \left[\delta t^2 \nu^{-1} \|\partial_{tt} \mathbf{u}\|_{L^2(0, t_n; H)}^2 + \nu^{-1} \|\partial_t \boldsymbol{\theta}_\pi\|_{L^2(0, t_n; H)}^2 \right. \\ &\quad \left. + \sum_{m=0}^{n-1} \delta t \left(\|\boldsymbol{\theta}_\pi^{m+1}\|_V^2 + \|y_\pi^{m+1}\|_Q^2 + \nu \|h^{-1} \boldsymbol{\theta}_\pi^{m+1}\|_H^2 + j(\Pi_h^l p(t_{m+1}), \Pi_h^l p(t_{m+1})) \right) \right]. \end{aligned}$$

Finally, the velocity error estimate is obtained using approximation (3.5) and the consistency of the pressure stabilization (3.3), which yields

$$\begin{aligned} &\|\boldsymbol{\theta}_h^n\|_H^2 + \sum_{m=0}^{n-1} \delta t \|(\boldsymbol{\theta}_h^{m+1}, y_h^{m+1})\|_h^2 \leq \|\boldsymbol{\theta}_h^0\|_H^2 \\ &\quad + C \left[\frac{\delta t^2}{\nu} \|\partial_{tt} \mathbf{u}\|_{L^2(0, t_n; H)}^2 + \frac{h^{2r_u}}{\nu} \|\partial_t \mathbf{u}\|_{L^2(0, t_n; H^{r_u}(\Omega))}^2 \right. \\ &\quad \left. + \nu h^{2(r_u-1)} \sum_{m=0}^{n-1} \delta t \|\mathbf{u}(t_{m+1})\|_{r_u, \Omega}^2 + \frac{h^{2s_p}}{\nu} \sum_{m=0}^{n-1} \delta t \|p(t_{m+1})\|_{s_p, \Omega}^2 \right]. \end{aligned}$$

For the pressure error estimate we first note that, from (5.1), it suffices to control $\|y_h^{n+1}\|_{0, \Omega}$. To this end, we use the modified inf-sup condition (3.8):

$$(5.6) \quad \beta \|y_h^{n+1}\|_Q \leq \sup_{\mathbf{v}_h \in [V_h^k]^d} \frac{|b(y_h^{n+1}, \mathbf{v}_h)|}{\|\mathbf{v}_h\|_V} + C j(y_h^{n+1}, y_h^{n+1})^{\frac{1}{2}}.$$

From (5.1) we get

$$b(y_h^{n+1}, \mathbf{v}_h) = -b(y_\pi^{n+1}, \mathbf{v}_h) + b(p(t_{n+1}) - p_h^{n+1}, \mathbf{v}_h).$$

The first term can be bounded, using the continuity of $b(\cdot, \cdot)$ (see (2.3)), which yields

$$b(y_\pi^{n+1}, \mathbf{v}_h) \leq \|y_\pi^{n+1}\|_Q \|\mathbf{v}_h\|_V.$$

On the other hand, using the modified Galerkin orthogonality (Lemma 5.1 with $q_h = 0$) we have

$$\begin{aligned} & b(p(t_{n+1}) - p_h^{n+1}, \mathbf{v}_h) \\ &= -a(\mathbf{u}(t_{n+1}) - \mathbf{u}_h^{n+1}, \mathbf{v}_h) - (\partial_t \mathbf{u}(t_{n+1}) - \bar{D}\mathbf{u}_h^{n+1}, \mathbf{v}_h) \\ &\leq C \|(\mathbf{u}(t_{n+1}) - \mathbf{u}_h^{n+1}, 0)\|_h \|\mathbf{v}_h\|_V + \|\partial_t \mathbf{u}(t_{n+1}) - \bar{D}\mathbf{u}_h^{n+1}\|_H \|\mathbf{v}_h\|_H. \end{aligned}$$

As a result, from the above estimations we have

$$\beta \|y_h^{n+1}\|_Q \leq C (\|y_\pi^{n+1}\|_Q + \|(\mathbf{u}(t_{n+1}) - \mathbf{u}_h^{n+1}, y_h^{n+1})\|_h) + \frac{C_P}{\nu^{\frac{1}{2}}} \|\partial_t \mathbf{u}(t_{n+1}) - \bar{D}\mathbf{u}_h^{n+1}\|_H.$$

Therefore,

$$\begin{aligned} \beta^2 \sum_{m=0}^{n-1} \delta t \|y_h^{m+1}\|_Q^2 &\leq C \sum_{m=0}^{n-1} \delta t \left(\|y_\pi^{m+1}\|_Q^2 + \|(\mathbf{u}(t_{m+1}) - \mathbf{u}_h^{m+1}, y_h^{m+1})\|_h^2 \right. \\ &\quad \left. + \nu^{-1} \|\partial_t \mathbf{u}(t_{m+1}) - \bar{D}\mathbf{u}_h^{m+1}\|_H^2 \right), \end{aligned}$$

and we conclude using approximation and the error estimate for the velocity. \square

We solve the problem of the pressure convergence by providing an error estimate for the time derivative of the velocity.

THEOREM 5.3. *Under the assumptions of Theorem 5.2, assuming that $p \in C^0([0, T]; H^s(\Omega))$, $\mathbf{u}_0 \in V \cap H_0(\text{div}; \Omega)$, and $\mathbf{u}_h^0 \stackrel{\text{def}}{=} P_h^k(\mathbf{u}_0, 0)$, for $1 \leq n \leq N$ we have*

$$\begin{aligned} & \sum_{m=0}^{n-1} \delta t \|\bar{D}\mathbf{u}_h^{m+1} - \partial_t \mathbf{u}(t_{m+1})\|_H^2 + \|(P_h^k(\mathbf{u}(t_n), p(t_n)) - \mathbf{u}_h^n, R_h^l(\mathbf{u}(t_n), p(t_n)) - p_h^n)\|_h^2 \\ & \leq C \left(\delta t^2 \|\partial_t \mathbf{u}\|_{L^2(0, T; H)}^2 + h^{2r_u} \|\partial_t \mathbf{u}\|_{L^2(0, T; H^{r_u}(\Omega))}^2 \right) + C \frac{h^{2s_p}}{\nu} \|p(0)\|_{s_p, \Omega}^2. \end{aligned}$$

Proof. In order to provide an optimal error estimate, we decompose the error in terms of the Ritz-projection operator (3.16) as follows:

(5.7)

$$\begin{aligned} \mathbf{u}(t_{n+1}) - \mathbf{u}_h^{n+1} &= \underbrace{\mathbf{u}(t_{n+1}) - P_h^k(\mathbf{u}(t_{n+1}), p(t_{n+1}))}_{\boldsymbol{\theta}_\pi^{n+1}} + \underbrace{P_h^k(\mathbf{u}(t_{n+1}), p(t_{n+1})) - \mathbf{u}_h^{n+1}}_{\boldsymbol{\theta}_h^{n+1}} \\ &= \boldsymbol{\theta}_\pi^{n+1} + \boldsymbol{\theta}_h^{n+1}, \\ p(t_{n+1}) - p_h^{n+1} &= \underbrace{p(t_{n+1}) - R_h^l(\mathbf{u}(t_{n+1}), p(t_{n+1}))}_{y_\pi^{n+1}} + \underbrace{R_h^l(\mathbf{u}(t_{n+1}), p(t_{n+1})) - p_h^{n+1}}_{y_h^{n+1}} \\ &= y_\pi^{n+1} + y_h^{n+1}. \end{aligned}$$

Using the triangle inequality, we then have

$$(5.8) \quad \sum_{m=0}^{n-1} \delta t \|\partial_t \mathbf{u}(t_{m+1}) - \bar{D}\mathbf{u}_h^{m+1}\|_H^2 \\ \leq C \sum_{m=0}^{n-1} \delta t (\|\partial_t \mathbf{u}(t_{m+1}) - \bar{D}\mathbf{u}(t_{m+1})\|_H^2 + \|\bar{D}\boldsymbol{\theta}_\pi^{m+1}\|_H^2 + \|\bar{D}\boldsymbol{\theta}_h^{m+1}\|_H^2).$$

For the first term, we proceed as in (5.5) using a Taylor expansion, which yields

$$\|\partial_t \mathbf{u}(t_{n+1}) - \bar{D}\mathbf{u}(t_{n+1})\|_H \leq \delta t^{\frac{1}{2}} \|\partial_{tt} \mathbf{u}(s)\|_{L^2((t_n, t_{n+1}); H)}.$$

For the second term, we have

$$(5.9) \quad \|\bar{D}\boldsymbol{\theta}_\pi^{n+1}\|_H = \frac{1}{\delta t} \int_{t_n}^{t_{n+1}} \|\partial_t \boldsymbol{\theta}_\pi(s)\|_H ds \leq \delta t^{-\frac{1}{2}} \|\partial_t \boldsymbol{\theta}_\pi\|_{L^2((t_n, t_{n+1}); H)}.$$

Finally, for the third term we use the modified Galerkin orthogonality (Lemma 5.1 with $q_h = 0$) and the definition of the Ritz-projection (3.16) to obtain

$$\begin{aligned} & \|\bar{D}\boldsymbol{\theta}_h^{n+1}\|_H^2 + a(\boldsymbol{\theta}_h^{n+1}, \bar{D}\boldsymbol{\theta}_h^{n+1}) + b(y_h^{n+1}, \bar{D}\boldsymbol{\theta}_h^{n+1}) \\ &= -(\bar{D}\boldsymbol{\theta}_\pi^{n+1}, \bar{D}\boldsymbol{\theta}_h^{n+1}) - a(\boldsymbol{\theta}_\pi^{n+1}, \bar{D}\boldsymbol{\theta}_h^{n+1}) \\ & \quad - b(y_\pi^{n+1}, \bar{D}\boldsymbol{\theta}_h^{n+1}) + (\bar{D}\mathbf{u}(t_{n+1}) - \partial_t \mathbf{u}(t_{n+1}), \bar{D}\boldsymbol{\theta}_h^{n+1}) \\ &= -(\bar{D}\boldsymbol{\theta}_\pi^{n+1}, \bar{D}\boldsymbol{\theta}_h^{n+1}) + (\bar{D}\mathbf{u}(t_{n+1}) - \partial_t \mathbf{u}(t_{n+1}), \bar{D}\boldsymbol{\theta}_h^{n+1}). \end{aligned}$$

Young's inequality yields

$$\begin{aligned} & \frac{1}{2} \|\bar{D}\boldsymbol{\theta}_h^{n+1}\|_H^2 + a(\boldsymbol{\theta}_h^{n+1}, \bar{D}\boldsymbol{\theta}_h^{n+1}) + b(y_h^{n+1}, \bar{D}\boldsymbol{\theta}_h^{n+1}) \\ & \leq C (\|\bar{D}\boldsymbol{\theta}_\pi^{n+1}\|_H^2 + \|\bar{D}\mathbf{u}(t_{n+1}) - \partial_t \mathbf{u}(t_{n+1})\|_H^2). \end{aligned}$$

In addition, for $0 \leq n \leq N$, testing (3.16) at the time level n with $\mathbf{v}_h = \mathbf{0}$, we have

$$(5.10) \quad b(q_h, P_h^k(\mathbf{u}(t_n), p(t_n))) = j(R_h^l(\mathbf{u}(t_n), p(t_n)), q_h).$$

On the other hand, for $1 \leq n \leq N$, testing (3.19) at the time level n with $\mathbf{v}_h = \mathbf{0}$ and since, by definition, $\mathbf{u}_h^0 \stackrel{\text{def}}{=} P_h^k(\mathbf{u}_0, 0)$, we have

$$(5.11) \quad b(q_h, \mathbf{u}_h^n) = j(p_h^n, q_h)$$

for all $q_h \in Q_h^l$ and $0 \leq n \leq N$ and where we have defined $p_h^0 \stackrel{\text{def}}{=} R_h^l(\mathbf{u}_0, 0)$. As a result, from (5.10)–(5.11), we have

$$b(q_h, \boldsymbol{\theta}_h^n) = j(y_h^n, q_h)$$

for all $q_h \in Q_h^l$ and $0 \leq n \leq N$. We therefore have, for $0 \leq n \leq N-1$,

$$b(y_h^{n+1}, \bar{D}\boldsymbol{\theta}_h^{n+1}) = j(\bar{D}y_h^{n+1}, y_h^{n+1}).$$

On the other hand, using the symmetry of a and j , we have

$$\begin{aligned} a(\boldsymbol{\theta}_h^{n+1}, \bar{D}\boldsymbol{\theta}_h^{n+1}) &= \frac{1}{2}\bar{D}a(\boldsymbol{\theta}_h^{n+1}, \boldsymbol{\theta}_h^{n+1}) + \frac{\delta t}{2}a(\bar{D}\boldsymbol{\theta}_h^{n+1}, \bar{D}\boldsymbol{\theta}_h^{n+1}), \\ j(y_h^{n+1}, \bar{D}y_h^{n+1}) &= \frac{1}{2}\bar{D}j(y_h^{n+1}, y_h^{n+1}) + \frac{\delta t}{2}j(\bar{D}y_h^{n+1}, \bar{D}y_h^{n+1}), \end{aligned}$$

so that

$$\begin{aligned} \frac{1}{2}\|\bar{D}\boldsymbol{\theta}_h^{n+1}\|_H^2 + \frac{1}{2}\bar{D}(a(\boldsymbol{\theta}_h^{n+1}, \boldsymbol{\theta}_h^{n+1}) + j(y_h^{n+1}, y_h^{n+1})) \\ \leq \|\bar{D}\boldsymbol{\theta}_\pi^{n+1}\|_H^2 + \|\bar{D}\mathbf{u}(t_{n+1}) - \partial_t\mathbf{u}(t_{n+1})\|_H^2. \end{aligned}$$

Thus, after multiplication by $2\delta t$ and summation over $0 \leq n \leq N - 1$, we have

$$\begin{aligned} (5.12) \quad \sum_{m=0}^{n-1} \delta t \|\bar{D}\boldsymbol{\theta}_h^{m+1}\|_H^2 + \|\boldsymbol{\theta}_h^n, y_h^n\|_h^2 \\ \leq \|\boldsymbol{\theta}_h^0, y_h^0\|_h^2 + C \sum_{m=0}^{n-1} \delta t (\|\bar{D}\boldsymbol{\theta}_\pi^{m+1}\|_H^2 + \|\bar{D}\mathbf{u}(t_{m+1}) - \partial_t\mathbf{u}(t_{m+1})\|_H^2). \end{aligned}$$

For the initial terms, we use the linearity of the Ritz-projection and its approximation properties (Lemma 3.4) to obtain

$$\begin{aligned} \|\boldsymbol{\theta}_h^0, y_h^0\|_h^2 &= \|(P_h^k(\mathbf{0}, p(0)), R_h^l(\mathbf{0}, p(0)))\|_h^2 \\ &\leq \frac{C}{\nu} h^{2s_p} \|p(0)\|_{s_p, \Omega}^2. \end{aligned}$$

Therefore, using (5.9) and (5.5), we have

$$\sum_{m=0}^{n-1} \delta t \|\bar{D}\boldsymbol{\theta}_h^{m+1}\|_H^2 + \|\boldsymbol{\theta}_h^n, y_h^n\|_h^2 \leq C \left(\frac{h^{2s_p}}{\nu} \|p(0)\|_{s_p, \Omega}^2 + \delta t^2 \|\partial_{tt}\mathbf{u}\|_{L^2(0, t_n; H)}^2 \right)$$

for $1 \leq n \leq N$. \square

Finally, for completeness, we here give a result of optimal convergence in the $L^\infty((0, T), H)$ -norm. For this we assume that the domain Ω is such that the optimal convergence in the H -norm holds for the Ritz-projection (see Lemma 3.4). This result is of importance since it shows that the initialization of the BDF2 method using one BDF1 step is justified (i.e., we keep error optimality in time).

THEOREM 5.4. *Assume that the domain Ω is sufficiently smooth so that the H -estimate (3.18) holds. Assume also that $\mathbf{u} \in H^1(0, T; [H_{\mathbf{u}}^r(\Omega)]^d) \cap H^2(0, T; [L^2(\Omega)]^d)$, $p \in C^0([0, T]; H^{s_p}(\Omega))$ with $r_{\mathbf{u}} \geq 2$, $s_p \geq 1$, $\mathbf{u}_0 \in V \cap H_0(\text{div}; \Omega)$, and $\mathbf{u}_h^0 \stackrel{\text{def}}{=} P_h^k(\mathbf{u}_0, 0)$. Then the following estimate holds for $1 \leq n \leq N$:*

$$\begin{aligned} \|\mathbf{u}(t_n) - \mathbf{u}_h^n\|_H \leq \frac{C}{\nu^{\frac{1}{2}}} \left(h^{r_{\mathbf{u}}} \|\mathbf{u}_0\|_{r_{\mathbf{u}}, \Omega} + h^{s_p+1} \|p(0)\|_{s_p, \Omega} \right. \\ \left. + h^{r_{\mathbf{u}}} \|\partial_t\mathbf{u}\|_{L^1(0, t_n; H^{r_{\mathbf{u}}}(\Omega))} + \delta t \|\partial_{tt}\mathbf{u}\|_{L^1(0, t_n; H)} \right). \end{aligned}$$

Proof. Since the proof is similar to that of Theorem 5.3 and we will give only the outline. Let $\boldsymbol{\theta}_h^{n+1}$ and y_h^{n+1} be defined as in (5.7). From (5.2) and (5.3), it follows that

$$(\bar{D}\boldsymbol{\theta}_h^{n+1}, \boldsymbol{\theta}_h^{n+1}) \leq -(\bar{D}\boldsymbol{\theta}_\pi^{n+1}, \boldsymbol{\theta}_h^{n+1}) + (\bar{D}\mathbf{u}(t_{n+1}) - \partial_t \mathbf{u}(t_{n+1}), \boldsymbol{\theta}_h^{n+1}).$$

Applying now the Cauchy–Schwarz inequality, we have

$$\|\boldsymbol{\theta}_h^{n+1}\|_H \leq \|\boldsymbol{\theta}_h^n\|_H + \delta t (\|\bar{D}\boldsymbol{\theta}_\pi^{n+1}\|_H + \|\bar{D}\mathbf{u}(t_{n+1}) - \partial_t \mathbf{u}(t_{n+1})\|_H),$$

and by summation over n , we get

$$\|\boldsymbol{\theta}_h^n\|_H \leq \|\boldsymbol{\theta}_h^0\|_H + \sum_{m=0}^{n-1} \delta t (\|\bar{D}\boldsymbol{\theta}_\pi^{m+1}\|_H + \|\bar{D}\mathbf{u}(t_{m+1}) - \partial_t \mathbf{u}(t_{m+1})\|_H)$$

for $1 \leq n \leq N$. The first term in the right-hand side can be estimated using Lemma 3.4 since, by definition,

$$(5.13) \quad \boldsymbol{\theta}_h^0 = P_h^k(\mathbf{u}(0), p(0)) - \mathbf{u}_h^0 = P_h^k(\mathbf{u}_0, p(0)) - P_h^k(\mathbf{u}_0, 0) = P_h^k(\mathbf{0}, p(0)).$$

Finally, for the finite difference consistency terms we use a standard argument (see, e.g., [34, Theorem 1.5, page 14]). \square

Remark 5.5. From (5.13), one could pretend to initialize the time-stepping procedure with $\mathbf{u}_h^0 = P_h^k(\mathbf{u}_0, p(0))$ (as in [33], for instance). In practice, however, the initial pressure is unknown, so that the choice $\mathbf{u}_h^0 = P_h^k(\mathbf{u}_0, 0)$ is more convenient. Lemma 3.4 shows that we can preserve optimality while keeping this choice (see also [4]).

Remark 5.6. Note that the above convergence proofs use only stability, Galerkin orthogonality, and the truncation error of the finite difference time approximation scheme. Hence the extension to the second order Crank–Nicholson or BDF2 scheme is straightforward. In particular we recall that the estimate of Theorem 5.4 shows that the initialization using one BDF1 step does not make the convergence deteriorate, provided the solution is sufficiently smooth under the first time step. Indeed, for smooth solutions we expect $\|\partial_{tt}\mathbf{u}\|_{L^1(0, \delta t; H)}$ to be $O(\delta t)$, and hence the global convergence will be second order in spite of the initial low order perturbation.

6. Numerical experiments. In this section we will consider some numerical examples using the CIP stabilization, described in subsection 3.1.1. We present computations demonstrating the optimal convergence using finite element spaces consisting of quadratic functions, for the space discretization, BDF1, BDF2, and the Crank–Nicholson scheme for the time discretization. We also verify numerically that, for small time steps, the pressure is unstable for initial data that are not discretely divergence free. All computations have been performed using FreeFem++ [26].

6.1. Convergence rate in time. We consider problem (2.1) in two dimensions, $\Omega = [0, 1] \times [0, 1]$ and $T = 1$, with nonhomogeneous boundary conditions. The right-hand side \mathbf{f} and the boundary and initial data are chosen in order to ensure that the exact solution is given by

$$\mathbf{u}(x, y, t) = g(t) \begin{pmatrix} \sin(\pi x - 0.7) \sin(\pi y + 0.2) \\ \cos(\pi x - 0.7) \cos(\pi y + 0.2) \end{pmatrix},$$

$$p(x, y, t) = g(t) (\sin(x) \cos(y) + (\cos(1) - 1) \sin(1)),$$

with $g(t) = 1 + t^5 + e^{-\frac{t}{10}} + \sin(t)$.

In order to illustrate the convergence rate in time of the discrete solution, we have used quadratic approximations in space and a mesh parameter $h = 0.01$. In this case, the stability condition (4.16) is always satisfied for the range of time steps considered. Thus, the choice of the Lagrange interpolant or of the Ritz-projection as approximation of the initial velocity give similar results.

In Figures 1(a)–(c) we report the convergences of the errors for the velocities ($\|\cdot\|_{L^\infty(0,T;L^2(\Omega))}$) and the pressures ($\|\cdot\|_{L^2(0,T;L^2(\Omega))}$) for the BDF1, Crank–Nicholson, and BDF2 schemes. In all the numerical examples, both the velocities and the pressures converge at the optimal rate ($O(\delta t)$ for BDF1 and $O(\delta t^2)$ for Crank–Nicholson and BDF2). The BDF2 scheme was initialized using one step of BDF1.

6.2. Behavior in the small time-step limit. In this subsection we illustrate the impact of the initial velocity approximation on the approximate pressures for small time steps. For nondiscrete divergence-free initial approximations, a pressure instability is predicted by Corollary 4.3 unless condition (4.16) is satisfied. In other words, pressure instabilities are expected for very small time steps.

We consider problem (2.1) in two dimensions and with nonhomogeneous boundary conditions. We set $\Omega = [0, 1] \times [0, 1]$, and the right-hand side \mathbf{f} and the boundary data are chosen in order to ensure that the exact (steady) solution is given by

$$\mathbf{u}(x, y, t) = \begin{pmatrix} \sin(\pi x - 0.7) \sin(\pi y + 0.2) \\ \cos(\pi x - 0.7) \cos(\pi y + 0.2) \end{pmatrix},$$

$$p(x, y, t) = \sin x \cos y + (\cos(1) - 1) \sin(1).$$

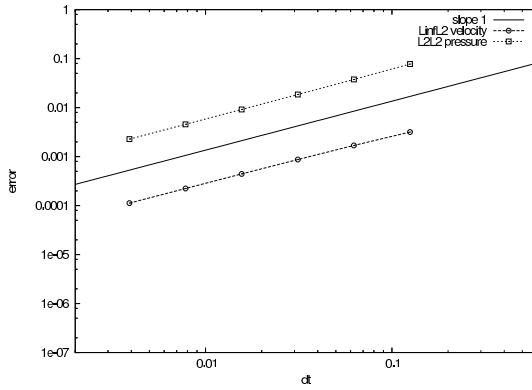
This numerical experiment is, in some degree, motivated by the work reported in [3] (see also [19]), where pressure instabilities, of a different nature, are illustrated for pressure stabilizations involving residuals of the PDEs (e.g., PSPG and GLS). Indeed, the time derivative involved in the residual perturbs the coercivity of the space semidiscrete operator, which leads to pressure instabilities for (sufficiently) small time steps (see [3]). Let us emphasize that, according to section 4, such instabilities do not appear here, in particular since the CIP pressure stabilization (and the other examples of subsection 3.1.1) are consistent without introducing the time derivative.

For different initial velocity approximations, we compare the behavior of the error in the pressure after one time step of the backward Euler scheme, i.e.,

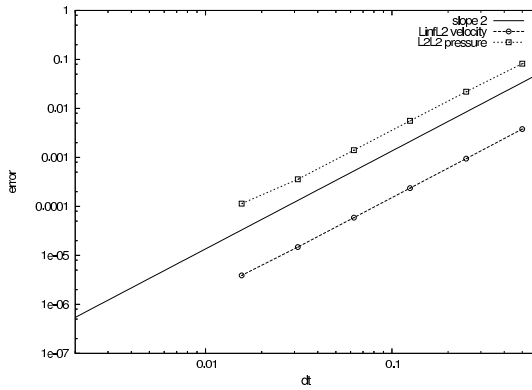
$$\delta t^{\frac{1}{2}} \|p(t_1) - p_h^1\|_Q.$$

We choose the initial data either as the Lagrange interpolant, $\mathbf{u}_h^0 = I_h^k \mathbf{u}_0$, or as the Ritz-projection, $\mathbf{u}_h^0 = P_h^k(\mathbf{u}_0, 0)$.

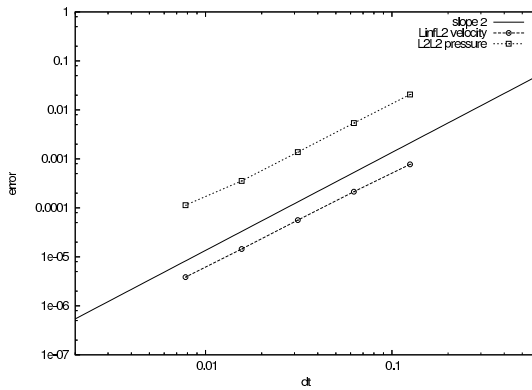
In Figure 2 we have reported the convergence history (in space) of the pressure error, at the first time step, using $\mathbb{P}_1/\mathbb{P}_1$ finite elements for different time step sizes. The pressure instability for small time steps is illustrated in Figure 4(a), where the initial velocity approximation is given in terms of the Lagrange interpolant. Indeed, we can observe that the pressure error has the right convergence rate in space, but it grows when the time step is decreased. On the other hand, as shown in Figure 4(b), the instability is eliminated when the initial velocity approximation is provided by the Ritz-projection, as stated in Corollary 4.3. In this case the error remains bounded (dominated by the space discretization) while reducing the time-step size.



(a) BDF1 scheme



(b) Crank-Nicholson



(c) BDF2 scheme

FIG. 1. Convergence history in time: $\mathbb{P}_2/\mathbb{P}_2$ CIP stabilized finite elements.

Similar results are found with $\mathbb{P}_2/\mathbb{P}_2$ finite elements, as shown in Figure 3. In particular, we can notice, from Figures 2(a) and 3(a), that for quadratic approximations the pressure instability shows up only for very small time steps. As a matter of fact, condition (4.16) is less restrictive for quadratic than for affine velocity approximations of smooth initial data. Finally, some pressure contours are reported in

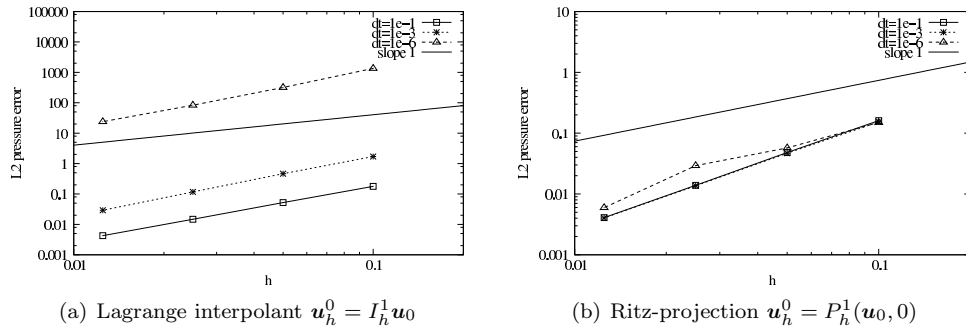


FIG. 2. Convergence history: $\mathbb{P}_1/\mathbb{P}_1$ finite elements.

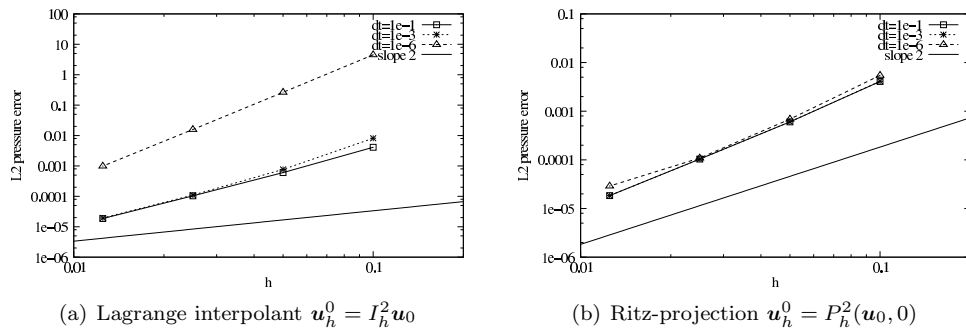


FIG. 3. Convergence history: $\mathbb{P}_2/\mathbb{P}_2$ finite elements.

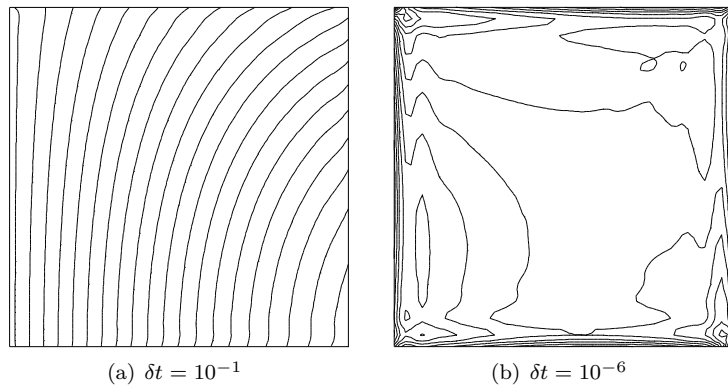


FIG. 4. Pressure contour lines with $\mathbb{P}_2/\mathbb{P}_2$ finite elements in a 40×40 mesh: $\mathbf{u}_h^0 = I_h^2 \mathbf{u}_0$.

Figure 4 for the Lagrange interpolation, and in Figure 5 for the Ritz-projection. The pressure degradation is clearly visible in Figure 4, whereas with the Ritz-projection initialization (Figure 5) the pressure remains unconditionally stable.

7. Conclusion. In this paper we have proved unconditional stability and optimal error estimates, in natural norms, for pressure stabilized finite element approximations of the transient Stokes problem. It should be noted that the extension of

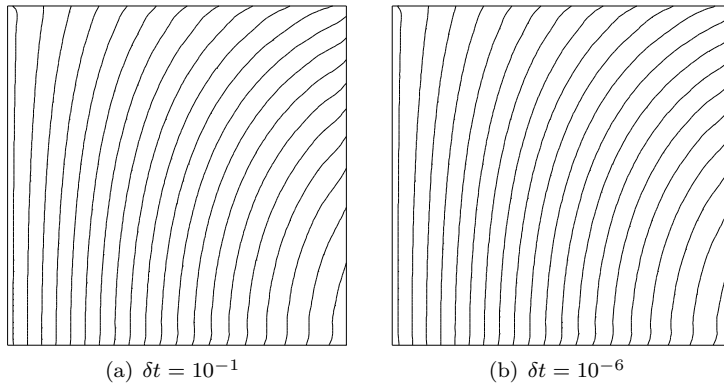


FIG. 5. Pressure contour lines with $\mathbb{P}_2/\mathbb{P}_2$ finite elements in a 40×40 mesh: $\mathbf{u}_h^0 = P_h^2(\mathbf{u}_0, 0)$.

the present results to mixed formulations of the Poisson problem is straightforward. We have shown that for small initial time steps the use of a pressure stabilization dependent Ritz-projection, for the initial data, is essential to avoid pressure instabilities, unless a condition between time and space discretization parameters is satisfied. From the analysis, we also conclude that a second order scheme (e.g., BDF2) can be initialized (without optimality loss) using a first step with BDF1, *provided* that the Ritz-projection (3.16) is used for the initial data.

It is interesting to note that for low order elements the weakly consistent stabilization operators still yield optimal convergence in time when used with a second order scheme. However, in the case when streamline upwind Petrov–Galerkin (SUPG)-type stabilization is used for the convective term, the convergence order in time will be lost unless full consistency is guaranteed in the stabilization term. This is why SUPG-type stabilizations prompt space time finite element formulations with discontinuous approximation in time.

Some of the methods described in subsection 3.1.1, on the other hand, may be extended to the case of Oseen’s equations, handling all Reynolds numbers, by applying the same type of stabilizing term for the convection (see [13, 6, 16, 7] for details).

REFERENCES

- [1] S. BADIA AND R. CODINA, *On a Multiscale Approach to the Transient Stokes Problem. Transient Subscales and Anisotropic Space-Time Discretization*, preprint, 2007.
- [2] R. BECKER AND M. BRAACK, *A finite element pressure gradient stabilization for the Stokes equations based on local projections*, *Calcolo*, 38 (2001), pp. 173–199.
- [3] P. B. BOCHEV, M. D. GUNZBURGER, AND R. B. LEHOUCQ, *On stabilized finite element methods for the Stokes problem in the small time step limit*, *Internat. J. Numer. Methods Fluids*, 53 (2007), pp. 573–597.
- [4] D. BOFFI AND L. GASTALDI, *Analysis of finite element approximation of evolution problems in mixed form*, *SIAM J. Numer. Anal.*, 42 (2004), pp. 1502–1526.
- [5] M. BOMAN, *Estimates for the L_2 -projection onto continuous finite element spaces in a weighted L_p -norm*, *BIT*, 46 (2006), pp. 249–260.
- [6] M. BRAACK AND E. BURMAN, *Local projection stabilization for the Oseen problem and its interpretation as a variational multiscale method*, *SIAM J. Numer. Anal.*, 43 (2006), pp. 2544–2566.
- [7] M. BRAACK, E. BURMAN, V. JOHN, AND G. LUBE, *Stabilized finite element methods for the generalized Oseen problem*, *Comput. Methods Appl. Mech. Engrg.*, 196 (2007), pp. 853–866.

- [8] J. H. BRAMBLE, J. E. PASCIAK, AND O. STEINBACH, *On the stability of the L^2 projection in $H^1(\Omega)$* , Math. Comp., 71 (2002), pp. 147–156.
- [9] F. BREZZI AND M. FORTIN, *A minimal stabilisation procedure for mixed finite element methods*, Numer. Math., 89 (2001), pp. 457–491.
- [10] F. BREZZI AND J. PITKÄRANTA, *On the stabilization of finite element approximations of the Stokes equations*, in Efficient Solutions of Elliptic Systems (Kiel, 1984), Notes Numer. Fluid Mech. 10, Vieweg, Braunschweig, 1984, pp. 11–19.
- [11] A. N. BROOKS AND T. J. R. HUGHES, *Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier–Stokes equations*, Comput. Methods Appl. Mech. Engrg., 32 (1982), pp. 199–259.
- [12] E. BURMAN, *Pressure projection stabilizations for Galerkin approximations of Stokes and Darcy’s problem*, Numer. Methods Partial Differential Equations, 24 (2008), pp. 127–143.
- [13] E. BURMAN, M. A. FERNÁNDEZ, AND P. HANSBO, *Continuous interior penalty finite element method for Oseen’s equations*, SIAM J. Numer. Anal., 44 (2006), pp. 1248–1274.
- [14] E. BURMAN AND P. HANSBO, *Edge stabilization for the generalized Stokes problem: A continuous interior penalty method*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 2393–2410.
- [15] P. CLÉMENT, *Approximation by finite element functions using local regularization*, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. RAIRO Analyse Numérique, 9 (1975), pp. 77–84.
- [16] R. CODINA, *Analysis of a stabilized finite element approximation of the Oseen equations using orthogonal subscales*, Appl. Numer. Math., 58 (2008), pp. 2413–2430.
- [17] R. CODINA AND J. BLASCO, *A finite element formulation for the Stokes problem allowing equal velocity-pressure interpolation*, Comput. Methods Appl. Mech. Engrg., 143 (1997), pp. 373–391.
- [18] R. CODINA AND J. BLASCO, *Stabilized finite element method for the transient Navier-Stokes equations based on a pressure gradient projection*, Comput. Methods Appl. Mech. Engrg., 182 (2000), pp. 277–300.
- [19] R. CODINA, J. PRINCIPE, O. GUASCH, AND S. BADIA, *Time dependent subscales in the stabilized finite element approximation of incompressible flow problems*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 2413–2430.
- [20] M. CROUZEIX AND V. THOMÉE, *The stability in L_p and W_p^1 of the L_2 -projection onto finite element function spaces*, Math. Comp., 48 (1987), pp. 521–532.
- [21] C. R. DOHRMANN AND P. B. BOCHEV, *A stabilized finite element method for the Stokes problem based on polynomial pressure projections*, Internat. J. Numer. Methods Fluids, 46 (2004), pp. 183–201.
- [22] A. ERN AND J.-L. GUERMOND, *Theory and Practice of Finite Elements*, Appl. Math. Sci. 159, Springer-Verlag, New York, 2004.
- [23] L. P. FRANCA AND S. L. FREY, *Stabilized finite element methods. II. The incompressible Navier–Stokes equations*, Comput. Methods Appl. Mech. Engrg., 99 (1992), pp. 209–233.
- [24] J. FREUND AND R. STENBERG, *On weakly imposed boundary conditions for second order problems*, in Proceedings of the Ninth International Conference on Finite Elements in Fluids, M. Morandi Cecchi et al., eds., Venice, 1995, pp. 327–336. Available online at <http://math.tkk.fi/~rstenber/Publications/Venice95.pdf>.
- [25] V. GIRAULT AND P. A. RAVIART, *Finite Element Methods for Navier–Stokes Equations*, Comput. Math. 5, Springer-Verlag, Berlin, 1986.
- [26] F. HECHT, O. PIRONNEAU, A. LE HYARIC, AND K. OHTSUKA, *FreeFem++ v. 2.11. User’s Manual*, Laboratoire J. L. Lions, University of Paris 6, Paris, France.
- [27] T. J. R. HUGHES, L. P. FRANCA, AND M. BALESTRA, *A new finite element formulation for computational fluid dynamics. V. Circumventing the Babuška-Brezzi condition: A stable Petrov-Galerkin formulation of the Stokes problem accommodating equal-order interpolations*, Comput. Methods Appl. Mech. Engrg., 59 (1986), pp. 85–99.
- [28] G. MATTHIES, P. SKRZYPACZ, AND L. TOBISKA, *A Unified Analysis for Local Projection Stabilisations Applied to the Oseen Problem*, Technical report, Otto-von-Guericke-Universität Magdeburg, Magdeburg, Germany, 2007.
- [29] J. NITSCHKE, *Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind*, Abh. Math. Sem. Univ. Hamburg, 36 (1971), pp. 9–15.
- [30] A. QUARTERONI AND A. VALLI, *Numerical Approximation of Partial Differential Equations*, Comput. Math. 23, Springer-Verlag, Berlin, 1994.
- [31] L. R. SCOTT AND S. ZHANG, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.

- [32] D. SILVESTER, *Optimal low order finite element methods for incompressible flow*, Comput. Methods Appl. Mech. Engrg., 111 (1994), pp. 357–368.
- [33] M. SURI, *Mixed Variational Principles for Time-Dependent Problems*, Ph.D. thesis, Carnegie-Mellon University, Pittsburgh, PA, 1983.
- [34] V. THOMÉE, *Galerkin Finite Element Methods for Parabolic Problems*, 2nd ed., Comput. Math. 25, Springer-Verlag, Berlin, 2006.

LOCALIZED LINEAR POLYNOMIAL OPERATORS AND QUADRATURE FORMULAS ON THE SPHERE*

Q. T. LE GIA[†] AND H. N. MHASKAR[‡]

Abstract. The purpose of this paper is to construct universal, auto-adaptive, localized, linear, polynomial (-valued) operators based on scattered data on the (hyper) sphere \mathbb{S}^q ($q \geq 2$). The approximation and localization properties of our operators are studied theoretically in deterministic as well as probabilistic settings. Numerical experiments are presented to demonstrate their superiority over traditional least squares and discrete Fourier projection polynomial approximations. An essential ingredient in our construction is the construction of quadrature formulas based on scattered data, exact for integrating spherical polynomials of (moderately) high degree. Our formulas are based on scattered sites; i.e., in contrast to such well-known formulas as Driscoll–Healy formulas, we need not choose the location of the sites in any particular manner. While the previous attempts to construct such formulas have yielded formulas exact for spherical polynomials of degree at most 18, we are able to construct formulas exact for spherical polynomials of degree 178.

Key words. quadrature formulas, localized kernels, polynomial quasi interpolation, learning theory on the sphere

AMS subject classifications. 65D32, 41A10, 41A25

DOI. 10.1137/060678555

1. Introduction. The problem of approximation of functions on the sphere arises in almost all applications involving modeling of data collected on the surface of the earth. More recent applications such as manifold matching and neural networks have led to the approximation of functions on the unit sphere \mathbb{S}^q embedded in the Euclidean space \mathbb{R}^{q+1} for integers $q \geq 3$ as well. Various applications in learning theory, meteorology, cosmology, and geophysics require analysis of *scattered data* collected on the sphere [9, 7, 8]. This means that the data is of the form $\{(\xi, f(\xi))\}$ for some unknown function $f : \mathbb{S}^q \rightarrow \mathbb{R}$, where one has no control on the choice of the sites ξ .

There are many methods to model such data: spherical splines, radial basis functions (called zonal function networks in this context), etc. However, the most traditional method is to approximate by spherical polynomials, i.e., restrictions of algebraic polynomials in $q+1$ variables to \mathbb{S}^q . Apart from tradition, some important advantages of polynomials are that they are eigenfunctions of many pseudodifferential operators which arise in practical applications and that they are infinitely smooth. Unlike in the case of spline approximation with a given degree of the piecewise component polynomials, global polynomial approximation does not exhibit a saturation property [2, section 2, Chapter 11]; i.e., for an arbitrary sequence $\delta_n \downarrow 0$, it is possible to find a continuous function on the sphere, not itself a polynomial, which can be approximated by spherical polynomials of degree at most n uniformly within δ_n , $n \geq 1$. In [21, 19], we have shown how a good polynomial approximation also yields a good zonal function

*Received by the editors December 27, 2006; accepted for publication (in revised form) August 6, 2008; published electronically December 5, 2008.

<http://www.siam.org/journals/sinum/47-1/67855.html>

[†]School of Mathematics, University of New South Wales, Sydney, NSW 2052, Australia (qlegia@unsw.edu.au). The research of this author was supported by the Australian Research Council under its Centres of Excellence Program.

[‡]Department of Mathematics, California State University, Los Angeles, CA 90032 (hmhaska@calstatela.edu). The research of this author was supported in part by grant DMS-0605209 from the National Science Foundation and grant W911NF-04-1-0339 from the U.S. Army Research Office.

network approximation. In [20], we have shown that the approximation spaces determined by zonal function network approximation are the same as those determined by polynomial approximations.

To illustrate the issues to be discussed in this paper, we consider an example in the case $q = 1$ or, equivalently, the case of 2π -periodic functions on the real line. In this discussion only, let $f(x) = |\cos x|^{1/4}$, $x \in \mathbb{R}$. In Figure 1 (left), we show the log-plot of the absolute errors between f and its (trigonometric) Fourier projection of order 31, where the Fourier coefficients are estimated by a 128 point DFT. In Figure 1 (right), we show a similar log-plot where the Fourier projection is replaced by a suitable summability operator (described more precisely in (3.1)), yielding again a trigonometric polynomial of order 31. It is clear that our summability operator is far more localized than the Fourier projection; i.e., the error in approximation decreases more rapidly as one goes away from the singularities at $\pi/2$ and $3\pi/2$. The maximum error on $[3\pi/4, 5\pi/4]$ is 0.0103 for the projection, and 0.0028 for our operator. Out of the 2048 points considered for the test, the error by the summability operator is less than 10^{-3} at 38.96% points, the corresponding percentage for the projection is only 4.88%. In contrast to free-knot spline approximation, our summability operator is universal; i.e., its construction (convolution with a kernel) does not require any a priori knowledge about the location of singularities of the target function. It yields a single, globally defined trigonometric polynomial, computed using global data. Nevertheless, it is auto-adaptive, in the sense that the error in approximation on different subintervals adjusts itself according to the smoothness of the target function on these subintervals. In [24, 25], we have given a very detailed analysis of the approximation properties of these operators in the case $q = 1$.

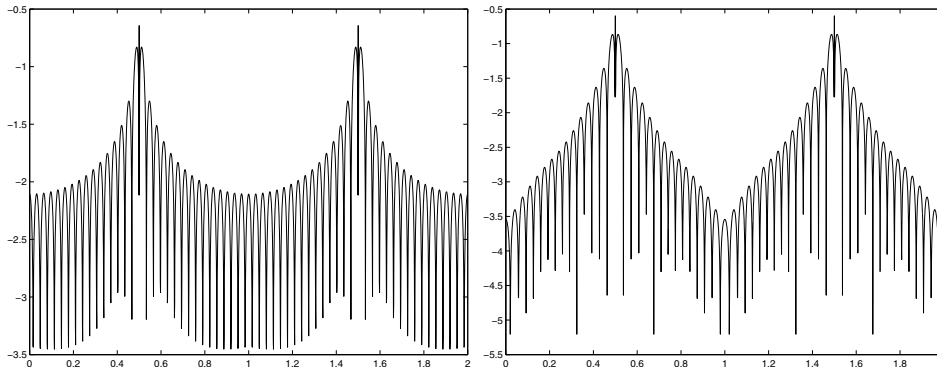


FIG. 1. The log-plot of the absolute error between the function $f(x) = |\cos x|^{1/2}$ and its Fourier projection of order 31 (left). The log-plot of the absolute error between the function f and its trigonometric polynomial approximation obtained by our summability operator (right). The numbers on the x axis are in multiples of π ; the actual absolute errors are 10^y .

Our computation based on a 128 point DFT implies that the values of the function are available at 128 equidistant points. If only scattered data is available, the following method is often used (especially in the context of approximation on the sphere) to estimate the values needed for the DFT. For each point ξ , we consider the nearest point of the form $2\pi k/128$ and imagine that the value of f at this point is $f(\xi)$, taking averages in the case of multiplicities and interpolating in the case of gaps. If we use our summability operator, estimating the Fourier coefficients in this way, then the maximum error on $[3\pi/4, 5\pi/4]$ is 0.0357, and the proportion of points where the

error is less than 10^{-3} is 7.08%. It is clear that a careful construction of quadrature formulas is essential to obtaining good approximation results.

The purpose of this paper is to construct universal, auto-adaptive, localized, linear, polynomial (ℓ^p -valued) operators based on scattered data on \mathbb{S}^q ($q \geq 2$) and to analyze their approximation properties. An essential ingredient in our construction is the construction of quadrature formulas based on scattered data, exact for integrating spherical polynomials of (moderately) high degree, and satisfying certain technical conditions known as the Marcinkiewicz–Zygmund (M–Z) conditions. Our construction is different from the usual construction of quadrature formulas (designs) studied in numerical analysis, where one has a choice of the placement of nodes. In [22, 23], we proved the existence of such quadrature formulas for scattered data. However, previous efforts to compute such formulas did not yield exactness beyond degree 18 polynomials. This was a severe limitation on the practical applications of our theoretical constructions. We will show that a very simple idea of solving a system of equations involving a Gram matrix yields surprisingly good results, in particular, quadrature formulas exact for integrating polynomials of degree as high as 178. Gram matrices are typically ill conditioned. However, we will show both theoretically and numerically that the ones which we use are, in fact, very well conditioned. We will introduce another algorithm of theoretical interest to compute data dependent orthogonal polynomials and use these to compute the quadrature formulas in a memory efficient manner. To the best of our knowledge, this is the first effort to extend the univariate constructions in Gautschi’s book [11] to a multivariate setting. Considering that computation of classical spherical harmonics is a very delicate task, requiring many tricks based on the special function properties of these polynomials for a stable computation, it is not expected that our computation of data dependent orthogonal polynomials with no such special function properties would be stable. In describing this algorithm, we hope to stimulate further research in this interesting direction. We note that even if this algorithm is not as stable for high degrees as the other algorithm, it yields satisfactory quadrature formulas exact for integrating polynomials of degree 32. Most importantly, our newfound ability to compute quadrature formulas for moderately high degrees allows us to offer our operators as a viable, practical method of approximation, even superior to the commonly used methods of least squares and Fourier projection as far as localized approximation is concerned.

An additional problem is when the available values of the target function are noisy. One may assume that the noise is an additive random variable with mean zero. It is also routine in learning theory to assume that the random variables have a bounded range. This assumption is usually satisfied with a high probability even if the random variables do not actually have a bounded range. However, one does not typically know the actual distribution of these random variables. We obtain probabilistic estimates in this setting on the global and local approximations by our operators. To underline the practical utility of our operators, we use them for modeling the MAGSAT data supplied to us by Dr. Thorsten Maier, obtaining results comparable to those obtained by other techniques.

In section 2, we review certain facts about spherical polynomials, the existence of quadrature formulas to integrate these, a few properties of the quadrature weights, and certain polynomial kernels which we will need throughout the paper. In section 3, we study the approximation properties of the linear polynomial operators. The new results here are Theorems 3.1 and 3.2. The first parts of these theorems were proved essentially in [18], but were not stated in the form given here. In order to apply these

operators in practice, one needs quadrature formulas exact for high degree spherical polynomials. Explicit algorithms to construct such formulas are described in section 4. The new results in this section are Theorems 4.1 and 4.2. Numerical results are presented in section 5, and the proofs of all new results are given in section 6.

2. Background. In this section, we review some known results regarding spherical polynomials and localized polynomial kernels.

2.1. Spherical polynomials. Let $q \geq 1$ be an integer, \mathbb{S}^q be the unit sphere embedded in the Euclidean space \mathbb{R}^{q+1} (i.e., $\mathbb{S}^q := \{(x_1, \dots, x_{q+1}) \in \mathbb{R}^{q+1} : \sum_{k=1}^{q+1} x_k^2 = 1\}$), and μ_q be its Lebesgue surface measure, normalized so that $\mu_q(\mathbb{S}^q) = 1$. The surface area of \mathbb{S}^q is $\frac{2\pi^{(q+1)/2}}{\Gamma((q+1)/2)}$. For $\delta > 0$, a spherical cap with radius δ and center $\mathbf{x}_0 \in \mathbb{S}^q$ is defined by

$$\mathbb{S}_\delta^q(\mathbf{x}_0) := \{\mathbf{x} \in \mathbb{S}^q : \arccos(\mathbf{x} \cdot \mathbf{x}_0) \leq \delta\}.$$

If $1 \leq p \leq \infty$, and $f : \mathbb{S}^q \rightarrow \mathbb{R}$ is measurable, we write

$$\|f\|_p := \begin{cases} \left\{ \int_{\mathbb{S}^q} |f(\mathbf{x})|^p d\mu_q(\mathbf{x}) \right\}^{1/p} & \text{if } 1 \leq p < \infty, \\ \text{ess sup}_{\mathbf{x} \in \mathbb{S}^q} |f(\mathbf{x})| & \text{if } p = \infty. \end{cases}$$

The space of all Lebesgue measurable functions on \mathbb{S}^q such that $\|f\|_p < \infty$ will be denoted by L^p , with the usual convention that two functions are considered equal as elements of this space if they are equal almost everywhere. The symbol $C(\mathbb{S}^q)$ denotes the class of all continuous, real-valued functions on \mathbb{S}^q , equipped with the norm $\|\circ\|_\infty$.

For a real number $x \geq 0$, let Π_x^q denote the class of all spherical polynomials of degree at most x . (This is the same as the class Π_n^q , where n is the largest integer not exceeding x . However, our extension of the notation allows us, for example, to use the simpler notation $\Pi_{n/2}^q$ rather than the more cumbersome notation $\Pi_{\lfloor n/2 \rfloor}^q$.) For a fixed integer $\ell \geq 0$, the restriction to \mathbb{S}^q of a homogeneous harmonic polynomial of exact degree ℓ is called a spherical harmonic of degree ℓ . Most of the following information is based on [26], [33, section IV.2], and [5, Chapter XI], although we use a different notation. The class of all spherical harmonics of degree ℓ will be denoted by \mathbf{H}_ℓ^q . The spaces \mathbf{H}_ℓ^q are mutually orthogonal relative to the inner product of L^2 . For any integer $n \geq 0$, we have $\Pi_n^q = \bigoplus_{\ell=0}^n \mathbf{H}_\ell^q$. The dimension of \mathbf{H}_ℓ^q is given by

$$(2.1) \quad d_\ell^q := \dim \mathbf{H}_\ell^q = \begin{cases} \frac{2\ell + q - 1}{\ell + q - 1} \binom{\ell + q - 1}{\ell} & \text{if } \ell \geq 1, \\ 1 & \text{if } \ell = 0 \end{cases}$$

and that of Π_n^q is $\sum_{\ell=0}^n d_\ell^q = d_n^{q+1}$. Furthermore, $L^2 = L^2\text{-closure}\{\bigoplus_{\ell=0}^\infty \mathbf{H}_\ell^q\}$. Hence, if we choose an orthonormal basis $\{Y_{\ell,k} : k = 1, \dots, d_\ell^q\}$ for each \mathbf{H}_ℓ^q , then the set $\{Y_{\ell,k} : \ell = 0, 1, \dots \text{ and } k = 1, \dots, d_\ell^q\}$ is a complete orthonormal basis for L^2 . One has the well-known addition formula [26] and [5, Chapter XI, Theorem 4]:

$$(2.2) \quad \sum_{k=1}^{d_\ell^q} Y_{\ell,k}(\mathbf{x}) Y_{\ell,k}(\zeta) = \frac{2^{q-1} \Gamma(q/2)^2}{\Gamma(q)} p_\ell(1) p_\ell(\mathbf{x} \cdot \zeta), \quad \ell = 0, 1, \dots,$$

where $p_\ell := p_\ell^{(q/2-1, q/2-1)}$ is the orthonormalized Jacobi polynomial with positive leading coefficient

$$\int_{-1}^1 p_\ell(t)p_k(t)(1-t^2)^{q/2-1}dt = \begin{cases} 1 & \text{if } \ell = k, \\ 0 & \text{otherwise.} \end{cases}$$

In particular, for $\mathbf{x} \in \mathbb{S}^q$, $\ell = 0, 1, \dots$,

$$(2.3) \quad \sum_{k=1}^{d_\ell^q} Y_{\ell,k}^2(\mathbf{x}) = \frac{2^{q-1}\Gamma(q/2)^2}{\Gamma(q)} p_\ell(1)^2 = \int_{\mathbb{S}^q} \sum_{k=1}^{d_\ell^q} Y_{\ell,k}^2(\zeta) d\mu_q(\zeta) = d_\ell^q.$$

2.2. Localized polynomial kernels. Let $h : [0, \infty) \rightarrow \mathbb{R}$ be a compactly supported function, and let $t > 0$. We define for $u \in \mathbb{R}$

$$(2.4) \quad \Phi_t(h; u) := \frac{2^{q-1}\Gamma(q/2)^2}{\Gamma(q)} \sum_{\ell=0}^\infty h\left(\frac{\ell}{t}\right) p_\ell(1)p_\ell(u)$$

and define $\Phi_t(h; u) = 0$ if $t \leq 0$.

In what follows, we adopt the following convention regarding constants. The letters c, c_1, \dots will denote generic, positive constants depending only on the dimension q and other fixed quantities in the discussion such as the function h , the different norms involved in the formula, etc. Their value will be different at different occurrences, even within the same formula. The symbol $A \sim B$ will mean $cA \leq B \leq c_1A$.

The following proposition summarizes some of the important properties of the kernels defined in (2.4).

PROPOSITION 2.1. *Let $S \geq q$ be an integer, $h : [0, \infty) \rightarrow \mathbb{R}$ be an S times iterated integral of a function of bounded variation, $h(x) = 1$ for $x \in [0, 1/2]$, $h(x) = 0$ for $x > 1$, and h be nonincreasing. Let $\mathbf{x} \in \mathbb{S}^q$. We have, for every integer $n \geq 0$, $\Phi_n(h; \circ \cdot \mathbf{x}) \in \Pi_n^q$ and*

$$(2.5) \quad \int \Phi_n(h; \mathbf{x} \cdot \zeta) P(\zeta) d\mu_q(\zeta) = P(\mathbf{x}), \quad P \in \Pi_{n/2}.$$

Further,

$$(2.6) \quad \begin{aligned} & \sup_{n \geq 1, \zeta \in \mathbb{S}^q} \int |\Phi_n(h; \zeta \cdot \xi)| d\mu_q(\xi) = \sup_{n \geq 1} \int |\Phi_n(h; \mathbf{x} \cdot \xi)| d\mu_q(\xi) \\ & = \frac{2\pi^{q/2}}{\Gamma(q/2)} \sup_{n \geq 1} \int_{-1}^1 |\Phi_n(h; u)| (1-u^2)^{q/2-1} du < \infty, \\ & \int |\Phi_n(h; \mathbf{x} \cdot \xi)|^2 d\mu_q(\xi) = \frac{2\pi^{q/2}}{\Gamma(q/2)} \int_{-1}^1 |\Phi_n(h; u)|^2 (1-u^2)^{q/2-1} du \\ (2.7) \quad & \sim n^q \sim \max_{\xi \in \mathbb{S}^q} |\Phi_n(h; \mathbf{x} \cdot \xi)| = |\Phi_n(h; 1)|, \end{aligned}$$

and for every $\xi \in \mathbb{S}^q$, $\xi \neq \mathbf{x}$,

$$(2.8) \quad |\Phi_n(h; \mathbf{x} \cdot \xi)| \leq cn^q \begin{cases} (n\sqrt{1-\mathbf{x} \cdot \xi})^{1/2-q/2-S} & \text{if } 0 \leq \mathbf{x} \cdot \xi < 1, \\ n^{-S} & \text{if } -1 \leq \mathbf{x} \cdot \xi < 0. \end{cases}$$

Except for (2.7), all parts of Proposition 2.1 have been proved and verified repeatedly in [17, 18, 12, 19]. We will sketch a proof of this proposition, mainly to reconcile notation.

Proof of Proposition 2.1. Equation (2.5) and the first two equations in (2.6) are clear. The last estimate in (2.6) follows from [17, Lemma 4.6] with the following choice of the parameters there: $\alpha = \beta = q/2 - 1$, $h_\nu = h(\nu/n)$, where we observe that by a repeated application of the mean value theorem,

$$\sum_{\nu=0}^{\infty} (\nu + 1)^s |\Delta^r h(\nu/n)| \leq cn^{s-r+1}, \quad s \in \mathbb{R}, r, n = 1, 2, \dots,$$

where Δ^r is the r th order forward difference applied with respect to ν . Similarly, the estimate (2.8) follows from [17, Lemma 4.10] with the same parameters as above, S in place of K in [17], and $y = \mathbf{x} \cdot \xi$ (cf. the appendix in [12]). We prove (2.7). The first equation is a consequence of the rotation invariance of μ_q . In view of the addition formula (2.2),

$$(2.9) \quad \Phi_n(h; \mathbf{x} \cdot \xi) = \sum_{\ell=0}^n h \left(\frac{\ell}{n} \right) \sum_{k=1}^{d_\ell^q} Y_{\ell,k}(\mathbf{x}) Y_{\ell,k}(\xi).$$

It follows, using (2.3) and the facts that $h(\ell/n) = 1$ for $\ell \leq n/2$ and $0 \leq h(t) \leq 1$ for $t \in [0, \infty)$, that

$$\int \Phi_n(h; \mathbf{x} \cdot \xi)^2 d\mu_q(\xi) = \sum_{\ell=0}^n h \left(\frac{\ell}{n} \right)^2 \sum_{k=1}^{d_\ell^q} Y_{\ell,k}(\mathbf{x})^2 = \sum_{\ell=0}^n h \left(\frac{\ell}{n} \right)^2 d_\ell^q \sim n^q.$$

Similarly, using the Schwarz inequality, (2.2), (2.3), and the fact that $h(\ell/n) \geq 0$,

$$\begin{aligned} \Phi_n(h; 1) &= |\Phi_n(h; \mathbf{x} \cdot \mathbf{x})| \leq \sup_{\xi \in \mathbb{S}^q} |\Phi_n(h; \mathbf{x} \cdot \xi)| \\ &\leq \sum_{\ell=0}^n h \left(\frac{\ell}{n} \right) \left\{ \sum_{k=1}^{d_\ell^q} Y_{\ell,k}(\mathbf{x})^2 \right\}^{1/2} \left\{ \sum_{k=1}^{d_\ell^q} Y_{\ell,k}(\xi)^2 \right\}^{1/2} = \sum_{\ell=0}^n h \left(\frac{\ell}{n} \right) d_\ell^q = \Phi_n(h; 1). \end{aligned}$$

Since $d_\ell^q \sim \ell^{q-1}$, $0 \leq h(\ell/n) \leq 1$, and $h(\ell/n) = 1$ for $\ell \leq n/2$, the above two estimates lead to (2.7). \square

In the remainder of this paper, h will denote a fixed function satisfying the conditions of Proposition 2.1.

2.3. Quadrature formulas. Let \mathcal{C} be a finite set of distinct points on \mathbb{S}^q . A quadrature formula based on \mathcal{C} has the form $\mathcal{Q}(f) = \sum_{\xi \in \mathcal{C}} w_\xi f(\xi)$, where w_ξ , $\xi \in \mathcal{C}$, are real numbers. For integer $n \geq 0$, the formula is exact for degree n if $\mathcal{Q}(P) = \int_{\mathbb{S}^q} P d\mu_q$ for all $P \in \Pi_n^q$. It is not difficult to verify that if $\mathcal{Q}_n(f) = \sum w_{\xi_n} f(\xi_n)$ is a sequence of quadrature formulas, with \mathcal{Q}_n being exact with degree n , then $\mathcal{Q}_n(f) \rightarrow \int f d\mu_q$ for every continuous function f on \mathbb{S}^q if and only if $\sum |w_{\xi_n}| \leq c$, with c being independent of n . In what follows, we will assume tacitly that \mathcal{C} is one of the members of a nested sequence of finite subsets of \mathbb{S}^q , whose union is dense in \mathbb{S}^q . All the constants may depend upon the whole sequence, but not on any individual member of this sequence. Thus, a formula \mathcal{Q} will be called a bounded variation

formula if $\sum_{\xi \in \mathcal{C}} |w_\xi| \leq c$, with the understanding that this is an abbreviation for the concept described above with a sequence of quadrature formulas.

DEFINITION 2.1. *Let $m \geq 0$ be an integer. The set \mathcal{C} admits an M-Z quadrature of order m if there exist weights w_ξ such that*

$$(2.10) \quad \int_{\mathbb{S}^q} P(\mathbf{x}) d\mu_q(\mathbf{x}) = \sum_{\xi \in \mathcal{C}} w_\xi P(\xi), \quad P \in \Pi_{2m}^q,$$

and

$$(2.11) \quad \left(\sum_{\xi \in \mathcal{C}} |w_\xi| |P(\xi)|^p \right)^{1/p} \leq c \|P\|_p, \quad P \in \Pi_{2m}^q, \quad 1 \leq p < \infty.$$

The weights w_ξ will be called M-Z weights of order m . The condition (2.11) will be referred to as the M-Z condition.

If \mathcal{C} admits an M-Z quadrature of order m , and $\{w_\xi\}$ are the weights involved, it is clear from using (2.11) with the polynomial identically equal to 1 in place of P that $\sum_{\xi \in \mathcal{C}} |w_\xi| \leq c$. Further, if $\zeta \in \mathcal{C}$, then applying (2.11) with $p = 2$ and $\Phi_m(h; \zeta \cdot \circ)$ in place of P , we obtain for M-Z weights of order m :

$$|w_\zeta| \Phi_m(h; 1)^2 \leq \sum_{\xi \in \mathcal{C}} |w_\xi| \Phi_m(h; \zeta \cdot \xi)^2 \leq c \int \Phi_m(h; \zeta \cdot \mathbf{x})^2 d\mu_q(\mathbf{x}).$$

The estimate (2.7) now implies that for all M-Z weights $\{w_\xi\}$ of order m ,

$$(2.12) \quad |w_\xi| \leq cm^{-q}, \quad \xi \in \mathcal{C}.$$

In [22, 23], we proved that every finite set $\mathcal{C} \subset \mathbb{S}^q$ admits an M-Z quadrature with an order depending upon how dense the set \mathcal{C} is. This density is measured in terms of the mesh norm. The mesh norm of \mathcal{C} with respect to a subset $K \subseteq \mathbb{S}^q$ is defined to be

$$(2.13) \quad \delta_{\mathcal{C}}(K) := \sup_{\mathbf{x} \in K} \text{dist}(\mathbf{x}, \mathcal{C}).$$

The following theorem summarizes the quadrature formula given in [22, 23].

THEOREM 2.1. *There exists a constant α_q with the following property. Let \mathcal{C} be a finite set of distinct points on \mathbb{S}^q , and let m be an integer with $m \leq \alpha_q (\delta_{\mathcal{C}}(\mathbb{S}^q))^{-1}$. Then \mathcal{C} admits an M-Z quadrature of order m , and the set $\{w_\xi\}$ of M-Z weights may be chosen to satisfy*

$$(2.14) \quad |\{\xi : w_\xi \neq 0\}| \sim m^q \sim \dim(\Pi_{2m}^q).$$

3. Polynomial operators. For $t > 0$, $f \in L^1$, we define the summability operator σ_t^* by the formula

$$(3.1) \quad \begin{aligned} \sigma_t^*(h; f, \mathbf{x}) &= \int_{\mathbb{S}^q} f(\zeta) \Phi_t(h; \mathbf{x} \cdot \zeta) d\mu_q(\zeta) \\ &= \sum_{\ell=0}^{\infty} h \left(\frac{\ell}{t} \right) \sum_{k=1}^{d_\ell^q} \hat{f}(\ell, k) Y_{\ell, k}(\mathbf{x}), \quad \mathbf{x} \in \mathbb{S}^q. \end{aligned}$$

(It is convenient, and customary in approximation theory, to use the notation $\sigma_t^*(h; f, \mathbf{x})$ rather than $\sigma_t^*(h; f)(\mathbf{x})$.) Although we defined the operator for L^1 to underline the fact that it is a universal operator, we will be interested only in its restriction to $C(\mathbb{S}^q)$. If $f : \mathbb{S}^q \rightarrow \mathbb{R}$ is a continuous function, the degree of approximation of f from Π_x^q is defined by

$$E_x(f) = \inf_{P \in \Pi_x^q} \|f - P\|_\infty.$$

It is well known [16, 18] that, for all integers $n \geq 1$ and $f \in C(\mathbb{S}^q)$,

$$(3.2) \quad E_n(f) \leq \|f - \sigma_n^*(h; f)\|_\infty \leq cE_{n/2}(f).$$

Following [18], we now define a discretized version of these operators.

If $\mathcal{C} \subset \mathbb{S}^q$ is a finite set and $\mathbf{W} = \{w_\xi\}_{\xi \in \mathcal{C}}$ and $\mathbf{Z} = \{z_\xi\}_{\xi \in \mathcal{C}}$ are sets of real numbers, we define the polynomial operator

$$(3.3) \quad \sigma_t(\mathcal{C}, \mathbf{W}; h; \mathbf{Z}, \mathbf{x}) := \sum_{\xi \in \mathcal{C}} w_\xi z_\xi \Phi_t(h; \mathbf{x} \cdot \xi), \quad t \in \mathbb{R}, \mathbf{x} \in \mathbb{S}^q.$$

If $f : \mathbb{S}^q \rightarrow \mathbb{R}$ and $z_\xi = f(\xi)$, $\xi \in \mathcal{C}$, we will write $\sigma_t(\mathcal{C}, \mathbf{W}; h; f, \mathbf{x})$ in place of $\sigma_t(\mathcal{C}, \mathbf{W}; h; \mathbf{Z}, \mathbf{x})$. In [18], we denoted these operators by $\sigma_t(\nu; h, f)$, where ν is the measure that associates the mass w_ξ with $\xi \in \mathcal{C}$. In this paper, we prefer to use the slightly expanded notation as in (3.3). If $n \geq 1$ is an integer, $\mathcal{C} \subset \mathbb{S}^q$ is a finite set that admits an M–Z quadrature of order n , and \mathbf{W} is the set of the corresponding M–Z weights, then it is shown in [18, Proposition 4.1] that

$$(3.4) \quad E_n(f) \leq \|f - \sigma_n(\mathcal{C}, \mathbf{W}; h; f)\|_\infty \leq cE_{n/2}(f), \quad f \in C(\mathbb{S}^q).$$

In this paper, we will be especially interested in the approximation of functions in the class \mathbb{W}_r , $r > 0$, composed of functions $f \in C(\mathbb{S}^q)$ for which $E_n(f) = \mathcal{O}(n^{-r})$, $n \geq 1$. A complete characterization of the classes \mathbb{W}_r in terms of such constructive properties of its members as the number of partial derivatives and their moduli of smoothness is well known [27, 16]. In view of (3.2), $f \in \mathbb{W}_r$ if and only if

$$\|f\|_{\mathbb{W}_r} := \|f\|_\infty + \sup_{n \geq 1} 2^{nr} \|\sigma_{2n}^*(h; f) - \sigma_{2n-1}^*(h; f)\|_\infty < \infty.$$

In practical applications, the data is contaminated with noise. Therefore, we wish to examine the behavior of our operators based on data of the form $\{(\xi, f(\xi) + \epsilon_\xi)\}$, where ϵ_ξ are independent random variables with unknown probability distributions, each with mean 0. If the range of these random variables is not bounded, one can still assume that the probability of the variables going out of a sufficiently large interval is small. Hence, it is customary in learning theory to assume that the variables ϵ_ξ have a bounded range, so that one may use certain technical inequalities of probability theory known as Bennett’s inequalities; see the proof of Lemma 6.2 below.

In the statements of the theorems below, we use three parameters. The symbol M denotes the number of points in the data set; we assume that the set admits an M–Z quadrature of order m , and the degree n of the polynomial approximant $\sigma_n(\mathcal{C}, \mathbf{W}; h; f)$ is determined in terms of m . For theoretical considerations where one is not concerned about the actual numerical constructions of the quadrature weights, one can imagine a data set \mathcal{C} with $M := |\mathcal{C}| \sim \delta_{\mathcal{C}}(\mathbb{S}^q)^{-q}$ and assume that the weights \mathbf{W} are as guaranteed by Theorem 2.1. If so, then we can take $M \sim m^q \sim \delta_{\mathcal{C}}(\mathbb{S}^q)^{-q}$ in

the discussion in this section. For example, the estimates (3.7) and (3.8) below can then be expressed in terms of the number of samples, respectively, as follows:

$$(3.5) \quad \|f - \sigma_n(\mathcal{C}, \mathbf{W}; h; f)\|_\infty \leq cM^{-r/q}, \quad \text{with some } n \sim M^{1/q},$$

and

$$(3.6) \quad \text{Prob} \left(\|\sigma_n(\mathcal{C}, \mathbf{W}; h; \mathbf{Z}) - f\|_\infty \geq c_1 \frac{(\log M)^c}{M^{r/(q+2r)}} \right) \leq c_2 M^{-c},$$

with some $n \sim (M/\log M)^{1/(2r+q)}$.

THEOREM 3.1. *Suppose that $m \geq 1$ is an integer and $\mathcal{C} = \{\xi_j\}_{j=1}^M$ admits an M -Z quadrature of order m , and let \mathbf{W} be the corresponding quadrature weights. Let $r > 0$, $f \in \mathbb{W}_r$, $\|f\|_{\mathbb{W}_r} = 1$.*

(a) *For integer $n \leq m$, we have*

$$(3.7) \quad \|f - \sigma_n(\mathcal{C}, \mathbf{W}; h; f)\|_\infty \leq cn^{-r}.$$

(b) *For $j = 1, \dots, M$, let ϵ_j be independent random variables with mean 0 and range $[-1, 1]$, and let $\mathbf{Z} = \{\epsilon_j + f(\xi_j)\}$. If $A > 0$ and $n \geq 1$ is the greatest integer with $(A+q)n^{2r+q} \log n \leq c_3 m^q$, $n \leq m$, then*

$$(3.8) \quad \text{Prob} \left(\|\sigma_n(\mathcal{C}, \mathbf{W}; h; \mathbf{Z}) - f\|_\infty \geq c_1 n^{-r} \right) \leq c_2 n^{-A}.$$

Here, the constants c_1, c_2, c_3 are independent of the distribution of the variables ϵ_j .

We now turn our attention to local approximation by our operators. In what follows, if $K \subseteq \mathbb{S}^q$, $f: K \rightarrow \mathbb{R}$, then $\|f\|_{\infty, K} := \sup_{\mathbf{x} \in K} |f(\mathbf{x})|$. If $\mathbf{x}_0 \in \mathbb{S}^q$, a function f is defined to be r -smooth at \mathbf{x}_0 if there is a spherical cap $\mathbb{S}_\delta^q(\mathbf{x}_0)$ such that $f\phi \in \mathbb{W}_r$ for every infinitely differentiable function ϕ supported on $\mathbb{S}_\delta^q(\mathbf{x}_0)$. We have proved in [18, Theorem 3.3] that f is r -smooth at a point \mathbf{x}_0 if and only if there is a cap $\mathbb{S}_\delta^q(\mathbf{x}_0)$ such that

$$\|\sigma_{2^n}^*(h; f) - \sigma_{2^{n-1}}^*(h; f)\|_{\infty, \mathbb{S}_\delta^q(\mathbf{x}_0)} = \mathcal{O}(2^{-nr}).$$

Accordingly, if K is a spherical cap, we may define the class $\mathbb{W}_r(K)$ as consisting of $f \in C(\mathbb{S}^q)$, for which

$$\|f\|_{\mathbb{W}_r(K)} := \|f\|_\infty + \sup_{n \geq 1} 2^{nr} \|\sigma_{2^n}^*(h; f) - \sigma_{2^{n-1}}^*(h; f)\|_{\infty, K} < \infty.$$

THEOREM 3.2. *Suppose that $m \geq 1$ is an integer and $\mathcal{C} = \{\xi_j\}_{j=1}^M$ admits an M -Z quadrature of order m , and let \mathbf{W} be the corresponding quadrature weights. Let $0 < r \leq S - q$, $K' \subset K$ be concentric spherical caps, $f \in C(\mathbb{S}^q)$, and $\|f\|_{\mathbb{W}_r(K)} = 1$.*

(a) *For integer n , $1 \leq n \leq m$,*

$$(3.9) \quad \|f - \sigma_n(\mathcal{C}, \mathbf{W}; h; f)\|_{\infty, K'} \leq cn^{-r}.$$

(b) *For $j = 1, \dots, M$, let ϵ_j be independent random variables with mean 0 and range contained in $[-1, 1]$, and let $\mathbf{Z} = \{\epsilon_j + f(\xi_j)\}$. If $A > 0$ and $n \geq 1$ is the greatest integer with $(A+q)n^{2r+q} \log n \leq c_3 m^q$, $n \leq m$, then*

$$(3.10) \quad \text{Prob} \left(\|\sigma_n(\mathcal{C}, \mathbf{W}; h; \mathbf{Z}) - f\|_{\infty, K'} \geq c_1 n^{-r} \right) \leq c_2 n^{-A}.$$

Here, the constants c_1, c_2, c_3 are independent of the distribution of the variables ϵ_j .

4. Construction of quadrature formulas. In this section, we describe two algorithms to obtain bounded variation quadrature formulas associated with a given finite set of points $\mathcal{C} \subset \mathbb{S}^q$. Both of these constructions can be described in a very general setting. Since this also simplifies the notation and ideas considerably by avoiding the use of real and imaginary parts of a doubly indexed polynomial $Y_{\ell,k}$, we will describe the algorithms in this generality.

Let Ω be a nonempty set; μ be a probability measure on Ω ; $\mathcal{C} \subset \Omega$, y_1, y_2, \dots , be a complete orthonormal basis for $L^2(\Omega, \mu)$, where $y_1 \equiv 1$; and V_k denote the span of y_1, \dots, y_k . Let ν be another measure on Ω , and $\langle \circ, \circ \rangle$ denote the inner product of $L^2(\Omega, \nu)$. For an integer $N \geq 1$, the Gram matrix G_N is an $N \times N$ matrix, defined by $(G_N)_{\ell,k} = \langle y_\ell, y_k \rangle = (G_N)_{k,\ell}$, $1 \leq k, \ell \leq N$. We wish to find a weight function W on Ω such that $\int_\Omega P d\mu = \int_\Omega PW d\nu$ for all $P \in V_N$ for an integer N for which G_N is positive definite.

For the applications to the case of quadrature formulas for the sphere, $\Omega = \mathbb{S}^q$, $\mu = \mu_q$, and y_k 's are the orthogonal spherical harmonics, arranged in a sequence, so that $y_1 \equiv 1$, and all polynomials of lower degree are listed before those of a higher degree. To include all polynomials in Π_n^q , we need $N = d_n^{q+1}$. There are many possibilities for defining the measure ν . The simplest is the measure ν^{MC} that associates the mass $1/|\mathcal{C}|$ with each point of \mathcal{C} . A more sophisticated way to define the measure ν is the following. We obtain a partition of \mathbb{S}^q into a dyadic triangulation such that each triangle contains at least one point of \mathcal{C} . We choose only one point in each triangle and, hence, assume that each triangle contains exactly one point of \mathcal{C} . We define the measure ν^{TR} to be the measure that associates with each $\xi \in \mathcal{C}$ the area of the triangle containing ξ .

One of the simplest ideas for computing the quadrature weights is the following. Let N be an integer for which G_N is positive definite. If $P = \sum_{j=1}^N a_j y_j$, then $\int_\Omega P d\mu = a_1$. Also, the vector $\mathbf{a} = (a_1, \dots, a_N)^T$ satisfies the matrix equation

$$G_N \mathbf{a} = (\langle P, y_1 \rangle, \dots, \langle P, y_N \rangle)^T,$$

so that

$$(4.1) \quad \int_\Omega P d\mu = \sum_{k=1}^N (G_N)_{1,k}^{-1} \langle P, y_k \rangle = \left\langle P, \sum_k (G_N)_{1,k}^{-1} y_k \right\rangle.$$

In the setting of the sphere, this gives the following quadrature formula:

$$(4.2) \quad \int_{\mathbb{S}^q} P d\mu_q = \sum_{\xi \in \mathcal{C}} P(\xi) \left\{ \nu(\{\xi\}) \sum_{k=1}^N (G_N)_{1,k}^{-1} y_k(\xi) \right\} =: \sum_{\xi \in \mathcal{C}} w_\xi^{LSQ} P(\xi).$$

We formulate this as the following algorithm.

ALGORITHM LSQ.

Input: The matrix $Y = (y_k(\xi))$, $k = 1, \dots, N$ (optional), and the vector $\mathbf{v} = (\nu(\{\xi\}))$.

1. Solve $Y \text{diag}(\mathbf{v}) Y^T \mathbf{b} = (1, 0, \dots, 0)^T$.
2. Return $w_\xi^{LSQ} = \nu(\{\xi\}) \sum_{k=1}^N b_k y_k(\xi)$.

We observe that $G_N = Y \text{diag}(\mathbf{v}) Y^T$. It is clear that the matrix G_N is always positive semidefinite; the assumption that it is positive definite is equivalent to the assumption that no nonzero element of V_N vanishes identically on \mathcal{C} . If \mathcal{C} and $\{\nu(\{\xi\})\}$

satisfy the M–Z inequalities, Theorem 4.1 shows that G_N is well conditioned. Assuming that the matrix Y is input, the time to compute G_N is $\mathcal{O}(N^2|\mathcal{C}|)$, and the space requirement is $\mathcal{O}(N^2)$. (In the case of the sphere \mathbb{S}^q , we need $N = d_n^{q+1} = \mathcal{O}(n^q)$ to compute formulas exact for degree n .) The vector \mathbf{b} in step 1 can be found using such iterative methods as the conjugate residual method. We refer the reader to [6] for a more detailed analysis of this method. Using this approach, the matrix Y and G_N need not be stored or precomputed, but the product of the matrix G_N with an arbitrary residual vector \mathbf{r} needs to be computed. This observation results in a substantial savings in the time and memory complexity of the algorithm when the results are desired only within a given accuracy. For the unit sphere \mathbb{S}^2 , when $N = (n + 1)^2$, the product $G_N \mathbf{r} = Y \text{diag}(\mathbf{v}) Y^T \mathbf{r}$ can be computed within an accuracy ϵ using a recent algorithm of Keiner [14] using $\mathcal{O}(n^2(\log n)^2 + \log(1/\epsilon)|\mathcal{C}|)$ operations, where ϵ is the accuracy of the method.

One way to interpret this algorithm is the following. Let $f : \mathbb{S}^q \rightarrow \mathbb{R}$, and P be the solution to the least squares problem

$$P = \arg \min \{ \langle f - Q, f - Q \rangle : Q \in V_N \}.$$

If \mathbf{f} is the vector $(\langle f, y_j \rangle)$, then $P = \sum_j (G_N^{-1} \mathbf{f})_j y_j$. The quadrature formula with weights w_ξ^{LSQ} thus offers $\int_{\mathbb{S}^q} P d\mu_q$ as the approximation to $\int_{\mathbb{S}^q} f d\mu_q$. The weights w_ξ^{LSQ} also satisfy a least squares property among all the possible quadrature formulas, as shown in Lemma 6.1(a). We summarize some of the properties of the weights w_ξ^{LSQ} in the following theorem.

THEOREM 4.1. *Let $n \geq 1$ be an integer, $N = d_n^{q+1}$, \mathcal{C} be a finite set of points on \mathbb{S}^q , and ν be a measure supported on \mathcal{C} . Let $v_\xi := \nu(\{\xi\})$, $\xi \in \mathcal{C}$, and*

$$(4.3) \quad c_1 \|P\|_p \leq \left\{ \sum_{\xi \in \mathcal{C}} v_\xi |P(\xi)|^p \right\}^{1/p} \leq c_2 \|P\|_p, \quad P \in \Pi_n^q, \quad 1 \leq p \leq \infty.$$

(a) *For the Gram matrix G_N , the lowest eigenvalue is $\geq c_1^2$, and the largest eigenvalue is $\leq c_2^2$, where c_1, c_2 are the constants in (4.3) with $p = 2$. In particular, G_N is positive definite. Moreover, $\sum_{\xi \in \mathcal{C}} |w_\xi^{LSQ}| \leq c$.*

(b) *If*

$$(4.4) \quad \left| \int P^2 d\nu - \int P^2 d\mu_q \right| \leq \frac{c}{n^q} \int P^2 d\mu_q, \quad P \in \Pi_n^q,$$

then $|w_\xi^{LSQ}| \leq cv_\xi$, $\xi \in \mathcal{C}$. In particular, the weights $\{w_\xi^{LSQ}\}$ satisfy the M–Z condition.

(c) *Let $M \geq 1$ be an integer, \mathcal{C} be an independent random sample of M points chosen from the distribution μ_q , and $A, \eta > 0$. Let $v_\xi = 1/M$, $\xi \in \mathcal{C}$. There exists a constant $c = c(A)$ such that if $n \geq 2$ is an integer with $M \geq cn^q \log n / \eta^2$, then*

$$(4.5) \quad \text{Prob} \left(\left| \int P^2 d\nu - \int P^2 d\mu_q \right| \geq \eta \int P^2 d\mu_q, \quad P \in \Pi_n^q \right) \leq c_1 n^{-A}.$$

In particular, if $M \geq cn^{3q} \log n$, then condition (4.4) is satisfied with probability exceeding $1 - c_1 n^{-A}$.

One disadvantage of Algorithm LSQ is that one needs to know the value of N in advance. We now describe an idea which has the potential to avoid this problem. In

the case when Ω is a subset of a Euclidean space and the y_j 's are polynomials, with y_1 denoting the constant polynomial, one can construct a system $\{t_k\}$ of orthonormalized polynomials with respect to ν using recurrence relations. Recurrence relations for orthogonal polynomials in several variables have been discussed in detail by Dunkl and Xu [4, Chapter 3]. In contrast to the viewpoint in [4], we may depend upon a specific enumeration but require the recurrence relation to have a specific form described in Theorem 4.2 below. This form allows us to generalize the ideas in Gautschi's book [11, Chapter 2] in our context.

To describe our ideas in general, let $\Omega \subset \mathbb{R}^{q+1}$, u_1, u_2, \dots , be an enumeration of the monomials in $q + 1$ variables, so that u_1 is the monomial identically equal to 1, the restrictions of u_k 's to Ω are linearly independent, all lower degree polynomials are listed before the higher degree ones, and $V_k = \text{span} \{u_1, \dots, u_k\}$. It is not difficult to see that, for every integer $k \geq 1$, there is a minimal index $p(k)$ such that there exists a monomial \tilde{f}_k of degree 1 with

$$(4.6) \quad \tilde{f}_k u_{p(k)} = u_{k+1}, \quad k = 1, 2, \dots$$

We now let, for each $k = 1, 2, \dots$, $\{y_1, \dots, y_k\}$ be a basis for V_k orthonormal with respect to μ , $N \geq 1$ be an integer for which the Gram matrix G_N is positive definite, and, for each $k = 1, \dots, N$, $\{t_1, \dots, t_k\}$ be a basis for V_k orthonormal with respect to ν . Clearly, any polynomial $P \in V_N$ can be written in the form

$$P(\mathbf{x}) = \int P(\zeta) \sum_k t_k(\mathbf{x}) t_k(\zeta) d\nu(\zeta),$$

and, consequently, one gets the "quadrature formula"

$$(4.7) \quad \int P(\mathbf{x}) d\mu(\mathbf{x}) = \int P(\zeta) \left\{ \sum_k \left(\int t_k(\mathbf{x}) d\mu(\mathbf{x}) \right) t_k(\zeta) \right\} d\nu(\zeta).$$

In this discussion only, let $t_k =: \sum_j c_{k,j} y_j$, and let the matrix $(c_{k,j})$ be denoted by C . The condition that t_1, \dots, t_N is an orthonormal system with respect to ν is equivalent to the condition that $CG_N C^T = I$, where I is the $N \times N$ identity matrix. Hence, $G_N^{-1} = C^T C$. Moreover, $\int t_k d\mu = c_{k,1}$ for $k = 1, \dots, N$, and, hence, we conclude that

$$\sum_k \left(\int t_k(\mathbf{x}) d\mu(\mathbf{x}) \right) t_k = \sum_j \sum_k c_{k,1} c_{k,j} y_j = \sum_j (G_N^{-1})_{1,j} y_j.$$

Thus, the quadrature weights in (4.7) are the same as those in (4.1).

First, we summarize the various recurrence relations in Theorem 4.2, although we will not use all of them. We will denote the (total) degree of u_k by D_k , and observe that D_k is also the degree of y_k and t_k , $D_j \leq j$, and $D_{p(k)} = D_{k+1} - 1$.

THEOREM 4.2. *There exist real numbers $s_{k,j}$, $\tilde{r}_{k,j}$, $A_k \geq 0$, and a linear polynomial f_k such that*

$$(4.8) \quad f_k y_{p(k)} = y_{k+1} - \sum_{\substack{D_{k+1}-2 \leq D_j \leq D_k \\ j \leq k}} \tilde{r}_{k,j} y_j, \quad f_k t_{p(k)} = A_k t_{k+1} - \sum_{\substack{D_{k+1}-2 \leq D_j \leq D_k \\ j \leq k}} s_{k,j} t_j.$$

More generally, if P is any linear polynomial, there exist real numbers $r_{k,j}(P)$ such that

$$(4.9) \quad P y_k = \sum_{D_k-1 \leq D_j \leq D_k+1} r_{k,j}(P) y_j.$$

We have $t_k = \sum_j c_{k,j} y_j$, where

$$(4.10) \quad A_k C_{k+1, \ell} = \left\{ \sum_{D_\ell - 1 \leq D_m \leq D_{\ell+1}} r_{\ell, m}(f_k) c_{p(k), m} + \sum_{D_{k+1} - 2 \leq D_j \leq D_k} s_{k, j} c_{j, \ell} \right\}.$$

In the context of the sphere \mathbb{S}^q , we will compute t_k 's using (4.8) and compute $\int t_k d\mu_q$ using a known quadrature formula. The resulting algorithm, Algorithm REC, in the context of the sphere is summarized below. This algorithm is similar to the Stieltjes method in Gautschi's book [11, section 2.2]. Even though it is feasible to carry out the algorithm for as large an N as the data allow and to find this value of N during run time, it is still desirable from the point of view of numerical stability to limit the largest N from the outset. Accordingly, in describing the following algorithm, we stipulate that the quadrature formula is to be computed to be exact only for polynomials in V_N for the largest possible $N \leq L$ for some integer $L \geq 1$. We assume further that we know another quadrature formula (for example, the Driscoll–Healy formula [3]) exact for polynomials in V_L :

$$(4.11) \quad \sum_{\zeta \in \mathcal{C}^*} \lambda_\zeta P(\zeta) = \int P d\mu_q, \quad P \in V_L.$$

ALGORITHM REC.

Input: An integer L ; the sequence $p(k)$, $k = 1, \dots, L$; sets \mathcal{C} , \mathcal{C}^* ; weights $(\lambda_\zeta)_{\zeta \in \mathcal{C}^*}$ so that (4.11) holds; the values $\{y_j(\xi)\}_{\xi \in \mathcal{C}}$, $\{y_j(\zeta)\}_{\zeta \in \mathcal{C}^*}$ for $j = 1, 2, 3, 4$; and the values $f_k(\xi)$, $f_k(\zeta)$, $k = 1, \dots, L$.

1. Using the Gram–Schmidt procedure, initialize t_1, t_2, t_3, t_4 , for points in both \mathcal{C} and \mathcal{C}^* , and initialize $N = 4$.
2. For $k = 1, \dots, 4$, let $\gamma_k = \sum_{\zeta \in \mathcal{C}^*} \lambda_\zeta t_k(\zeta)$.
3. For each $\xi \in \mathcal{C}$, initialize $w_\xi = \sum_{k=1}^4 \gamma_k t_k(\xi)$.
4. For $k = 4, 5, \dots$ (so that the degrees are at least 0 for all polynomials entering in the recursions) and while $N \leq L$, repeat steps 5–8 below.
5. For j with $D_{k+1} - 2 \leq D_j \leq D_k$, set

$$s_{k, j} = \langle f_k t_{p(k)}, t_j \rangle.$$

6. Define T_{k+1} by

$$T_{k+1} = f_k t_{p(k)} - \sum_{D_{k+1} - 2 \leq D_j \leq D_k} s_{k, j} t_j$$

for points in both \mathcal{C} and \mathcal{C}^* . If $I_{k+1} = \langle T_{k+1}, T_{k+1} \rangle = 0$, then stop and set $N = k$. Otherwise, define $t_{k+1} = T_{k+1} / I_{k+1}^{1/2}$.

7. Set $\gamma_{k+1} = \sum_{\zeta \in \mathcal{C}^*} \lambda_\zeta t_{k+1}(\zeta)$.
8. For each $\xi \in \mathcal{C}$, $w_\xi = w_\xi + \gamma_{k+1} t_{k+1}(\xi)$, $k = k + 1$, $N = N + 1$.

In the case of the sphere \mathbb{S}^q , we take $L = d_{\tilde{n}}^{q+1}$ for some integer $\tilde{n} \geq 1$. The number of j 's with $D_{k+1} - 2 \leq D_j \leq D_k$, $1 \leq j, k \leq L$, is $\mathcal{O}(\tilde{n}^{q-1})$. In this discussion only, let $M = |\mathcal{C}| + |\mathcal{C}^*|$. Consequently, steps 5 and 6 require $\mathcal{O}(M\tilde{n}^{q-1})$ operations. Since the remaining two steps in the loop take $\mathcal{O}(M)$ operations, the loop starting at step 4 requires $\mathcal{O}(M\tilde{n}^{2q-1})$ operations. Finally, we observe that in implementing the above algorithm, one need not keep the whole matrix $t_k(\xi)$; only the rows corresponding to three degrees are required in any step. In particular, the memory requirement of this algorithm is $\mathcal{O}(M\tilde{n}^{q-1})$.

5. Numerical experiments. The objective of this section is to demonstrate and supplement the theoretical results presented in sections 3 and 4.

Our first set of experiments illustrates Algorithms LSQ and REC. The experiments were conducted over a long period of time, many of them long before we started to write the paper. Hence, the normalizations for the spherical polynomials $Y_{\ell,k}$ in Tables 1 and 2 are somewhat different from those in the rest of the paper. This is reflected in the sum of the absolute values of the weights, but has no effect on the various results other than scaling.

First, we report on Algorithm LSQ. Each of the experiments in this case was repeated 30 times with data sets chosen randomly from the distribution μ_2 on \mathbb{S}^2 . To test our algorithms, we computed the computed Gram matrix G^{COM} given by

$$G_{\ell,m}^{COM} = \sum_{\xi \in \mathcal{C}} w_{\xi}^{LSQ} y_{\ell}(\xi) y_m(\xi), \quad \ell, m < \lfloor n/2 \rfloor.$$

The average maximum matrix norm of the difference between G^{COM} and the identity matrix of the same size indicates the error of the quadrature formulas. The results are shown in Table 1. Based on these results we conjecture that in order to obtain stable quadrature formulas (i.e., with small condition number for the original Gram matrix G_N) exact for degree $n \geq 1$, one has to use at most $4d_n^{q+1}$ uniformly distributed points. In contrast, the theoretical guarantee in Theorem 4.1(c) requires $\mathcal{O}(n^{3q} \log n)$ points.

TABLE 1

The statistics for the experiments with Algorithm LSQ. $M = |\mathcal{C}|$; $n - 2$ is the degree of spherical polynomials for which exact quadrature formulas were computed; $N = n^2$; pos stands for the number of positive weights; and $\kappa(G_N)$, λ_{\min} , and λ_{\max} are the condition number, the maximum eigenvalue, and the minimum eigenvalue of the matrix G_N , respectively.

M	n	Error	$\sum w_{\xi} $	$\min w_{\xi}$	$\max w_{\xi}$	pos	$\kappa(G_N)$	λ_{\min}	λ_{\max}
8192	16	$2.41 * 10^{-15}$	3.5449	$2.29 * 10^{-4}$	$7.88 * 10^{-4}$	8192	2.43	0.607	1.4730
	44	$4.32 * 10^{-15}$	3.5714	$-5.06 * 10^{-4}$	0.0029	8039	37.52	0.078	2.8047
	64	$6.15 * 10^{-15}$	5.5575	-0.00664	0.0073	6068	1695.1	0.003	3.9315
	84	$9.73 * 10^{-12}$	82.152	-0.16274	0.1551	4431	$3.52 * 10^6$	$2.6 * 10^{-6}$	5.4851
16384	44	$4.43 * 10^{-15}$	3.5449	$-9.50 * 10^{-6}$	$8.82 * 10^{-4}$	16382	9.19	0.24036	2.1590
	64	$5.25 * 10^{-15}$	3.5787	$-3.75 * 10^{-4}$	0.0015	16014	51.9	0.05964	2.9150
	84	$7.10 * 10^{-15}$	4.4757	-0.0024	0.0032	13361	944.86	0.00612	3.8457
	100	$1.94 * 10^{-15}$	9.1325	-0.0077	0.0072	10625	11896.1	$4.8 * 10^{-4}$	4.6008
32768	44	$6.02 * 10^{-15}$	3.5449	$3.11 * 10^{-5}$	$2.90 * 10^{-4}$	32768	4.270	0.4157	1.7652
	64	$7.09 * 10^{-15}$	3.5450	$-1.79 * 10^{-5}$	$5.23 * 10^{-4}$	32761	7.977	0.8208	5.2519
	84	$7.71 * 10^{-15}$	3.5574	$-1.43 * 10^{-4}$	$7.92 * 10^{-4}$	32410	42.97	0.0716	2.7967
	100	$7.62 * 10^{-15}$	3.6777	$-4.28 * 10^{-4}$	$9.96 * 10^{-4}$	30819	145.6	0.0250	3.2967

As can be seen from the table, for a fixed degree n , the condition number $\kappa(G_N)$ decreases as the number of points increases. For $n > 140$ and various sets of randomly generated points on the sphere, we do not obtain good numerical results. This might be due to a defect in the built-in numerical procedures used by MATLAB in computing the spherical harmonics of high degree at values close to -1 or 1 . The situation was much better for the dyadic points, i.e., the centers of the dyadic triangles.

For dyadic points on the sphere, the best result we obtained so far is $n = 178$ with 131072 points. As a further verification of this quadrature, we considered the following data. The data is constructed using coefficients $\{a_{\ell,k}\}$ for spherical polynomials up to degree 90, taken from model MF4 used for modeling the lithospheric field. The model, based on CHAMP satellite data, is computed by geophysicists at

GeoForschungsZentrum Potsdam in Germany. We use those coefficients to construct the samples of a function $f = \sum_{\ell,k} a_{\ell,k} Y_{\ell,k}$ at the centers of $8 * 4^7$ dyadic triangles. We then use our precomputed quadrature based at these centers which can integrate spherical polynomials up to degree 178 to compute the Fourier coefficients $\widehat{a}_{\ell,k}$. The maximum difference between the vector $\{\widehat{a}_{\ell,k}\}$ and the vector $\{a_{\ell,k}\}$ was found to be $6.66 * 10^{-15}$.

Next, we considered Algorithm REC. In the context of spherical polynomials, the recurrence relations have to be chosen very carefully using the special function properties of the spherical harmonics $Y_{\ell,k}$ in order to get stable results [29]. In the present situation, the polynomials t_k have no special structure. Therefore, it turns out that Algorithm REC is not very stable for high degrees. However, when we took the centers of 8192 dyadic triangles as the quadrature nodes and used the measure ν^{TR} as the starting measure, then we were able to obtain satisfactory quadrature formulas for degree 32. We note an interesting feature here that all the weights obtained by this algorithm are positive. These results are summarized in Table 2.

TABLE 2
Quadrature constructed using REC on 8192 dyadic points.

n	Error	$\min(w_\xi)$	$\max(w_\xi)$	$\sum w_\xi$
16	$4.196643 * 10^{-14}$	$5.181468 * 10^{-4}$	$2.538441 * 10^{-3}$	12.56637
22	$5.302425 * 10^{-13}$	$5.175583 * 10^{-4}$	$2.543318 * 10^{-3}$	12.56637
32	$9.240386 * 10^{-11}$	$5.154855 * 10^{-4}$	$2.544376 * 10^{-3}$	12.56637
42	$4.434868 * 10^{-8}$	$5.086157 * 10^{-4}$	$2.544141 * 10^{-3}$	12.56637
44	$2.320896 * 10^{-5}$	$5.094771 * 10^{-4}$	$2.562948 * 10^{-3}$	12.56638

Our second set of experiments demonstrates the local approximation properties of the operators $\sigma_n(\mathcal{C}, \mathbf{W}; h)$ for a smooth function h . For this purpose, we consider the following benchmark functions (5.1), considered by various authors [32, 31, 15, 10]. Using the notation $\mathbf{x} = (x_1, x_2, x_3)$, the functions are defined by

$$\begin{aligned}
 g_1(\mathbf{x}) &= (x_1 - 0.9)_+^{3/4} + (x_3 - 0.9)_+^{3/4}, \\
 g_2(\mathbf{x}) &= [0.01 - (x_1^2 + x_2^2 + (x_3 - 1)^2)]_+ + \exp(x_1 + x_2 + x_3), \\
 g_3(\mathbf{x}) &= 1/(101 - 100x_3), \\
 g_4(\mathbf{x}) &= 1/(|x_1| + |x_2| + |x_3|), \\
 g_5(\mathbf{x}) &= \begin{cases} \cos^2\left(\frac{3\pi}{2}\text{dist}(\mathbf{x}, \mathbf{x}_0)\right) & \text{if } \text{dist}(\mathbf{x}, \mathbf{x}_0) < 1/3, \\ 0 & \text{if } \text{dist}(\mathbf{x}, \mathbf{x}_0) \geq 1/3, \end{cases}
 \end{aligned}$$

(5.1) where $\mathbf{x}_0 = (-1/2, -1/2, 1/\sqrt{2})$.

In order to define the function h , we recall first that the B spline B_m of order m is defined recursively [1, p. 131] by

$$(5.2) \quad B_m(x) := \begin{cases} 1 & \text{if } m = 1, 0 < x \leq 1, \\ 0 & \text{if } m = 1, x \in \mathbb{R} \setminus (0, 1], \\ \frac{x}{m-1} B_{m-1}(x) + \frac{m-x}{m-1} B_{m-1}(x-1) & \text{if } m > 1, x \in \mathbb{R}. \end{cases}$$

The function B_m is an $m-1$ times iterated integral of a function of bounded variation. We will choose h to be

$$(5.3) \quad h_m(x) = \sum_{k=-m}^m B_m(2mx - k),$$

for different values of m , in order to illustrate the effect of the smoothness of h_m on the quality of local approximation. If $m \geq 3$, the function h_m satisfies the conditions in Proposition 2.1 with $S = m - 1$. We note that the discretized Fourier projection operator $\sigma_{63}(\mathcal{C}, \mathbf{W}; h_1)$ has been called the hyperinterpolation operator [30].

One example of the localization properties of our operators is given in Table 3, where we show the error in approximation of g_1 on the whole sphere and on the cap $K = \mathbb{S}_{0.4510}^2((-1/\sqrt{2}, 0, -1/\sqrt{2}))$. The operators were constructed using the Driscoll–Healy quadrature formulas [3] based on $4(n+1)^2$ points, exact for integrating polynomials of degree $2n$. The maximum error on the whole sphere, given in columns 2 and 3, is estimated by the error at 10000 randomly chosen points; the maximum error on the cap, given in columns 4 and 5, is estimated by the error at 1000 randomly chosen points on the cap. It is clear that even though the maximum error on the whole sphere is slightly better for the (discretized) Fourier projection than for our summability operator, the singularities of g_1 continue to dominate the error in the Fourier projection on a cap away from these singularities; the performance of our summability operator is far superior.

TABLE 3

$S2errh1 = \max_{\mathbf{x} \in \mathbb{S}^2} |g_1(\mathbf{x}) - \sigma_n(\mathcal{C}, \mathbf{W}; h_1, g_1, \mathbf{x})|$, $S2errh5 = \max_{\mathbf{x} \in \mathbb{S}^2} |g_1(\mathbf{x}) - \sigma_n(\mathcal{C}, \mathbf{W}; h_5, g_1, \mathbf{x})|$,
 $Kerrh1 = \max_{\mathbf{x} \in K} |g_1(\mathbf{x}) - \sigma_n(\mathcal{C}, \mathbf{W}; h_1, g_1, \mathbf{x})|$, $Kerrh5 = \max_{\mathbf{x} \in K} |g_1(\mathbf{x}) - \sigma_n(\mathcal{C}, \mathbf{W}; h_5, g_1, \mathbf{x})|$,
 and $(\mathcal{C}, \mathbf{W})$ are given by the Driscoll–Healy formulas.

n	$S2errh1$	$S2errh5$	$Kerrh1$	$Kerrh5$
63	0.0097	0.0112	$3.4351 * 10^{-4}$	$6.5926 * 10^{-7}$
127	0.0044	0.0055	$8.0596 * 10^{-5}$	$6.5240 * 10^{-8}$
255	0.0033	0.0038	$1.4170 * 10^{-5}$	$1.1816 * 10^{-8}$

Theorem 3.2 points out another way to demonstrate the superior localization of our summability operator without a priori knowledge of the locations of the singularities. Since each of the test functions is infinitely differentiable on large caps of different sizes, Theorem 3.2 suggests that the more localized the method, the greater the probability that the approximation error would be smaller than a given number. To demonstrate also how our ability to construct quadrature formulas based on scattered data helps us to analyze the approximation properties of our summability operators, we took for the set \mathcal{C} a randomly generated sample of 65536 points. For these points, the weights \mathbf{W} computed by Algorithm LSQ yield a quadrature formula exact for integrating spherical polynomials of degree 126. We compared three approximation methods, the least squares approximation from Π_{63}^2 , the approximation given by the operator $\sigma_{63}(\mathcal{C}, \mathbf{W}; h_1)$, and the approximation given by $\sigma_{63}(\mathcal{C}, \mathbf{W}; h_5)$. For each function, we computed the absolute value of the difference between the approximate value computed by each of the three methods and the true value of the function at 20000 randomly chosen points on the sphere. The percentage of points where the value of this difference is less than 10^{-x} is reported in Table 4 for $x = 2 : 10$. It is very obvious that $\sigma_{63}(\mathcal{C}, \mathbf{W}; h_5)$ gives a performance far superior to those of the other methods, due to its localization properties.

TABLE 4

Percentages of error less than 10^{-x} for different functions: $S1 =$ error with $\sigma_{63}(\mathcal{C}, \mathbf{W}; h_1)$, $LS =$ least squares, $S5 =$ error with $\sigma_{63}(\mathcal{C}, \mathbf{W}; h_5)$. For example, for the function g_3 , $S5$ was less than 10^{-7} for 82.22% of the 20000 randomly selected points, while $S1$ (respectively, LS) was less than 10^{-7} for 1.26% (respectively, 1.49%) points.

$x \rightarrow$	10	9	8	7	6	5	4	3	2	
g_1	$S1$	0	0	0.005	0.02	0.42	4.44	39.43	94.79	100
	LS	0	0	0	0.04	0.56	5.32	46.38	95.45	100
	$S5$	0.02	0.19	1.87	16.89	59.36	68.34	79.01	93.09	99.97
g_2	$S1$	0	0.01	0.09	0.74	7.94	84.99	99.19	99.99	100
	LS	0	0.01	0.15	1.29	13.29	86.09	99.28	100	100
	$S5$	0.39	3.34	41.95	90.78	94.52	97.19	99.18	99.97	100
g_3	$S1$	0	0.01	0.11	1.26	12.02	91.87	99.86	100	100
	LS	0	0.01	0.10	1.49	16.26	93.12	99.87	100	100
	$S5$	0.51	5.43	51.08	82.22	91.90	95.79	98.49	99.87	100
g_4	$S1$	0	0	0	0.01	0.18	1.91	18.28	83.81	99.97
	LS	0	0	0.01	0.02	0.25	2.16	21.24	86.43	99.98
	$S5$	0.01	0.01	0.04	0.36	3.47	17.48	40.98	80.06	99.88
g_5	$S1$	0	0	0.01	0.09	1.12	11.84	88.94	99.45	100
	LS	0	0.01	0.01	0.15	1.42	15.23	90.47	99.75	100
	$S5$	0.08	0.64	5.73	66.82	83.54	88.74	92.95	96.78	97.64

TABLE 5

Percentages of error less than 10^{-x} for $\epsilon = 0.01$, $S1 =$ error with $\sigma_{63}(\mathcal{C}, \mathbf{W}; h_1)$, $LS =$ least squares, $S5 =$ error with $\sigma_{63}(\mathcal{C}, \mathbf{W}; h_5)$. The random noise in the left half comes from the uniform distribution in $[-\epsilon, \epsilon]$; that in the right half is from the normal distribution with mean 0 and standard deviation ϵ .

$x \rightarrow$	5	4	3	2	3.0	2.75	2.5	2.25
$S1$	0.05	0.635	9.93	97.45	0	7.97	92.75	100.00
LS	0	0	0	100	0	0.04	30.93	99.07
$S5$	0.085	1.015	10.03	97.49	0.24	51.97	99.87	100.00

Next, we illustrate the stability of our operators under noise. Since our operators are linear operators, we assume for this part of the study that the target function f is the zero function contaminated either by uniform random noise in the range $[-\epsilon, \epsilon]$ or by a normally distributed random variable with mean 0 and standard deviation ϵ . We let \mathcal{C} be a set of 65536 random points and computed corresponding weights \mathbf{W} that integrate exactly polynomial up to degree 126. These were used in calculating $\sigma_{63}(\mathcal{C}, \mathbf{W}; h_1)$ and $\sigma_{63}(\mathcal{C}, \mathbf{W}; h_5)$ at each point of a test data set consisting of 20000 random samples from the distribution μ_2 . For each value of $\epsilon = 0.1; 0.01; 0.001; 0.0001$, the experiment was repeated 50 times, and the errors were then averaged over the number of repetitions. The percentage of points at which the absolute computed value was less than 10^{-x} is reported in Table 5 in the case when $\epsilon = 0.01$. The results for the other values of ϵ were consistent with the linearity of the operator. We observe that in each case, both $\sigma_{63}(\mathcal{C}, \mathbf{W}; h_1)$ and $\sigma_{63}(\mathcal{C}, \mathbf{W}; h_5)$ yield better results than the least squares approximation, while $\sigma_{63}(\mathcal{C}, \mathbf{W}; h_5)$ is slightly superior to $\sigma_{63}(\mathcal{C}, \mathbf{W}; h_1)$.

Finally, we used our operator $\sigma_{22}(\mathcal{C}, \mathbf{W}; h_7)$ with the MAGSAT data. Our purpose here was only to test how our methods work on “real life” data. This data, kindly supplied to us by Dr. Thorsten Maier, measures the magnetic field of the earth in nano-Tesla as a vector field. It was derived from vectorial MAGSAT morning data that has been processed by Nils Olsen of the Danish Space Research Institute. The measurements are averaged on a longitude-latitude grid with $\Delta\phi = 4^\circ$ and $\Delta\theta = 2^\circ$ in geomagnetic coordinates. The radial variations of the MAGSAT satellite have

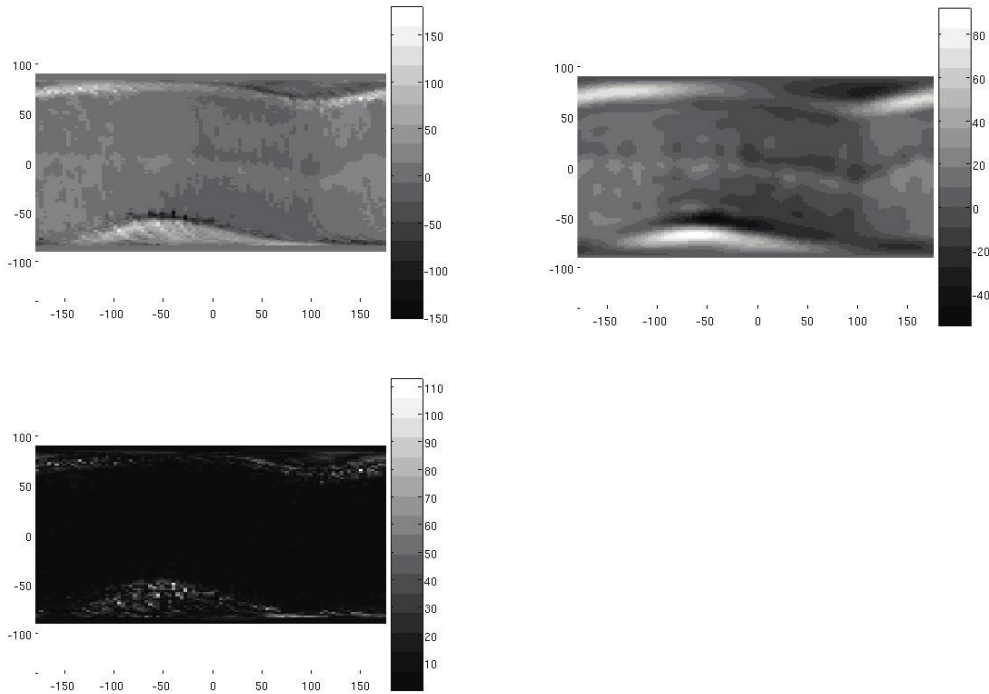


FIG. 2. From left to right: The original data, its reconstruction using $\sigma_{22}(\mathcal{C}, \mathbf{W}; h_7)$, and the error in the approximation, $|\sigma_{22}(\mathcal{C}, \mathbf{W}; h_7) - y|$.

been neglected in the data set and, therefore, prior to the averaging process, the GSFC(12/83) reference potential model has been subtracted. The data results from one month of measurements, centered at March 21, 1980. We extract the east-west component of the vectorial data as a scalar-valued function on the sphere. In total, there are 8190 data sites. A quadrature of degree 44 was computed based on those sites. Figure 2 shows the original data, its reconstruction using $\sigma_{22}(\mathcal{C}, \mathbf{W}; h_7)$, and the error in the approximation, $|\sigma_{22}(\mathcal{C}, \mathbf{W}; h_7) - y|$, as a map in the longitude-latitude plane. As can be seen from the figures, the reconstruction preserves the key features of the original data.

6. Proofs. In the interest of organization, we will prove the various new results in the paper in the following order. We will prove Theorem 4.2 first, since its proof does not require any preparation. We will then use Proposition 2.1 to prove Theorems 3.1(a) and 3.2(a). Next, we will prove Lemma 6.1 and use it to prove parts (a) and (b) of Theorem 4.1. The remaining results involve probabilities. We prove Lemma 6.2 next, estimating the probability that the supremum norm of a sum of random spherical polynomials exceeds a given number. This lemma will be used immediately to prove Theorem 4.1(c). Finally, we will prove Theorems 3.1(b) and 3.2(b).

Proof of Theorem 4.2. It is convenient to prove (4.9) first. Since Py_k is a polynomial of degree $D_k + 1$, there exist real numbers $r_{k,j}(P)$ such that

$$Py_k = \sum_{D_j \leq D_k + 1} r_{k,j}(P)y_j.$$

Since the system $\{y_k\}$ is orthonormal with respect to μ ,

$$r_{k,j}(P) = \int_{\Omega} P y_k y_j d\mu.$$

If $D_j < D_k - 1$, then the degree of $P y_j$ is less than D_k . Because of the lexicographic ordering where lower degree polynomials appear before the higher degree ones, this implies that $P y_j \in V_{k-1}$. Since y_k is orthogonal to V_{k-1} , it follows that $r_{j,k}(P) = 0$ if $D_j < D_k - 1$. This completes the proof of (4.9).

We observe that $y_{p(k)} \in \text{span}\{u_1, \dots, u_{p(k)}\}$. Thus, there exists a constant α such that $\alpha y_{p(k)} - u_{p(k)} \in V_{p(k)-1}$. Thus, $\alpha \tilde{f}_k y_{p(k)} - \tilde{f}_k u_{p(k)} = \alpha \tilde{f}_k y_{p(k)} - u_{k+1}$ is a linear combination of terms of the form $\tilde{f}_k u_j$, $1 \leq j \leq p(k) - 1$. Since $p(k)$ is the minimal index for which there exists a monomial \tilde{f}_k with $\tilde{f}_k u_{p(k)} \in V_{k+1}$, each of the terms $\tilde{f}_k u_j$, $1 \leq j \leq p(k) - 1$, is in V_k . It follows that $\alpha \tilde{f}_k y_{p(k)} - u_{k+1} \in V_k$. Again, there exists a constant α' such that $\alpha' u_{k+1} - y_{k+1} \in V_k$. Therefore, writing $f_k = \alpha \alpha' \tilde{f}_k$, we conclude that $f_k y_{p(k)} - y_{k+1} = \alpha' (\alpha \tilde{f}_k y_{p(k)} - u_{k+1}) + \alpha' u_{k+1} - y_{k+1} \in V_k$; i.e., $f_k y_{p(k)} = y_{k+1} - \sum_{D_j \leq D_k} \tilde{r}_{k,j} y_j$. The first equation in (4.8) is now proved in view of (4.9), applied with $p(k)$ in place of k and the fact that $D_{p(k)} = D_{k+1} - 1$. We note that f_k is a constant multiple of the monomial \tilde{f}_k . The second equation in (4.8) is proved in the same way.

Using the second equation in (4.8) and (4.9), we obtain from the definition of $c_{k,j}$'s that

$$\begin{aligned} A_k c_{k+1,\ell} &= A_k \int_{\Omega} t_{k+1} y_{\ell} d\mu \\ &= \int_{\Omega} f_k t_{p(k)} y_{\ell} d\mu + \sum_{D_{k+1}-2 \leq D_j \leq D_k} s_{k,j} \int_{\Omega} t_j y_{\ell} d\mu \\ &= \sum_{D_{\ell}-1 \leq D_m \leq D_{\ell}+1} r_{\ell,m}(f_k) \int_{\Omega} t_{p(k)} y_m d\mu + \sum_{D_{k+1}-2 \leq D_j \leq D_k} s_{k,j} c_{j,\ell} \\ &= \sum_{D_{\ell}-1 \leq D_m \leq D_{\ell}+1} r_{\ell,m}(f_k) c_{p(k),m} + \sum_{D_{k+1}-2 \leq D_j \leq D_k} s_{k,j} c_{j,\ell}. \end{aligned}$$

This proves (4.10). \square

Next, we use Proposition 2.1 to prove Theorems 3.1(a) and 3.2(a).

Proof of Theorem 3.1(a). To prove part (a), we assume without loss of generality that $n \geq 8$, and let $\ell \geq 1$ be the largest integer with $2^{\ell+2} \leq n$. In view of (3.4),

$$\begin{aligned} \|f - \sigma_n^*(\mathcal{C}, \mathbf{W}; h; f)\|_{\infty} &\leq c E_{n/2}(f) \leq c E_{2^{\ell+1}}(f) \leq c \|f - \sigma_{2^{\ell+1}}^*(h; f)\|_{\infty} \\ &\leq c \sum_{k=\ell+1}^{\infty} \|\sigma_{2^{k+1}}^*(h; f) - \sigma_{2^k}^*(h; f)\|_{\infty} \leq c 2^{-r\ell} \leq c n^{-r}. \end{aligned}$$

This proves part (a). \square

Proof of Theorem 3.2(a). Let K'' be a spherical cap, concentric with K , K' , and having radius equal to the average of the radii of K , K' . Let ψ be a fixed C^{∞} function that is equal to 1 on K'' and equal to 0 outside of K . Without loss of generality, we

can assume that $n \geq 8$, and let $\ell \geq 1$ be the largest integer such that $2^{\ell+2} \leq n$. The direct theorem of approximation theory (cf. [27]) implies that there exists $P \in \Pi_{2^\ell}^q$ such that

$$\|\psi - P\|_\infty \leq c2^{-\ell S}.$$

Therefore, using the definition of $\|f\|_{\mathbb{W}_r(K)}$, we conclude that

$$\begin{aligned} E_{2^{\ell+1}}(f\psi) &\leq \|f\psi - P\sigma_{2^\ell}^*(h; f)\|_\infty \leq \|(f - \sigma_{2^\ell}^*(h; f))\psi\|_\infty + \|(\psi - P)\sigma_{2^\ell}^*(h; f)\|_\infty \\ &\leq c\{\|f - \sigma_{2^\ell}^*(h; f)\|_{\infty, K} + 2^{-nS}\|f\|_\infty\} \\ &\leq c\left\{\sum_{k=\ell+1}^\infty \|\sigma_{2^{k+1}}^*(h; f) - \sigma_{2^k}^*(h; f)\|_{\infty, K} + 2^{-nS}\|f\|_\infty\right\} \leq c2^{-r\ell}. \end{aligned}$$

In view of (3.4),

$$\begin{aligned} \|f - \sigma_n(\mathcal{C}, \mathbf{W}; h; f\psi)\|_{\infty, K'} &= \|f\psi - \sigma_n(\mathcal{C}, \mathbf{W}; h; f\psi)\|_{\infty, K'} \\ &\leq \|f\psi - \sigma_n(\mathcal{C}, \mathbf{W}; h; f\psi)\|_\infty \\ (6.1) \qquad \qquad \qquad &\leq cE_{n/2}(f\psi) \leq E_{2^{\ell+1}}(f\psi) \leq c2^{-r\ell} \leq cn^{-r}. \end{aligned}$$

Since $1 - \psi(\zeta) = 0$ for $\zeta \in K''$, we can use (2.8) to deduce that, for $\mathbf{x} \in K'$,

$$\begin{aligned} |\sigma_n(\mathcal{C}, \mathbf{W}; h; (1 - \psi)f, \mathbf{x})| &= \left| \sum_{\xi \in \mathcal{C} \setminus K''} w_\xi f(\xi)(1 - \psi(\xi))\Phi_n(h; \mathbf{x} \cdot \xi) \right| \\ &\leq \frac{c(K, K', K'')}{n^{S-q}} \|(1 - \psi)f\|_\infty \sum_{\xi \in \mathcal{C}} |w_\xi| \leq \frac{c(K, K', K'')}{n^{S-q}}. \end{aligned}$$

Together with (6.1) and the fact that $r \leq S - q$, this implies (3.9). □

Next, we prove Lemma 6.1, describing certain extremal properties for the weights w_ξ^{LSQ} . These will be used in the proof of parts (a) and (b) of Theorem 4.1.

LEMMA 6.1. *Let $n \geq 1$ be an integer, $N = d_n^{q+1}$, \mathcal{C} be a finite set of points on \mathbb{S}^q , and ν be a measure supported on \mathcal{C} . Let $v_\xi := \nu(\{\xi\})$, $\xi \in \mathcal{C}$.*

(a) *If the Gram matrix is positive definite, then the weights w_ξ^{LSQ} are solutions of the extremal problem to minimize $\sum_{\xi \in \mathcal{C}} w_\xi^2/v_\xi$ subject to the conditions that $\sum_{\xi \in \mathcal{C}} w_\xi y_\ell(\xi) = \delta_{1,\ell}$.*

(b) *If (4.3) holds, there exist real numbers W_ξ , $\xi \in \mathcal{C}$, such that $|W_\xi| \leq v_\xi$ for $\xi \in \mathcal{C}$ and $\sum_{\xi \in \mathcal{C}} W_\xi P(\xi) = \int P d\mu_q$ for all $P \in \Pi_n^q$.*

Proof. In this proof, we will write G in place of G_N . The Lagrange multiplier method for solving the minimization problem sets up parameters λ_ℓ and minimizes

$$\sum_{\xi \in \mathcal{C}} w_\xi^2/v_\xi - 2 \sum_\ell \lambda_\ell \left(\sum_\xi w_\xi y_\ell(\xi) - \delta_{1,\ell} \right).$$

Setting the gradient (with respect to w_ξ) equal to 0, we get $w_\xi = v_\xi \sum_\ell \lambda_\ell y_\ell(\xi)$. Writing, in this proof only, $Q = \sum_\ell \lambda_\ell y_\ell$, we see that $w_\xi = v_\xi Q(\xi)$. Substituting back

in the linear constraints, this reduces to $\sum_{\xi \in \mathcal{C}} v_\xi Q(\xi) y_\ell(\xi) = \delta_{1,\ell}$. These conditions determine Q uniquely; indeed, $Q = \sum_j G_{1,j}^{-1} y_j$. This proves part (a).

Part (b) is proved essentially in [22, 23], but since it is not stated in this manner, we sketch a proof again. During this proof, different constants will retain their values. Let $M = |\mathcal{C}|$, \mathbb{R}^M be equipped with the norm $\|\mathbf{r}\| = \sum_{\xi \in \mathcal{C}} v_\xi |r_\xi|$. In this proof only, let \mathcal{S} be the operator defined on Π_n^q by $\mathcal{S}(P) = (P(\xi))_{\xi \in \mathcal{C}} \in \mathbb{R}^M$, and let \mathbb{V} be the range of \mathcal{S} . The estimate

$$(6.2) \quad \int |P| d\mu_q \leq c_1 \sum_{\xi \in \mathcal{C}} v_\xi |P(\xi)|$$

implies that the operator $\mathcal{S} : \Pi_n^q \rightarrow \mathbb{V}$ is invertible. We may now define a linear functional on \mathbb{V} by

$$x^*(\mathbf{r}) = \int \mathcal{S}^{-1}(\mathbf{r}) d\mu_q, \quad \mathbf{r} \in \mathbb{V}.$$

It is clear from (6.2) that the norm of x^* is bounded above by c_1 . The Hahn–Banach theorem yields a norm-preserving extension of this functional to the whole space \mathbb{R}^M . Identifying this functional with the vector $(W_\xi)_{\xi \in \mathcal{C}}$, the extension property implies that $\sum_{\xi \in \mathcal{C}} W_\xi P(\xi) = \int P d\mu_q$ for all $P \in \Pi_n^q$, while the norm preservation property implies that $|W_\xi| \leq c_1 v_\xi$ for $\xi \in \mathcal{C}$. \square

Proof of Theorem 4.1(a), (b). Let $N = d_n^{q+1}$, $\mathbf{r} \in \mathbb{R}^N$, and $P = \sum_\ell r_\ell y_\ell$. In this proof only, we write G in place of G_N . Then

$$\mathbf{r}^T G \mathbf{r} = \sum_{\ell, m} r_\ell \left\{ \sum_{\xi \in \mathcal{C}} v_\xi y_\ell(\xi) y_m(\xi) \right\} r_m = \sum_{\xi \in \mathcal{C}} v_\xi P(\xi)^2,$$

and $\mathbf{r}^T \mathbf{r} = \|P\|_2^2$.

Therefore, (4.3) with $p = 2$ implies that $c_1^2 \mathbf{r}^T \mathbf{r} \leq \mathbf{r}^T G \mathbf{r} \leq c_2^2 \mathbf{r}^T \mathbf{r}$ for all $\mathbf{r} \in \mathbb{R}^N$. The statements about the eigenvalues of G are an immediate consequence of the Raleigh–Ritz theorem [13, Theorem 4.2.2]. Using Lemma 6.1(b), we obtain weights W_ξ such that $\sum_{\xi \in \mathcal{C}} W_\xi y_\ell(\xi) = \delta_{1,\ell}$, $\ell = 1, \dots, N$, and $|W_\xi| \leq c v_\xi$, $\xi \in \mathcal{C}$. During the remainder of this proof, we write $w_\xi = w_\xi^{LSQ}$. In view of Lemma 6.1(a), we have

$$\sum_{\xi \in \mathcal{C}} |w_\xi| \leq \left\{ \sum_{\xi} v_\xi \right\}^{1/2} \left\{ \sum_{\xi \in \mathcal{C}} \frac{w_\xi^2}{v_\xi} \right\}^{1/2} \leq \left\{ \sum_{\xi} v_\xi \right\}^{1/2} \left\{ \sum_{\xi \in \mathcal{C}} \frac{W_\xi^2}{v_\xi} \right\}^{1/2} \leq c \sum_{\xi \in \mathcal{C}} v_\xi \leq c_1.$$

This completes the proof of part (a).

In order to prove part (b), we adopt the following notation during this proof only. Let I denote the $N \times N$ identity matrix. For any $N \times N$ matrix H , let $\|H\|$ denote $\sup \|H\mathbf{r}\|$, $\|\mathbf{r}\| = 1$, $\mathbf{r} \in \mathbb{R}^N$. We note that $\|H\|$ is the largest singular value of H . If H is a symmetric, positive definite matrix, then it is also the largest eigenvalue of H , and, moreover, $|\mathbf{r}_1^T H \mathbf{r}_2| \leq \|H\| \|\mathbf{r}_1\| \|\mathbf{r}_2\|$, $\mathbf{r}_1, \mathbf{r}_2 \in \mathbb{R}^N$. Using (4.4), it is easy to conclude using the Raleigh–Ritz theorem that $\|G - I\| \leq cn^{-q}$, $\|G^{-1}\| \leq c$, and, hence,

$$\|G^{-1} - I\| = \|G^{-1}(I - G)\| \leq c \|G^{-1}\| \|G - I\| \leq cn^{-q}.$$

Let $\mathbf{y}(\mathbf{x})$ denote the vector $(y_1(\mathbf{x}), \dots, y_N(\mathbf{x}))^T$ for $\mathbf{x} \in \mathbb{S}^q$. In view of the addition formula, $\|\mathbf{y}(\mathbf{x})\|^2$ is independent of \mathbf{x} , and, hence,

$$\|\mathbf{y}(\mathbf{x})\|^2 = \int_{\mathbb{S}^q} \sum_{j=1}^N y_j(\mathbf{x})^2 d\mu_q(\mathbf{x}) = d_n^{q+1} \leq cn^q, \quad \mathbf{x} \in \mathbb{S}^q.$$

Consequently, we have

$$\begin{aligned} \frac{|w_\xi^{LSQ}|}{v_\xi} &= \left| \int \mathbf{y}(\mathbf{x})^T G^{-1} \mathbf{y}(\xi) d\mu_q(\mathbf{x}) \right| \\ &\leq \left| \int \mathbf{y}(\mathbf{x})^T (G^{-1} - I) \mathbf{y}(\xi) d\mu_q(\mathbf{x}) \right| + \left| \int \mathbf{y}(\mathbf{x})^T \mathbf{y}(\xi) d\mu_q(\mathbf{x}) \right| \\ &\leq \|G^{-1} - I\| \int \|\mathbf{y}(\mathbf{x})\| \|\mathbf{y}(\xi)\| d\mu_q + |(1, 0, \dots, 0)^T \mathbf{y}(\xi)| \leq cn^{-q} n^q + c \leq c. \end{aligned}$$

This completes the proof of part (b). \square

The proof of the remaining new results in the paper are based on the following lemma, which gives a recipe for estimating the probabilities involving polynomial-valued random variables.

LEMMA 6.2. *Let $n, M \geq 1$ be integers, let $\{\omega_j\}_{j=1}^M$ be independent random variables, and, for $j = 1, \dots, M$, let $Z_j = Z(\omega_j, \circ) \in \Pi_n^q$ have mean equal to 0 according to ω_j . Let $B, R > 0$, $\max_{1 \leq j \leq M, \mathbf{x} \in \mathbb{S}^q} |Z_j(\mathbf{x})| \leq Rn^q$, and the sum of the variances of Z_j be bounded by Bn^q uniformly on \mathbb{S}^q . If $A > 0$ and $12R^2(A + q)n^q \log n \leq B$, then*

$$(6.3) \quad \text{Prob} \left(\left\| \sum_{j=1}^M Z_j \right\|_\infty \geq \sqrt{12B(A + q)n^q \log n} \right) \leq c_1 n^{-A}.$$

Here, the positive constant c_1 is independent of M and the distributions of ω_j .

Proof. The proof depends upon Bennett's inequality [28, p. 192]. In this proof only, we adopt a slightly different meaning for the symbols L, V, η . Let L, V, η be positive numbers, and let $X_j, j = 1, \dots, M$, be independent random variables. According to Bennett's inequality, if the mean of each X_j is 0, the range of each X_j is a subset of $[-L, L]$, and V exceeds the sum of the variances of X_j , then, for $\eta > 0$,

$$(6.4) \quad \text{Prob} \left(\left| \sum_{j=1}^M X_j \right| \geq \eta \right) \leq 2 \exp \left(-\frac{V}{L^2} g \left(\frac{L\eta}{V} \right) \right),$$

where, in this proof only, $g(t) := (1+t) \log(1+t) - t$. We observe that $g(t) = \int_0^t \int_0^u (1+w)^{-1} dw du$. Therefore, if $0 \leq t \leq 1/2$, then for $0 \leq w \leq u \leq t$, $(1+w)^{-1} \geq 2/3$ and hence, $g(t) \geq t^2/3$. Consequently, if $L\eta \leq V/2$, then

$$(6.5) \quad \text{Prob} \left(\left| \sum_{j=1}^M X_j \right| \geq \eta \right) \leq 2 \exp \left(\frac{-\eta^2}{(3V)} \right).$$

Now, let $\mathbf{x} \in \mathbb{S}^q$. We apply (6.5) with $Z_j(\mathbf{x})$ in place of X_j , Rn^q in place of L , Bn^q in place of V , and $\eta = \sqrt{3B(A + q)n^q \log n}$. Our condition on n ensures that

$L\eta/V \leq 1/2$ with these choices. Therefore,

$$(6.6) \quad \text{Prob} \left(\left| \sum_{j=1}^M Z_j(\mathbf{x}) \right| \geq \sqrt{3B(A+q)n^q \log n} \right) \leq 2n^{-A-q}.$$

Next, in the proof only, let $P^* = \sum_{j=1}^M Z_j$, $\mathbf{x}^* \in \mathbb{S}^q$ be chosen so that $|P^*(\mathbf{x}^*)| = \|P^*\|_\infty$, and let $\mathcal{C} \subset \mathbb{S}^q$ be chosen so that $|\mathcal{C}| \sim cn^q$ and $\delta_{\mathcal{C}}(\mathbb{S}^q) \leq 1/(2n)$. Then we may find $\xi^* \in \mathcal{C}$ such that $\text{dist}(\mathbf{x}^*, \xi^*) \leq 1/(2n)$. Since $P^* \in \Pi_n^q$, its restriction to the great circle through \mathbf{x}^* and ξ^* is a trigonometric polynomial of order at most n . In view of the Bernstein inequality for these polynomials [2, Chapter 4, inequality (1.1)],

$$|P^*(\xi^*) - P^*(\mathbf{x}^*)| \leq n\|P^*\|_\infty \text{dist}(\xi^*, \mathbf{x}^*) \leq (1/2)|P^*(\mathbf{x}^*)|.$$

We deduce that

$$\left\| \sum_{j=1}^M Z_j \right\|_\infty \leq 2 \max_{\mathbf{x} \in \mathcal{C}} \left| \sum_{j=1}^M Z_j(\mathbf{x}) \right|.$$

Therefore, the event $\|\sum_{j=1}^M Z_j\|_\infty \geq 2\sqrt{3B(A+q)n^q \log n}$ is a subset of the union of the $|\mathcal{C}|$ events $|\sum_{j=1}^M Z_j(\mathbf{x})| \geq \sqrt{3B(A+q)n^q \log n}$, $\mathbf{x} \in \mathcal{C}$. Hence, the estimate (6.6) implies (6.3) with $c_1 = 2c$. \square

We are now in a position to prove Theorem 4.1(c).

Proof of Theorem 4.1(c). Let $\mathbf{x} \in \mathbb{S}^q$. In this proof only, let $Z_\xi = \Phi_{4n}(h; \mathbf{x} \cdot \xi) - \int \Phi_{4n}(h; \mathbf{x} \cdot \zeta) d\mu_q(\zeta)$. Then the mean of each Z_ξ is 0, and its variance can be estimated by

$$\int Z_\xi^2 d\mu_q(\xi) \leq \int (\Phi_{4n}(h; \mathbf{x} \cdot \xi))^2 d\mu_q(\xi) \leq cn^q.$$

Finally, $|Z_\xi| \leq cn^q$ for each ξ . Hence, we may use Lemma 6.2, with cM in place of B and c in place of R , to conclude that

$$\text{Prob} \left(\sup_{\mathbf{x} \in \mathbb{S}^q} \left| \frac{1}{M} \sum_{\xi \in \mathcal{C}} \Phi_{4n}(h; \mathbf{x} \cdot \xi) - \int \Phi_{4n}(h; \mathbf{x} \cdot \zeta) d\mu_q(\zeta) \right| \geq c_2 \sqrt{\frac{n^q \log n}{M}} \right) \leq cn^{-A}$$

and, with $M \geq cn^q \log n / \eta^2$,

$$\text{Prob} \left(\sup_{\mathbf{x} \in \mathbb{S}^q} \left| \frac{1}{M} \sum_{\xi \in \mathcal{C}} \Phi_{4n}(h; \mathbf{x} \cdot \xi) - \int \Phi_{4n}(h; \mathbf{x} \cdot \zeta) d\mu_q(\zeta) \right| \geq \eta \right) \leq cn^{-A}.$$

Since any $P \in \Pi_{2n}^q$ can be written in the form

$$P(\zeta) = \int P(\mathbf{x}) \Phi_{4n}(h; \mathbf{x} \cdot \zeta) d\mu_q(\mathbf{x}),$$

we see that, with probability exceeding $1 - cn^{-A}$,

$$\begin{aligned}
 & \left| \frac{1}{M} \sum_{\xi \in \mathcal{C}} P(\xi) - \int P(\zeta) d\mu_q(\zeta) \right| \\
 &= \left| \frac{1}{M} \sum_{\xi \in \mathcal{C}} \int P(\mathbf{x}) \Phi_{4n}(h; \xi \cdot \mathbf{x}) d\mu_q(\mathbf{x}) \right. \\
 &\quad \left. - \int \int P(\mathbf{x}) \Phi_{4n}(h; \mathbf{x} \cdot \zeta) d\mu_q(\mathbf{x}) d\mu_q(\zeta) \right| \\
 &\leq \int |P(\mathbf{x})| \left| \frac{1}{M} \sum_{\xi \in \mathcal{C}} \Phi_{4n}(h; \xi \cdot \mathbf{x}) - \int \Phi_{4n}(h; \mathbf{x} \cdot \zeta) d\mu_q(\zeta) \right| d\mu_q(\mathbf{x}) \\
 &\leq \eta \int |P(\mathbf{x})| d\mu_q(\mathbf{x}).
 \end{aligned}$$

For $P \in \Pi_n^q$, we may now apply this estimate with $P^2 \in \Pi_{2n}^q$. \square

Another immediate consequence of Lemma 6.2 is the following lemma, describing the probabilistic behavior of the operator $\sigma_n(\mathcal{C}, \mathbf{W}; h)$.

LEMMA 6.3. *Suppose that $m \geq 1$ is an integer and that $\mathcal{C} = \{\xi_j\}_{j=1}^M$ admits an M-Z quadrature of order m , and let \mathbf{W} be the corresponding quadrature weights. Let $R, V > 0$, and, for $j = 1, \dots, M$, let ϵ_j be independent random variables with mean 0, variance not exceeding V , and range $[-R, R]$. Let $g \in C(\mathbb{S}^q)$, $\|g\|_\infty \leq 1$, and $\mathbf{E} = \{\epsilon_j g(\xi_j)\}_{\xi_j \in \mathcal{C}}$. Then, for integer $n \geq 1$ with $(R^2/V)(A + q)n^q \log n \leq c_3 m^q$,*

$$(6.7) \quad \text{Prob} \left(\|\sigma_n(\mathcal{C}, \mathbf{W}; h; \mathbf{E})\|_\infty \geq c_1 \sqrt{\frac{V(A + q)n^q \log n}{m^q}} \right) \leq c_2 n^{-A}.$$

Here the positive constants c_1, c_2, c_3 depend only on q but not on M and the distributions of ϵ_j .

Proof. In this proof only, if $\xi = \xi_j \in \mathcal{C}$, we will write ϵ_ξ for ϵ_j and w_ξ for the weight in \mathbf{W} corresponding to ξ . We use Lemma 6.2 with $E_\xi = m^q w_\xi \epsilon_\xi g(\xi) \Phi_n(h; \xi \cdot \circ)$, $\xi \in \mathcal{C}$. We note that the random variable ω_j in Lemma 6.2 is ϵ_ξ in this case. It is clear that the mean of each E_ξ is 0. Since (2.12) implies that $|w_\xi| \leq cm^{-q}$, (2.7) shows that $\|E_\xi\|_\infty \leq cRn^q$. Moreover, for any $\mathbf{x} \in \mathbb{S}^q$, the variance of $E_\xi(\mathbf{x})$ does not exceed $Vm^{2q} w_\xi^2 \Phi_n(h; \xi \cdot \mathbf{x})^2$. In view of the fact that w_ξ are M-Z quadrature weights, (2.12) and (2.7) imply that

$$\begin{aligned}
 \sum_{\xi \in \mathcal{C}} m^{2q} w_\xi^2 \Phi_n(h; \xi \cdot \mathbf{x})^2 &\leq cm^q \sum_{\xi \in \mathcal{C}} |w_\xi| \Phi_n^2(h; \xi \cdot \mathbf{x}) \\
 &\leq cm^q \int_{\mathbb{S}^q} \Phi_n^2(h; \zeta \cdot \mathbf{x}) d\mu_q(\zeta) \leq cm^q n^q.
 \end{aligned}$$

Thus, we may choose B in Lemma 6.2 to be cVm^q . The estimate (6.7) now follows as a simple consequence of (6.3). \square

We are now in a position to prove the probabilistic assertions of Theorems 3.1 and 3.2.

Proof of Theorem 3.1(b). To prove part (b), we use Lemma 6.3 with $g \equiv 1$. Since the range of ϵ_j 's is contained in $[-1, 1]$, we can take $R = V = 1$ and obtain from (6.7) that

$$\text{Prob} \left(\|\sigma_n(\mathcal{C}, \mathbf{W}; h; \mathbf{E})\|_\infty \geq c_4 \sqrt{\frac{(A+q)n^q \log n}{m^q}} \right) \leq c_2 n^{-A}.$$

The choice of n with an appropriate c_3 ensures that $(A+q)n^q \log n / m^q \leq n^{-2r}$. Therefore,

$$\text{Prob} (\|\sigma_n(\mathcal{C}, \mathbf{W}; h; \mathbf{E})\|_\infty \geq c_4 n^{-r}) \leq c_2 n^{-A}.$$

The estimate (3.8) is now clear in view of (3.4) and the linearity of the operators $\sigma_n(\mathcal{C}, \mathbf{W}; h)$. \square

Proof of Theorem 3.2(b). We apply Lemma 6.3 again with $g \equiv 1$. As before, we may choose $R = V = 1$. The choice of n with an appropriate c_3 ensures that $(A+q)n^q \log n / m^q \leq n^{-2r}$. Therefore, (6.7) with these choices implies that

$$\text{Prob} (\|\sigma_n(\mathcal{C}, \mathbf{W}; h; \mathbf{E})\|_\infty \geq c_4 n^{-r}) \leq c_2 n^{-A}.$$

Together with inequality (3.9) and the linearity of the operators $\sigma_n(\mathcal{C}, \mathbf{W}; h)$, this leads to (3.10). \square

7. Conclusion. We have described a construction of linear operators yielding spherical polynomial approximations based on scattered data on a Euclidean sphere. While the operators can be defined for arbitrary continuous functions on the sphere, without any a priori knowledge about the location and nature of its singularities, they are auto-adaptive in the sense that the approximation properties of these globally defined polynomials adapt themselves on the different parts of the sphere according to the smoothness of the target function on these parts. While the theoretical properties of these operators and their localization were studied in [18], a bottleneck in their numerical construction was the construction of quadrature formulas based on scattered data, exact for integrating moderately high degree spherical polynomials. Until now, it has been possible only to compute quadrature formulas exact for at most degree 18 polynomials. We show that a simple construction involving a Gram matrix is surprisingly well conditioned, and yields the necessary quadrature rules, up to degree 178. Using these newly constructed quadrature formulas, we are able to demonstrate that our constructions yield approximation properties superior to those of more traditional techniques of least squares and Fourier projection, in the sense that the presence of singularities in some parts of the sphere affects the degree of approximation by our operators on other parts far less than in the case of these other traditional techniques. We give probabilistic estimates on the local and global degrees of approximation by our operators in the presence of noise and demonstrate its use in the modeling of a “real life” data set. We also describe a theoretical algorithm for construction of data dependent multivariate orthogonal polynomials and their use in the construction of quadrature formulas, analogous to the univariate algorithms in Gautschi’s book [11].

Acknowledgments. This paper is a result of a long process, involving discussions with a number of mathematicians. In particular, it is our pleasure to acknowledge the support and encouragement of Mahadevan Ganesh, Thorsten Maier, Volker Michel, Dominik Michel, Ian Sloan, and Joe Ward. We are also grateful to the two referees and Fred Hickernell for their many useful suggestions for the improvement of the first draft of this paper.

REFERENCES

- [1] C. DE BOOR, *A Practical Guide to Splines*, Springer-Verlag, New York, 1978.
- [2] R. A. DEVORE AND G. G. LORENTZ, *Constructive Approximation*, Springer-Verlag, Berlin, 1993.
- [3] J. R. DRISCOLL AND D. M. HEALY, *Computing Fourier transforms and convolutions on the 2-sphere*, Adv. in Appl. Math., 15 (1994), pp. 202–250.
- [4] C. F. DUNKL AND Y. XU, *Orthogonal Polynomials of Several Variables*, Cambridge University Press, Cambridge, UK, 2001.
- [5] A. ERDÉLYI, W. MAGNUS, F. OBERHETTINGER, AND F. G. TRICOMI, *Higher Transcendental Functions*, Vol. II, California Institute of Technology, Bateman Manuscript Project, McGraw-Hill, New York, Toronto, London, 1953.
- [6] B. FISCHER, *From orthogonal polynomials to iteration schemes for linear systems: CG and CR revisited*, in Wavelet Analysis and Applications, Proceedings of the International Workshop in Delhi, 1999, P. K. Jain, M. Krishnan, H. N. Mhaskar, J. Prestin, and D. Singh, eds., Narosa Publishing, New Delhi, India, 2001, pp. 225–247.
- [7] W. FREEDEN, T. GERVENES, AND M. SCHREINER, *Constructive Approximation on the Sphere, with Applications to Geomathematics*, The Clarendon Press, Oxford University Press, New York, 1998.
- [8] W. FREEDEN AND V. MICHEL, *Multiscale Potential Theory, with Applications to Geoscience*, Birkhäuser Boston, Boston, 2004.
- [9] W. FREEDEN, M. SCHREINER, AND R. FRANKE, *A survey on spherical spline approximation*, Surveys Math. Indust., 7 (1997), pp. 29–85.
- [10] M. GANESH AND H. N. MHASKAR, *Matrix-free interpolation on the sphere*, SIAM J. Numer. Anal., 44 (2006), pp. 1314–1331.
- [11] W. GAUTSCHI, *Orthogonal Polynomials: Computation and Approximation*, Oxford University Press, New York, 2004.
- [12] K. HESSE, H. N. MHASKAR, AND I. H. SLOAN, *Quadrature in Besov spaces on the Euclidean sphere*, J. Complexity, 23 (2007), pp. 528–552.
- [13] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
- [14] J. KEINER, *Fast Spherical Fourier Transforms and Applications*, Diplomarbeit, Universität zu Lübeck, Lübeck, Germany, 2005.
- [15] Q. T. LE GIA AND H. N. MHASKAR, *Polynomial operators and local approximation of solutions of pseudo-differential equations on the sphere*, Numer. Math., 103 (2006), pp. 299–322.
- [16] P. I. LIZORKIN AND KH. P. RUSTAMOV, *Nikolskii-Besov spaces on a sphere that are associated with approximation theory*, Tr. Mat. Inst. Steklova, 204 (1993), pp. 172–201 (in Russian); Proc. Steklov Inst. Math., 3 (1994), pp. 149–172 (in English).
- [17] H. N. MHASKAR, *Polynomial operators and local smoothness classes on the unit interval*, J. Approx. Theory, 131 (2004), pp. 243–267.
- [18] H. N. MHASKAR, *On the representation of smooth functions on the sphere using finitely many bits*, Appl. Comput. Harmon. Anal., 18 (2005), pp. 215–233.
- [19] H. N. MHASKAR, *Weighted quadrature formulas and approximation by zonal function networks on the sphere*, J. Complexity, 22 (2006), pp. 348–370.
- [20] H. N. MHASKAR, F. NARCOWICH, J. PRESTIN, AND J. D. WARD, *L^p Bernstein Estimates and Approximation by Spherical Basis Functions*, manuscript; available online from <http://www.arxiv.org/abs/0810.5075>.
- [21] H. N. MHASKAR, F. J. NARCOWICH, AND J. D. WARD, *Approximation properties of zonal function networks using scattered data on the sphere*, Adv. Comput. Math., 11 (1999), pp. 121–137.
- [22] H. N. MHASKAR, F. J. NARCOWICH, AND J. D. WARD, *Spherical Marcinkiewicz-Zygmund inequalities and positive quadrature*, Math. Comp., 70 (2001), pp. 1113–1130.
- [23] H. N. MHASKAR, F. J. NARCOWICH, AND J. D. WARD, *Corrigendum to: “Spherical Marcinkiewicz-Zygmund inequalities and positive quadrature”* [Math. Comp., 70 (2001), pp. 1113–1130], Math. Comp., 71 (2002), pp. 453–454.
- [24] H. N. MHASKAR AND J. PRESTIN, *On the detection of singularities of a periodic function*, Adv. Comput. Math., 12 (2000), pp. 95–131.
- [25] H. N. MHASKAR AND J. PRESTIN, *On local smoothness classes of periodic functions*, J. Fourier Anal. Appl., 11 (2005), pp. 353–373.
- [26] C. MÜLLER, *Spherical Harmonics*, Lecture Notes in Math. 17, Springer-Verlag, Berlin, 1966.
- [27] S. PAWELKE, *Über die Approximationsordnung bei Kugelfunktionen und algebraischen Polynomen*, Tôhoku Math. J. (2), 24 (1972), pp. 473–486.

- [28] D. POLLARD, *Convergence of Stochastic Processes*, Springer-Verlag, New York, 1984.
- [29] D. POTTS, G. STEIDL, AND M. TASCHE, *Fast algorithms for discrete polynomial transforms*, *Math. Comp.*, 67 (1998), pp. 1577–1590.
- [30] I. H. SLOAN, *Polynomial interpolation and hyperinterpolation over general regions*, *J. Approx. Theory*, 83 (1995), pp. 238–254.
- [31] I. H. SLOAN AND A. SOMMARIVA, *Approximation on the sphere using radial basis functions plus polynomials*, *Adv. Comput. Math.*, 29 (2008), pp. 147–177.
- [32] I. H. SLOAN AND R. S. WOMERSLEY, *Extremal systems of points and numerical integration on the sphere*, *Adv. Comput. Math.*, 21 (2004), pp. 107–125.
- [33] E. M. STEIN AND G. WEISS, *Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton, NJ, 1971.

ON THE INTERPOLATION ERROR ESTIMATES FOR \mathcal{Q}_1 QUADRILATERAL FINITE ELEMENTS*

SHIPENG MAO[†], SERGE NICAISE[‡], AND ZHONG-CI SHI[†]

Abstract. In this paper, we study the relation between the error estimate of the bilinear interpolation on a general quadrilateral and the geometric characters of the quadrilateral. Some explicit bounds of the interpolation error are obtained based on some sharp estimates of the integral over $\frac{1}{|J|^{p-1}}$ for $1 \leq p \leq \infty$ on the reference element, where J is the Jacobian of the nonaffine mapping. This allows us to introduce weak geometric conditions (depending on p) leading to interpolation error estimates in the $W^{1,p}$ norm, for any $p \in [1, \infty)$, which can be regarded as a generalization of the *regular decomposition property* (RDP) condition introduced in [G. Acosta and R. G. Durán, *SIAM J. Numer. Anal.*, 38 (2000), pp. 1073–1088] for $p = 2$ and new RDP conditions (NRDP) for $p \neq 2$. We avoid the use of the reference family elements, which allows us to extend the results to a larger class of elements and to introduce the NRDP condition in a more unified way. As far as we know, the mesh condition presented in this paper is weaker than any other mesh conditions proposed in the literature for any p with $1 \leq p \leq \infty$.

Key words. error estimates, quadrilateral elements, isoparametric finite elements, maximal angle condition

AMS subject classifications. 65N30, 65N15

DOI. 10.1137/070700486

1. Introduction. Quadrilateral finite elements, particularly low order quadrilateral elements, are widely used in engineering computations due to their flexibility and simplicity. However, numerical accuracy cannot be achieved over arbitrary quadrilateral meshes, and certain geometric conditions are indispensable to guarantee the optimal convergence error estimates. It is known that the \mathcal{Q}_1 quadrilateral finite element is one of the most widely used quadrilateral elements. In order to obtain its optimal interpolation error, many mesh conditions have been introduced in the literature; let us give a review of them.

Denoting by \mathcal{Q} the Lagrange interpolation operator and using the standard notation for Sobolev spaces (cf. [11]), the first interpolation error estimate for the operator \mathcal{Q} goes back to Ciarlet and Raviart in [14], where the regular quadrilateral is supposed to satisfy

$$(1.1) \quad h_K / \bar{h}_K \leq \mu_1$$

and

$$(1.2) \quad |\cos \theta_K| \leq \mu_2 < 1$$

for all angles θ_K of the quadrilateral K ; here h_K is the diameter of K , and \bar{h}_K is the length of the shortest side of K . Under the above so-called nondegenerate condition,

*Received by the editors August 20, 2007; accepted for publication (in revised form) August 11, 2008; published electronically December 5, 2008. This research was supported by the National Basic Research Program of China under grant 2005CB321701.

<http://www.siam.org/journals/sinum/47-1/70048.html>

[†]LSEC, Institute of Computational Mathematics, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing, 100190, People's Republic of China (maosp@lsec.cc.ac.cn, shi@lsec.cc.ac.cn).

[‡]Université de Valenciennes et du Hainaut Cambrésis, LAMAV, ISTV, F-59313 - Valenciennes Cedex 9, France (snicaise@univ-valenciennes.fr, <http://www.univ-valenciennes.fr/lamav/nicaise/accueil.htm>).

Ciarlet and Raviart proved the following interpolation error estimate:

$$(1.3) \quad |u - \mathcal{Q}u|_{1,K} \leq Ch_K |u|_{2,K}.$$

On the other hand, error estimates for degenerate elements have attracted much attention since the works by Babuška and Aziz [9] and by Jamet [19]; interested readers are also referred to the works [5, 7, 8, 12, 13, 15, 17, 18, 21, 23, 24, 22, 32], the book [6] by Apel, the ICM report [16] by Durán, and the references therein. For triangular elements, the constant C in the estimate (1.3) depends only on the maximal angle of the element. For quadrilaterals, the situation may be different since the maximal angle condition is not necessary due to [20]. In such a case, the term “degenerate” is used in the following two situations: one refers to elements which are close to a triangle, while the other refers to narrow elements or anisotropic elements; interested readers are referred to Jamet [20] for the first case and to Ženíšek and Vanmaele [29, 30] and Apel [5] for the second.

In the first case, Jamet [20] considered a quadrilateral that can degenerate into a regular triangle and proved the error estimate (1.3) under the condition that there exists a constant σ such that

$$h_K/\rho_K \leq \sigma,$$

where ρ_K denotes the diameter of the maximum circle contained in quadrilateral K .

In view of this result, one may believe that the maximal angle condition is not necessary for the optimal interpolation error of the \mathcal{Q}_1 Lagrange interpolation operator. Recently, Acosta and Durán [2] made a great contribution in this regard and obtained the optimal interpolation error of \mathcal{Q}_1 interpolation under the regular decomposition property (RDP) condition (cf. the definition in section 2), which states that, if we divide the quadrilateral into two triangles by one diagonal, the ratio of the length of the other diagonal to that of the first is bounded, and both of the divided triangles satisfy the maximal angle condition. The above RDP condition is so weak that almost all of the degenerate quadrilateral conditions proposed before fall into this scope. Furthermore, the authors of [2] assert that this condition is necessary and state it as an open problem in the conclusion of their paper. For related papers concerning this assertion, we refer the reader to [25, 26, 31]. More recently, interpolation error estimates have been extended to the L^p setting; more precisely, the estimate

$$|u - \mathcal{Q}u|_{1,p,K} \leq Ch_K |u|_{2,p,K}$$

was proved by Acosta and Monzón [4] in the case $1 \leq p < 3$. In addition, the authors of [4] introduced the *double angle condition* (DAC), which is indeed equivalent to (1.2), and showed that this condition is a sufficient condition for the optimal interpolation error in the $W^{1,p}$ norm with $p \geq 3$. Though the DAC condition is much stronger than the RDP condition, so far, it is the weakest mesh condition for the optimal interpolation error in the $W^{1,p}$ norm with $p \geq 3$. One of the key techniques employed in [2] and [4] is to introduce an appropriate affine change of variables, which reduces the problem to a reference family of elements.

In this paper, we revisit the optimal error estimates of \mathcal{Q}_1 isoparametric Lagrange interpolation for degenerate quadrilaterals. Our motivation comes from the observation that, if we divide the quadrilateral into two triangles by the longer diagonal, when the two triangles have comparable areas, we should impose the maximal angle condition for both triangles. Otherwise, we may need to impose the maximal angle condition only for the big triangle T_1 , and because the error on the small triangle

T_3 contributes little to the interpolation error on the global quadrilateral, its maximal angle may become very large as $\frac{|T_3|}{|T_1|}$ approaches zero. Based on this observation, we introduce a generalized RDP condition which involves the ratio between the area of the two divided triangles in the mesh condition and show that under this generalized RDP condition the estimate (1.3) is valid. The interpolation error of \mathcal{Q}_1 Lagrange interpolation in the $W^{1,p}$ norm with $1 \leq p \leq \infty$ is proved in the same spirit. More precisely, we have found the weakest known geometric condition on a quadrilateral K (that depends on p) such that

$$(1.4) \quad |u - \mathcal{Q}u|_{1,p,K} \leq Ch_K |u|_{2,p,K}, \quad p \in [1, \infty),$$

holds. One of the key points in the proof of this estimate is to bound the integral

$$\int_{\widehat{K}} \frac{1}{|J|^{p-1}},$$

where J is the Jacobian of the mapping sending the reference element \widehat{K} to K . It turns out that to obtain such a bound, the cases $p \in [1, 2)$, $p = 2$, $p \in (2, 3)$, $p \in [3, \frac{7}{2}]$, $p \in (\frac{7}{2}, 4]$, and $p > 4$ have to be distinguished, leading to different geometric hypotheses. Note that, for $p \geq 3$, the proposed condition is much weaker than the DAC condition proposed in [4]. The technique developed in this paper is a combination of those in [2] and [20].

The rest of the paper is organized as follows. In section 2, we present our motivation for the geometric condition by revisiting a simple example considered in [2], and based on some observations we propose our generalized RDP (GRDP) condition for the optimal H^1 interpolation error. In section 3, following the techniques developed in [2] and [20], we prove the optimal interpolation error in the H^1 norm for the \mathcal{Q}_1 quadrilateral finite element. In section 4, we prove the above-mentioned technical bounds for different values of p . Then in section 5 we introduce our different geometric conditions depending on the values of $p \geq 1$ in the above-mentioned intervals and prove the interpolation error estimate (1.4) for all $p \geq 1$.

2. The generalized regular decomposition property. In this section, we will introduce a mesh condition that is sufficient for (1.3). This can be regarded as a generalization of the *regular decomposition property* presented by Acosta and Durán in [2].

We will adopt the following notation. Let K be a general quadrilateral with its vertices M_1, M_2, M_3, M_4 enumerated in counterclockwise order. In order to define the isoparametric elements on K , if $\widehat{K} = [0, 1]^2$ denotes the reference element, then there exists a bijective mapping $F_K : \widehat{K} \rightarrow K$ which is defined as

$$(2.1) \quad M = F_K(\widehat{M}) = \sum_{i=1}^4 M_i \widehat{\phi}_i(\xi, \eta) \quad \forall \widehat{M} = (\xi, \eta) \in \widehat{K},$$

where $\widehat{\phi}_i$ is the bilinear basis function associated with the vertex \widehat{M}_i , i.e., $\widehat{\phi}_i(\widehat{M}_j) = \delta_i^j, i = 1, 2, 3, 4$ and $j = 1, 2, 3, 4$.

Let the basis functions on the general quadrilateral K be defined as $\phi_i(M) = \widehat{\phi}_i(\widehat{M}) = \widehat{\phi}_i(F_K^{-1}(M))$ for any point $M \in K$. Then the \mathcal{Q}_1 isoparametric interpolation operator is defined by

$$\mathcal{Q}u(M) = \widehat{\mathcal{Q}}\widehat{u}(\widehat{M}) \quad \text{for any point } M \in K,$$

and $\widehat{\mathcal{Q}}$ is the bilinear Lagrange interpolation operator on \widehat{K} .

There are several mesh conditions in the literature that lead to (1.3). Among them, the RDP condition proposed by Acosta and Durán in [2] is the weakest. It is defined as follows.

DEFINITION 2.1. *A quadrilateral or a triangle verifies the maximal angle condition (MAC) with constant $\psi < \pi$, or $MAC(\psi)$, if the interior angles of K are less than or equal to ψ .*

DEFINITION 2.2. *Let K be a convex quadrilateral. We say that K satisfies the regular decomposition property with constants $N \in \mathbb{R}_+$ and $0 < \psi < \pi$, or $RDP(N, \psi)$, if we can divide K into two triangles along one of its diagonals, always called d_1 , the other being denoted by d_2 in such a way that $|d_2|/|d_1| \leq N$ and both triangles satisfy $MAC(\psi)$.*

In order to motivate our mesh condition introduced below, we first analyze the following examples. Let $K = K(a, b, \tilde{a}, \tilde{b})$ be the convex quadrilateral with vertices $M_1 = (0, 0), M_2 = (a, 0), M_3 = (\tilde{a}, \tilde{b}), M_4 = (0, b)$. Consider the case $K(1, a, a, a)$ (cf. the left-hand side of Figure 2.1) and take $u = x^2$. Straightforward computations show that

$$\left\| \frac{\partial(u - \mathcal{Q}u)}{\partial y} \right\|_{0,K}^2 \geq Ca \ln(a^{-1}) \quad \text{and} \quad |u|_{2,K}^2 \leq Ca.$$

Then the constant on the right-hand side of (1.3) cannot be bounded when a approaches zero. This is just the counterexample given in [2].

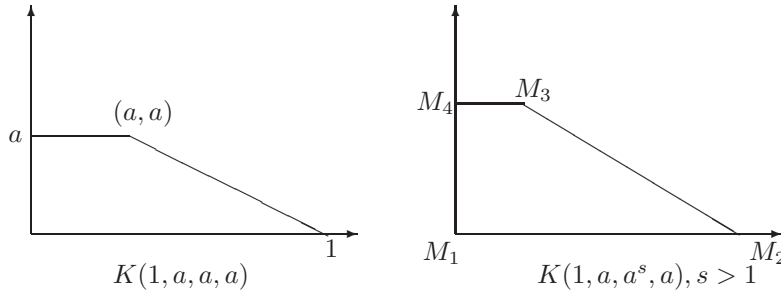


FIG. 2.1.

If we consider the case $K(1, a, a^s, a)$ with $s > 1$ (the right-hand side of Figure 2.1), we have

$$\left\| \frac{\partial(u - \mathcal{Q}u)}{\partial y} \right\|_{0,K}^2 \leq Ca^{2s-1} \ln(a^{-1}), \quad |u|_{2,K}^2 \geq Ca.$$

However, in this case the error constant

$$\frac{\left\| \frac{\partial(u - \mathcal{Q}u)}{\partial y} \right\|_{0,K}^2}{|u|_{2,K}^2} \leq Ca^{2s-2} \ln(a^{-1})$$

can be bounded with a constant independent of a . Both cases do not satisfy the RDP condition since the MAC of $\triangle M_2 M_3 M_4$ is violated if we divide the quadrilateral by the diagonal $M_2 M_4$.

What is the difference between these two examples? One reasonable interpretation is that for $s > 1$ the ratio $\frac{|\triangle M_2 M_3 M_4|}{|\triangle M_1 M_2 M_4|} = a^s$ for $K(1, a, a^s, a)$ is much smaller than the

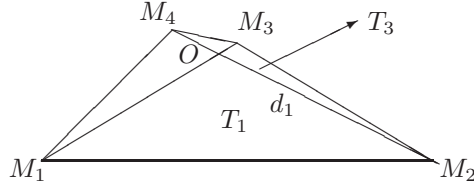


FIG. 2.2. A general convex quadrilateral K .

ratio for $K(1, a, a, a)$. This suggests relaxing the MAC on $\triangle M_2M_3M_4$ because the error on $\triangle M_2M_3M_4$ contributes less compared to that on $\triangle M_2M_1M_4$.

Based on these considerations, we introduce the following geometric condition, which can be regarded as a GRDP condition and in section 3 will be proved to be sufficient for the optimal interpolation error estimate for the \mathcal{Q}_1 Lagrange interpolation.

DEFINITION 2.3. *Let K be a convex quadrilateral (illustrated by Figure 2.2). We say that K satisfies the generalized regular decomposition property with constant $N \in \mathbb{R}_+$ and $0 < \psi < \pi$, or $GRDP(N, \psi)$, if we can divide K into two triangles along one of its diagonals, always called d_1 , in such a way that the big triangle satisfies $MAC(\psi)$ and that*

$$(2.2) \quad \frac{h_K}{|d_1| \sin \alpha} \left(\frac{|T_3|}{|T_1|} \ln \frac{|T_1|}{|T_3|} \right)^{\frac{1}{2}} \leq N,$$

where the big triangle will always be called T_1 , the other is denoted by T_3 , h_K denotes the diameter of the quadrilateral K , and α is the maximal angle of T_3 .

Remark 2.4. Noticing that the term $\frac{|T_3|}{|T_1|} = \frac{|a_3|}{|a_1|}$, where $a_3 = d_2 \cap T_3$ and $a_1 = d_2 \cap T_1$ denote the two parts of the diagonal d_2 divided by the diagonal d_1 , the condition (2.2) can be easily checked in practical computations; in particular, if we choose the longest diagonal for d_1 , the condition (2.2) becomes $\frac{1}{\sin \alpha} \left(\frac{|a_3|}{|a_1|} \ln \frac{|a_1|}{|a_3|} \right)^{\frac{1}{2}} \leq N$ since the big triangle satisfies $MAC(\psi)$. Note that the above constant N is a generic constant and may be different from that in (2.2).

Remark 2.5. It is easy to see that if a quadrilateral K satisfies the RDP condition, then it also satisfies the GRDP condition. However, the converse is not true, as shown by the example $K(1, a, a^s, a)$ with $s > 2$. Note that the elements $K(1, a, a^s, a)$ do not satisfy the RDP condition if $s > 1$, while they satisfy the GRDP condition if and only if $s > 2$.

Remark 2.6. In fact, if one divides the quadrilateral into two triangles by the longest diagonal and if the small triangle is much smaller compared to the big one, then the quadrilateral K is almost degenerated into the big triangle, and the MAC of the small triangle should be relaxed under the control of $\frac{|T_3|}{|T_1|}$ because the error on T_3 contributes little to the interpolation error on the global quadrilateral. This is just our motivation for the presentation of the GRDP condition.

Remark 2.7. Let us finish this section by showing that for some particular examples the condition in Definition 2.3 that the big triangle satisfy the MAC is necessary. Indeed, consider the family of quadrilaterals Q_α of vertices $M_1 = (-1 + \cos \alpha, -\sin \alpha)$, $M_2 = (1, 0)$, $M_3 = (1 - \cos \alpha, \sin \alpha)$, and $M_4 = (-1, 0)$, with the parameter $\alpha \in (\frac{\pi}{2}, \pi)$; see Figure 2.3. Since the angles at M_2 and M_4 are larger than α , any triangle obtained

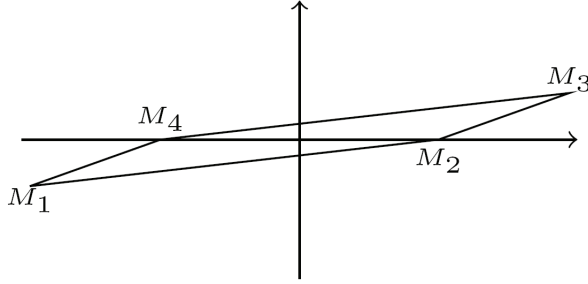


FIG. 2.3. The quadrilateral Q_α .

by subdividing Q_α by one diagonal does not satisfy $\text{MAC}(\psi)$ for some $\psi < \pi$ independent on α . If we consider $u(x, y) = x^2$, we directly see that

$$|u|_{2, Q_\alpha} = 2|Q_\alpha|^{1/2}.$$

On the other hand, by using the affine transformation that maps Q_α into the reference \hat{K} , we check that

$$\begin{aligned} |u - \mathcal{Q}u|_{1, Q_\alpha}^2 &\geq \left| \frac{\partial}{\partial y}(u - \mathcal{Q}u) \right|_{0, Q_\alpha}^2 = \left| \frac{\partial \mathcal{Q}u}{\partial y} \right|_{0, Q_\alpha}^2 \\ &\geq \frac{|Q_\alpha|}{4(\sin \alpha)^2} \int_{\hat{K}} (2 \cos \alpha \eta + 2(1 - \cos \alpha)\xi + (1 - \cos \alpha)^2) d\xi d\eta. \end{aligned}$$

As α goes to π , we see that

$$\int_{\hat{K}} (2 \cos \alpha \eta + 2(1 - \cos \alpha)\xi + (1 - \cos \alpha)^2) d\xi d\eta \rightarrow \int_{\hat{K}} (-2\eta + 4\xi + 4) d\xi d\eta = C^2 > 0.$$

Hence there exists $\beta_0 > 0$ small enough such that for all $\alpha \in (\pi - \beta_0, \pi)$

$$\int_{\hat{K}} (2 \cos \alpha \eta + 2(1 - \cos \alpha)\xi + (1 - \cos \alpha)^2) d\xi d\eta \geq \frac{C^2}{2}.$$

This finally shows that for all $\alpha \in (\pi - \beta_0, \pi)$ one has

$$\frac{|u - \mathcal{Q}u|_{1, Q_\alpha}}{|u|_{2, Q_\alpha}} \geq \frac{C}{4 \sin \alpha},$$

and hence the ratio $\frac{|u - \mathcal{Q}u|_{1, Q_\alpha}}{|u|_{2, Q_\alpha}}$ goes to infinity as α tends to π .

3. Interpolation error estimate in H^1 for \mathcal{Q}_1 elements. In this section, we shall prove the optimal order error estimate for \mathcal{Q}_1 Lagrange elements satisfying the GRDP condition by following the idea from [2] and [20]. Let Π be the conforming P_1 Lagrange interpolation operator on the big triangle T_1 ; i.e., Πu is the linear function which admits the same values with the function u at the three vertices M_1, M_2 , and M_4 . Then we have

$$|u - \mathcal{Q}u|_{1, K} \leq |\Pi u - \mathcal{Q}u|_{1, K} + |u - \Pi u|_{1, K}.$$

Because $\Pi u - \mathcal{Q}u$ belongs to the isoparametric finite element space and vanishes at M_1, M_2 , and M_4 , it follows that

$$(\Pi u - \mathcal{Q}u)(x) = (\Pi u - u)(M_3)\phi_3(x),$$

where ϕ_3 is the basis function corresponding to M_3 . Hence we obtain

$$(3.1) \quad |u - \mathcal{Q}u|_{1,K} \leq |(\Pi u - u)(M_3)| |\phi_3|_{1,K} + |u - \Pi u|_{1,K}.$$

The goal of the rest of this section is to estimate the two terms of the right-hand side of (3.1). We first give an estimate for the term $|\phi_3|_{1,K}$ following the idea developed in [20] and leave the terms $|(\Pi u - u)(M_3)|$ and $|u - \Pi u|_{1,K}$ for the end.

In order to estimate $|\phi_3|_{1,K}$, we start with a new bound for the term $\int_{\hat{K}} \frac{1}{|J|} d\xi d\eta$, where J is the Jacobian of the mapping F_K .

LEMMA 3.1. *Let K be a general convex quadrilateral with consecutive vertices M_1, M_2, M_3 , and M_4 (cf. Figure 2.2). Let θ be the angle of the two diagonals M_1M_3 (denoted by d_2) and M_2M_4 (denoted by d_1), and let O be the point at which they intersect. Let $a_i = |OM_i|$ with $a_i > 0$ for $i = 1, 2, 4$ and $a_3 \geq 0$. Let α, s be the maximal angle and the shortest edge of the triangle T_3 , respectively. Without loss of generality, we can assume that $M_3M_4 = s$. Then we have*

$$(3.2) \quad \int_{\hat{K}} \frac{1}{|J|} d\xi d\eta < \frac{4}{|d_1||s| \sin \alpha} \frac{|T_3|}{|T_1|} \left(2 + \ln \frac{|T_1|}{|T_3|} \right).$$

Proof. Let $(O\tilde{x}, O\tilde{y})$ be two auxiliary axes oriented along the vectors M_1M_3 and M_2M_4 . Let \tilde{J} be the Jacobian of the affine mapping $(\xi, \eta) \rightarrow (\tilde{x}, \tilde{y})$ and let J_1 be the Jacobian of the affine mapping $(\tilde{x}, \tilde{y}) \rightarrow (x, y)$. Then we have $J = \tilde{J}J_1$ with $|J_1| = \sin \theta$. It follows from (2.1) that

$$\begin{cases} \tilde{x} = -(1 - \xi)(1 - \eta)a_1 + \xi\eta a_3, \\ \tilde{y} = -\xi(1 - \eta)a_2 + (1 - \xi)\eta a_4. \end{cases}$$

First we assume that $a_2 \geq a_4$. It is easy to see that

$$\begin{aligned} \tilde{J} &= a_2a_3\xi + a_3a_4\eta + a_1a_4(1 - \xi) + a_1a_2(1 - \eta) \\ &\geq a_2a_3\xi + a_1a_2(1 - \eta) \geq 0. \end{aligned}$$

Then we can derive that

$$\begin{aligned} \int_{\hat{K}} \frac{1}{|\tilde{J}|} d\xi d\eta &\leq \int_{\hat{K}} \frac{1}{a_2a_3\xi + a_1a_2(1 - \eta)} d\xi d\eta \\ &= \frac{1}{a_2a_3} \int_0^1 \ln \left(1 + \frac{a_3}{a_1(1 - \eta)} \right) d\eta \\ &= \frac{1}{a_2a_3} \left(\int_0^{\frac{a_3}{a_1}} + \int_{\frac{a_3}{a_1}}^1 \right) \ln \left(1 + \frac{a_3}{a_1t} \right) dt \\ &< \frac{1}{a_2a_3} \left(\int_0^{\frac{a_3}{a_1}} \sqrt{\frac{a_3}{a_1t}} dt + \int_{\frac{a_3}{a_1}}^1 \frac{a_3}{a_1t} dt \right) \\ &= \frac{1}{a_1a_2} \left(2 - \ln \frac{a_3}{a_1} \right), \end{aligned}$$

where we have used the inequalities $\ln(1+x) < \sqrt{x}$ for all $x \in [1, \infty)$ and $\ln(1+x) < x$ for all $x \in [\frac{a_3}{a_1}, 1]$.

On the other hand, an application of the sin theorem in the triangle M_3OM_4 yields

$$(3.3) \quad \sin \theta = \frac{\sin \angle M_2M_4M_3|s|}{a_3}.$$

Furthermore, it can be easily proved that $\sin \angle M_2M_4M_3 \geq \frac{1}{2} \sin \alpha$. Indeed, if $\alpha = \angle M_2M_4M_3$, the assertion is obvious; otherwise, $\alpha = \angle M_2M_3M_4$, and then

$$\sin \angle M_2M_4M_3 = \sin(\angle M_2M_4M_3 + \angle M_4M_2M_3) \leq 2 \sin \angle M_2M_4M_3$$

because $\angle M_4M_2M_3 \leq \angle M_2M_4M_3$. Therefore,

$$\int_{\widehat{K}} \frac{1}{|J|} d\xi d\eta < \frac{2}{a_2|s| \sin \alpha} \frac{a_3}{a_1} \left(2 - \ln \frac{a_3}{a_1} \right),$$

together with the fact that $\frac{a_3}{a_1} = \frac{|T_3|}{|T_1|}$ and $a_2 \geq \frac{1}{2}|d_1|$, implies (3.2).

In the case $a_2 < a_4$, we just use the inequality $\tilde{J} \geq a_3a_4\eta + a_1a_4(1 - \xi)$ and prove the assertion by the same argument. \square

LEMMA 3.2. *Let K be a general convex quadrilateral with the same hypotheses as Lemma 3.1. Then we have*

$$(3.4) \quad \left| \phi_3 \right|_{1,K} \leq \frac{8h_K}{(|d_1||s| \sin \alpha)^{\frac{1}{2}}} \left(\frac{|T_3|}{|T_1|} \left(2 + \ln \frac{|T_1|}{|T_3|} \right) \right)^{\frac{1}{2}}.$$

Proof. By Lemma 2.2 in [20], we have

$$\left| \phi_3 \right|_{1,K} \leq 4h_K \left(\int_{\widehat{K}} \frac{1}{|J|} d\xi d\eta \right)^{\frac{1}{2}} \left| \widehat{\phi}_3 \right|_{1,\infty,\widehat{K}}.$$

The conclusion follows from Lemma 3.1 and the fact that $|\widehat{\phi}_3|_{1,\infty,\widehat{K}} = 1$. \square

Remark 3.3. As mentioned in [2], the error estimate of the term $|\phi_3|_{1,K}$ is the most technical one. It is estimated therein by introducing an appropriate affine change of variables that reduces the problem to a reference family of elements. Here we did not adopt the technique developed in [2] because we have not imposed the MAC on the small triangle. Meanwhile, the estimate of $|\phi_3|_{1,K}$ in [2] (see Lemma 4.6 of [2]) can be easily recovered under the assumption that $\frac{|d_2|}{|d_1|}$ is bounded.

Remark 3.4. Note that (3.4) gives a sharp estimate of the term $|\phi_3|_{1,K}$ up to a generic constant. In fact, one can just consider the example of the quadrilateral $K(1, b, a, b)$ under the assumption $0 < a, b \ll 1$. Some immediate calculations yield

$$\begin{aligned} \left| \phi_3 \right|_{1,K} &\geq \left\| \frac{\partial \phi_3}{\partial y} \right\|_{0,K} \geq C \frac{1}{\sqrt{b(1-a)}} \left(\ln \frac{1}{a} \right)^{\frac{1}{2}} \\ &\geq C \frac{h_K}{(|d_1||s| \sin \alpha)^{\frac{1}{2}}} \left(\frac{|T_3|}{|T_1|} \left(2 + \ln \frac{|T_1|}{|T_3|} \right) \right)^{\frac{1}{2}} \end{aligned}$$

since $|s| = a$, $\sin \alpha > b$, and $\frac{|T_3|}{|T_1|} = a$.

The next lemma gives an estimate for $|(u - \Pi u)(M_3)|$.

LEMMA 3.5. *Let K be a general convex quadrilateral; then we have*

$$(3.5) \quad |(u - \Pi u)(M_3)| \leq \left(\frac{4|s|}{|d_1| \sin \alpha} \right)^{\frac{1}{2}} \left\{ |u - \Pi u|_{1,T_3} + h_K |u|_{2,T_3} \right\}.$$

Proof. The proof is just the same as that of Lemma 4.2 in [2], which exploited a trace theorem with a sharp dependence of the constant given by [28]; we therefore omit it here. \square

Remark 3.6. The result of Lemma 3.5 gives a sharp estimate up to a generic constant. Consider the example of the quadrilateral $K(1, b, a, b)$ under the assumption $0 < a, b \ll 1$ and the function $u(x, y) = x^2$. We then see that $\Pi u = x$ and therefore

$$|(\Pi u - u)(M_3)| = a(1 - a) \geq C \left(\frac{|s|}{|d_1| \sin \alpha} \right)^{\frac{1}{2}} \left\{ |u - \Pi u|_{1,T_3} + h_K |u|_{2,T_3} \right\}$$

since $|s| = a$, $\sin \alpha > b$, and $|u - \Pi u|_{1,T_3} + h_K |u|_{2,T_3} \leq \sqrt{ab}$.

It remains to bound the term $|u - \Pi u|_{1,K}$. This is the goal of the following lemma.

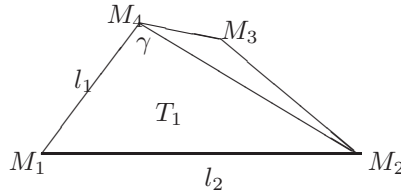


FIG. 3.1.

LEMMA 3.7. *Let K be a general convex quadrilateral and Π be the linear Lagrange interpolation operator defined on T_1 ; then*

$$(3.6) \quad |u - \Pi u|_{1,K} \leq \frac{4}{\sin \gamma} \left(1 + \frac{2}{\pi} \right) \left(\frac{2|K|}{|T_1|} \right)^{\frac{1}{2}} h_K |u|_{2,K},$$

where γ is the maximal angle of T_1 .

Proof. Without loss of generality, we can assume that $\angle M_1 M_3 M_2 = \gamma$ is the maximal angle of T_1 and adopt the notation of Figure 3.1. Let v_1, v_2 be the directions of the edges l_1 and l_2 , respectively. Then we have

$$|u - \Pi u|_{1,K} \leq \frac{1}{\sin \gamma} \left(\|\nabla(u - \Pi u) \cdot v_1\|_{0,K} + \|\nabla(u - \Pi u) \cdot v_2\|_{0,K} \right).$$

Consider $A = \nabla(u - \Pi u) \cdot v_1 = \frac{\partial(u - \Pi u)}{\partial v_1}$, and let A_K be the mean value of A on K . The well-known Poincaré inequality gives

$$(3.7) \quad \|A - A_K\|_{0,K} \leq \frac{h_K}{\pi} |A|_{1,K}.$$

Note that for a general convex domain the constant in the above Poincaré inequality can be taken explicitly and independent of its shape (i.e., it depends only on its diameter). However, the original proof in [27] contains a mistake, and recently [10] gives a corrected proof; fortunately, the optimal constant $\frac{1}{\pi}$ in the Poincaré inequality remains valid.

Now we bound $\|A_K\|_{0,K}$ following an idea from [2]. By the sharp trace theorem of [28], the Cauchy–Schwarz inequality, and the fact that $\int_{l_1} A dv_1 = 0$, we have

$$\begin{aligned} \|A_K\|_{0,K} &= |K|^{\frac{1}{2}} |A_K| = \frac{|K|^{\frac{1}{2}}}{|l_1|} \left| \int_{l_1} (A - A_K) dv_1 \right| \\ &\leq \left(\frac{2|K|}{|T_1|} \right)^{\frac{1}{2}} \left(\|A - A_K\|_{0,K} + h_K |A|_{1,K} \right), \end{aligned}$$

which, together with (3.7), gives

$$\|\nabla(u - \Pi u) \cdot v_1\|_{0,K} \leq \left(1 + \frac{2}{\pi} \right) \left(\frac{2|K|}{|T_1|} \right)^{\frac{1}{2}} h_K |u|_{2,K}.$$

The term $\|\nabla(u - \Pi u) \cdot v_2\|_{0,K}$ is estimated similarly, and the proof of the lemma follows. \square

Collecting all of the above lemmas, we can obtain the main theorem of this section, which gives the optimal error estimate in the H^1 norm for convex quadrilaterals.

THEOREM 3.8. *Let K be a convex quadrilateral satisfying $GRDP(N, \psi)$; then we have*

$$(3.8) \quad |u - \mathcal{Q}u|_{1,K} \leq Ch_K |u|_{2,K},$$

with $C > 0$ depending only on N and ψ .

Proof. A combination of (3.4) and (3.5) yields

$$\begin{aligned} |(\Pi u - u)(M_3)| |\phi_3|_{1,K} &\leq \frac{16h_K}{|d_1| \sin \alpha} \left(\frac{|T_3|}{|T_1|} \left(2 + \ln \frac{|T_1|}{|T_3|} \right) \right)^{\frac{1}{2}} \\ &\quad \times \left\{ |u - \Pi u|_{1,T_3} + h_K |u|_{2,T_3} \right\}, \end{aligned}$$

which, together with (3.1) and (3.6), gives

$$\begin{aligned} |u - \mathcal{Q}u|_{1,K} &\leq \left\{ \frac{32h_K}{|d_1| \sin \alpha} \left(\frac{|T_3|}{|T_1|} \left(2 + \ln \frac{|T_1|}{|T_3|} \right) \right)^{\frac{1}{2}} + 1 \right\} \\ (3.9) \quad &\quad \times \frac{4}{\sin \gamma} \left(1 + \frac{2}{\pi} \left(\frac{2|K|}{|T_1|} \right)^{\frac{1}{2}} \right) h_K |u|_{2,K}. \end{aligned}$$

Since T_1 satisfies the MAC and $|T_1| \geq \frac{1}{2}|K|$, (3.8) follows from the assumption (2.2) and (3.9). \square

Remark 3.9. In fact, (3.9) gives an explicit error bound for the bilinear interpolation operator. If the two divided triangles have comparable areas, i.e., $\frac{|T_3|}{|T_1|} = O(1)$, then the results of [2] can be recovered from (3.9) with the RDP condition. Otherwise, if the quadrilateral is nearly degenerated into the triangle T_1 , i.e., $\frac{|T_3|}{|T_1|} \rightarrow 0$, we can see that the interpolation error of the \mathcal{Q}_1 element is dominated by that of the P_1 Lagrange interpolation operator on T_1 . Obviously, this is a quite reasonable conclusion. This reinforces the fact that the MAC imposed on T_1 cannot be relaxed

because it is also a necessary condition for the optimal interpolation error of the P_1 Lagrange interpolation operator (cf. [6, 9]).

Remark 3.10. One may ask whether the GRDP condition is necessary in the sense that, given a family of elements that does not satisfy the GRDP condition, the interpolation error estimate (3.8) cannot be uniformly bounded. Indeed we have already shown in Remark 2.7 that the MAC on the biggest triangle seems to be necessary. In its full generality we cannot hope to show that (2.2) is also necessary. From Remarks 3.4 and 3.6 we present a family of quadrilaterals that satisfies the MAC but not the condition (2.2) and for which the interpolation error estimate (3.8) is not uniformly satisfied. In that sense our condition (2.2) is almost necessary. For that purpose, we take the example from Remark 3.4 and the function $u(x, y) = x^2$. We then see that $\Pi u(x, y) = x$ and therefore

$$(\Pi u - u)(M_3) = a(1 - a), \quad |u|_{2,K} \leq C\sqrt{b}.$$

By the triangular inequality, we then have

$$|u - \mathcal{Q}u|_{1,K} \geq |(\Pi u - u)(M_3)| |\phi_3|_{1,K} - |u - \Pi u|_{1,K}.$$

Hence by Lemma 3.7, we have

$$\begin{aligned} \frac{|u - \mathcal{Q}u|_{1,K}}{h_K |u|_{2,K}} &\geq \frac{|(\Pi u - u)(M_3)| |\phi_3|_{1,K}}{h_K |u|_{2,K}} - \frac{|u - \Pi u|_{1,K}}{h_K |u|_{2,K}} \\ &\geq C_1 \frac{a\sqrt{|\ln a|}}{b} - C_2 \end{aligned}$$

for some positive constants C_1, C_2 independent on a and b . By choosing $b = a|\ln a|^\alpha$, with $\alpha \geq 0$, the above right-hand side tends to infinity for $a \rightarrow 0$ and therefore (3.8) is not uniformly satisfied. Furthermore, we easily check that this family of quadrilaterals satisfies the MAC but not the condition (2.2). In our proof of (3.8), the point where we make an overestimation is when we use the estimates $\|u - \Pi u\|_{1,T_3} \leq \|u - \Pi u\|_{1,K}$ and $|u|_{2,T_3} \leq |u|_{2,K}$. In many cases (for instance, in the case $a = b^2$), the right-hand sides are much larger than the corresponding left-hand sides.

4. Some technical bounds. Until now, we have studied the interpolation error estimate of the \mathcal{Q}_1 element in the H^1 norm. In this section and the next, we will extend our mesh condition for the error estimate in $W^{1,p}$ for $p \geq 1$. As suggested by section 3, the key point is to estimate $\frac{|\phi_3|_{1,p,K}}{h_K}$. But noticing that $|\frac{\partial \phi_3}{\partial x}| \leq \frac{2h_K}{|J|}$, $|\frac{\partial \phi_3}{\partial y}| \leq \frac{2h_K}{|J|}$, we have

$$(4.1) \quad |\phi_3|_{1,p,K} \leq 4h_K \left(\int_{\hat{K}} \frac{1}{|J|^{p-1}} d\xi d\eta \right)^{\frac{1}{p}},$$

and we are reduced to estimating the quantity

$$\int_{\hat{K}} \frac{1}{|J|^{p-1}} d\xi d\eta.$$

The remainder of this section is devoted to finding an explicit bound of this quantity for different values of p .

Let us start with the case $1 \leq p < 3$.

LEMMA 4.1. *Let K be a general convex quadrilateral; then we have*

$$(4.2) \quad \int_{\widehat{K}} \frac{1}{|J|^{p-1}} d\xi d\eta < \frac{4^{p-1}}{(2-p)(|d_1||s|\sin\alpha)^{p-1}} \left(\frac{|T_3|}{|T_1|}\right)^{p-1} \quad \text{for } p \in [1, 2),$$

$$(4.3) \quad \int_{\widehat{K}} \frac{1}{|J|^{p-1}} d\xi d\eta < \frac{4^{p-1}}{(p-2)(3-p)(|d_1||s|\sin\alpha)^{p-1}} \left(\frac{|T_3|}{|T_1|}\right)^{p-1} \quad \text{for } p \in (2, 3).$$

Proof. We adopt the notation introduced in Lemma 3.1. By direct computations, if $a_2 \geq a_4$, we can derive

$$\begin{aligned} \int_{\widehat{K}} \frac{1}{|\tilde{J}|^{p-1}} d\xi d\eta &\leq \int_{\widehat{K}} \frac{1}{(a_2 a_3 \xi + a_1 a_2 (1-\eta))^{p-1}} d\xi d\eta \\ &= \frac{1}{(2-p)(a_1 a_2)^{p-1}} \int_0^1 \left(\left(1 + \frac{a_3}{a_1} \xi\right)^{2-p} - \xi^{2-p} \right) d\xi \\ &= \frac{1}{(2-p)(3-p)a_1^{p-2} a_2^{p-1} a_3} \left(\left(1 + \frac{a_3}{a_1}\right)^{3-p} - \left(\frac{a_3}{a_1}\right)^{3-p} - 1 \right). \end{aligned}$$

If $p \in [1, 2)$, it can be easily proved that $(1+x)^{3-p} < 1 + (3-p)x + x^{3-p}$ for all $x \in (0, 1]$. Then we have

$$\int_{\widehat{K}} \frac{1}{|\tilde{J}|^{p-1}} d\xi d\eta < \frac{1}{(2-p)a_1^{p-1} a_2^{p-1}}.$$

Recalling that $|J| = |\tilde{J}| \sin\theta$ and (3.3), we get

$$\int_{\widehat{K}} \frac{1}{|J|^{p-1}} d\xi d\eta < \frac{4^{p-1}}{(2-p)(|d_1||s|\sin\alpha)^{p-1}} \left(\frac{a_3}{a_1}\right)^{p-1},$$

which implies (4.2).

Otherwise, if $p \in (2, 3)$, we have

$$\begin{aligned} \int_{\widehat{K}} \frac{1}{|\tilde{J}|^{p-1}} d\xi d\eta &\leq \frac{1}{(2-p)(a_2 a_3)^{p-1}} \int_0^1 \left(\left(1 + \frac{a_1}{a_3}(1-\eta)\right)^{2-p} - \left(\frac{a_1}{a_3}(1-\eta)\right)^{2-p} \right) d\eta \\ &= \frac{1}{(p-2)(3-p)a_3^{p-2} a_2^{p-1} a_1} \left(\left(\frac{a_1}{a_3}\right)^{3-p} + 1 - \left(1 + \frac{a_1}{a_3}\right)^{3-p} \right). \end{aligned}$$

On the other hand, we directly see that

$$(1+x)^{3-p} \geq x^{3-p} \quad \forall x > 0.$$

Then we have

$$\int_{\widehat{K}} \frac{1}{|\tilde{J}|^{p-1}} d\xi d\eta < \frac{1}{(p-2)(3-p)a_3^{p-2} a_2^{p-1} a_1}$$

and

$$\int_{\widehat{K}} \frac{1}{|J|^{p-1}} d\xi d\eta < \frac{4^{p-1}}{(p-2)(3-p)(|d_1||s|\sin\alpha)^{p-1}} \frac{a_3}{a_1},$$

which implies (4.3).

The case $a_2 < a_4$ is treated similarly. The proof is complete. \square

Remark 4.2. Comparing the upper bounds in (4.2), (4.3) with the one from (3.2), we can make the following conclusions: the right-hand sides of (4.2) and (4.3) are not valid for $p = 2$. Hence it is reasonable to have added a log factor in (3.2). Indeed, in the proof of Lemma 4.1, passing to the limit $p \rightarrow 2$ in the estimates of $\int_{\hat{K}} \frac{1}{|\tilde{J}|^{p-1}} d\xi d\eta$, we can show that

$$\begin{aligned} \int_{\hat{K}} \frac{1}{|\tilde{J}|} d\xi d\eta &= \lim_{p \rightarrow 2} \int_{\hat{K}} \frac{1}{|\tilde{J}|^{p-1}} d\xi d\eta \\ &\leq \lim_{p \rightarrow 2} \frac{1}{(p-2)(3-p)a_3^{p-2}a_2^{p-1}a_1} \left(\left(\frac{a_1}{a_3}\right)^{3-p} + 1 - \left(1 + \frac{a_1}{a_3}\right)^{3-p} \right) \\ &= \frac{1}{a_1 a_2} \left(\frac{a_1}{a_3} \ln \frac{a_1}{a_3} - \left(1 + \frac{a_1}{a_3}\right) \ln \left(1 + \frac{a_1}{a_3}\right) \right). \end{aligned}$$

This leads to (2.2) because $\left(\frac{a_1}{a_3} \ln \frac{a_1}{a_3} - \left(1 + \frac{a_1}{a_3}\right) \ln \left(1 + \frac{a_1}{a_3}\right)\right) \leq 2 + \ln \frac{a_1}{a_3}$.

On the other hand, the right-hand side of (4.3) is not valid for the critical point $p = 3$, so it is natural to make another analysis of the bound for $p \geq 3$. This will be discussed below.

Remark 4.3. When p approaches 2, the estimate (4.2) can be improved in order to avoid a blowup. In fact, for $p < 2$, $|\phi_3|_{1,p,K}$ can be bounded directly from (3.2). Indeed, taking into account $\frac{1}{p-1} > 1$ and applying Hölder’s inequality in the right-hand side of (4.1), it holds that

$$\begin{aligned} |\phi_3|_{1,p,K} &\leq 4h_K \left(\int_{\hat{K}} \frac{1}{|\tilde{J}|} d\xi d\eta \right)^{1-\frac{1}{p}} \\ &\leq \frac{4^{1+\frac{1}{p}} h_K}{(|d_1| |s| \sin \alpha)^{1-\frac{1}{p}}} \left(\frac{|T_3|}{|T_1|} \right)^{1-\frac{1}{p}} \left(2 + \ln \frac{|T_1|}{|T_3|} \right)^{1-\frac{1}{p}}. \end{aligned}$$

Let us proceed with the case $p \geq 3$.

LEMMA 4.4. *Let K be a general convex quadrilateral, and let $T_i = \triangle M_{i-1} M_i M_{i+1}$, $i = 1, 2, 3, 4$, with $M_{i\pm 4} = M_i$. Without loss of generality we may assume that $|T_3| = \min_{i=1,3} |T_i|$, $|T_4| = \min_{i=2,4} |T_i|$. Then we have*

$$(4.4) \quad \int_{\hat{K}} \frac{1}{|\tilde{J}|^{p-1}} d\xi d\eta \leq \frac{2^{1+\frac{1}{p}} M^{1-\frac{3}{p}+\frac{1}{2p}}}{|T_1|^{\frac{1}{p}} |T_3|^{1-\frac{2}{p}}} \quad \text{for } p \in \left[3, \frac{7}{2} \right],$$

$$(4.5) \quad \int_{\hat{K}} \frac{1}{|\tilde{J}|^{p-1}} d\xi d\eta \leq \frac{2^{1+\frac{2}{p}} M^{\frac{1}{p}}}{(p-2)^{\frac{1}{p}} |T_1|^{\frac{1}{p}} |T_3|^{1-\frac{2}{p}}} \quad \text{for } p \in \left(\frac{7}{2}, 4 \right],$$

$$(4.6) \quad \int_{\hat{K}} \frac{1}{|\tilde{J}|^{p-1}} d\xi d\eta < \frac{(p-3)^{\frac{1}{p}} 2^{1+\frac{2}{p}} M^{1-\frac{3}{p}}}{(p-2)^{\frac{1}{p}} |T_1|^{\frac{1}{p}} |T_3|^{1-\frac{2}{p}}} \quad \text{for } p > 4,$$

where $M = \max\{1, \frac{|T_3|}{|T_4|}\}$.

Proof. First consider the case $p = 3$. In view of the results in Lemma 3.1, we have

$$J = \tilde{J}J_1 = 2(|T_1| + (|T_2| - |T_1|)\xi + (|T_4| - |T_1|)\eta).$$

Noticing the fact

$$|T_1| + |T_3| = |T_2| + |T_4| = |K|,$$

we can derive that

$$\begin{aligned} \int_{\widehat{K}} \frac{1}{|J|^2} d\xi d\eta &= \frac{1}{4} \int_{\widehat{K}} \frac{1}{(|T_1| + (|T_2| - |T_1|)\xi + (|T_4| - |T_1|)\eta)^2} d\xi d\eta \\ &= \frac{1}{4(|T_1| - |T_2|)} \int_0^1 \left(\frac{1}{|T_2| + (|T_4| - |T_1|)\eta} - \frac{1}{|T_1| + (|T_4| - |T_1|)\eta} \right) d\eta \\ &= \frac{1}{4(|T_1| - |T_2|)(|T_4| - |T_1|)} \ln \frac{|T_1||T_3|}{|T_2||T_4|} \\ &= \frac{1}{4(|T_1||T_3| - |T_2||T_4|)} \ln \frac{|T_1||T_3|}{|T_2||T_4|}. \end{aligned}$$

Now let us discuss the above result. If $|T_3| \leq |T_4|$, then we have

$$(4.7) \quad |T_1| \geq |T_2| \quad \text{and} \quad |T_1||T_3| \leq |T_2||T_4|,$$

which implies

$$(4.8) \quad \int_{\widehat{K}} \frac{1}{|J|^2} d\xi d\eta \leq \frac{1}{4|T_1||T_3|},$$

where we have used the inequality $\frac{\ln x}{1-\frac{x}{2}} \leq 1$ for all $x \in (0, 1]$.

On the other hand, if $|T_3| > |T_4|$, then

$$(4.9) \quad |T_2| > |T_1| \quad \text{and} \quad |T_1||T_3| > |T_2||T_4|;$$

thus we can derive

$$\begin{aligned} \int_{\widehat{K}} \frac{1}{|J|^2} d\xi d\eta &= \frac{1}{4(|T_1||T_3| - |T_2||T_4|)} \ln \left(1 + \frac{|T_1||T_3| - |T_2||T_4|}{|T_2||T_4|} \right) \\ &\leq \frac{1}{4(\sqrt{|T_1||T_3|} + \sqrt{|T_2||T_4|})\sqrt{|T_2||T_4|}} \\ (4.10) \quad &< \frac{1}{4\sqrt{|T_1||T_3||T_2||T_4|}} < \frac{1}{4|T_1||T_3|} \left(\frac{|T_3|}{|T_4|} \right)^{\frac{1}{2}}, \end{aligned}$$

where we have used the inequality $\ln(1+x) \leq \sqrt{x}$ for all $x \in (0, \infty)$.

Then a combination of (4.8), (4.10), and (4.1) gives (4.4) for the case $p = 3$.

For $p \in (3, \frac{7}{2}]$, we write

$$\begin{aligned} \int_{\widehat{K}} \frac{1}{|J|^{p-1}} d\xi d\eta &= \int_{\widehat{K}} \frac{1}{|J|^{p-3}} \frac{1}{|J|^2} d\xi d\eta \\ &\leq \left(\frac{M}{2|T_3|} \right)^{p-3} \int_{\widehat{K}} \frac{1}{|J|^2} d\xi d\eta, \end{aligned}$$

where $M = \max\{1, \frac{|T_3|}{|T_4|}\}$, since we notice that

$$\min_{\widehat{K}} |J| = 2 \min\{|T_1|, |T_2|, |T_3|, |T_4|\} = 2 \min\{|T_3|, |T_4|\} = \frac{2|T_3|}{M}.$$

By the estimate (4.8) and (4.10), we deduce that

$$\int_{\widehat{K}} \frac{1}{|J|^{p-1}} d\xi d\eta \leq \left(\frac{M}{2|T_3|}\right)^{p-3} \frac{M^{\frac{1}{2}}}{4|T_1||T_3|}.$$

This estimate yields

$$\left|\phi_3\right|_{1,p,K} \leq \frac{2^{1+\frac{1}{p}} h_K}{|T_1|^{\frac{1}{p}} |T_3|^{1-\frac{2}{p}}} M^{1-\frac{3}{p}+\frac{1}{2p}}.$$

Now we come to the case $\frac{7}{2} < p$. An immediate computation gives that

$$\begin{aligned} \int_{\widehat{K}} \frac{1}{|J|^{p-1}} d\xi d\eta &= \frac{1}{(2-p)2^{p-1}(|T_2| - |T_1|)} \left(\int_0^1 \left(\frac{1}{|T_2| + (|T_4| - |T_1|)\eta} \right)^{2-p} d\eta \right. \\ &\quad \left. - \int_0^1 \left(\frac{1}{|T_1| + (|T_4| - |T_1|)\eta} \right)^{2-p} d\eta \right) \\ (4.11) \quad &= \frac{|T_1|^{3-p} + |T_3|^{3-p} - |T_2|^{3-p} - |T_4|^{3-p}}{(p-2)(p-3)2^{p-1}(|T_1| - |T_2|)(|T_1| - |T_4|)}. \end{aligned}$$

If $|T_3| > |T_4|$, using (4.9) and the fact that $|T_2| - |T_1| = |T_3| - |T_4|$, and by applying Cauchy's mean value theorem repeatedly, we have

$$\begin{aligned} &|T_1|^{3-p} + |T_3|^{3-p} - |T_2|^{3-p} - |T_4|^{3-p} \\ &= \frac{(|T_2||T_4|)^{p-3}(|T_1|^{p-3} + |T_3|^{p-3}) - (|T_1||T_3|)^{p-3}(|T_2|^{p-3} + |T_4|^{p-3})}{|T_1|^{p-3}|T_2|^{p-3}|T_3|^{p-3}|T_4|^{p-3}} \\ &= \frac{(|T_2||T_4|)^{p-3}(|T_1|^{p-3} + |T_3|^{p-3} - |T_2|^{p-3} - |T_4|^{p-3})}{|T_1|^{p-3}|T_2|^{p-3}|T_3|^{p-3}|T_4|^{p-3}} \\ &\quad + \frac{((|T_2||T_4|)^{p-3} - (|T_1||T_3|)^{p-3})(|T_2|^{p-3} + |T_4|^{p-3})}{|T_1|^{p-3}|T_2|^{p-3}|T_3|^{p-3}|T_4|^{p-3}} \\ &= \frac{(p-3)(|T_2||T_4|)^{p-3}(m_{34}^{p-4} - m_{12}^{p-4})(|T_2| - |T_1|)}{|T_1|^{p-3}|T_2|^{p-3}|T_3|^{p-3}|T_4|^{p-3}} \\ &\quad + \frac{(p-3)(|T_2||T_4| - |T_1||T_3|)m_{1324}^{p-4}(|T_2|^{p-3} + |T_4|^{p-3})}{|T_1|^{p-3}|T_2|^{p-3}|T_3|^{p-3}|T_4|^{p-3}} \\ &= \frac{(p-3)(p-4)(|T_2||T_4|)^{p-3}m_{1234}^{p-5}(m_{12} - m_{34})(|T_1| - |T_2|)}{|T_1|^{p-3}|T_2|^{p-3}|T_3|^{p-3}|T_4|^{p-3}} \\ (4.12) \quad &\quad + \frac{(p-3)(|T_1| - |T_2|)(|T_1| - |T_4|)m_{1324}^{p-4}(|T_2|^{p-3} + |T_4|^{p-3})}{|T_1|^{p-3}|T_2|^{p-3}|T_3|^{p-3}|T_4|^{p-3}}; \end{aligned}$$

here the constants $m_{12}, m_{34}, m_{1324}, m_{1234}$ are constants produced by the mean value theorem that satisfy $|T_1| \leq m_{12} \leq |T_2|, |T_4| \leq m_{34} \leq |T_3|, |T_2||T_4| \leq m_{1324} \leq |T_1||T_3|,$ and $|T_{34}| \leq m_{1234} \leq |T_{12}|.$

Since, for $\frac{7}{2} < p \leq 4,$ we have

$$(p - 4)(|T_2||T_4|)^{p-3}m_{1234}^{p-5}(m_{12} - m_{34}) \leq 0$$

and

$$m_{1324}^{p-4} \leq (|T_2||T_4|)^{p-4},$$

by (4.11) and (4.12), it holds that

$$\begin{aligned} \int_{\hat{K}} \frac{1}{|J|^{p-1}} d\xi d\eta &\leq \frac{|T_4|^{p-4}|T_2|^{2p-7}}{(p - 2)2^{p-2}|T_1|^{p-3}|T_2|^{p-3}|T_3|^{p-3}|T_4|^{p-3}} \\ (4.13) \qquad \qquad \qquad &= \frac{1}{(p - 2)2^{p-2}|T_1||T_3|^{p-2}} \frac{|T_3|}{|T_4|}. \end{aligned}$$

When $p > 4,$ further noticing that

$$\begin{aligned} |T_4|^{p-3}m_{1234}^{p-5} &\leq \begin{cases} |T_4|^{2p-8}, & 4 < p \leq 5, \\ |T_2|^{p-5}|T_4|^{p-3}, & p > 5, \end{cases} \\ &\leq (|T_1||T_3|)^{p-4} \end{aligned}$$

and

$$m_{12} - m_{34} \leq |T_2| - |T_4| \leq 2(|T_1| - |T_4|),$$

we can derive

$$\begin{aligned} \int_{\hat{K}} \frac{1}{|J|^{p-1}} d\xi d\eta &\leq \frac{(p - 3)(|T_1||T_3|)^{p-4}|T_2|^{p-3}}{(p - 2)2^{p-2}|T_1|^{p-3}|T_2|^{p-3}|T_3|^{p-3}|T_4|^{p-3}} \\ (4.14) \qquad \qquad \qquad &\leq \frac{(p - 3)}{(p - 2)2^{p-2}|T_1||T_3|^{p-2}} \left(\frac{|T_3|}{|T_4|} \right)^{p-3}. \end{aligned}$$

In the case $|T_3| \leq |T_4|,$ we proved similarly that (4.13) and (4.14) hold by (4.7). This yields (4.5) and (4.6). \square

Remark 4.5. The technique developed in Lemma 4.4 renders the continuity of the upper bounds at the turning points $p = 3, \frac{7}{2}, 4.$

5. Interpolation error estimates in $W^{1,p}.$ According to the bounds from the previous section, we can make the following definition.

DEFINITION 5.1. *Let K be a convex quadrilateral. We say that K satisfies the new RDP with constants $N \in \mathbb{R}_+, 0 < \psi < \pi,$ and $p \in [1, \infty) \setminus \{2\},$ or NRDP(N, ψ, p), if we can divide K into two triangles along one of its diagonals, always called $d_1,$ in such a way that the big triangle satisfies MAC(ψ) and that*

$$(5.1) \qquad \frac{h_K}{(2 - p)^{\frac{1}{p}}|d_1| \sin \alpha} \left(\frac{|T_3|}{|T_1|} \right)^{1 - \frac{1}{p}} \leq N \quad \text{for } p \in [1, 2),$$

$$(5.2) \quad \frac{h_K}{(p-2)^{\frac{1}{p}}(3-p)^{\frac{1}{p}}|d_1|\sin\alpha} \left(\frac{|T_3|}{|T_1|}\right)^{\frac{1}{p}} \leq N \quad \text{for } p \in (2, 3),$$

$$(5.3) \quad \frac{h_K M^{1-\frac{3}{p}+\frac{1}{2p}}}{|d_1|\sin\alpha} \left(\frac{|T_3|}{|T_1|}\right)^{\frac{1}{p}} \leq N \quad \text{for } p \in \left[3, \frac{7}{2}\right],$$

$$(5.4) \quad \frac{h_K M^{\frac{1}{p}}}{|d_1|\sin\alpha} \left(\frac{|T_3|}{|T_1|}\right)^{\frac{1}{p}} \leq N \quad \text{for } p \in \left(\frac{7}{2}, 4\right],$$

$$(5.5) \quad \frac{h_K M^{1-\frac{3}{p}}}{|d_1|\sin\alpha} \left(\frac{|T_3|}{|T_1|}\right)^{\frac{1}{p}} \leq N \quad \text{for } p > 4,$$

where the big triangle will always be called T_1 , the other is called T_3 , h_K denotes the diameter of the quadrilateral K , α denotes the maximal angle of T_3 , and s denotes the smallest edge of T_3 .

Remark 5.2. In view of Remark 4.3, when $p < 2$, the condition (5.1) could be replaced by

$$\frac{h_K}{|d_1|\sin\alpha} \left(\frac{|T_3|}{|T_1|} \ln \frac{|T_1|}{|T_3|}\right)^{1-\frac{1}{p}} \leq N.$$

This condition is more advantageous than (5.1) for p close to 2, while it is the converse for p far from 2.

We first state an estimate for $|(u - \Pi u)(M_3)|$, which follows from the proof of Lemma 5.2 of [4].

LEMMA 5.3. *Let K be a general convex quadrilateral and Π be the linear Lagrange interpolation operator defined on T_1 ; then*

$$(5.6) \quad |(u - \Pi u)(M_3)| \leq \left(\frac{2^p |s|^{p-1}}{|d_1|\sin\alpha}\right)^{\frac{1}{p}} \left\{ |u - \Pi u|_{1,p,T_3} + h_K |u|_{2,p,T_3} \right\}, \quad p \geq 1.$$

Remark 5.4. Since K is convex, it is well known that the Poincaré inequality holds for general $p \geq 1$; i.e., there exists a constant C_p depending only on p such that

$$(5.7) \quad \|v\|_{0,p,K} \leq C_p h |v|_{1,p,K}$$

for any $v \in W^{1,p}(K)$ with vanishing average on K .

Now we are in a position to bound the term $|u - \Pi u|_{1,p,K}$.

LEMMA 5.5. *Let K be a general convex quadrilateral and Π be the linear Lagrange interpolation operator defined on T_1 ; then*

$$|u - \Pi u|_{1,p,K} \leq (1 + 2C_p) \frac{2^{2-\frac{1}{p}}}{\sin\gamma} \left(\frac{|K|}{|T_1|}\right)^{\frac{1}{p}} h_K |u|_{2,p,K},$$

where γ is the maximal angle of T_1 .

Proof. The proof is just a combination of the arguments of Lemma 3.7 in section 3 (using here (5.7)) and of Lemma 5.3 of [4]. \square

Now we come to the main theorem of this section.

THEOREM 5.6. *Let K be a convex quadrilateral satisfying NRDP(N, ψ, p) with $p \in [1, \infty) \setminus \{2\}$; then we have*

$$(5.8) \quad |u - \mathcal{Q}u|_{1,p,K} \leq Ch_K |u|_{2,p,K},$$

with $C > 0$ depending only on N, p , and ψ .

Proof. A combination of (4.1), Definition 5.1, and (5.6) yields

$$|(\Pi u - u)(M_3)| |\phi_3|_{1,p,K} \leq C_p N \left\{ |u - \Pi u|_{1,p,T_3} + h_K |u|_{2,p,T_3} \right\},$$

where $C_p > 0$ depends only on p .

Now invoking Lemma 5.5, we get

$$\begin{aligned} |u - \mathcal{Q}u|_{1,p,K} &\leq |(\Pi u - u)(M_3)| |\phi_3|_{1,p,K} + |u - \Pi u|_{1,p,K} \\ &\leq \kappa(K, p) C'_p h_K |u|_{2,p,K}, \end{aligned}$$

where $C'_p > 0$ depends only on p and $\kappa(K, p)$ is defined by

$$\kappa(K, p) = N + (1 + C_p N) \frac{1}{\sin \gamma} \left(\frac{|K|}{|T_1|} \right)^{\frac{1}{p}}.$$

Hence we will obtain (5.8) if $\kappa(K, p)$ is bounded uniformly. By the definition of T_1 implying that $|K| \leq 2|T_1|$, we have

$$\left(\frac{|K|}{|T_1|} \right)^{\frac{1}{p}} \leq 2^p.$$

We therefore conclude by using the assumption on γ from Definition 5.1. □

Let us finish our paper by comparing our geometrical conditions with the *double angle condition* (DAC) introduced in [4], which we recall here.

DEFINITION 5.7. *Let K be a convex quadrilateral. We say that K satisfies the DAC with constants ψ_m, ψ_M , or $DAC(\psi_m, \psi_M)$, if the interior angles ω of K verify $0 < \psi_m \leq \omega \leq \psi_M < \pi$.*

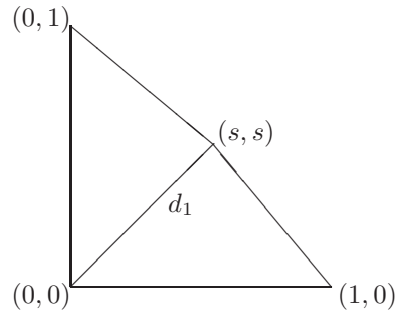
Obviously, the above $DAC(\psi_m, \psi_M)$ is equivalent to (1.2). In [4], the authors proved that it is a sufficient condition for the optimal error estimate in $W^{1,p}$ with $p \geq 3$ and showed that the restriction on the maximal angle cannot be relaxed by some counterexamples. In fact, the $DAC(\psi_m, \psi_M)$ is a quite strong geometric condition, and the following elementary implications hold:

$$DAC(\psi_m, \psi_M) \implies MAC(\psi_M) \implies RDP(N, \psi_M) \implies GRDP(N, \psi_M).$$

One can see that the above NRDP condition for $p \geq 3$ is much weaker than the DAC. In fact, if the quadrilateral K satisfies $DAC(\psi_m, \psi_M)$, it can be easily proved that $\frac{|T_3|}{|T_4|} \leq C(\psi_m, \psi_M)$ and $d_1 = O(h_K)$; hence there exists a constant $N = N(\psi_m, \psi_M)$ such that K satisfies $NRDP(N(\psi_m, \psi_M), \psi_M, p)$ for all $p \geq 3$.

If K satisfies $\frac{|T_3|}{|T_4|} \leq C$ which is satisfied in many cases, the above NRDP condition for $p \geq 3$ is even weaker than the RDP condition. However, there are some examples such that K satisfies the RDP condition but not the above NRDP condition for $p \geq 3$, e.g., $K(1, 1, s, s)$ with $s \rightarrow \frac{1}{2}$ (cf. Figure 5.1), which is employed as a counterexample in [4]. In this sense we may say that our NRDP conditions for $p \geq 3$ are as weak as the RDP condition.

If K satisfies $\frac{|T_3|}{|T_4|} \rightarrow \infty$, the quadrilateral K is almost degenerated into the triangle T_2 . In such a case, the constant of the interpolation error in the $W^{1,p}$ norm ($p > 3$) may not be bounded even if K satisfies the RDP condition. This can be partly

FIG. 5.1. $K(1, 1, s, s)$.

interpreted by the fact that $\max\{1, \frac{|T_3|}{|T_4|}\}$ appears as a factor in our NRDP condition. Taking $K(1, 1, s, s)$ with $s \rightarrow \frac{1}{2}$ (see Figure 5.1) as an example, if one chooses d_1 to divide K into two triangles, K does satisfy the RDP condition, but (5.8) does not hold because $\frac{|T_3|}{|T_4|} \rightarrow \infty$, which violates the NRDP condition for $p \geq 3$.

6. Conclusion. Interpolation error estimates of the finite elements play an important role in the finite element literature. In this paper, we have introduced a generalized RDP (GRDP) condition for $p = 2$ and new RDP (NRDP) conditions for $p \neq 2$, which permit us to prove some interpolation error estimates in the $W^{1,p}$ norm with $1 \leq p < \infty$ for the \mathcal{Q}_1 isoparametric finite elements. As far as we know, our NRDP conditions presented here are weaker than any other mesh conditions proposed in the literature for the same p with $1 \leq p < \infty$.

The results of this paper are valid only for bilinear elements, and it seems difficult to extend them to higher order Lagrange quadrilateral elements or other quadrilateral elements, e.g., mixed elements and nonconforming elements. Some preliminary results for some mixed elements and nonconforming elements are already obtained and will be investigated in our future work.

Acknowledgments. We warmly thank the referees for their helpful remarks that allowed us to considerably improve the presentation of our results. We also would like to thank Professor Thomas Apel, Professor Roland Becker, and Professor Ricardo Durán for some useful discussions.

REFERENCES

- [1] G. ACOSTA AND R. G. DURÁN, *The maximum angle condition for mixed and nonconforming elements: Application to the Stokes equations*, SIAM. J. Numer. Anal., 37 (1999), pp. 18–36.
- [2] G. ACOSTA AND R. G. DURÁN, *Error estimates for \mathcal{Q}_1 isoparametric elements satisfying a weak angle condition*, SIAM. J. Numer. Anal., 38 (2000), pp. 1073–1088.
- [3] G. ACOSTA AND R. G. DURÁN, *An optimal Poincaré inequality in L_1 for convex domains*, Proc. Amer. Math. Soc., 132 (2004), pp. 195–202.
- [4] G. ACOSTA AND G. MONZÓN, *Interpolation error estimates in $W^{1,p}$ for degenerate \mathcal{Q}_1 isoparametric elements*, Numer. Math., 104 (2006), pp. 129–150.
- [5] T. APEL, *Anisotropic interpolation error estimates for isoparametric quadrilateral finite elements*, Computing, 60 (1998), pp. 157–174.
- [6] T. APEL, *Anisotropic Finite Elements: Local Estimates and Applications*, Adv. Numer. Math., B. G. Teubner, Stuttgart, Germany, 1999.
- [7] T. APEL AND S. NICAISE, *The finite element method with anisotropic mesh grading for elliptic problems in domains with corners and edges*, Math. Methods Appl. Sci., 21 (1998), pp. 519–549.

- [8] T. APEL, S. NICAISE, AND J. SCHÖBERL, *Crouzeix-Raviart type finite elements on anisotropic meshes*, Numer. Math., 89 (2001), pp. 193–223.
- [9] I. BABUŠKA AND A. K. AZIZ, *On the angle condition in the finite element method*, SIAM J. Numer. Anal., 13 (1976), pp. 214–226.
- [10] M. BEBENDORF, *A note on the Poincaré inequality for convex domains*, Z. Anal. Anwendungen, 22 (2003), pp. 751–756.
- [11] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.
- [12] W. CAO, *On the error of linear interpolation and the orientation, aspect ratio, and internal angles of a triangle*, SIAM. J. Numer. Anal., 43 (2005), pp. 19–40.
- [13] S. C. CHEN, D. Y. SHI, AND Y. C. ZHAO, *Anisotropic interpolation and quasi-Wilson element for narrow quadrilateral meshes*, IMA J. Numer. Anal., 24 (2004), pp. 77–95.
- [14] P. G. CIARLET AND P. A. RAVIART, *Interpolation theory over curved elements, with applications to finite element methods*, Comput. Methods Appl. Mech. Engrg., 1 (1972), pp. 217–249.
- [15] R. G. DURÁN, *Error estimates for narrow 3D finite elements*, Math. Comp., 68 (1999), pp. 187–199.
- [16] R. G. DURÁN, *Error estimates for anisotropic finite elements and applications*, in Proceedings of the International Congress of Mathematicians, European Mathematical Society, Zürich, 2006, pp. 1181–1200.
- [17] R. G. DURÁN AND A. L. LOMBARDI, *Error estimates on anisotropic Q_1 elements for functions in weighted Sobolev spaces*, Math. Comp., 74 (2005), pp. 1679–1706.
- [18] L. FORMAGGIA AND S. PEROTTO, *New anisotropic a priori error estimates*, Numer. Math., 89 (2001), pp. 641–667.
- [19] P. JAMET, *Estimations d’erreur pour des éléments finis droits presque dégénérés*, RAIRO Anal. Numér., 10 (1976), pp. 43–61.
- [20] P. JAMET, *Estimation of the interpolation error for quadrilateral finite elements which can degenerate into triangles*, SIAM J. Numer. Anal., 14 (1977), pp. 925–930.
- [21] M. KŘÍŽEK, *On the maximal angle condition for linear tetrahedral elements*, SIAM J. Numer. Anal., 29 (1992), pp. 513–520.
- [22] S. P. MAO AND S. C. CHEN, *Accuracy analysis of Adini’s non-conforming plate element on anisotropic meshes*, Comm. Numer. Methods Engrg., 22 (2006), pp. 433–440.
- [23] S. P. MAO, S. C. CHEN, AND H. X. SUN, *A quadrilateral, anisotropic, superconvergent nonconforming double set parameter element*, Appl. Numer. Math., 27 (2006), pp. 937–961.
- [24] S. P. MAO AND Z.-C. SHI, *Nonconforming rotated Q_1 element on non-tensor product degenerate meshes*, Sci. China Ser. A, 49 (2006), pp. 1363–1375.
- [25] P. MING AND Z.-C. SHI, *Quadrilateral mesh revisited*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 5671–5682.
- [26] P. MING AND Z.-C. SHI, *Quadrilateral mesh*, Chinese Ann. Math. Ser. B., 23 (2002), pp. 235–252.
- [27] L. E. PAYNE AND H. F. WEINBERGER, *An optimal Poincaré inequality for convex domains*, Arch. Rational Mech. Anal., 5 (1960), pp. 286–292.
- [28] R. VERFÜRTH, *Error estimates for some quasi-interpolation operator*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 695–713.
- [29] A. ŽENISEK AND M. VANMAELE, *The interpolation theorem for narrow quadrilateral isoparametric finite elements*, Numer. Math., 72 (1995), pp. 123–141.
- [30] A. ŽENISEK AND M. VANMAELE, *Applicability of the Bramble-Hilbert lemma in interpolation problems of narrow quadrilateral isoparametric finite elements*, J. Comput. Appl. Math., 65 (1995), pp. 109–122.
- [31] Z. ZHANG, *Polynomial preserving gradient recovery and a posteriori estimate for bilinear element on irregular quadrilaterals*, Int. J. Numer. Anal. Model., 1 (2004), pp. 1–24.
- [32] J. ZHANG AND F. KIKUCHI, *Interpolation error estimates of a modified 8-node serendipity finite element*, Numer. Math., 85 (2000), pp. 503–524.

REHABILITATION OF THE LOWEST-ORDER RAVIART–THOMAS ELEMENT ON QUADRILATERAL GRIDS*

PAVEL B. BOCHEV[†] AND DENIS RIDZAL[‡]

Abstract. A recent study [D. N. Arnold, D. Boffi, and R. S. Falk, *SIAM J. Numer. Anal.*, 42 (2005), pp. 2429–2451] reveals that convergence of finite element methods using $H(\operatorname{div}, \Omega)$ -compatible finite element spaces deteriorates on nonaffine quadrilateral grids. This phenomena is particularly troublesome for the lowest-order Raviart–Thomas elements, because it implies loss of convergence in some norms for finite element solutions of mixed and least-squares methods. In this paper we propose reformulation of finite element methods, based on the natural mimetic divergence operator [M. Shashkov, *Conservative Finite Difference Methods on General Grids*, CRC Press, Boca Raton, FL, 1996], which restores the order of convergence. Reformulations of mixed Galerkin and least-squares methods for the Darcy equation illustrate our approach. We prove that reformulated methods converge optimally with respect to a norm involving the mimetic divergence operator. Furthermore, we prove that standard and reformulated versions of the mixed Galerkin method lead to *identical* linear systems, but the two versions of the least-squares method are veritably different. The surprising conclusion is that the degradation of convergence in the mixed method on nonaffine quadrilateral grids is superficial, and that the lowest-order Raviart–Thomas elements are safe to use in this method. However, the breakdown in the least-squares method is real, and there one should use our proposed reformulation.

Key words. Raviart–Thomas, quadrilateral, mixed methods, least-squares methods, mimetic methods

AMS subject classifications. 65F10, 65F30, 78A30

DOI. 10.1137/070704265

1. Introduction. We consider finite element solution of the elliptic boundary value problem

$$(1.1) \quad \begin{cases} \nabla \cdot \mathbf{u} + \sigma \Theta_0 p = f, & \text{in } \Omega \quad \text{and} \quad p = 0 \quad \text{on } \Gamma_D, \\ \nabla p + \Theta_1^{-1} \mathbf{u} = 0, & \mathbf{n} \cdot \mathbf{u} = 0 \quad \text{on } \Gamma_N, \end{cases}$$

where $\Omega \subset \mathbb{R}^2$ has a Lipschitz-continuous boundary $\partial\Omega = \Gamma_D \cup \Gamma_N$, \mathbf{n} is the unit outward normal to $\partial\Omega$, Θ_1 is a symmetric tensor, Θ_0 is a real valued function, and σ is a nondimensional parameter that is either 0 or 1. Regarding Θ_1 and Θ_0 , we will assume that there exists a constant $\alpha > 0$ such that for every $\mathbf{x} \in \Omega$ and $\boldsymbol{\xi} \in \mathbb{R}^2$,

$$(1.2) \quad \frac{1}{\alpha} \boldsymbol{\xi}^T \boldsymbol{\xi} \leq \boldsymbol{\xi}^T \Theta_1(\mathbf{x}) \boldsymbol{\xi} \leq \alpha \boldsymbol{\xi}^T \boldsymbol{\xi} \quad \text{and} \quad \frac{1}{\alpha} \leq \Theta_0(\mathbf{x}) \leq \alpha.$$

The equations (1.1) are often called the Darcy problem and provide a simplified model of a single phase flow in porous media. In this context, p is the pressure, \mathbf{u} is the

*Received by the editors October 1, 2007; accepted for publication (in revised form) July 2, 2008; published electronically December 19, 2008. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94-AL85000. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

<http://www.siam.org/journals/sinum/47-1/70426.html>

[†]Applied Mathematics and Applications, Sandia National Laboratories, P. O. Box 5800, MS 1320, Albuquerque, NM 87185-1320 (pbboche@sandia.gov).

[‡]Optimization and Uncertainty Quantification, Sandia National Laboratories, P. O. Box 5800, MS 1320, Albuquerque, NM 87185-1320 (dridzal@sandia.gov).

Darcy velocity, and Θ_1 is the permeability tensor divided by the viscosity. The use of this first-order system as a basis for a finite element method stems from the fact that in porous media flow the vector variable \mathbf{u} is more important than the pressure p . In such cases numerical methods that compute accurate, locally conservative velocity approximations are favored.

Two such methods are the mixed Galerkin method [11] and the locally conservative least-squares method [7, 8, 13]. The main focus of this paper will be on implementations of these two methods with the lowest-order quadrilateral Raviart–Thomas elements (RT_0) [19, 11]. Several reasons motivate our interest in these elements. Quadrilateral grids are widely used in the petroleum industry for porous media flow simulations and there are connections between conservative finite difference methods for (1.1) and mixed methods implemented with the lowest-order $H(\text{div}, \Omega)$ -compatible spaces; see [1, 5, 22] and the references therein. Our study is also prompted by the recent work of Arnold, Boffi, and Falk [4], whose paper asserts that the accuracy of $H(\text{div}, \Omega)$ -conforming finite element spaces deteriorates on nonaffine quadrilateral grids, which in turn leads to reduced orders of convergence in finite element methods. Arnold, Boffi, and Falk [4] support this assertion by examples that show reduced convergence in $H(\text{div}, \Omega)$ of the vector variable in the mixed method, and examples which suggest that in the least-squares method loss of accuracy also spreads to pressure approximations.

These examples are particularly damning for low-order elements because for them the degradation of accuracy in the methods takes the form of a total loss of convergence in some norms for one or both variables. The main goal of this paper is to restore confidence in RT_0 elements and show that with some simple modifications in the finite element methods they can be safely used on general, shape-regular, but not necessarily affine quadrilateral grids.

The proposed reformulation of the mixed and least-squares methods is motivated by mimetic finite difference methods [20]. A mimetic discretization of (1.1) uses the so-called *natural mimetic* divergence, DIV , and *derived* gradient, $\overline{\text{GRAD}}$, operators; see [15, 16]. Of particular interest to us is DIV , which is constructed using the coordinate-invariant definition [2, p. 188]

$$(1.3) \quad \nabla \cdot \mathbf{u}(\mathbf{x}) = \lim_{\kappa \ni \mathbf{x}; \mu(\kappa) \rightarrow 0} \frac{\int_{\partial \kappa} \mathbf{u} \cdot \mathbf{n} \, dS}{\mu(\kappa)}$$

of the divergence operator.¹ The result is a discrete operator² that maps face-based values (the fluxes of \mathbf{u}) onto cell-based constants. Because DIV acts on the same set of degrees of freedom as used to define the lowest-order Raviart–Thomas space, its action can be extended to that space in a natural way. This is the key to our *mimetic* reformulation of finite element methods, in which the main idea is to

¹In this definition κ is a bounded region and $\mu(\kappa)$ denotes its measure. The mimetic approximation of $\nabla \cdot \mathbf{u}$ on an element κ , belonging to a finite element partition \mathcal{T}_h of Ω , is defined by the right-hand side in this formula, assuming that \mathbf{u} and \mathbf{n} are constant on the faces of κ .

²For brevity we call this operator “natural divergence.”

replace³ the analytic divergence $\nabla \cdot$ by the natural divergence DIV .

A somewhat unexpected byproduct of our analysis is a theorem which shows that the mimetic reformulation of the mixed method is actually equivalent to its standard version, in the sense that the two methods generate identical linear algebraic systems with identical solutions. Since in the mimetic reformulation $\text{DIV}(\mathbf{u}^h)$ converges to the divergence of the exact solution, it follows that the same must be true for the solution of the standard mixed method. In other words, the flux degrees of freedom in the mixed Galerkin solution do contain accurate information about the divergence of the exact solution. The reason $\nabla \cdot$ fails to recover this information on nonaffine quads is that it acts on the flux data *indirectly* via basis functions defined by the Piola transform, which makes the result dependent upon the element shape.⁴ In contrast, DIV is able to always recover accurate divergence approximation because it acts *directly* on the flux degrees of freedom, which makes its action independent of the element shape. It follows that the loss of convergence in the mixed method is superficial and that this method can be safely used on nonaffine quadrilateral grids.

Unlike the mixed method, mimetic reformulation of the least-squares method turns out to be veritably different from its standard finite element realization, and the loss of convergence in this method, reported in [4], is genuine. We refine the conclusions of [4] by showing that for Darcy problems that include a “reaction” term ($\sigma = 1$) the loss of accuracy does not spread to the pressure approximation. However, the “information content” of the velocity approximation is ruined, and using DIV in lieu of $\nabla \cdot$ to extract divergence information does not help much. Thus, the breakdown in the least-squares method is real, and for general quadrilateral grids one should use our proposed reformulation.

This paper is organized as follows. Section 2 reviews notation and definitions of finite element spaces. Section 3 discusses the natural divergence operator, its properties, and extension to the lowest-order Raviart–Thomas elements. Section 4 presents mimetic reformulations of mixed and least-squares methods. Section 5 contains analyses of these methods. Numerical results are collected in section 6.

2. Notation and quotation of results. For $p > 0$, $H^p(\Omega)$ denotes the Sobolev space of order p with norm and inner product denoted by $\|\cdot\|_p$ and $(\cdot, \cdot)_p$, respectively. When $p = 0$, we use the standard notation $L^2(\Omega)$. The symbol $|\cdot|_k$, $0 \leq k \leq p$, denotes the k th seminorm on $H^p(\Omega)$, while $H_D^1(\Omega)$ is the subspace of $H^1(\Omega)$ consisting of all functions that vanish on Γ_D . The set $H(\text{div}, \Omega) = \{\mathbf{u} \in (L^2(\Omega))^2 \mid \nabla \cdot \mathbf{u} \in L^2(\Omega)\}$ and its subset $H_N(\text{div}, \Omega) = \{\mathbf{v} \in H(\text{div}, \Omega) \mid \mathbf{v} \cdot \mathbf{n} = 0 \text{ on } \Gamma_N\}$ are Hilbert spaces when equipped with the graph norm $\|\mathbf{u}\|_{\text{div}} = (\|\mathbf{u}\|_0^2 + \|\nabla \cdot \mathbf{u}\|_0^2)^{1/2}$.

Throughout this paper \mathcal{T}_h is a partition of Ω into convex quadrilateral elements κ , \mathcal{N}_h is the set of nodes \mathbf{x}_i in \mathcal{T}_h , and \mathcal{F}_h is the set of oriented faces \mathbf{f}_i in \mathcal{T}_h . A face is oriented by choosing a unit normal $\mathbf{n}_{\mathbf{f}}$, and an element is oriented by choosing a unit normal \mathbf{n}_{κ} to its boundary $\partial\kappa$. By default, all elements are oriented as sources

³A perfectly valid alternative solution is to divide each element into two affine triangles and simply use an RT_0 space on triangles [17]. Nonetheless, quadrilateral elements may still be favored for the following reasons. When a quadrilateral grid is transformed into a triangular one by the above procedure, the number of faces increases by a number equal to the number of elements in the original mesh. Because in the RT_0 space each face is associated with a degree of freedom, this means that the size of the discretized problem will also increase by the same number without formally increasing its accuracy. Second, for problems with advection, quadrilateral grids are easier to align with the flow, which reduces the amount of artificial numerical diffusion.

⁴This is also the reason why formal finite element analysis fails to recognize that the mixed Galerkin solution does contain accurate divergence information.

so that \mathbf{n}_κ is the outer unit normal to $\partial\kappa$.

We assume that the elements in \mathcal{T}_h satisfy the usual conditions required of finite element partitions; see [12, pp. 38–51]. In what follows we restrict our attention to shape-regular partitions \mathcal{T}_h where each κ is a bilinear image of the reference square $\hat{\kappa} = [-1, 1]^2$. We recall that for such partitions there exists a positive α such that

$$(2.1) \quad \frac{1}{\alpha} \mu(\kappa) \leq \| \det D\Phi_\kappa \|_{\infty, \hat{\kappa}} \leq \alpha \mu(\kappa) \quad \forall \kappa \in \mathcal{T}_h;$$

see [14, p. 105]. In (2.1) $D\Phi_\kappa(\hat{\mathbf{x}})$ is the derivative of the bilinear function $\Phi_\kappa(\hat{\mathbf{x}})$ that maps $\hat{\kappa}$ to a given quadrilateral κ . When the range of Φ_κ is clear from the context we will skip the subscript κ . There also hold (see [14, p. 105])

$$(2.2) \quad \det D\Phi_\kappa(\hat{\mathbf{x}}) > 0 \quad \forall \hat{\mathbf{x}} \in \hat{\kappa} \quad \text{and} \quad \mu(\kappa) = \det D\Phi_\kappa(0, 0) \mu(\hat{\kappa}).$$

The first property follows from the convexity of each κ .

$P_{qr}(V)$ denotes polynomial functions on a region $V \subset \mathfrak{R}^2$, whose degree in x and y does not exceed q and r , respectively. Thus, $P_{00}(V)$ is the set of constant polynomials on V ; P_{11} is the set of bilinear polynomials on V ; and so on.

Since our focus is on low-order methods, for the mixed Galerkin method we consider pressure approximations by the piecewise constant space

$$(2.3) \quad Q_0 = \{ p^h \in L^2(\Omega) \mid p^h|_\kappa \in P_{00}(\kappa) \quad \forall \kappa \in \mathcal{T}_h \}$$

and velocity approximations by the lowest-order Raviart–Thomas space

$$(2.4) \quad RT_0 = \{ \mathbf{u}^h \in H(\text{div}, \Omega) \mid \mathbf{u}^h|_\kappa = \mathcal{P}_\kappa \circ \hat{\mathbf{u}}^h; \quad \hat{\mathbf{u}}^h \in P_{10}(\hat{\kappa}) \times P_{01}(\hat{\kappa}) \quad \forall \kappa \in \mathcal{T}_h \},$$

where $\mathcal{P}_\kappa = \det(D\Phi(\hat{\mathbf{x}}))^{-1} D\Phi(\hat{\mathbf{x}})$ is the Piola transform; see [11, p. 97]. The least-squares method uses the same space for the velocity, and the C^0 Lagrangian space

$$(2.5) \quad Q_1 = \{ p^h \in C^0(\bar{\Omega}); p^h|_\kappa = \hat{p} \circ \Phi_\kappa^{-1}; \quad \hat{p} \in P_{11}(\hat{\kappa}) \quad \forall \kappa \in \mathcal{T}_h \}$$

for the pressure approximation.

Finite element spaces are restricted by boundary conditions. RT_0^N is the subspace of RT_0 such that $\mathbf{u}^h \cdot \mathbf{n} = 0$ on Γ_N , and Q_1^D is the subspace of Q_1 such that $p^h = 0$ on Γ_D . No boundary conditions are imposed on Q_0 .

Remark 1. The mapping Φ_κ is affine if and only if κ is a parallelogram. Therefore, in general, RT_0 and Q_1 are not piecewise polynomial spaces.

The unisolvent set of Q_0 consists of the element averages

$$(2.6) \quad \Lambda(Q_0) = \left\{ l_\kappa \mid l_\kappa(p) = \int_\kappa p dx; \quad \kappa \in \mathcal{T}_h \right\},$$

the unisolvent set of Q_1 is given by the nodal values

$$(2.7) \quad \Lambda(Q_1) = \left\{ l_x \mid l_x(p) = \int_\Omega \delta(\mathbf{x}) p dx; \quad \mathbf{x} \in \mathcal{N}_h \right\},$$

and the unisolvent set for RT_0 is the average flux across element faces

$$(2.8) \quad \Lambda(RT_0) = \left\{ l_f \mid l_f(\mathbf{v}) = \int_f \mathbf{v} \cdot \mathbf{n} dS; \quad \mathbf{f} \in \mathcal{F}_h \right\}.$$

The symbols $\{p_\kappa\}$, $\{p_x\}$, and $\{\mathbf{u}_f\}$ stand for the basis sets of Q_0 , Q_1 , and RT_0 , which are dual to (2.6), (2.7), and (2.8), respectively; see [12, 11] for further details.

\mathcal{I}_{Q_0} , \mathcal{I}_{Q_1} , and \mathcal{I}_{RT_0} are the interpolation operators into Q_0 , Q_1 , and RT_0 induced by the degrees of freedom in (2.6)–(2.8). Domains of \mathcal{I}_{Q_1} and \mathcal{I}_{RT_0} consist of those functions in $H^1(\Omega)$ and $H(\text{div}, \Omega)$ for which the functionals in (2.7) and (2.8) are meaningful. For the domain of \mathcal{I}_{RT_0} we will use the space

$$(2.9) \quad W(\Omega) = \{\mathbf{u} \in (L^s(\Omega))^2 \mid \nabla \cdot \mathbf{u} \in L^2(\Omega); \quad s > 2\}.$$

With this choice \mathcal{I}_{RT_0} is uniformly bounded as an operator $W \mapsto RT_0$ (see [11, p. 125]):

$$(2.10) \quad \|\mathcal{I}\mathbf{u}\|_{\text{div}} \leq C\|\mathbf{u}\|_W.$$

When the range of the interpolation operator is clear from this type of argument we skip the space designation and simply write \mathcal{I} .

Approximation properties of interpolation operators are as follows. The L^2 projection \mathcal{I}_{Q_0} is first-order accurate (see [14, p. 108]):

$$(2.11) \quad \|p - \mathcal{I}_{Q_0}p\|_0 \leq Ch\|p\|_1 \quad \forall p \in H^1(\Omega).$$

On shape-regular quadrilateral grids \mathcal{I}_{RT_0} is first-order⁵ accurate in L^2 :

$$(2.12) \quad \|\mathbf{u} - \mathcal{I}_{RT_0}\mathbf{u}\|_0 \leq Ch\|\mathbf{u}\|_1 \quad \forall \mathbf{u} \in H^1(\Omega)^2;$$

see [4, Theorem 4.1]. The nodal interpolant \mathcal{I}_{Q_1} satisfies the error bound

$$(2.13) \quad \|p - \mathcal{I}_{Q_1}p\|_0 + h\|\nabla(p - \mathcal{I}_{Q_1}p)\|_0 \leq Ch\|p\|_2 \quad \forall p \in H^2(\Omega);$$

see [14, p. 107] and [3]. The next lemma states an important property of the divergence operator that will be needed later.

LEMMA 2.1. *Divergence is a surjective mapping $H_N(\text{div}, \Omega) \cap W(\Omega) \mapsto L^2(\Omega)$ with a continuous lifting from $L^2(\Omega)$ into $H_N(\text{div}, \Omega) \cap W(\Omega)$; that is, for every $q \in L^2(\Omega)$ there exists $\mathbf{u}_q \in H_N(\text{div}, \Omega) \cap W(\Omega)$ such that*

$$(2.14) \quad q = \nabla \cdot \mathbf{u}_q \quad \text{and} \quad \|\mathbf{u}_q\|_W \leq C\|q\|_0.$$

For details we refer the reader to [11, p. 136].

3. Extension of DIV to RT_0 . Definition of the natural divergence DIV is based on the coordinate-independent characterization of $\nabla \cdot \mathbf{u}$ in (1.3), applied to each cell $\kappa \in \mathcal{T}_h$. Let \mathcal{F}_h^* and \mathcal{T}_h^* denote the duals of \mathcal{F}_h and \mathcal{T}_h , i.e., collections of real numbers $\{F_f\}$, $\{K_\kappa\}$ associated with the oriented faces and cells in the mesh. Clearly, \mathcal{F}_h^* and \mathcal{T}_h^* are isomorphic⁶ to RT_0 and Q_0 , respectively; therefore, we denote their elements by the same symbols.

The natural divergence is a mapping $\text{DIV} : \mathcal{F}_h^* \mapsto \mathcal{T}_h^*$ defined by

$$(3.1) \quad \text{DIV}(\mathbf{u}^h)|_\kappa = \frac{1}{\mu(\kappa)} \sum_{f \in \mathcal{F}_h(\kappa)} \sigma_f F_f \quad \forall \kappa \in \mathcal{T}_h,$$

⁵On nonaffine grids the divergence error of Raviart–Thomas spaces drops by one order. As a result, $\nabla \cdot \mathcal{I}_{RT_0}(\mathbf{u})$ does not converge to $\nabla \cdot \mathbf{u}$; see [4, Theorem 4.2]. However, as we shall see, the natural divergence of the interpolant is first-order accurate.

⁶This is the key reason why many conservative finite difference methods for (1.1) can be related to low-order implementations of the mixed method—both types of schemes share the same set of degrees of freedom.

where $\mathbf{u}^h \in \mathcal{F}_h^*$, $\mathcal{F}_h(\kappa)$ is the set of oriented faces of κ and

$$\sigma_{\mathbf{f}} = \begin{cases} 1 & \text{if } \mathbf{n}_{\mathbf{f}} = \mathbf{n}_{\kappa}, \\ -1 & \text{if } \mathbf{n}_{\mathbf{f}} = -\mathbf{n}_{\kappa}. \end{cases}$$

Note that $F_{\mathbf{f}}$ are also the degrees of freedom that define vector fields in RT_0 :

$$\mathbf{u}^h = \sum_{\mathbf{f} \in \mathcal{F}_h} F_{\mathbf{f}} \mathbf{u}_{\mathbf{f}} \quad \forall \mathbf{u}^h \in RT_0.$$

Therefore, the action of DIV can be extended to RT_0 vector fields by simply adopting formula (3.1) to compute the discrete divergence of $\mathbf{u}^h \in RT_0$. This defines an operator $\text{DIV} : RT_0 \mapsto Q_0$ which we will use to reformulate mixed and least-squares methods. It is easy to see that for the basis $\{\mathbf{u}_{\mathbf{f}}\}$ of RT_0 ,

$$(3.2) \quad \text{DIV}(\mathbf{u}_{\mathbf{f}}) = \frac{\sigma_{\mathbf{f}}}{\mu(\kappa)} \quad \forall \mathbf{f} \in \mathcal{F}_h.$$

The next lemma states an important property of the natural divergence.

LEMMA 3.1. *The natural divergence DIV has a pointwise Commuting Diagram Property (CDP)*

$$(3.3) \quad \begin{array}{ccc} W(\Omega) & \xrightarrow{\nabla \cdot} & L^2(\Omega) \\ \mathcal{I}_{RT_0} \downarrow & & \downarrow \mathcal{I}_{Q_0} \\ RT_0 & \xrightarrow{\text{DIV}} & Q_0 \end{array}$$

Proof. We need to prove that $\text{DIV}(\mathcal{I}_{RT_0} \mathbf{u}) = \mathcal{I}_{Q_0}(\nabla \cdot \mathbf{u})$ for all $\mathbf{u} \in W(\Omega)$. From definition (3.1) and equation (3.2) it follows that

$$\text{DIV}(\mathcal{I}_{RT_0} \mathbf{u})|_{\kappa} = \text{DIV} \sum_{\mathbf{f} \in \mathcal{F}_h(\kappa)} F_{\mathbf{f}} \mathbf{u}_{\mathbf{f}} = \frac{1}{\mu(\kappa)} \sum_{\mathbf{f} \in \mathcal{F}_h(\kappa)} \sigma_{\mathbf{f}} F_{\mathbf{f}}.$$

On the other hand, from (2.6) and the divergence theorem

$$\mathcal{I}_{Q_0}(\nabla \cdot \mathbf{u})|_{\kappa} = \frac{1}{\mu(\kappa)} \int_{\kappa} \nabla \cdot \mathbf{u} \, dx = \frac{1}{\mu(\kappa)} \int_{\partial\kappa} \mathbf{u} \cdot \mathbf{n} \, dS = \frac{1}{\mu(\kappa)} \sum_{\mathbf{f} \in \mathcal{F}_h(\kappa)} \int_{\mathbf{f}} \mathbf{u} \cdot \mathbf{n} \, dS.$$

CDP follows from the identity

$$\int_{\mathbf{f}} \mathbf{u} \cdot \mathbf{n} \, dS = \sigma_{\mathbf{f}} F_{\mathbf{f}}. \quad \square$$

A discrete version of Lemma 2.1 holds for the natural divergence.

LEMMA 3.2. *The natural divergence is a surjective mapping $RT_0 \mapsto Q_0$ with a continuous lifting from Q_0 into RT_0 ; that is, for every $q^h \in Q_0$ there exists $\mathbf{u}_q^h \in RT_0$ such that*

$$(3.4) \quad q^h = \text{DIV}(\mathbf{u}_q^h) \quad \text{and} \quad \|\mathbf{u}_q^h\|_0 + \|\text{DIV}(\mathbf{u}_q^h)\|_0 \leq C \|q^h\|_0.$$

Proof. To show that DIV is surjective we use CDP and the fact that analytic divergence is a surjective mapping $W(\Omega) \mapsto L^2(\Omega)$ (Lemma 2.1). Any $q^h \in Q_0$ is also

in $L^2(\Omega)$, and so there exists $\mathbf{u}_q \in H_N(\text{div}, \Omega) \cap W(\Omega)$ such that $\nabla \cdot \mathbf{u}_q = q^h$. Let $\mathbf{u}_q^h = \mathcal{I}_{RT_0}(\mathbf{u}_q)$. From CDP it follows that $\text{DIV}(\mathbf{u}_q^h) = \mathcal{I}_{Q_0}(\nabla \cdot \mathbf{u}_q) = \mathcal{I}_{Q_0}(q^h) = q^h$.

Because by construction $\text{DIV}(\mathbf{u}_q^h) = q^h$, to prove that the lifting of DIV from Q_0 into RT_0 is continuous it suffices to show that $\|\mathbf{u}_q^h\|_0 = \|\mathcal{I}_{RT_0}(\mathbf{u}_q)\|_0 \leq \|q^h\|_0$. Using (2.10) (uniform boundedness of \mathcal{I}_{RT_0}) and (2.14) in Lemma 2.1, we see that

$$\|\mathcal{I}_{RT_0}(\mathbf{u}_q)\|_0 \leq \|\mathcal{I}_{RT_0}(\mathbf{u}_q)\|_{\text{div}} \leq C\|\mathbf{u}_q\|_W \leq C\|q^h\|_0.$$

This proves the lemma. \square

Remark 2. According to Lemma 2.1, $\nabla \cdot$ is surjection $H_N(\text{div}, \Omega) \mapsto L^2(\Omega)$. In the mixed method the domain and the range of this operator are approximated by RT_0 and Q_0 elements, respectively. However, on nonaffine quadrilateral grids, $\nabla \cdot RT_0 \neq Q_0$, and the surjective property connecting the domain and the range of $\nabla \cdot$ is lost.⁷ By replacing the analytic divergence by DIV , surjectivity is restored. As a result, if RT_0 is to approximate the domain of the divergence and Q_0 its range, then the approximation of $\nabla \cdot$, which is compatible with its surjective property, is given by DIV rather than $\nabla \cdot$. In other words, DIV provides a better approximation of $\nabla \cdot$ on RT_0 than the usual finite element practice of restricting the analytic operator to the finite element space. This fact validates the mimetic reformulation strategy presented in the next two sections.

Remark 3. The surjective property of DIV , and its lack thereof in $\nabla \cdot$, is a direct consequence of the way these operators act on the flux degrees of freedom. As we have already noted in section 1, DIV operates directly on these degrees of freedom, whereas the action of $\nabla \cdot$ is indirect via basis functions defined by the Piola transform. As a result, the divergence approximation computed by DIV depends only on the flux data and not on the element shape.

The following two lemmas will prove useful later.

LEMMA 3.3. *For every $\mathbf{u}^h \in RT_0$ there holds*

$$(3.5) \quad \int_{\kappa} \nabla \cdot \mathbf{u}^h \, dx = \int_{\kappa} \text{DIV}(\mathbf{u}^h) \, dx, \quad \kappa \in \mathcal{T}_h.$$

Proof. It is enough to show (3.5) for a basis function \mathbf{u}_f associated with a face $f \in \partial\kappa$. Using (3.2) and the definition of the basis functions,

$$\int_{\kappa} \text{DIV}(\mathbf{u}_f) \, dx = \frac{\sigma_f}{\mu(\kappa)} \int_{\kappa} dx = \sigma_f = \int_{\partial\kappa} \mathbf{n} \cdot \mathbf{u}_f \, dS = \int_{\kappa} \nabla \cdot \mathbf{u}_f \, dx. \quad \square$$

LEMMA 3.4. *Assume that \mathcal{T}_h is shape-regular. There is a positive constant C_D such that*

$$(3.6) \quad \|\nabla \cdot \mathbf{u}^h\|_0 \leq C_D \|\text{DIV}(\mathbf{u}^h)\|_0.$$

Proof. It suffices to show (3.6) for one element κ and one basis function \mathbf{u}_f with $f \in \partial\kappa$. After changing variables and noting that the Jacobian is positive (see (2.2)),

$$\|\nabla \cdot \mathbf{u}_f\|_{0,\kappa}^2 = \int_{\kappa} (\nabla \cdot \mathbf{u}_f)(\nabla \cdot \mathbf{u}_f) \, dx = \int_{\hat{\kappa}} (\nabla_{\hat{\mathbf{x}}} \cdot \hat{\mathbf{u}}_f)(\nabla_{\hat{\mathbf{x}}} \cdot \hat{\mathbf{u}}_f)(\det D\Phi)^{-1} \, d\hat{x},$$

⁷The reason why stability of the mixed method is not ruined on such grids is that the following weak CDP holds for $\nabla \cdot$: $\mathcal{I}_{Q_0}(\nabla \cdot \mathcal{I}_{RT_0}(\mathbf{u})) = \mathcal{I}_{Q_0}(\nabla \cdot \mathbf{u})$, i.e.,

$$\int_{\Omega} q^h \nabla \cdot \mathbf{u} \, dx = \int_{\Omega} q^h \nabla \cdot \mathcal{I}_{RT_0}(\mathbf{u}) \, dx.$$

According to Fortin’s lemma this is enough for the inf-sup condition to hold; see [11, p. 138].

where $\hat{\mathbf{f}}$ is one of the faces of the reference element $\hat{\kappa}$. From (2.4) and (2.8) it follows that

$$\hat{\mathbf{u}}_{\hat{\mathbf{f}}} = \frac{1}{4} \begin{bmatrix} 1 \pm x \\ 0 \end{bmatrix} \quad \text{or} \quad \hat{\mathbf{u}}_{\hat{\mathbf{f}}} = \frac{1}{4} \begin{bmatrix} 0 \\ 1 \pm y \end{bmatrix}, \quad \text{and} \quad \nabla_{\hat{\mathbf{x}}} \cdot \hat{\mathbf{u}}_{\hat{\mathbf{f}}} = 1/4.$$

Using the mean value theorem, the lower bound in (2.1), (3.2), and $\mu(\hat{\kappa}) = 4$, we have

$$\|\nabla \cdot \mathbf{u}_{\mathbf{f}}\|_{0,\kappa}^2 = \frac{1}{16} \int_{\hat{\kappa}} (\det D\Phi)^{-1} d\hat{x} = \frac{\mu(\hat{\kappa})}{16 \det D\Phi(\hat{x}^*)} \leq \frac{\alpha\mu(\hat{\kappa})}{16\mu(\kappa)} = \frac{\alpha}{4} \|\text{DIV}(\mathbf{u}_{\mathbf{f}})\|_{0,\kappa}^2.$$

Thus, (3.6) holds with $C_D = \alpha/4$. \square

4. Mimetic reformulation of finite element methods. We begin with a brief summary of the standard mixed method [11] and the locally conservative least-squares method [8]. For further information about related least-squares methods, we refer the reader to [6, 13, 7] and the references cited therein.

4.1. Standard methods. The standard mixed finite element method for (1.1) solves the following variational problem: seek $\mathbf{u}^h \in RT_0^N$ and $p^h \in Q_0$ such that

$$(4.1) \quad \begin{cases} \int_{\Omega} \mathbf{u}^h \Theta_1^{-1} \mathbf{v}^h dx - \int_{\Omega} p^h \nabla \cdot \mathbf{v}^h dx = 0 & \forall \mathbf{v}^h \in RT_0^N, \\ \int_{\Omega} \nabla \cdot \mathbf{u}^h q^h dx + \sigma \int_{\Omega} p^h \Theta_0 q^h dx = \int_{\Omega} f q^h dx & \forall q^h \in Q_0. \end{cases}$$

The second method in our study is a compatible least-squares method for (1.1). In this method the finite element approximation is determined by seeking the minimizer of the least-squares quadratic functional

$$(4.2) \quad J(p^h, \mathbf{u}^h; f) = \|\Theta_0^{-1/2}(\nabla \cdot \mathbf{u}^h + \sigma \Theta_0 p^h - f)\|_0^2 + \|\Theta_1^{1/2}(\nabla p^h + \Theta_1^{-1} \mathbf{u}^h)\|_0^2$$

in $U^h = Q_1^D \times RT_0^N$. The standard finite element implementation of this method solves the following variational equation: seek $\{p^h, \mathbf{u}^h\} \in Q_1^D \times RT_0^N$ such that

$$(4.3) \quad \begin{cases} \int_{\Omega} (\nabla p^h + \Theta_1^{-1} \mathbf{u}^h) \Theta_1 (\nabla q^h + \Theta_1^{-1} \mathbf{v}^h) dx \\ \quad + \int_{\Omega} (\nabla \cdot \mathbf{u}^h + \sigma \Theta_0 p^h) \Theta_0^{-1} (\nabla \cdot \mathbf{v}^h + \sigma \Theta_0 q^h) dx \\ = \int_{\Omega} f \Theta_0^{-1} (\nabla \cdot \mathbf{v}^h + \sigma \Theta_0 q^h) dx & \forall q^h \in Q_1^D, \forall \mathbf{v}^h \in RT_0^N. \end{cases}$$

The following theorem from [8] provides additional information about the standard least-squares method. Specifically, it asserts that (4.3) is locally conservative.

THEOREM 4.1. *Assume that the reaction term is present in (1.1), i.e., $\sigma = 1$. Then, the least-squares equation (4.3) decouples into independent problems for the velocity: seek $\mathbf{u}^h \in RT_0^N$ such that*

$$(4.4) \quad \int_{\Omega} \mathbf{u}^h \Theta^{-1} \mathbf{v}^h dx + \int_{\Omega} \nabla \cdot \mathbf{u}^h \Theta_0^{-1} \nabla \cdot \mathbf{v}^h dx = \int_{\Omega} f \Theta_0^{-1} \nabla \cdot \mathbf{v}^h dx \quad \forall \mathbf{v}^h \in RT_0^N;$$

and the pressure: seek $p^h \in Q_1^D$ such that

$$(4.5) \quad \int_{\Omega} \nabla p^h \Theta_1 \nabla q^h dx + \int_{\Omega} p^h \Theta_0 q^h dx = \int_{\Omega} f q^h dx \quad \forall q^h \in Q_1^D.$$

If the grid is such that the analytic divergence is a surjective map $RT_0 \mapsto Q_0$, then the solution of the weak problem (4.4) coincides with the velocity approximation in the mixed method (4.1).

For proof of this theorem we refer the reader to [8]. On the positive side, Theorem 4.1 implies that for problems with a reaction term the deterioration of accuracy should not spread to the pressure approximation. This follows from the fact that (4.5) defines the Ritz–Galerkin method for (1.1) which retains optimal orders of convergence on general quadrilateral grids; see [12].

On the negative side, for nonaffine quadrilateral elements the analytic divergence does not map RT_0 onto Q_0 (see Remark 2), and so the solution of (4.4) will not coincide with the velocity approximation in the mixed method. Considering that Theorem 5.1 will show that the mixed method produces accurate velocities, this spells potential trouble for the least-squares velocity.

Of course, in the absence of a reaction term ($\sigma = 0$), the least-squares equation remains coupled. In this case we can expect deterioration of accuracy in both variables. Numerical tests in [4] confirm this conjecture. Section 6 will provide further computational evidence to corroborate these conclusions.

4.2. Reformulated methods. We obtain mimetic reformulations of (4.1) and (4.3) by swapping the analytic divergence with DIV . The reformulated mixed method is as follows: seek $\mathbf{u}^h \in RT_0^N$ and $p^h \in Q_0$ such that

$$(4.6) \quad \begin{cases} \int_{\Omega} \mathbf{u}^h \Theta_1^{-1} \mathbf{v}^h \, dx - \int_{\Omega} p^h \text{DIV}(\mathbf{v}^h) \, dx = 0 & \forall \mathbf{v}^h \in RT_0^N, \\ \int_{\Omega} \text{DIV}(\mathbf{u}^h) q^h \, dx + \sigma \int_{\Omega} p^h \Theta_0 q^h \, dx = \int_{\Omega} f q^h \, dx & \forall q^h \in Q_0. \end{cases}$$

We make the usual identifications:

$$a^h(\mathbf{u}^h, \mathbf{v}^h) = \int_{\Omega} \mathbf{u}^h \Theta_1^{-1} \mathbf{v}^h \, dx \quad \text{and} \quad b^h(\mathbf{u}^h, p^h) = \int_{\Omega} p^h \text{DIV}(\mathbf{u}^h) \, dx.$$

Note that $a^h(\cdot, \cdot)$ and $b^h(\cdot, \cdot)$ are defined only for finite element functions.

The method for reformulation of the least-squares method is as follows: seek $\{p^h, \mathbf{u}^h\} \in Q_1^D \times RT_0^N$ such that

$$(4.7) \quad \begin{cases} \int_{\Omega} (\nabla p^h + \Theta_1^{-1} \mathbf{u}^h) \Theta_1 (\nabla q^h + \Theta_1^{-1} \mathbf{v}^h) \, dx \\ \quad + \int_{\Omega} (\text{DIV}(\mathbf{u}^h) + \sigma \Theta_0 p^h) \Theta_0^{-1} (\text{DIV}(\mathbf{v}^h) + \sigma \Theta_0 q^h) \, dx \\ = \int_{\Omega} f \Theta_0^{-1} (\text{DIV}(\mathbf{v}^h) + \sigma \Theta_0 q^h) \, dx & \forall q^h \in Q_1^D, \forall \mathbf{v}^h \in RT_0^N. \end{cases}$$

Remark 4. An existing finite element program for the standard mixed or the least-squares method can be trivially converted to its mimetic reformulation by changing just a few lines of code. From (3.1), (2.2), and $\mu(\hat{\kappa}) = 4$, it follows that

$$\text{DIV}(\mathbf{u}_f)|_{\kappa} = \frac{\sigma f}{\mu(\kappa)} = \frac{\sigma f}{4 \det(D\Phi(0, 0))}.$$

As a result, the conversion to mimetic reformulations amounts to replacing multiple calls to the function that computes $\nabla \cdot \mathbf{u}_f(\mathbf{x})$ at quadrature points, along with the

computation of $\det D\Phi$ at those points, by a single call to compute $\det(D\Phi(0,0))$ combined with a few Boolean operations related to the orientation choice σ_f .

Remark 5. An alternative approach that also restores the first-order convergence in the divergence error has been proposed in [21]. The idea is to “correct” the standard RT_0 basis on $\hat{\kappa}$ so that the basis functions on any $\kappa \in \mathcal{T}_h$ have constant divergence. Correction is affected by adding a vector field defined with the help of $\det D\Phi_\kappa$ which makes the reference basis dependent upon the elements in \mathcal{T}_h . This should be contrasted with our approach, where the definition of the RT_0 basis on $\hat{\kappa}$ is unchanged and remains independent of $\kappa \in \mathcal{T}_h$; instead one changes the definition of the divergence operator on κ . Connection between a mixed method implemented with the modified RT_0 space and a mimetic finite difference scheme for (1.1) is shown in [10].

5. Properties of reformulated methods. This section examines stability and convergence of the reformulated methods. We begin with the analysis of the reformulated mixed method.

5.1. The mixed method. The following theorem shows that (4.1) and (4.6) are equivalent.

THEOREM 5.1. *The standard mixed method (4.1) and its mimetic reformulation (4.6) give rise to identical linear systems of equations for the unknown coefficients of $\mathbf{u}^h \in RT_0^N$ and $p^h \in Q_0$; i.e., their solutions coincide.*

Proof. The mixed problem (4.1) and its mimetic reformulation (4.6) reduce to the linear systems of equations

$$\begin{bmatrix} \mathbf{M}_u & \mathbf{D}^T \\ \mathbf{D} & \mathbf{M}_p \end{bmatrix} \begin{bmatrix} \vec{\mathbf{u}} \\ \vec{p} \end{bmatrix} = \begin{bmatrix} \vec{0} \\ \vec{f} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \mathbf{M}_u & \tilde{\mathbf{D}}^T \\ \tilde{\mathbf{D}} & \mathbf{M}_p \end{bmatrix} \begin{bmatrix} \vec{\mathbf{u}} \\ \vec{p} \end{bmatrix} = \begin{bmatrix} \vec{0} \\ \vec{f} \end{bmatrix},$$

respectively, for the unknown coefficients $\vec{\mathbf{u}}, \vec{p}$ of \mathbf{u}^h and p^h . Here \mathbf{M}_u and \mathbf{M}_p are the consistent mass matrices for RT_0 and Q_0 finite element spaces, respectively. The matrices \mathbf{D} and $\tilde{\mathbf{D}}$ are given by their respective entries

$$\mathbf{D}_{f,\kappa} = \int_{\Omega} p_{\kappa} \nabla \cdot \mathbf{u}_f \, dx \quad \text{and} \quad \tilde{\mathbf{D}}_{f,\kappa} = \int_{\Omega} p_{\kappa} \text{DIV}(\mathbf{u}_f) \, dx, \quad \mathbf{f} \in \mathcal{F}_h, \kappa \in \mathcal{T}_h.$$

The theorem will follow if we can show that $\mathbf{D}_{f,\kappa} = \tilde{\mathbf{D}}_{f,\kappa}$. Let κ be fixed and \mathbf{f} be one of its faces. The basis function $p_{\kappa} = 1/\mu(\kappa)$ is constant on κ and $p_{\kappa} = 0$ on all other elements. Therefore, using (3.5) from Lemma 3.3 it follows that

$$\begin{aligned} \mathbf{D}_{f,\kappa} &= \int_{\Omega} p_{\kappa} \nabla \cdot \mathbf{u}_f \, dx = \frac{1}{\mu(\kappa)} \int_{\kappa} \nabla \cdot \mathbf{u}_f \, dx \\ &= \frac{1}{\mu(\kappa)} \int_{\kappa} \text{DIV}(\mathbf{u}_f) \, dx = \int_{\Omega} p_{\kappa} \text{DIV}(\mathbf{u}_f) \, dx = \tilde{\mathbf{D}}_{f,\kappa}. \quad \square \end{aligned}$$

Remark 6. Boffi brought to our attention a similar equivalence result [9] for two modifications of a primal finite element method for Maxwell’s eigenvalue problem⁸ defined by using a local L^2 projection and reduced integration, respectively. From (3.3) in Lemma 3.1 $\text{DIV}(\mathbf{u}^h) = \mathcal{I}_{Q_0}(\nabla \cdot \mathbf{u}^h)$ for all $\mathbf{u}^h \in RT_0$, from which it follows that the first approach of [9] is equivalent to our mimetic reformulation.

⁸Finite element solution of this problem in two dimensions requires “rotated” RT elements, which suffer from the same accuracy problems as standard RT elements on nonaffine quadrilateral grids.

In light of this theorem we could in principle skip a formal stability analysis of the reformulated mixed method because we already know that the discrete system in the standard mixed method is well behaved. However, a separate stability proof for (4.6) will be very convenient for the error estimates in which we will work with the discrete norm $\|\mathbf{u}^h\|_{\text{DIV}} = (\|\mathbf{u}^h\|_0^2 + \|\text{DIV}(\mathbf{u}^h)\|_0^2)^{1/2}$. The proofs are stated for the case $\sigma = 0$. The extension to $\sigma = 1$ is straightforward.

LEMMA 5.2. *Let $Z^h = \{\mathbf{u}^h \in RT_0 \mid \text{DIV}(\mathbf{u}^h) = 0\}$ denote the null-space of DIV . The form $a^h(\cdot, \cdot)$ is coercive on $Z^h \times Z^h$:*

$$(5.1) \quad C_a \|\mathbf{v}^h\|^2 \leq a^h(\mathbf{v}^h, \mathbf{v}^h) \quad \forall \mathbf{v}^h \in Z^h.$$

The form $b^h(\cdot, \cdot)$ satisfies a discrete inf-sup condition:

$$(5.2) \quad \sup_{\mathbf{v}^h \in RT_0} \frac{b^h(\mathbf{v}^h, p^h)}{\|\mathbf{v}^h\|_{\text{DIV}}} \geq C_b \|p^h\|_0 \quad \forall p^h \in Q_0.$$

Proof. The first statement is a direct consequence of the definition of Z^h and condition (1.2) on the tensor Θ_1 . To prove the inf-sup condition we proceed as follows. Let $p^h \in Q_0$ be arbitrary. From Lemma 3.2 we know that there exists a $\mathbf{u}_p^h \in RT_0$ such that (3.4) holds. Therefore,

$$\sup_{\mathbf{v}^h \in RT_0} \frac{b^h(\mathbf{v}^h, p^h)}{\|\mathbf{v}^h\|_{\text{DIV}}} \geq \frac{b^h(\mathbf{u}_p^h, p^h)}{\|\mathbf{u}_p^h\|_{\text{DIV}}} \geq \frac{\|p^h\|_0^2}{\|\mathbf{u}_p^h\|_{\text{DIV}}} \geq C \|p^h\|_0. \quad \square$$

Lemma 5.2 directly implies the following stability result.

THEOREM 5.3. *Define the discrete bilinear operator*

$$Q^h(\mathbf{u}^h, p^h; \mathbf{v}^h, q^h) = a^h(\mathbf{u}^h, \mathbf{v}^h) - b^h(\mathbf{v}^h, p^h) + b^h(\mathbf{u}^h, q^h).$$

There exists a positive constant C_Q such that

$$(5.3) \quad \sup_{(\mathbf{v}^h, q^h) \in RT_0^N \times Q_0} \frac{Q^h(\mathbf{u}^h, p^h; \mathbf{v}^h, q^h)}{\|\mathbf{v}^h\|_{\text{DIV}} + \|q^h\|_0} \geq C_Q (\|\mathbf{u}^h\|_{\text{DIV}} + \|p^h\|_0).$$

We can now prove optimal error estimates for (4.6).

THEOREM 5.4. *Assume that (2.1) holds for the finite element partition \mathcal{T}_h and that the exact solution of (1.1) is such that $p \in H_D^1(\Omega)$ and $\mathbf{u} \in H_N(\text{div}, \Omega) \cap (H^2(\Omega))^2$. Solution of the reformulated mixed problem (4.6) satisfies the error bound*

$$(5.4) \quad \|\nabla \cdot \mathbf{u} - \text{DIV}(\mathbf{u}^h)\|_0 + \|\mathbf{u} - \mathbf{u}^h\|_0 + \|p - p^h\|_0 \leq Ch (\|\mathbf{u}\|_2 + \|p\|_1).$$

Proof. To avoid tedious technical details, we limit the proof to the case $\Theta_1 = \mathbb{I}$. We begin by splitting the left-hand side in (5.4) into interpolation error and discrete error:

$$\begin{aligned} & \|\nabla \cdot \mathbf{u} - \text{DIV}(\mathbf{u}^h)\|_0 + \|\mathbf{u} - \mathbf{u}^h\|_0 + \|p - p^h\|_0 \\ & \leq (\|\nabla \cdot \mathbf{u} - \text{DIV}\mathcal{I}(\mathbf{u})\|_0 + \|\mathbf{u} - \mathcal{I}\mathbf{u}\|_0 + \|p - \mathcal{I}p\|_0) \\ & \quad + (\|\text{DIV}\mathcal{I}(\mathbf{u}) - \text{DIV}(\mathbf{u}^h)\|_0 + \|\mathcal{I}\mathbf{u} - \mathbf{u}^h\|_0 + \|\mathcal{I}p - p^h\|_0) = E_{\mathcal{I}} + E_h. \end{aligned}$$

The next step is to estimate the discrete error E_h in terms of the interpolation error $E_{\mathcal{I}}$. We make use of the fact that for $(\mathbf{v}^h, q^h) \in RT_0^N \times Q_0$,

$$(5.5) \quad \int_{\Omega} \mathbf{u}^h \mathbf{v}^h \, dx - \int_{\Omega} p^h \text{DIV}(\mathbf{v}^h) \, dx = 0 = \int_{\Omega} \mathbf{u} \mathbf{v}^h \, dx - \int_{\Omega} p \nabla \cdot \mathbf{v}^h \, dx$$

and

$$(5.6) \quad \int_{\Omega} q^h \operatorname{DIV}(\mathbf{u}^h) dx = \int_{\Omega} f q^h dx = \int_{\Omega} q^h \nabla \cdot \mathbf{u} dx.$$

For brevity we switch to inner product notation. Adding and subtracting $\mathcal{I}\mathbf{u}$ and $\mathcal{I}p$ in (5.5), we obtain

$$(\mathbf{u} - \mathcal{I}\mathbf{u}, \mathbf{v}^h) + (\mathcal{I}\mathbf{u} - \mathbf{u}^h, \mathbf{v}^h) - (p - \mathcal{I}p, \nabla \cdot \mathbf{v}^h) - (\mathcal{I}p, \nabla \cdot \mathbf{v}^h) + (p^h, \operatorname{DIV}(\mathbf{v}^h)) = 0.$$

As $\mathcal{I}p$ is constant, by Lemma (3.3) we can replace $(\mathcal{I}p, \nabla \cdot \mathbf{v}^h)$ by $(\mathcal{I}p, \operatorname{DIV}(\mathbf{v}^h))$; thus

$$(5.7) \quad (\mathbf{u}^h - \mathcal{I}\mathbf{u}, \mathbf{v}^h) + (\mathcal{I}p - p^h, \operatorname{DIV}(\mathbf{v}^h)) = (\mathbf{u} - \mathcal{I}\mathbf{u}, \mathbf{v}^h) + (\mathcal{I}p - p, \nabla \cdot \mathbf{v}^h).$$

Adding and subtracting $\mathcal{I}(\nabla \cdot \mathbf{u})$ in (5.6) yields

$$(\nabla \cdot \mathbf{u} - \mathcal{I}(\nabla \cdot \mathbf{u}), q^h) + (\mathcal{I}(\nabla \cdot \mathbf{u}) - \operatorname{DIV}(\mathbf{u}^h), q^h) = 0.$$

Using CDP, we obtain

$$(5.8) \quad (\operatorname{DIV}(\mathbf{u}^h - \mathcal{I}\mathbf{u}), q^h) = (\nabla \cdot \mathbf{u} - \operatorname{DIV}(\mathcal{I}\mathbf{u}), q^h).$$

Substituting $(\mathbf{u}^h - \mathcal{I}\mathbf{u}, \mathcal{I}p - p^h)$ into the inf-sup result (5.3), we get

$$\begin{aligned} Q^h(\mathbf{u}^h - \mathcal{I}\mathbf{u}, \mathcal{I}p - p^h; \mathbf{v}^h, q^h) \\ \geq C_Q (\|\mathbf{u}^h - \mathcal{I}\mathbf{u}\|_0 + \|\operatorname{DIV}(\mathbf{u}^h - \mathcal{I}\mathbf{u})\|_0 + \|p^h - \mathcal{I}p\|_0) \\ \times (\|\mathbf{v}^h\|_0 + \|\operatorname{DIV}(\mathbf{v}^h)\|_0 + \|q^h\|_0). \end{aligned}$$

On the other hand, due to the definition of Q^h , (5.7), (5.8), Cauchy inequalities, and Lemma 3.4,

$$\begin{aligned} Q^h(\mathbf{u}^h - \mathcal{I}\mathbf{u}, \mathcal{I}p - p^h; \mathbf{v}^h, q^h) \\ = (\mathbf{u} - \mathcal{I}\mathbf{u}, \mathbf{v}^h) + (\mathcal{I}p - p, \nabla \cdot \mathbf{v}^h) + (\nabla \cdot \mathbf{u} - \operatorname{DIV}(\mathcal{I}\mathbf{u}), q^h) \\ \leq (\|\mathbf{u} - \mathcal{I}\mathbf{u}\|_0 + \|p - \mathcal{I}p\|_0 + \|\nabla \cdot \mathbf{u} - \operatorname{DIV}(\mathcal{I}\mathbf{u})\|_0) \\ \times (\|\mathbf{v}^h\|_0 + C_D \|\operatorname{DIV}(\mathbf{v}^h)\|_0 + \|q^h\|_0) \end{aligned}$$

for a positive constant C_D . It is safe to assume $C_D \geq 1$ (without loss of generality), and thus $(\|\mathbf{v}^h\|_0 + C_D \|\operatorname{DIV}(\mathbf{v}^h)\|_0 + \|q^h\|_0) \leq C_D (\|\mathbf{v}^h\|_0 + \|\operatorname{DIV}(\mathbf{v}^h)\|_0 + \|q^h\|_0)$, which, combined with the previous two estimates of Q^h , yields

$$\begin{aligned} E_h &= \|\mathbf{u}^h - \mathcal{I}\mathbf{u}\|_0 + \|\operatorname{DIV}(\mathbf{u}^h - \mathcal{I}\mathbf{u})\|_0 + \|p^h - \mathcal{I}p\|_0 \\ &\leq \frac{C_D}{C_Q} (\|\mathbf{u} - \mathcal{I}\mathbf{u}\|_0 + \|p - \mathcal{I}p\|_0 + \|\nabla \cdot \mathbf{u} - \operatorname{DIV}(\mathcal{I}\mathbf{u})\|_0) = \frac{C_D}{C_Q} E_{\mathcal{I}}. \end{aligned}$$

Therefore,

$$(5.9) \quad \|\nabla \cdot \mathbf{u} - \operatorname{DIV}(\mathbf{u}^h)\|_0 + \|\mathbf{u} - \mathbf{u}^h\|_0 + \|p - p^h\|_0 \leq \left(1 + \frac{C_D}{C_Q}\right) E_{\mathcal{I}}.$$

The remainder of the proof follows from CDP, (2.11), and (2.12). We have

$$\|\nabla \cdot \mathbf{u} - \text{DIV}(\mathcal{I}\mathbf{u})\|_0 = \|\nabla \cdot \mathbf{u} - \mathcal{I}(\nabla \cdot \mathbf{u})\|_0 \leq Ch\|\nabla \cdot \mathbf{u}\|_1$$

and

$$\|\mathbf{u} - \mathcal{I}\mathbf{u}\|_0 + \|p - \mathcal{I}p\|_0 \leq Ch(\|\mathbf{u}\|_1 + \|p\|_1);$$

i.e.,

$$E_{\mathcal{I}} \leq Ch(\|\nabla \cdot \mathbf{u}\|_1 + \|\mathbf{u}\|_1 + \|p\|_1) \leq Ch(\|\mathbf{u}\|_2 + \|p\|_1),$$

which establishes (5.4). \square

Remark 7. The presence of the constant C_D in (5.9) indicates that the size of the approximation error is directly related to assumption (2.1) on the shape-regularity of the finite element partition \mathcal{T}_h .

5.2. The least-squares method. It is easy to see that the equivalence result of Theorem 5.1 cannot be extended to the least-squares method. To convince ourselves that the mimetic reformulation (4.7) of this method is genuinely different from its standard version (4.3) let us examine the term

$$\int_{\Omega} (\text{DIV}(\mathbf{u}^h) + \sigma\Theta_0 p^h)\Theta_0^{-1}(\text{DIV}(\mathbf{v}^h) + \sigma\Theta_0 q^h)dx$$

from (4.7). It is clear that, for the same reasons as stated in Remark 2, on a nonaffine quadrilateral element,

$$\int_{\Omega} \text{DIV}(\mathbf{u}^h)\Theta_0^{-1}\text{DIV}(\mathbf{v}^h)dx \neq \int_{\Omega} \nabla \cdot \mathbf{u}^h\Theta_0^{-1}\nabla \cdot \mathbf{v}^h dx.$$

The cross terms also do not match because in the least-squares method $p^h \in Q_1$ is not constant and cannot be pulled out of the integral as in Theorem 5.1. Thus,

$$\int_{\Omega} q^h\text{DIV}(\mathbf{u}^h)dx \neq \int_{\Omega} q^h\nabla \cdot \mathbf{u}^h dx.$$

Another difference between the two versions of the least-squares method is that the splitting in Theorem 4.1 does not extend to the mimetic reformulation (4.7). This would require the discrete Green's identity

$$\int_{\Omega} p^h\text{DIV}(\mathbf{u}^h) dx + \int_{\Omega} \mathbf{u}^h\nabla p^h dx = 0 \quad \forall \mathbf{u}^h \in RT_0^N, \quad \forall p^h \in Q_1^D,$$

which in general does not hold. Therefore, velocity computed by (4.7) is not locally conservative in the sense described in [8]. This can be fixed by using the flux-correction procedure defined in [8].

The following theorem asserts stability of the reformulated least-squares method.

THEOREM 5.5. *Assume that (2.1) holds. There is a positive constant C such that*

$$(5.10) \quad \begin{aligned} & C(\|\text{DIV}(\mathbf{u}^h)\|_0 + \|\mathbf{u}^h\|_0 + \|p^h\|_1) \\ & \leq \|\Theta_0^{-1/2}(\text{DIV}(\mathbf{u}^h) + \Theta_0 p^h)\|_0 + \|\Theta_1^{1/2}(\nabla p^h + \Theta_1^{-1}\mathbf{u}^h)\|_0 \end{aligned}$$

for every $\mathbf{u}^h \in RT_0^N$ and $p^h \in Q_1^D$.

Proof. To avoid simple but tedious technical details, we state the proof for $\Theta_1 = \mathbb{I}$ and $\Theta_0 = 1$. In this case, the right-hand side in (5.10) expands into

$$\begin{aligned} \|\text{DIV}(\mathbf{u}^h) + p^h\|_0^2 + \|\nabla p^h + \mathbf{u}^h\|_0^2 &= 2 \left(\int_{\Omega} p^h \text{DIV}(\mathbf{u}^h) dx + \int_{\Omega} \mathbf{u}^h \nabla p^h dx \right) \\ &+ \|\mathbf{u}^h\|_0^2 + \|\text{DIV}(\mathbf{u}^h)\|_0^2 + \|p^h\|_0^2 + \|\nabla p^h\|_0^2. \end{aligned}$$

We switch to inner product notation. Adding and subtracting the projection of p^h onto Q_0 , using Green's identity, (3.5) in Lemma 3.3, and Cauchy's inequality, we get

$$\begin{aligned} &(\text{DIV}(\mathbf{u}^h), p^h) + (\nabla p^h, \mathbf{u}^h) \\ &= (\text{DIV}(\mathbf{u}^h), p^h - \mathcal{I}_{Q_0} p^h) + (\text{DIV}(\mathbf{u}^h), \mathcal{I}_{Q_0} p^h) - (\nabla \cdot \mathbf{u}^h, p^h) \\ &= (\text{DIV}(\mathbf{u}^h), p^h - \mathcal{I}_{Q_0} p^h) + (\nabla \cdot \mathbf{u}^h, \mathcal{I}_{Q_0} p^h - p^h) \\ &\leq \|\text{DIV}(\mathbf{u}^h)\|_0 \|p^h - \mathcal{I}_{Q_0} p^h\|_0 + \|\nabla \cdot \mathbf{u}^h\|_0 \|p^h - \mathcal{I}_{Q_0} p^h\|_0. \end{aligned}$$

Using (3.6) from Lemma 3.4, the approximation result (2.11), and the inequality $2ab \leq a^2 + b^2$, we get

$$\begin{aligned} &2((\text{DIV}(\mathbf{u}^h), p^h) + (\nabla p^h, \mathbf{u}^h)) \\ &\leq 2(1 + C_D)Ch \|\text{DIV}(\mathbf{u}^h)\|_0 \|\nabla p^h\|_0 \leq (1 + C_D)Ch (\|\text{DIV}(\mathbf{u}^h)\|_0^2 + \|\nabla p^h\|_0^2). \end{aligned}$$

As a result, for sufficiently small h ,

$$\begin{aligned} &\|\text{DIV}(\mathbf{u}^h) + p^h\|_0^2 + \|\nabla p^h + \mathbf{u}^h\|_0^2 \\ &\geq (1 - (1 + C_D)Ch) (\|\mathbf{u}^h\|_0^2 + \|\text{DIV}(\mathbf{u}^h)\|_0^2 + \|p^h\|_0^2 + \|\nabla p^h\|_0^2) \\ &\geq \frac{1}{2} (\|\mathbf{u}^h\|_0^2 + \|\text{DIV}(\mathbf{u}^h)\|_0^2 + \|p^h\|_1^2). \quad \square \end{aligned}$$

Theorem 5.5 in conjunction with the Lax–Milgram lemma implies that the reformulated least-squares problem has a unique solution. Using this theorem, we can also prove optimal error estimates for the solution of (4.7).

THEOREM 5.6. *Assume that (2.1) holds for the finite element partition \mathcal{T}_h and that the exact solution of (1.1) is such that $p \in H_D^1(\Omega) \cap H^2(\Omega)$ and $\mathbf{u} \in H_N(\text{div}, \Omega) \cap (H^2(\Omega))^2$. Solution of the reformulated least-squares problem (4.7) satisfies the error bound*

$$(5.11) \quad \|\nabla \cdot \mathbf{u} - \text{DIV}(\mathbf{u}^h)\|_0 + \|\mathbf{u} - \mathbf{u}^h\|_0 + \|p - p^h\|_1 \leq Ch (\|p\|_2 + \|\mathbf{u}\|_2).$$

Proof. For clarity we state the proof using the same setting as in the proof of Theorem 5.5. We begin by splitting the left-hand side in (5.11) into interpolation error and discrete error:

$$\begin{aligned} &\|\nabla \cdot \mathbf{u} - \text{DIV}(\mathbf{u}^h)\|_0 + \|\mathbf{u} - \mathbf{u}^h\|_0 + \|p - p^h\|_1 \\ &\leq (\|\nabla \cdot \mathbf{u} - \text{DIV}\mathcal{I}(\mathbf{u})\|_0 + \|\mathbf{u} - \mathcal{I}\mathbf{u}\|_0 + \|p - \mathcal{I}p\|_1) \\ &\quad + (\|\text{DIV}\mathcal{I}(\mathbf{u}) - \text{DIV}(\mathbf{u}^h)\|_0 + \|\mathcal{I}\mathbf{u} - \mathbf{u}^h\|_0 + \|\mathcal{I}p - p^h\|_1) = E_{\mathcal{I}} + E_h. \end{aligned}$$

The next step is to estimate E_h in terms of the interpolation error. For this purpose, we use that $f = \nabla \cdot \mathbf{u} + p$ and $0 = \nabla p + \mathbf{u}$ to write (4.7) as

$$\begin{aligned} & (\text{DIV}(\mathbf{u}^h) + p^h, \text{DIV}(\mathbf{v}^h) + q^h) + (\nabla p^h + \mathbf{u}^h, \nabla q^h + \mathbf{v}^h) \\ &= (\nabla \cdot \mathbf{u} + p, \text{DIV}(\mathbf{v}^h) + q^h) + (\nabla p + \mathbf{u}, \nabla q^h + \mathbf{v}^h). \end{aligned}$$

Subtracting the interpolants of p and \mathbf{u} from both sides of this identity and using Cauchy’s inequality gives

$$\begin{aligned} & (\text{DIV}(\mathbf{u}^h - \mathcal{I}\mathbf{u}) + p^h - \mathcal{I}p, \text{DIV}(\mathbf{v}^h) + q^h) + (\nabla(p^h - \mathcal{I}p) + \mathbf{u}^h - \mathcal{I}\mathbf{u}, \nabla q^h + \mathbf{v}^h) \\ &= (\nabla \cdot \mathbf{u} - \text{DIV}(\mathcal{I}\mathbf{u}) + p - \mathcal{I}p, \text{DIV}(\mathbf{v}^h) + q^h) + (\nabla(p - \mathcal{I}p) + \mathbf{u} - \mathcal{I}\mathbf{u}, \nabla q^h + \mathbf{v}^h) \\ &\leq C(\|\nabla \cdot \mathbf{u} - \text{DIV}(\mathcal{I}\mathbf{u})\|_0 + \|\mathbf{u} - \mathcal{I}\mathbf{u}\|_0 + \|p - \mathcal{I}p\|_1) (\|\text{DIV}(\mathbf{v}^h)\|_0 + \|\mathbf{v}^h\|_0 + \|q^h\|_1) \\ &\leq CE_{\mathcal{I}} \times (\|\text{DIV}(\mathbf{v}^h)\|_0 + \|\mathbf{v}^h\|_0 + \|q^h\|_1). \end{aligned}$$

Then we set $\mathbf{v}^h = \mathbf{u}^h - \mathcal{I}\mathbf{u}$ and $q^h = p^h - \mathcal{I}p$ and use the stability bound (5.10):

$$\begin{aligned} E_h^2 &\leq C (\text{DIV}(\mathbf{u}^h - \mathcal{I}\mathbf{u}) + p^h - \mathcal{I}p, \text{DIV}(\mathbf{u}^h - \mathcal{I}\mathbf{u}) + p^h - \mathcal{I}p) \\ &\quad + (\nabla(p^h - \mathcal{I}p) + \mathbf{u}^h - \mathcal{I}\mathbf{u}, \nabla(p^h - \mathcal{I}p) + \mathbf{u}^h - \mathcal{I}\mathbf{u}) \leq CE_{\mathcal{I}} \times E_h. \end{aligned}$$

Therefore, $E_h \leq CE_{\mathcal{I}}$ and

$$\|\nabla \cdot \mathbf{u} - \text{DIV}(\mathbf{u}^h)\|_0 + \|\mathbf{u} - \mathbf{u}^h\|_0 + \|p - p^h\|_1 \leq (1 + C)E_{\mathcal{I}}.$$

To complete the proof we estimate $E_{\mathcal{I}}$ as follows. Using CDP (3.3) and (2.11),

$$\|\nabla \cdot \mathbf{u} - \text{DIV}\mathcal{I}(\mathbf{u})\|_0 = \|\nabla \cdot \mathbf{u} - \mathcal{I}(\nabla \cdot \mathbf{u})\|_0 \leq Ch\|\nabla \cdot \mathbf{u}\|_1,$$

while from (2.12) and (2.13) we have that

$$\|\mathbf{u} - \mathcal{I}\mathbf{u}\|_0 + \|p - \mathcal{I}p\|_1 \leq Ch(\|\mathbf{u}\|_1 + \|p\|_2).$$

Therefore,

$$E_{\mathcal{I}} \leq Ch(\|p\|_2 + \|\mathbf{u}\|_2).$$

This establishes (5.11). \square

6. Numerical results. Computational experiments in this section illustrate the properties of standard and reformulated finite element methods using three different partitions of $\Omega = [0, 1]^2$ into quadrilateral elements; see Figure 6.1. We refer to the leftmost partition in this figure as the “trapezoidal grid.” This grid was used by Arnold, Boffi, and Falk [4] to demonstrate loss of accuracy in div-conforming elements and is characterized by a high degree of “nonaffinity.” The middle partition corresponds to a randomly perturbed⁹ uniform grid which provides a more realistic

⁹This grid was suggested by one of the anonymous referees and is defined as follows. We start with a uniform partition of Ω into square elements with side lengths h and draw a circle of radius $h/4$ around each node. All internal nodes are then randomly repositioned inside these circles, corner nodes are held fixed, and the rest of the nodes on the boundary are moved randomly along the sides of Ω within $\pm h/4$ of their original locations.

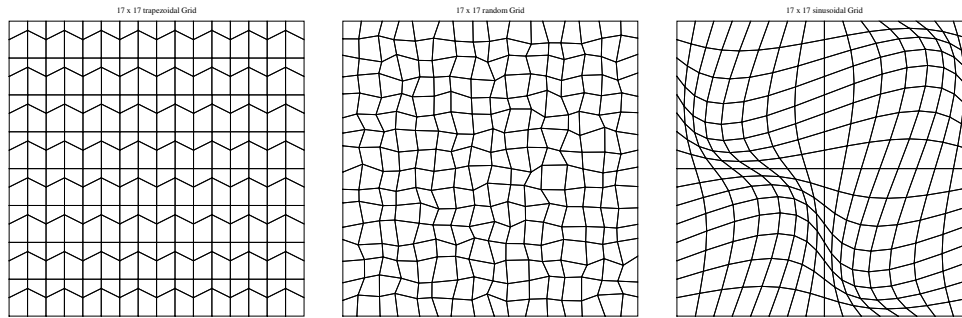


FIG. 6.1. *Quadrilateral grids used in the computational experiments. From left to right: Trapezoidal grid [4], randomly perturbed grid, and sinusoidal grid [18].*

example of a highly nonaffine quadrilateral grid. The rightmost partition in Figure 6.1 provides an example of a smooth nonorthogonal grid¹⁰ which may become quite distorted while retaining “near-affinity” of most of its quadrilateral elements. This grid is used to underscore the fact that the root cause for the loss of convergence is nonaffinity of the grid rather than the level of its distortion.

We show that on the first two partitions, i.e., on trapezoidal and random grids,

- the natural divergence of the velocity approximation in the standard mixed method is first-order accurate; i.e., it is optimally accurate;
- the deterioration of accuracy in the least-squares method does not spread to the pressure approximation if the reaction term is present; however, the velocity approximation is worse than in the mixed method;
- without the reaction term, the deterioration of accuracy affects all variables in the least-squares method;
- the mimetic reformulation of the least-squares method solves all of these problems and yields optimally accurate pressure and velocity approximations.

As far as the last partition is concerned, we show that thanks to the almost affine nature of the sinusoidal grid there is virtually no degradation of accuracy in mixed and least-squares methods.

The linear systems are assembled using 2×2 Gauss quadrature and solved “exactly” using direct solvers. The order of convergence study solves (1.1) with $\Gamma_N = \emptyset$,

$$\Theta_1 = \begin{bmatrix} \exp((x+y)/2) & \sin(2\pi x) \\ \sin(2\pi x) & \exp((x/2+y)/2) \end{bmatrix}, \quad \Theta_0 = 1,$$

and the right-hand side and boundary data generated from the manufactured solution

$$p = -\exp(x) \sin(y), \quad \text{and} \quad \mathbf{u} = -\Theta_1 \nabla p.$$

Orders of convergence are estimated using data on 33×33 , 65×65 , and 129×129 grids with 1024, 4096, and 16384 elements, respectively.

The maximum anisotropy of Θ_1 is attained at the top right corner of Ω and equals $4 \exp(3/2) \approx 18$. The nonconstant full tensor permeability is used only in order to

¹⁰The nodal positions in this grid are defined by

$$x(\xi, \eta, t) = \xi + \alpha(t) \sin(2\pi\xi) \sin(2\pi\eta) \quad \text{and} \quad y(\xi, \eta, t) = \eta + \alpha(t) \sin(2\pi\xi) \sin(2\pi\eta),$$

respectively, where $\alpha(t) \leq 0.1$ and t is a real parameter between 0 and 1; see [18]. The grid shown in Figure 6.1 corresponds to $t = 0.5$ and $\alpha(t) = t/5$.

TABLE 6.1

Error data and estimated orders of convergence for the standard mixed Galerkin (MG) method and its mimetic reformulation (RMG) on trapezoidal grids.

Error	Method	33 × 33	65 × 65	129 × 129	Order
$\ p - p^h\ _0$	MG	0.179E-01	0.893E-02	0.447E-02	0.9999
	RMG	0.179E-02	0.893E-02	0.447E-02	0.9999
$\ \mathbf{u} - \mathbf{u}^h\ _0$	MG	0.652E-01	0.325E-01	0.162E-01	1.0029
	RMG	0.652E-01	0.325E-01	0.162E-01	1.0029
$\ \nabla \cdot \mathbf{u} - \nabla \cdot \mathbf{u}^h\ _0$	MG	0.117E+01	0.110E+01	0.109E+01	0.0211
	RMG	0.117E+01	0.110E+01	0.109E+01	0.0211
$\ \nabla \cdot \mathbf{u} - \text{DIV}(\mathbf{u}^h)\ _0$	MG	0.440E+00	0.220E+00	0.110E+00	1.0000
	RMG	0.440E+00	0.220E+00	0.110E+00	1.0000

TABLE 6.2

Error data and estimated orders of convergence for the standard mixed Galerkin (MG) method and its mimetic reformulation (RMG) on randomly perturbed grids.

Error	Method	33 × 33	65 × 65	129 × 129	Order
$\ p - p^h\ _0$	MG	0.170E-01	0.848E-02	0.424E-02	1.000
	RMG	0.170E-01	0.848E-02	0.424E-02	1.000
$\ \mathbf{u} - \mathbf{u}^h\ _0$	MG	0.611E-01	0.308E-01	0.155E-01	1.000
	RMG	0.611E-01	0.308E-01	0.155E-01	1.000
$\ \nabla \cdot \mathbf{u} - \nabla \cdot \mathbf{u}^h\ _0$	MG	0.975E+00	0.890E+00	0.875E+00	0.025
	RMG	0.975E+00	0.890E+00	0.875E+00	0.025
$\ \nabla \cdot \mathbf{u} - \text{DIV}(\mathbf{u}^h)\ _0$	MG	0.464E+00	0.232E+00	0.116E+00	1.000
	RMG	0.464E+00	0.232E+00	0.116E+00	1.000

TABLE 6.3

Error data and estimated orders of convergence for the standard mixed Galerkin method on sinusoidal grids.

Error	33 × 33	65 × 65	129 × 129	Order
$\ p - p^h\ _0$	0.184E-01	0.902E-02	0.449E-02	1.007
$\ \mathbf{u} - \mathbf{u}^h\ _0$	0.638E-01	0.318E-01	0.159E-01	1.001
$\ \nabla \cdot \mathbf{u} - \nabla \cdot \mathbf{u}^h\ _0$	0.576E+00	0.288E+00	0.144E+00	0.999
$\ \nabla \cdot \mathbf{u} - \text{DIV}(\mathbf{u}^h)\ _0$	0.541E+00	0.271E+00	0.135E+00	0.999

make the tests more “realistic” and is not necessary to elicit the loss of convergence in the two standard methods. The latter can be observed even in the trivial case of $\Theta_1 = \mathbf{I}$, where \mathbf{I} is a 2×2 unit matrix; see [4].

The mixed method and its reformulation. The presence of the reaction term in (1.1) or lack thereof does not affect the overall behavior of the computed error. For brevity, results without this term ($\sigma = 0$) are omitted. Error data and estimated convergence rates for (4.1) and its mimetic reformulation (4.6) on trapezoidal and randomly perturbed grids are summarized in Tables 6.1–6.2. The tables show identical¹¹ errors for both versions of the mixed method, which confirms the assertion of

¹¹To avoid data variations caused by the randomness of the grid, for each grid size the two methods were run on the same instance of the random mesh.

TABLE 6.4

Error data and estimated orders of convergence for the standard least-squares method (LS) and its mimetic reformulation (RLS) on trapezoidal grids: Problem (1.1) with reaction term ($\sigma = 1$).

Error	Method	33×33	65×65	129×129	Order
$\ p - p^h\ _0$	LS	0.863E-04	0.216E-04	0.541E-05	1.998
	RLS	0.876E-04	0.219E-04	0.549E-05	1.998
$\ \nabla(p - p^h)\ _0$	LS	0.1592E-01	0.799E-02	0.401E-02	0.997
	RLS	0.1592E-01	0.799E-02	0.401E-02	0.997
$\ \mathbf{u} - \mathbf{u}^h\ _0$	LS	0.675E-01	0.362E-01	0.225E-01	0.683
	RLS	0.652E-01	0.325E-01	0.162E-01	1.003
$\ \nabla \cdot \mathbf{u} - \nabla \cdot \mathbf{u}^h\ _0$	LS	0.115E+01	0.109E+01	0.107E+01	0.021
	RLS	–	–	–	–
$\ \nabla \cdot \mathbf{u} - \text{DIV}(\mathbf{u}^h)\ _0$	LS	0.479E+00	0.285E+00	0.210E+00	0.439
	RLS	0.440E+00	0.220E+00	0.110E+00	1.000

Theorem 5.1 that their solutions coincide.

As predicted by Theorem 5.1, when the divergence error of the velocity approximation is measured *directly* by DIV instead of *indirectly* by $\nabla \cdot$, the order of convergence improves to 1. This validates our assertion that the loss of convergence in the mixed method is superficial rather than real. It follows that a standard implementation of this method with the lowest-order Raviart–Thomas element is safe to use on general quadrilateral grids, as long as one remembers to extract the divergence information using DIV.

Table 6.3 shows error data and estimated convergence rates for the standard mixed method on the sinusoidal grid. Owing to the fact that this grid is nearly affine, the rates of convergence measured by using the analytic and the mimetic divergence operators are identical despite the small variations in their values.

The least-squares method and its reformulation. Theorem 4.1 suggests that the reaction term could be very important for the standard least-squares method. This turns out to be the case. Tables 6.4–6.5 show error data for (4.3) and (4.7) with the term ($\sigma = 1$) on trapezoidal and randomly perturbed grids, respectively. For the velocity in the standard method on both grids we see a reduced order of convergence in the L^2 -norm and an almost complete loss of convergence in the divergence error. It is worth pointing out that using DIV to extract the divergence information from the standard least-squares solution does not help much. Nevertheless, the order of convergence in the divergence error is somewhat improved.

As predicted by Theorem 4.1, when the reaction term is present the loss of accuracy does not spread to the pressure approximation in the standard method. Tables 6.4–6.5 show the expected second- and first-order convergence for the L^2 - and H^1 -seminorm errors of this variable, respectively.

In the absence of the reaction term, the standard least-squares method fares much worse. Tables 6.6–6.7 show that the loss of accuracy on trapezoidal and randomly perturbed grids when $\sigma = 0$ affects both variables. We see that, without the reaction term, the L^2 order of convergence of the pressure is completely ruined, and the H^1 -seminorm error is severely reduced. These results are consistent with the numerical data on trapezoidal grids presented in [4] and confirm that, unlike in the mixed method, the loss of accuracy in the least-squares method is real. Inclusion of the reaction term helps to stem the deterioration of the pressure approximation, but, as

TABLE 6.5

Error data and estimated orders of convergence for the standard least-squares method (LS) and its mimetic reformulation (RLS) on randomly perturbed grids: Problem (1.1) with reaction term ($\sigma = 1$).

Error	Method	33×33	65×65	129×129	Order
$\ p - p^h\ _0$	LS	0.650E-04	0.167E-04	0.414E-05	2.008
	RLS	0.733E-04	0.200E-04	0.505E-05	1.984
$\ \nabla(p - p^h)\ _0$	LS	0.146E-01	0.734E-02	0.366E-02	1.003
	RLS	0.148E-01	0.742E-02	0.372E-02	0.996
$\ \mathbf{u} - \mathbf{u}^h\ _0$	LS	0.627E-01	0.326E-01	0.185E-01	0.815
	RLS	0.611E-01	0.308E-01	0.154E-01	0.999
$\ \nabla \cdot \mathbf{u} - \nabla \cdot \mathbf{u}^h\ _0$	LS	0.942E+00	0.882E+00	0.858E+00	0.040
	RLS	-	-	-	-
$\ \nabla \cdot \mathbf{u} - \text{DIV}(\mathbf{u}^h)\ _0$	LS	0.484E+00	0.271E+00	0.190E+00	0.511
	RLS	0.463E+00	0.232E+00	0.116E+00	1.001

TABLE 6.6

Error data and estimated orders of convergence for the standard least-squares method (LS) and its mimetic reformulation (RLS) on trapezoidal grids: Problem (1.1) without reaction term ($\sigma = 0$).

Error	Method	33×33	65×65	129×129	Order
$\ p - p^h\ _0$	LS	0.175E-02	0.150E-02	0.144E-02	0.060
	RLS	0.378E-03	0.947E-04	0.237E-04	1.999
$\ \nabla(p - p^h)\ _0$	LS	0.201E-01	0.136E-01	0.114E-01	0.254
	RLS	0.161E-01	0.801E-02	0.401E-02	0.999
$\ \mathbf{u} - \mathbf{u}^h\ _0$	LS	0.676E-01	0.363E-01	0.227E-01	0.678
	RLS	0.652E-01	0.325E-01	0.162E-01	1.003
$\ \nabla \cdot \mathbf{u} - \nabla \cdot \mathbf{u}^h\ _0$	LS	0.115E+01	0.109E+01	0.107E+01	0.021
	RLS	-	-	-	-
$\ \nabla \cdot \mathbf{u} - \text{DIV}(\mathbf{u}^h)\ _0$	LS	0.479E+00	0.285E+00	0.211E+00	0.437
	RLS	0.440E+00	0.220E+00	0.110E+00	1.000

TABLE 6.7

Error data and estimated orders of convergence for the standard least-squares method (LS) and its mimetic reformulation (RLS) on random grids: Problem (1.1) without reaction term ($\sigma = 0$).

Error	Method	33×33	65×65	129×129	Order
$\ p - p^h\ _0$	LS	0.123E-02	0.101E-02	0.951E-03	0.082
	RLS	0.396E-03	0.973E-04	0.251E-04	1.956
$\ \nabla(p - p^h)\ _0$	LS	0.169E-01	0.104E-01	0.793E-02	0.388
	RLS	0.152E-01	0.745E-02	0.373E-02	0.997
$\ \mathbf{u} - \mathbf{u}^h\ _0$	LS	0.624E-01	0.328E-01	0.186E-01	0.816
	RLS	0.610E-01	0.308E-01	0.153E-01	1.005
$\ \nabla \cdot \mathbf{u} - \nabla \cdot \mathbf{u}^h\ _0$	LS	0.930E+00	0.874E+00	0.861E+00	0.021
	RLS	-	-	-	-
$\ \nabla \cdot \mathbf{u} - \text{DIV}(\mathbf{u}^h)\ _0$	LS	0.489E+00	0.273E+00	0.186E+00	0.553
	RLS	0.463E+00	0.233E+00	0.116E+00	1.003

a whole, the standard version of the least-squares method cannot be deemed robust enough for general quadrilateral grids.

TABLE 6.8

Error data and estimated orders of convergence for the standard least-squares method (LS) on sinusoidal grids: Problem (1.1) with ($\sigma = 1$) and without ($\sigma = 0$) reaction term.

Error	Method	33×33	65×65	129×129	Order
$\ p - p^h\ _0$	$\sigma = 0$	0.598E-03	0.151E-03	0.379E-04	1.996
	$\sigma = 1$	0.158E-03	0.400E-04	0.100E-04	1.996
$\ \nabla(p - p^h)\ _0$	$\sigma = 0$	0.184E-01	0.902E-02	0.449E-02	1.007
	$\sigma = 1$	0.179E-01	0.896E-02	0.448E-02	1.000
$\ \mathbf{u} - \mathbf{u}^h\ _0$	$\sigma = 0$	0.638E-01	0.318E-01	0.159E-01	1.001
	$\sigma = 1$	0.638E-01	0.318E-01	0.159E-01	1.001
$\ \nabla \cdot \mathbf{u} - \nabla \cdot \mathbf{u}^h\ _0$	$\sigma = 0$	0.576E+00	0.288E+00	0.144E+00	1.000
	$\sigma = 1$	0.576E+00	0.288E+00	0.144E+00	1.000
$\ \nabla \cdot \mathbf{u} - \text{DIV}(\mathbf{u}^h)\ _0$	$\sigma = 0$	0.541E+00	0.271E+00	0.135E+00	1.000
	$\sigma = 1$	0.541E+00	0.271E+00	0.135E+00	1.000

As expected, the mimetic reformulation of the least-squares method completely eliminates these problems. From the data in Tables 6.4–6.7 we see that the reformulation restores the optimal order of convergence for all variables regardless of whether or not the reaction terms are included.

Finally, Table 6.8 shows error and convergence data for the standard least-squares method on sinusoidal grids. We see that despite the highly distorted nature of this grid, the fact that most of its elements remain close to affine quads is enough to restore convergence rates for all variables.

7. Conclusions. The mimetic reformulation proposed in this paper is a simple yet effective approach to restoring convergence of finite element methods that employ the lowest-order quadrilateral Raviart–Thomas elements.

By proving that the reformulation of the mixed method is equivalent to its standard version, we establish that the loss of convergence in this method is benign and can be avoided by using DIV to compute the divergence of the velocity approximation.

Our results also show that the deterioration of accuracy in the least-squares method is real. For problems with a reaction term it is confined to the velocity approximation, but without this term, the loss of convergence spreads to both variables. The mimetic reformulation of the least-squares method mitigates convergence problems and should be used whenever computations with this method involve nonaffine quadrilateral grids.

Acknowledgment. We thank the anonymous referees for several suggestions that helped to improve this paper and prompted us to include more informative numerical examples and data.

REFERENCES

- [1] T. ARBOGAST, C. DAWSON, P. T. KEENAN, M. F. WHEELER, AND I. YOTOV, *Enhanced cell-centered finite differences for elliptic equations on general geometry*, SIAM J. Sci. Comput., 19 (1998), pp. 404–425.
- [2] V. ARNOLD, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, New York, 1989.
- [3] D. N. ARNOLD, D. BOFFI, AND R. S. FALK, *Approximation by quadrilateral finite elements*, Math. Comp., 71 (2002), pp. 909–922.
- [4] D. N. ARNOLD, D. BOFFI, AND R. S. FALK, *Quadrilateral $H(\text{div})$ finite elements*, SIAM J. Numer. Anal. 42 (2005), pp. 2429–2451.

- [5] M. BERNDT, K. LIPNIKOV, M. SHASHKOV, M. F. WHEELER, AND I. YOTOV, *Superconvergence of the velocity in mimetic finite difference methods on quadrilaterals*, SIAM J. Numer. Anal., 43 (2005), pp. 1728–1749.
- [6] P. B. BOCHEV AND M. D. GUNZBURGER, *Finite element methods of least-squares type*, SIAM Rev., 40 (1998), pp. 789–837.
- [7] P. B. BOCHEV AND M. D. GUNZBURGER, *On least-squares finite element for the Poisson equation and their connection to the Dirichlet and Kelvin principles*, SIAM J. Numer. Anal., 43 (2005), pp. 340–362.
- [8] P. BOCHEV AND M. GUNZBURGER, *A locally conservative least-squares method for Darcy flows*, Comm. Numer. Methods Engrg., 24 (2008), pp. 97–110.
- [9] D. BOFFI, F. KIKUCHI, AND J. SCHBERL, *Edge element computation of Maxwell's eigenvalues on general quadrilateral meshes*, Math. Models Methods Appl. Sci., 16 (2006), pp. 265–273.
- [10] D. BOFFI AND L. GASTALDI, *Some Remarks on Quadrilateral Mixed Finite Elements*, submitted to the Fifth MIT Conference on Computational Fluid and Solid Mechanics.
- [11] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
- [12] P. CIARLET, *The Finite Element Method for Elliptic Problems*, Classics in Applied Mathematics 40, SIAM, Philadelphia, 2002.
- [13] G. FIX, M. GUNZBURGER, AND R. NICOLAIDES, *On finite element methods of the least-squares type*, Comput. Math. Appl., 5 (1979), pp. 87–98.
- [14] V. GIRAULT AND P. RAVIART, *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms*, Springer-Verlag, Berlin, 1986.
- [15] J. HYMAN AND M. SHASHKOV, *Natural discretizations for the divergence, gradient, and curl on logically rectangular grids*, Comput. Math. Appl., 33 (1997), pp. 81–104.
- [16] J. HYMAN AND M. SHASHKOV, *Adjoint operators for the natural discretizations of the divergence, gradient and curl on logically rectangular grids*, Appl. Numer. Math., 25 (1997), pp. 413–442.
- [17] YU. KUZNETSOV AND S. REPIN, *Convergence analysis and error estimates for mixed finite element method on distorted meshes*, J. Numer. Math., 13 (2005), pp. 33–51.
- [18] L. G. MARGOLIN AND M. SHASHKOV, *Second-order sign-preserving conservative interpolation (remapping) on general grids*, J. Comput. Phys., 184 (2003), pp. 266–298.
- [19] P. A. RAVIART AND J. M. THOMAS, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of Finite Element Method, I. Galligani and E. Magenes, eds., Lecture Notes in Math. 606, Springer, Berlin, 1977, pp. 292–315.
- [20] M. SHASHKOV, *Conservative Finite Difference Methods on General Grids*, CRC Press, Boca Raton, FL, 1996.
- [21] J. SHEN, *Mixed Finite Element Methods on Distorted Rectangular Grids*, Technical report ISC-94-13-MATH, Texas A & M University, College Station, TX, 1994.
- [22] M. WHEELER AND I. YOTOV, *A cell-centered finite difference method on quadrilaterals*, in Compatible Spatial Discretizations, D. Arnold, P. Bochev, R. Lehoucq, R. Nicolaides, M. Shashkov, eds., IMA Vol. Math. Appl. 142, Springer, New York, 2006, pp. 189–207.

LOW ORDER DISCONTINUOUS GALERKIN METHODS FOR SECOND ORDER ELLIPTIC PROBLEMS*

E. BURMAN[†] AND B. STAMM[‡]

Abstract. We consider DG-methods for second order scalar elliptic problems using piecewise affine approximation in two or three space dimensions. We prove that both the symmetric and the nonsymmetric versions of the DG-method have regular system matrices without penalization of the interelement solution jumps provided boundary conditions are imposed in a certain weak manner. Optimal convergence is proved for sufficiently regular meshes and data. We then propose a DG-method using piecewise affine functions enriched with quadratic bubbles. Using this space we prove optimal convergence in the energy norm for both a symmetric and nonsymmetric DG-method without stabilization. All of these proposed methods share the feature that they conserve mass locally independent of the penalty parameter.

Key words. discontinuous Galerkin, elliptic equation, Crouzeix–Raviart approximation, interior penalty, local mass conservation

AMS subject classifications. 65M160, 65M15

DOI. 10.1137/070685105

1. Introduction. The discontinuous Galerkin (DG) method for $(2n)$ th order elliptic problems was introduced by Baker [2] with special focus on the fourth order case. In parallel, the interior penalty method of Douglas and Dupont for second order elliptic problems [9] led to the symmetric interior penalty DG-method (SIPG-method) proposed by Wheeler [15] and Arnold [1]. In the SIPG-method, a penalty term on the solution jumps between adjacent elements and has to be introduced to ensure coercivity of the bilinear form.

In the nineties, Babuska, Baumann, and Oden proposed a nonsymmetric method for elliptic problems with a less stiff penalty term [12] (NIPG). The DG-methods for second order elliptic problems have been further analyzed in the works by Girault, Rivière, and Wheeler [13] and Larson and Niklasson [11]. In these papers the authors proved that in the nonsymmetric case, when using high order polynomial approximation, optimal convergence can be obtained without any penalization of the solution jumps. In a recent paper, Brezzi and Marini [6] showed that enriching the piecewise affine discontinuous finite element space by some nonconforming quadratic bubbles yields a space which is the smallest one for which the nonsymmetric version converges optimally without interior penalty.

For a review of discontinuous Galerkin methods for elliptic problems, we refer the reader to Arnold [1], and for a review of stabilization mechanisms in discontinuous Galerkin methods, we refer the reader to Brezzi et al. [5].

One of the advantages of the DG-method is that it has enhanced local conservation compared to the continuous Galerkin method. However, only in the case of the

*Received by the editors March 13, 2007; accepted for publication (in revised form) July 23, 2008; published electronically December 19, 2008. This project received financial support from the Swiss National Science Foundation under grant 200021-113304.

<http://www.siam.org/journals/sinum/47-1/68510.html>

[†]Department of Mathematics, University of Sussex, BN1 9RF Brighton, United Kingdom (e.n.burman@sussex.ac.uk).

[‡]Institute of Analysis and Scientific Computing, Swiss Institute of Technology, Lausanne, CH-1015, Switzerland (benjamin.stamm@epfl.ch).

nonsymmetric DG-method, without penalization of the solution jumps, is the conservation independent of the stabilization parameter. The nonsymmetric formulation, however, is not adjoint consistent and one may not analyze L^2 -convergence using the Nitsche trick.

In this paper we discuss the relation between stabilization, existence of discrete solution, and optimal convergence, in the case of scalar second order elliptic problems. The aim is to design a low order DG-method that

1. has optimal convergence in the DG-energy norm (including both the broken H^1 semi-norm and the solution jumps over element faces) and the L^2 -norm,
2. is locally massconservative independently of the penalty parameter.

We will show that for the symmetric DG-method the only thing required to guarantee the existence of a solution to the discrete system is to impose boundary conditions in the same way as for the Crouzeix–Raviart nonconforming finite elements. No interelement penalization of the solution jumps is needed. Under some assumptions on the mesh and on the data, optimal convergence is obtained as well. Indeed, the condition is either that the mesh is uniform in the asymptotic limit and that the right-hand side is smooth enough or that the mesh satisfies a certain macroelement property.

If these conditions are not satisfied, then the convergence of the solution jumps can be perturbed by the appearance of a checkerboard mode that vanishes too slowly in the absence of interior penalty. We exemplify the checkerboard mode numerically and show how it is quenched by penalization.

To reduce the constraints on the mesh we enrich the space with quadratic nonconforming bubble functions. These stabilizing bubbles eliminate the checkerboard mode. We prove optimal convergence in the energy norm without stabilization both in the symmetric and nonsymmetric case. In the symmetric case we obtain additionally optimal convergence in the L^2 -norm. In both cases the analysis relies on a discrete inf-sup condition drawing from earlier ideas on minimal stabilization for DG-methods in [8, 7, 10, 11].

2. The problem setting. Let Ω be an open, bounded, and convex polygon (polyhedron in three space dimensions) in \mathbb{R}^d , $d = 2, 3$, with outer unit normal n . Let \mathcal{K} be a subdivision of $\Omega \subset \mathbb{R}^d$ into nonoverlapping d -simplices κ . An element $\kappa \in \mathcal{K}$ is assumed to be an closed set. We consider the following elliptic problem with homogeneous Dirichlet boundary conditions.

Find $u : \Omega \rightarrow \mathbb{R}$ such that

$$(2.1) \quad \begin{cases} -\nabla \cdot \sigma \nabla u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

where $f \in L^2(\Omega)$ and with a diffusion coefficient that is piecewise constant on each element $\sigma(x)|_\kappa = \sigma_\kappa \in \mathbb{R}$ for all $\kappa \in \mathcal{K}$ and $\sigma(x) > \sigma_0 > 0$. We assume that there exists a constant $\rho > 0$ such that $\sigma|_{\kappa_1} \leq \rho \sigma|_{\kappa_2}$ for any two elements satisfying $\partial\kappa_1 \cap \partial\kappa_2 \neq \emptyset$. The fact that the boundary conditions are of homogeneous Dirichlet type is not a limitation of the presented methods but rather to avoid technical details.

Let \mathcal{F}_i denote the set of interior faces ($(d - 1)$ -manifolds) of the mesh; i.e., the set of faces that are not included in the boundary $\partial\Omega$. The set \mathcal{F}_e denotes the faces that are included in $\partial\Omega$ and defines $\mathcal{F} = \mathcal{F}_i \cup \mathcal{F}_e$. Note that for an element $\kappa \in \mathcal{K}$, $\mathcal{F}(\kappa)$ denotes the set of faces of κ . Furthermore, denote by Γ the skeleton of the mesh, i.e., the set of points belonging to faces, $\Gamma = \{x \in \overline{\Omega} : \exists F \in \mathcal{F} \text{ s.t. } x \in F\}$.

Assume that \mathcal{K} is shape-regular, does not contain any hanging node, and covers $\overline{\Omega}$ exactly. Suppose that each $\kappa \in \mathcal{K}$ is an affine image of the reference element $\hat{\kappa}$; i.e., for each element κ there exists an affine transformation $T_\kappa : \hat{\kappa} \rightarrow \kappa$. For an element $\kappa \in \mathcal{K}$, h_κ denotes its diameter. Set $h = \max_{\kappa \in \mathcal{K}} h_\kappa$, and let \tilde{h}, \tilde{m} be the functions such that $\tilde{h}|_\kappa = h_\kappa$ (resp., $\tilde{m}|_\kappa = \text{meas}_d(\kappa)$). We will say that a family of subdivisions $\{\mathcal{K}\}_h$ is asymptotically ζ -uniform with some $\zeta > 0$ if there exists a constant $c > 0$ such that for every mesh \mathcal{K} and every $F \in \mathcal{F}_i$ there holds $|\tilde{m}|_{\kappa_1} - \tilde{m}|_{\kappa_2}| \leq c \text{meas}_{d-1}(F)^\zeta \min(\tilde{m}|_{\kappa_1}, \tilde{m}|_{\kappa_2})$, where $F = \partial\kappa_1 \cap \partial\kappa_2$ with $\kappa_1, \kappa_2 \in \mathcal{K}$.

For a face $F \in \mathcal{F}$, h_F denotes its diameter, and let \tilde{h}_F be the function such that $\tilde{h}_F|_F = h_F$.

For a nonempty subdomain $R \subset \Omega$ or $R \subset \Gamma$, $(\cdot, \cdot)_R$ denotes the $L^2(R)$ -scalar product, $\|\cdot\|_R = (\cdot, \cdot)_R^{1/2}$ the corresponding norm, and $\|\cdot\|_{s,R}$ the $H^s(R)$ -norm. The elementwise counterparts will be distinguished using the discrete partition as subscript, for example $(\cdot, \cdot)_\mathcal{K} = \sum_{\kappa \in \mathcal{K}} (\cdot, \cdot)_\kappa$. For $s \geq 1$, let $H^s(\mathcal{K})$ be the space of piecewise Sobolev H^s -functions and denote its norm by $\|\cdot\|_{s,\mathcal{K}}$.

For $v \in H^1(\mathcal{K})$, $\tau \in [H^1(\mathcal{K})]^d$ and an interior face $F = \kappa_1 \cap \kappa_2 \in \mathcal{F}_i$, where κ_1 and κ_2 are two distinct elements of \mathcal{K} with respective outer unit normals n_1 and n_2 , define the jump and average by

$$\begin{aligned} [v] &= (v|_{\kappa_1} n_1 + v|_{\kappa_2} n_2), & \{v\} &= \frac{1}{2} (v|_{\kappa_1} + v|_{\kappa_2}), \\ [\tau] &= (\tau|_{\kappa_1} \cdot n_1 + \tau|_{\kappa_2} \cdot n_2), & \{\tau\} &= \frac{1}{2} (\tau|_{\kappa_1} + \tau|_{\kappa_2}). \end{aligned}$$

Additionally we define on each face $F \in \mathcal{F}$ the unit normal n_F in an arbitrary but fixed manner.

On outer faces $F = \partial\kappa \cap \partial\Omega \in \mathcal{F}_e$, for some $\kappa \in \mathcal{K}$ with outer unit normal n , the jump and the average are defined as $[v] = v|_F n$ and $\{v\} = v|_F$ (resp., $[\tau] = \tau|_F \cdot n$ and $\{\tau\} = \tau|_F$). The projection of $\{v\}$ and $[v]$ onto the space of facewise constant functions are denoted by $\overline{\{v\}} \in \mathbb{R}$ and $\overline{[v]} \in \mathbb{R}^d$ and defined by

$$\int_F \overline{\{v\}} \, ds = \int_F \{v\} \, ds \quad \text{and} \quad \int_F \overline{[v]} \, ds = \int_F [v] \, ds$$

for all $F \in \mathcal{F}$.

The shape-regularity implies that there exists a constant $c > 0$ independent of the mesh size h such that on any face $F \in \mathcal{F}$,

$$h_F \leq \{\tilde{h}\} \leq c h_F.$$

In this paper $c > 0$ denotes a generic constant and can change at each occurrence, while an indexed constant stays fixed. Any constant is independent of the mesh size h but not necessarily of σ .

3. Finite element spaces. We will consider two low order finite element spaces, the space of piecewise affine discontinuous functions, and the space of piecewise affine discontinuous functions enriched with nonconforming quadratic bubbles. We show that every function in the former space can be written as a sum of a midpoint continuous function (in the Crouzeix–Raviart space) and a “midpoint discontinuous” function in a space that will be specified later. The motivation for this decomposition is that such a choice of basis in the symmetric DG-bilinear form results in a block diagonal matrix, and hence the continuous and the discontinuous contributions may be analyzed separately.

Define the piecewise affine discontinuous finite element space by

$$V_h^1 = \{v_h \in L^2(\Omega) : v_h|_\kappa \in \mathbb{P}_1(\kappa) \forall \kappa \in \mathcal{K}\}.$$

Then, introduce the enriched space

$$V_h^b = V_h^1 \oplus V^b,$$

with

$$V^b = \{v \in L^2(\Omega) : v(x)|_\kappa = \alpha_\kappa x \cdot x, \text{ where } \alpha_\kappa \in \mathbb{R}\},$$

where $x = (x_1, \dots, x_d)$ denotes the physical variables. Observe that V_h^b is the space introduced by Brezzi and Marini in [6]. Additionally, we define

$$V_{h,0}^1 = \left\{ v_h \in V_h^1 : \int_F v_h ds = 0 \forall F \in \mathcal{F}_e \right\},$$

the space of piecewise affine elements where the homogeneous Dirichlet boundary conditions are imposed on the midpoints of each exterior face.

3.1. Splitting of the finite element space $V_{h,0}^1$. The idea is to split $V_{h,0}^1$ into a midpoint continuous space, the Crouzeix–Raviart space, and a midpoint discontinuous space, similar to what was proposed in the one-dimensional case in [10]. Recall the definition of the Crouzeix–Raviart space

$$V^C = \left\{ v_h \in V_{h,0}^1 : \int_F [v_h] ds = 0 \forall F \in \mathcal{F}_i \right\}.$$

Its “midpoint discontinuous” counterpart is defined by

$$V^D = \left\{ v_h \in V_{h,0}^1 : \int_F \{v_h\} ds = 0 \forall F \in \mathcal{F}_i \right\}.$$

Denote by N_{int} the number of interior faces of the mesh \mathcal{K} . Let us denote $\{\phi_i^c\}_{i=1}^{N_{int}}$, the Crouzeix–Raviart basis defined such that

$$\int_{F_j} \{\phi_i^c\} ds = \delta_{i,j} \text{meas}_{d-1}(F_i) \quad \forall 1 \leq i, j \leq N_{int}.$$

This defines a basis for the space V^C . Denote by $x_F \in F$ the midpoint of $F \in \mathcal{F}$ for a one-dimensional face and the barycenter (which coincides with the Gauss point) of F for a two-dimensional face. Recall that the midpoint integration rule on F is exactly of order two in the one-dimensional case and of order one in the two-dimensional case. Thus $v_c \in V^C$ is midpoint continuous, i.e.,

$$[v_c](x_F) = \frac{1}{\text{meas}_{d-1}(F)} \int_F [v_c] ds = 0 \quad \forall F \in \mathcal{F},$$

in two and three space dimensions. Now, let us define a basis for the space V^D . For each face $F_i \in \mathcal{F}_i$ consider the basis function ϕ_i^d defined by

$$\phi_i^d = \frac{1}{2} \phi_i^c \frac{\nabla \phi_i^c \cdot n_{F_i}}{|\nabla \phi_i^c \cdot n_{F_i}|}$$

with n_{F_i} a fixed but arbitrary unit normal associated to the face F_i . It is easy to verify that $\phi_i^d \in V^D$ and that

$$\int_{F_j} [\phi_i^d] \cdot n_{F_j} ds = \delta_{i,j} \text{meas}_{d-1}(F_i) \quad \forall F_i \in \mathcal{F}.$$

Now we are ready to prove the following lemma.

LEMMA 3.1. *The splitting of $V_{h,0}^1$ into V^C and V^D is a direct sum; i.e., $V_{h,0}^1 = V^C \oplus V^D$. Any function v_h in $V_{h,0}^1$ can be written as*

$$v_h(x) = \sum_{i=1}^{N_{int}} c_i \phi_i^c(x) + d_i \phi_i^d(x),$$

where $c_i = \frac{1}{\text{meas}_{d-1}(F_i)} \int_{F_i} \{v_h\} ds$ and $d_i = \frac{1}{\text{meas}_{d-1}(F_i)} \int_{F_i} [v_h] \cdot n_{F_i} ds$.

Remark 3.2. Note that, for $v \in H^1(\Omega)$ the function

$$(3.1) \quad i_c v(x) = \sum_{i=1}^{N_{int}} c_i \phi_i^c(x)$$

with $c_i = \frac{1}{\text{meas}_{d-1}(F_i)} \int_{F_i} v ds$ is the Crouzeix–Raviart interpolant and has optimal approximation properties.

Proof. First, assume that $v_h \in V^C \cap V^D$ and show that $v_h \equiv 0$. This is an immediate consequence of the properties of V^C and V^D . One may then show that $V^C \oplus V^D$ covers $V_{h,0}^1$ entirely. This is easily proven by taking $v_h \in V_{h,0}^1$ and defining $w_h \in V^C \oplus V^D$ by

$$w_h(x) = \sum_{i=1}^{N_{int}} c_i \phi_i^c(x) + d_i \phi_i^d(x)$$

with $c_i = \frac{1}{\text{meas}_{d-1}(F_i)} \int_{F_i} \{v_h\} ds$ and $d_i = \frac{1}{\text{meas}_{d-1}(F_i)} \int_{F_i} [v_h] \cdot n_{F_i} ds$. Using the properties of the basis functions $\{\phi_i^c\}_{i=1}^{N_{int}}$ and $\{\phi_i^d\}_{i=1}^{N_{int}}$ it follows that $w_h \equiv v_h$. \square

LEMMA 3.3 (asymptotic L^2 -orthogonality between V^C and V^D). *Assume that the mesh is asymptotically ζ -uniform with some $\zeta > 0$, then the spaces V^C and V^D satisfy the following weak L^2 -orthogonality property: there exists a constant $c > 0$ independent of h , such that*

$$|(v_c, v_d)_\mathcal{K}| \leq ch^\zeta \|v_c\|_\mathcal{K} \|v_d\|_\mathcal{K} + r(d) \|\tilde{h} \nabla v_c\|_\mathcal{K} \|\tilde{h} \nabla v_d\|_\mathcal{K} \quad \forall v_c \in V^C, v_d \in V^D,$$

where $r(2) = 0$ and $r(3) = c$.

Proof. In the two-dimensional case we can proceed as follows. Take $v_c \in V^C$ and $v_d \in V^D$ and develop

$$(3.2) \quad |(v_c, v_d)_\mathcal{K}| = \left| \sum_{\kappa \in \mathcal{K}} (v_c, v_d)_\kappa \right| = \frac{1}{3} \left| \sum_{\kappa \in \mathcal{K}} \sum_{F \in \mathcal{F}(\kappa)} \tilde{m}|_\kappa v_c(x_F) v_d(x_F) \right|$$

with numerical integration on κ using the midpoints/barycenters of its faces as integration points (x_F denotes the midpoint/barycenter of the face F). This numerical

integration on κ is exact of order 2 in the two-dimensional case but not in the three-dimensional one. Since $v_c(x_F) = v_d(x_F) = 0$ for exterior faces $F \in \mathcal{F}_e$ and since v_c is midpoint continuous we can rearrange the sum

$$|(v_c, v_d)_\mathcal{K}| = \frac{2}{3} \left| \sum_{F \in \mathcal{F}_i} v_c(x_F) \{ \tilde{m} v_d \}(x_F) \right|.$$

Using the equality $\{wv\} = \{w\}\{v\} + \frac{1}{4}[w] \cdot [v]$ and the fact that $\{v_d\}(x_F) = 0$ for all interior faces yields

$$|(v_c, v_d)_\mathcal{K}| \leq c \sum_{F \in \mathcal{F}_i} |v_c(x_F)| |[\tilde{m}](x_F) \cdot n_F| |[v_d](x_F) \cdot n_F|.$$

The regularity assumption on the mesh implies that $|[\tilde{m}] \cdot n_F| \leq ch_F^\zeta \{ \tilde{m} \}$ with $\zeta > 0$, and since $\frac{1}{2}|[v_d](x_F) \cdot n_F| = |v_d|_{\kappa_i}(x_F)|$, for $i = 1, 2$, we can rearrange the sum again,

$$|(v_c, v_d)_\mathcal{K}| \leq ch^\zeta \sum_{\kappa \in \mathcal{K}} (|v_c|, |v_d|)_\kappa \leq ch^\zeta \|v_c\|_\mathcal{K} \|v_d\|_\mathcal{K}.$$

In the three-dimensional case, since the numerical integration is no longer of order two, we introduce the local midpoint interpolation $i_h^\kappa : H^2(\kappa) \rightarrow \mathbb{P}_1(\kappa)$ for each element $\kappa \in \mathcal{K}$ by $i_h^\kappa v(x_F) = v(x_F)$ for all $F \in \mathcal{F}(\kappa)$. Then using the triangle inequality yields

$$(3.3) \quad |(v_c, v_d)_\mathcal{K}| \leq \left| \sum_{\kappa \in \mathcal{K}} \int_\kappa i_h^\kappa(v_c v_d) dx \right| + \left| \sum_{\kappa \in \mathcal{K}} \int_\kappa (v_c v_d - i_h^\kappa(v_c v_d)) dx \right|.$$

The first term of the right-hand side of (3.3) can be developed as in (3.2) since now for the local midpoint interpolation i_h^κ the above defined numerical integration rule is exact. For the second term of the right-hand side of (3.3), one can show using standard interpolation results that

$$\left| \sum_{\kappa \in \mathcal{K}} \int_\kappa (v_c v_d - i_h^\kappa(v_c v_d)) dx \right| \leq c \|\tilde{h} \nabla v_c\|_\mathcal{K} \|\tilde{h} \nabla v_d\|_\mathcal{K}. \quad \square$$

3.2. Properties of the enriched space V_h^b . The motivation for the particular form of the enriched space is given in the following lemma. The key idea is that the gradient of a function in V_h^b restricted to an element is locally in the Raviart–Thomas space. Let RT_0 denote the space of Raviart–Thomas elements of order zero.

LEMMA 3.4. *For all $w_h \in V_h^b$ there holds*

$$\nabla w_h|_\kappa \in RT_0(\kappa),$$

and for all $r_h \in RT_0(\kappa)$ there exists $w_h \in V_h^b$ such that $\nabla w_h|_\kappa = r_h$ for all $\kappa \in \mathcal{K}$.

Proof. Let $w_h \in V_h^b$, restricting w_h to an arbitrary element κ we can write

$$w_h|_\kappa(x) = \alpha x \cdot x + \beta \cdot x + \gamma,$$

where $\alpha, \gamma \in \mathbb{R}$ and $\beta \in \mathbb{R}^d$ are the local degrees of freedom. Then

$$\nabla w_h|_\kappa(x) = 2\alpha x + \beta.$$

To show that this function lies in the Raviart–Thomas finite element space we have to map it on the reference element using the *Piola* transformation. But let us first introduce the *affine* transformation T_κ between the reference element $\hat{\kappa}$ defined by its vertices $a_i = e_i$ for $i = 1, \dots, d$ and $a_{d+1} = \mathcal{O}$ and the physical element κ . The vectors e_i denote the unit vectors corresponding to the i th coordinate. The affine transformation may be written as

$$T_\kappa(\hat{x}) = J_\kappa \hat{x} + t_\kappa,$$

where $\hat{x} = (\hat{x}_1, \dots, \hat{x}_d)^\top \in \hat{\kappa}$ denotes the variable in the reference element. Then we denote by ψ_κ the Piola transformation between the physical element and the reference element defined by

$$\psi_\kappa(v)(\hat{x}) = |J_\kappa| J_\kappa^{-1} v(T_\kappa(\hat{x})).$$

Thus

$$\psi_\kappa(\nabla w_h|_\kappa)(\hat{x}) = |J_\kappa| J_\kappa^{-1} (2\alpha T_\kappa(\hat{x}) + \beta) = |J_\kappa| (2\alpha \hat{x} + J_\kappa^{-1}(\beta + 2\alpha t_\kappa)),$$

and this function is clearly an element of the Raviart–Thomas finite element space on the reference element.

On the other hand if $r_h \in RT_0$, then $\psi_\kappa \circ r_h|_\kappa$ is of the form

$$\psi_\kappa(r_h|_\kappa)(\hat{x}) = a\hat{x} + b,$$

where $a \in \mathbb{R}$ and $b = (b_1, \dots, b_d)^\top \in \mathbb{R}^d$. Thus

$$r_h|_\kappa(x) = \frac{1}{|J_\kappa|} (ax + J_\kappa b - t_\kappa).$$

Defining locally,

$$w_h|_\kappa(x) = \frac{1}{|J_\kappa|} \left(\frac{a}{2} x \cdot x + (J_\kappa b - t_\kappa) \cdot x \right)$$

yields that

$$\nabla w_h|_\kappa(x) = r_h|_\kappa(x). \quad \square$$

4. Poincaré inequalities. The analysis of the model problem (2.1) relies on the Poincaré inequality. In this section we state the Poincaré inequalities that hold for the two spaces V^C and V^D , separately. Then we prove a stronger Poincaré inequality for the space V^D under some assumptions on the mesh.

LEMMA 4.1. *There is a constant $c > 0$ independent of h such that for all $v_h \in V_h^b$ there holds*

$$c \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} [v_h]\|_{\mathcal{F}}^2 \leq \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} \overline{[v_h]}\|_{\mathcal{F}}^2 + \|\sigma^{\frac{1}{2}} \nabla v_h\|_{\mathcal{K}}^2,$$

where $\omega|_F = \max(\sigma|_{\kappa_1}, \sigma|_{\kappa_2})$ for $F = \partial\kappa_1 \cap \partial\kappa_2$.

Proof. The proof is completed immediately by the approximation properties of the average jump, a discrete trace inequality, and the bounded variation of σ over faces. \square

COROLLARY 4.2. *The following Poincaré inequality for broken H^1 -spaces holds for all $v_h \in V_h^b$:*

$$c \|\sigma^{\frac{1}{2}} v_h\|_{\mathcal{K}}^2 \leq \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} \overline{[v_h]}\|_{\mathcal{F}}^2 + \|\sigma^{\frac{1}{2}} \nabla v_h\|_{\mathcal{K}}^2.$$

Proof. An immediate consequence of the previous lemma and the Poincaré inequality is

$$(4.1) \quad c \|\sigma^{\frac{1}{2}} v_h\|_{\mathcal{K}}^2 \leq \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} [v_h]\|_{\mathcal{F}}^2 + \|\sigma^{\frac{1}{2}} \nabla v_h\|_{\mathcal{K}}^2$$

proved by Brenner [3]. \square

PROPOSITION 4.3 (Poincaré inequality for V^C). *There exists a constant $c > 0$ depending only on Ω such that, for all $h < 1$,*

$$\forall v_c \in V^C, \quad c \|v_c\|_{\mathcal{K}} \leq \|\nabla v_c\|_{\mathcal{K}}.$$

Proof. See Temam [14] for the proof. \square

PROPOSITION 4.4 (Poincaré inequality for V^D). *There exists a constant $c > 0$ depending only on Ω such that, for all $h < 1$,*

$$\forall v_d \in V^D, \quad c \|v_d\|_{\mathcal{K}} \leq \|\nabla v_d\|_{\mathcal{K}}.$$

Proof. Let $v_d \in V^D$ be fixed. Then, define the splitting of Ω into two parts \mathcal{K}_1 and \mathcal{K}_2 by

$$\begin{aligned} \mathcal{K}_1 &= \{\kappa \in \mathcal{K} : \exists x \in \kappa \text{ s.t. } v_d(x) = 0\}, \\ \mathcal{K}_2 &= \{\kappa \in \mathcal{K} : v_d(x) \neq 0 \forall x \in \kappa\}. \end{aligned}$$

First, prove the inequality for the region \mathcal{K}_1 . Fix an element $\kappa_1 \in \mathcal{K}_1$ and define

$$Z(\kappa_1) = \{x \in \kappa_1 : v_d(x) = 0\}.$$

Since $v_d|_{\kappa_1} \in \mathbb{P}_1(\kappa_1)$ we may write

$$v_d(x) = \nabla v_d \cdot (x - x^*) \quad \text{with } x^* \in Z(\kappa_1).$$

Thus we conclude immediately that

$$(4.2) \quad \|v_d\|_{\mathcal{K}_1} \leq \|\tilde{h} \nabla v_d\|_{\mathcal{K}_1}.$$

Second, split \mathcal{K}_2 in maximal subsets $\{\mathcal{K}_2^j\}_{j=1}^m$ in order that each

$$\Omega_2^j = \left(\overset{\circ}{\bigcup}_{\kappa \in \mathcal{K}_2^j} \kappa \right)$$

is connected. Fix a subset \mathcal{K}_2^j and observe that $|v_d|$ is midpoint continuous on interior faces of \mathcal{K}_2^j . In consequence, $|v_d|$ lies in the Crouzeix–Raviart space over the domain Ω_2^j , and we may proceed analogously to the proof of Proposition 4.3. Details are left to the reader. \square

In case the mesh has a certain macroelement structure we may prove a stronger Poincaré inequality for the space V^D .

PROPOSITION 4.5 (strong Poincaré inequality for V^D). *Let \mathcal{K} be a mesh. Assume that there exists a coarse mesh \mathcal{T} covering $\bar{\Omega}$ such that each macroelement (d -simplex) $T \in \mathcal{T}$ contains exactly $d+1$ elements $\kappa_1, \dots, \kappa_{d+1}$ of \mathcal{K} and such that $\kappa_i \cap \kappa_j \cap \overset{\circ}{T} \neq \emptyset$ for all $1 \leq i, j \leq d+1$. Then the following inequality holds:*

$$\forall v_d \in V^D, \quad c \|v_d\|_{\mathcal{K}} \leq \|\tilde{h} \nabla v_d\|_{\mathcal{K}}.$$

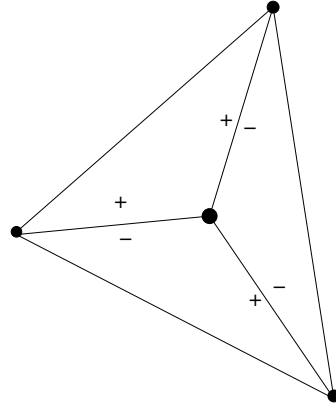


FIG. 1. Illustration of the macroelement argument of Lemma 4.5. The “+” and “-” signs refer to the sign of v_d at the midpoints of the faces.

Proof. Let $v_d \in V^D$ and fix an element $\kappa \in \mathcal{K}$. If there exists $x^* \in \kappa$ such that $v_d(x^*) = 0$, then we conclude analogous to (4.2) that

$$\|v_d\|_\kappa \leq \|h_\kappa \nabla v_d\|_\kappa.$$

Otherwise there exists in the same macroelement $T \in \mathcal{T}$ a neighbor element $\kappa' \in \mathcal{K}$ such that there exists $x^* \in \kappa'$ with $v_d(x^*) = 0$. Indeed assume that $v_d(x) \neq 0$ for all $x \in T$. Observe that v_d changes sign in the midpoint of each face since it lies in V^D and hence $\int_F \{v_d\} = 0$. Consider all three elements of T in two dimensions and an arbitrary selection of three elements containing κ' in three dimensions. The solution changes sign over each face. However, in the three elements the sign has to change four times. Hence it has to change sign within one element, which leads to a contradiction. See the illustration in Figure 1. Thus there exists at least one element $\kappa^* \in \mathcal{K}$ of the macroelement T such that there exists a point $x^* \in \kappa^*$ with $v_d(x^*) = 0$. Since $\kappa \cap \kappa^* \neq \emptyset$, we conclude that

$$\|v_d\|_\kappa \leq c \|h_\kappa \nabla v_d\|_\kappa. \quad \square$$

Remark 4.6. Observe that the above defined macroelement property is sufficient but not necessary for the strong Poincaré inequality to hold. The sufficient and necessary condition on the mesh is: For each element $\kappa \in \mathcal{K}$ there exists a path in an h -neighborhood of κ , starting and ending at κ , passing an odd number of faces.

5. Discontinuous Galerkin methods. Define the following bilinear form:

$$(5.1) \quad a_s(u_h, v_h) = (\sigma \nabla u_h, \nabla v_h)_\mathcal{K} - (\{\sigma \nabla u_h\}, [v_h])_\mathcal{F} - s (\{\sigma \nabla v_h\}, [u_h])_\mathcal{F}$$

for $s \in \{-1, 1\}$ and the stabilization term

$$j(u_h, v_h) = (\omega \tilde{h}_\mathcal{F}^{-1} \overline{[u_h]}, \overline{[v_h]})_{\mathcal{F}_i},$$

where $\omega|_F = \max(\sigma|_{\kappa_1}, \sigma|_{\kappa_2})$ for $F = \partial\kappa_1 \cap \partial\kappa_2$. Note that we only penalize the average value of the jumps in the spirit of Lemma 4.1. Let us define two methods to approximate the solution of (2.1).

Reduced interior penalty (RIP-) method. Find $u_h^1 \in V_{h,0}^1$ such that

$$(5.2) \quad a_s(u_h^1, v_h) + \gamma j(u_h^1, v_h) = (f, v_h)_\mathcal{K} \quad \forall v_h \in V_{h,0}^1,$$

for some $\gamma \in \mathbb{R}$ and $s \in \{-1, 1\}$.

Bubble stabilized discontinuous Galerkin (BSDG-) method. Find $u_h^b \in V_h^b$ such that, with $s \in \{-1, 1\}$,

$$(5.3) \quad a_s(u_h^b, v_h) = (f, v_h)_{\mathcal{K}} \quad \forall v_h \in V_h^b.$$

Remark 5.1 (local mass conservation property). The solutions u_h^1, u_h^b of (5.2) (resp., (5.3)) satisfy

$$\begin{aligned} - \int_{\partial\kappa} \{\sigma \nabla u_h^1\} \cdot n_\kappa \, ds + \gamma \int_{\partial\kappa} \omega \tilde{h}_{\mathcal{F}}^{-1} [\overline{u_h^1}] \cdot n_\kappa \, ds &= \int_\kappa f \, dx, \\ - \int_{\partial\kappa} \{\sigma \nabla u_h^b\} \cdot n_\kappa \, ds &= \int_\kappa f \, dx. \end{aligned}$$

LEMMA 5.2 (consistency of methods). *If the exact solution u of problem (2.1) satisfies $u \in H^2(\Omega)$, then the formulations defined by (5.2) and (5.3) are consistent (and adjoint consistent if the bilinear form is symmetric). Moreover, the following Galerkin orthogonalities hold:*

$$\begin{aligned} a_s(u - u_h^1, v_h) + \gamma j(u - u_h^1, v_h) &= 0 & \forall v_h \in V_{h,0}^1, \\ a_s(u - u_h^b, v_h) &= 0 & \forall v_h \in V_h^b, \end{aligned}$$

where $u_h^1 \in V_{h,0}^1$ and $u_h^b \in V_h^b$ denote the discrete solutions of (5.2) (resp., (5.3)).

Proof. Since $u \in H^2(\Omega)$ and the form $a_s(\cdot, \cdot)$ coincides with the SIPG formulation for $s = 1$ (resp., NIPG for $s = -1$) it is consistent. Moreover, for $s = 1$ the method is adjoint consistent. Observe that for $u \in H^1(\Omega)$ there holds

$$j(u, v_h) = (\omega \tilde{h}_{\mathcal{F}}^{-1} [\overline{u}], [\overline{v_h}])_{\mathcal{F}_i} = 0. \quad \square$$

6. Analysis of the RIP-method. For the analysis it is useful to introduce the following norms:

$$\begin{aligned} \|v\|^2 &= \|\sigma^{\frac{1}{2}} \nabla v\|_{\mathcal{K}}^2 + \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} [v]\|_{\mathcal{F}}^2, \\ \|w\|_c^2 &= \|\sigma^{\frac{1}{2}} \nabla w\|_{\mathcal{K}}^2 + \|\tilde{h}_{\mathcal{F}}^{\frac{1}{2}} \{\sigma \nabla w\}\|_{\mathcal{F}_i}^2 \end{aligned}$$

for all $v \in H^1(\mathcal{K})$, $w \in H^2(\mathcal{K})$. We have the following standard approximability results that we state without proof.

LEMMA 6.1 (approximability in V^C). *Let $u \in H^\beta(\Omega)$, with $\beta \in \{1, 2\}$, and $i_c u \in V^C$ denote the Crouzeix–Raviart-interpolant of u onto V^C defined by (3.1), then there holds*

$$(6.1) \quad \|u - i_c u\|_{\mathcal{K}} \leq c h^\beta \|u\|_{\beta, \mathcal{K}}.$$

If $u \in H^2(\Omega)$, then

$$(6.2) \quad \| \|u - i_c u\| \| \leq c h \|u\|_{2, \mathcal{K}},$$

$$(6.3) \quad \| \|u - i_c u\|_c \| \leq c h \|u\|_{2, \mathcal{K}}.$$

6.1. Stability. In this section we will use the orthogonality properties of V^C and V^D to obtain coercivity results for the unstabilized method also. These results ensure the existence of the discrete solution.

LEMMA 6.2 (orthogonality relations). *The bilinear forms $a_s(\cdot, \cdot)$ and $j(\cdot, \cdot)$ satisfy the following orthogonality relations:*

$$\begin{aligned} a_s(v_c, v_d) &= 0 & \forall v_c \in V^C, \forall v_d \in V^D, \\ a_s(v_d, v_c) &= (1-s)(\sigma \nabla v_d, \nabla v_c)_\mathcal{K} & \forall v_c \in V^C, \forall v_d \in V^D, \\ j(v_c, v_d) &= j(v_d, v_c) = 0 & \forall v_c \in V^C, \forall v_d \in V^D. \end{aligned}$$

Remark 6.3. The spaces V^C and V^D are orthogonal with respect to the symmetric bilinear form $a_1(\cdot, \cdot)$.

Proof. Let $v_c \in V^C$ and $v_d \in V^D$. Since $\int_F [v_c] ds = 0$ for all interior faces $F \in \mathcal{F}_i$, it follows directly that

$$j(v_c, v_d) = j(v_d, v_c) = 0,$$

and that

$$a_s(v_c, v_d) = (\sigma \nabla v_c, \nabla v_d)_\mathcal{K} - (\{\sigma \nabla v_c\}, [v_d])_{\mathcal{F}_i}.$$

An integration by parts yields

$$a_s(v_c, v_d) = -(\nabla \cdot \sigma \nabla v_c, v_d)_\mathcal{K} + ([\sigma \nabla v_c], \{v_d\})_{\mathcal{F}} = 0,$$

since $\int_F \{v_d\} ds = 0$ for all faces $F \in \mathcal{F}$ and $\nabla \cdot \sigma \nabla v_c|_\kappa = 0$ for all $\kappa \in \mathcal{K}$. Analogously we prove that

$$a_s(v_d, v_c) = (\sigma \nabla v_d, \nabla v_c)_\mathcal{K} - s(\{\sigma \nabla v_c\}, [v_d])_{\mathcal{F}_i} = (1-s)(\sigma \nabla v_d, \nabla v_c)_\mathcal{K}. \quad \square$$

LEMMA 6.4. *The bilinear forms $a_s(\cdot, \cdot)$ and $j(\cdot, \cdot)$ satisfy the following relations:*

$$\begin{aligned} a_s(u_c, v_c) &= (\sigma \nabla u_c, \nabla v_c)_\mathcal{K} & \forall u_c, v_c \in V^C, \\ a_s(u_d, v_d) &= -s(\sigma \nabla u_d, \nabla v_d)_\mathcal{K} & \forall u_d, v_d \in V^D, \\ j(u_c, v_c) &= 0 & \forall u_c, v_c \in V^C. \end{aligned}$$

Proof. The proof is similar to the one of Lemma 6.2 and uses the properties of the spaces V^C and V^D . \square

LEMMA 6.5 (splitting of the RIP-method). *The first method defined by (5.2) is equivalent to: Find $u_c \in V^C$, $u_d \in V^D$ such that*

$$(6.4) \quad (\sigma \nabla u_c, \nabla v_c)_\mathcal{K} + (1-s)(\sigma \nabla u_d, \nabla v_c)_\mathcal{K} = (f, v_c)_\mathcal{K} \quad \forall v_c \in V^C,$$

$$(6.5) \quad -s(\sigma \nabla u_d, \nabla v_d)_\mathcal{K} + \gamma(\omega \tilde{h}_{\mathcal{F}}^{-1} [\overline{u_d}], \overline{[v_d]})_{\mathcal{F}_i} = (f, v_d)_\mathcal{K} \quad \forall v_d \in V^D.$$

Remark 6.6. Observe that for $s = 1$, (6.4) is the Crouzeix–Raviart method for problem (2.1). As a consequence the stability and convergence analysis is known. It follows that the midpoint continuous part of u_h is independent of the parameter γ .

Moreover, note that (6.5) is independent of (6.4). Hence we can solve first for the discontinuous field u_d and then for the continuous field u_c also in the case $s = -1$.

Proof. Let $v_h \in V_{h,0}^1$. Since $V_{h,0}^1 = V^C \oplus V^D$ we can write $v_h = v_c + v_d$ with $v_c \in V^C$ and $v_d \in V^D$. Analogously we can write $u_h^1 = u_c + u_d$. Testing in (5.2) with v_c and v_d separately yields the problem: Find $u_c \in V^C$ and $u_d \in V^D$ such that

$$\begin{aligned} a_s(u_c + u_d, v_c) + \gamma j(u_c + u_d, v_c) &= (f, v_c)_K & \forall v_c \in V^C, \\ a_s(u_c + u_d, v_d) + \gamma j(u_c + u_d, v_d) &= (f, v_d)_K & \forall v_d \in V^D. \end{aligned}$$

Applying Lemma 6.2 leads directly to: Find $u_c \in V^C$ and $u_d \in V^D$ such that

$$(6.6) \quad a_s(u_c + u_d, v_c) = (f, v_c)_K \quad \forall v_c \in V^C,$$

$$(6.7) \quad a_s(u_d, v_d) + \gamma j(u_d, v_d) = (f, v_d)_K \quad \forall v_d \in V^D.$$

Note that the equivalences between the problems (6.6) and (6.4) (resp., (6.7) and (6.5)) follow directly from Lemma 6.4. \square

LEMMA 6.7 (coercivity of the RIP-method). *A unique solution to the discrete problem (6.5) exists for all $s\gamma \leq 0$ and $s\gamma > C_{stab}$, where $C_{stab} > 0$ is a certain constant independent of h . Moreover, the following coercivity bound holds:*

$$C(\gamma) \| \|u_d\| \|^2 \leq |a_s(u_d, u_d) + \gamma j(u_d, u_d)|,$$

where

$$C(\gamma) = \begin{cases} c \min(1, |\gamma|) & \text{if } s\gamma < 0, \\ c(\gamma - C_{stab}) & \text{if } s\gamma > C_{stab}. \end{cases}$$

On general meshes for $\gamma = 0$ there holds

$$c \|\sigma^{\frac{1}{2}} u_d\|_K^2 \leq |a_s(u_d, u_d)|.$$

On meshes described in Proposition 4.5 there holds, for $\gamma = 0$,

$$(6.8) \quad c \| \|u_d\| \|^2 \leq |a_s(u_d, u_d)|.$$

Proof. Let us prove first the regularity of (6.5) for $s\gamma < 0$. Observe, using Lemma 4.1, that

$$\begin{aligned} c \min(1, |\gamma|) \| \|u_d\| \|^2 &\leq \|\sigma^{\frac{1}{2}} \nabla u_d\|_K^2 + |\gamma| \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} \overline{[u_d]}\|_{\mathcal{F}_i}^2 = -s a_s(u_d, u_d) + |\gamma| j(u_d, u_d) \\ &= |a_s(u_d, u_d) + \gamma j(u_d, u_d)| \end{aligned}$$

since $\|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} \overline{[u_d]}\|_{\mathcal{F}_e}^2 = 0$.

For $s\gamma > 0$ observe that using the inverse and trace inequalities yields

$$\| \|u_d\| \|^2 = \|\sigma^{\frac{1}{2}} \nabla u_d\|_K^2 + \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} [u_d]\|_{\mathcal{F}}^2 \leq c \|\tilde{h}^{-1} \sigma^{\frac{1}{2}} u_d\|_K^2.$$

On the other hand, by norm equivalence on discrete spaces there exists a constant $c_\star > 0$, independent of the mesh size h , such that

$$\|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} \overline{[u_d]}\|_{\mathcal{F}_i}^2 \geq c_\star \|\tilde{h}^{-1} \sigma^{\frac{1}{2}} u_d\|_K^2,$$

since $\omega|_F \geq \sigma|_{\kappa_i}$, $i = 1, 2$, for $F = \partial\kappa_1 \cap \partial\kappa_2$. Thus, using the inverse inequality with constant c_{ie} yields

$$\begin{aligned} |a_s(u_d, u_d) + \gamma j(u_d, u_d)| &\geq -\|\sigma^{\frac{1}{2}} \nabla u_d\|_{\mathcal{K}}^2 + s\gamma \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} \overline{[u_d]}\|_{\mathcal{F}_i}^2 \\ &\geq (s\gamma c_\star - c_{ie}) \|\tilde{h}^{-1} \sigma^{\frac{1}{2}} u_d\|_{\mathcal{K}}^2. \end{aligned}$$

Observe that coercivity holds under the assumption that $s\gamma = |\gamma| > \frac{c_{ie}}{c_\star} =: C_{stab}$.

For $\gamma = 0$ on general meshes observe that

$$c \|\sigma^{\frac{1}{2}} u_d\|_{\mathcal{K}}^2 = \|\sigma^{\frac{1}{2}} \nabla u_d\|_{\mathcal{K}}^2 = -s a_s(u_d, u_d) = |a_s(u_d, u_d)|$$

using the Poincaré inequality noted in Proposition 4.4.

For $\gamma = 0$ on meshes described in Proposition 4.5, we have

$$\begin{aligned} \|u_d\|^2 &= \|\sigma^{\frac{1}{2}} \nabla u_d\|_{\mathcal{K}}^2 + \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} [u_d]\|_{\mathcal{F}}^2 \leq \|\sigma^{\frac{1}{2}} \nabla u_d\|_{\mathcal{K}}^2 + c \|\tilde{h}^{-1} \sigma^{\frac{1}{2}} u_d\|_{\mathcal{K}}^2 \\ &\leq c \|\sigma^{\frac{1}{2}} \nabla u_d\|_{\mathcal{K}}^2 = c |a_s(u_d, u_d)| \end{aligned}$$

using the trace inequality and the strong Poincaré inequality noted in Proposition 4.5. \square

6.2. Convergence. We will now address the question of optimal convergence for different values of the stabilization parameter. In the case where the stabilization parameter is set to zero, the lack of continuity of the bilinear form may perturb convergence. However, if the mesh has the macroelement structure of Proposition 4.5 optimal convergence is recovered.

THEOREM 6.8. *Let $u \in H^2(\Omega)$ be the solution of (2.1), and let u_h^1 be the solution of (5.2) with $s\gamma < 0$, $s\gamma > C_{stab}$, or $\gamma = 0$ on the meshes defined in Proposition 4.5, then there holds*

$$\|u - u_h^1\| \leq ch \|u\|_{2,\mathcal{K}}.$$

Remark 6.9. On general meshes, in the particular case $\gamma = 0$, this theorem is no longer valid. In this case an optimal convergence result can be shown under some restrictive regularity assumptions on f and the mesh; see Theorem 6.10.

Proof. First, note that for the bilinear form $a_s(\cdot, \cdot)$ the following continuity holds for all $w \in H_0^1(\Omega)$, $w_c \in V^C$, and $v_h \in V_h^1$ by the Cauchy–Schwarz inequality:

$$(6.9) \quad a_s(w - w_c, v_h) \leq \|w - w_c\|_c \|v_h\|.$$

1. Since $u_h^1 = u_c + u_d$ decomposes the error in two midpoint-continuous parts and one midpoint-discontinuous part,

$$(6.10) \quad \|u - u_h^1\| \leq \|u - i_c u\| + \|u_c - i_c u\| + \|u_d\|.$$

2. Observe that by Lemmas 6.4, 5.2, and 6.2 and (6.9),

$$\begin{aligned} \|u_c - i_c u\|^2 &= a_s(u_c - i_c u, u_c - i_c u) = a_s(u - i_c u - u_d, u_c - i_c u) \\ &\leq c (\|u - i_c u\|_c + \|u_d\|) \|u_c - i_c u\|, \end{aligned}$$

since $u_c - i_c u + u - u_h^1 = u - i_c u - u_d$ and $a_s(u - u_h^1, i_c u - u_c) = 0$. Thus

$$(6.11) \quad \|u - u_h^1\| \leq c (\|u - i_c u\| + \|u - i_c u\|_c + \|u_d\|).$$

3. Use Lemma 6.1 to bound the first two terms of the right-hand side of (6.11),

$$\| \|u - i_c u\| \| + \| \|u - i_c u\|_c \leq ch \|u\|_{2,\mathcal{K}}.$$

4. For the third term of (6.11) use the coercivity noted in Lemma 6.7,

$$\| \|u_d\| \|^2 \leq \frac{c}{C(\gamma)} |a_s(u_d, u_d) + \gamma j(u_d, u_d)|.$$

In the particular case $\gamma = 0$, the constant $C(0)$ denotes the constant of (6.8).

5. Use the consistency of the bilinear form noted in Lemma 5.2,

$$\| \|u_d\| \|^2 \leq \frac{c}{C(\gamma)} |a_s(u - u_c, u_d)|,$$

since $u_d + u - u_h^1 = u - u_c$ and $a_s(u - u_h^1, u_d) - \gamma j(u_h^1, u_d) = 0$.

6. Conclude by applying the continuity (6.9) and the approximation result (6.3) that

$$\| \|u - u_h^1\| \| \leq ch \|u\|_{2,\mathcal{K}}. \quad \square$$

Under some restrictions we can show optimal convergence also in the particular case of $\gamma = 0$ for the symmetric version on meshes without the macroelement property.

THEOREM 6.10. *Let $u \in H^2(\Omega)$ be the solution of (2.1), and let u_h^1 be the solution of (5.2) with $s = 1$ and $\gamma = 0$. Assume that $f \in H^\beta(\Omega)$, with $\beta \in \{1, 2\}$, and that the mesh is asymptotically ζ -uniform with some $\zeta > 0$, then there holds*

$$\| \sigma^{\frac{1}{2}} \nabla(u - u_h^1) \|_{\mathcal{K}} \leq c(h \|u\|_{2,\mathcal{K}} + (h^\zeta + r(d) h^2) \|f\|_{1,\mathcal{K}} + h^\beta \|f\|_{\beta,\mathcal{K}}),$$

where $r(2) = 0$ and $r(3) = c$.

Proof. Using the triangle inequality we can split

$$\| \sigma^{\frac{1}{2}} \nabla(u - u_h^1) \|_{\mathcal{K}} \leq \| \sigma^{\frac{1}{2}} \nabla(u - u_c) \|_{\mathcal{K}} + \| \sigma^{\frac{1}{2}} \nabla u_d \|_{\mathcal{K}}.$$

Since we only consider the symmetric version ($s = 1$), u_c is the standard Crouzeix–Raviart solution, and the first term of the right-hand side of the previous equation can be bounded by

$$\| \sigma^{\frac{1}{2}} \nabla(u - u_c) \|_{\mathcal{K}} \leq ch \|u\|_{2,\mathcal{K}}.$$

From (6.5) we can write

$$(6.12) \quad \| \sigma^{\frac{1}{2}} \nabla u_d \|^2 = |(f, u_d)_{\mathcal{K}}| \leq |(f - i_c f, u_d)_{\mathcal{K}}| + |(i_c f, u_d)_{\mathcal{K}}|,$$

where i_c is the Crouzeix–Raviart interpolant introduced in Remark 3.2. The first term of the right-hand side of (6.12) can be bounded by

$$(6.13) \quad |(f - i_c f, u_d)_{\mathcal{K}}| \leq \|f - i_c f\|_{\mathcal{K}} \|u_d\|_{\mathcal{K}} \leq ch^\beta \|f\|_{\beta,\mathcal{K}} \| \sigma^{\frac{1}{2}} \nabla u_d \|_{\mathcal{K}}$$

by optimal approximation properties of the Crouzeix–Raviart interpolant and by the Poincaré inequality for V^D noted in Proposition 4.4.

For the second term of the right-hand side of (6.12), we use Lemma 3.3, the Poincaré inequality for V^D , Proposition 4.4, and the stability of the Crouzeix–Raviart interpolant $\|i_c f\|_{\mathcal{K}} \leq \|f\|_{1,\mathcal{K}}$ (resp., $\|\nabla i_c f\|_{\mathcal{K}} \leq \|f\|_{1,\mathcal{K}}$):

$$(6.14) \quad \begin{aligned} |(i_c f, u_d)_{\mathcal{K}}| &\leq c(h^\zeta \|i_c f\|_{\mathcal{K}} \|u_d\|_{\mathcal{K}} + r(d) h^2 \|\nabla i_c f\|_{\mathcal{K}} \| \sigma^{\frac{1}{2}} \nabla u_d \|_{\mathcal{K}}) \\ &\leq c(h^\zeta + r(d) h^2) \|f\|_{1,\mathcal{K}} \| \sigma^{\frac{1}{2}} \nabla u_d \|_{\mathcal{K}}. \end{aligned}$$

Combining (6.13) and (6.14) completes the proof. \square

Remark 6.11. If $\beta = 2$ and $\zeta = 2$, then optimal convergence can be shown in the triple norm $\|\cdot\|$ since in this case

$$\|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} [u_d]\|_{\mathcal{F}}^2 \leq c \|\tilde{h}^{-1} \nabla u_d\|_{\mathcal{K}} \leq ch \|f\|_{2,\mathcal{K}}.$$

Remark 6.12. Observe that on uniform meshes the convergence is only limited by the regularity of f .

We will now show that we have optimal L^2 -convergence for the symmetric version thanks to the adjoint consistency. For the nonsymmetric version, the L^2 -convergence rate depends on the regularity of the mesh and the right-hand side, as pointed out in [10] in the one-dimensional case.

THEOREM 6.13. *Let $u \in H^2(\Omega)$ with $\|u\|_{2,\mathcal{K}} \leq c \|f\|_{\mathcal{K}}$ be the solution of (2.1), and let u_h^1 be the solution of (5.2) with $s\gamma < 0$, $s\gamma > C_{stab}$, or $\gamma = 0$ on meshes as described in Proposition 4.5, then the following hold.*

(a) *If $s = 1$, then*

$$\|u - u_h^1\|_{\mathcal{K}} \leq ch^2 \|u\|_{2,\mathcal{K}}.$$

(b) *If $s = -1$, assuming that $f \in H^\beta(\Omega)$, with $\beta \in \{1, 2\}$, and that the mesh is asymptotically ζ -uniform with some $\zeta > 0$, then*

$$\|u - u_h^1\|_{\mathcal{K}} \leq c(h^2 \|u\|_{2,\mathcal{K}} + (h^\zeta + r(d)h^2) \|f\|_{1,\mathcal{K}} + h^\beta \|f\|_{\beta,\mathcal{K}}),$$

where $r(2) = 0$ and $r(3) = c$.

Proof. Let $e = u - u_h$ and consider the dual problem: Find $\phi \in H_0^1(\Omega)$ such that

$$(\sigma \nabla \phi, \nabla z)_{\mathcal{K}} = (e, z)_{\mathcal{K}} \quad \forall z \in H_0^1(\Omega).$$

Under the regularity assumptions on u we have $\|\phi\|_{1,\mathcal{K}} \leq c \|e\|_{\mathcal{K}}$ and $\|\phi\|_{2,\mathcal{K}} \leq c \|e\|_{\mathcal{K}}$. It follows that

$$-(\nabla \cdot \sigma \nabla \phi, z)_{\mathcal{K}} = (e, z)_{\mathcal{K}} \quad \forall z \in L^2(\Omega).$$

Then we have by the dual consistency of Lemma 5.2,

$$\begin{aligned} \|e\|_{\mathcal{K}}^2 &= a_1(e, \phi) = a_s(e, \phi) + (s - 1)(\{\sigma \nabla \phi\}, [e])_{\mathcal{F}} \\ &= a_s(e, \phi - i_c \phi) - (s - 1)(\{\sigma \nabla \phi\}, \overline{[u_d]})_{\mathcal{F}_i} - (s - 1)(\{\sigma \nabla(\phi - i_c \phi)\}, [e] - \overline{[e]})_{\mathcal{F}}. \end{aligned}$$

First, observe that using the Cauchy–Schwarz inequality, the trace inequality for nondiscrete functions, and the approximation properties of the DG-solution, Theorem 6.8, (resp., the interpolation property of the Crouzeix–Raviart interpolant from (6.2)), implies that

$$\begin{aligned} a_s(e, \phi - i_c \phi) &\leq c \left(\|e\|^2 + \|\tilde{h}_{\mathcal{F}}^{\frac{1}{2}} \{\sigma \nabla e\}\|_{\mathcal{F}}^2 \right)^{\frac{1}{2}} \left(\|\phi - i_c \phi\|^2 + \|\tilde{h}_{\mathcal{F}}^{\frac{1}{2}} \{\sigma \nabla(\phi - i_c \phi)\}\|_{\mathcal{F}}^2 \right)^{\frac{1}{2}} \\ &\leq c \left(\|e\|^2 + h^2 \|u\|_{2,\mathcal{K}}^2 \right)^{\frac{1}{2}} \left(\|\phi - i_c \phi\|^2 + h^2 \|\phi\|_{2,\mathcal{K}}^2 \right)^{\frac{1}{2}} \\ &\leq ch^2 \|u\|_{2,\mathcal{K}} \|\phi\|_{2,\mathcal{K}} \leq ch^2 \|u\|_{2,\mathcal{K}} \|e\|_{\mathcal{K}}. \end{aligned}$$

Second, observe that

$$\begin{aligned} (\{\sigma \nabla \phi\}, \overline{[u_d]})_{\mathcal{F}_i} &\leq c \|\tilde{h}_{\mathcal{F}}^{\frac{1}{2}} \{\sigma \nabla \phi\}\|_{\mathcal{F}} \|\tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} \omega^{\frac{1}{2}} [u_d]\|_{\mathcal{F}} \leq c \|\phi\|_{1,\mathcal{K}}^{\frac{1}{2}} \|\phi\|_{2,\mathcal{K}}^{\frac{1}{2}} \|u_d\| \\ &\leq c \|e\|_{\mathcal{K}} \|u_d\| \end{aligned}$$

using a multiplicative trace inequality, see [4], and third, applying the approximability of the average jump and the trace inequality for nondiscrete functions yields

$$\begin{aligned} (\{\sigma \nabla \phi - \sigma \nabla i_c \phi\}, [e] - \overline{[e]})_{\mathcal{F}} &\leq c \|\tilde{h}_{\mathcal{F}}^{\frac{1}{2}} \{\sigma \nabla (\phi - i_c \phi)\}\|_{\mathcal{F}} \|\tilde{h}_{\mathcal{F}}^{\frac{1}{2}} [\nabla e]_t\|_{\mathcal{F}} \\ &\leq c (\|\nabla (\phi - i_c \phi)\|_{\mathcal{K}} + h \|\phi\|_{2,\mathcal{K}}) (\|e\| + h \|u\|_{2,\mathcal{K}}) \\ &\leq c h^2 \|\phi\|_{2,\mathcal{K}} \|u\|_{2,\mathcal{K}} \leq c h^2 \|e\|_{\mathcal{K}} \|u\|_{2,\mathcal{K}}, \end{aligned}$$

where $[\nabla e]_t$ denotes the tangential jump defined by $[\nabla e]_t|_F = \nabla e|_{\kappa_1} \times n_1 + \nabla e|_{\kappa_2} \times n_2$ for $F = \partial \kappa_1 \cap \partial \kappa_2 \in \mathcal{F}$. We conclude that

$$\|e\|_{\mathcal{K}} \leq c h^2 \|u\|_{2,\mathcal{K}} + c |s - 1| (\|u_d\| + h^2 \|u\|_{2,\mathcal{K}}).$$

For the symmetric case the result follows immediately since $|s - 1| = 0$.

Consider now the nonsymmetric case for which $s = -1$. By the coercivity noted in Lemma 6.7, it follows that

$$C(\gamma) \| \|u_d\| \|^2 \leq |a_{-1}(u_d, u_d) + \gamma j(u_d, u_d)| = |(f, u_d)_{\mathcal{K}}|.$$

Using Lemma 3.3, (6.1), and the stability of the Crouzeix–Raviart interpolant, we may conclude that

$$\begin{aligned} C(\gamma) \| \|u_d\| \|^2 &\leq |(f - i_c f, u_d)_{\mathcal{K}}| + |(i_c f, u_d)_{\mathcal{K}}| \\ &\leq (\|f - i_c f\|_{\mathcal{K}} + c h^\zeta \|i_c f\|_{\mathcal{K}}) \|u_d\|_{\mathcal{K}} + r(d) h^2 \|\nabla i_c f\|_{\mathcal{K}} \|\nabla u_d\|_{\mathcal{K}} \\ &\leq (c h^\beta \|f\|_{\beta,\mathcal{K}} + c (h^\zeta + r(d) h^2) \|f\|_{1,\mathcal{K}}) \| \|u_d\| \|. \quad \square \end{aligned}$$

Remark 6.14. For $\gamma = 0$, in the particular case of Remark 6.11 optimal convergence in the L^2 -norm can be shown also on regular meshes without the macroelement property. The details are left to the reader.

6.3. Numerical tests. Observe that the only difference between the standard SIPG/NIPG-method and the RIP-method is that in the latter case the stabilization term is composed by the facewise L^2 -projection of order 0 of the jumps. From an implementational viewpoint this can be realized by reducing the order of the quadrature formula for the numerical integration on the faces; i.e., applying the midpoint integration rules for the computation of the stabilization term.

6.3.1. Test problems. Let us briefly present the test problems used for the numerical tests.

(i) *Problem with smooth solution.* We consider problem (2.1) with $\sigma = 1$ and $f(x, y) = 2(2 - x^2 - y^2)$ on the square $\Omega = (-1, 1)^2$. The analytic exact solution is given by $u(x, y) = (x^2 - 1)(y^2 - 1) \in C^\infty(\overline{\Omega})$. A sequence of unstructured meshes is considered.

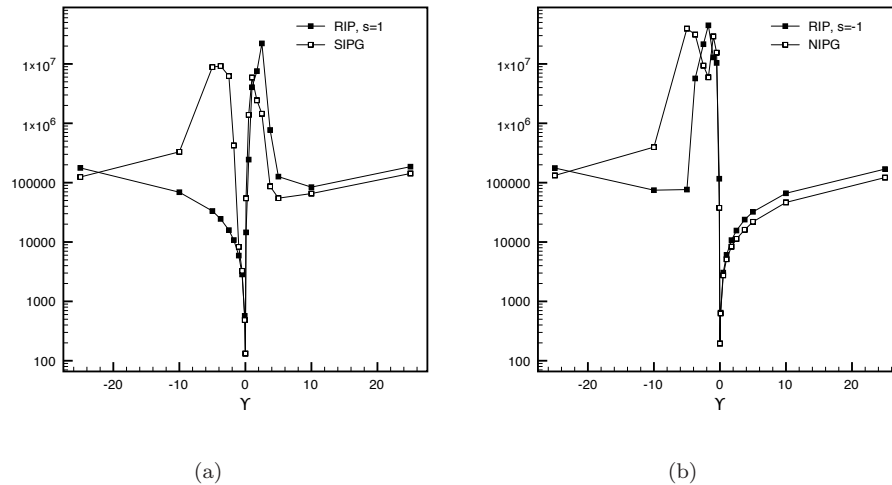


FIG. 2. Comparison of the condition number of the matrices corresponding to the symmetric version (a) and nonsymmetric version (b) of the RIP-method and the SIPG- (a), respectively, NIPG-method (b) for the test problem (i) with smooth solution.

(ii) *Problem with irregular solution.* Now choose the following L -shaped domain: $\Omega = ([-1, 1] \times [-1, 0] \cup [0, 1]^2)^\circ$. We consider problem (2.1) with $\sigma = 1$ and $f \equiv 0$ and nonhomogeneous boundary conditions such that the solution is

$$u(x, y) = (x^2 + y^2)^{\frac{1}{3}} \sin\left(\frac{2}{3} \arctan_*\left(\frac{x}{y}\right)\right),$$

where \arctan_* is chosen in such a manner that it is a continuous function at points with $y = 0$. One can prove that $u \notin H^2(\Omega)$. Therefore Theorems 6.8, 6.13, 7.4, and 7.5 are no longer valid. A sequence of unstructured meshes is considered.

(iii) *Problem with checkerboard mode.* We consider problem (2.1) with $\sigma = 1$ and $f(x, y) = -1 + 2\chi_{x>y}$, where χ denotes the characteristic function on the square $\Omega = (-1, 1)^2$. A sequence of structured meshes is considered.

6.3.2. Robustness with respect to the stabilization parameter. Let us consider the test problem (i) with smooth solution. We compare the robustness of the symmetric RIP-method with the SIPG-method; respectively, the nonsymmetric RIP-method with the NIPG-method.

In Figure 2 we give comparisons of the condition number of the corresponding matrices. We define the condition number of a square matrix (not necessarily symmetric positive definite) as the ratio of the largest singular value of the matrix to the smallest one. Since the continuous and the discontinuous part of the approximation decouples for our formulation, we may also consider negative values of the penalization parameter. One readily verifies from the graphics that the approximate solution degenerates for values of the stability parameter that do not satisfy the hypothesis $s\gamma \leq 0$ or $s\gamma > C_{stab}$ given in Lemma 6.7. Observe in particular for the symmetric methods that the RIP-method is stable for negative stabilization parameters whereas the SIPG-method is not, which is due to the fact that the stabilization in the former case affects only the discontinuous subspace.

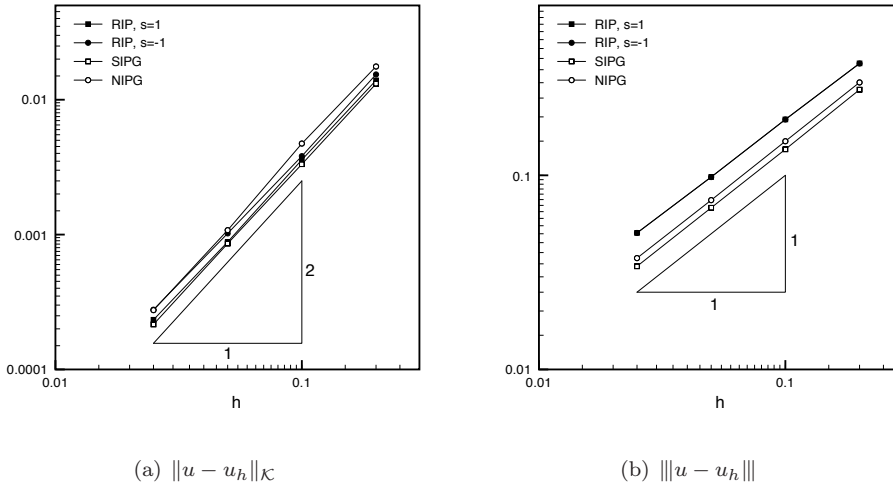


FIG. 3. L^2 -error (a) and energy-error (b) for h -refinement for the test problem (i) with smooth solution using stabilization parameters $\gamma = 0$ for the RIP-method, $\gamma = 10$ for the SIPG-method, and $\gamma = 1$ for the NIPG-method.

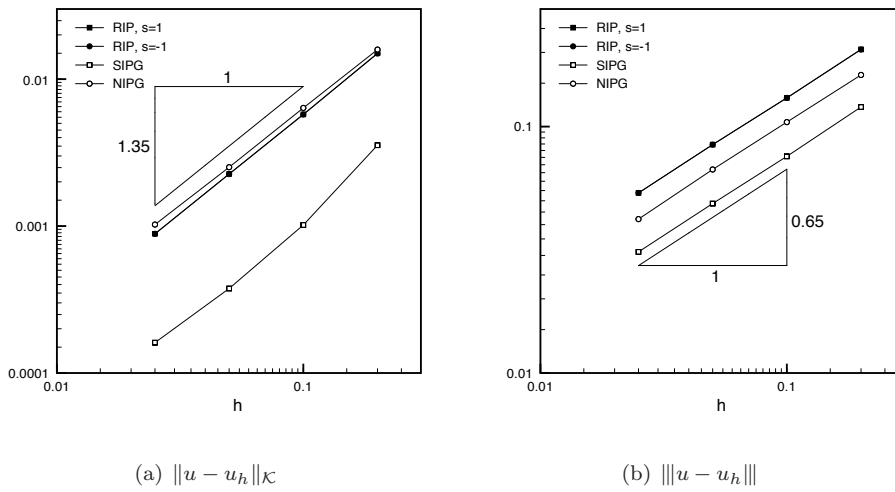


FIG. 4. L^2 -error (a) and energy-error (b) for h -refinement for the test problem (ii) with irregular solution using stabilization parameters $\gamma = 0$ for the RIP-method, $\gamma = 10$ for the SIPG-method, and $\gamma = 1$ for the NIPG-method.

6.3.3. Convergence. The convergence rates of the RIP-method with stabilization parameter $\gamma = 0$ are compared to those of the standard SIPG- and NIPG-method once for the problem (i) with regular solution and once for the problem (ii) with irregular solution.

Note that in several plots the curves of two different methods may have exactly the same error and the two curves are indistinguishable.

Figure 3 shows the optimal convergence rates of the error in the approximations of the solution of the smooth problem measured in the L^2 - and energy-norm. The

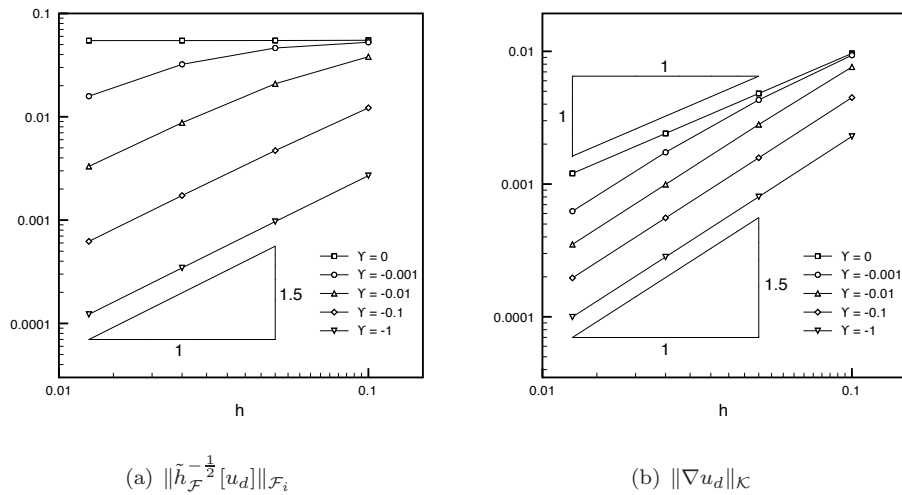


FIG. 5. Different norms of the discontinuous field u_d of the RIP-method for h -refinement and for different stabilization parameters γ for the test problem (ii).

symmetric versions have slightly better convergence rates in the L^2 -norm, which can be justified by Theorem 6.13.

Figure 4 shows the accuracy of the methods when solving the problem with an irregular solution. The SIPG-method has a smaller L^2 -error than the other methods; however, the convergence rates are the same.

The test problem (iii) is chosen so as to give rise to a checkerboard mode in the discontinuous field. The convergence of the jump term for different values of the stabilization parameter is given in Figure 5(a), and the convergence of the broken H^1 seminorm of u_d is given in Figure 5(b). Clearly the broken H^1 seminorm of u_d converges for the case without stabilization even though u_d does not converge in the norm $\|\cdot\|$ including the jumps. This lack of convergence of the interelement jumps is caused by the checkerboard mode in the field u_d . In Figure 6 we give plots of the u_d field for various values of the penalization parameter γ . This clearly illustrates how the penalization localizes the perturbation caused by the discontinuous data, and hence enhances convergence for $\gamma \neq 0$.

7. Analysis of the BSDG-method. In the previous section we saw that for the unstabilized symmetric DG-method the appearance of a checkerboard mode for rough data destroyed convergence of the solution jumps. Optimal convergence is recovered if the mesh has a certain macroelement structure. In the framework of the BSDG-method this structure is replaced by a bubble enrichment of the space. The motivation for the DG-method using the enriched space is to obtain local mass conservation independent of the stabilization parameter for a low order DG-method while keeping optimal convergence properties in the general case.

7.1. Projection. In order to prove stability of the method we first need to define the following projection.

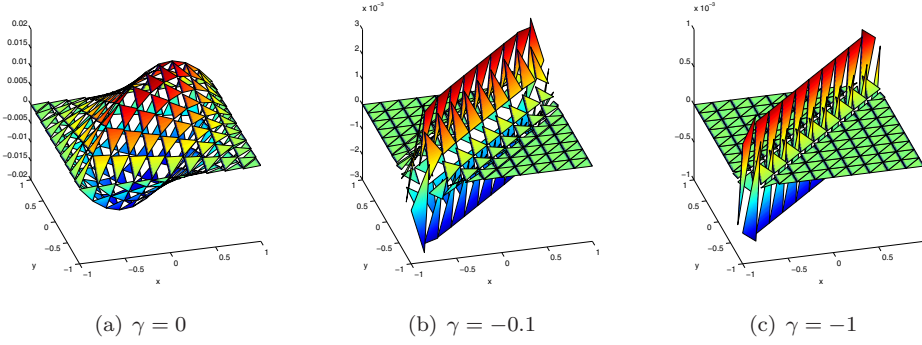


FIG. 6. The discontinuous field u_d of the RIP-method computed using different values of the stabilization parameter for the test problem (ii) generating a checkerboard mode.

LEMMA 7.1 (interpolant). Let $u_h \in V_h^b$ be a fixed function. Then there exists a unique $w_h \in V_h^b$ such that

$$(7.1) \quad \int_{\kappa} w_h \, dx = 0 \quad \forall \kappa \in \mathcal{K},$$

$$(7.2) \quad \{\sigma \nabla w_h\}|_F \cdot n_F = -s\omega h_F^{-1} \overline{[u_h]}|_F \cdot n_F \quad \forall F \in \mathcal{F},$$

$$(7.3) \quad \overline{[w_h]}|_F = 0 \quad \forall F \in \mathcal{F}_i.$$

In addition the following a priori estimate holds:

$$(7.4) \quad \|\sigma^{\frac{1}{2}} \nabla w_h\|_{\mathcal{K}} \leq c \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} \overline{[u_h]}\|_{\mathcal{F}},$$

where $\omega = \max(\sigma|_{\kappa_1}, \sigma|_{\kappa_2})$ for $F = \partial\kappa_1 \cap \partial\kappa_2$.

Proof. Let us first observe that the number of conditions on the projection is equal to the number of unknowns. The dimension of the finite element space V_h^b is $(d+2)N_{el}$, where N_{el} denotes the number of elements in the mesh. On the other hand condition (7.1) enforces N_{el} constraints whereas conditions (7.2) and (7.3) demand $N_f + N_{int}$ constraints where N_f and N_{int} denote the number of the number of faces, respectively, the number of interior faces of the mesh. Observing that $N_f + N_{int} = (d+1)N_{el}$ implies directly a square linear system to determine the projection, let us now establish the a priori estimate

$$(7.5) \quad \|\sigma^{\frac{1}{2}} \nabla w_h\|_{\mathcal{K}} \leq c \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} \overline{[u_h]}\|_{\mathcal{F}}.$$

Since w_h has zero mean over each element, it satisfies the following strong Poincaré inequality:

$$(7.6) \quad \|\sigma^{\frac{1}{2}} w_h\|_{\mathcal{K}} \leq c \|\tilde{h} \sigma^{\frac{1}{2}} \nabla w_h\|_{\mathcal{K}}.$$

Hence, using a trace inequality yields

$$(7.7) \quad \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} \overline{[w_h]}\|_{\mathcal{F}}^2 \leq c \|\sigma^{\frac{1}{2}} \tilde{h}^{-1} w_h\|_{\mathcal{K}}^2 \leq c \|\sigma^{\frac{1}{2}} \nabla w_h\|_{\mathcal{K}}^2.$$

Moreover, integrating by parts and using the properties of w_h , it follows that

$$\begin{aligned} \|\sigma^{\frac{1}{2}}\nabla w_h\|_{\mathcal{K}}^2 &= -\underbrace{(\nabla \cdot \sigma \nabla w_h, w_h)_{\mathcal{K}}}_{=0} + \underbrace{(\{\sigma \nabla w_h\}, [w_h])_{\mathcal{F}}}_{=0} + \underbrace{([\sigma \nabla w_h], \{w_h\})_{\mathcal{F}_i}}_{=0} \\ &= -s(\omega \tilde{h}_{\mathcal{F}}^{-1} \overline{[u_h]}, [w_h])_{\mathcal{F}} \leq \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} \overline{[u_h]}\|_{\mathcal{F}} \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} [w_h]\|_{\mathcal{F}} \end{aligned}$$

since $\{\sigma \nabla w_h\}|_{F \cdot n_F}$ and $[\sigma \nabla w_h]|_F$ are constant along each face $F \in \mathcal{F}$; see Lemma 3.4. Applying further (7.7) proves (7.4):

$$\|\sigma^{\frac{1}{2}}\nabla w_h\|_{\mathcal{K}} \leq c \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} \overline{[u_h]}\|_{\mathcal{F}}.$$

Since we consider a square linear system, existence and uniqueness of a solution of the linear system are equivalent. Uniqueness follows by the a priori estimate. \square

COROLLARY 7.2. *Let $u_h \in V_h^b$ be a fixed function. Then there exists a unique $y_h \in V_h^b$ such that*

$$\begin{aligned} \int_{\kappa} y_h \, dx &= \int_{\kappa} (f - \nabla \cdot \sigma \nabla u_h) \, dx && \forall \kappa \in \mathcal{K}, \\ \{\sigma \nabla y_h\}|_{F \cdot n_F} &= -s\omega h_F^{-1} \overline{[u_h]}|_{F \cdot n_F} && \forall F \in \mathcal{F}, \\ \overline{\{y_h\}}|_F &= [\sigma \nabla u_h]|_F && \forall F \in \mathcal{F}_i. \end{aligned}$$

Proof. Since the matrix associated to the above defined projection w_h has zero kernel, y_h exists and is unique. \square

7.2. Stability. Although we do not explicitly penalize the solution jumps, control of the solution jumps in the energy-norm is recovered by an inf-sup argument as shown in this section.

THEOREM 7.3 (discrete inf-sup condition). *There exists a constant $c > 0$ independent of h such that for all $u_h \in V_h^b$ there holds*

$$c \| \|u_h\| \| \leq \sup_{v_h \in V_h^b} \frac{a_s(u_h, v_h)}{\| \|v_h\| \|}$$

for $s \in \{-1, 1\}$.

Proof. Let us prove this theorem in four steps.

Step 1. First, we take $v_h = u_h$ in a standard fashion

$$(7.8) \quad a_s(u_h, u_h) = \|\sigma^{\frac{1}{2}}\nabla u_h\|_{\mathcal{K}}^2 - (1+s)(\{\sigma \nabla u_h\}, \overline{[u_h]})_{\mathcal{F}}$$

since $\{\sigma \nabla u_h\}|_{F \cdot n_F}$ is constant along each face $F \in \mathcal{F}$. Applying a trace inequality and the inverse inequality for the second term of the right-hand side of (7.8) followed by an arithmetic-geometric inequality, there exists a constant $c_u > 0$ independent on the mesh size h such that

$$(7.9) \quad a_s(u_h, u_h) \geq \frac{1}{2} \|\sigma^{\frac{1}{2}}\nabla u_h\|_{\mathcal{K}}^2 - c_u(1+s)^2 \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} \overline{[u_h]}\|_{\mathcal{F}}^2.$$

Step 2. Second, by Lemma 7.1 there exists $w_h \in V_h^b$ such that

1. $\int_{\kappa} w_h \, dx = 0 \forall \kappa \in \mathcal{K}$,
2. $\{\sigma \nabla w_h\} \cdot n_F = -s\omega h_F^{-1} \overline{[u_h]} \cdot n_F$ on each face $F \in \mathcal{F}$,

3. $\int_F \{w_h\} = 0$ on each face $F \in \mathcal{F}_i$.

An immediate consequence is that

$$\begin{aligned} a_s(u_h, w_h) &= -(\nabla \cdot \sigma \nabla u_h, w_h)_{\mathcal{K}} + ([\sigma \nabla u_h], \{w_h\})_{\mathcal{F}_i} - s(\{\sigma \nabla w_h\}, [u_h])_{\mathcal{F}} \\ (7.10) \quad &= \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} \overline{[u_h]}\|_{\mathcal{F}}^2. \end{aligned}$$

Step 3. Combining the results (7.9) and (7.10), we may take

$$v_h = u_h + \left(\frac{1}{2} + c_u(1 + s)^2\right)w_h$$

to obtain

$$a_s(u_h, v_h) \leq \frac{1}{2} (\|\sigma^{\frac{1}{2}} \nabla u_h\|_{\mathcal{K}}^2 + \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} \overline{[u_h]}\|_{\mathcal{F}}^2) \leq c \|u_h\|^2$$

by Lemma 4.1.

Step 4. To conclude, it remains to show that there exists $c > 0$ independent of h such that

$$\|v_h\| \leq c \|u_h\|.$$

This follows by straightforward estimation,

$$\|v_h\| = \|u_h + cw_h\| \leq \|u_h\| + c \|w_h\|.$$

Consider the second term of the right-hand side,

$$\|w_h\|^2 = \|\sigma^{\frac{1}{2}} \nabla w_h\|_{\mathcal{K}}^2 + \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} [w_h]\|_{\mathcal{F}}^2 = I_1 + I_2.$$

It follows by (7.4) that

$$I_1 \leq c \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} \overline{[u_h]}\|_{\mathcal{F}}^2 \leq c \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} [u_h]\|_{\mathcal{F}}^2,$$

and by the trace inequality, the strong Poincaré inequality (7.6), and by (7.4) that

$$I_2 \leq c \|\sigma^{\frac{1}{2}} \tilde{h}^{-1} w_h\|_{\mathcal{K}}^2 \leq c \|\sigma^{\frac{1}{2}} \nabla w_h\|_{\mathcal{K}}^2 \leq c \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} \overline{[u_h]}\|_{\mathcal{F}}^2 \leq c \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} [u_h]\|_{\mathcal{F}}^2. \quad \square$$

7.3. Convergence. Using the previously derived inf-sup condition optimal convergence is proved in a standard fashion.

THEOREM 7.4. *Let $u \in H^2(\Omega)$ be the solution of (2.1) and u_h^b the solution of (5.3). Then there holds*

$$\|u - u_h^b\| \leq ch \|u\|_{2, \mathcal{K}}.$$

Proof. First, note that for the bilinear form $a_s(\cdot, \cdot)$ the following continuity holds for all $v \in H^1(\Omega)$, $v_c \in V^C$, and $v_h \in V_h^b$ by Cauchy–Schwarz and trace inequalities

$$(7.11) \quad a_s(v - v_c, v_h) \leq \|v - v_c\|_c \|v_h\|.$$

1. Decompose the error in a weakly continuous and a discrete part

$$\|u - u_h^b\| \leq \|u - i_c u\| + \|i_c u - u_h^b\|.$$

Recall that i_c denotes the Crouzeix–Raviart interpolant onto V^C and observe that the convergence of the continuous part follows by Lemma 6.1.

2. Use the inf-sup condition on the discrete part and the consistency of the bilinear form (Lemma 5.2)

$$c \|i_c u - u_h^b\| \leq \sup_{v_h \in V_h^b} \frac{a_s(i_c u - u_h^b, v_h)}{\|v_h\|} \leq \sup_{v_h \in V_h^b} \frac{a_s(i_c u - u, v_h)}{\|v_h\|}.$$

3. Conclude by applying the continuity (7.11) and the approximation result of Lemma 6.1 that

$$\|u - u_h^1\| \leq ch \|u\|_{2, \mathcal{K}}.$$

Finally, using Lemma 4.1 and Corollary 4.2 proves the result. \square

The following optimal L^2 -convergence for the symmetric version may readily be proved for the symmetric version of the BSDG-method thanks to the adjoint consistency.

THEOREM 7.5. *Let $u \in H^2(\Omega)$ with $\|u\|_{2, \mathcal{K}} \leq c \|f\|_{\mathcal{K}}$ be the solution of (2.1) and u_h^b the solution of (5.3) with $s = 1$, then there holds*

$$\|u - u_h^b\|_{\mathcal{K}} \leq c h^2 \|u\|_{2, \mathcal{K}}.$$

Additionally, this method has some interesting properties as pointed out in the following remark.

Remark 7.6. Let $u_h^b \in V_h^b$ be the solution of (5.3). If the right-hand side f is elementwise constant, then there holds

$$\|\nabla \cdot \sigma \nabla u_h^b - f\|_{\mathcal{K}}^2 + \|[\sigma \nabla u_h^b]\|_{\mathcal{F}_i}^2 + \|\omega^{\frac{1}{2}} \tilde{h}_{\mathcal{F}}^{-\frac{1}{2}} \overline{[u_h^b]}\|_{\mathcal{F}}^2 = 0.$$

Indeed, choosing the function y_h defined in Corollary 7.2 in (5.3) and applying an integration by parts leads to the result. As a consequence, since the flux $\sigma \nabla u_h^b$ is continuous across faces, the solution u_h^b satisfies the following local mass conservation property:

$$-\int_{\partial \kappa} \sigma \nabla u_h^b \cdot n_{\kappa} ds = \int_{\kappa} f dx.$$

7.4. Numerical tests. We consider the same three test problems as in section 6.3.1 and give the convergence rates for those test problems. Note that in the case of the BSDG-method the local mass conservation property is satisfied independently of the stabilization parameter. We get optimal convergence of the error in the L^2 - and energy-norms, and similar convergence curves as for the SIPG- and NIPG-methods for the case of smooth exact solution (Figure 7). When the solution presents a singularity (Figure 8) we once again observe a larger constant for the BSDG-method than for the SIPG-method, in particular in the L^2 -norm. For the test problem (ii) we compare the error of the jumps of the solution of the RIP-method using various stabilization parameters and the same error of BSDG-method (Figure 9). As predicted by Theorem 7.4 the solution of BSDG-method shows optimal convergence also in this case. Hence the checkerboard mode is not present in the solution.

8. Conclusion. In this paper we discussed low order discontinuous Galerkin methods for second order scalar elliptic problems in two and three space dimensions. The main results are given in the following points.

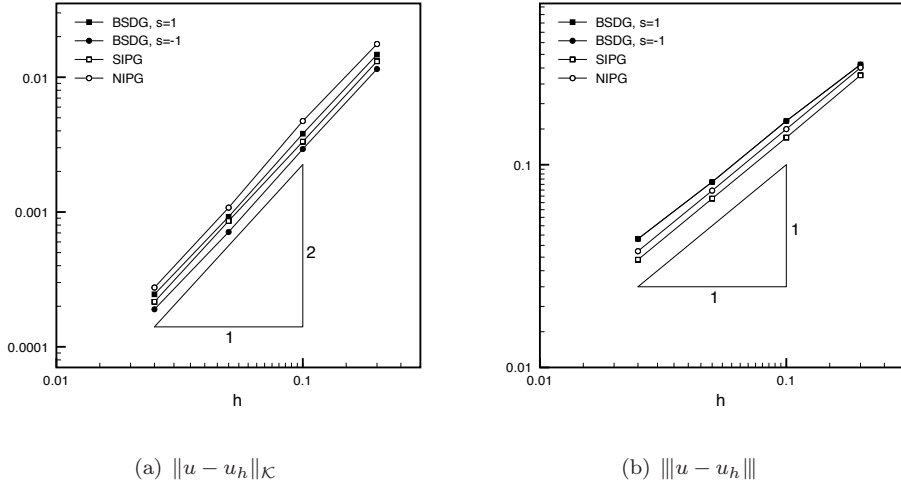


FIG. 7. L^2 -error (a) and energy-error (b) for h -refinement for the test problem (ii) with smooth solution. For the SIPG- and NIPG-methods a stabilization parameter of $\gamma = 10$, respectively, $\gamma = 1$ is chosen.

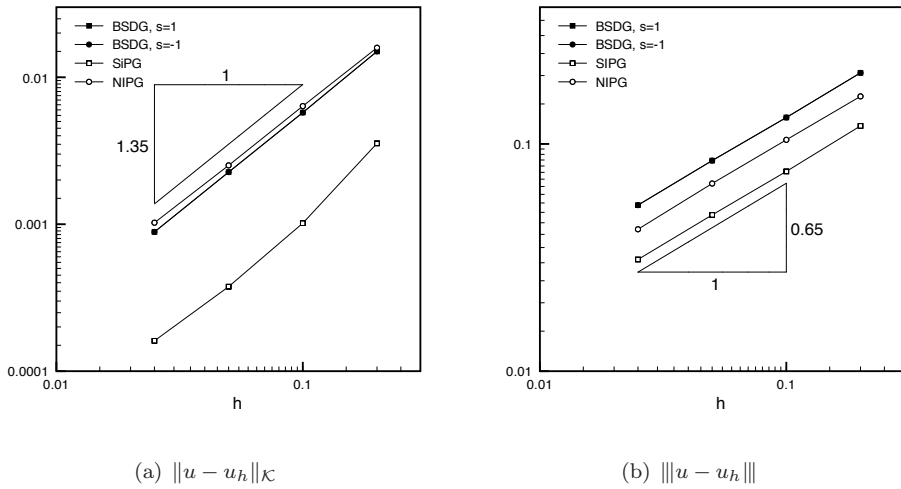


FIG. 8. L^2 -error (a) and energy-error (b) for h -refinement for the test problem (i) with irregular solution. For the SIPG- and NIPG-methods a stabilization parameter of $\gamma = 10$, respectively, $\gamma = 1$ is chosen.

(i) Midpoint imposition of Dirichlet boundary conditions is sufficient to assure existence of a discrete solution for the symmetric DG-formulation using piecewise affine approximation (no interior stabilization is needed).

(ii) The symmetric version of RIP-method without stabilization has optimal convergence in the energy-norm and in the L^2 -norm provided the meshes and data are sufficiently regular or satisfy the macroelement property of Proposition 4.5.

(iii) For irregular data and general meshes a checkerboard mode destroys convergence for the unstabilized DG-method when using piecewise affine approximations.

(iv) Enriching the space with nonconforming quadratic bubbles leads to a DG-method where the symmetric and nonsymmetric versions are stable without stabilization and optimally convergent in the energy-norm. Moreover, the methods are locally

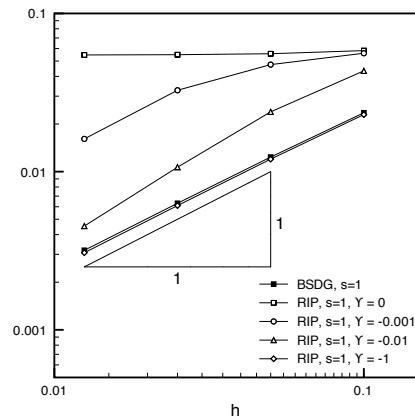
(a) $\|\tilde{h}^{-\frac{1}{2}}[u_h]\|_{\mathcal{F}}$

FIG. 9. Error of the jumps $\|\tilde{h}^{-\frac{1}{2}}[u_h]\|_{\mathcal{F}}$ of the BSDG-method compared to the RIP-method for h -refinement and the test problem (ii).

mass conservative independently of the penalty parameter.

(v) The symmetric DG-method on the enriched space has additionally optimal convergence in the L^2 -norm.

The aim of this work was to construct a symmetric DG-method that enjoys optimal convergence and local mass conservation independently of the penalty parameter. This goal has been realized in the framework of low order approximation in the symmetric version of the BSDG-method (see (5.3)) or, in the symmetric version of the RIP-method with $\gamma = 0$ on meshes having the macroelement property of Lemma 4.5.

Acknowledgments. The authors thank the anonymous reviewers for their comments that helped improve this manuscript.

REFERENCES

- [1] D. N. ARNOLD, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.
- [2] G. A. BAKER, *Finite element methods for elliptic equations using nonconforming elements*, Math. Comp., 31 (1977), pp. 45–59.
- [3] S. C. BRENNER, *Poincaré–Friedrichs inequalities for piecewise H^1 functions*, SIAM J. Numer. Anal., 41 (2003), pp. 306–324.
- [4] S. C. BRENNER AND L. R. SCOTT, *The mathematical theory of finite element methods*, Texts in Applied Mathematics 15, Springer-Verlag, New York, 1994.
- [5] F. BREZZI, B. COCKBURN, L. D. MARINI, AND E. SÜLI, *Stabilization mechanisms in discontinuous Galerkin finite element methods*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 3293–3310.
- [6] F. BREZZI AND L. D. MARINI, *Bubble stabilization of discontinuous Galerkin methods*, in Advances in Numerical Mathematics, Proceedings of the International Conference on the Occasion of the 60th Birthday of Y. A. Kuznetsov, 2005, W. Fitzgibbon, R. Hoppe, J. Periaux, O. Pironneau, and Y. Vassilevski, eds., Institute of Numerical Mathematics of The Russian Academy of Sciences, Moscow, 2006, pp. 25–36.
- [7] E. BURMAN AND B. STAMM, *Minimal stabilization for discontinuous Galerkin finite element methods for hyperbolic problems*, J. Sci. Comput., 33 (2007), pp. 183–208.
- [8] E. BURMAN AND B. STAMM, *Symmetric and non-symmetric discontinuous Galerkin methods stabilized using bubble enrichment*, C. R. Math. Acad. Sci. Paris, 346 (2008), pp. 103–106.

- [9] J. DOUGLAS AND T. DUPONT, *Interior penalty procedures for elliptic and parabolic Galerkin methods*, in Computing Methods in Applied Sciences (Second Internat. Sympos., Versailles, 1975), Lecture Notes in Phys. 58, Springer, Berlin, 1976, pp. 207–216
- [10] M. G. LARSON AND A. J. NIKLASSON, *Analysis of a family of discontinuous Galerkin methods for elliptic problems: The one dimensional case*, Numer. Math., 99 (2004), pp. 113–130.
- [11] M. G. LARSON AND A. J. NIKLASSON, *Analysis of a nonsymmetric discontinuous Galerkin method for elliptic problems: Stability and energy error estimates*, SIAM J. Numer. Anal., 42 (2004), pp. 252–264.
- [12] J. T. ODEN, I. BABUŠKA, AND C. E. BAUMANN, *A discontinuous hp finite element method for diffusion problems*, J. Comput. Phys., 146 (1998), pp. 491–519.
- [13] B. RIVIÈRE, M. F. WHEELER, AND V. GIRAULT, *A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems*, SIAM J. Numer. Anal., 39 (2001), pp. 902–931.
- [14] R. TEMAM, *Navier-Stokes equations*, Theory and Numerical Analysis (reprint of the 1984 edition), AMS Chelsea Publishing, Providence, RI, 2001.
- [15] M. F. WHEELER, *An elliptic collocation-finite element method with interior penalties*, SIAM J. Numer. Anal., 15 (1978), pp. 152–161.

CONVERGENCE ANALYSIS OF AN ADAPTIVE INTERIOR PENALTY DISCONTINUOUS GALERKIN METHOD*

R. H. W. HOPPE[†], G. KANSCHAT[‡], AND T. WARBURTON[§]

Abstract. We study the convergence of an adaptive interior penalty discontinuous Galerkin (IPDG) method for a two-dimensional model second order elliptic boundary value problem. Based on a residual-type a posteriori error estimator, we prove that after each refinement step of the adaptive scheme we achieve a guaranteed reduction of the global discretization error in the mesh-dependent energy norm associated with the IPDG method. In contrast to recent work on adaptive IPDG methods [O. Karakashian and F. Pascal, *Convergence of Adaptive Discontinuous Galerkin Approximations of Second-order Elliptic Problems*, preprint, University of Tennessee, Knoxville, TN, 2007], the convergence analysis does not require multiple interior nodes for refined elements of the triangulation. In fact, it will be shown that bisection of the elements is sufficient. The main ingredients of the proof of the error reduction property are the reliability and a perturbed discrete local efficiency of the estimator, a bulk criterion that takes care of a proper selection of edges and elements for refinement, and a perturbed Galerkin orthogonality property with respect to the energy inner product. The results of numerical experiments are given to illustrate the performance of the adaptive method.

Key words. discontinuous Galerkin, adaptive methods, interior penalty, error estimates

AMS subject classifications. 65N30, 65N50

DOI. 10.1137/070704599

1. Introduction. During the past decade, discontinuous Galerkin (DG) methods have emerged as a powerful algorithmic tool in the numerical solution of boundary and initial boundary value problems for partial differential equations (PDE) (cf., e.g., [15, 17] and the references therein). For second order elliptic problems, one may distinguish between primal schemes and mixed methods. Primal schemes rely on augmenting the elliptic operator by an appropriate penalization of the discontinuous nodal shape functions. On the other hand, in mixed methods the second order PDE is reformulated as a system of first order PDEs for which suitable numerical fluxes are designed. The most prominent primal schemes are interior penalty discontinuous Galerkin (IPDG) methods, whereas a widely used class of mixed techniques is given by the local discontinuous Galerkin (LDG) methods. Both IPDG and LDG methods have been intensively studied with regard to an a priori error analysis in terms of error estimates for the global discretization error (see, e.g., [2, 3, 12, 26]).

The a posteriori analysis of finite element methods (FEM) is in some state of maturity, as documented by a series of monographs that have been published in recent years [1, 4, 6, 19, 32, 37]. As far as DG methods are concerned, a posteriori error estimators have been developed and analyzed for elliptic problems in H^1 in [7, 25, 27, 33, 34], for elliptic problems in $H(\text{curl})$ in [22, 23], and for the Stokes problem in [24].

*Received by the editors October 5, 2007; accepted for publication (in revised form) July 31, 2008; published electronically December 31, 2008.

<http://www.siam.org/journals/sinum/47-1/70459.html>

[†]Department of Mathematics, University of Houston, Houston, TX 77204-3008, and Institute of Mathematics, University of Augsburg, D-86159 Augsburg, Germany (rohop@math.uh.edu).

[‡]Department of Mathematics, Texas A&M University, College Station, TX 77843 (kanschat@tamu.edu). This author's research was supported by NSF grant DMS-0713829.

[§]CAAM, Rice University, Houston, TX 77005-1892 (timwar@caam.rice.edu). This author's research was supported by NSF grants DMS-0512673 and CNS-0514002 and by AFOSR FA9550-05-1-0473DMS-0713829.

In this paper, we will be concerned with a convergence analysis of an adaptive IPDG method in the sense that for a two-dimensional (2D) second order elliptic model problem we will prove guaranteed error reduction with respect to the mesh-dependent energy norm. We note that for standard conforming P1 approximations of elliptic problems the convergence analysis of adaptive finite element methods (AFEM) has been initiated in [5] and further studied in [18, 29, 30, 31], whereas the issue of optimal order of convergence has been addressed in [8] and [36]. Nonstandard finite element techniques such as mixed and nonconforming methods and edge element discretizations of Maxwell’s equations have been recently investigated in [9, 10, 11]. In the recent paper [28], a convergence analysis of symmetric IPDG methods has been provided. In contrast to [28], our analysis does not require multiple interior nodes for refined elements of the triangulation. In fact, we show that it suffices to refine by bisection.

The paper is organized as follows: In section 2, we briefly introduce the IPDG method. Section 3 describes the adaptive loop consisting of the basic steps SOLVE, ESTIMATE, MARK, and REFINE and states the main convergence result. Section 4 recalls the reliability of the estimator from [27] and establishes a perturbed discrete local efficiency, whereas section 5 is devoted to the proof of the error reduction property. Finally, section 6 contains a documentation of the results of numerical experiments that illustrate the performance of the adaptive IPDG (AIPDG).

2. The IPDG method. We assume $\Omega \subset \mathbb{R}^2$ to be a bounded, polygonal domain with boundary $\Gamma = \partial\Omega, \Gamma = \Gamma_D \cup \Gamma_N, \Gamma_D \cap \Gamma_N = \emptyset$. We adopt standard notation from Sobolev space theory and refer to $(\cdot, \cdot)_{k,D}$ and $\|\cdot\|_{k,D}, k \in \mathbb{N}_0, D \subseteq \Omega$, as the $H^k(D)$ -inner product and associated norm, respectively.

As a model problem, for given $f \in L^2(\Omega), u^D \in H^{1/2}(\Gamma_D), u^N \in L^2(\Gamma_N)$, we consider Poisson’s equation with inhomogeneous Dirichlet and Neumann boundary data

$$\begin{aligned} (2.1a) \quad & -\Delta u = f \quad \text{in } \Omega, \\ (2.1b) \quad & u = u^D \quad \text{on } \Gamma_D, \\ (2.1c) \quad & \partial_{n_{\Gamma_N}} u = u^N \quad \text{on } \Gamma_N, \end{aligned}$$

whose variational formulation amounts to the computation of a solution $u \in V := \{v \in H^1(\Omega) \mid v|_{\Gamma_D} = u^D\}$ such that

$$(2.2) \quad a(u, v) = (f, v)_\Omega + \langle u^N, v \rangle_{\Gamma_N}, \quad v \in H_{0,\Gamma_D}^1(\Omega),$$

where $a(u, v) := \int_\Omega \nabla u \cdot \nabla v dx$.

For the DG approximation of (2.2), we further assume that $\mathcal{T}_H(\Omega)$ is a simplicial triangulation of Ω which aligns with Γ_D, Γ_N on the boundary Γ . For $D \subseteq \bar{\Omega}$, we denote by $|D|$ the volume of D and by $\Pi_p(D), p \in \mathbb{N}_0$, the linear space of polynomials of degree p on D , and we refer to $\mathcal{N}_H(D), \mathcal{E}_H(D)$, and $\mathcal{T}_H(D)$ as the sets of vertices, edges, and elements, respectively, in D . For $T \in \mathcal{T}_H(\Omega)$, h_T stands for the diameter of T , whereas for $E \in \mathcal{E}_H(\bar{\Omega})$, we denote by h_E the length of E . Moreover, for an interior edge $E \in \mathcal{E}_H(\Omega)$ such that $E = T_+ \cap T_-, T_\pm \in \mathcal{T}_H(\Omega)$, we refer to $\omega_E := T_+ \cup T_-$ as the patch formed by the union of the elements sharing E as a common edge. Finally, for a function $g \in L^2(D), D \subset \bar{\Omega}$, the quantity \hat{g}_D stands for the integral mean of g with respect to D , i.e., $\hat{g}_D := |D|^{-1} \int_D g dx$.

We define the product space $V_H := \prod_{T \in \mathcal{T}_H(\Omega)} \Pi_p(T)$, $p \in \mathbb{N}$, and introduce the bilinear form $a_H(\cdot, \cdot) : V_H \times V_H \rightarrow \mathbb{R}$ according to

$$(2.3) \quad a_H(u_H, v_H) := \sum_{T \in \mathcal{T}_H(\Omega)} (\nabla u_H, \nabla v_H)_T \\ - \sum_{E \in \mathcal{E}_H(\bar{\Omega})} \left((\{\partial_{n_E} u_H\}, [v_H])_E + ([u_H]_E, \{\partial_{n_E} v_H\})_E \right) \\ + \alpha \sum_{E \in \mathcal{E}_H(\bar{\Omega})} h_E^{-1} ([u_H]_E, [v_H]_E)_E,$$

where the normal vector on E points from T_+ to T_- and with $v_H^\pm := v_H|_{T_\pm}$ on E ,

$$\begin{aligned} [v_H]_E &:= v_H^+ - v_H^-, & E \in \mathcal{E}_H(\Omega), \\ [v_H]_E &:= v_H|_E, & E \in \mathcal{E}_H(\Gamma), \\ \{v_H\}_E &:= \frac{1}{2} (v_H^+ + v_H^-), & E \in \mathcal{E}_H(\Omega), \\ \{v_H\}_E &:= v_H|_E, & E \in \mathcal{E}_H(\Gamma), \end{aligned}$$

and $\alpha > 0$ stands for a properly chosen penalization parameter.

Then, the interior penalty method in its symmetric formulation amounts to the computation of $u_H \in V_H$ such that

$$(2.4) \quad a_H(u_H, v_H) = \ell(v_H), \quad v_H \in V_H,$$

where

$$(2.5) \quad \ell(v_H) := (f, v_H)_\Omega + (u^N, v_H)_{\Gamma_N} - \sum_{E \subset \Gamma_D} (u^D, \partial_n v_H - \alpha h_E^{-1} v_H)_E.$$

On V_H , we introduce the mesh-dependent H^1 -norm

$$(2.6) \quad \|v_H\|_{1,H,\Omega} := \left(\sum_{T \in \mathcal{T}_H(\Omega)} \|\nabla v_H\|_T^2 + \sum_{E \in \mathcal{E}_H(\bar{\Omega})} (h_E \|\{\partial_{n_E} v_H\}\|_E^2 + h_E^{-1} \|[v_H]\|_E^2) \right)^{1/2}.$$

As has been shown in [27], the bilinear form $a_H(\cdot, \cdot)$ is bounded

$$(2.7) \quad |a_H(u_H, v_H)| \leq (1 + \alpha) \|u_H\|_{1,H,\Omega} \|v_H\|_{1,H,\Omega}, \quad u_H, v_H \in V_H,$$

and for sufficiently large α coercive with respect to the $\|\cdot\|_{1,H,\Omega}$ -norm, i.e., there exist positive constants α_{min} and γ such that for $\alpha \geq \alpha_{min}$

$$(2.8) \quad |a_H(v_H, v_H)| \geq \gamma \|v_H\|_{1,H,\Omega}^2, \quad v_H \in V_H.$$

It follows from (2.7) and (2.8) that for $\alpha \geq \alpha_{min}$ the IPDG (2.4) admits a unique

solution $u_H \in V_H$. Moreover, for such α the mesh-dependent energy norm

$$(2.9) \quad |||v_H|||_{H,\Omega} := a_H(v_H, v_H)^{1/2}, \quad v_H \in V_H,$$

is equivalent to the $\|\cdot\|_{1,H,\Omega}$ -norm

$$(2.10) \quad \gamma \|v_H\|_{1,H,\Omega}^2 \leq |||v_H|||_{H,\Omega}^2 \leq (1 + \alpha) \|v_H\|_{1,H,\Omega}^2, \quad v_H \in V_H.$$

For a subset $D_H \subset \mathcal{T}_H(\Omega)$ of the triangulation, $\|\cdot\|_{1,H,D_H}$ and $|||\cdot|||_{H,D_H}$ are defined analogously.

Remark 2.1. We have chosen the 2D model problem (2.1a)–(2.1c) to focus on the essential ingredients for the proof of the error reduction property and not to overload the convergence analysis with too many technicalities. Using the tools from [29], we think that the results can be extended to more general elliptic differential operators, thus including advection-diffusion problems. We further believe that the ideas presented in this paper can be also adopted to hybridized DG methods [16] where the number of degrees of freedom is comparable to standard finite element discretizations.

3. The adaptive loop and the main convergence result. An adaptive FEM for the IPDG (2.4) consists of successive loops of the following sequence:

$$(3.1) \quad \text{SOLVE} \rightarrow \text{ESTIMATE} \rightarrow \text{MARK} \rightarrow \text{REFINE}.$$

Here, SOLVE stands for the numerical solution of (2.4) with respect to the given triangulation $\mathcal{T}_H(\Omega)$. We remark that for this purpose efficient preconditioned iterative solvers have been developed, analyzed, and implemented (cf., e.g., [20, 21, 25]).

The following residual-type a posteriori error estimator η_H has been introduced and analyzed in [27]:

$$(3.2) \quad \eta_H^2 := \sum_{T \in \mathcal{T}_H(\Omega)} \eta_T^2 + \sum_{E \in \mathcal{E}_H(\Omega)} \eta_E^2.$$

Here, η_T stands for the element residual

$$(3.3) \quad \eta_T := h_T \|f + \Delta u_H\|_T, \quad T \in \mathcal{T}_H(\Omega).$$

On the other hand, η_E summarizes the edge residuals

$$(3.4) \quad \eta_E^2 := \eta_{E,1}^2 + \eta_{E,2}^2 + \eta_{E,N}^2 + \eta_{E,D}^2$$

given by

$$(3.5a) \quad \eta_{E,1} := h_E^{1/2} \|[\partial_{n_E} u_H]\|_E, \quad E \in \mathcal{E}_H(\Omega),$$

$$(3.5b) \quad \eta_{E,2} := h_E^{-1/2} \|[u_H]\|_E, \quad E \in \mathcal{E}_H(\Omega),$$

$$(3.5c) \quad \eta_{E,N} := h_E^{1/2} \|u^N - \partial_{n_E} u_H\|_E, \quad E \in \mathcal{E}_H(\Gamma_N),$$

$$(3.5d) \quad \eta_{E,D} := h_E^{-1/2} \|u^D - u_H\|_E, \quad E \in \mathcal{E}_H(\Gamma_D).$$

The convergence analysis further invokes the data oscillations

$$(3.6) \quad \text{osc}_H^2 := \text{osc}_H^2(f) + \text{osc}_H^2(u^D) + \text{osc}_H^2(u^N),$$

where

$$(3.7a) \quad \text{osc}_H^2(f) := \sum_{T \in \mathcal{T}_H(\Omega)} \text{osc}_T^2(f),$$

$$\text{osc}_T(f) := h_T \|f - \hat{f}_T\|_T,$$

$$(3.7b) \quad \text{osc}_H^2(u^D) := \sum_{E \in \mathcal{E}_H(\Gamma_D)} \text{osc}_E^2(u^D),$$

$$\text{osc}_E(u^D) := h_E^{-1/2} \|u^D - \hat{u}_E^D\|_E,$$

$$(3.7c) \quad \text{osc}_H^2(u^N) := \sum_{E \in \mathcal{E}_H(\Gamma_N)} \text{osc}_E^2(u^N),$$

$$\text{osc}_E(u^N) := h_E^{1/2} \|u^N - \hat{u}_E^N\|_E.$$

In the step MARK of the adaptive loop, given a universal constant $0 < \Theta \leq 1$, we choose subsets $\mathcal{M}_T \subset \mathcal{T}_H(\Omega)$ and $\mathcal{M}_E \subset \mathcal{M}_H(\bar{\Omega})$ such that the following bulk criterion is satisfied:

$$(3.8a) \quad \Theta \sum_{T \in \mathcal{T}_H(\Omega)} \eta_T^2 \leq \sum_{T \in \mathcal{M}_T} \eta_T^2,$$

$$(3.8b) \quad \Theta \sum_{E \in \mathcal{E}_H(\bar{\Omega})} \eta_E^2 \leq \sum_{E \in \mathcal{M}_E} \eta_E^2.$$

The bulk criterion can be realized by a greedy algorithm.

As far as the data oscillations are concerned, for simplicity we assume that the set \mathcal{M}_T selected by (3.8a) is already rich enough such that there exists a constant $0 \leq \rho_2 < 1$ such that

$$(3.9) \quad \text{osc}_h^2 \leq \rho_2 \text{osc}_H^2.$$

We note that the data oscillations may be included in the bulk criterion as well to guarantee (3.9). We refer to [30, 31] for details.

The refinement strategy in the final step REFINE of the adaptive loop is as follows: If an element $T \in \mathcal{T}_H(\Omega)$ has been marked for refinement, it will be refined by longest edge bisection. If an edge $E \in \mathcal{E}_H(\Omega)$, $E = T^+ \cap T^-$ (resp., $E \in \mathcal{T}_H(\Gamma)$, $E = \partial T \cap \Gamma$) has been marked, the triangles T^\pm (resp., the triangle T) will be refined by bisection. We note that this refinement is different from that used in [28] where the refinement of a triangle requires multiple interior nodes based on subsequent regular refinements.

The main result of this paper is a guaranteed error reduction of the global discretization error measured in the mesh-dependent energy norm associated with the IPDG method.

THEOREM 3.1. *Let $u \in V$ be the solution of (2.2), and suppose that $u_H \in V_H$ and $u_h \in V_h$ are the solutions of IPDG (2.4) with respect to the triangulation $\mathcal{T}_H(\Omega)$ and the adaptively refined triangulation $\mathcal{T}_h(\Omega)$ generated according to the refinement rules described before. Assume that (3.9) holds true. Then, for sufficiently large penalization parameter α there exist positive constants $\rho_1 < 1$ and C which depend*

only on α, Θ and the shape regularity of the triangulations such that for $e_H := u - u_H$ and $e_h := u - u_h$ there holds

$$(3.10) \quad \begin{pmatrix} a_h(e_h, e_h) \\ \text{osc}_h^2 \end{pmatrix} \leq \begin{pmatrix} \rho_1 & C \\ 0 & \rho_2 \end{pmatrix} \begin{pmatrix} a_H(e_H, e_H) \\ \text{osc}_H^2 \end{pmatrix}.$$

The proof of Theorem 3.1 will be given in section 5 based on the reliability and a perturbed discrete local efficiency of the estimator (3.2), which will be studied in the following section.

4. Reliability and perturbed discrete local efficiency. The reliability of the residual-type a posteriori error estimator (3.2) has been established in [27] using standard techniques from AFEM [37]. Here, we prove that it is also locally efficient in a relaxed way. We will derive the main lemmas for the case of the newest edge bisection [13, 14, 35].

THEOREM 4.1. *Let $u \in V$ and $u_H \in V_H$ be the solution of (2.2) and its IPDG approximation (2.4), and let η_H and osc_H be the residual error estimator and the data oscillations as given by (3.2) and (3.6), respectively. Then, for $e_H := u - u_H$ there holds*

$$(4.1) \quad a_H(e_H, e_H) \lesssim \eta_H^2.$$

Discrete local efficiency means that up to data oscillations the local contributions of the estimator can be bounded from above by the energy norm of the difference between the fine mesh and coarse mesh approximations on a refined triangle and the patch ω_E associated with a refined edge, respectively [18, 30]. In the framework of the IPDG approximations under consideration, we can prove only a perturbed discrete local efficiency in the sense that the upper bounds involve additional quantities in terms of the fine mesh approximation. In particular, the following result holds true.

THEOREM 4.2. *Let $u \in V$ and $u_H \in V_H, u_h \in V_h$, be the solution of (2.2) and its IPDG approximations (2.4) with respect to $\mathcal{T}_H(\Omega)$ and $\mathcal{T}_h(\Omega)$, respectively. Moreover, let η_H and osc_H be the residual error estimator (3.2) and the data oscillations (3.6), respectively. Then, there holds*

$$(4.2) \quad \begin{aligned} \sum_{T \in \mathcal{M}_T} \eta_T^2 + \sum_{E \in \mathcal{M}_E} \eta_E^2 &\lesssim a_h(u_h - u_H, u_h - u_H) \\ &+ \alpha \sum_{E' \in \mathcal{E}_h(\mathcal{M}_E \setminus \Gamma_D)} h_{E'}^{-1} \| [u_h] \|_{E'}^2 \\ &+ \alpha \sum_{E' \in \mathcal{E}_h(\mathcal{M}_E \cap \Gamma_D)} h_{E'}^{-1} \| u^D - u_h \|_{E'}^2 + \text{osc}_H^2. \end{aligned}$$

Proof. The proof of (4.2) follows by collecting the estimates from the subsequent series of lemmas. \square

LEMMA 4.3. *Let $T \in \mathcal{T}_H(\Omega)$ be a refined triangle such that $T = T_+ \cup T_-, T_{\pm} \in \mathcal{T}_h(\Omega)$. Then, there holds*

$$(4.3) \quad \begin{aligned} h_T^2 \| f + \Delta u_H \|_T^2 &\lesssim a_h|_T(u_h - u_H, u_h - u_H) + \text{osc}_T^2(f) \\ &+ \alpha \sum_{E \in \mathcal{E}_H(\partial T \cap \Omega)} \eta_{E,2}^2 + \alpha \sum_{E \in \mathcal{E}_H(\partial T \cap \Gamma_D)} \eta_{E,D}^2 + \sum_{E \in \mathcal{E}_H(\partial T \cap \Gamma_N)} \text{osc}_E^2(u^N). \end{aligned}$$

Proof. We denote by $CR_p(\Omega; \mathcal{T}_h(\Omega))$, $p \in \mathbb{N}$, the nonconforming Crouzeix–Raviart finite element space, where $v_h|_{T'} \in \Pi_p(T')$, $T' \in \mathcal{T}_h(\Omega)$, is uniquely determined by the degrees of freedom

$$\begin{aligned} \int_E v_h q_E ds, \quad q_E \in \Pi_{p-1}(E), \quad E \in \mathcal{E}_h(T'), \\ \int_T v_h q_{T'} dx, \quad q_{T'} \in \Pi_{p-3}(T') \quad (p \geq 3). \end{aligned}$$

We choose $\varphi_h \in V_h$ with $\varphi_h|_{T_\pm} \in \Pi_p(T_\pm)$ and $\varphi_h|_{T'} \equiv 0$, $T' \in \mathcal{T}_h(\Omega) \setminus \{T\}$, such that

$$(4.4a) \quad h_{T_\pm}^2 \left\| \hat{f}_T + \Delta u_H \right\|_{T_\pm}^2 = \left(\hat{f}_T + \Delta u_H, \varphi_h \right)_{T_\pm},$$

$$(4.4b) \quad \|\varphi_h\|_{T_\pm}^2 \lesssim h_{T_\pm}^4 \left\| \hat{f}_T + \Delta u_H \right\|_{T_\pm}^2,$$

$$(4.4c) \quad (q_h, \varphi_h)_E = 0, \quad q_h \in \Pi_{p-1}(E), \quad E \in \mathcal{E}_h(\partial T).$$

In particular, in the case $p \leq 2$ we choose φ_h as a linear combination of the basis functions associated with the interior edge $E \in \mathcal{E}_h(\text{int}(T))$, whereas for $p \geq 3$ we choose φ_h as a linear combination of the basis functions associated with $\text{int}(T_\pm)$. Using (4.4a), Green’s formula, and setting $T_1 := T_+$, $T_2 := T_-$, we obtain

$$(4.5) \quad h_T^2 \left\| \hat{f}_T + \Delta u_H \right\|_T^2 = \sum_{i=1}^2 \left(\hat{f}_T + \Delta u_H, \varphi_h \right)_{T_i} \\ = \sum_{i=1}^2 \left(-(\nabla u_H, \nabla \varphi_h)_{T_i} + (f, \varphi_h)_{T_i} + \left(\hat{f}_T - f, \varphi_h \right)_{T_i} \right),$$

where we have used that due to (4.4c) for $T \in \mathcal{T}_H$ there holds

$$(4.6a) \quad (\partial_{n_E} u_H, \varphi_h)_E = 0, \quad E \in \mathcal{E}_h(\partial T), \quad p \geq 1,$$

$$(4.6b) \quad (\partial_{n_E} u_H, [\varphi_h])_E = 0, \quad E \in \mathcal{E}_h(\text{int}(T)), \quad p \geq 1.$$

On the other hand, φ_h is an admissible test function in the fine grid equation (2.4) whence

$$(4.7) \quad \sum_{i=1}^2 \left((\nabla u_h, \nabla \varphi_h)_{T_i} - (f, \varphi_h)_{T_i} \right) \\ + \sum_{E \in \mathcal{E}_h(\partial T \cap \Gamma_D)} (u^D, \partial_{n_E} \varphi_h - \alpha h_E^{-1} \varphi_h)_E \\ - \sum_{E \in \mathcal{E}_h(\partial T \cap \Gamma_N)} (u^N, \varphi_h)_E \\ - \sum_{E \in \mathcal{E}_h(T)} \left((\{\partial_{n_E} u_h\}, [\varphi_h])_E + ([u_h], \{\partial_{n_E} \varphi_h\})_E \right) \\ + \alpha \sum_{E \in \mathcal{E}_h(T)} h_E^{-1} ([u_h], [\varphi_h])_E = 0.$$

Adding (4.5) and (4.7) and observing again (4.6a)–(4.6b) as well as $[u_H] = 0$ on $E \in \mathcal{E}_h(\text{int}(T))$, it follows that

$$\begin{aligned}
 (4.8) \quad & h_T^2 \left\| \hat{f}_T + \Delta u_H \right\|_T^2 \\
 &= \sum_{i=1}^2 \left((\nabla(u_h - u_H), \nabla \varphi_h)_{T_i} + \left(\hat{f}_T - f, \varphi_h \right)_{T_i} \right) \\
 &\quad - \sum_{E \in \mathcal{E}_h(T)} \left((\{\partial_{n_E}(u_h - u_H)\}, [\varphi_h])_E \right. \\
 &\quad \quad \left. + ([u_h - u_H], \{\partial_{n_E} \varphi_h\})_E \right) \\
 &\quad - \sum_{E \in \mathcal{E}_h(\partial T \cap \Omega)} ([u_H], \{\partial_{n_E} \varphi_h\})_E \\
 &\quad + \sum_{E \in \mathcal{E}_h(\partial T \cap \Gamma_D)} (u^D - u_H, \partial_{n_E} \varphi_h - \alpha h_E^{-1} \varphi_h)_E \\
 &\quad - \sum_{E \in \mathcal{E}_h(\partial T \cap \Gamma_N)} (u^N - \hat{u}_E^N, \varphi_h)_E \\
 &\quad + \alpha \sum_{E \in \mathcal{E}_h(T)} h_E^{-1} ([u_h - u_H], [\varphi_h])_E \\
 &\quad + \alpha \sum_{E \in \mathcal{E}_h(\partial T \setminus \Gamma_D)} h_E^{-1} ([u_H], [\varphi_h])_E.
 \end{aligned}$$

In view of (4.4b), the inverse inequality and the trace inequalities imply that for $1 \leq i \leq 4$

$$(4.9a) \quad \|\nabla \varphi_h\|_{T_i}^2 \lesssim h_{T_i}^2 \|\hat{f}_T + \Delta u_H\|_{T_i}^2,$$

$$(4.9b) \quad \|\varphi_h\|_E^2 \lesssim h_E^3 \|\hat{f}_T + \Delta u_H\|_{T_i}^2, \quad E \in \mathcal{E}_h(\partial T_i),$$

$$(4.9c) \quad \|\{\partial_{n_E} \varphi_h\}\|_E^2 \lesssim h_E \|\hat{f}_T + \Delta u_H\|_{T_i}^2, \quad E \in \mathcal{E}_h(\partial T_i).$$

Then, using (4.4b) and (4.9a)–(4.9c), straightforward estimation of the terms on the right-hand side in (4.8) gives the assertion. \square

LEMMA 4.4. *Let $E \in \mathcal{E}_H(\Omega)$, $E = T_+ \cap E_-$, $T_\pm \in \mathcal{T}_H(\Omega)$, be a refined edge and $\omega_E := T_+ \cup T_-$. Then, there holds*

$$\begin{aligned}
 (4.10) \quad & h_E \|\{\partial_{n_E} u_H\}\|_E^2 \lesssim \sum_{T_\pm \in \mathcal{T}_H(\omega_E)} \eta_{T_\pm}^2 + \alpha \sum_{E' \in \mathcal{E}_H(\omega_E \cap \Omega)} \eta_{E',2}^2 \\
 & + \alpha \sum_{E \in \mathcal{E}_H(\partial \omega_E \cap \Gamma_D)} \eta_{E',D}^2 + \sum_{E \in \mathcal{E}_H(\partial \omega_E \cap \Gamma_N)} \text{osc}_{E'}^2(u^N).
 \end{aligned}$$

For a refined edge $E \in \mathcal{E}_H(\Gamma_N)$ with $E = \partial T \cap \Gamma_N$, $T \in \mathcal{T}_H(\Omega)$, we have

$$(4.11) \quad h_E \|u^N - \partial_{n_E} u_H\|_E^2 \lesssim \eta_T^2 + \alpha \sum_{E' \in \mathcal{E}_H(T \cap \Omega)} \eta_{E',2}^2.$$

Proof. For the proof of (4.10) let us assume that $E = T_+ \cap T_-$, $T_\pm \in \mathcal{T}_H(\Omega)$. We choose $\varphi_H \in CR_p(\Omega; \mathcal{T}_H(\Omega))$ as a linear combination of the basis functions associated

with the edge E such that

$$(4.12a) \quad h_E \|[\partial_{n_E} u_H]\|_E^2 = ([\partial_{n_E} u_H], \varphi_H)_E,$$

$$(4.12b) \quad \|\varphi_H\|_{T_\pm} \lesssim h_E^{3/2} \|[\partial_{n_E} u_H]\|_E,$$

$$(4.12c) \quad (q_{E'}, \varphi_H)_{E'} = 0, \quad q_{E'} \in \Pi_{p-1}(E')$$

for any edge E' . Using the definition of $\partial_{n_E} u_H$, it follows that

$$(4.13) \quad \begin{aligned} & ([\partial_{n_E} u_H], \varphi_H)_E \\ &= (\nu_E^+ \cdot \nabla u_H^+, \varphi_H)_E + (\nu_E^- \cdot \nabla u_H^-, \varphi_H)_E. \end{aligned}$$

By Green's formula, we find

$$(4.14) \quad \begin{aligned} & (\nabla u_H, \nabla \varphi_H)_{T_\pm} \\ &= -(\Delta u_H, \varphi_H)_{T_\pm} + \sum_{E' \in \mathcal{E}_H(\partial T_\pm)} (\partial_{n_{E'}} u_H, \varphi_H)_{E'}. \end{aligned}$$

By (4.12c) we have

$$(4.15) \quad (\partial_{n_{E'}} u_H, \varphi_H)_{E'} = 0, \quad E' \in \mathcal{E}_H(\partial \omega_E),$$

whence

$$(4.16) \quad \begin{aligned} & h_E \|[\partial_{n_E} u_H]\|_E^2 \\ &= ([\partial_{n_E} u_H], \varphi_H)_E = (\nabla u_H, \nabla \varphi_H)_{\omega_E} + (\Delta u_H, \varphi_H)_{\omega_E}. \end{aligned}$$

On the other hand, since φ_H is an admissible test function in (2.4), we have

$$(4.17) \quad \begin{aligned} & (\nabla u_H, \nabla \varphi_H)_{\omega_E} = (f, \varphi_H)_{\omega_E} + \sum_{E' \in \mathcal{E}_H(\partial \omega_E \cap \Gamma_N)} (u^N, \varphi_H)_{E'} \\ & - \sum_{E' \in \mathcal{E}_H(\partial \omega_E \cap \Gamma_D)} (u^D, \partial_{n_{E'}} \varphi_H - \alpha h_{E'}^{-1} \varphi_H)_{E'} \\ & + \sum_{E' \in \mathcal{E}_H(\omega_E)} ([u_H], \{\partial_{n_{E'}} \varphi_H\})_{E'} - \alpha \sum_{E' \in \mathcal{E}_H(\partial \omega_E)} h_{E'}^{-1} ([u_H], [\varphi_H])_{E'}, \end{aligned}$$

where we have used (4.15) and (4.12c) on E . Combining (4.16) and (4.17) results in

$$(4.18) \quad \begin{aligned} & h_E \|[\partial_{n_E} u_H]\|_E^2 = (f + \Delta u_H, \varphi_H)_{\omega_E} \\ & + \sum_{E' \in \mathcal{E}_H(\omega_E \setminus \Gamma_D)} ([u_H], \{\partial_{n_{E'}} \varphi_H\})_{E'} \\ & - \alpha \sum_{E' \in \mathcal{E}_H(\partial \omega_E \setminus \Gamma_D)} h_{E'}^{-1} ([u_H], [\varphi_H])_{E'} \\ & + \sum_{E' \in \mathcal{E}_H(\partial \omega_E \cap \Gamma_N)} (u^N - \hat{u}_{E'}^N, \varphi_H)_{E'} \\ & - \sum_{E' \in \mathcal{E}_H(\partial \omega_E \cap \Gamma_D)} (u^D - u_H, \partial_{n_{E'}} \varphi_H - \alpha h_{E'}^{-1} \varphi_H)_{E'}. \end{aligned}$$

Observing (4.12b), the trace inequalities yield

$$(4.19a) \quad \|\varphi_H\|_{E'} \lesssim h_E \|[\partial_{n_E} u_H]\|_E, \quad E' \in \mathcal{E}_H(\partial \omega_E),$$

$$(4.19b) \quad \|[\partial_{n_{E'}} \varphi_H]\|_{E'} \lesssim \|[\partial_{n_E} u_H]\|_E, \quad E' \in \mathcal{E}_H(\partial \omega_E).$$

Taking advantage of (4.12b), (4.19a), and (4.19b), the assertion can be deduced by straightforward estimation of the terms on the right-hand side in (4.18). The proof of (4.11) follows by similar arguments. \square

LEMMA 4.5. *Let $E \in \mathcal{E}_H(\Omega)$, $E = T_+ \cap T_-$, $T_\pm \in \mathcal{T}_H(\Omega)$, be a refined edge and $\omega_E := T_+ \cup T_-$. Then, there holds*

$$(4.20) \quad \alpha h_E^{-1} \|[u_H]|_E\|_E^2 \lesssim a_h|_{\omega_E}(u_h - u_H, u_h - u_H) + \alpha \sum_{E' \in \mathcal{E}_h(E)} h_{E'}^{-1} \|[u_h]|_{E'}^2.$$

Likewise, if $E \in \mathcal{E}_H(\Gamma_D)$ is a refined edge such that $E = \partial T \cap \Gamma_D$, $T \in \mathcal{T}_H(\Omega)$, there holds

$$(4.21) \quad \alpha h_E^{-1} \|u^D - u_H\|_E^2 \lesssim a_h|_T(u_h - u_H, u_h - u_H) + \alpha \sum_{E' \in \mathcal{E}_h(E)} h_{E'}^{-1} \|u^D - u_h\|_{E'}^2 + \text{osc}_E^2(u^D).$$

Proof. For the proof of (4.20), choose $\psi_H^\pm \in CR_p(\Omega; \mathcal{T}_H(\Omega))$ with $\text{supp}(\psi_H^\pm) = T_\pm$ as a linear combination of basis functions associated with E such that

$$(4.22a) \quad ([u_H], \psi_H^\pm)_E = \pm \frac{1}{2} \|[u_H]|_E\|_E^2,$$

$$(4.22b) \quad \|\psi_H^\pm\|_{T_\pm} \lesssim h_E^{1/2} \|[u_H]|_E\|_E.$$

We define $\varphi_H \in V_H$ by $\varphi_H|_{T_\pm} = \psi_H^\pm$ and $\varphi_H|_T \equiv 0$, $T \in \mathcal{T}_H(\Omega) \setminus \{\omega_E\}$. Then, it follows from (4.22a) that

$$(4.23) \quad \alpha h_E^{-1} \|[u_H]|_E\|_E^2 = \alpha h_E^{-1} ([u_H], [\varphi_H])_E.$$

Since φ_H is an admissible test function in (2.4), we have

$$(4.24) \quad \begin{aligned} & \alpha h_E^{-1} ([u_H], [\varphi_H])_E = \\ & - (\nabla u_H, \nabla \varphi_H)_{\omega_E} + (f, \varphi_H)_{\omega_E} \\ & + \sum_{E' \in \mathcal{E}_H(\partial \omega_E \cup \{E\})} (\{\partial_{n_{E'}} u_H\}, [\varphi_H])_{E'} \\ & + \sum_{E' \in \mathcal{E}_H(\partial \omega_E) \cup \{E\}} ([u_H], \{\partial_{n_{E'}} \varphi_H\})_{E'} \\ & - \alpha \sum_{E' \in \mathcal{E}_H(\partial \omega_E)} h_{E'}^{-1} ([u_H], [\varphi_H])_{E'} \\ & + \sum_{E' \in \mathcal{E}_H(\partial \omega_E \cap \Gamma_N)} (u^N, \varphi_H)_{E'} \\ & - \alpha \sum_{E' \in \mathcal{E}_H(\partial \omega_E \cap \Gamma_D)} (u^D, \partial_{n_{E'}} \varphi_H - \alpha h_{E'}^{-1} \varphi_H)_{E'}. \end{aligned}$$

On the other hand, $(\varphi_H|_{T'})_{T' \in \mathcal{T}_h(\Omega)}$ is an admissible test function in the fine grid

equation (2.4). Hence, observing $[\varphi_H] = 0$ and $[u_H] = 0$ on $E' \in \mathcal{E}_h(\text{int}(T_\pm))$, we get

$$\begin{aligned}
 (4.25) \quad 0 &= (\nabla u_h, \nabla \varphi_H)_{\omega_E} - (f, \varphi_H)_{\omega_E} \\
 &\quad - \sum_{E' \in \mathcal{E}_h(\partial\omega_E \cup \{E\})} (\{\partial_{n_{E'}} u_h\}, [\varphi_H])_{E'} \\
 &\quad - \sum_{E' \in \mathcal{E}_h(\partial\omega_E \cup \{E\})} ([u_h], \{\partial_{n_{E'}} \varphi_H\})_{E'} \\
 &\quad - \sum_{E' \in \mathcal{E}_h(\text{int}(\omega_E) \setminus \{E\})} ([u_h - u_H], \{\partial_{n_{E'}} \varphi_H\})_{E'} \\
 &\quad + \alpha \sum_{E' \in \mathcal{E}_h(\partial\omega_E \cup \{E\})} h_{E'}^{-1} ([u_h], [\varphi_H])_{E'} \\
 &\quad - \sum_{E' \in \mathcal{E}_h(\partial\omega_E \cap \Gamma_N)} (u^N, \varphi_H)_{E'} \\
 &\quad + \alpha \sum_{E' \in \mathcal{E}_h(\partial\omega_E \cap \Gamma_D)} (u^D, \partial_{n_{E'}} \varphi_H - \alpha h_{E'}^{-1} \varphi_H)_{E'}.
 \end{aligned}$$

In view of (4.22b), the inverse inequality and the trace inequalities imply

$$(4.26a) \quad \|\nabla \psi_H^\pm\|_{T_\pm} \lesssim h_E^{-1/2} \|[u_H]\|_E,$$

$$(4.26b) \quad \|\psi_H^\pm\|_{E'} \lesssim \|[u_H]\|_E, \quad E' \in \mathcal{E}_H(\omega_E),$$

$$(4.26c) \quad \|\partial_{n_{E'}} \psi_H^\pm\|_{E'} \lesssim h_E^{-1} \|[u_H]\|_E, \quad E' \in \mathcal{E}_H(\omega_E).$$

Combining (4.24) and (4.25) and using (4.22b) and (4.26a)–(4.26c), straightforward estimation gives the assertion. The proof of (4.21) can be established similarly. \square

Remark 4.6. The proof of the perturbed discrete local efficiency is carried out under the assumption of geometrically conforming meshes. However, the fact that in Lemmas 4.4 and 4.5 the admissible test functions for the fine grid equation are chosen as linear combinations of the coarse grid Crouzeix–Raviart basis functions allows the handling of hanging nodes as well.

5. Proof of the error reduction property. In the convergence analysis of standard FEMs [18, 30], the proof of the error reduction property makes essential use of Galerkin orthogonality which in the framework of IPDG reads as follows:

$$(5.1) \quad a_h(u_h - u_H, u_h - u_H) = a_h(e_H, e_H) - a_h(e_h, e_h).$$

However, we measure the error e_H with respect to the mesh-dependent energy norm $a_H(\cdot, \cdot)$ associated with the coarse mesh $\mathcal{T}_H(\Omega)$, and, hence, (5.1) cannot be used directly. It is known from the convergence analysis of adaptive nonconforming finite elements [10] or of mixed finite elements [11] that in the absence of Galerkin orthogonality convergence can be established provided some sort of perturbed Galerkin orthogonality holds true. For the IPDG under consideration, we can rewrite (5.1) according to

$$(5.2) \quad a_h(u_h - u_H, u_h - u_H) = (1 + \delta_{h,H}(e_H)) a_H(e_H, e_H) - a_h(e_h, e_h),$$

where in the case $a_H(e_H, e_H) \neq 0$ the perturbation term $\delta_{h,H}(e_H)$ is given by

$$(5.3) \quad \delta_{h,H}(e_H) := \frac{a_h(e_H, e_H) - a_H(e_H, e_H)}{a_H(e_H, e_H)}.$$

We would be able to conclude, if we can show that $\delta_{h,H}(e_H)$ can be made sufficiently small.

LEMMA 5.1 (perturbed Galerkin orthogonality). *There exists a positive constant C_1 depending only on the local geometry of the triangulations such that for the perturbation term $\delta_{h,H}(e_H)$ there holds*

$$(5.4) \quad \delta_{h,H}(e_H) \leq \frac{C_1}{\alpha}.$$

Proof. Following the reasoning in [28, Proposition 4.1], we can easily show

$$a_h(e_H, e_H) \leq a_H(e_H, e_H) + c_1 \alpha \left(\sum_{E \in \mathcal{E}_H(\Omega)} h_E^{-1} \| [u_H] \|_E^2 + \sum_{E \in \mathcal{E}_H(\Gamma_D)} h_E^{-1} \| u^D - u_H \|_E^2 \right),$$

where $c_1 > 0$ is a constant depending only on the local geometry of the triangulations. On the other hand, the local efficiency of the residual estimator (cf. [27]) tells us that there exists another positive constant c_2 which also depends only on the local geometry of the triangulations such that

$$(5.5) \quad \sum_{E \in \mathcal{E}_H(\Omega)} h_E^{-1} \| [u_H] \|_E^2 + \sum_{E \in \mathcal{E}_H(\Gamma_D)} h_E^{-1} \| u^D - u_H \|_E^2 \leq \frac{c_2}{\alpha^2} a_H(e_H, e_H).$$

Combining the two preceding estimates allows us to conclude with $C_1 := c_1 c_2$. \square

Proof of Theorem 3.1. The reliability, the bulk criterion, and the discrete local efficiency infer the existence of a positive constant C_2 depending only on γ, Θ , and the local geometry of the triangulations such that

$$a_H(e_H, e_H) \leq C_2 \left(a_h(u_h - u_H, u_h - u_H) + \text{osc}_H^2 + \alpha \sum_{E \in \mathcal{E}_h(\Omega)} h_E^{-1} \| [u_h] \|_E^2 + \alpha \sum_{E \in \mathcal{E}_h(\Gamma_D)} h_E^{-1} \| u_D - u_h \|_E^2 \right).$$

Using (5.2), (5.4), and (5.5) with h instead of H , we obtain the existence of a positive constant C_3 such that

$$a_H(e_H, e_H) \leq C_2 \left(1 + \frac{2C_1}{\alpha} \right) a_H(e_H, e_H) - \left(C_2 - \frac{C_3}{\alpha} \right) a_h(e_h, e_h) + C_2 \text{osc}_H^2,$$

from which we deduce

$$a_h(e_h, e_h) \leq \left(C_2 - \frac{C_3}{\alpha}\right)^{-1} \left[\left(\left(1 + \frac{2C_1}{\alpha}\right) C_2 - 1 \right) a_H(e_H, e_H) + C_2 \operatorname{osc}_H^2 \right].$$

For $\alpha > 2C_1C_2 + C_3$, the error reduction property (3.10) results with $\rho_1 := (C_2 - \frac{C_3}{\alpha})^{-1}((1 + \frac{2C_1}{\alpha})C_2 - 1) < 1$. \square

Remark 5.2. We note that $a_h(\cdot, \cdot)$ is not coercive on the energy space. However, it can be shown (cf. Proposition 4.2 in [28]) that $\|e_h\|_{1,h,\Omega}^2 \lesssim a_h(e_h, e_h)$.

Remark 5.3. The coefficients $C_i, 1 \leq i \leq 3$, depend on the local geometry of the triangulations which is determined by the initial coarse triangulation and the refinement process. Moreover, C_2 depends on γ in (2.8) and on $0 < \Theta \leq 1$ in (3.8a) and (3.8b). For an appropriate initial triangulation, we can expect $C_i = O(1), 1 \leq i \leq 3$, so that the requirement $\alpha > C_1C_2 + C_3$ results in values of α of approximately the same magnitude as required for the coercivity of the mesh-dependent bilinear forms. This is confirmed by the numerical results in section 6.

6. Computational results. In the following numerical experiments, we used the bisection algorithm, for all test cases, derived from the *AFEM@Matlab* implementation [14].

We verify the suitability of our theoretical results using standard test cases (see [10]). They are studies of the behavior of the algorithm in the case of the standard singularities induced by a reentrant corner of the domain. The right-hand side is chosen to be zero, and, therefore, data oscillations are present only on the boundary edges not adjacent to the singularity, where values of the analytical solutions are prescribed. The penalty parameter α has been chosen according to $\alpha = 15(p + 1)^2$ as dictated by the coercivity requirement (2.8) (cf. also Remark 5.3).

First, we study the L-shaped domain with Dirichlet data $u^D = 0$ on the two edges adjacent to the reentrant corner and Neumann data on the remaining boundary. The refinement parameter is chosen as $\Theta = 0.6$. Table 6.1 shows a decline of the energy norm for this case by a factor of about 2/3 in each refinement step. This factor is only slightly better for quartic shape functions, confirming that the reduction rate depends mostly on Θ . Nevertheless, the meshes for P_4 are growing much slower, and in both cases we obtain the optimal approximation rates in terms of N_{dof} , namely, $N_{\text{dof}}^{-1/2}$ and N_{dof}^{-2} . Data oscillation occurs only at the outer boundary and is negligible in this case.

TABLE 6.1

Decline of the energy norm and data oscillation in terms of the refinement step, polynomial degrees 1 and 4.

l	P_1			P_4		
	N_{dof}	$\ e_l\ _A$	osc_l	N_{dof}	$\ e_l\ _A$	osc_l
0	36	2.81e-1	9.32e-2	180	6.07e-2	9.32e-2
1	114	1.98e-1	6.83e-2	570	3.83e-2	6.83e-2
2	252	1.39e-1	3.35e-2	960	2.60e-2	5.38e-2
3	630	9.53e-2	1.66e-2	1440	1.75e-2	3.35e-2
4	1428	6.48e-2	1.07e-2	2280	1.23e-2	2.28e-2
5	3180	4.42e-2	5.85e-3	3000	7.72e-3	1.71e-2
6	6714	3.00e-2	4.10e-3	4020	4.86e-3	1.15e-2
7	14076	2.08e-2	2.46e-3	5355	3.06e-3	8.41e-3
8	28368	1.43e-2	1.51e-3	6330	1.93e-3	6.20e-3
9	58461	9.91e-3	9.18e-4	7620	1.22e-3	4.31e-3

TABLE 6.2

Decline of the energy and data oscillation in terms of the refinement step, polynomial degrees 1 and 4.

l	P_1			P_4		
	N_{dof}	$\ e_l\ _A$	osc_l	N_{dof}	$\ e_l\ _A$	osc_l
0	18	8.83e-1	2.32e-1	90	5.16e-1	2.37e-1
1	48	6.90e-1	2.09e-1	165	4.39e-1	2.20e-1
2	138	5.76e-1	1.80e-1	285	3.61e-1	1.95e-1
3	219	4.86e-1	1.61e-1	690	3.03e-1	1.83e-1
...						
18	54804	4.42e-2	3.59e-2	32400	2.27e-2	4.69e-2
19	76809	3.73e-2	3.21e-2	39630	1.91e-2	4.22e-2
20	106821	3.16e-2	2.88e-2	52605	1.59e-2	3.79e-2
21	149829	2.67e-2	2.57e-2	63480	1.34e-2	3.41e-2

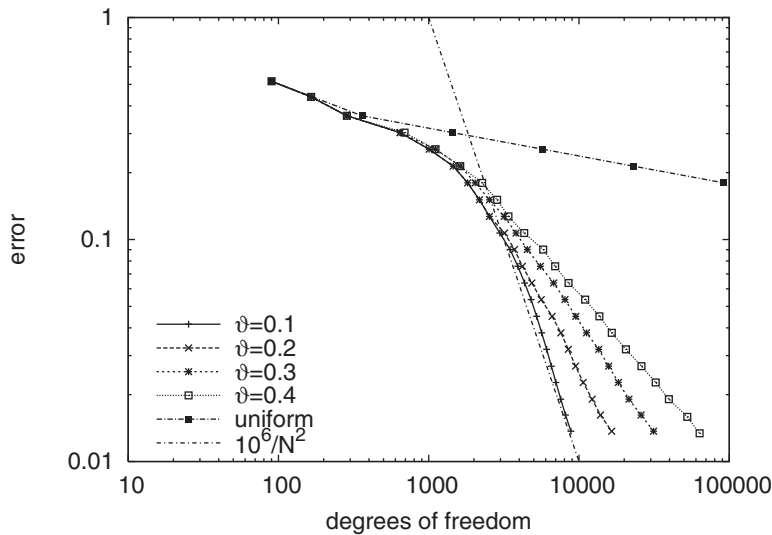


FIG. 6.1. Error versus number of degrees of freedom for the slit domain, quartic polynomials.

Next, we study the higher singularity of the slit domain. Here, Dirichlet boundary conditions are used on the whole boundary. Table 6.2 shows that for $\Theta = 0.4$ we obtain again constant error reduction rates.

The solution in this example is highly singular, and we expect that at least for higher order polynomials the refinement should be very local. Indeed, Figure 6.1 shows that the parameter Θ must be chosen carefully in order to obtain the optimal approximation with respect to the degrees of freedom, confirming results from [36] for standard AFEM. Only the very small value of $\Theta = 0.1$ is able to reproduce the optimal convergence order of N^{-2} . Figure 6.2 shows that this corresponds asymptotically to adding only 1/16 of the current number of cells in each step.

Even with the small size of $\Theta = 0.1$, the optimal N -term approximation rate is obtained only after several thousand degrees of freedom. Comparing Figures 6.1 and 6.2, we note that this corresponds to the fact that the bulk criterion refines much faster than the asymptotic rate in its initialization phase. On the other hand, this fast refinement allows the method to reach the asymptotic regime in only about 10 steps. Figure 6.3 shows that the refinement for $\Theta = 0.1$ is much more concentrated

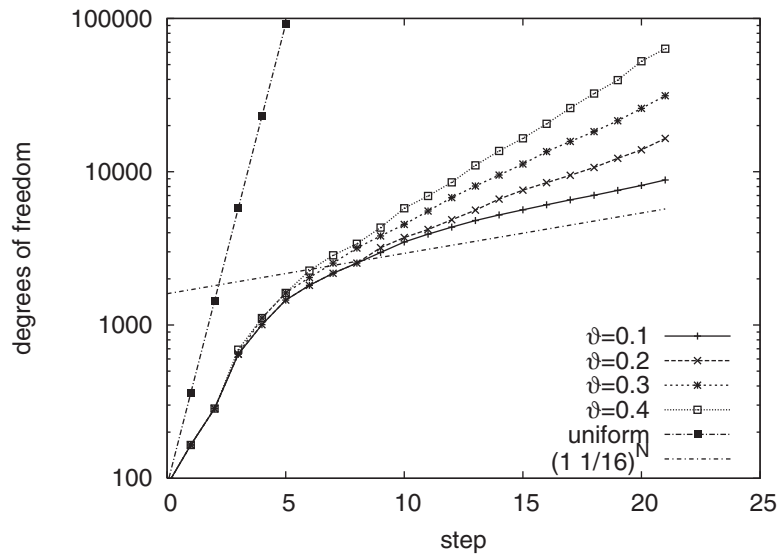


FIG. 6.2. Development of mesh sizes during adaptive refinement, quartic polynomials.

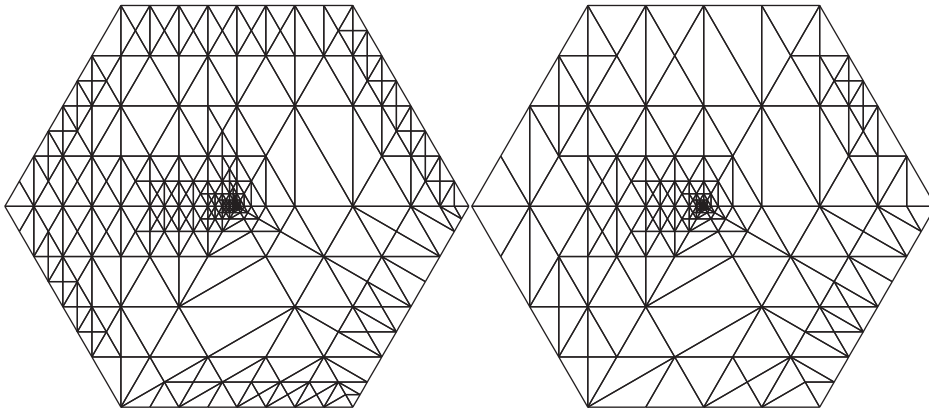


FIG. 6.3. Meshes for the slit domain, polynomial degree 4, $\Theta = 0.4$ (left) and $\Theta = 0.1$ (right).

at the central singularity, while $\Theta = 0.4$ puts more weight in reducing the boundary projection errors.

REFERENCES

- [1] M. AINSWORTH AND J.T. ODEN, *A Posteriori Error Estimation in Finite Element Analysis*, Wiley, Chichester, UK, 2000.
- [2] D.N. ARNOLD, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.
- [3] D.N. ARNOLD, F. BREZZI, B. COCKBURN, AND L.D. MARINI, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.
- [4] I. BABUSKA AND T. STROUBOULIS, *The Finite Element Method and its Reliability*, Clarendon Press, Oxford, 2001.
- [5] I. BABUSKA AND M. VOGELIUS, *Feedback and adaptive finite element solution of one-dimensional boundary value problems*, Numer. Math., 44 (1984), pp. 75–102.

- [6] W. BANGERTH AND R. RANNACHER, *Adaptive Finite Element Methods for Differential Equations*, Lectures Math. ETH Zürich Birkhäuser, Basel, Switzerland, 2003.
- [7] R. BECKER, P. HANSBO, AND M.G. LARSON, *Energy norm a posteriori error estimation for discontinuous Galerkin methods*, *Comput. Methods Appl. Mech. Engrg.*, 192 (2003), pp. 723–733.
- [8] P. BINEV, W. DAHMEN, AND R. DEVORE, *Adaptive finite element methods with convergence rates*, *Numer. Math.*, 97 (2004), pp. 219–268.
- [9] C. CARSTENSEN AND R.H.W. HOPPE, *Convergence analysis of an adaptive edge finite element method for the 2d eddy current equations*, *J. Numer. Math.*, 13 (2005), pp. 19–32.
- [10] C. CARSTENSEN AND R.H.W. HOPPE, *Convergence analysis of an adaptive nonconforming finite element method*, *Numer. Math.*, 103 (2006), pp. 251–266.
- [11] C. CARSTENSEN AND R.H.W. HOPPE, *Error reduction and convergence for an adaptive mixed finite element method*, *Math. Comp.*, 75 (2006), pp. 1033–1042.
- [12] P. CASTILLO, B. COCKBURN, I. PERUGIA, AND D. SCHÖTZAU, *An a priori error analysis of the local discontinuous Galerkin method for elliptic problems*, *SIAM J. Numer. Anal.*, 38 (2000), pp. 1676–1706.
- [13] L. CHEN, *Short Bisection Implementation in MATLAB*, preprint, Department of Mathematics, University of Maryland, College Park, MD, 2006.
- [14] L. CHEN AND C. ZHANG, *AFEM@matlab: A MATLAB Package of Adaptive Finite Element Methods*, Technical report, University of Maryland, College Park, MD, 2006.
- [15] B. COCKBURN, *Discontinuous Galerkin methods*, *ZAMM Z. Angew. Math. Mech.*, 83 (2003), pp. 731–754.
- [16] B. COCKBURN, J. GOPALAKRISHNAN, AND R. LAZAROV, *Unified Hybridization of Discontinuous Galerkin, Mixed and Continuous Galerkin Methods for Second Order Elliptic Problems*, *SIAM J. Numer. Anal.*, to appear.
- [17] B. COCKBURN, G.E. KARNIADAKIS, AND C.-W. SHU, EDS., *Discontinuous Galerkin Methods*, Lecture Notes in Comput. Sci. Engrg. 11, Springer, Berlin, Heidelberg, New York, 2000.
- [18] W. DÖRFLER, *A convergent adaptive algorithm for Poisson's equation*, *SIAM J. Numer. Anal.*, 33 (1996), pp. 1106–1124.
- [19] K. ERIKSSON, D. ESTEP, P. HANSBO, AND C. JOHNSON, *Computational Differential Equations*, Cambridge University Press, Cambridge, 1995.
- [20] J. GOPALAKRISHNAN AND G. KANSCHAT, *Multi-level preconditioners for the interior penalty method*, in *Numerical Mathematics and Advanced Applications: Proceedings of the ENUMATH 2001*, F. Brezzi et al. eds., Springer, Milan, 2003, pp. 795–804.
- [21] J. GOPALAKRISHNAN AND G. KANSCHAT, *A multilevel discontinuous Galerkin method*, *Numer. Math.*, 95 (2003), pp. 527–550.
- [22] P. HOUSTON, I. PERUGIA, AND D. SCHÖTZAU, *Energy norm a posteriori error estimation for mixed discontinuous Galerkin approximations of the Maxwell operator*, *Comput. Methods Appl. Mech. Engrg.*, 194 (2005), pp. 499–510.
- [23] P. HOUSTON, I. PERUGIA, AND D. SCHÖTZAU, *A posteriori error estimation for discontinuous Galerkin discretizations of $H(\text{curl})$ -elliptic partial differential equations*, *IMA J. Numer. Anal.*, to appear.
- [24] P. HOUSTON, D. SCHÖTZAU, AND T. WIHLER, *Energy norm a posteriori error estimation for mixed discontinuous Galerkin approximations of the Stokes problem*, *J. Sci. Comput.*, 22 (2005), pp. 357–380.
- [25] G. KANSCHAT, *Discontinuous Galerkin Methods for Viscous Incompressible Flow*, Deutscher Universitätsverlag, Wiesbaden, Germany, 2007.
- [26] G. KANSCHAT AND R. RANNACHER, *Local error analysis of the interior penalty discontinuous Galerkin method for second order problems*, *J. Numer. Math.*, 10 (2002), pp. 249–274.
- [27] O.A. KARAKASHIAN AND F. PASCAL, *A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems*, *SIAM J. Numer. Anal.*, 41 (2003), pp. 2374–2399.
- [28] O. KARAKASHIAN AND F. PASCAL, *Convergence of Adaptive Discontinuous Galerkin Approximations of Second-order Elliptic Problems*, *SIAM J. Numer. Anal.*, 45 (2007), pp. 641–665.
- [29] K. MEKCHAY AND R.H. NOCHETTO, *Convergence of adaptive finite element methods for general second order linear elliptic PDEs*, *SIAM J. Numer. Anal.*, 43 (2005), pp. 1803–1827.
- [30] P. MORIN, R.H. NOCHETTO, AND K.G. SIEBERT, *Data oscillation and convergence of adaptive FEM*, *SIAM J. Numer. Anal.*, 38 (2000), pp. 466–488.
- [31] P. MORIN, R.H. NOCHETTO, AND K.G. SIEBERT, *Convergence of adaptive finite element methods*, *SIAM Rev.*, 44 (2002), pp. 631–658.
- [32] P. NEITTAANMÄKI AND S. REPIN, *Reliable Methods for Mathematical Modelling. Error Control and a Posteriori Estimates*, Elsevier, New York, 2004.

- [33] B. RIVIÈRE, M.F. WHEELER, AND V. GIRAULT, *Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems, Part I.*, *Comput. Geom.*, 3 (1999), pp. 337–360.
- [34] B. RIVIÈRE AND M.F. WHEELER, *A posteriori error estimates and mesh adaptation strategy for discontinuous Galerkin methods applied to diffusion problems*, *Comput. Math. Appl.*, 46 (2003), pp. 141–163.
- [35] A. SCHMIDT AND K.G. SIEBERT, *Design of Adaptive Finite Element Software: The Finite Element Toolbox ALBERTA*, *Lecture Notes in Comput. Sci. Engrg.* 42, Springer, Berlin, 2005.
- [36] R. STEVENSON, *Optimality of a standard adaptive finite element method*, *Found. Comput. Math.*, 7 (2007), pp. 245–269.
- [37] R. VERFÜRTH, *A Review of A Posteriori Estimation and Adaptive Mesh-Refinement Techniques*, Wiley-Teubner, New York, Stuttgart, 1996.

MODELING OF THERMALLY ASSISTED MAGNETODYNAMICS*

LUBOMÍR BAÑAS[†], ANDREAS PROHL[‡], AND MARIÁN SLODIČKA[§]

Abstract. Thermomagnetic recording uses local heating of ferromagnetic media to locally decrease coercivity and change saturation magnetization of the material and alleviate magnetization reversal. A modified Landau–Lifshitz equation is proposed as a phenomenological model to allow for changes in magnetization magnitude and is studied both analytically and numerically.

Key words. temperature dependent Landau–Lifshitz equation, full-discretization, numerical analysis, convergence

AMS subject classifications. 65M12, 74F15

DOI. 10.1137/070694995

1. Introduction. The application of thermal energy to enable recording on extremely high anisotropy magnetic media at room temperature is a viable means of extending the density of stored information on hard disk drives; cf. [14, 20, 21, 23]. In order to guarantee long-term data/magnetic regions thermal stability, a usable signal-to-noise ratio has to exist to find, follow, and read the smallest bits. Stability depends on material properties like saturation magnetization $\tilde{M}_s = \tilde{M}_s(\tau)$ and uniaxial magnetocrystalline anisotropy $\tilde{K} = \tilde{K}(\tau)$, where both (in magnitude) monotonically drop towards zero as the material temperature τ is increased toward the Curie temperature τ_C —the temperature where thermal energy overcomes electronic exchange forces in ferromagnets and produces a randomizing effect, leading to total disorder, and hence zero saturation magnetization; cf. Figure 1. Thermomagnetic recording uses local heating close to or above Curie temperature of the ferromagnetic medium during recording to locally decrease $K(\tau)$ to thus lower the energy barrier for reversal, and allow one to write data at available magnetic fields from recording heads. The written bits rapidly freeze during the cooling process and are stable at room temperature. Throughout this process in ferromagnetic films with tailored anisotropy, saturation magnetization behavior, and Curie temperature (see, e.g., [28], [21]), the heat deposition should be minimum and concentrated at the recording site, with optimized heating profile and cooling rate [20], to avoid lost of neighboring information and damage of ferromagnetic material at peak cycling temperatures; see Figure 1.

Isothermal gyromagnetic dynamics of single magnetic moment particles are described by the Landau–Lifshitz (LL) equation: Let $M_s(t, \mathbf{x}) \equiv \tilde{M}_s(\tau(t, \mathbf{x}))$ for $(t, \mathbf{x}) \in \Omega_T$. Given $\mathbf{m}(0, \cdot) = \mathbf{m}_0$, for $|\mathbf{m}_0| = M_s \geq 0$, solve

$$(1.1) \quad \mathbf{m}_t = -\gamma \mathbf{m} \times \mathbf{h}_{\text{eff}} - \gamma \delta \frac{\mathbf{m}}{M_s} \times (\mathbf{m} \times \mathbf{h}_{\text{eff}}) \quad \text{in } \Omega_T := (0, T) \times \Omega,$$

*Received by the editors June 21, 2007; accepted for publication (in revised form) August 11, 2008; published electronically December 31, 2008.

<http://www.siam.org/journals/sinum/47-1/69499.html>

[†]Department of Mathematics, Heriot-Watt University, EH14 4AS Edinburgh, United Kingdom (l.banas@hw.ac.uk, <http://www.ma.hw.ac.uk/~lubomir>). This author’s research was supported by EPSRC grant EP/C548973/1.

[‡]Mathematisches Institut, Universität Tübingen, Auf der Morgenstelle 10, D-72076 Tübingen, Germany (prohl@na.uni-tuebingen.de).

[§]Department of Mathematical Analysis, Faculty of Engineering, Ghent University, Galglaan 2, B-9000 Gent, Belgium (marian.slodicka@ugent.be, <http://cage.ugent.be/~ms>). This author’s research was supported by BOF/GOA-project 01G00607, Ghent University, Belgium.

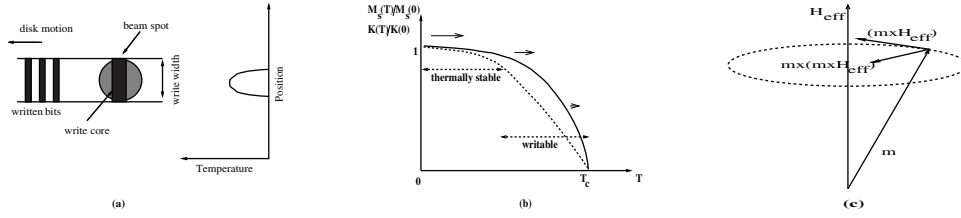


FIG. 1. (a) *thermally assisted magnetic recording*, (b) *temperature-dependent saturation magnetization \tilde{M}_s resp. anisotropy \tilde{K}* (dotted), and (c) *precession of magnetization*.

with the first term on the right-hand side representing magnetic moment precession (with gyromagnetic ratio $\gamma > 0$), and the second term on the right-hand side describing damping of the moment motion (with damping coefficient $\delta > 0$). Here, $\mathbf{h}_{\text{eff}} = -D\mathcal{E}(\mathbf{m})$ is the effective field, with Landau energy

$$(1.2) \quad \mathcal{E}(\mathbf{m}) = \alpha \int_{\Omega} |\nabla \mathbf{m}|^2 \, d\mathbf{x} + \int_{\Omega} \varphi(\mathbf{m}) \, d\mathbf{x} + \frac{\mu_0}{2} \int_{\mathbb{R}^n} |\nabla u|^2 \, d\mathbf{x} - \mu_0 \int_{\Omega} \langle \mathbf{h}_{\text{ext}}, \mathbf{m} \rangle \, d\mathbf{x},$$

which ensembles exchange, anisotropy, magnetostatic, and Zeeman's energy, respectively. Here, α denotes the exchange constant and μ_0 stands for the vacuum permeability. For uniaxial anisotropy with easy axis $\mathbf{e} \in \mathbb{S}^2$, it is common to choose $\varphi(\mathbf{m}) = K |\langle \mathbf{m}, \mathbf{e}_{\perp} \rangle|^2$, with $\tilde{K}(t, \mathbf{x}) \equiv K(\tau(t, \mathbf{x})) \geq 0$ for all $(t, \mathbf{x}) \in \Omega_T$ introduced before; see [19] for a recent survey.

Precessional motion of magnetization is based on quantum mechanics, where the mean value \mathbf{s}_{aver} of the spin operator $\mathbf{s} \in \mathbb{S}^2$ evolves according to Schrödinger's equation

$$\frac{d}{dt} \mathbf{s}_{\text{aver}} = -\frac{\gamma}{\mu_0} \mathbf{s}_{\text{aver}} \times \mathbf{b},$$

with magnetic induction $\mathbf{b} = \mu_0 \mathbf{h}_{\text{eff}}$; if the magnetization $\mathbf{m} : \Omega_T \rightarrow \mathbb{R}^3$ is understood to be the dipole of spins per unit volume, we arrive at

$$(1.3) \quad \mathbf{m}_t = -\gamma \mathbf{m} \times \mathbf{h}_{\text{eff}}.$$

Experimental evidence suggests an additional damping term to be responsible for alignment of magnetization with the applied field \mathbf{h}_{eff} , which is the reason for the additional damping term in (1.2) introduced by Landau and Lifshitz. An important feature of both (1.1) and (1.3) is conservation of length of initial magnetization; i.e., $\frac{1}{2} \frac{d}{dt} |\mathbf{m}|^2 = 0$ almost everywhere in Ω_T . Later on, Gilbert augmented \mathbf{h}_{eff} by an Ohmic dissipation term, by inserting the modified $\tilde{\mathbf{h}}_{\text{eff}} = \mathbf{h}_{\text{eff}} - \beta \mathbf{m}_t$ into (1.3). For different constants, this Landau–Lifshitz–Gilbert equation can again be restated in the form (1.1); see [1, 8, 9, 13, 15, 25] and [2, 3, 4, 5, 6, 7, 10, 12, 15, 18, 19, 22] for further details. Despite the model's success, the underlying physics of damping behavior is still not clear to rigorously justify the corresponding term in (LL); moreover, the saturation magnetization $\tilde{M}_s = \tilde{M}_s(\tau)$ varies for changing temperature, for which Landau proposes the following phenomenological power-law behavior for all $\tau < \tau_C$,

$$(1.4) \quad \tilde{M}_s(\tau) = \tilde{M}_0 \left(1 - \frac{\tau}{\tau_C}\right)^{\beta},$$

where the exponent $\beta > 0$ is found by experimental or theoretical evidence. This law agrees very well with experimental observations away from τ_C .

In this work, we propose and study an extended Landau–Lifshitz equation allowing for changes in the saturation magnetization. The subsequent modified Landau–Lifshitz model uses mutual orthogonality of \mathbf{m} , $\mathbf{m} \times \mathbf{h}_{\text{eff}}$, and $\mathbf{m} \times (\mathbf{m} \times \mathbf{h}_{\text{eff}})$, for every $(t, \mathbf{x}) \in \Omega_T$, to describe temperature-dependent gyroscopic precession: For a given $0 < M_s \in C^2(\Omega_T)$ solve (with $\mathbf{h}_{\text{eff}} = \Delta \mathbf{m}$)

$$(1.5) \quad \mathbf{m}_t = \kappa \mathbf{m} - \gamma \mathbf{m} \times \mathbf{h}_{\text{eff}} - \gamma \delta \frac{\mathbf{m}}{M_s} \times (\mathbf{m} \times \mathbf{h}_{\text{eff}}) \quad \text{in } \Omega_T,$$

$$(1.6) \quad \partial_{\mathbf{n}} \mathbf{m} = 0 \quad \text{on } \partial \Omega_T := \partial \Omega \times [0, T],$$

$$(1.7) \quad \mathbf{m}(0, \cdot) = \mathbf{m}_0 \quad \text{in } \Omega.$$

Let $\kappa(t, \mathbf{x}) \equiv \tilde{\kappa}(\tau(t, \mathbf{x}), \tau_t(t, \mathbf{x}))$, for all $(t, \mathbf{x}) \in \Omega_T$, which is chosen in such a way that $|\mathbf{m}(t, \mathbf{x})| = M_s(t, \mathbf{x})$ holds in Ω_T . We make the following assumptions: $\tilde{\kappa} \in C^2([0, \tau_C) \times \mathbb{R})$ satisfies

1. $\tilde{\kappa}(\tau, \tau_t) = 0$ for $\tau_t = 0$,
2. $\tilde{\kappa}(\tau, \tau_t) \leq 0$ for $\tau_t \geq 0$, and $\tilde{\kappa}(\tau, \tau_t) > 0$ for $\tau_t < 0$.

A scalar multiplication of (1.5) by \mathbf{m} yields to

$$(1.8) \quad \frac{d}{dt} M_s^2 = 2\kappa M_s^2 \quad \text{in } \Omega_T.$$

For a given temperature distribution $\tau : \Omega_T \rightarrow \mathbb{R}^+$, we obtain by the chain rule,

$$(1.9) \quad \tilde{\kappa}(\tau, \tau_t) = \frac{\tau_t}{2} \frac{d}{d\tau} \ln \left[\tilde{M}_s(\tau) \right]^2 = \tau_t \frac{d}{d\tau} \ln \tilde{M}_s(\tau),$$

which satisfies the above requirements (1) and (2). Shrinking, extension, and conservation of magnetization saturation by means of heating, cooling, as well as thermal equilibrium as further effects may then be described by (1.9), once $M_s \in C^2([0, \tau_C])$ is provided, which is either by experiment or theory (e.g., mean field theory). For $0 \leq t_1 \leq t_2 \leq T$ we compute

$$(1.10) \quad \tilde{M}_s^2(\tau(t_2, \cdot)) = \tilde{M}_s^2(\tau(t_1, \cdot)) \exp \left(-2 \int_{t_1}^{t_2} \tilde{\kappa}(\tau(s, \cdot), \tau_t(s, \cdot)) \, ds \right) \quad \text{in } \Omega.$$

In passing, a possible explicit dependence of $\tilde{\kappa}$ on M_s may also be assumed, which seems advantageous in a neighborhood of τ_C . Moreover, if we adopt the phenomenological power-law (1.4), then the relation (1.9) will take the form

$$(1.11) \quad \tilde{\kappa}(\tau, \tau_t) = -\frac{\tau_t \beta}{\tau_C - \tau}.$$

In the literature, to verify solvability of (1.1) is often based on the reformulation in Landau–Lifshitz–Gilbert form, which exploits the property $|\mathbf{m}| = \text{const}$ almost everywhere in Ω_T . In the present case, the target for solutions of (1.5)–(1.6) depends on $(t, \mathbf{x}) \in \Omega_T$, and the authors are not aware of any analytical studies where the target of solutions is allowed to (smoothly) vary in space-time. Our first contribution in this work is to verify the existence of locally strong (section 3) and globally weak (section 5) solutions, for $\Omega \subset \mathbb{R}^d$, $d = 2, 3$ a bounded Lipschitz domain. While the first result uses abstract results, where locally strong solutions are constructed as proper

limits of a sequence of smooth solutions, whose existence follows from general inverse function theorem (here, we follow [26, 27], and [18], where (LL) with $M_s \equiv \text{const}$ is studied), weak solutions are constructed as proper limits of iterates of a practical discretization in space-time (see Scheme B).

From a numerical viewpoint, to construct convergent, fully practical numerical schemes (including discretization in time and space, numerical integration, and a simple fixed point scheme to solve arising nonlinear algebraic problems), where iterates respect the constraint $|\mathbf{m}| = M_s > 0$ in a proper sense is a nontrivial endeavor. Over the last decade, projection strategies have been shown to converge in the context of (locally existing) strong solutions for (LL), where $M_s \equiv \text{const}$, and optimal convergence rates have been verified in this case [22]. Unfortunately, convergence of these methods in the case of only weak solutions is still not clear; it is only recently that space-time discretizations of (LL) and for $M_s = \text{const}$ were found, where iterates satisfy $|\mathbf{M}^j| = \text{const}$ at all nodes of the triangulation ($0 \leq j \leq J$), and construct weak solutions in the limit when all discretization parameters tend to zero. Interestingly, both ansatzes use different formulations of the problem in the continuous setting, leading to different schemes, numerical analysis, and properties as indicated.

The present case of $0 < M_s \equiv M_s(t, \mathbf{x})$ makes the construction of stable, convergent discretizations, which satisfy $|\mathbf{M}^j| = M_s > 0$ even only at mesh-points challenging ($0 \leq j \leq J$). More specifically, we study two discretizations of (1.5)–(1.7), which are also based on two formulations (1.5)–(1.7), and (5.1), (1.6)–(1.7) of the problem, which are equivalent for strong solutions:

- Scheme A: At the end of each iterative step, vectors are projected to the sphere of a given radius $M_s = M_s(t_j, \mathbf{z})$, for every $1 \leq j \leq J$, and every node of the mesh $\mathbf{z} \in \mathcal{N}_h$. The scheme is based on a reformulation of (1.5) in the form (3.1). Iterates converge to (locally existing) strong solutions of (1.5)–(1.7) at optimal rates.
- Scheme B: The scheme is a discretization in space-time of (5.1), (1.6)–(1.7), which uses trapezoidal rule, reduced integration (in space), and discrete Laplacian. Its motivation comes from (LL) with $M_s = \text{const}$, where conservation of the “sphere-constraint” at mesh-points $\mathbf{z} \in \mathcal{N}_h$, and convergence to weak solutions is known [6, 7]. As will be shown in section 5, $M_s = M_s(t_j, \mathbf{z})$ at nodes $\mathbf{z} \in \mathcal{N}_h$ holds only approximately, but is attained in the limit of vanishing discretization parameters. Iterates are shown to converge to weak solutions of (the reformulation) (5.1), (1.6)–(1.7).

Sections 3 and 4 briefly recall arguments to verify corresponding results on locally existing strong solutions for $M_s = \text{const}$ from [18, 10], and on optimal rates of convergence for iterates of a projection method (Scheme A) [22, 12]. Main results are stated in section 5, where weak solutions to (5.1), (1.6)–(1.7) are constructed as proper limits of iterates of Scheme B. The paper closes with section 7, where 2D computational academic and applied examples compare Schemes A and B, and evidence interesting dynamics for the new model of targets varying in space-time for (LL). Benchmark problems from [29] are adopted and studied for the case of smoothly varying spheres in space-time, and comparative studies using Schemes A and B are discussed.

2. Preliminaries. Throughout this paper we assume that \mathcal{T}_h is a quasiuniform triangulation of the polygonal or polyhedral bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$ into triangles or tetrahedra of maximal diameter h for $d = 2$ or $d = 3$, respectively. We let $V_h \subset W^{1,2}(\Omega)$ denote the lowest order finite element space on \mathcal{T}_h ; i.e., $\phi_h \in V_h$ if and only if $\phi_h \in C(\bar{\Omega})$ and $\phi_h|_K$ is affine for each $K \in \mathcal{T}_h$. Given the set of all nodes

(or vertexes) \mathcal{N}_h in \mathcal{T}_h and letting $\{\varphi_{\mathbf{z}} : \mathbf{z} \in \mathcal{N}_h\}$ denote the nodal basis in V_h , we define the nodal interpolation operator $\mathcal{I}_h : C(\overline{\Omega}) \rightarrow V_h$ by $\mathcal{I}_h \psi := \sum_{\mathbf{z} \in \mathcal{N}_h} \psi(\mathbf{z}) \varphi_{\mathbf{z}}$, for $\psi \in C(\overline{\Omega})$, and $P_h : L^2(\Omega) \rightarrow V_h$ the $L^2(\Omega)$ -projection. We use boldface notation to indicate vectorial quantities, like $\mathbf{V}_h = [V_h]^3$, for example. We write $(\mathbf{f}, \mathbf{g}) = \int_{\Omega} \langle \mathbf{f}, \mathbf{g} \rangle \, d\mathbf{x}$ for $\mathbf{f}, \mathbf{g} \in L^2(\Omega, \mathbb{R}^3)$ and abbreviate $\|\mathbf{f}\| = \|\mathbf{f}\|_{L^2(\Omega)}$. For functions $\mathbf{f}, \mathbf{g} \in C(\overline{\Omega}, \mathbb{R}^3)$ a discrete inner product is defined by

$$(2.1) \quad (\mathbf{f}, \mathbf{g})_h := \int_{\Omega} \mathcal{I}_h[\langle \mathbf{f}, \mathbf{g} \rangle] \, d\mathbf{x} = \sum_{\mathbf{z} \in \mathcal{N}_h} \beta_{\mathbf{z}} \langle \mathbf{f}(\mathbf{z}), \mathbf{g}(\mathbf{z}) \rangle,$$

where $\beta_{\mathbf{z}} = \int_{\Omega} \varphi_{\mathbf{z}} \, d\mathbf{x}$ for all $\mathbf{z} \in \mathcal{N}_h$; we define $\|\boldsymbol{\psi}\|_h^2 := (\boldsymbol{\psi}, \boldsymbol{\psi})_h$. We remark that there holds

$$(2.2) \quad \|\boldsymbol{\psi}\| \leq \|\boldsymbol{\psi}\|_h \leq (d+2)^{1/2} \|\boldsymbol{\psi}\| \quad \forall \boldsymbol{\psi} \in \mathbf{V}_h.$$

Basic interpolation estimates yield (cf., e.g., [7]) that

$$(2.3) \quad |(\boldsymbol{\phi}, \boldsymbol{\psi})_h - (\boldsymbol{\phi}, \boldsymbol{\psi})| \leq Ch \|\boldsymbol{\phi}\| \|\nabla \boldsymbol{\psi}\| \quad \forall \boldsymbol{\phi}, \boldsymbol{\psi} \in \mathbf{V}_h,$$

where here and throughout this paper $C > 0$ denotes a (h, k) -independent generic constant, which may change from place to place. We define a discrete Laplace operator $\tilde{\Delta}_h : W^{1,2}(\Omega, \mathbb{R}^3) \rightarrow \mathbf{V}_h$ by requiring

$$\left(-\tilde{\Delta}_h \boldsymbol{\phi}, \boldsymbol{\chi}\right)_h = (\nabla \boldsymbol{\phi}, \nabla \boldsymbol{\chi}) \quad \forall \boldsymbol{\chi} \in \mathbf{V}_h.$$

We note that there exists a constant $c_1 > 0$ such that for all $\boldsymbol{\phi} \in \mathbf{V}_h$ there holds (cf. [6])

$$(2.4) \quad \|\tilde{\Delta}_h \boldsymbol{\phi}\|_h \leq c_1 h^{-2} \|\boldsymbol{\phi}\|_h \quad \text{and} \quad \|\tilde{\Delta}_h \boldsymbol{\phi}\|_{L^\infty} \leq c_1 h^{-2} \|\boldsymbol{\phi}\|_{L^\infty}.$$

Given a time-step size $k > 0$ and a sequence $\{\varphi^j\}_{j \geq 0}$ in some vector space X , we set

$$d_t \varphi^{j+1} := k^{-1} (\varphi^{j+1} - \varphi^j) \quad \text{and} \quad \varphi^{j+1/2} := \frac{\varphi^j + \varphi^{j+1}}{2},$$

for $j \geq 0$. Note that there holds $\langle d_t \varphi^{j+1}, \varphi^{j+1/2} \rangle_X = \frac{1}{2} d_t \|\varphi^{j+1}\|_X^2$, if X is a Hilbert space. Finally, we recall some elementary properties of vector products: Let $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4 \in \mathbb{R}^3$ there holds

$$(2.5) \quad \begin{aligned} \langle \mathbf{a}_1 \times \mathbf{a}_2, \mathbf{a}_3 \rangle &= -\langle \mathbf{a}_2, \mathbf{a}_1 \times \mathbf{a}_3 \rangle, \\ \mathbf{a}_1 \times (\mathbf{a}_2 \times \mathbf{a}_3) &= \langle \mathbf{a}_1, \mathbf{a}_3 \rangle \mathbf{a}_2 - \langle \mathbf{a}_1, \mathbf{a}_2 \rangle \mathbf{a}_3, \\ \langle \mathbf{a}_1 \times \mathbf{a}_2, \mathbf{a}_3 \times \mathbf{a}_4 \rangle &= \langle \mathbf{a}_1, \mathbf{a}_3 \rangle \langle \mathbf{a}_2, \mathbf{a}_4 \rangle - \langle \mathbf{a}_2, \mathbf{a}_3 \rangle \langle \mathbf{a}_1, \mathbf{a}_4 \rangle. \end{aligned}$$

3. Local existence of strong solutions for (1.5)–(1.7). We start with a reformulation of (1.5)–(1.7). For a given $0 < M_s \in L^\infty(0, T; W^{2,2}(\Omega))$, solve

$$(3.1) \quad \mathbf{m}_t - \gamma \delta M_s \Delta \mathbf{m} - \kappa \mathbf{m} = \frac{\gamma \delta}{M_s} |\nabla \mathbf{m}|^2 \mathbf{m} - \frac{\gamma \delta}{2M_s} (\Delta M_s^2) \mathbf{m} - \gamma \mathbf{m} \times \Delta \mathbf{m},$$

together with (1.6), (1.7), where $|\mathbf{m}_0| = M_s(0, \cdot)$. In the following, for given $T > 0$ we call $\mathbf{m} \in C([0, T]; W^{1,2}(\Omega, \mathbb{R}^3)) \cap L^2(0, T; W^{2,2}(\Omega, \mathbb{R}^3))$ that solves (1.5) in

distributional sense, and (1.6)–(1.7), a strong solution to (1.5)–(1.7). We use the same notion for (3.1), (1.6)–(1.7) below.

LEMMA 3.1. *Let $0 < M_s \in L^\infty(0, T; W^{2,2}(\Omega))$ be given, with $\partial_n M_s = 0$ on $\partial\Omega_T$. The function $\mathbf{m} : \Omega_T \rightarrow \mathbb{R}^3$ is a strong solution of (1.5)–(1.7) if and only if it is a strong solution to (3.1), (1.6)–(1.7).*

Proof. We use the vector cross product formula (2.5)₂ and (1.5), and $|\mathbf{m}|^2 = M_s^2$ in Ω_T to obtain

$$\begin{aligned} \mathbf{m} \times (\mathbf{m} \times \Delta \mathbf{m}) &= \langle \mathbf{m}, \Delta \mathbf{m} \rangle \mathbf{m} - |\mathbf{m}|^2 \Delta \mathbf{m} \\ (3.2) \qquad \qquad \qquad &= -|\nabla \mathbf{m}|^2 \mathbf{m} + \frac{1}{2} (\Delta M_s^2) \mathbf{m} - M_s^2 \Delta \mathbf{m}. \end{aligned}$$

Inserting this identity into (1.5) proves that a strong solution $\mathbf{m} : \Omega_T \rightarrow \mathbb{R}^3$ of (1.5)–(1.7) solves (3.1) in a distributional sense, satisfying (1.6)–(1.7).

Let now $\mathbf{m} : \Omega_T \rightarrow \mathbb{R}^3$ be a strong solution to (3.1), (1.6)–(1.7). Setting $W(t, \mathbf{x}) = |\mathbf{m}(t, \mathbf{x})|^2$, and observing

$$W_t = 2\langle \mathbf{m}, \mathbf{m}_t \rangle, \quad \nabla W = 2\langle \mathbf{m}, \nabla \mathbf{m} \rangle, \quad \Delta W = 2\langle \mathbf{m}, \Delta \mathbf{m} \rangle + 2|\nabla \mathbf{m}|^2,$$

implies a.e. in Ω_T ,

$$W_t - \gamma \delta M_s \Delta W - 2\kappa W = \frac{2\gamma \delta}{M_s} |\nabla \mathbf{m}|^2 (W - M_s^2) - \frac{\gamma \delta}{M_s} (\Delta M_s^2) W.$$

Setting $Z = W - M_s^2$ leads to

$$Z_t - \gamma \delta M_s \Delta Z - 2\kappa Z = (2\kappa M_s^2 - \partial_t M_s^2) + \frac{\gamma \delta}{M_s} [2|\nabla \mathbf{m}|^2 - (\Delta M_s^2)] Z.$$

The first bracket on the right-hand side vanishes, owing to (1.8). Multiplying with Z in L^2 sense, and observing that $\partial_n Z = 0$ on $\partial\Omega_T$ gives

$$\left(\frac{Z_t}{M_s}, Z \right) + \gamma \delta \|\nabla Z\|^2 = \left(\frac{2\kappa}{M_s} + \frac{2\gamma \delta}{M_s^2} |\nabla \mathbf{m}|^2 - \frac{\gamma \delta}{M_s^2} (\Delta M_s^2), Z^2 \right).$$

Owing to $c_1 > M_s > 0$, rearranging terms then leads to

$$\frac{d}{dt} \|Z\|^2 + \|\nabla Z\|^2 \leq \tilde{C} \|Z\|^2,$$

for some finite $0 \leq \tilde{C} \equiv \tilde{C}(|\nabla \mathbf{m}|, |(\Delta M_s^2)|)$. Integration in time, observing $Z(0, \cdot) = 0$, and Gronwall’s inequality then shows that $Z \equiv 0$ in Ω_T . \square

In order to verify existence of strong solutions to (3.1), (1.6)–(1.7), we either suppose smallness of initial energies $E(\mathbf{m}_0) \equiv \frac{1}{2} \|\nabla \mathbf{m}_0\|^2$ to conclude global existence of strong solutions, or local existence for bounded initial energies. The following arguments are taken from [18, 26], and [27, section 1.4], and are mostly sketched, in case their elaboration can be taken from these references on a line-by-line basis. First, we verify a priori bounds, under the assumption that smooth solutions $\mathbf{m} : \Omega_T \rightarrow \mathbb{R}^3$ of (3.1) exist.

LEMMA 3.2. *Let $T > 0$. There exists a constant $0 < C = C(M_s, \gamma, \delta) < \infty$ such that for any smooth solution to (3.1), (1.6)–(1.7) there holds for all $0 \leq t \leq T$,*

$$\frac{\delta^2}{2(1 + \delta^2)} \int_0^t \|\mathbf{m}_t(s, \cdot)\|^2 ds + \frac{\gamma \delta}{2} E(\mathbf{m}(t, \cdot)) \leq C(1 + E(\mathbf{m}_0)).$$

Proof. Multiply (3.1) by \mathbf{m}_t and integrate over $[0, T] \times \Omega$. We compute

$$\begin{aligned}
 -(M_s \Delta \mathbf{m}, \mathbf{m}_t) &= \frac{1}{2} \frac{d}{dt} (M_s, |\nabla \mathbf{m}|^2) - \frac{1}{2} (\partial_t M_s, |\nabla \mathbf{m}|^2) + (\nabla M_s^\top \nabla \mathbf{m}, \mathbf{m}_t), \\
 2(\kappa \mathbf{m}, \mathbf{m}_t) &= \frac{d}{dt} (\kappa, M_s^2) - (\partial_t \kappa, M_s^2), \\
 (3.3) \quad \left(\frac{|\nabla \mathbf{m}|^2}{2M_s}, \partial_t M_s^2 \right) &= (|\nabla \mathbf{m}|^2, \partial_t M_s), \\
 \left(\frac{\Delta M_s^2}{4M_s}, \partial_t M_s^2 \right) &= \frac{1}{2} (\Delta M_s^2, \partial_t M_s).
 \end{aligned}$$

Next, we bound $|(\mathbf{m} \times \Delta \mathbf{m}, \mathbf{m}_t)|$: take the cross product of (3.1) with \mathbf{m} , and use (3.2) to obtain

$$\begin{aligned}
 \mathbf{m} \times \mathbf{m}_t &= \gamma \delta M_s \mathbf{m} \times \Delta \mathbf{m} - \gamma \mathbf{m} \times (\mathbf{m} \times \Delta \mathbf{m}) \\
 (3.4) \quad &= \gamma \delta M_s \mathbf{m} \times \Delta \mathbf{m} + \gamma M_s^2 \Delta \mathbf{m} - \frac{\gamma}{2} (\Delta M_s^2) \mathbf{m} + \gamma |\nabla \mathbf{m}|^2 \mathbf{m} \\
 &= \gamma M_s \left(\delta + \frac{1}{\delta} \right) \mathbf{m} \times \Delta \mathbf{m} - \frac{M_s}{\delta} \kappa \mathbf{m} + \frac{M_s}{\delta} \mathbf{m}_t.
 \end{aligned}$$

Multiplying this equation with $\frac{\delta}{M_s} \mathbf{m}_t$ leads to

$$0 = \gamma (\delta^2 + 1) (\mathbf{m} \times \Delta \mathbf{m}, \mathbf{m}_t) - \frac{1}{2} (M_s \kappa, \partial_t M_s) + \|\mathbf{m}_t\|^2,$$

and hence

$$\gamma \int_0^T |(\mathbf{m} \times \Delta \mathbf{m}, \mathbf{m}_t)| \, ds \leq C + \frac{1}{1 + \delta^2} \|\mathbf{m}_t\|^2.$$

Putting things together, we arrive at

$$\begin{aligned}
 \frac{1}{2} \left(1 - \frac{1}{1 + \delta^2} \right) \int_0^T \|\mathbf{m}_t(s, \cdot)\|^2 \, ds &+ \frac{\gamma \delta}{2} (M_s(T, \cdot) |\nabla \mathbf{m}(T, \cdot)|^2) \\
 &\leq \frac{\gamma \delta}{2} (M_s(0, \cdot) |\nabla \mathbf{m}(0, \cdot)|^2) + \tilde{C}_1 \int_0^T [|\nabla \mathbf{m}(s, \cdot)|^2] \, ds + \tilde{C}_2,
 \end{aligned}$$

for some positive, bounded

$$\tilde{C}_1 \equiv \tilde{C}_2 (\|M_s\|_{W^{1,\infty}(\Omega_T)}, \delta), \quad \tilde{C}_2 \equiv \tilde{C}_2 (\|M_s\|_{W^{2,\infty}(\Omega_T)}).$$

The assertion of the lemma then follows from applying Gronwall’s inequality. \square

In the sequel, we limit ourselves to $\Omega \subset \mathbb{R}^2$. Let $E(\mathbf{m}; \omega) = \int_\omega |\nabla \mathbf{m}|^2 \, d\mathbf{y}$, for $\omega \subset \Omega$, and $B_R(\mathbf{x}) \subset \mathbb{R}^2$ be a ball around $\mathbf{x} \in \mathbb{R}^2$ of radius $R > 0$. The following technical lemma holds for functions from $L^\infty(0, T; W^{1,2}(\Omega, \mathbb{R}^3)) \cap L^2(0, T; W^{2,2}(\Omega, \mathbb{R}^3))$, and is taken from [27, p. 274]; see also [18, Lemma 3.1].

LEMMA 3.3. *Let $T > 0$. There exist bounded, positive constants $C, R_0 > 0$ such that for any $\varphi \in L^\infty(0, T; W^{1,2}(\Omega, \mathbb{R}^3)) \cap L^2(0, T; W^{2,2}(\Omega, \mathbb{R}^3))$, and any $R \in (0, R_0]$ there holds*

$$\begin{aligned}
 &\int_0^T \|\nabla \varphi(s, \cdot)\|_{L^4}^4 \, ds \\
 &\leq C \left(\operatorname{ess\,sup}_{(t,\mathbf{x}) \in \Omega_T} \int_{B_R(\mathbf{x})} |\nabla \varphi(t, \cdot)|^2 \, d\mathbf{x} \right) \left(\int_0^T \|\nabla^2 \varphi\|^2 \, dt + R^{-2} \int_0^T \|\nabla \varphi(s, \cdot)\|^2 \, ds \right).
 \end{aligned}$$

Like in [26, Lemma 1.5], local energies at a fixed time can be controlled in terms of those at earlier times. The proof on the subsequent lemma uses Lemma 3.2 and localizes the argumentation around (3.4). A proof for (3.1) for $M_s = \text{const}$ is given in [18, Lemma 3.4], and can easily be adopted to the present case.

LEMMA 3.4. *Let $T > 0$. There exists a constant $C \equiv C(M_s, \gamma, \delta) > 0$ such that for any smooth solution of (3.1), any $R \in (0, R_0]$, and any $(t, \mathbf{x}) \in \Omega_T$ there holds the estimate*

$$E(\mathbf{m}(t, \cdot); B_R(\mathbf{x}_0)) \leq E(\mathbf{m}_0; B_{2R}(\mathbf{x}_0)) + C \frac{t}{R^2} (1 + E(\mathbf{m}_0)).$$

We can now follow the general argumentation from [26, p. 274], which we include for the convenience of the reader: For $\mathbf{m}_0 \in W^{1,2}(\Omega, \mathbb{R}^3)$ and any given $\varepsilon_0 > 0$, let $R_0 > 0$ be the maximal number such that

$$\sup_{\mathbf{x}_0 \in \Omega} E(\mathbf{m}_0; B_{2R_0}(\mathbf{x}_0)) < \varepsilon_0,$$

and let $T_0 > 0$ be the number such that any smooth solution $\mathbf{m} : \Omega_T \rightarrow \mathbb{R}^3$ of (3.1) taking initial value $\mathbf{m}_0 : \Omega \rightarrow \mathbb{R}^3$ satisfies

$$\sup_{\mathbf{x}_0 \in \Omega} \sup_{0 \leq t \leq T_0} E(\mathbf{m}(t, \cdot); B_{R_0}(\mathbf{x}_0)) < 2\varepsilon_0.$$

Note that in view of Lemma 3.4, we may let $T_0 = \frac{\varepsilon_0 R_0^2}{CE(\mathbf{m}_0)}$. Let $\{\phi_i\}$ be smooth cut-off functions subordinate to a cover of $\Omega \subset \mathbb{R}^2$ by balls $B_{2R_0}(\mathbf{x}_i)$ with finite overlap, and such that $0 \leq \phi_i \leq 1$, and $|\nabla \phi_i| \leq \frac{2}{R_1}$, such that $\sum_i \phi_i^2 = 1$. We may then compute

$$\begin{aligned} \|\nabla \mathbf{m}(t, \cdot)\|_{L^4}^4 &= \sum_i \int_{\Omega} |\nabla \mathbf{m}(t, \cdot)|^4 \phi_i^2 \, d\mathbf{x} \\ (3.5) \quad &\leq C \sup_i E(\mathbf{m}(t, \cdot); B_{2R_0}(\mathbf{x}_i)) \left(\int_{\Omega} |\nabla^2 \mathbf{m}(t, \cdot)|^2 \, d\mathbf{x} + R_0^{-2} E(\mathbf{m}_0) \right) \\ &\leq C\varepsilon_0 \left(\int_{\Omega} |\nabla^2 \mathbf{m}(t, \cdot)|^2 \, d\mathbf{x} + R_0^{-2} E(\mathbf{m}_0) \right). \end{aligned}$$

We are now in a position to verify the following bound.

LEMMA 3.5. *For sufficiently small $\varepsilon_0 > 0$, and any $R \in (0, R_0]$, there exists $T_0 > 0$ such that an existing smooth solution of (3.1) satisfies*

$$\int_0^t \|\nabla^2 \mathbf{m}(s, \cdot)\|^2 \, ds \leq C (E(\mathbf{m}_0) + 1) \left(1 + \frac{t}{R^2}\right) \quad \forall 0 \leq t \leq T_0.$$

Proof. Multiply (3.1) with $-\Delta \mathbf{m}$ to find

$$\frac{1}{2} \frac{d}{dt} \|\nabla \mathbf{m}\|^2 + \frac{\gamma \delta}{2} \|\sqrt{M_s} \Delta \mathbf{m}\|^2 \leq C (1 + \|\nabla \mathbf{m}\|^2 + |(\nabla \kappa \otimes \mathbf{m}, \nabla \mathbf{m})| + \|\nabla \mathbf{m}\|_{L^4}^4).$$

For sufficiently small $\varepsilon_0 > 0$, by (3.5) we can control the last term on the right-hand side by the second one on the left-hand side. Integration in time and Lemma 3.2 then yields the assertion. \square

Summing up, for sufficiently small $\varepsilon_0 > 0$ we have locally in time bounds for $\mathbf{m} : \Omega_{T_0} \rightarrow \mathbb{R}^3$ in

$$V(\Omega_{T_0}; \mathbb{R}^3) := \left\{ \varphi : \Omega_{T_0} \rightarrow \mathbb{R}^3 : \text{ess sup}_{0 \leq t \leq T_0} E(\varphi) + \int_0^{T_0} \|\nabla^2 \varphi(s, \cdot)\|^2 + \|\varphi_t(s, \cdot)\|^2 ds < \infty \right\}.$$

Then it is not difficult to verify uniqueness of smooth solutions $\mathbf{m} : \Omega_T \rightarrow \mathbb{R}^3$ to (3.1), (1.6)–(1.7) from these bounds; see [18, Theorem 3.9] for the simplified case $M_s \equiv 1$.

In order to prove the local existence in time of smooth solutions to (3.1), (1.6)–(1.7), we may then exploit strong parabolicity of our problem and follow the arguments given in [18, Theorem 3.12 and Lemma 3.10] for $M_s \equiv \text{const}$. Let us note that all results from [18] are valid for an equation of the type

$$\partial_t \mathbf{u} + \mathbf{u} \times \Delta \mathbf{u} + \mathbf{u} \times (\mathbf{u} \times \Delta \mathbf{u}) = \mathbf{0}.$$

Our equation (3.1) represents a slight modification of this by adding an additional term of the type \mathbf{u} , due to variable M_s . This can be handled in an easier way than other terms (note that other terms contain derivatives) and therefore we skip unnecessary details. According to [18, Theorem 3.13], we may say that:

THEOREM 3.1. *Let $\mathbf{m}_0 : \Omega \rightarrow \mathbb{R}^3$ be a smooth map and $0 < M_s \in L^\infty(0, T; W^{2,2}(\Omega))$. There exists $T_0 > 0$, and a unique, smooth mapping $\mathbf{m} : \Omega_{T_0} \rightarrow \mathbb{R}^3$ solving (3.1), (1.6)–(1.7). Let $\ell \in \mathbb{N}$, and $\mathbf{m}_{0,\ell} \in C^\infty(\Omega, \mathbb{R}^3)$ be such that*

$$\sup_{(t, \mathbf{x}_0) \in \Omega_{T_0}} E(\mathbf{m}_\ell(t, \cdot); B_R(\mathbf{x}_0)) \leq \varepsilon_0 \quad \forall R \in (0, R_0],$$

for any $\varepsilon_0 > 0$, with $\mathbf{m}_{0\ell} \rightarrow \mathbf{m}_0$ in $W^{1,2}(\Omega, \mathbb{R}^3)$ for $\ell \rightarrow \infty$. For every $\ell \in \mathbb{N}$, let $\mathbf{m}_\ell : \Omega_{T_0} \rightarrow \mathbb{R}^3$ be the smooth, unique local solution to (3.1), (1.6) starting from $\mathbf{m}_{0\ell}$. Similar to [18, Lemma 3.7], a consequence from Lemmas 3.2 and 3.5, and the Aubin–Lions lemma, there holds for $\ell \rightarrow \infty$,

$$(3.6) \quad \begin{aligned} \partial_t \mathbf{m}_\ell &\rightharpoonup \partial_t \mathbf{m} && \text{in } L^2(0, T_0; L^2(\Omega, \mathbb{R}^3)), \\ \Delta \mathbf{m}_\ell &\rightharpoonup \Delta \mathbf{m} && \text{in } L^2(0, T_0; L^2(\Omega, \mathbb{R}^3)), \\ \mathbf{m}_\ell &\rightarrow \mathbf{m} && \text{in } L^2(0, T_0; W^{1,2}(\Omega, \mathbb{R}^3)) \cap C([0, T_0]; L^2(\Omega, \mathbb{R}^3)). \end{aligned}$$

Take any smooth vector field ϕ and any $t \in (0, T_0)$. The solution \mathbf{m}_ℓ to (3.1) satisfies

$$\begin{aligned} &(\mathbf{m}_\ell(t) - \mathbf{m}_{0\ell}, \phi) - \gamma\delta \int_0^t (M_s \Delta \mathbf{m}_\ell, \phi) - \kappa \int_0^t (\mathbf{m}_\ell, \phi) \\ &= \int_0^t \left(\frac{\gamma\delta}{M_s} |\nabla \mathbf{m}_\ell|^2 \mathbf{m}_\ell, \phi \right) - \int_0^t \left(\frac{\gamma\delta}{2M_s} (\Delta M_s^2) \mathbf{m}_\ell, \phi \right) - \gamma \int_0^t (\mathbf{m}_\ell \times \Delta \mathbf{m}_\ell, \phi). \end{aligned}$$

Using (3.6) and recalling that a product of a weak and a strong convergent sequence is strongly convergent, we can pass to the limit for $\ell \rightarrow \infty$ and we get

$$\begin{aligned} &(\mathbf{m}(t) - \mathbf{m}(0), \phi) - \gamma\delta \int_0^t (M_s \Delta \mathbf{m}, \phi) - \kappa \int_0^t (\mathbf{m}, \phi) \\ &= \int_0^t \left(\frac{\gamma\delta}{M_s} |\nabla \mathbf{m}|^2 \mathbf{m}, \phi \right) - \int_0^t \left(\frac{\gamma\delta}{2M_s} (\Delta M_s^2) \mathbf{m}, \phi \right) - \gamma \int_0^t (\mathbf{m} \times \Delta \mathbf{m}, \phi). \end{aligned}$$

Differentiating this with respect to the time variable we see that $\mathbf{m} \in V(\Omega_{T_0}, \mathbb{R}^3)$ may be identified to be a strong solution of (3.1), (1.6)–(1.7) for $t \in (0, T_0)$. Hence we verified:

THEOREM 3.2. *Let $\mathbf{m}_0 \in W^{1,2}(\Omega, \mathbb{R}^3)$, $0 < M_s \in L^\infty(0, T; W^{2,2}(\Omega))$ be given. Then there exists a unique, local in time strong solution to (3.1), (1.6)–(1.7), which—by Lemma 3.1—is also a local in time strong solution to (1.5)–(1.7).*

According to the imbedding $W^{2,2}(\Omega) \subset C(\bar{\Omega})$ and the assumption $0 < M_s \in L^\infty(0, T; W^{2,2}(\Omega))$, we see that M_s always remains strictly positive. There arises a natural question: what will happen if $M_s \rightarrow 0$? The constants obtained from a priori estimates depend on M_s in such a way that they blow up if $M_s \rightarrow 0$. Therefore we do not allow M_s to vanish.

4. Convergence with optimal rates for Scheme A. A space-time discretization of (3.1) may produce iterates $\{\mathbf{M}^j\} \subset \mathbf{V}_h$, with $|\mathbf{M}^j(\mathbf{z})| \neq M_s(t_j, \mathbf{z})$ for $1 \leq j \leq J$ and $\mathbf{z} \in \mathcal{N}_h$. This is the motivation for the following projection scheme, which adopts the corresponding one for $M_s = \text{const}$ in [22, p. 139] to the present case.

Scheme A: Let $\mathbf{M}^0 \equiv \tilde{\mathbf{M}}^0 \in \mathbf{V}_h$. 1. For $j = 1, 2, \dots, J - 1$ find $\mathbf{M}^j \in \mathbf{V}_h$ such that for all $\Phi \in \mathbf{V}_h$ there holds

$$\begin{aligned} & \frac{1}{k} \left(\tilde{\mathbf{M}}^j - \mathbf{M}^{j-1}, \Phi \right)_h + \gamma \delta \left(M_s(t_j, \cdot) \nabla \tilde{\mathbf{M}}^j, \nabla \Phi \right) = \gamma \delta \left(\nabla \tilde{\mathbf{M}}^j, \nabla M_s(t_j, \cdot) \otimes \Phi \right) \\ & + \left(\left[\frac{\gamma \delta}{M_s(t_j, \cdot)} \left(|\nabla \tilde{\mathbf{M}}^{j-1}|^2 - \frac{1}{2} \Delta M_s^2(t_j, \cdot) \right) + \kappa(t_j, \cdot) \right] \tilde{\mathbf{M}}^j, \Phi \right) \\ & + \gamma \left(\tilde{\mathbf{M}}^{j-1} \times \nabla \tilde{\mathbf{M}}^j, \nabla \Phi \right). \end{aligned}$$

2. Compute $\mathbf{M}^j \in \mathbf{V}_h$ according to

$$\mathbf{M}^j(\mathbf{z}) = M_s(t_j, \mathbf{z}) \frac{\tilde{\mathbf{M}}^j(\mathbf{z})}{|\tilde{\mathbf{M}}^j(\mathbf{z})|} \quad \forall \mathbf{z} \in \mathcal{N}_h.$$

If we combine both steps, we obtain the following equation for iterates $\tilde{\mathbf{M}}^j : \Omega \rightarrow \mathbb{R}^3$,

$$\begin{aligned} & (d_t \tilde{\mathbf{M}}^j, \Phi)_h + \gamma \delta \left(M_s(t_j, \cdot) \nabla \tilde{\mathbf{M}}^j, \nabla \Phi \right) = \gamma \delta \left(\nabla \tilde{\mathbf{M}}^j, \nabla M_s(t_j, \cdot) \otimes \Phi \right) \\ (4.1) \quad & - \frac{1}{k} \left(\left[1 - \frac{M_s(t_{j-1}, \cdot)}{|\tilde{\mathbf{M}}^{j-1}|} \right] \tilde{\mathbf{M}}^{j-1}, \Phi \right)_h + \left(\left[\frac{\gamma \delta}{M_s(t_j, \cdot)} \left(|\nabla \tilde{\mathbf{M}}^{j-1}|^2 \right. \right. \right. \\ & \left. \left. \left. - \frac{1}{2} \Delta M_s^2(t_j, \cdot) \right) + \kappa(t_j, \cdot) \right] \tilde{\mathbf{M}}^j, \Phi \right) + \gamma \left(\tilde{\mathbf{M}}^{j-1} \times \nabla \tilde{\mathbf{M}}^j, \nabla \Phi \right) \quad \forall \Phi \in \mathbf{V}_h. \end{aligned}$$

As is now evident from the leading term in the second line, the projection method may be regarded as a semi-explicit penalization method. The following result asserts optimal convergence rates for solutions of Scheme A. Its proof uses an inductive argument, which has been developed in [22, Chapter 4] ($d = 2$), and used in [12] ($d = 3$) for the case $M_s = \text{const}$; we skip a proof of the following.

THEOREM 4.1. *Let $t_J \leq T_0$, where $[0, T_0)$ is the interval, where a strong solution to (3.1), (1.6)–(1.7) exists. For $k \leq k_0(T_0, M_s, \gamma \delta)$ sufficiently small, solutions $\{\tilde{\mathbf{M}}^j\}_{j=1}^J$ of Scheme A exist. Suppose that $\|\mathbf{M}^0 - \mathbf{m}_0\|_{W^{1,2}} \leq C(k + h)$. There exists*

a constant $C > 0$ independent from $k, h \geq 0$, such that

$$(4.2) \quad \max_{1 \leq j \leq J} \|\mathbf{M}(t_j, \cdot) - \tilde{\mathbf{M}}^j\|_{L^2} + \left(k \sum_{j=1}^J \|\mathbf{M}(t_j, \cdot) - \tilde{\mathbf{M}}^j\|_{W^{1,2}}^2 \right)^{1/2} \leq C(k + h).$$

Moreover, there holds

$$(4.3) \quad \max_{1 \leq j \leq J} \|M_s(t_j, \cdot) - |\tilde{\mathbf{M}}^j|\|_{L^2} \leq C(k + h).$$

Remark 4.1 (Extension of results to 3D). Let $\Omega \subset \mathbb{R}^3$. Carbou and Fabrie [10] have proved the local existence of a regular solution of (1.1), (1.6)–(1.7) (for $\mathbf{h}_{\text{eff}} = \Delta \mathbf{m}$) in the case $M_s = \text{const}$. For variable, strictly positive $M_s \in C^2(\overline{\Omega}_T)$, we may proceed accordingly then to generalize their result to 3D.

5. Construction of weak solutions. Scheme B. Let $\Omega \subset \mathbb{R}^3$. We consider a reformulation of (1.5), which has been proposed by Gilbert for the case $M_s = \text{const}$; cf. [19]. As in (3.4), we take the cross product of (3.1) with \mathbf{m} , and hence restate (1.5) (with $\mathbf{h}_{\text{eff}} = \Delta \mathbf{m}$) as

$$(5.1) \quad \mathbf{m}_t = \kappa \mathbf{m} - \gamma (1 + \delta^2) \mathbf{m} \times \Delta \mathbf{m} + \delta \frac{\mathbf{m}}{M_s} \times \mathbf{m}_t.$$

We introduce the notion of weak solutions of (5.1), (1.6)–(1.7), which generalizes strong solutions introduced in section 3. Let $\kappa \in C^1(\Omega_T)$.

DEFINITION 5.1. Let $\mathbf{m}_0 \in L^\infty(\Omega, \mathbb{R}^3) \cap W^{1,2}(\Omega, \mathbb{R}^3)$, such that $|\mathbf{m}_0(\cdot)| = M_s(0, \cdot)$ almost everywhere in Ω . A function $\mathbf{m} : \Omega_T \rightarrow \mathbb{R}^3$ is called a weak solution of (5.1), (1.6)–(1.7) if

- (1) $\partial_t \mathbf{m} \in L^2(\Omega_T, \mathbb{R}^3)$, and $\mathbf{m} \in L^\infty(0, T; W^{1,2}(\Omega, \mathbb{R}^3))$, with $\mathbf{m}(0, \cdot) = \mathbf{m}_0$ in the sense of traces,
- (2) $|\mathbf{m}| \in L^\infty(\Omega_T)$, and satisfies (1.8) almost everywhere in Ω_T ,
- (3) for all $\phi \in C^\infty(\overline{\Omega}_T, \mathbb{R}^3)$, there holds

$$(5.2) \quad \int_{\Omega_T} \left[\langle \mathbf{m}_t, \phi \rangle - \kappa \langle \mathbf{m}, \phi \rangle - \frac{\delta}{M_s} \langle \mathbf{m} \times \mathbf{m}_t, \phi \rangle - \gamma (1 + \delta^2) \langle \mathbf{m} \times \nabla \mathbf{m}, \nabla \phi \rangle \right] dx dt = 0.$$

In the remainder of this section, we verify existence of weak solutions of (5.1), (1.6)–(1.7), by using a practical finite-element based scheme.

Scheme B: Let $\kappa \in C(\overline{\Omega}_T)$ and $\mathbf{M}^0 \in \mathbf{V}_h$. For $j = 0, 1, 2, \dots, J - 1$ find $\mathbf{M}^{j+1} \in \mathbf{V}_h$ such that for all $\Phi \in \mathbf{V}_h$ there holds

$$\begin{aligned} & (d_t \mathbf{M}^{j+1}, \Phi)_h - \left(\kappa_{j+1} \mathbf{M}^{j+1/2}, \Phi \right)_h - \delta \left(\frac{\mathbf{M}^j}{M_s^{j+1/2}} \times d_t \mathbf{M}^{j+1}, \Phi \right)_h \\ & = -\gamma (1 + \delta^2) \left(\mathbf{M}^{j+1/2} \times \tilde{\Delta}_h \mathbf{M}^{j+1/2}, \Phi \right)_h. \end{aligned}$$

Remark 5.1. 1. The linear second term in Scheme B is motivated from the identity

$$\mathbf{M}^j \times d_t \mathbf{M}^{j+1} = \left(\mathbf{M}^{j+1/2} - \frac{k}{2} d_t \mathbf{M}^{j+1} \right) \times d_t \mathbf{M}^{j+1} = \mathbf{M}^{j+1/2} \times d_t \mathbf{M}^{j+1}.$$

2. As is detailed in [6] for the case $M_s \equiv \text{const}$, solutions to Scheme B satisfy a perturbed version of a discretization of (3.1), which is due to the competition of local

and nonlocal aspects in fully discrete finite-element based discretizations of (3.1) and (5.1), respectively. First, we verify the existence of solutions to Scheme B.

LEMMA 5.1. *Let $T > 0$, $\kappa \in C(\overline{\Omega}_T)$, $J + 1 = \lceil T/k \rceil$, and $\mathbf{M}^0 \in \mathbf{V}_h$ be given. Then, for a sufficiently small $k \leq k_0(T)$, there exists $\{\mathbf{M}^{j+1}\}_{j=0}^J \subset \mathbf{V}_h$, which solves Scheme B.*

Proof. Let $\mathbf{M}^j \in \mathbf{V}_h$ be given. We define a continuous functional $\mathbf{F} : \mathbf{V}_h \rightarrow \mathbf{V}_h$ by setting for $\mathbf{W} \in \mathbf{V}_h$

$$\mathbf{F}(\mathbf{W}) = \frac{2}{k} \{ \mathbf{W} - \mathbf{M}^j \} - \mathcal{I}_h \left[\kappa_{j+1} \mathbf{W} + \frac{2\delta}{kM_s^{j+1/2}} \mathbf{W} \right. \\ \left. \times \left[\mathbf{W} - \mathbf{M}^j \right] - \gamma(1 + \delta^2) \mathbf{W} \times \tilde{\Delta}_h \mathbf{W} \right].$$

For $k^{-1} > \max_{\Omega_T} \kappa$, and all $\mathbf{W} \in \mathbf{V}_h$ such that $(1 - k \max_{\Omega_T} \kappa) \|\mathbf{W}\|_h \geq \|\mathbf{M}^j\|$, we have

$$\begin{aligned} (\mathbf{F}(\mathbf{W}), \mathbf{W})_h &= \frac{2}{k} \left(\|\mathbf{W}\|_h^2 - (\mathbf{M}^j, \mathbf{W})_h \right) - \left(\kappa(t_{j+1}, \cdot) \mathbf{W}, \mathbf{W} \right)_h \\ &\geq \frac{2}{k} \|\mathbf{W}\|_h \left((1 - k \max_{\Omega_T} \kappa) \|\mathbf{W}\|_h - \|\mathbf{M}^j\|_h \right) \\ &\geq 0. \end{aligned}$$

Hence, Brouwer’s fixed point theorem (see also Evans [17, p. 493]) implies the existence of $\mathbf{W}^* \in \mathbf{V}_h$ such that $\mathbf{F}(\mathbf{W}^*) = 0$. Then, $\mathbf{M}^{j+1} := 2\mathbf{W}^* - \mathbf{M}^j$ solves (5.3). \square

We study stability properties of discrete solutions from Scheme B. This is established in the next lemmas.

LEMMA 5.2. *Let the assumptions of Lemma 5.1 be fulfilled, and $\kappa \in C^2(\Omega_T)$. Then*

- (i) $\max_{\mathbf{z} \in \mathcal{N}_h} \max_{0 \leq j \leq J+1} |\mathbf{M}^j(\mathbf{z})|^2 + \max_{\mathbf{z} \in \mathcal{N}_h} \max_{1 \leq j \leq J+1} |d_t |\mathbf{M}^j(\mathbf{z})|^2| \leq C,$
- (ii) $\|\nabla \mathbf{M}^{j+1}\|^2 + k \sum_{j=0}^J \frac{\delta}{1 + \delta^2} \|d_t \mathbf{M}^{j+1}\|_h^2 \leq C,$
- (iii) $k \sum_{j=0}^J \|\mathbf{M}^{j+1/2} \times \tilde{\Delta}_h \mathbf{M}^{j+1/2}\|_h^2 \leq C,$

where $C \equiv C(t_J, \gamma, \delta, M_s)$.

Proof. To verify assertion (i), choose $\Phi = \mathbf{M}^{j+1/2}(\mathbf{z})\varphi_{\mathbf{z}}$ for Scheme B, and observe Remark 5.1 to find

$$(5.3) \quad d_t |\mathbf{M}^{j+1}(\mathbf{z})|^2 = 2\kappa(t_{j+1}, \mathbf{z}) |\mathbf{M}^{j+1/2}(\mathbf{z})|^2 \\ \leq C \left(|\mathbf{M}^{j+1}(\mathbf{z})|^2 + |\mathbf{M}^j(\mathbf{z})|^2 \right).$$

Summation of (5.3) over j followed by an application of the discrete version of Gronwall’s lemma yields to

$$\max_{\mathbf{z} \in \mathcal{N}_h} \max_{0 \leq j \leq J+1} |\mathbf{M}^j(\mathbf{z})|^2 \leq C.$$

The rest of assertion (i) now follows from the last inequality and (5.3).

In order to verify (ii), we first choose $\Phi = -\tilde{\Delta}_h \mathbf{M}^{j+1/2}$. We obtain

$$\begin{aligned} & -(\kappa(t_{j+1}, \cdot) \mathbf{M}^{j+1/2}, -\tilde{\Delta}_h \mathbf{M}^{j+1/2})_h = -(\mathcal{I}_h[\kappa(t_{j+1}, \cdot) \mathbf{M}^{j+1/2}], \tilde{\Delta}_h \mathbf{M}^{j+1/2})_h - I \\ & \geq -C \|\kappa(t_{j+1}, \cdot)\|_{C(\bar{\Omega})} \|\nabla \mathbf{M}^{j+1/2}\|^2 \\ & - \left| (\mathbf{M}^{j+1/2} \otimes \nabla \kappa(t_{j+1}, \cdot), \nabla \mathbf{M}^{j+1/2}) \right| - I + II, \end{aligned}$$

with error terms I, II , which can be controlled by standard interpolation estimates, using $\nabla^2 \mathbf{M}^{j+1/2}|_K = 0$, for all $K \in \mathcal{T}_h$,

$$\begin{aligned} I & := \left([\mathbf{Id} - \mathcal{I}_h](\kappa(t_{j+1}, \cdot) \mathbf{M}^{j+1/2}), \tilde{\Delta}_h \mathbf{M}^{j+1/2} \right)_h \\ & \leq Ch^2 \|\kappa(t_{j+1}, \cdot)\|_{C^2(\Omega)} \|\mathbf{M}^{j+1/2}\|_{W^{1,2}(\Omega)} \|\Delta_h \mathbf{M}^{j+1/2}\|, \\ II & := \left(\nabla[\mathbf{Id} - \mathcal{I}_h](\kappa(t_{j+1}, \cdot) \mathbf{M}^{j+1/2}), \nabla \mathbf{M}^{j+1/2} \right) \\ & \leq Ch \|\kappa(t_{j+1}, \cdot)\|_{C^2(\Omega)} \|\mathbf{M}^{j+1/2}\|_{W^{1,2}(\Omega)}^2. \end{aligned}$$

Putting things together, by inverse estimate we find for some $C \equiv C(\kappa, \gamma, \delta) > 0$,

$$(5.4) \quad \frac{1}{2} d_t \|\nabla \mathbf{M}^{j+1}\|^2 \leq C(1+h) \|\nabla \mathbf{M}^{j+1}\|^2 - \delta \left(\frac{\mathbf{M}^j}{M_s^{j+1/2}} \times d_t \mathbf{M}^{j+1}, \tilde{\Delta}_h \mathbf{M}^{j+1/2} \right)_h.$$

Choosing $\Phi = \mathcal{I}_h \left[\frac{d_t \mathbf{M}^{j+1}}{M_s^{j+1/2}} \right]$ in Scheme B yields

$$\begin{aligned} \frac{\delta}{\gamma(1+\delta^2)} \left\| \frac{d_t \mathbf{M}^{j+1}}{\sqrt{M_s^{j+1/2}}} \right\|_h^2 & = \frac{\delta}{\gamma(1+\delta^2)} \left(\kappa(t_{j+1}, \cdot) \frac{\mathbf{M}^{j+1/2}}{M_s^{j+1/2}}, d_t \mathbf{M}^{j+1} \right)_h \\ & - \delta \left(\frac{\mathbf{M}^{j+1/2}}{M_s^{j+1/2}} \times \tilde{\Delta}_h \mathbf{M}^{j+1/2}, d_t \mathbf{M}^{j+1} \right)_h. \end{aligned}$$

Now, adding (5.4) to this relation, thanks to Remark 5.1 and Young's inequality, we find

$$d_t \|\nabla \mathbf{M}^{j+1}\|^2 + \frac{\delta}{\gamma(1+\delta^2)} \left\| \frac{d_t \mathbf{M}^{j+1}}{\sqrt{M_s^{j+1/2}}} \right\|_h^2 \leq C(1+h) \left(1 + \|\nabla \mathbf{M}^{j+1}\|^2 \right).$$

We employ discrete Gronwall's inequality, and assertion (ii) follows. Assertion (iii) is now an immediate consequence from Scheme B. \square

Note that (5.3) is the discrete version of property (1.8) for solutions of Scheme A at every node $\mathbf{z} \in \mathcal{N}_h$; hence, we can only expect that $|\mathbf{M}^j(\mathbf{z})|^2 \approx M_s^2(t_j, \mathbf{z})$, for all $1 \leq j \leq J$, and all $\mathbf{z} \in \mathcal{N}_h$.

DEFINITION 5.2. For $(t, \mathbf{x}) \in [t_j, t_{j+1}) \times \Omega$ we define

$$\begin{aligned} \mathcal{M}_{k,h}(t, \mathbf{x}) & := \frac{t-t_j}{k} \mathbf{M}^{j+1}(\mathbf{x}) + \frac{t_{j+1}-t}{k} \mathbf{M}^j(\mathbf{x}) = \mathbf{M}^j(\mathbf{x}) + (t-t_j) d_t \mathbf{M}^{j+1}(\mathbf{x}), \\ \mathcal{M}_{k,h}^-(t, \mathbf{x}) & := \mathbf{M}^j(\mathbf{x}), \quad \mathcal{M}_{k,h}^+(t, \mathbf{x}) := \mathbf{M}^{j+1}(\mathbf{x}), \quad \overline{\mathcal{M}}_{k,h}(t, \mathbf{x}) := \mathbf{M}^{j+1/2}(\mathbf{x}). \end{aligned}$$

For almost every $T' > 0$, assertion (ii) of Lemma 5.2 may be rewritten as

$$(5.5) \quad \|\nabla \mathcal{M}_{k,h}^+(T', \cdot)\|_{L^2}^2 + \frac{\delta}{1+\delta^2} \int_0^{T'} \|(\mathcal{M}_{k,h})_t(s, \cdot)\|_h^2 dt \leq C.$$

This bound yields the existence of some $\mathbf{m} \in W^{1,2}(\Omega_T, \mathbb{R}^3)$, which is the weak limit (as $k, h \rightarrow 0$) of a subsequence $\{\mathbf{M}_{k,h}\}$ such that

$$(5.6) \quad \begin{aligned} \mathcal{M}_{k,h} &\rightharpoonup \mathbf{m} && \text{in } W^{1,2}(\Omega_T, \mathbb{R}^3), \\ \nabla \mathcal{M}_{k,h}^+, \nabla \mathcal{M}_{k,h}^-, \nabla \overline{\mathcal{M}}_{k,h} &\rightharpoonup \nabla \mathbf{m} && \text{in } L^2(\Omega_T, \mathbb{R}^3), \\ \mathcal{M}_{k,h}^-, \mathcal{M}_{k,h}^+, \overline{\mathcal{M}} &\rightarrow \mathbf{m} && \text{in } L^2(\Omega_T, \mathbb{R}^3), \end{aligned}$$

where we use the Aubin–Lions compactness result to obtain (5.6)₃. Dropping the index $\{k, h\}$ on $\mathcal{M}_{k,h}$, Scheme B may be rewritten in the following form: taking $\Phi(t, \cdot) := \mathcal{I}_h \phi(t, \cdot)$ for $\phi \in C^\infty(\Omega_T, \mathbb{R}^3)$ there holds

$$(5.7) \quad \int_0^T \left[(\mathcal{M}_t, \Phi)_h - (\kappa^+ \overline{\mathcal{M}}, \Phi)_h - \delta \left(\mathcal{M}^- \times \frac{\mathcal{M}_t}{M_s}, \Phi \right)_h + \gamma (1 + \delta^2) \left(\overline{\mathcal{M}} \times \tilde{\Delta}_h \overline{\mathcal{M}}, \Phi \right)_h \right] dt = 0.$$

THEOREM 5.1. *Suppose that the assumptions of Lemma 5.2 are valid, and*

$$\mathbf{M}^0 \rightarrow \mathbf{m}_0 \quad \text{in } W^{1,2}(\Omega, \mathbb{R}^3), \quad \text{and} \quad |\mathbf{M}^0(\mathbf{z})| \rightarrow M_s(0, \mathbf{z}) \quad \forall \mathbf{z} \in \mathcal{N}_h \quad (h \rightarrow 0).$$

For $k, h \rightarrow 0$ there exists $\mathbf{m} \in W^{1,2}(\Omega_T, \mathbb{R}^3)$ such that iterates $\{\mathcal{M}_{k,h}\}$ of Scheme B subconverge to \mathbf{m} in $W^{1,2}(\Omega_T, \mathbb{R}^3)$, and \mathbf{m} is a weak solution to (5.1), (1.6)–(1.7).

Remark 5.2. It is not known if iterates of Scheme B converge at optimal rates to (locally existing) strong solutions of (1.5)–(1.7): this shortcoming reflects the known, but seemingly confusing situation even in the simpler case $M_s = \text{const}$ in [6, 7], where (a simplified version of) Scheme B was developed to construct weak solutions in the limit, but where convergence with rates in the presence of existing strong solutions is still an open problem.

Proof. Step 1. Verification of (i), (iii) of Definition 5.1. Effects of reduced integration need to be controlled by using (2.3): For almost all $t \in (0, T)$, we have

$$|(\mathcal{M}_t, \Phi) - (\mathcal{M}_t, \Phi)_h| \leq Ch \|\mathcal{M}_t\|_{L^2(\Omega)} \|\nabla \Phi\|_{L^2(\Omega)}.$$

This bound, and (5.5), together with standard Lagrange interpolation results yields to

$$(5.8) \quad \int_0^T (\mathcal{M}_t, \Phi)_h dt \rightarrow \int_0^T (\mathbf{m}_t, \phi) dt \quad (k, h \rightarrow 0).$$

In a similar fashion, we obtain

$$(5.9) \quad \int_0^T (\kappa^+ \overline{\mathcal{M}}, \Phi)_h dt \rightarrow \int_0^T (\kappa \mathbf{m}, \phi) dt \quad (k, h \rightarrow 0).$$

The convergence

$$(5.10) \quad \int_0^T \left(\frac{\overline{\mathcal{M}}}{M_s} \times \mathcal{M}_t, \Phi \right)_h dt \rightarrow \int_0^T \left(\frac{\mathbf{m}}{M_s} \times \mathbf{m}_t, \phi \right) dt \quad (k, h \rightarrow 0)$$

again uses (2.3), (5.5), and Lemma (5.2), (i) to control effects due to reduced integration, and (5.6)₃ together with (5.6)₂.

The last term in (5.7) requires a more detailed study. We write

$$\begin{aligned}
 - \left(\overline{\mathcal{M}} \times \tilde{\Delta}_h \overline{\mathcal{M}}, \Phi \right)_h &= \left(\overline{\mathcal{M}} \times \Phi, \tilde{\Delta}_h \overline{\mathcal{M}} \right)_h = \left([\text{Id} - \mathcal{I}_h] \left(\overline{\mathcal{M}} \times \Phi \right), \tilde{\Delta}_h \overline{\mathcal{M}} \right)_h \\
 + \left(\nabla [\text{Id} - \mathcal{I}_h] \left(\overline{\mathcal{M}} \times \Phi \right), \nabla \overline{\mathcal{M}} \right) &- \left(\nabla \left[\overline{\mathcal{M}} \times \Phi \right], \nabla \overline{\mathcal{M}} \right) =: I + II - III.
 \end{aligned}$$

Control of I uses the bound $\|\tilde{\Delta}_h \Psi\| \leq Ch^{-1} \|\nabla \Psi\|_{L^2}$, and estimates of nodal interpolation,

$$\begin{aligned}
 I &\leq Ch^2 h^{-1} \sum_{K \in \mathcal{T}} \|D^2(\overline{\mathcal{M}} \times \Phi)\|_{L^2(K)} \|\nabla \overline{\mathcal{M}}\|_{L^2(K)} \\
 &\leq Ch \|\nabla \overline{\mathcal{M}}\| \|\Phi\|_{C^2} \|\nabla \overline{\mathcal{M}}\|.
 \end{aligned}$$

Similarly, we obtain

$$II \leq Ch \sum_{K \in \mathcal{T}} \|D^2(\overline{\mathcal{M}} \times \Phi)\|_{L^2(K)} \|\nabla \overline{\mathcal{M}}\|_{L^2(K)} \leq Ch \|\nabla \overline{\mathcal{M}}\| \|\Phi\|_{C^2} \|\nabla \overline{\mathcal{M}}\|.$$

For the third term, we use the vector identity $\langle \nabla \mathbf{a}, \nabla[\mathbf{a} \times \mathbf{b}] \rangle = \langle \nabla \mathbf{a}, \mathbf{a} \times \nabla \mathbf{b} \rangle$, for $\mathbf{a}, \mathbf{b} \in W^{1,2}(\Omega, \mathbb{R}^3)$ to verify

$$III = \left(\nabla \left[\overline{\mathcal{M}} \times \Phi \right], \nabla \overline{\mathcal{M}} \right) = \left(\overline{\mathcal{M}} \times \nabla \Phi, \nabla \overline{\mathcal{M}} \right).$$

Putting things together then yields

$$\begin{aligned}
 (5.11) \quad \int_0^T \left(\overline{\mathcal{M}} \times \tilde{\Delta}_h \overline{\mathcal{M}}, \Phi \right)_h ds &\rightarrow - \int_0^T (\mathbf{m} \times \nabla \phi, \nabla \mathbf{m}) dt \\
 &= - \int_0^T (\nabla[\mathbf{m} \times \phi], \nabla \phi) ds \quad (k, h \rightarrow 0).
 \end{aligned}$$

We may now insert (5.8)–(5.11) into (5.7) to validate (ii) of Definition 5.1 for $\mathbf{m} : \Omega_T \rightarrow \mathbb{R}^3$.

Step 2. Verification of (ii) of Definition 5.1. The function $|\mathbf{m}|$ solves

$$(5.12) \quad \frac{d}{dt} |\mathbf{m}|^2 = 2\kappa |\mathbf{m}|^2 \quad \text{a.e. in } \Omega_T.$$

Its verification is split into two parts:

A) *Verification of (5.12) for $\mathbf{z} \in \mathcal{N}_h$.* For all $t \in [t_{m-1}, t_m)$, there hold

$$(5.13) \quad |\mathcal{M}_{k,h}(t, \mathbf{z})|^2 = |\mathcal{M}_{k,h}^-(t, \mathbf{z})|^2 + 2(t - t_{m-1})\kappa^+(t, \mathbf{z}) |\overline{\mathcal{M}}_{k,h}(t, \mathbf{z})|^2 \quad \forall \mathbf{z} \in \mathcal{N}_h,$$

$$(5.14) \quad |\mathcal{M}_{k,h}(t, \mathbf{x})|^2 = \left| \sum_{\mathbf{z}_i \in \mathcal{N}_h \cap \overline{K}} \mathcal{M}_{k,h}(t, \mathbf{z}_i) \varphi_{\mathbf{z}_i}(\mathbf{x}) \right|^2 \quad \forall \mathbf{x} \in K.$$

For every $\mathbf{z} \in \mathcal{N}_h$, consider $\{|\mathcal{M}_{k,h}(\cdot, \mathbf{z})|^2\}_k$. Let $0 \leq s \leq t \leq T$ such that $t \in$

$[t_{m-1}, t_m]$, and $s \in [t_{j-1}, t_j]$, for some $t_j \leq t_m$. Then

$$\begin{aligned}
 & |\mathcal{M}_{k,h}(t, \mathbf{z})|^2 - |\mathcal{M}_{k,h}(s, \mathbf{z})|^2 = |\mathcal{M}_{k,h}(t, \mathbf{z})|^2 - |\mathbf{M}^{m-1}(\mathbf{z})|^2 \\
 & \quad + \sum_{\ell=j}^{m-1} \{ |\mathbf{M}^\ell(\mathbf{z})|^2 - |\mathbf{M}^{\ell-1}(\mathbf{z})|^2 \} + |\mathbf{M}^{j-1}(\mathbf{z})|^2 - |\mathcal{M}_{k,h}(s, \mathbf{z})|^2 \\
 (5.15) \quad & = 2(t - t_{m-1}) \left(\kappa|_{[t_{m-1}, t_m]} \right)^+ (\mathbf{z}) |\overline{\mathbf{M}}_{k,h}(t, \mathbf{z})|^2 + 2k \sum_{\ell=j}^{k-1} \kappa(t_\ell, \mathbf{z}) |\mathbf{M}^{\ell-1/2}(\mathbf{z})|^2 \\
 & \quad + 2(t_{j-1} - s) \left(\kappa|_{[t_{j-1}, t_j]} \right)^+ (\mathbf{z}) |\overline{\mathbf{M}}_{k,h}(s, \mathbf{z})|^2,
 \end{aligned}$$

and uniform (Lipschitz-)equicontinuity of $\{|\mathcal{M}_{k,h}(\cdot, \mathbf{z})|^2\}_k$, for every $\mathbf{z} \in \mathcal{N}_h$ then follows from the computation

$$\begin{aligned}
 & \| |\mathcal{M}_{k,h}(t, \mathbf{z})|^2 - |\mathcal{M}_{k,h}(s, \mathbf{z})|^2 \|_{C([0, T])} \leq C \{ (t - t_{m-1}) + (t_{m-1} - t_j) + (t_j - s) \} \\
 (5.16) \quad & \leq C |t - s| \quad \forall \mathbf{z} \in \mathcal{N}_h.
 \end{aligned}$$

By the Arzela–Ascoli theorem, for every $\mathbf{z} \in \mathcal{N}_h$ there exist a subsequence $\{|\mathcal{M}_{k,h}(\cdot, \mathbf{z})|^2\}_k$, and $\chi^2(\cdot, \mathbf{z}) \in C([0, T])$, such that for $k \rightarrow 0$,

$$\max_{\mathbf{z} \in \mathcal{N}_h} \max_{t \in [0, T]} \left| |\mathcal{M}_{k,h}(t, \mathbf{z})|^2 - \chi^2(t, \mathbf{z}) \right| \rightarrow 0.$$

We need to verify that $t \mapsto \chi^2(t, \mathbf{z})$ solves (1.8), for all $\mathbf{z} \in \mathcal{N}_h$. Let $t \in [t_{m-1}, t_m]$. By construction, there holds

$$\begin{aligned}
 & |\mathcal{M}_{k,h}(t, \mathbf{z})|^2 = |\mathbf{M}^0(\mathbf{z})|^2 + 2 \int_{t_{m-1}}^t \left\{ \kappa^+(s, \mathbf{z}) |\mathbf{M}^{m-1/2}(\mathbf{z})|^2 - \kappa(s, \mathbf{z}) |\mathcal{M}_{k,h}(s, \mathbf{z})|^2 \right\} ds \\
 (5.17) \quad & + 2 \sum_{\ell=1}^{m-1} \int_{t_{\ell-1}}^{t_\ell} \left\{ \kappa^+(s, \mathbf{z}) |\mathbf{M}^{\ell-1/2}(\mathbf{z})|^2 - \kappa(s, \mathbf{z}) |\mathcal{M}_{k,h}(s, \mathbf{z})|^2 \right\} ds \\
 & + 2 \int_0^t \kappa(s, \mathbf{z}) |\mathcal{M}_{k,h}(s, \mathbf{z})|^2 ds.
 \end{aligned}$$

By uniform equicontinuity of $\{|\mathcal{M}_{k,h}(\cdot, \mathbf{z})|^2\}_k$, Lipschitz continuity of $t \mapsto \kappa(t, \mathbf{z})$, for every $\mathbf{z} \in \mathcal{N}_h$, and parallelogram identity together with the bound of Lemma 5.1, (iii) to conclude for all $\varepsilon > 0$, and sufficiently small $k = k(\varepsilon)$,

$$\sum_{\ell=1}^m \int_{t_{\ell-1}}^{t_\ell} \left| \kappa^+(s, \mathbf{z}) |\mathbf{M}^{\ell-1/2}(\mathbf{z})|^2 - \kappa(s, \mathbf{z}) |\mathcal{M}_{k,h}(s, \mathbf{z})|^2 \right| ds \leq \varepsilon.$$

Together with uniform convergence of $|\mathcal{M}_{k,h}(\cdot, \mathbf{z})|^2 \rightarrow \chi^2(\cdot, \mathbf{z})$ ($k \rightarrow 0$) on $[0, T]$, (1.8) holds for the limit $\chi^2(\cdot, \mathbf{z})$, for every $\mathbf{z} \in \mathcal{N}_h$.

B) *Verification of (5.12) for almost all $(t, \mathbf{x}) \in \Omega_T$.* Let $\mathbf{x} \in \overline{K} = \text{conv}(\mathbf{z}_1, \dots, \mathbf{z}_4)$, and fix one $\mathbf{z}_* = \min_i \{ |\mathbf{z}_i - \mathbf{x}| \}$. There exist $\xi_* \equiv \xi_*(\mathbf{x}, \mathbf{z}_*) \in \mathbb{S}^2$, and $0 \leq h_*(\mathbf{x}, \mathbf{z}_*) < h$, such that $\mathcal{M}_{k,h}(t, \mathbf{x}) = \mathcal{M}_{k,h}(t, \mathbf{z}_*) + h_* \nabla \mathcal{M}_{k,h}(t, \cdot) \xi_*$ owing $\nabla \mathcal{M}_{k,h}(t, \cdot)|_K = \text{const.}$, and

$$(5.18) \quad |\mathcal{M}_{k,h}(t, \mathbf{x})|^2 = |\mathcal{M}_{k,h}(t, \mathbf{z}_*)|^2 + 2h_* \left\langle \mathcal{M}_{k,h}(t, \mathbf{z}_*), \nabla \mathcal{M}_{k,h}(t, \cdot) \xi_* \right\rangle + h_*^2 |\nabla \mathcal{M}_{k,h}(t, \cdot) \xi_*|^2.$$

For the leading term on the right-hand side, we have (5.13). Therefore, we modify argument (5.17) in the way that instead of adding and subtracting $\int_0^t \kappa(s, \mathbf{z}) |\mathcal{M}_{k,h}(s, \mathbf{z})|^2 ds$ on the right-hand side, we choose $\pm \int_0^t \kappa(s, \mathbf{x}) |\mathcal{M}_{k,h}(s, \mathbf{x})|^2 ds$. For all $\mathbf{x} \in \Omega$, there holds

$$\begin{aligned}
 & |\mathcal{M}_{k,h}(t, \mathbf{x})|^2 - |\mathbf{M}^0(\mathbf{x})|^2 - 2 \int_0^t \kappa(s, \mathbf{x}) |\mathcal{M}_{k,h}(s, \mathbf{x})|^2 ds \\
 &= [|\mathbf{M}^0(\mathbf{z}_*)|^2 - |\mathbf{M}^0(\mathbf{x})|^2] + 2 \int_{t_{m-1}}^t \left\{ \kappa^+(s, \mathbf{z}) [|\overline{\mathcal{M}}_{k,h}(s, \mathbf{z})|^2 \right. \\
 &\quad \left. - |\mathcal{M}_{k,h}(s, \mathbf{x})|^2] - [\kappa(s, \mathbf{x}) - \kappa^+(s, \mathbf{z})] |\mathcal{M}_{k,h}(s, \mathbf{x})|^2 \right\} ds \\
 (5.19) \quad &+ 2 \sum_{\ell=1}^{m-1} \int_{t_{\ell-1}}^{t_\ell} \left\{ \kappa^+(s, \mathbf{z}) |\overline{\mathcal{M}}_{k,h}(s, \mathbf{z})|^2 - \kappa(s, \mathbf{x}) |\mathcal{M}_{k,h}(s, \mathbf{x})|^2 \right\} ds \\
 &+ 2h_* \langle \mathcal{M}_{k,h}(t, \mathbf{z}_*), \nabla \mathcal{M}_{k,h}(t, \cdot) \xi_* \rangle + h_*^2 |\nabla \mathcal{M}_{k,h}(t, \cdot) \xi_*|^2 \\
 &\leq Ch_* \left[\|\nabla \mathbf{M}^0\|_{L^\infty} \max_{\mathbf{z}_i \in \mathcal{N}_h \cap \overline{K}} |\mathbf{M}^0(\mathbf{z}_i)| + \|\kappa\|_{L^\infty(\Omega_T)} \|\mathcal{M}_{k,h}\|_{L^\infty(\Omega_T)} \right. \\
 &\quad \left. \int_{t_{m-1}}^t |\nabla \overline{\mathcal{M}}_{k,h}(s, \cdot)| ds \right] \\
 &+ (k+h) \|\kappa\|_{W^{1,\infty}(\Omega_T)} \|\mathcal{M}_{k,h}\|_{L^\infty(\Omega_T)}^2 + Ck \|\mathcal{M}_{k,h}\|_{L^\infty(\Omega_T)} \\
 &\quad \int_{t_{m-1}}^t |\partial_t \mathcal{M}_{k,h}(s, \mathbf{x})| ds \\
 &+ Ch \|\mathcal{M}_{k,h}\|_{L^\infty(\Omega_T)} \|\nabla \mathcal{M}_{k,h}(t, \cdot)\|_{L^\infty(\Omega)} + h^2 \|\nabla \mathcal{M}_{k,h}(t, \cdot)\|_{L^\infty(\Omega)}^2,
 \end{aligned}$$

where we use binomial formula, differentiability, and boundedness of $\kappa : \Omega_T \rightarrow \mathbb{R}$, the estimate

$$\|\mathcal{M}_{k,h}^+ - \mathcal{M}_{k,h}\| + \|\overline{\mathcal{M}}_{k,h} - \mathcal{M}_{k,h}\| \leq 2k \|\partial_t \mathcal{M}_{k,h}\|,$$

and Lemma 5.2. Integration over Ω_T of (5.19) then leads to

$$\int_{\Omega_T} \left| |\mathcal{M}_{k,h}(s, \mathbf{x})|^2 - |\mathbf{M}^0(\mathbf{x})|^2 - 2 \int_0^s \kappa(\tilde{s}, \mathbf{x}) |\mathcal{M}_{k,h}(\tilde{s}, \mathbf{x})|^2 d\tilde{s} \right| dx ds \rightarrow 0 \quad (k, h \rightarrow 0).$$

Because of binomial formula and (5.6)₂, we may pass to limits in every term, which verifies (1.8) almost everywhere in Ω_T for $\mathbf{m} : \Omega_T \rightarrow \mathbb{R}^3$. \square

6. Solution of the nonlinear system in Scheme B. We use a simple fixed-point iterative algorithm to solve the nonlinear system in each step of Scheme B. Similar fixed-point algorithms were employed in the context of the LLG and Maxwell-LLG equations; see [5, 6]. We solve the nonlinear system for $\mathbf{W}_h := \mathbf{M}_h^{j+1/2}$. The time derivative $d_t \mathbf{m}_h^{j+1}$ is replaced by $\frac{2}{k} (\mathbf{W}_h - \mathbf{M}_h^j)$. After linearization of the nonlinear term $(\mathbf{W}_h \times \tilde{\Delta}_h \mathbf{W}_h, \Phi)_h$ and using the identity $\mathbf{M}_h^j \times d_t \mathbf{M}_h^{j+1} = -\frac{2}{k} \mathbf{M}_h^{j+1/2} \times \mathbf{M}_h^j$, we obtain the following algorithm.

ALGORITHM 6.1. Set $\mathbf{W}_h^0 := \mathbf{M}_h^j$ and $\ell := 0$.

(i) Compute $\mathbf{W}_h^{\ell+1} \in \mathbf{V}_h$ such that for all $\Phi \in \mathbf{V}_h$ there holds

$$(6.1) \quad \frac{2}{k} (\mathbf{W}_h^{\ell+1}, \Phi)_h - (\kappa^{j+1} \mathbf{W}_h^{\ell+1}, \Phi)_h - \frac{2\delta}{k} \left(\frac{\mathbf{M}_h^j}{M_s^{j+1/2}} \times \mathbf{W}_h^{\ell+1}, \Phi \right)_h + \gamma(1 + \delta^2) \left(\mathbf{W}_h^{\ell+1} \times \tilde{\Delta}_h \mathbf{W}_h^\ell, \Phi \right)_h = \frac{2}{k} \left(\mathbf{M}_h^j, \Phi \right)_h.$$

(ii) For fixed $\varepsilon > 0$, stop and set $\mathbf{M}_h^{j+1} := 2\mathbf{W}_h^{\ell+1} - \mathbf{M}_h^j$, once

$$(6.2) \quad \|\mathbf{W}_h^{\ell+1} - \mathbf{W}_h^\ell\|_h \leq \varepsilon.$$

(iii) Set $\ell := \ell + 1$ and go to (i).

We give a condition, which is sufficient for convergence for every $1 \leq j \leq J$.

LEMMA 6.1. Suppose that $\|\mathbf{M}_h^j\|_{L^\infty} \leq C_0$. For all $\ell \geq 0$ there exists a unique solution to (6.1). Further, there holds

$$(6.3) \quad \|\mathbf{W}_h^{\ell+1} - \mathbf{W}_h^\ell\|_h \leq \Theta \|\mathbf{W}_h^\ell - \mathbf{W}_h^{\ell-1}\|_h,$$

with $\Theta < 1$ provided that $k < \frac{(2-k\|\kappa^{j+1}\|_{L^\infty})h^2}{\gamma(1+\delta^2)C_0C}$ is sufficiently small, where $C > 0$ only depends on the geometry of the mesh. For given $\mathbf{M}_h^j \in \mathbf{V}_h$ such that $|\mathbf{M}_h^j(\mathbf{z})| = M_s(t_j, \mathbf{z})$ for all $\mathbf{z} \in \mathcal{N}_h$, by Banach fixed point theorem, the contraction property (6.3) then implies the existence of a unique $\mathbf{M}_h^{j+1} \in \mathbf{V}_h$, which solves Scheme B.

Proof. In order to prove the existence of $\mathbf{W}_h^{\ell+1} \in \mathbf{V}_h$ for $\ell \geq 0$ it suffices to show that the bilinear form $a : \mathbf{V}_h \times \mathbf{V}_h \rightarrow \mathbb{R}$ defined by the left-hand side of (6.1) is positive definite. In order to do so, we set $\Phi = \mathbf{W}_h^{\ell+1}$ in (6.1), and use the properties of κ to obtain for $k < 2/\|\kappa^{j+1}\|_{L^\infty}$,

$$a(\mathbf{W}_h^{\ell+1}, \mathbf{W}_h^{\ell+1}) \geq \left(\frac{2}{k} - \|\kappa^{j+1}\|_{L^\infty} \right) \|\mathbf{W}_h^{\ell+1}\|_h^2 > 0.$$

Next, we choose $\Phi = \mathbf{W}_h^{\ell+1}(\mathbf{z})\varphi_{\mathbf{z}}$ in (6.1) and get (for sufficiently small k)

$$(6.4) \quad |\mathbf{W}_h^{\ell+1}(\mathbf{z})| \leq \frac{2}{(2-k\|\kappa\|_{L^\infty})} |\mathbf{M}_h^j(\mathbf{z})| \quad (\ell \geq 0).$$

Finally, we subtract two subsequent iterations of (6.1), set $\Phi = \mathbf{e}_h^{\ell+1} := \mathbf{W}_h^{\ell+1} - \mathbf{W}_h^\ell$, and use (6.4) together with $\|\tilde{\Delta}_h \mathbf{e}_h^\ell\|_{L^2} \leq Ch^{-2}\|\mathbf{e}_h^\ell\|_h$ to conclude

$$\begin{aligned} \left(\frac{2-k\|\kappa^{j+1}\|_{L^\infty}}{k} \right) \|\mathbf{e}_h^{\ell+1}\|_h^2 &\leq \gamma(1 + \delta^2) \left(\mathbf{W}_h^{\ell+1} \times \tilde{\Delta}_h \mathbf{e}_h^\ell, \mathbf{e}_h^{\ell+1} \right)_h \\ &\leq \gamma(1 + \delta^2) Ch^{-2} \|\mathbf{W}_h^{\ell+1}\|_{L^\infty} \|\mathbf{e}_h^\ell\|_h \|\mathbf{e}_h^{\ell+1}\|_h \\ &\leq \gamma(1 + \delta^2) C_0 Ch^{-2} \|\mathbf{e}_h^\ell\|_h \|\mathbf{e}_h^{\ell+1}\|_h; \end{aligned}$$

i.e., the algorithm converges for k such that $\frac{\gamma(1+\delta^2)C_0C}{(2-k\|\kappa^{j+1}\|_{L^\infty})} \frac{k}{h^2} < 1$. \square

7. Computational experiments. We study temperature effects on the discrete finite-time blow-up problem suggested in [6]. Our computational code was built on the finite element package ALBERT [24].

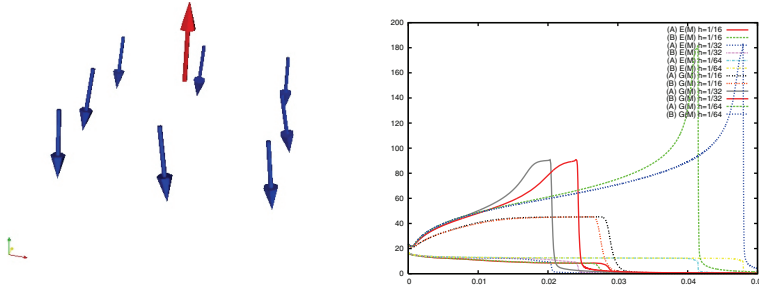


FIG. 2. Detail of the magnetization near the blow-up time (left). $M_s \equiv 1$: Plot of $t_j \mapsto E(\mathbf{M}_h^j)$, and $t_j \mapsto \|\nabla \mathbf{M}_h^j\|_{L^\infty}$ for Schemes A,B ($h = \frac{1}{16}, h = \frac{1}{32}, h = \frac{1}{64}$) (right).

For $\Omega = (0, 1)^2$, let $\mathbf{m}^0 : \Omega \rightarrow \mathbb{S}^2$ be defined by

$$\mathbf{m}^0(\mathbf{x}) = \begin{cases} (0, 0, -M_s(\mathbf{x})) & \text{for } |\mathbf{x}^*| \leq 1/2 \\ M_s(\mathbf{x}) (2\mathbf{x}^* A, A^2 - |\mathbf{x}^*|^2) / (A^2 + |\mathbf{x}^*|^2) & \text{for } |\mathbf{x}^*| \geq 1/2 \end{cases}$$

with $\mathbf{x}^* = (x_1 - 0.5, x_2 - 0.5)$, $\mathbf{x} = (x_1, x_2) \in \Omega$, and $A = (1 - 2|\mathbf{x}^*|)^4/4$. We set $\mathbf{M}_h^0 = I^h \mathbf{m}^0$. The above choice with $M_s \equiv 1$ leads to discrete finite-time blow-up of the solution with singularity at $\mathbf{x} = (0.5, 0.5)$ as reported in [6]. Shortly before the blow-up takes part, the magnetization vector at $\mathbf{x} = (0.5, 0.5)$ points in the $(0, 0, 1)$ direction, with all the surrounding vectors pointing antiparallel, i.e., in the $(0, 0, -1)$ direction; see Figure 2 (left) for detail of the magnetization at $\mathbf{x} = (0.5, 0.5)$ near the blow-up time. For comparison, we display the evolution of the discrete energy, and $\|\nabla \mathbf{M}_h\|_{L^\infty}$ in Figure 2 (right), respectively, for different values of h .

The computational domain Ω was partitioned into uniform squares with side h , each square subdivided into four equal triangles.

For the evolution of the saturation magnetization M_s we adopt the law (1.4) with $\beta = 0.5$, and the function κ given in (1.11). Further we set $\alpha = \delta = \gamma = 1$, $\tau_C = 1$ in all our computations.

For Scheme B, we choose the time step using the following formula $k = 0.0256 h^2$. Given the mesh parameter h we set $\varepsilon = 4 \times 10^{-12} h^2$ in (6.2). With the former choice of the parameter ε the fixed-point iterations in (6.1) always converged after at most 14 iterations with average iteration count of about 5. We were able to use time steps independent of h with the projection scheme (Scheme A) (cf. [4]). The presented results for the projection scheme are computed with $k = 10^{-4}$ for $h = 1/16, 1/32$. For $h = 1/64$, in order to obtain results that are in better agreement with the corresponding results computed by the Scheme B, we take $k = 2.5 \times 10^{-5}$ for the projection scheme.

Example 7.1. The first example demonstrates the effects of constant in time, varying in space temperature $\tau \equiv \tau(\mathbf{x})$. The temperature is defined as

$$\tau(\mathbf{x}) = 0.99 e^{B|\mathbf{x}^*|^2}.$$

We take $B = -25$. The saturation magnetization $M_s(\mathbf{x})$ is now space dependent with values in the interval $[0.1, 1]$.

The discrete energy and $t_j \mapsto \|\nabla \mathbf{M}_h^j\|_{L^\infty}$ are displayed for both Schemes A and B, in Figure 3 (left) and Figure 3 (right), respectively. The discrete energy decreases monotonically for both methods.

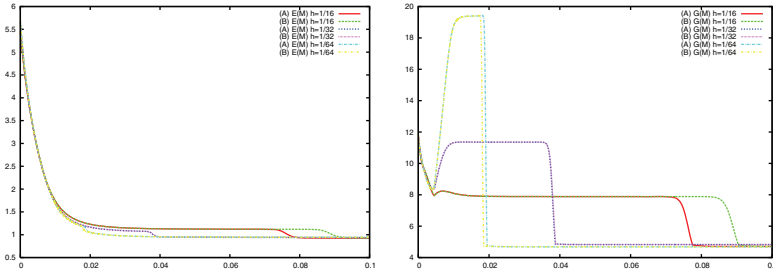


FIG. 3. Example 7.1: Plot of $t_j \mapsto E(\mathbf{M}_h^j)$ for Schemes A,B ($h = \frac{1}{16}, \frac{1}{32}, \frac{1}{64}$) (left). Example 7.1: Plot of $t_j \mapsto E(\mathbf{M}_h^j)$ for Schemes A,B ($h = \frac{1}{16}, \frac{1}{32}, \frac{1}{64}$) (right).

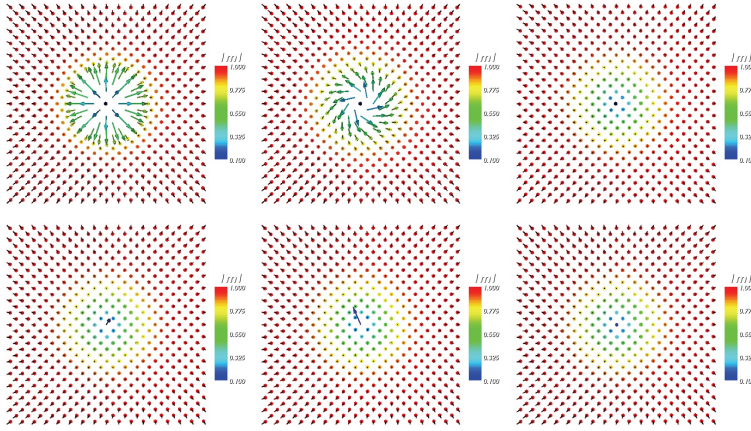


FIG. 4. Example 7.1: Magnetization at times $t = 0, 0.0068, 0.033, 0.073, 0.076, 0.1$ (from left to right, from top to bottom).

Snapshots of the magnetization at different time levels computed by the Scheme A can be found in Figure 4; the vectors are colored according to the modulus of the magnetization.

Note that the constraint $|\mathbf{M}_h^j(\mathbf{z})| = M_s(\mathbf{z})$ for \mathbf{M}_h^j computed by Scheme B was satisfied up to the error of the iterative algebraic solver for all nodes $\mathbf{z} \in \mathcal{N}_h$.

Example 7.2. We study the influence of a space-time varying $M_s(t, \mathbf{x})$ on the evolution of the blow-up example. The temperature profile is defined as

$$\tau(t, \mathbf{x}) = \begin{cases} 0.99 e^{B|\mathbf{x}^*|^2} e^{A_1(t-t_{\max})^2} & t < t_{\max}, \\ 0.99 e^{B|\mathbf{x}^*|^2} e^{A_2(t-t_{\max})^2} & t > t_{\max}, \end{cases}$$

with $B = -25$, $A_1 = -100000$, $A_2 = -10000$, and $t_{\max} = 0.02$. The temperature attains its maximum at time $t = t_{\max}$.

The results for both Schemes A and B are displayed in Figure 5 (left) and (right), respectively. In the case of time varying M_s , the monotonic decrease of the discrete energy is not to be expected as can be seen from the graphs. The results are similar for both methods. The effects of the temperature on the magnetization cause a slight delay of the blow-up for $h = 1/16, 1/32$, when compared to $M_s \equiv 1$ (Figure 2 (right)). The difference is quite significant for $h = 1/64$, when the blow-up occurs earlier than in the case of constant M_s .

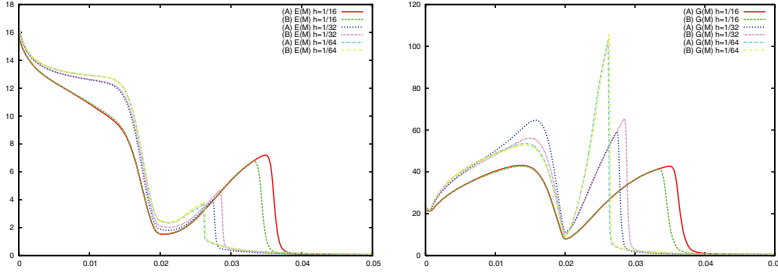


FIG. 5. *Example 7.2: Plot of $t_j \mapsto E(\mathbf{M}_h^j)$ for Schemes A,B ($h = \frac{1}{16}, \frac{1}{32}, \frac{1}{64}$) (left). Example 7.2: Plot of $t_j \mapsto \|\nabla \mathbf{M}_h^j\|_{L^\infty}$ for Schemes A,B ($h = \frac{1}{16}, \frac{1}{32}, \frac{1}{64}$) (right).*

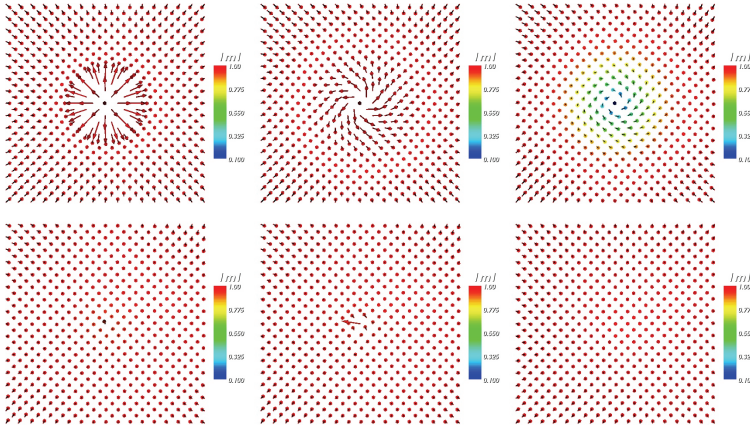


FIG. 6. *Example 7.2: Magnetization at times $t = 0, 0.0135, 0.02, 0.0345, 0.0356, 0.05$ (from left to right, from top to bottom).*

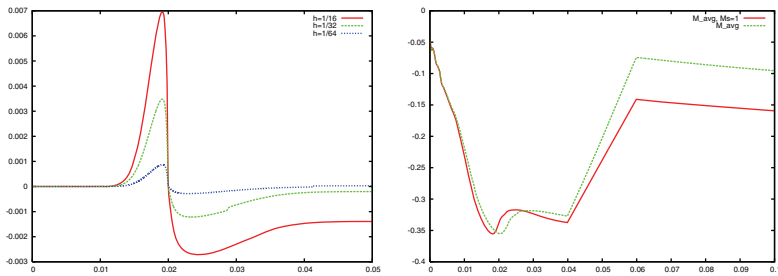


FIG. 7. *Example 7.2: Plot of $t_j \mapsto \max_{\mathbf{z} \in \mathcal{N}_h} \|\mathbf{M}_h^j(\mathbf{z}) - M_s(t_j, \mathbf{z})\|$ for Scheme B ($h = \frac{1}{16}, h = \frac{1}{32}, h = \frac{1}{64}$) (left). Example 7.3: The evolution of M_y^{avg} with and without temperature effects (right).*

Snapshots of the magnetization at different time levels computed by the Scheme A can be found in Figure 6; the vectors are colored according to the modulus of the magnetization.

In the case of time-varying M_s , Scheme B no longer satisfies the constraint $\|\mathbf{M}_h^j(\mathbf{z})\| = M_s(t_j, \mathbf{z})$. This is because of the time integration error due to the discretization of the function κ in time and the nonconsistency of (5.3) with (1.10). In Figure 7 (left) we depict the deviation in the magnitude of the magnetization from

the prescribed values; i.e., $\max_{\mathbf{z} \in \mathcal{N}_h} \|\mathbf{M}_h^j(\mathbf{z}) - M_s(t_j, \mathbf{z})\|$. We observe that the error decreases with decreasing time step.

Example 7.3. The following example is to demonstrate the combined effects of a constant applied magnetic field and temperature on the evolution of the magnetization; i.e., we consider equation (1.1) with $\mathbf{h}_{eff} = \mathbf{h}_{app} + \Delta \mathbf{m}$.

Assume that the temperature solves the linear heat equation,

$$(7.1) \quad \begin{aligned} \tau_t - K \Delta \tau &= f & \text{in} & \quad \Omega, \\ \partial_{\mathbf{n}} \tau &= 0 & \text{on} & \quad \partial \Omega_T, \\ \tau(0, \mathbf{x}) &= \tau_0(\mathbf{x}), \end{aligned}$$

where $f \equiv f(t, \mathbf{x})$ is a prescribed function to represent the heating source (e.g., a laser beam) and $K > 0$ is a constant, which depends on the thermal conductivity, the density, and the heat capacity of the material. We take the initial condition $\tau_0 \equiv 0$, $\alpha \equiv 1$, and f as follows:

$$f(t, \mathbf{x}) = \begin{cases} \begin{cases} 1 & |\mathbf{x}^*| < 0.5, \\ 0, & \text{elsewhere,} \end{cases} & t \leq 0.04 \\ 0 & t > 0.04. \end{cases}$$

After we heat the domain we apply for short time a constant magnetic field in the direction of the y -axes. This is represented by the constant vector field \mathbf{h}_{app} which is defined as

$$\mathbf{h}_{app}(t) = \begin{cases} (0, 10, 0) & \text{for } 0.04 \leq t \leq 0.06, \\ (0, 0, 0) & \text{for } t < 0.04 \text{ or } t > 0.06. \end{cases}$$

We compute the temperature from (7.1) by using backward Euler discretization in time, and standard H^1 -conforming linear finite elements in space. We denote the computed temperature at time t_j by $\tau_h^j(\cdot) \approx \tau(t_j, \cdot)$.

This example was computed using Scheme B, since the projection scheme involves to evaluate ΔM_s , which is not directly available when solving (7.1) by linear finite elements. The extension of Scheme B to the case $\mathbf{h}_{eff} = \mathbf{h}_{app} + \Delta \mathbf{m}$ is straightforward since \mathbf{h}_{app} is a constant vector field. The coefficient κ_{j+1} in Scheme B was computed in the following way:

$$\kappa_{j+1} = -\frac{\beta}{k} \left(\frac{\tau_h^{j+1} - \tau_h^j}{\tau_C - \tau_h^{j+1}} \right),$$

i.e., we use $(\tau_h)_t(t_{j+1}, \cdot) \approx \frac{\tau_h^{j+1} - \tau_h^j}{k}$.

We study the influence of the applied magnetic field \mathbf{h} on the evolution of the magnetization. For this purpose we define the function

$$(7.2) \quad M_y^{avg} = \frac{1}{\#\mathcal{N}_h} \sum_{\mathbf{z} \in \mathcal{N}_h} \frac{[\mathbf{M}_h(\mathbf{z})]_y}{|\mathbf{M}_h(\mathbf{z})|},$$

with $\mathbf{M}_h = ([\mathbf{M}_h]_x, [\mathbf{M}_h]_y, [\mathbf{M}_h]_z)$. The quantity $M_y^{avg} \in [-M_s, M_s]$ represents the “relative alignment” of the magnetization in the direction of the applied field; i.e., the greater the effect of the applied field on the magnetization is, the greater is the value of M_y^{avg} .

We display the time evolution of M_y^{avg} in Figure 7 (right) for the problem with temperature effects, and for $M_s \equiv 1$. The results indicate that the applied magnetic

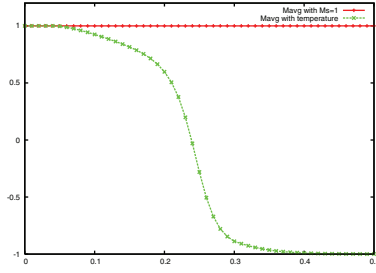


FIG. 8. Example 7.4: The evolution of M_y^{avg} with and without temperature effects.

field has greater effects on the evolution of the magnetization when temperature is taken into account.

Example 7.4. The last example illustrates the effect of temperature on the magnetization switching by a constant applied field in materials with high anisotropy. We take

$$\mathbf{h}_{eff} = K(\mathbf{m} \cdot \mathbf{p})\mathbf{p} + \mathbf{h}_{app} + \Delta\mathbf{m},$$

where K is the anisotropy constant and $\mathbf{p} = \mathbf{e}_\perp$ is a unit vector of anisotropy direction (with $\mathbf{e} \in \mathbb{S}^2$ the easy axis). The temperature is uniform in space and only varies in time as

$$\tau(t) = \begin{cases} 0.9999 e^{-10000(t-0.04)^2} & t < 0.04, \\ 0.9999 & t > 0.04. \end{cases}$$

We take $K = -1000$ and $\mathbf{p} = (1, 0, 0)$. The choice of $K < 0$ results in an in-plane anisotropy; i.e., at the steady state magnetization lies in the plane perpendicular to the vector \mathbf{p} , in the present case the yz -plane. The magnetization is initially aligned in the y -direction, i.e.,

$$\mathbf{m}_0 = (0, 1, 0).$$

At time $t = 0.04$, after the temperature attains its maximum, a constant magnetic field is applied for a period of time, in order to switch the magnetization to the opposite direction. This is modeled by the following choice of \mathbf{h}_{app} :

$$\mathbf{h}_{app}(t) = \begin{cases} (20, -20, 0) & 0.04 < t < 0.3, \\ (0, 0, 0) & \text{elsewhere.} \end{cases}$$

The example was computed using the Scheme B. The fixed-point algorithm (6.1) is adopted to include the additional anisotropy term; this is done by simply adding the linearized term $\gamma(1 + \delta^2)\mathbf{W}_h^{l+1} \times [K(\mathbf{W}_h^l \cdot \mathbf{p})\mathbf{p}]$ to the left-hand side of (6.1).

In Figure 8 we display the graphs of the computed evolutions of M_y^{avg} from (7.2) for the case with temperature effects and with constant $M_s = 1$. We observe that the applied field \mathbf{h}_{app} has no effect on the magnetization for the problem with constant M_s . On the other hand, when the effects of the temperature are included, the magnetization is reversed to the $(0, -1, 0)$ direction. The computed results support the physical observations from magnetic recording applications that lower strength of the magnetic field is needed for writing if local heating is employed. The practical need for lower writing fields leads to potential improvements in the storage density, faster writing times, and higher reliability of magnetic storage; cf. [21].

Acknowledgment. The project started when A.P. visited Ghent University in February 2006, and was finished when L.B. and M.S. visited Universität Tübingen in April and July 2007, respectively. The hospitality of the hosting institutes is gratefully acknowledged. L.B. was supported by the EPSRC grant EP/C548973/1 and M.S. was supported by the BOF/GOA-project no. 01G00607, Ghent University, Belgium.

REFERENCES

- [1] A. AHARONI, *Introduction to the Theory of Ferromagnetism*, Oxford University Press, Oxford, 1996.
- [2] F. ALOUGES AND A. SOYEUR, *On global weak solutions for Landau-Lifshitz equation: Existence and nonuniqueness*, *Nonlinear Anal., Theory, Meth. Appl.*, 18 (1992), pp. 1071–1084.
- [3] F. ALOUGES AND P. JAISSON, *Convergence of a finite elements discretization for the Landau-Lifshitz equations*, *Math. Models Methods Appl. Sci.*, 16 (2006), pp. 299–316.
- [4] Ľ. BAÑAS, *Numerical methods for the Landau-Lifshitz-Gilbert equation*, *Lect. Notes Comput. Sci.*, 3401 (2005), pp. 158–165.
- [5] Ľ. BAÑAS, S. BARTELS, AND A. PROHL, *A convergent implicit discretization of the Maxwell-Landau-Lifshitz-Gilbert equation*, *SIAM J. Numer. Anal.*, 46 (2008), pp. 1399–1422.
- [6] S. BARTELS AND A. PROHL, *Convergence of an implicit finite element method for the Landau-Lifshitz-Gilbert equation*, *SIAM J. Numer. Anal.*, 44 (2006), pp. 1405–1419.
- [7] S. BARTELS AND A. PROHL, *Constraint preserving implicit finite element discretization of harmonic map flow into spheres*, *Math. Comp.*, 76 (2007), pp. 1847–1859.
- [8] G. BERTOTTI, *Hysteresis in Magnetism*, Academic Press, San Diego, CA, 1998.
- [9] W.F. BROWN, *Magnetostatic Principles in Ferromagnetism*, Springer, New York, 1966.
- [10] G. CARBOU AND P. FABRIE, *Regular solutions for Landau-Lifshitz equation in a bounded domain*, *J. Differential Integral Equations*, 3 (2001), pp. 213–229.
- [11] Y. CHEN AND M.C. HONG, *Existence and partial regularity results for the heat flow for harmonic maps*, *Math. Z.*, 201 (1989), pp. 83–103.
- [12] I. CIMRÁK, *Error estimates for a semi-implicit numerical scheme solving the Landau-Lifshitz equation with an exchange field*, *IMA J. Numer. Anal.*, 25 (2005), pp. 611–634.
- [13] A. HUBERT AND R. SCHÄFER, *Magnetic Domains*, Springer, Berlin, 1998.
- [14] T.W. MCDANIEL, *Ultimate limits to thermally assisted magnetic recording*, *J. Phys. Condens. Matter*, 17 (2005), pp. R315–332.
- [15] W. E AND X.-P. WANG, *Numerical methods for the Landau-Lifshitz equation*, *SIAM J. Numer. Analysis*, 38 (2000), pp. 1647–1665.
- [16] V. YU, M.A. EGOROV, R.V. SHUBIN, (EDS.), GAMKRELIDZE, *Partial differential equations I. Foundations of the classical theory. (Transl. from the Russian by R. Cooke*, in *Encyclopedia of Mathematical Sciences*, 30, Springer, Berlin, 1992.
- [17] L.C. EVANS, *Partial differential equations*, AMS, Providence, RI, 1998.
- [18] B. GUO AND M.-C. HONG, *The Landau-Lifshitz equation of the ferromagnetic spin chain and harmonic maps*, *Calc. Var.*, 1 (1993), pp. 311–334.
- [19] M. KRUŽÍK AND A. PROHL, *Recent developments in modeling, analysis and numerics of ferromagnetism*, *SIAM Rev.*, 48 (2006), pp. 439–483.
- [20] A. LYBERATOS AND K.Y. GUSLIENKO, *Thermal stability of the magnetization following thermomagnetic writing in perpendicular media*, *J. Appl. Phys.*, 94 (2003), pp. 1119–1129.
- [21] K. MATSUMOTO, A. INOMATA, AND S. HASEGAWA, *Thermally Assisted Magnetic Recording*, *Fujitsu Sci. Tech. J.*, 42 (2006), pp. 158–167.
- [22] A. PROHL, *Computational Micromagnetism*, Teubner, Leipzig, 2001.
- [23] J. RUIGROK, *Thermally assisted magnetodynamics*, slides, downloadable at: www.evsf2.science.ru.nl/masterclass/Amsterdam-Masterclass/Ruigrok.pdf
- [24] A. SCHMIDT AND K.G. SIEBERT, *ALBERT—software for scientific computations and applications*, *Acta Math. Univ. Comenian. (N.S.)*, 70 (2000), pp. 105–122.
- [25] A.W. SPARGO, *Finite element analysis of magnetization reversal in granular thin films*, Ph.D. thesis, U. Wales, Bangor, 2002.
- [26] M. STRUWE, *On the evolution of harmonic maps of Riemannian surfaces*, *Comm. Math. Helv.*, 60 (1985), pp. 558–581.
- [27] M. STRUWE, *Geometric evolution problems*, *IAS/Park City Math. Series*, 2 (1996), pp. 259–339.
- [28] J.U. THIELE, K.R. COFFEY, M.F. TONEY, J.A. HEDSTROM, AND A.J. KELLOCK, *Temperature dependent magnetic properties of highly chemically ordered $Fe_{55-x}Ni_xL_{10}$ films*, *J. Appl. Phys.*, 91 (2002), pp. 6595–6600.
- [29] webpage address: <http://www.ctcms.nist.gov/~rdm/mumag.org.html>.

LOCAL ANISOTROPIC INTERPOLATION ERROR ESTIMATES BASED ON DIRECTIONAL DERIVATIVES ALONG EDGES*

U. HETMANIUK[†] AND P. KNUPP[†]

Abstract. We present new local anisotropic error estimates for the Lagrangian finite element interpolation. The bounds apply to affine equivalent elements and use information from directional derivatives of the function to interpolate along a set of adjacent edges. These new bounds do not require any geometric limitation but may vary, in some cases, with the node ordering. Several existing results are recovered from the new bounds. Examples compare the asymptotic behavior of the new and existing bounds when the diameter of the element goes to zero. For some elements with small or large angles, our new bound exhibits the same asymptotic behavior as the norm of the interpolation error while existing results do not have the correct asymptotic behavior.

Key words. affine equivalent elements, anisotropic estimates, finite element interpolation error, Lagrange interpolation

AMS subject classifications. 65N15, 65N30

DOI. 10.1137/060666524

1. Introduction. Interpolation operators, which associate a function v with a function in the finite element space, are key components for the analysis of the finite element method. In particular, measuring the interpolation error is a crucial step to prove the convergence of the finite element method.

For nodal Lagrangian finite elements, the first bounds of the interpolation error appeared in the 1970s. These early estimates focused on bounding seminorms of the interpolation error over one element K with shape and size characteristics of K and with the norm of a Fréchet derivative for the function v . Examples are given by Babuška and Aziz [2], Ciarlet [8], Jamet [14], Zlámal [18], and in the references cited therein. Such bounds are called isotropic as they treat equally the partial derivatives of v .

However, engineering applications often generate solutions with different scales of variations along different directions. For these problems, it is important to derive bounds, called anisotropic bounds, which take into account these variations. Early works studied the relation between the error when approximating a quadratic function by a linear polynomial and the geometric characteristics of a simplex (see, for example, Nadler [15, 16] and Rippa [17]). For the nodal Lagrangian linear interpolation, Berzins [4] derived an exact expression for the L^2 -norm of the linear interpolation error on a tetrahedron in terms of directional derivatives along the edges. Bank and Smith [3] studied the $W^{1,2}$ seminorm of the linear interpolation error on a triangle. For convex quadratic functions, D’Azevedo and Simpson [9] and Chen [7] derived similar expressions for, respectively, the L^∞ -norm on a triangle and the L^p -norm on a simplex.

For an arbitrary function v , Apel [1] studied local anisotropic interpolation error estimates that treat distinctly partial derivatives of v along coordinate directions.

*Received by the editors August 1, 2006; accepted for publication (in revised form) September 2, 2008; published electronically December 31, 2008. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000.

<http://www.siam.org/journals/sinum/47-1/66652.html>

[†]Sandia National Laboratories, P.O. Box 5800, MS 1320, Albuquerque, NM 87185-1320 (ulhetma@sandia.gov, pknupp@sandia.gov).

Unfortunately, his results require a priori geometric limitations on the element: a limitation on the maximum angle and a coordinate system condition. For triangles, this coordinate system condition limits the angle between the longest side and the x_1 -axis. Chen, Sun, and Xu [6] presented optimal linear interpolation error estimates in L^p -norm for simplicial meshes that satisfy a geometric assumption; i.e., they are quasi-uniform under a metric based on the Hessian of v . Formaggia and Perotto [10], Georgoulis, Hall, and Houston [12], and Huang [13] presented local anisotropic bounds that do not require any geometric limitation. However, for right-angled triangles with a small angle or for rectangles with a large aspect ratio, their estimates do not have the correct asymptotic behavior when the element diameter goes to zero.

Therefore, the goal of this paper is to present new local anisotropic bounds that do not require any geometric limitation and to assess their asymptotic behavior when the diameter of the element goes to zero. We develop our results for affine equivalent elements and for Lagrangian finite elements. Section 2 describes the notations used in this paper. Section 3 presents the new bounds and their proof. Finally, in section 4, we compare the new estimates with existing results and we assess the asymptotic behavior of these bounds when the diameter of the element goes to zero.

2. Notation. Throughout this paper, we adopt the following notations.

- \mathbb{R} is the set of real numbers.
- $d = 2$ or 3 denotes the space dimension.
- (x_1, \dots, x_d) denotes a global Cartesian coordinate system and the vectors $(\mathbf{e}_1, \dots, \mathbf{e}_d)$ are the associated unit basis vectors.
- \mathbf{x} the vector position in \mathbb{R}^d .
- For each integer $k \geq 0$, \mathbb{P}_k is the space of polynomials of degree smaller than k in (x_1, \dots, x_d) . \mathbb{Q}_k is the space of all polynomials of degree smaller than k with respect to each one of the d variables (x_1, \dots, x_d) .
- For any integer $m > 0$, $\alpha = (\alpha_1, \dots, \alpha_m)$ is a multi-index where each index α_i belongs to $\{1, 2, \dots, d\}$. I_m is the set of multi-indices $\{1, 2, \dots, d\}^m$.
- For any smooth function v defined in a domain $\Omega \subset \mathbb{R}^d$, Dv denotes the first Fréchet derivative and $D^m v$ the Fréchet derivative of order m . $D^m v \cdot (\mathbf{w}_1, \dots, \mathbf{w}_m)$ is the partial derivative of order m in the direction of the m vectors $\mathbf{w}_1, \dots, \mathbf{w}_m$ and, at a point $\mathbf{x} \in \Omega$, its value is denoted $D^m v(\mathbf{x}) \cdot (\mathbf{w}_1, \dots, \mathbf{w}_m)$.
- For $1 \leq p < \infty$, $L^p(\Omega)$ is the Sobolev space of real functions whose p -power is integrable over Ω . Each space $L^p(\Omega)$ is endowed with the usual norm, denoted $\|v\|_{0,p,\Omega}$. $L^\infty(\Omega)$ is the Sobolev space of essentially bounded real functions. We denote by $\|v\|_{0,\infty,\Omega}$ the usual norm in $L^\infty(\Omega)$.
- For $m > 0$ and $1 \leq p \leq \infty$, the spaces $W^{m,p}(\Omega)$ contain functions whose weak derivatives up to order m are in $L^p(\Omega)$. Each space $W^{m,p}(\Omega)$ is endowed with the norm $\|v\|_{m,p,\Omega}$ defined by

$$\left(\int_{\Omega} |v(\mathbf{x})|^p + \sum_{j=1}^m \sum_{\alpha \in I_j} |D^j v(\mathbf{x}) \cdot (\mathbf{e}_{\alpha_1}, \dots, \mathbf{e}_{\alpha_j})|^p d\mathbf{x} \right)^{\frac{1}{p}} \quad \text{for } 1 \leq p < \infty$$

and, when $p = \infty$,

$$\max \left(\sup_{\mathbf{x} \in \Omega} |v(\mathbf{x})|, \max_{1 \leq j \leq m} \max_{\alpha \in I_j} \sup_{\mathbf{x} \in \Omega} |D^j v(\mathbf{x}) \cdot (\mathbf{e}_{\alpha_1}, \dots, \mathbf{e}_{\alpha_j})| \right).$$

- For $m > 0$, we will also use the seminorm $|v|_{m,p,\Omega}$ defined by

$$\left(\sum_{\alpha \in I_m} \int_{\Omega} |D^{\alpha} v(\mathbf{x}) \cdot (\mathbf{e}_{\alpha_1}, \dots, \mathbf{e}_{\alpha_m})|^p d\mathbf{x} \right)^{1/p} \quad \text{for } 1 \leq p < \infty$$

and, when $p = \infty$,

$$\max_{\alpha \in I_m} \sup_{\mathbf{x} \in \Omega} |D^{\alpha} v(\mathbf{x}) \cdot (\mathbf{e}_{\alpha_1}, \dots, \mathbf{e}_{\alpha_m})|.$$

- The letter C is used to denote a generic positive constant.

We consider the element \hat{K} to be either

- the unit right triangle: $\{(x_1, x_2) \in \mathbb{R}^2 \mid 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1 - x_1\}$,
- the unit square: $\{(x_1, x_2) \in \mathbb{R}^2 \mid 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1\}$,
- the unit right tetrahedron: $\{(x_1, x_2, x_3) \in \mathbb{R}^3 \mid 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1 - x_1, 0 \leq x_3 \leq 1 - x_1 - x_2\}$,
- the unit cube: $\{(x_1, x_2, x_3) \in \mathbb{R}^3 \mid 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1, 0 \leq x_3 \leq 1\}$.

We introduce now the Lagrangian finite element $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ such that

$$\begin{aligned} \hat{\mathcal{P}} &= \begin{cases} \mathbb{P}_k & \text{when } \hat{K} \text{ is a simplex} \\ \mathbb{Q}_k & \text{when } \hat{K} \text{ is a square or a cube} \end{cases} \\ \hat{\mathcal{N}} &= \left\{ \hat{N}_{i,\hat{K}}, 1 \leq i \leq n \mid \hat{N}_{i,\hat{K}}(\hat{v}) = \hat{v}(\hat{\mathbf{x}}_i) \right\}, \end{aligned}$$

where $(\hat{\mathbf{x}}_i)_{1 \leq i \leq n}$ is the classical set of nodes for \hat{K} . k defines the polynomial degree of the shape functions $(\hat{\phi}_{i,\hat{K}})_{1 \leq i \leq n}$. For any smooth function \hat{v} defined on \hat{K} , its nodal interpolant is defined by

$$(2.1) \quad \hat{\Pi} \hat{v} = \sum_{i=1}^n \hat{N}_{i,\hat{K}}(\hat{v}) \hat{\phi}_{i,\hat{K}}.$$

Given an invertible matrix \mathbf{A} of size d and a vector \mathbf{a} of length d , we define the affine-equivalent Lagrangian finite element $(K, \mathcal{P}, \mathcal{N})$, where the element K satisfies

$$(2.2) \quad K = \left\{ \mathbf{x} \in \mathbb{R}^d \mid \mathbf{x} = \mathbf{A}\hat{\mathbf{x}} + \mathbf{a}, \forall \hat{\mathbf{x}} \in \hat{K} \right\}.$$

Following the classical approach (see, for example, Ciarlet [8, section 2.3]), we introduce, for any smooth function v defined on K , its nodal interpolant

$$(2.3) \quad \Pi v = \sum_{i=1}^n N_{i,K}(v) \phi_{i,K}.$$

In the remainder of this document, any smooth function v defined on K is associated with a smooth function \hat{v} on \hat{K} , such that

$$\hat{v}(\hat{\mathbf{x}}) = v(\mathbf{A}\hat{\mathbf{x}} + \mathbf{a}).$$

3. New local anisotropic bounds. In this section, we derive the new local anisotropic bounds on one element. On an affine-equivalent element K , let $(\mathbf{b}_j)_{1 \leq j \leq d}$ denote a basis of \mathbb{R}^d composed of adjacent edge vectors of K . Several choices of basis are available. For instance, each vertex of a simplex is associated to a different basis.

A different ordering of edges results also in a different basis. For our new results, a different basis may result in a different anisotropic bound. So we will comment on the behavior of the upper bound when using different choices of adjacent edges.

First, the result on the L^q -norm of the interpolation error is stated. Its proof is given in section 3.1.

THEOREM 3.1. *Let v be a function on K belonging to $W^{l,p}(K)$. The integral index l and the real numbers p and q satisfy*

$$(3.1) \quad 1 \leq p, q \leq \infty, \quad 0 \leq l \leq k + 1.$$

We assume that the continuous embeddings hold

$$(3.2) \quad W^{l,p}(K) \hookrightarrow \mathcal{C}(K) \quad \text{and} \quad W^{l,p}(K) \hookrightarrow L^q(K)$$

Let Π be the nodal interpolation (2.3) to the finite element space with polynomials of degree k . Under these assumptions, the following anisotropic bounds hold. When p is finite, we have

$$(3.3) \quad \|v - \Pi v\|_{0,q,K} \leq C |K|^{\frac{1}{q} - \frac{1}{p}} \left[\sum_{\beta \in I_l} \int_K |D^l v(\mathbf{x}) \cdot (\mathbf{b}_{\beta_1}, \dots, \mathbf{b}_{\beta_l})|^p dx \right]^{\frac{1}{p}}.$$

When $p = \infty$, we have

$$(3.4) \quad \|v - \Pi v\|_{0,q,K} \leq C |K|^{\frac{1}{q}} \max_{\beta \in I_l} \left[\sup_{\mathbf{x} \in K} |D^l v(\mathbf{x}) \cdot (\mathbf{b}_{\beta_1}, \dots, \mathbf{b}_{\beta_l})| \right].$$

The constants C depend only on \hat{K} , l , p , q , and d .

The final bound is a simple reformulation of existing results from Ciarlet [8], Apel [1], Formaggia and Perotto [11], and Huang [13]. However, this formulation highlights the importance of directional derivatives along edges of the element. The constraint (3.2) is classical when working with the Lagrangian interpolation operator (see Ciarlet [8, Theorem 3.1.5] or Apel [1, section 2.1.1]). Besides the affine-equivalent assumption on K , this theorem does not require further geometric assumption on K . When using different choices of adjacent edges, the resulting upper bounds have the same asymptotic behavior when the diameter of K goes to zero.

Then, norms of partial derivatives of the interpolation error are bounded. The result is proved in section 3.2.

THEOREM 3.2. *Let $m > 0$ and α be in I_m . Let v be a continuous function belonging to $W^{l,p}(K)$, where $1 \leq p \leq \infty$ and $1 \leq l \leq k + 1$. Let $1 \leq q \leq \infty$, such that $W^{l-m,p}(K)$ has a continuous embedding into $L^q(K)$ and the following constraints are satisfied:*

$$(3.5a) \quad l > m \quad \text{when } \alpha \in \{\{1\}^m, \{2\}^m\} \quad \text{when } K \subset \mathbb{R}^2$$

$$(3.5b) \quad \begin{cases} l > m & \text{when } \alpha \in \{\{1, 2\}^m, \{1, 3\}^m, \{2, 3\}^m\} \\ p > 2 & \text{when } m = l - 1 \text{ and } \alpha \in \{\{1\}^m, \{2\}^m, \{3\}^m\} \end{cases} \quad \text{when } K \subset \mathbb{R}^3.$$

Let Π be the nodal interpolation (2.3) to the finite element space with polynomials of degree k . Let \mathbf{B} denote the matrix whose columns are the edge vectors $(\mathbf{b}_j)_{1 \leq j \leq d}$. The entries of matrix \mathbf{B}^{-1} are denoted $B_{i,j}^{(-1)}$.

Under these assumptions, the following anisotropic bounds hold. When p is finite, we have

$$(3.6) \quad \|D^m(v - \Pi v) \cdot (\mathbf{e}_{\alpha_1}, \dots, \mathbf{e}_{\alpha_m})\|_{0,q,K} \leq C |K|^{\frac{1}{q} - \frac{1}{p}} \left\{ \sum_{j_1, \dots, j_m=1}^d |B_{j_1, \alpha_1}^{(-1)}|^q \dots |B_{j_m, \alpha_m}^{(-1)}|^q \left[\sum_{\beta \in I_{l-m}} \int_K |D^l v(\mathbf{x}) \cdot (\mathbf{b}_{j_1}, \dots, \mathbf{b}_{j_m}, \mathbf{b}_{\beta_1}, \dots, \mathbf{b}_{\beta_{l-m}})|^p d\mathbf{x} \right]^{\frac{q}{p}} \right\}^{\frac{1}{q}}.$$

When $p = \infty$, we have

$$(3.7) \quad \|D^m(v - \Pi v) \cdot (\mathbf{e}_{\alpha_1}, \dots, \mathbf{e}_{\alpha_m})\|_{0,q,K} \leq C |K|^{\frac{1}{q}} \left\{ \sum_{j_1, \dots, j_m=1}^d |B_{j_1, \alpha_1}^{(-1)}|^q \dots |B_{j_m, \alpha_m}^{(-1)}|^q \max_{\beta \in I_{l-m}} \left[\sup_{\mathbf{x} \in K} |D^l v(\mathbf{x}) \cdot (\mathbf{b}_{j_1}, \dots, \mathbf{b}_{j_m}, \mathbf{b}_{\beta_1}, \dots, \mathbf{b}_{\beta_{l-m}})| \right]^q \right\}^{\frac{1}{q}}.$$

The constants C depend only on \hat{K} , l , m , p , q , and d .

Besides the affine-equivalent assumption on K , Theorem 3.2 does not require further geometric assumption on K . The constraints (3.5) are explained in the proof. When choosing a different ordering of the adjacent edges, the resulting upper bounds will be unchanged. When using different choices of adjacent edges or different node numbering, the resulting upper bounds may have different asymptotic behaviors when the diameter of K goes to zero. Such an example will be presented in section 4.

3.1. Proof of Theorem 3.1. We give the proof when p and q are finite. The other cases are proved in a similar fashion. The proof technique is classical and consists of deriving estimates on \hat{K} and of applying an affine coordinate transformation between K and \hat{K} . First, we recall bounds on \hat{K} in the following lemma.

LEMMA 3.3. *Let \hat{v} be a function on \hat{K} belonging to $W^{l,p}(\hat{K})$. The integral index l and the real numbers p and q satisfy*

$$(3.8) \quad 1 \leq p, q \leq \infty, \quad 0 \leq l \leq k + 1.$$

We assume that the continuous embeddings hold

$$(3.9) \quad W^{l,p}(\hat{K}) \hookrightarrow C(\hat{K}) \quad \text{and} \quad W^{l,p}(\hat{K}) \hookrightarrow L^q(\hat{K}).$$

Then the following bound holds:

$$(3.10) \quad \|\hat{v} - \hat{\Pi}\hat{v}\|_{0,q,\hat{K}} \leq C |\hat{v}|_{l,p,\hat{K}},$$

where the constant C depends only on l , p , q , and \hat{K} .

The proof for this lemma is given in Ciarlet [8, section 3.1] and Apel [1, Chapter 2]. We recall that the element \hat{K} is either the unit right triangle, the unit square, the unit right tetrahedron, or the unit cube. The constraints (3.2) are equivalent to the constraints (3.9). \mathbf{B} denotes the matrix whose columns are the edge vectors $(\mathbf{b}_j)_{1 \leq j \leq d}$. \mathbf{B} is the Jacobian matrix of an affine map between \hat{K} and K .

When q is finite, we have

$$\begin{aligned} \|v - \Pi v\|_{0,q,K} &= \left(\int_K |(v - \Pi v)(\mathbf{x})|^q d\mathbf{x} \right)^{\frac{1}{q}} \\ &= \left(\int_{\hat{K}} |(\hat{v} - \hat{\Pi} \hat{v})(\hat{\mathbf{x}})|^q |\det \mathbf{B}| d\hat{\mathbf{x}} \right)^{\frac{1}{q}} \\ &\leq C |\det \mathbf{B}|^{\frac{1}{q}} \left(\sum_{\beta \in I_l} \int_{\hat{K}} |D^\beta \hat{v}(\hat{\mathbf{x}}) \cdot (\mathbf{e}_{\beta_1}, \dots, \mathbf{e}_{\beta_l})|^p d\hat{\mathbf{x}} \right)^{\frac{1}{p}}, \end{aligned}$$

where we have used the result of Lemma 3.3. Using the differentiation rule for composition of functions (see Ciarlet [8, p. 118]), we note that

$$D^l \hat{v}(\hat{\mathbf{x}}) \cdot (\mathbf{e}_{\beta_1}, \dots, \mathbf{e}_{\beta_l}) = D^l v(\mathbf{x}) \cdot (\mathbf{b}_{\beta_1}, \dots, \mathbf{b}_{\beta_l}).$$

Consequently, we obtain

$$\|v - \Pi v\|_{0,q,\hat{K}} \leq C |\det \mathbf{B}|^{\frac{1}{q}} \left(\sum_{\beta \in I_l} \int_K |D^l v(\mathbf{x}) \cdot (\mathbf{b}_{\beta_1}, \dots, \mathbf{b}_{\beta_l})|^p |\det \mathbf{B}^{-1}| d\mathbf{x} \right)^{\frac{1}{p}},$$

and conclude by using the relation

$$|\det \mathbf{B}| = |K| / |\hat{K}|.$$

3.2. Proof of Theorem 3.2. We give the proof when p and q are finite. The other cases are proved in a similar fashion. As previously, we start with local anisotropic bounds on \hat{K} and, then, apply an affine coordinate transformation between K and \hat{K} .

LEMMA 3.4. *Let α be in I_m and \hat{v} be a continuous function on \hat{K} satisfying $D^m \hat{v} \cdot (\mathbf{e}_{\alpha_1}, \dots, \mathbf{e}_{\alpha_m}) \in W^{l-m,p}(\hat{K})$. The indices l, m , and p satisfy*

$$(3.11) \quad 1 \leq p \leq \infty, \quad l, m \in \mathbb{N}, \quad 0 < m \leq l \leq k + 1,$$

and the constraints defined by

$$(3.12a) \quad l > m \quad \text{when } \alpha \in \{\{1\}^m, \{2\}^m\} \quad \text{when } \hat{K} \subset \mathbb{R}^2$$

$$(3.12b) \quad \begin{cases} l > m & \text{when } \alpha \in \{\{1, 2\}^m, \{1, 3\}^m, \{2, 3\}^m\} \\ p > 2 & \text{when } m = l - 1 \text{ and } \alpha \in \{\{1\}^m, \{2\}^m, \{3\}^m\} \end{cases} \quad \text{when } \hat{K} \subset \mathbb{R}^3.$$

For $1 \leq q \leq \infty$, such that $W^{l-m,p}(\hat{K})$ has a continuous embedding into $L^q(\hat{K})$, the following bound holds:

$$(3.13) \quad \left\| D^m(\hat{v} - \hat{\Pi} \hat{v}) \cdot (\mathbf{e}_{\alpha_1}, \dots, \mathbf{e}_{\alpha_m}) \right\|_{0,q,\hat{K}} \leq C |D^m \hat{v} \cdot (\mathbf{e}_{\alpha_1}, \dots, \mathbf{e}_{\alpha_m})|_{l-m,p,\hat{K}},$$

where the constant C depends only on l, m, p, q , and \hat{K} .

These bounds were proved by Apel [1] (see Lemmas 2.4, 2.6, 2.10, and 2.18, respectively, for triangles, tetrahedra, quadrilaterals, and hexahedrals). We emphasize that these bounds are on \hat{K} and that \hat{K} is either the unit right triangle, the unit square, the unit right tetrahedron, or the unit cube. In his book, Apel comments about the necessity of the constraints (3.12) to derive anisotropic bounds on \hat{K} . The constraints (3.5) match the constraints (3.12).

We represent the vectors \mathbf{e}_j in the basis of edge vectors $(\mathbf{b}_i)_{1 \leq i \leq d}$,

$$\mathbf{e}_j = \sum_{i=1}^d B_{i,j}^{(-1)} \mathbf{b}_i,$$

where the coefficients $B_{i,j}^{(-1)}$ are the entries of matrix \mathbf{B}^{-1} .

When q is finite, we have

$$\begin{aligned} & \|D^m(v - \Pi v) \cdot (\mathbf{e}_{\alpha_1}, \dots, \mathbf{e}_{\alpha_m})\|_{0,q,K} \\ &= \left(\int_K \left| \sum_{j_1, \dots, j_m=1}^d B_{j_1, \alpha_1}^{(-1)} \dots B_{j_m, \alpha_m}^{(-1)} D^m(v - \Pi v)(\mathbf{x}) \cdot (\mathbf{b}_{j_1}, \dots, \mathbf{b}_{j_m}) \right|^q d\mathbf{x} \right)^{\frac{1}{q}}. \end{aligned}$$

We recall the generalized mean inequality for a set of real positive numbers $(a_i)_{1 \leq i \leq N}$:

$$\left(\sum_{i=1}^N a_i \right)^q \leq N^{q-1} \left(\sum_{i=1}^N a_i^q \right).$$

Consequently, there exists a constant $C = d^{\frac{m(q-1)}{q}}$ such that

$$\begin{aligned} & \|D^m(v - \Pi v) \cdot (\mathbf{e}_{\alpha_1}, \dots, \mathbf{e}_{\alpha_m})\|_{0,q,K} \\ & \leq C \left(\sum_{j_1, \dots, j_m=1}^d |B_{j_1, \alpha_1}^{(-1)}|^q \dots |B_{j_m, \alpha_m}^{(-1)}|^q \int_K |D^m(v - \Pi v)(\mathbf{x}) \cdot (\mathbf{b}_{j_1}, \dots, \mathbf{b}_{j_m})|^q d\mathbf{x} \right)^{\frac{1}{q}}. \end{aligned}$$

The differentiation rule for composition of functions gives

$$D^m(v - \Pi v)(\mathbf{x}) \cdot (\mathbf{b}_{j_1}, \dots, \mathbf{b}_{j_m}) = D^m(\hat{v} - \hat{\Pi}\hat{v})(\hat{\mathbf{x}}) \cdot (\mathbf{e}_{j_1}, \dots, \mathbf{e}_{j_m}).$$

Using a change of variables, we obtain

$$\begin{aligned} & \|D^m(v - \Pi v) \cdot (\mathbf{e}_{\alpha_1}, \dots, \mathbf{e}_{\alpha_m})\|_{0,q,K} \leq C |\det \mathbf{B}|^{\frac{1}{q}} \\ & \left(\sum_{j_1, \dots, j_m=1}^d |B_{j_1, \alpha_1}^{(-1)}|^q \dots |B_{j_m, \alpha_m}^{(-1)}|^q \int_K |D^m(\hat{v} - \hat{\Pi}\hat{v})(\hat{\mathbf{x}}) \cdot (\mathbf{e}_{j_1}, \dots, \mathbf{e}_{j_m})|^q d\mathbf{x} \right)^{\frac{1}{q}}. \end{aligned}$$

The result of Lemma 3.4 now implies

$$\begin{aligned} & \|D^m(v - \Pi v) \cdot (\mathbf{e}_{\alpha_1}, \dots, \mathbf{e}_{\alpha_m})\|_{0,q,K} \leq C |\det \mathbf{B}|^{\frac{1}{q}} \\ & \left(\sum_{j_1, \dots, j_m=1}^d |B_{j_1, \alpha_1}^{(-1)}|^q \dots |B_{j_m, \alpha_m}^{(-1)}|^q |D^m \hat{v} \cdot (\mathbf{e}_{j_1}, \dots, \mathbf{e}_{j_m})|_{l-m,p,\hat{K}}^q \right)^{\frac{1}{q}}. \end{aligned}$$

The seminorm,

$$|D^m \hat{v} \cdot (\mathbf{e}_{j_1}, \dots, \mathbf{e}_{j_m})|_{l-m,p,\hat{K}}^q = \left(\sum_{\beta \in I_{l-m}} \int_{\hat{K}} |D^m \hat{v}(\hat{\mathbf{x}}) \cdot (\mathbf{e}_{j_1}, \dots, \mathbf{e}_{j_m}, \mathbf{e}_{\beta_1}, \dots, \mathbf{e}_{\beta_{l-m}})|^p d\hat{\mathbf{x}} \right)^{\frac{q}{p}},$$

becomes, after a change of variables,

$$|D^m \hat{v} \cdot (\mathbf{e}_{j_1}, \dots, \mathbf{e}_{j_m})|_{l-m,p,\hat{K}}^q = |\det \mathbf{B}|^{-\frac{q}{p}} \left(\sum_{\beta \in I_{l-m}} \int_K |D^m v(\mathbf{x}) \cdot (\mathbf{b}_{j_1}, \dots, \mathbf{b}_{j_m}, \mathbf{b}_{\beta_1}, \dots, \mathbf{b}_{\beta_{l-m}})|^p d\mathbf{x} \right)^{\frac{q}{p}}.$$

We conclude as previously

$$\begin{aligned} & \|D^m(v - \Pi v) \cdot (\mathbf{e}_{\alpha_1}, \dots, \mathbf{e}_{\alpha_m})\|_{0,q,K} \\ & \leq C |K|^{\frac{1}{q} - \frac{1}{p}} \left\{ \sum_{j_1, \dots, j_m=1}^d |B_{j_1, \alpha_1}^{(-1)}|^q \dots |B_{j_m, \alpha_m}^{(-1)}|^q \right. \\ & \left. \left[\sum_{\beta \in I_{l-m}} \int_K |D^l v(\mathbf{x}) \cdot (\mathbf{b}_{j_1}, \dots, \mathbf{b}_{j_m}, \mathbf{b}_{\beta_1}, \dots, \mathbf{b}_{\beta_{l-m}})|^p d\mathbf{x} \right]^{\frac{q}{p}} \right\}^{\frac{1}{q}}. \end{aligned}$$

Lemma 3.4 was proved only when \hat{K} is the unit right triangle, the unit square, the unit right tetrahedron, or the unit cube. We were not able to generalize Lemma 3.4 to the unit equilateral triangle or the unit equilateral tetrahedron. As the unit right triangle and tetrahedron are not invariant with respect to node numbering, the resulting bound on the affine-equivalent element K may vary with respect to node numbering. Further work would be required to guarantee such an invariance on simplices or to characterize the basis $(\mathbf{b}_j)_{1 \leq j \leq d}$ giving the sharpest upper bound.

3.3. Expression for particular cases. In this section, we write the new estimates for the $W^{m,q}$ seminorm and for square integrable functions. In particular, we present the expression for the L^2 -norm and for the $W^{1,2}$ seminorm.

Expression for the $W^{m,q}$ seminorm. We bound the $W^{m,q}$ seminorm by combining the results of Theorem 3.2 for all the partial derivatives of order m .

COROLLARY 3.5. *Let v be a continuous function belonging to $W^{l,p}(K)$, where $1 \leq p \leq \infty$ and $1 \leq l \leq k + 1$. Let $0 < m < l$ and $1 \leq q \leq \infty$, such that $W^{l-m,p}(K)$ has a continuous embedding into $L^q(K)$ and the following constraints are satisfied:*

(3.14a) $p > 2$ if $l = 1$ when $K \subset \mathbb{R}^2$,

(3.14b) $p > 2$ if $m = l - 1$ when $K \subset \mathbb{R}^3$.

Under these assumptions, the following anisotropic bounds hold. When p is finite, we have

$$(3.15) \quad |v - \Pi v|_{m,q,K} \leq C |K|^{\frac{1}{q} - \frac{1}{p}} \left\{ \sum_{\alpha \in I_m} \sum_{j_1, \dots, j_m=1}^d \left| B_{j_1, \alpha_1}^{(-1)} \right|^q \cdots \left| B_{j_m, \alpha_m}^{(-1)} \right|^q \left[\sum_{\beta \in I_{l-m}} \int_K |D^l v(\mathbf{x}) \cdot (\mathbf{b}_{j_1}, \dots, \mathbf{b}_{j_m}, \mathbf{b}_{\beta_1}, \dots, \mathbf{b}_{\beta_{l-m}})|^p d\mathbf{x} \right]^{\frac{q}{p}} \right\}^{\frac{1}{q}}.$$

When $p = \infty$, we have

$$(3.16) \quad |v - \Pi v|_{m,q,K} \leq C |K|^{\frac{1}{q}} \left\{ \sum_{\alpha \in I_m} \sum_{j_1, \dots, j_m=1}^d \left| B_{j_1, \alpha_1}^{(-1)} \right|^q \cdots \left| B_{j_m, \alpha_m}^{(-1)} \right|^q \max_{\beta \in I_{l-m}} \left[\sup_{\mathbf{x} \in K} |D^l v(\mathbf{x}) \cdot (\mathbf{b}_{j_1}, \dots, \mathbf{b}_{j_m}, \mathbf{b}_{\beta_1}, \dots, \mathbf{b}_{\beta_{l-m}})| \right]^q \right\}^{\frac{1}{q}}.$$

The constants C depend only on \hat{K} , l , m , p , q , and d .

The proof consists of summing the bounds (3.6) for all the multi-indices α belonging to I_m and of combining all the associated constraints (3.5).

Expression for the L^2 -norm in \mathbb{R}^2 . For the L^2 -norm, we set $m = 0$, $p = q = 2$, and $l = 2$. We have for any continuous function v belonging to $W^{2,2}(K)$:

$$(3.17) \quad \|v - \Pi v\|_{0,2,K}^2 \leq C \left(\int_K |D^2 v(\mathbf{x}) \cdot (\mathbf{b}_1, \mathbf{b}_1)|^2 d\mathbf{x} + \int_K |D^2 v(\mathbf{x}) \cdot (\mathbf{b}_2, \mathbf{b}_2)|^2 d\mathbf{x} + 2 \int_K |D^2 v(\mathbf{x}) \cdot (\mathbf{b}_1, \mathbf{b}_2)|^2 d\mathbf{x} \right).$$

In matrix notation, the bound becomes

$$\|v - \Pi v\|_{0,2,K}^2 \leq C \int_K \|\mathbf{B}^T D^2 v(\mathbf{x}) \mathbf{B}\|_F^2 d\mathbf{x},$$

where $D^2 v(\mathbf{x})$ denotes the Hessian matrix for v and $\|\cdot\|_F$ is the Frobenius norm. We note that the formula in matrix notation also holds in \mathbb{R}^3 . This anisotropic bound is well known and appeared, for instance, in the case of simplices, in Formaggia and Perotto [11] and in Huang [13].

Expression for the $W^{1,2}$ seminorm in \mathbb{R}^2 . For the $W^{1,2}$ seminorm, we set $m = 1$, $p = q = 2$, and $l = 2$. For any continuous function v in $W^{2,2}(K)$, we have

$$\begin{aligned} & |v - \Pi v|_{1,2,K}^2 \\ & \leq C \left[\left(\left| B_{1,1}^{(-1)} \right|^2 + \left| B_{1,2}^{(-1)} \right|^2 \right) \left(\int_K |D^2 v(\mathbf{x}) \cdot (\mathbf{b}_1, \mathbf{b}_1)|^2 d\mathbf{x} + \int_K |D^2 v(\mathbf{x}) \cdot (\mathbf{b}_1, \mathbf{b}_2)|^2 d\mathbf{x} \right) \right. \\ & \left. + \left(\left| B_{2,1}^{(-1)} \right|^2 + \left| B_{2,2}^{(-1)} \right|^2 \right) \left(\int_K |D^2 v(\mathbf{x}) \cdot (\mathbf{b}_2, \mathbf{b}_1)|^2 d\mathbf{x} + \int_K |D^2 v(\mathbf{x}) \cdot (\mathbf{b}_2, \mathbf{b}_2)|^2 d\mathbf{x} \right) \right]. \end{aligned}$$

In matrix notation, the bound becomes

$$(3.18) \quad |v - \Pi v|_{1,2,K}^2 \leq C \int_K \left\| \begin{bmatrix} \|\mathbf{B}^{-1}(1, \cdot)\|_2 & 0 \\ 0 & \|\mathbf{B}^{-1}(2, \cdot)\|_2 \end{bmatrix} \mathbf{B}^T D^2 v(\mathbf{x}) \mathbf{B} \right\|_F^2 dx,$$

where $\|\mathbf{B}^{-1}(1, \cdot)\|_2$ is the 2-norm for the first row vector of \mathbf{B}^{-1} .

4. Applications. In this section, the new bounds are compared to existing results. First, we derive existing results as upper bounds of our new bounds. This derivation indicates that our new bounds are, at worst, equivalent to existing results. Then, for some elements with small or large angles, we compare the asymptotic behavior of the new bounds with the behavior of the existing results, when the diameter of the element goes to zero. For these simple cases, our new bound exhibits the same asymptotic behavior as the norm of the interpolation error while existing results do not have the correct asymptotic behavior. This difference indicates that our new bounds are not equivalent to existing results and, in some cases, the new bounds are strictly sharper than the existing results of Ciarlet [8], Formaggia and Perotto [10], Georgoulis et al. [12], and Huang [13].

4.1. Recovering existing results. In this section, we recover, from Theorem 3.2, local bounds on K derived by Jamet [14], Formaggia and Perotto [11], and Apel [1].

Result of Jamet [14] on triangles. For triangular elements, Jamet proved

$$(4.1) \quad |v - \Pi v|_{m,p,K} \leq C \frac{h^{k+1-m}}{(\cos \theta)^m} |v|_{k+1,p,K},$$

where h is the diameter of K and θ is half the largest angle of triangle K . We will show that we can recover Jamet’s result from (3.15).

First we recall the decomposition of \mathbf{e}_j :

$$\mathbf{e}_j = \sum_{i=1}^2 B_{i,j}^{(-1)} \mathbf{b}_i = \sum_{i=1}^2 B_{i,j}^{(-1)} \|\mathbf{b}_i\|_2 \frac{\mathbf{b}_i}{\|\mathbf{b}_i\|_2}.$$

Using Lemma 2.4 of Jamet [14], we have

$$(4.2) \quad \sum_{j=1}^2 \left| B_{j,i}^{(-1)} \right| \|\mathbf{b}_j\|_2 \leq \frac{1}{\cos \theta}.$$

We recall also

$$(4.3) \quad \max_{1 \leq i \leq 2} \|\mathbf{b}_i\|_2 \leq \|\mathbf{B}\|_F \leq C \frac{h}{\hat{\rho}},$$

where $\hat{\rho}$ is the inner radius for the reference element \hat{K} . For a finite value of $p = q$ and for $l = k + 1$, we describe how to recover Jamet’s result from (3.15):

$$|v - \Pi v|_{m,p,K} \leq C \left[\sum_{\alpha \in I_m} \sum_{j_1, \dots, j_m=1}^2 \left| B_{j_1, \alpha_1}^{(-1)} \right|^p \cdots \left| B_{j_m, \alpha_m}^{(-1)} \right|^p \sum_{\beta \in I_{k+1-m}} \int_K |D^{k+1} v(\mathbf{x}) \cdot (\mathbf{b}_{j_1}, \dots, \mathbf{b}_{j_m}, \mathbf{b}_{\beta_1}, \dots, \mathbf{b}_{\beta_{k+1-m}})|^p dx \right]^{\frac{1}{p}}.$$

We now incorporate (4.3) and normalize the edge vectors:

$$|v - \Pi v|_{m,p,K} \leq C \frac{h^{k+1-m}}{\hat{\rho}^{k+1-m}} \left[\sum_{\alpha \in I_m} \sum_{j_1, \dots, j_m=1}^2 \left| B_{j_1, \alpha_1}^{(-1)} \right|^p \|\mathbf{b}_{j_1}\|_2^p \cdots \left| B_{j_m, \alpha_m}^{(-1)} \right|^p \|\mathbf{b}_{j_m}\|_2^p \right. \\ \left. \sum_{\beta \in I_{k+1-m}} \int_K \left| D^{k+1} v(\mathbf{x}) \cdot \left(\frac{\mathbf{b}_{j_1}}{\|\mathbf{b}_{j_1}\|_2}, \dots, \frac{\mathbf{b}_{j_m}}{\|\mathbf{b}_{j_m}\|_2}, \frac{\mathbf{b}_{\beta_1}}{\|\mathbf{b}_{\beta_1}\|_2}, \dots, \frac{\mathbf{b}_{\beta_{k+1-m}}}{\|\mathbf{b}_{\beta_{k+1-m}}\|_2} \right) \right|^p d\mathbf{x} \right]^{\frac{1}{p}}.$$

We can bound the integral term by $C |v|_{k+1,p,K}^p$:

$$|v - \Pi v|_{m,p,K} \leq C \frac{h^{k+1-m}}{\hat{\rho}^{k+1-m}} |v|_{k+1,p,K} \left[\sum_{\alpha \in I_m} \sum_{j_1, \dots, j_m=1}^2 \left| B_{j_1, \alpha_1}^{(-1)} \right|^p \|\mathbf{b}_{j_1}\|_2^p \cdots \left| B_{j_m, \alpha_m}^{(-1)} \right|^p \|\mathbf{b}_{j_m}\|_2^p \right]^{\frac{1}{p}}.$$

With the result (4.2), we conclude

$$|v - \Pi v|_{m,p,K} \leq C \left[\sum_{\alpha \in I_m} \sum_{j_1, \dots, j_m=1}^2 \left| B_{j_1, \alpha_1}^{(-1)} \right|^p \cdots \left| B_{j_m, \alpha_m}^{(-1)} \right|^p \right. \\ \left. \sum_{\beta \in I_{k+1-m}} \int_K \left| D^{k+1} v(\mathbf{x}) \cdot (\mathbf{b}_{j_1}, \dots, \mathbf{b}_{j_m}, \mathbf{b}_{\beta_1}, \dots, \mathbf{b}_{\beta_{k+1-m}}) \right|^p d\mathbf{x} \right]^{\frac{1}{p}} \\ \leq \frac{C}{(\cos \theta)^m} \frac{h^{k+1-m}}{\hat{\rho}^{k+1-m}} |v|_{k+1,p,K},$$

where the constants C depend only on \hat{K} , k , m , and p . This bound is isotropic as it does not distinguish the partial derivatives of v .

Bound of Formaggia and Perotto [11, Lemma 2] for smooth functions in \mathbb{R}^2 . For smooth functions ($v \in W^{2,2}(K)$), their result bounds the $W^{1,2}$ seminorm of the error for piecewise linear elements on triangles. For piecewise linear interpolation in \mathbb{R}^2 , our result writes as follows:

$$(4.4) \quad |v - \Pi v|_{1,2,K}^2 \leq C \int_K \left\| \begin{bmatrix} \|\mathbf{B}^{-1}(1, \cdot)\|_2 & 0 \\ 0 & \|\mathbf{B}^{-1}(2, \cdot)\|_2 \end{bmatrix} \mathbf{B}^T D^2 v(\mathbf{x}) \mathbf{B} \right\|_F^2 d\mathbf{x}.$$

We have

$$\left\| \begin{bmatrix} \|\mathbf{B}^{-1}(1, \cdot)\|_2 & 0 \\ 0 & \|\mathbf{B}^{-1}(2, \cdot)\|_2 \end{bmatrix} \mathbf{B}^T D^2 v(\mathbf{x}) \mathbf{B} \right\|_F^2 \leq \|\mathbf{B}^{-1}\|_F^2 \|\mathbf{B}^T D^2 v(\mathbf{x}) \mathbf{B}\|_F^2.$$

For any 2×2 matrix \mathbf{B} ,

$$(4.5) \quad \|\mathbf{B}^{-1}\|_F^2 \leq 2 \|\mathbf{B}^{-1}\|_2^2.$$

Consequently, we obtain

$$|v - \Pi v|_{1,2,K}^2 \leq C \|\mathbf{B}^{-1}\|_2^2 \int_K \|\mathbf{B}^T D^2 v(\mathbf{x}) \mathbf{B}\|_F^2 d\mathbf{x}.$$

We introduce the singular decomposition of $\mathbf{B} = \mathbf{R}\mathbf{\Sigma}\mathbf{P}^T$ where \mathbf{R} and \mathbf{P} are orthogonal matrices associated with the left and right singular vectors. $\mathbf{\Sigma}$ is a diagonal matrix containing the singular values. The column vectors of \mathbf{R} and \mathbf{P} are written

$$\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2] \quad \text{and} \quad \mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2].$$

We remark that $\mathbf{B}\mathbf{r}_i = \sigma_i\mathbf{p}_i$.

The Frobenius norm being invariant by rotation, we now write

$$|v - \Pi v|_{1,2,K}^2 \leq C \|\mathbf{B}^{-1}\|_2^2 \int_K \|\mathbf{\Sigma}\mathbf{R}^T D^2 v(\mathbf{x})\mathbf{R}\mathbf{\Sigma}\|_F^2 dx,$$

which is exactly the estimate of Formaggia and Perotto [11, Lemma 2]. This derivation shows also that we have

$$\begin{aligned} |v - \Pi v|_{1,2,K}^2 &\leq C \int_K \left\| \begin{bmatrix} \|\mathbf{B}^{-1}(1, \cdot)\|_2 & 0 \\ 0 & \|\mathbf{B}^{-1}(2, \cdot)\|_2 \end{bmatrix} \mathbf{B}^T D^2 v(\mathbf{x})\mathbf{B} \right\|_F^2 dx \\ &\leq C \|\mathbf{B}^{-1}\|_2^2 \int_K \|\mathbf{\Sigma}\mathbf{R}^T D^2 v(\mathbf{x})\mathbf{R}\mathbf{\Sigma}\|_F^2 dx. \end{aligned}$$

Georgoulis et al. [12] proved recently an extension for the result of Formaggia and Perotto to quadrilaterals in the $W^{m,2}$ spaces. We will show in the following section that these bounds do not have the correct asymptotic behavior when K is a right-angled triangle with a small angle or a rectangle with large aspect ratio and the diameter of K goes to zero.

Bounds of Apel [1, Theorem 2.1] on a triangle. Without any loss of generality, we restrict the derivation to a triangle K and finite values for p and q . We denote by h_1 the length of the longest edge and by $h_2 = |\det \mathbf{B}|/h_1$ the thickness. We assume that the element K satisfies a maximal angle condition; i.e., all the interior angles of K are uniformly bounded by a constant $\theta_* < \pi$, and a coordinate system condition; i.e., the angle γ between the longest edge and the x_1 -axis is bounded by $|\sin \gamma| \leq Ch_2/h_1$. Under these assumptions, Apel [1, Lemma 2.5] proved that

$$(4.6) \quad |B_{i,j}| \leq C \min(h_i, h_j) \quad \text{and} \quad |B_{i,j}^{(-1)}| \leq C \min(h_i^{-1}, h_j^{-1}),$$

where the constants C do not depend on K .

The new bound of Theorem 3.2 writes as follows:

$$\begin{aligned} |v - \Pi v|_{m,q,K} &\leq C |\det \mathbf{B}|^{\frac{1}{q} - \frac{1}{p}} \left\{ \sum_{\alpha \in I_m} \sum_{j_1, \dots, j_m=1}^d |B_{j_1, \alpha_1}^{(-1)}|^q \cdots |B_{j_m, \alpha_m}^{(-1)}|^q \right. \\ &\quad \left. \left[\sum_{\beta \in I_{l-m}} \int_K |D^l v(\mathbf{x}) \cdot (\mathbf{b}_{j_1}, \dots, \mathbf{b}_{j_m}, \mathbf{b}_{\beta_1}, \dots, \mathbf{b}_{\beta_{l-m}})|^p dx \right]^{\frac{q}{p}} \right\}^{\frac{1}{q}}. \end{aligned}$$

Using (4.6), we have

$$|B_{j_1, \alpha_1}^{(-1)}|^q \cdots |B_{j_m, \alpha_m}^{(-1)}|^q \leq C \left(\prod_{i=1}^m h_{j_i}^{-1} \right)^q.$$

When we represent an edge vector \mathbf{b}_{j_i} in the basis of $(\mathbf{e}_k)_{1 \leq k \leq d}$, we have

$$\begin{aligned} & D^l v(\mathbf{x}) \cdot (\mathbf{b}_{j_1}, \dots, \mathbf{b}_{j_m}, \mathbf{b}_{\beta_1}, \dots, \mathbf{b}_{\beta_{l-m}}) \\ &= \sum_{k_1, \dots, k_m=1}^d B_{k_1, j_1} \cdots B_{k_m, j_m} D^l v(\mathbf{x}) \cdot (\mathbf{e}_{k_1}, \dots, \mathbf{e}_{k_m}, \mathbf{b}_{\beta_1}, \dots, \mathbf{b}_{\beta_{l-m}}) \end{aligned}$$

and, consequently,

$$\begin{aligned} & |D^l v(\mathbf{x}) \cdot (\mathbf{b}_{j_1}, \dots, \mathbf{b}_{j_m}, \mathbf{b}_{\beta_1}, \dots, \mathbf{b}_{\beta_{l-m}})| \\ & \leq C \left(\prod_{i=1}^m h_{j_i} \right) \sum_{k_1, \dots, k_m=1}^d |D^l v(\mathbf{x}) \cdot (\mathbf{e}_{k_1}, \dots, \mathbf{e}_{k_m}, \mathbf{b}_{\beta_1}, \dots, \mathbf{b}_{\beta_{l-m}})|. \end{aligned}$$

As v is a smooth function, we can change the order of derivatives in the right-hand side to obtain

$$\begin{aligned} & \int_K |D^l v(\mathbf{x}) \cdot (\mathbf{b}_{j_1}, \dots, \mathbf{b}_{j_m}, \mathbf{b}_{\beta_1}, \dots, \mathbf{b}_{\beta_{l-m}})|^p d\mathbf{x} \\ & \leq C \left(\prod_{i=1}^m h_{j_i} \right) |D^{l-m} v \cdot (\mathbf{b}_{\beta_1}, \dots, \mathbf{b}_{\beta_{l-m}})|_{m,p,K}^p. \end{aligned}$$

The bound on the seminorm becomes now

$$|v - \Pi v|_{m,q,K} \leq C |\det \mathbf{B}|^{\frac{1}{q} - \frac{1}{p}} \left[\sum_{\beta \in I_{l-m}} |D^{l-m} v \cdot (\mathbf{b}_{\beta_1}, \dots, \mathbf{b}_{\beta_{l-m}})|_{m,p,K}^p \right]^{\frac{1}{p}}.$$

Therefore, we obtain

$$\begin{aligned} (4.7) \quad & |v - \Pi v|_{m,q,K} \\ & \leq C |\det \mathbf{B}|^{\frac{1}{q} - \frac{1}{p}} \left[\sum_{s \in I_{l-m}} \left(\prod_{i=1}^{l-m} h_{s_i}^p \right) |D^{l-m} v \cdot (\mathbf{e}_{s_1}, \dots, \mathbf{e}_{s_{l-m}})|_{m,p,K}^p \right]^{\frac{1}{p}}, \end{aligned}$$

which corresponds to the estimate of Apel [1, Theorem 2.1]. This derivation shows that, when assuming the maximal angle condition and the coordinate system condition on K , we have

$$\begin{aligned} & |v - \Pi v|_{m,q,K} \leq C |\det \mathbf{B}|^{\frac{1}{q} - \frac{1}{p}} \left\{ \sum_{\alpha \in I_m} \sum_{j_1, \dots, j_m=1}^d |B_{j_1, \alpha_1}^{(-1)}|^q \cdots |B_{j_m, \alpha_m}^{(-1)}|^q \right. \\ & \quad \left. \left[\sum_{\beta \in I_{l-m}} \int_K |D^l v(\mathbf{x}) \cdot (\mathbf{b}_{j_1}, \dots, \mathbf{b}_{j_m}, \mathbf{b}_{\beta_1}, \dots, \mathbf{b}_{\beta_{l-m}})|^p d\mathbf{x} \right]^{\frac{q}{p}} \right\}^{\frac{1}{q}} \\ & \leq C |\det \mathbf{B}|^{\frac{1}{q} - \frac{1}{p}} \left[\sum_{s \in I_{l-m}} \left(\prod_{i=1}^{l-m} h_{s_i}^p \right) |D^{l-m} v \cdot (\mathbf{e}_{s_1}, \dots, \mathbf{e}_{s_{l-m}})|_{m,p,K}^p \right]^{\frac{1}{p}}. \end{aligned}$$

The estimates of Apel [1, Theorem 2.1] and of Theorem 3.2 start from the same bound on the element \hat{K} . We recall that the bounds on \hat{K} in Lemma 3.4 do not require further geometric assumption. After exploiting the affine transformation between K and \hat{K} , the estimates of Apel and of Theorem 3.2 differ on the physical element K . Apel uses, in the right-hand side, partial derivatives along the coordinate axis direction. To bound the affine transformations from K to \hat{K} and from \hat{K} to K with (4.6), Apel requires geometric limitations on the element K : a maximum angle condition and a coordinate system condition. On the other hand, the bounds in this paper emphasize directional derivatives of the function along adjacent edges. They do not require further geometric assumption on K . These new bounds use element-related directions which are more appropriate for mesh adaptation techniques.

4.2. Behavior with small and large angles. In this section, we study the behavior of our bounds when small or large angles are present. We focus on elements in \mathbb{R}^2 , but the comments hold also for elements in \mathbb{R}^3 .

4.2.1. Right-angled triangle with a small angle. Let us consider the triangle

$$K = \{(x_1, x_2) \in \mathbb{R}^2 \mid 0 < x_1 < h \text{ and } 0 < x_2 < h^{\beta-1}(h - x_1)\},$$

where h is a real positive number, smaller than 1, that will converge to 0. β is a real positive number greater than 1. This triangle exhibits a small angle at the node $(h, 0)$ and two angles close to $\pi/2$ at the other nodes. We choose the matrix \mathbf{B} for the affine map to be

$$\mathbf{B} = \begin{bmatrix} h & 0 \\ 0 & h^\beta \end{bmatrix}.$$

Suppose we want to approximate the function $v(x_1, x_2) = x_1^2$. The nodal interpolant with \mathbb{P}_1 functions is $\Pi v(x_1, x_2) = hx_1$. The $W^{1,2}$ seminorm of the interpolation error satisfies

$$(4.8) \quad |v - \Pi v|_{1,2,K} = \left(\int_K |\nabla v - \nabla \Pi v|^2 \right)^{1/2} = \frac{1}{\sqrt{6}} h^{\frac{3+\beta}{2}}.$$

As the function v depends only on the variable x_1 , we have

$$\mathbf{B}^T D^2 v \mathbf{B} = \begin{bmatrix} 2h^2 & 0 \\ 0 & 0 \end{bmatrix}.$$

So the upper bound for the $W^{1,2}$ seminorm becomes

$$\left(\int_K \left\| \begin{bmatrix} \|\mathbf{B}^{-1}(1, \cdot)\|_2 & 0 \\ 0 & \|\mathbf{B}^{-1}(2, \cdot)\|_2 \end{bmatrix} \mathbf{B}^T D^2 v(\mathbf{x}) \mathbf{B} \right\|_F^2 dx \right)^{\frac{1}{2}} = \sqrt{2} h^{\frac{3+\beta}{2}},$$

which has the correct asymptotic behavior when h goes to zero.

Ciarlet [8, Theorem 3.1.2], Formaggia and Perotto [10, Proposition 2.1], and Huang [13, Lemma 2.1] use, as a starting point for their estimates, the following isotropic bound:

$$(4.9) \quad |v - \Pi v|_{1,2,K} \leq C \sqrt{\det \mathbf{B}} \|\mathbf{B}^{-1}\|_2 \left| \hat{v} - \hat{\Pi} \hat{v} \right|_{1,2,\hat{K}}.$$

We will show that this bound does not have the correct asymptotic behavior when the diameter of K goes to zero. On the reference triangle, we have

$$\hat{v}(\hat{x}_1, \hat{x}_2) = h^2 \hat{x}_1^2 \quad \text{and} \quad \left| \hat{v} - \hat{\Pi} \hat{v} \right|_{1,2,\hat{K}} = \frac{h^2}{\sqrt{6}}.$$

So the bound (4.9) becomes

$$\sqrt{\det \mathbf{B}} \|\mathbf{B}^{-1}\|_2 \left| \hat{v} - \hat{\Pi} \hat{v} \right|_{1,2,\hat{K}} = h^{\frac{1+\beta}{2}} h^{-\beta} \frac{h^2}{\sqrt{6}} = \frac{1}{\sqrt{6}} h^{\frac{5-\beta}{2}}.$$

The upper bound (4.9) is not sharp because of the term $\|\mathbf{B}^{-1}\|_2$. The isotropic estimate (4.9) can even explode to infinity when $\beta > 5$ while the interpolation error converges to zero. As the asymptotic behavior of (4.9) is not correct, the derived bounds of Ciarlet [8], Formaggia and Perotto [10], and Huang [13] will not have the correct asymptotic behavior. For this particular element and this choice of basis $(\mathbf{b}_j)_{1 \leq j \leq d}$, our new bound is sharper than the three existing results of Ciarlet [8], Formaggia and Perotto [10], and Huang [13].

Effect of node numbering. For this triangle, we can define three different matrices \mathbf{B} when we assume that the edges \mathbf{b}_1 and \mathbf{b}_2 are ordered counter-clockwise. We will assess the effect of these different matrices on the upper bound for $|v - \Pi v|_{1,2,K}$.

If we choose the matrix \mathbf{B} to be

$$\mathbf{B} = \begin{bmatrix} h & 0 \\ 0 & h^\beta \end{bmatrix},$$

the previous analysis shows that the upper bound has the correct asymptotic behavior. When the matrix \mathbf{B} is

$$\mathbf{B} = \begin{bmatrix} 0 & h \\ -h^\beta & -h^\beta \end{bmatrix},$$

the same conclusion holds. However, when we choose

$$\mathbf{B} = \begin{bmatrix} -h & -h \\ h^\beta & 0 \end{bmatrix},$$

the upper bound becomes

$$\left(\int_K \left\| \begin{bmatrix} \|\mathbf{B}^{-1}(1, \cdot)\|_2 & 0 \\ 0 & \|\mathbf{B}^{-1}(2, \cdot)\|_2 \end{bmatrix} \mathbf{B}^T D^2 v(\mathbf{x}) \mathbf{B} \right\|_F^2 d\mathbf{x} \right)^{\frac{1}{2}} = \sqrt{4h^{3+\beta} + 8h^{5-\beta}},$$

which does not have the correct asymptotic behavior when h goes to zero. This example illustrates that the bound (3.6) is not invariant with the numbering of vertices. We expect that the same conclusion holds for tetrahedra. Finally, we note that the upper bound (4.9) gives in all three cases the wrong asymptotic behavior when h goes to zero.

4.2.2. Rectangle with a large aspect ratio. Let us consider the rectangle

$$K = \{(x_1, x_2) \in \mathbb{R}^2 \mid 0 < x_1 < h \text{ and } 0 < x_2 < h^\beta\},$$

where h is a real positive number, smaller than 1, that will converge to 0. β is a real positive number greater than 1. We choose the matrix \mathbf{B} for the affine map to be

$$\mathbf{B} = \begin{bmatrix} h & 0 \\ 0 & h^\beta \end{bmatrix}.$$

We want to approximate the function $v(x_1, x_2) = x_1^2$. The nodal interpolant with \mathbb{Q}_1 functions is $\Pi v(x_1, x_2) = hx_1$. The $W^{1,2}$ seminorm of the interpolation error satisfies

$$(4.10) \quad |v - \Pi v|_{1,2,K} = \left(\int_K |\nabla v - \nabla \Pi v|^2 \right)^{1/2} = \frac{1}{\sqrt{3}} h^{\frac{3+\beta}{2}}.$$

We remark that

$$|v|_{2,2,K} = 2h^{\frac{1+\beta}{2}} \text{ and } |v - \Pi v|_{1,2,K} = \frac{h}{2\sqrt{3}} |v|_{2,2,K}.$$

Similarly to section 4.2.1, we can show that the upper bound for the $W^{1,2}$ seminorm is

$$\left(\int_K \left\| \begin{bmatrix} \|\mathbf{B}^{-1}(1, \cdot)\|_2 & 0 \\ 0 & \|\mathbf{B}^{-1}(2, \cdot)\|_2 \end{bmatrix} \mathbf{B}^T D^2 v(\mathbf{x}) \mathbf{B} \right\|_F^2 dx \right)^{\frac{1}{2}} = 2h^{\frac{3+\beta}{2}},$$

which has the correct asymptotic behavior when h goes to zero. However, the isotropic bound (4.9) satisfies

$$\sqrt{\det \mathbf{B}} \|\mathbf{B}^{-1}\|_2 |\hat{v} - \hat{\Pi} \hat{v}|_{1,2,\hat{K}} = h^{\frac{1+\beta}{2}} h^{-\beta} \frac{h^2}{\sqrt{6}} = \frac{1}{\sqrt{6}} h^{\frac{5-\beta}{2}},$$

which does not exhibit the correct asymptotic behavior when h goes to zero. For rectangular elements, Georgoulis et al. [12] and Huang [13] use the isotropic bound (4.9) as a starting point for their estimates. As the asymptotic behavior of (4.9) is not correct, the derived bounds of Georgoulis et al. [12] and Huang [13] will not have the correct asymptotic behavior. For this particular element, our new bound is sharper than the existing results of Georgoulis et al. [12] and Huang [13]. Mesh adaptation techniques based on (4.9) or on derived bounds could result in overmeshing.

Effect of node numbering. For this rectangle, we can define four different matrices \mathbf{B} when we assume that the edges \mathbf{b}_1 and \mathbf{b}_2 are ordered counter-clockwise. Here, the upper bound for $|v - \Pi v|_{1,2,K}$ remains unchanged for all the matrices. Indeed, for a rectangle or a parallelogram, the edge directions remain the same. In the upper bounds, making different choices of adjacent edges result in permuting the same terms, which will not change the asymptotic behavior of the upper bound. The same remark holds for parallelepipeds. Finally, we note that the upper bound (4.9) gives in all four cases the wrong asymptotic behavior when h goes to zero.

4.2.3. Rotating triangle with a large angle. We consider a rotating triangle K_α with vertices $(-\cos \alpha/2, -\sin \alpha/2)$, $(\cos \alpha/2, \sin \alpha/2)$, and $(-h \sin \alpha, h \cos \alpha)$. h is a real positive number that will go to 0. This triangle exhibits a large angle at the third vertex $(-h \sin \alpha, h \cos \alpha)$. We choose the matrix \mathbf{B} for the affine map to be

$$\mathbf{B} = \begin{bmatrix} \cos \alpha & -h \sin \alpha + \frac{\cos \alpha}{2} \\ \sin \alpha & h \cos \alpha + \frac{\sin \alpha}{2} \end{bmatrix}.$$

The shape functions are

$$\begin{aligned} \phi_{1,K_\alpha}(x_1, x_2) &= \frac{1}{2} - \left(\cos \alpha - \frac{\sin \alpha}{2h} \right) x_1 - \left(\sin \alpha + \frac{\cos \alpha}{2h} \right) x_2, \\ \phi_{2,K_\alpha}(x_1, x_2) &= \frac{1}{2} + \left(\cos \alpha + \frac{\sin \alpha}{2h} \right) x_1 + \left(\sin \alpha - \frac{\cos \alpha}{2h} \right) x_2, \\ \phi_{3,K_\alpha}(x_1, x_2) &= -\frac{\sin \alpha}{h} x_1 + \frac{\cos \alpha}{h} x_2. \end{aligned}$$

We want to approximate the function $v(x_1, x_2) = x_1^2$. The nodal interpolant with \mathbb{P}_1 functions is

$$\begin{aligned} \Pi v(x_1, x_2) &= \frac{1}{4} \cos^2 \alpha [\phi_{1,K_\alpha}(x_1, x_2) + \phi_{2,K_\alpha}(x_1, x_2)] + h^2 \sin^2 \alpha \phi_{3,K_\alpha}(x_1, x_2), \\ \Pi v(x_1, x_2) &= \frac{1}{4} \cos^2 \alpha - x_1 \left(h \sin^3 \alpha - \frac{\cos^2 \alpha \sin \alpha}{4h} \right) + x_2 \left(h \cos \alpha \sin^2 \alpha - \frac{\cos^3 \alpha}{4h} \right). \end{aligned}$$

We have the following integrals:

$$\begin{aligned} \int_{K_\alpha} \left[\frac{\partial}{\partial x_1} (v - \Pi v) \right]^2 &= h^3 \left(\frac{1}{3} \sin^2 \alpha - \frac{2}{3} \sin^4 \alpha + \frac{1}{2} \sin^6 \alpha \right) \\ &\quad + \frac{h}{12} (\cos^2 \alpha + 2 \cos^2 \alpha \sin^2 \alpha - 3 \cos^2 \alpha \sin^4 \alpha) + \frac{\cos^4 \alpha \sin^2 \alpha}{32h} \end{aligned}$$

and

$$\int_{K_\alpha} \left[\frac{\partial}{\partial x_2} (v - \Pi v) \right]^2 = \frac{\cos^6 \alpha}{32h} - \frac{h}{4} \cos^4 \alpha \sin^2 \alpha + \frac{h^3}{2} \cos^2 \alpha \sin^4 \alpha.$$

For $\alpha = 0$, the $W^{1,2}$ seminorm of the interpolation error is

$$|v - \Pi v|_{1,2,K_0} = \sqrt{\frac{h}{12} + \frac{1}{32h}}.$$

The interpolation error becomes unbounded when h goes to zero. This example was introduced by Babuška and Aziz [2] to justify that all interior angles of a triangle should be bounded away from π . The upper bound (3.18) becomes

$$\left(\int_{K_0} \left\| \begin{bmatrix} \|\mathbf{B}^{-1}(1, \cdot)\|_2 & 0 \\ 0 & \|\mathbf{B}^{-1}(2, \cdot)\|_2 \end{bmatrix} \mathbf{B}^T D^2 v(\mathbf{x}) \mathbf{B} \right\|_F^2 d\mathbf{x} \right)^{\frac{1}{2}} = \sqrt{\frac{5h}{2} + \frac{5}{4h}},$$

which has the correct asymptotic behavior when h goes to zero. We can also bound separately the partial derivative in x_1

$$\int_{K_0} \left[\frac{\partial}{\partial x_1} (v - \Pi v) \right]^2 \leq C \int_{K_0} \left\| \begin{bmatrix} |B_{1,1}^{(-1)}| & 0 \\ 0 & |B_{2,1}^{(-1)}| \end{bmatrix} \mathbf{B}^T D^2 v \mathbf{B} \right\|_F^2 = \frac{5Ch}{2},$$

which has the correct asymptotic behavior. The upper bound for the partial derivative in x_2 has also the correct asymptotic behavior.

For $\alpha = \pi/2$, the $W^{1,2}$ seminorm of the interpolation error is

$$|v - \Pi v|_{1,2,K_{\pi/2}} = \sqrt{\frac{h^3}{6}}.$$

Here the interpolation error converges to 0 when h goes to zero even though a large angle is present in the triangle $K_{\pi/2}$. This example illustrates the importance of the element orientation with respect to the function v . When studying the linear interpolation error for quadratic functions, Cao [5] highlighted also the importance of the element orientation with respect to the function when a large angle is present in the triangle.

The upper bound (3.18) becomes

$$\left(\int_{K_{\pi/2}} \left\| \begin{bmatrix} \|\mathbf{B}^{-1}(1, \cdot)\|_2 & 0 \\ 0 & \|\mathbf{B}^{-1}(2, \cdot)\|_2 \end{bmatrix} \mathbf{B}^T D^2 v(\mathbf{x}) \mathbf{B} \right\|_F^2 dx \right)^{\frac{1}{2}} = \sqrt{2} h^{\frac{3}{2}},$$

which has the correct asymptotic behavior when h goes to zero. We can also bound separately the partial derivatives in x_1

$$\int_{K_{\pi/2}} \left[\frac{\partial}{\partial x_1} (v - \Pi v) \right]^2 \leq C \int_{K_{\pi/2}} \left\| \begin{bmatrix} |B_{1,1}^{(-1)}| & 0 \\ 0 & |B_{2,1}^{(-1)}| \end{bmatrix} \mathbf{B}^T D^2 v \mathbf{B} \right\|_F^2 = 2Ch^3,$$

and in x_2

$$\int_{K_{\pi/2}} \left[\frac{\partial}{\partial x_2} (v - \Pi v) \right]^2 \leq C \int_{K_{\pi/2}} \left\| \begin{bmatrix} |B_{1,2}^{(-1)}| & 0 \\ 0 & |B_{2,2}^{(-1)}| \end{bmatrix} \mathbf{B}^T D^2 v \mathbf{B} \right\|_F^2 = 0.$$

Both upper bounds have the correct asymptotic behavior when h goes to zero.

Effect of node numbering. For this example, we can define three different matrices \mathbf{B} when we assume that the edges \mathbf{b}_1 and \mathbf{b}_2 are ordered counter-clockwise.

When α is 0 and if we choose the matrix \mathbf{B} to be

$$\mathbf{B} = \begin{bmatrix} 1 & \frac{1}{2} \\ 0 & h \end{bmatrix} \quad \text{or} \quad \mathbf{B} = \begin{bmatrix} -\frac{1}{2} & -1 \\ h & 0 \end{bmatrix},$$

the upper bound remains unchanged and it has the correct asymptotic behavior. When $\alpha = 0$ and the matrix \mathbf{B} is

$$\mathbf{B} = \begin{bmatrix} -\frac{1}{2} & \frac{1}{2} \\ -h & -h \end{bmatrix},$$

the upper bound becomes

$$\left(\int_{K_0} \left\| \begin{bmatrix} \|\mathbf{B}^{-1}(1, \cdot)\|_2 & 0 \\ 0 & \|\mathbf{B}^{-1}(2, \cdot)\|_2 \end{bmatrix} \mathbf{B}^T D^2 v(\mathbf{x}) \mathbf{B} \right\|_F^2 dx \right)^{\frac{1}{2}} = \sqrt{\frac{h}{2} + \frac{1}{8h}},$$

which still has the correct asymptotic behavior when h goes to zero.

The same conclusion holds when $\alpha = \pi/2$. For this example, the bound (3.18) keeps the same asymptotic behavior.

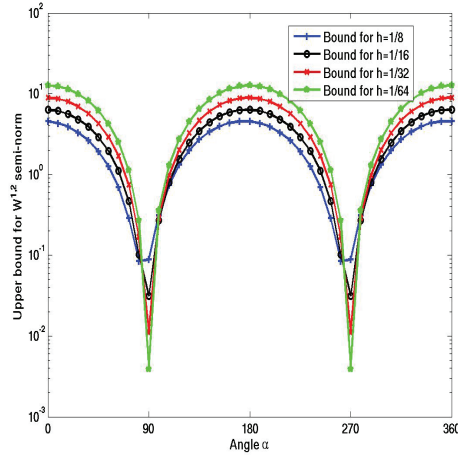


FIG. 4.1. Behavior of upper bound (3.18) for different heights.

Effect of α . Finally, in Figure 4.1, we present the behavior of the upper bound (3.18) as a function of α for several heights h . The upper bound increases when h goes to zero except when α is equal to $\pi/2$ and $3\pi/2$. This final example shows that the upper bound (3.18) reproduces the correct asymptotic behavior for every angle α . Therefore, the validity of (3.18) is not subject to a limitation on the angle between the largest edge and the x_1 -axis.

4.2.4. Parallelogram with two large angles. Let us consider the parallelogram

$$K = \left\{ (x_1, x_2) \in \mathbb{R}^2 \mid -\frac{1}{2} < x_1 < \frac{1}{2} \text{ and } |x_2| < h(1 - 2|x_1|) \right\},$$

where h is a real positive number, smaller than 1, that will converge to 0. This parallelogram exhibits two large angles at the nodes $(0, -h)$ and $(0, h)$ and two small angles at the other vertices. We choose the matrix \mathbf{B} for the affine map to be

$$\mathbf{B} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ h & h \end{bmatrix}.$$

We want to approximate the function $v(x_1, x_2) = x_1^2$. The nodal interpolant with \mathbb{Q}_1 functions is

$$\Pi v(x_1, x_2) = \frac{1}{4} \left[2x_1^2 - \frac{x_2^2}{2h^2} + \frac{1}{2} \right].$$

The $W^{1,2}$ seminorm of the interpolation error satisfies

$$(4.12) \quad |v - \Pi v|_{1,2,K} = \left(\int_K |\nabla v - \nabla \Pi v|^2 \right)^{1/2} = \sqrt{\frac{h}{24} + \frac{1}{96h}}.$$

The upper bound for the $W^{1,2}$ seminorm is

$$\left(\int_K \left\| \begin{bmatrix} \|\mathbf{B}^{-1}(1, \cdot)\|_2 & 0 \\ 0 & \|\mathbf{B}^{-1}(2, \cdot)\|_2 \end{bmatrix} \mathbf{B}^T D^2 v(\mathbf{x}) \mathbf{B} \right\|_F^2 dx \right)^{\frac{1}{2}} = \sqrt{h + \frac{1}{4h}},$$

which has the correct asymptotic behavior when h goes to zero.

We want to approximate the function $v(x_1, x_2) = x_2^2$. The nodal interpolant with \mathbb{Q}_1 functions is

$$\Pi v(x_1, x_2) = h^2 \left[-2x_1^2 + \frac{x_2^2}{2h^2} + \frac{1}{2} \right].$$

The $W^{1,2}$ seminorm of the interpolation error satisfies

$$(4.13) \quad |v - \Pi v|_{1,2,K} = \left(\int_K |\nabla v - \nabla \Pi v|^2 \right)^{1/2} = \sqrt{\frac{2h^5}{3} + \frac{h^3}{6}}.$$

The upper bound for the $W^{1,2}$ seminorm is

$$\left(\int_K \left\| \begin{bmatrix} \|\mathbf{B}^{-1}(1, \cdot)\|_2 & 0 \\ 0 & \|\mathbf{B}^{-1}(2, \cdot)\|_2 \end{bmatrix} \mathbf{B}^T D^2 v(\mathbf{x}) \mathbf{B} \right\|_F^2 dx \right)^{\frac{1}{2}} = \sqrt{4h^3 + 16h^5},$$

which has also the correct asymptotic behavior when h goes to zero.

When two large angles are present in a parallelogram, this example illustrates the importance of the element orientation with respect to the function v . In both cases, the upper bound (3.18) has the correct asymptotic behavior.

5. Conclusion. We have presented new anisotropic error estimates for the Lagrangian finite element interpolation on affine equivalent elements. The bounds use information from the directional derivatives of the function to interpolate along adjacent edges. Besides the affine equivalent assumption on K , these local estimates do not require further geometric assumption. For simplices, the asymptotic behavior of the bounds, when the diameter of the element goes to zero, may vary with the choice of node ordering. However, for parallelograms and parallelepipeds, the asymptotic behavior does not vary with the choice of node ordering. Finally, we showed that existing results can be derived from our new bounds. For some simple cases, the new bounds are sharper than some existing results.

REFERENCES

- [1] T. APEL, *Anisotropic Finite Elements: Local Estimates and Applications*, Teubner, Leipzig, 1999.
- [2] I. BABUŠKA AND A. K. AZIZ, *On the angle condition in the finite element method*, SIAM J. Numer. Anal., 13 (1976), pp. 214–226.
- [3] R. BANK AND R. SMITH, *Mesh smoothing using a posteriori error estimates*, SIAM J. Numer. Anal., 34 (1997), pp. 979–997.
- [4] M. BERZINS, *A solution-based triangular and tetrahedral mesh quality indicator*, SIAM J. Sci. Comput., 19 (1998), pp. 2051–2060.
- [5] W. CAO, *On the error of linear interpolation and the orientation, aspect ratio, and internal angles of a triangle*, SIAM J. Numer. Anal., 43 (2005), pp. 19–40.
- [6] L. CHEN, P. SUN, AND J. XU, *Optimal anisotropic meshes for minimizing interpolation errors in L^p -norm*, Math. Comput., 76 (2007), pp. 179–204.
- [7] L. CHEN, *On minimizing the linear interpolation error of convex quadratic functions and the optimal simplex*, East J. Approx., 14 (2008), pp. 271–284.
- [8] P. CIARLET, *The Finite Element Method for Elliptic Problems*, SIAM, Philadelphia, 2002.
- [9] E. D’AZEVEDO AND R. SIMPSON, *On optimal interpolation triangle incidences*, SIAM J. Sci. Comput., 10 (1989), pp. 1063–1075.
- [10] L. FORMAGGIA AND S. PEROTTO, *New anisotropic a priori error estimates*, Numer. Math., 89 (2001), pp. 641–667.
- [11] L. FORMAGGIA AND S. PEROTTO, *Anisotropic error estimates for elliptic problems*, Numer. Math., 94 (2003), pp. 67–92.

- [12] E. GEORGIOULIS, E. HALL, AND P. HOUSTON, *Discontinuous Galerkin methods for advection-diffusion-reaction problems on anisotropically refined meshes*, SIAM J. Sci. Comput., 30 (2007), pp. 246–271.
- [13] W. HUANG, *Measuring mesh qualities and application to variational mesh adaptation*, SIAM J. Sci. Comput., 26 (2005), pp. 1643–1666.
- [14] P. JAMET, *Estimation d'erreur pour des éléments finis droits presque dégénérés*, RAIRO Anal. Numer., 10 (1976), pp. 43–61.
- [15] E. NADLER, *Piecewise linear approximation on triangulations of a planar region*, Ph.D. thesis, Brown University, Providence, RI, 1985.
- [16] E. NADLER, *Piecewise linear best L_2 approximation on triangulations*, Approximation Theory, Vol. V, C. Chui, L. Schumaker, and J. Ward, eds., Academic Press, New York, 1986, pp. 499–502.
- [17] S. RIPPA, *Long and thin triangles can be good for linear interpolation*, SIAM J. Numer. Anal., 29 (1992), pp. 257–270.
- [18] M. ZLÁMAL, *On the finite element method*, Numer. Math., 12 (1968), pp. 394–409.

POSTPROCESSING FINITE-ELEMENT METHODS FOR THE NAVIER–STOKES EQUATIONS: THE FULLY DISCRETE CASE*

JAVIER DE FRUTOS[†], BOSCO GARCÍA-ARCHILLA[‡], AND JULIA NOVO[§]

Abstract. An accuracy-enhancing postprocessing technique for finite-element discretizations of the Navier–Stokes equations is analyzed. The technique had been previously analyzed only for semidiscretizations, and fully discrete methods are addressed in the present paper. We show that the increased spatial accuracy of the postprocessing procedure is not affected by the errors arising from any convergent time-stepping procedure. Further refined bounds are obtained when the time-stepping procedure is either the backward Euler method or the two-step backward differentiation formula.

Key words. Navier–Stokes equations, mixed finite-element methods, time-stepping methods, optimal regularity, error estimates, backward Euler method, two-step BDF

AMS subject classifications. 65M60, 65M20, 65M15, 65M12

DOI. 10.1137/070707580

1. Introduction. The purpose of the present paper is to study a postprocessing technique for fully discrete mixed finite-element (MFE) methods for the incompressible Navier–Stokes equations

$$(1.1) \quad u_t - \Delta u + (u \cdot \nabla)u + \nabla p = f,$$

$$(1.2) \quad \operatorname{div}(u) = 0,$$

in a bounded domain $\Omega \subset \mathbb{R}^d$ ($d = 2, 3$) with smooth boundary subject to homogeneous Dirichlet boundary conditions $u = 0$ on $\partial\Omega$. In (1.1), u is the velocity field, p the pressure, and f a given force field. We assume that the fluid density and viscosity have been normalized by an adequate change of scale in space and time.

For semidiscrete MFE methods the postprocessing technique has been studied in [2, 3, 18] and is as follows. In order to approximate the solution u and p corresponding to a given initial condition

$$(1.3) \quad u(\cdot, 0) = u_0,$$

at a time $t \in (0, T]$, $T > 0$, consider first standard MFE approximations u_h and p_h to the velocity and pressure, respectively, solutions at time $t \in (0, T]$ of the corresponding discretization of (1.1)–(1.3). Then compute MFE approximations \tilde{u}_h and \tilde{p}_h to the

*Received by the editors November 8, 2007; accepted for publication (in revised form) September 5, 2008; published electronically December 31, 2008.

<http://www.siam.org/journals/sinum/47-1/70758.html>

[†]Departamento de Matemática Aplicada, Universidad de Valladolid, 47011 Valladolid, Spain (frutos@mac.uva.es). This author's research was financed by DGI-MEC under project MTM2007-60528 (cofinanced by FEDER funds) and Junta de Castilla y León under grants VA079A06 and VA045A06.

[‡]Departamento de Matemática Aplicada II, Universidad de Sevilla, 41002 Sevilla, Spain (bosco@esi.us.es). This author's research was supported by DGI-MEC under project MTM2006-00847.

[§]Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049 Madrid, Spain (julia.novo@uam.es). This author's research was financed by DGI-MEC under project MTM2007-60528 (cofinanced by FEDER funds) and Junta de Castilla y León under grants VA079A06 and VA045A06.

solution \tilde{u} and \tilde{p} of the following Stokes problem,

$$(1.4) \quad -\Delta \tilde{u} + \nabla \tilde{p} = f - \frac{d}{dt}u_h - (u_h \cdot \nabla)u_h \quad \text{in } \Omega,$$

$$(1.5) \quad \operatorname{div}(\tilde{u})=0 \quad \text{in } \Omega,$$

$$(1.6) \quad \tilde{u} = 0 \quad \text{on } \partial\Omega.$$

The MFE on this postprocessing step can be either the same MFE over a finer grid or a higher-order MFE over the same grid. In [2, 18] it is shown that if the errors in the velocity (in the H^1 norm) and the pressure of the standard MFE approximations u_h and p_h are $O(t^{-(r-2)/2}h^{r-1})$, $r = 2, 3, 4$, for $t \in (0, T]$, then those of the postprocessed approximations \tilde{u}_h and \tilde{p}_h are $O(t^{-(r-1)/2}h^r |\log(h)|)$, that is, an $O(h |\log(h)|)$ improvement with respect the standard MFE error bound (see precise statement on Theorem 2.2 below), and if $r \geq 3$ (finite elements of degree at least two), the $O(h |\log(h)|)$ improvement is also obtained in the L^2 norm of the velocity.

In practice, however, the finite-element approximations u_h and p_h can rarely be computed exactly, and one has to compute approximations $U_h^{(n)} \approx u_h(t_n)$ and $P_h^{(n)} \approx p_h(t_n)$ at some time levels $0 = t_0 < t_1 \cdots < t_N = T$, by means of a time integrator. Consequently, instead of the postprocessed approximations $\tilde{u}_h(t_n)$ and $\tilde{p}_h(t_n)$, one obtains $\tilde{U}_h^{(n)}$ and $\tilde{P}_h^{(n)}$ as solutions of a system similar to (1.4)–(1.6) but with u_h on the right-hand side of (1.4) replaced by $U_h^{(n)}$ and \dot{u}_h replaced by an appropriate approximation $d_t^*U_h^{(n)}$.

In the present paper we analyze the errors $u(t_n) - \tilde{U}_h^{(n)}$ and $p(t_n) - \tilde{P}_h^{(n)}$. We show that, if any convergent time stepping procedure is used to integrate the standard MFE approximation, then the error of the fully discrete postprocessed approximation, $u(t_n) - \tilde{U}_h^{(n)}$, is that of the semidiscrete postprocessed approximation $u(t_n) - \tilde{u}_h$ plus a term \tilde{e}_n whose norm is proportional to that of the time-discretization error $e_n = u_h(t_n) - U_h^{(n)}$ of the MFE method, and, furthermore, we show $\tilde{e}_n = e_n$ plus higher-order terms for two particular time integration methods, the backward Euler method and the two-step backward differentiation formula (BDF) [9] (see also [25, section III.1]). We remark that the fact that \tilde{e}_n is asymptotically equivalent to e_n has proved its relevance when developing a posteriori error estimators for dissipative problems [17] (see also [15, 16]). To prove $\tilde{e}_n \approx e_n$ we perform first a careful error analysis of the backward Euler method and the two-step BDF. This allows us to obtain error estimates for the pressure that improve by a factor of the time step k those in the literature [10, 34].

It must be noticed that the backward Euler method and the two-step BDF are the only G-stable methods (see, e.g., [26, section V.6]) in the BDF family of methods. G-stability makes it easier the use of energy methods in the analysis, and this has proved crucial in obtaining our error bounds. At present we ignore if error bounds similar to that obtained in the present paper can be obtained without resource to energy methods, so that error bounds for higher-order methods in the BDF family of methods can be proved.

The analysis of fully discrete postprocessed methods may be less trivial than it may seem at first sight, since although many results for postprocessed semidiscrete methods can be found in the literature (see next paragraph) as well as numerical experiments (carried out with fully discrete methods) showing an increase in accuracy similar to that theoretically predicted in semidiscrete methods [3, 11, 14, 12, 20, 22], the only analysis of postprocessed fully discrete methods is that by Yan [44]. There, for a semilinear parabolic equation of the type $u_t - \Delta u = F(u)$, Δ being the Laplacian

operator and F a smooth and bounded function, the postprocess of a finite-element (FE) approximation when integrated in time with the backward Euler method with fixed stepsize k is analyzed (higher-order time-stepping methods are also considered, but only for linear homogeneous parabolic equations). Error estimates are obtained where an $O(k(1 + h^2))$ term is added to the bounds previously obtained for the postprocessed semidiscrete approximation. It must be remarked, though, that in [44] no attempt is made to analyze methods for equations with convective terms. In fact, in [44], it is stated that “It is not quite clear how it is possible to generalize our method to deal with a nonlinear convection term”. This is precisely what we do in the present paper.

The postprocess technique considered here was first developed for spectral methods in [20, 21]. Later it was extended to methods based on Chebyshev and Legendre polynomials [11], spectral element methods [12, 13], and finite element methods [22, 14]. In these works, numerical experiments show that, if the postprocessed approximation is computed at the final time T , the postprocessed method is computationally more efficient than the method to which it is applied. Similar results are obtained in the numerical experiments in [2, 3] for MFE methods. Due to this better practical performance, the postprocessing technique has been applied to the study of nonlinear shell vibrations [37], as well as to stochastic differential parabolic equations [38]. Also, it has been effectively applied to reduce the order of practical engineering problems modeled by nonlinear differential systems [42, 43].

The postprocess technique can be seen as a two-level method, where the postprocessed (or fine-mesh) approximations \tilde{u}_h and \tilde{p}_h are an improvement of the previously computed (coarse mesh) approximations u_h and p_h . Recent research on two-level finite-element methods for the transient Navier–Stokes equations can be found in [23, 27, 28, 40] (see also [30, 29, 36, 39] for spectral discretizations), where the fully nonlinear problem is dealt with on the coarse mesh, and a linear problem is solved on the fine mesh.

The rest of the paper is as follows. In section 2 we introduce some standard material and the methods to be studied. In section 3 we analyze the fully discrete postprocessed method. In section 4 we prove some technical results and, finally, section 5 is devoted to analyze the time discretization errors of the MFE approximation when integrated with the backward Euler method or the two-step BDF.

2. Preliminaries and notations.

2.1. The continuous solution. We will assume that Ω is a bounded domain in \mathbb{R}^d , $d = 2, 3$, of class \mathcal{C}^m , for $m \geq 2$, and we consider the Hilbert spaces

$$\begin{aligned} H &= \{u \in L^2(\Omega)^d \mid \operatorname{div}(u) = 0, u \cdot n|_{\partial\Omega} = 0\}, \\ V &= \{u \in H_0^1(\Omega)^d \mid \operatorname{div}(u) = 0\}, \end{aligned}$$

endowed with the inner product of $L^2(\Omega)^d$ and $H_0^1(\Omega)^d$, respectively. For $l \geq 0$ integer and $1 \leq q \leq \infty$, we consider the standard Sobolev spaces, $W^{l,q}(\Omega)^d$, of functions with derivatives up to order l in $L^q(\Omega)$, and $H^l(\Omega)^d = W^{l,2}(\Omega)^d$. We will denote by $\|\cdot\|_l$ the norm in $H^l(\Omega)^d$, and $\|\cdot\|_{-l}$ will represent the norm of its dual space. We consider also the quotient spaces $H^l(\Omega)/\mathbb{R}$ with norm $\|p\|_{H^l/\mathbb{R}} = \inf\{\|p + c\|_l \mid c \in \mathbb{R}\}$.

Let $\Pi : L^2(\Omega)^d \longrightarrow H$ be the $L^2(\Omega)^d$ projection onto H . We denote by A the Stokes operator on Ω :

$$A : \mathcal{D}(A) \subset H \longrightarrow H, \quad A = -\Pi\Delta, \quad \mathcal{D}(A) = H^2(\Omega)^d \cap V.$$

Applying Leray’s projector to (1.1), the equations can be written in the form

$$u_t + Au + B(u, u) = \Pi f \quad \text{in } \Omega,$$

where $B(u, v) = \Pi(u \cdot \nabla)v$ for u, v in $H_0^1(\Omega)^d$.

We shall use the trilinear form $b(\cdot, \cdot, \cdot)$ defined by

$$b(u, v, w) = (F(u, v), w) \quad \forall u, v, w \in H_0^1(\Omega)^d,$$

where

$$F(u, v) = (u \cdot \nabla)v + \frac{1}{2}(\nabla \cdot u)v \quad \forall u, v \in H_0^1(\Omega)^d.$$

It is straightforward to verify that b enjoys the skew-symmetry property

$$(2.1) \quad b(u, v, w) = -b(u, w, v) \quad \forall u, v, w \in H_0^1(\Omega)^d.$$

Let us observe that $B(u, v) = \Pi F(u, v)$ for $u \in V, v \in H_0^1(\Omega)^d$.

We shall assume that u is a strong solution up to time $t = T$, so that

$$(2.2) \quad \|u(t)\|_1 \leq M_1, \quad \|u(t)\|_2 \leq M_2, \quad 0 \leq t \leq T,$$

for some constants M_1 and M_2 . We shall also assume that there exists another constant \tilde{M}_2 such that

$$(2.3) \quad \|f\|_1 + \|f_t\|_1 + \|f_{tt}\|_1 \leq \tilde{M}_2, \quad 0 \leq t \leq T.$$

Let us observe, however, that if for $k \geq 2$

$$\sup_{0 \leq t \leq T} \|\partial_t^{\lfloor k/2 \rfloor} f\|_{k-1-2\lfloor k/2 \rfloor} + \sum_{j=0}^{\lfloor (k-2)/2 \rfloor} \sup_{0 \leq t \leq T} \|\partial_t^j f\|_{k-2j-2} < +\infty,$$

and if Ω is of class C^k , then, according to Theorems 2.4 and 2.5 in [32], there exist positive constants M_k and K_k such that the following bounds hold:

$$(2.4) \quad \|u(t)\|_k + \|u_t(t)\|_{k-2} + \|p(t)\|_{H^{k-1}/\mathbb{R}} \leq M_k \tau(t)^{1-k/2},$$

$$(2.5) \quad \int_0^t \sigma_{k-3}(s) (\|u(s)\|_k^2 + \|u_s(s)\|_{k-2}^2 + \|p(s)\|_{H^{k-1}/\mathbb{R}}^2 + \|p_s(s)\|_{H^{k-3}/\mathbb{R}}^2) ds \leq K_k^2,$$

where $\tau(t) = \min(t, 1)$ and $\sigma_n = e^{-\alpha(t-s)}\tau^n(s)$ for some $\alpha > 0$. Observe that for $t \leq T < \infty$, we can take $\tau(t) = t$ and $\sigma_n(s) = s^n$. For simplicity, we will take these values of τ and σ_n .

We note that although the results in the present paper require only (2.2) and (2.3) to hold, those in [18] that we summarize in section 2.3 require that for $r = 3, 4$, (2.4)–(2.5) hold for $k = r + 2$.

2.2. The spatial discretization. Let $\mathcal{T}_h = (\tau_i^h, \phi_i^h)_{i \in I_h}, h > 0$ be a family of partitions of suitable domains Ω_h , where h is the maximum diameter of the elements $\tau_i^h \in \mathcal{T}_h$, and ϕ_i^h are the mappings of the reference simplex τ_0 onto τ_i^h . We restrict ourselves to quasi-uniform and regular meshes \mathcal{T}_h .

Let $r \geq 3$, we consider the finite-element spaces

$$S_{h,r} = \left\{ \chi_h \in \mathcal{C}(\bar{\Omega}_h) \mid \chi_h|_{\tau_i^h} \circ \phi_i^h \in P^{r-1}(\tau_0) \right\} \subset H^1(\Omega_h), \quad S_{h,r}^0 = S_{h,r} \cap H_0^1(\Omega_h),$$

where $P^{r-1}(\tau_0)$ denotes the space of polynomials of degree at most $r - 1$ on τ_0 . Since we are assuming that the meshes are quasi-uniform, the following inverse inequality holds for each $v_h \in (S_{h,r}^0)^d$ (see, e.g., [7, Theorem 3.2.6])

$$(2.6) \quad \|v_h\|_{W^{m,q}(\tau)^d} \leq Ch^{l-m-d(\frac{1}{q'}-\frac{1}{q})} \|v_h\|_{W^{l,q'}(\tau)^d},$$

where $0 \leq l \leq m \leq 1$, $1 \leq q' \leq q \leq \infty$, and τ is an element in the partition \mathcal{T}_h .

We shall denote by $(X_{h,r}, Q_{h,r-1})$ the so-called Hood–Taylor element [5, 35], when $r \geq 3$, where

$$X_{h,r} = (S_{h,r}^0)^d, \quad Q_{h,r-1} = S_{h,r-1} \cap L^2(\Omega_h)/\mathbb{R}, \quad r \geq 3,$$

and the so-called mini-element [6] when $r = 2$, where $Q_{h,1} = S_{h,2} \cap L^2(\Omega_h)/\mathbb{R}$, and $X_{h,2} = (S_{h,2}^0)^d \oplus \mathbb{B}_h$. Here, \mathbb{B}_h is spanned by the bubble functions b_τ , $\tau \in \mathcal{T}_h$, defined by $b_\tau(x) = (d + 1)^{d+1} \lambda_1(x) \cdots \lambda_{d+1}(x)$, if $x \in \tau$ and 0 elsewhere, where $\lambda_1(x), \dots, \lambda_{d+1}(x)$ denote the barycentric coordinates of x . For these elements a uniform inf-sup condition is satisfied (see [5]), that is, there exists a constant $\beta > 0$ independent of the mesh grid size h such that

$$(2.7) \quad \inf_{q_h \in Q_{h,r-1}} \sup_{v_h \in X_{h,r}} \frac{(q_h, \nabla \cdot v_h)}{\|v_h\|_1 \|q_h\|_{L^2/\mathbb{R}}} \geq \beta.$$

The approximate velocity belongs to the discretely divergence-free space

$$V_{h,r} = X_{h,r} \cap \{ \chi_h \in H_0^1(\Omega_h)^d \mid (q_h, \nabla \cdot \chi_h) = 0 \quad \forall q_h \in Q_{h,r-1} \},$$

which is not a subspace of V . We shall frequently write V_h instead of $V_{h,r}$ whenever the value of r plays no particular role.

Let $\Pi_h : L^2(\Omega)^d \rightarrow V_{h,r}$ be the discrete Leray’s projection defined by

$$(\Pi_h u, \chi_h) = (u, \chi_h) \quad \forall \chi_h \in V_{h,r}.$$

We will use the following well-known bounds

$$(2.8) \quad \|(I - \Pi_h)u\|_j \leq Ch^{l-j} \|u\|_l, \quad 1 \leq l \leq 2, \quad j = 0, 1.$$

We will denote by $A_h : V_h \rightarrow V_h$ the discrete Stokes operator defined by

$$(\nabla v_h, \nabla \phi_h) = (A_h v_h, \phi_h) = \left(A_h^{1/2} v_h, A_h^{1/2} \phi_h \right) \quad \forall v_h, \phi_h \in V_h.$$

Let $(u, p) \in (H^2(\Omega)^d \cap V) \times (H^1(\Omega)/\mathbb{R})$ be the solution of a Stokes problem with right-hand side g , we will denote by $s_h = S_h(u) \in V_h$ the so-called Stokes projection (see [33]) defined as the velocity component of solution of the following Stokes problem: find $(s_h, q_h) \in (X_{h,r}, Q_{h,r-1})$ such that

$$(2.9) \quad (\nabla s_h, \nabla \phi_h) + (\nabla q_h, \phi_h) = (g, \phi_h) \quad \forall \phi_h \in X_{h,r},$$

$$(2.10) \quad (\nabla \cdot s_h, \psi_h) = 0 \quad \forall \psi_h \in Q_{h,r-1}.$$

Obviously, $s_h = S_h(u)$. The following bound holds for $2 \leq l \leq r$:

$$(2.11) \quad \|u - s_h\|_0 + h \|u - s_h\|_1 \leq Ch^l (\|u\|_l + \|p\|_{H^{l-1}/\mathbb{R}}).$$

The proof of (2.11) for $\Omega = \Omega_h$ can be found in [33]. For the general case, Ω_h must be such that the value of $\delta(h) = \max_{x \in \partial\Omega_h} \text{dist}(x, \partial\Omega)$ satisfies

$$(2.12) \quad \delta(h) = O\left(h^{2(r-1)}\right).$$

This can be achieved if, for example, $\partial\Omega$ is piecewise of class $\mathcal{C}^{2(r-1)}$, and superparametric approximation at the boundary is used [1]. Under the same conditions, the bound for the pressure is [24]

$$(2.13) \quad \|p - q_h\|_{L^2/\mathbb{R}} \leq C_\beta h^{l-1} (\|u\|_l + \|p\|_{H^{l-1}/\mathbb{R}}),$$

where the constant C_β depends on the constant β in the inf-sup condition (2.7).

In the sequel we will apply the above estimates to the particular case in which (u, p) is the solution of the Navier–Stokes problem (1.1)–(1.3). In that case s_h is the discrete velocity in problem (2.9)–(2.10) with $g = f - u_t - (u \cdot \nabla u)$. Note that the temporal variable t appears here merely as a parameter and then, taking the time derivative, the error bounds (2.11) and (2.13) can also be applied to the time derivative of s_h changing u, p by u_t, p_t , respectively.

Since we are assuming that Ω is of class \mathcal{C}^m and $m \geq 2$, from (2.11) and standard bounds for the Stokes problem [1, 19], we deduce that

$$(2.14) \quad \|(A^{-1}\Pi - A_h^{-1}\Pi_h) f\|_j \leq Ch^{2-j} \|f\|_0 \quad \forall f \in L^2(\Omega)^d, \quad j = 0, 1.$$

In our analysis we shall frequently use the following relation, which is a consequence of (2.14) and the fact that any $f_h \in V_h$ vanishes on $\partial\Omega$. For some $c \geq 1$,

$$(2.15) \quad \frac{1}{c} \|A_h^{s/2} f_h\|_0 \leq \|f_h\|_s \leq c \|A_h^{s/2} f_h\|_0 \quad \forall f_h \in V_h, \quad s = 1, -1.$$

Finally, we will use the following inequalities whose proof can be obtained applying [32, Lemma 4.4]

$$(2.16) \quad \|v_h\|_\infty \leq C \|A_h v_h\|_0 \quad \forall v_h \in V_h,$$

$$(2.17) \quad \|\nabla v_h\|_{L^3} \leq C \|\nabla v_h\|_0^{1/2} \|A_h v_h\|_0^{1/2} \quad \forall v_h \in V_h.$$

We consider the finite-element approximation (u_h, p_h) to (u, p) , solution of (1.1)–(1.3). That is, given $u_h(0) = \Pi_h u_0$, we compute $u_h(t) \in X_{h,r}$ and $p_h(t) \in Q_{h,r-1}$, $t \in (0, T]$, satisfying

$$(2.18) \quad (\dot{u}_h, \phi_h) + (\nabla u_h, \nabla \phi_h) + b(u_h, u_h, \phi_h) + (\nabla p_h, \phi_h) = (f, \phi_h) \quad \forall \phi_h \in X_{h,r},$$

$$(2.19) \quad (\nabla \cdot u_h, \psi_h) = 0 \quad \forall \psi_h \in Q_{h,r-1}.$$

For convenience, we rewrite this problem in the following way,

$$(2.20) \quad \dot{u}_h + A_h u_h + B_h(u_h, u_h) = \Pi_h f, \quad u_h(0) = \Pi_h u_0,$$

where $B_h(u, v) = \Pi_h F(u, v)$.

For $r = 2, 3, 4, 5$, provided that (2.11)–(2.13) hold for $l \leq r$, and (2.4)–(2.5) hold for $k = r$, then we have

$$(2.21) \quad \|u(t) - u_h(t)\|_0 + h \|u(t) - u_h(t)\|_1 \leq C \frac{h^r}{t^{(r-2)/2}}, \quad 0 \leq t \leq T,$$

(see, e.g., [18, 32, 33]), and also,

$$(2.22) \quad \|p(t) - p_h(t)\|_{L^2/\mathbb{R}} \leq C \frac{h^{r-1}}{t^{(r'-2)/2}}, \quad 0 \leq t \leq T,$$

where $r' = r$ if $r \leq 4$ and $r' = r + 1$ if $r = 5$. Results in [18] hold for h sufficiently small. In the rest of the paper we assume h to be small enough for (2.21)–(2.22) to hold.

Observe that from (2.21) and (2.2) it follows that $\|u_h(t)\|_1$ is bounded for $0 \leq t \leq T$. However, further bounds for $u_h(t)$ will be needed in the present paper, so we recall the following result, which, since we are considering finite times $0 < T < +\infty$, it is a rewriting of [34, Proposition 3.2].

PROPOSITION 2.1. *Let the forcing term f in (1.1) satisfy (2.3). Then, there exists a constant $\tilde{M}_3 > 0$, depending only on \tilde{M}_2 , $\|A_h u_h(0)\|_0$ and $\sup_{0 \leq t \leq T} \|u_h(t)\|_1$, such that the following bounds hold for $0 \leq t \leq T$:*

$$(2.23) \quad F_{0,2}(t) \equiv \|A_h u_h(t)\|_0^2 \leq \tilde{M}_3^2,$$

$$(2.24) \quad F_{1,r}(t) \equiv t^r \|A_h^{r/2} \dot{u}_h(t)\|_0^2 \leq \tilde{M}_3^2, \quad r = 0, 1, 2,$$

$$(2.25) \quad F_{2,r}(t) \equiv t^{r+2} \|A_h^{r/2} \ddot{u}_h(t)\|_0^2 \leq \tilde{M}_3^2, \quad r = -1, 0, 1,$$

$$(2.26) \quad I_{1,r}(t) \equiv \int_0^t s^{r-1} \|A_h^{r/2} \dot{u}_h(s)\|_0^2 ds \leq \tilde{M}_3^2, \quad r = 1, 2,$$

$$(2.27) \quad I_{2,r}(t) \equiv \int_0^t s^{r+1} \|A_h^{r/2} \ddot{u}_h(s)\|_0^2 ds \leq \tilde{M}_3^2, \quad r = -1, 0, 1.$$

2.3. The postprocessed method. This method obtains for any $t \in (0, T]$ an improved approximation by solving the following discrete Stokes problem: find $(\tilde{u}_h(t), \tilde{p}_h(t)) \in (\tilde{X}, \tilde{Q})$ satisfying

$$(2.28) \quad \begin{aligned} (\nabla \tilde{u}_h(t), \nabla \tilde{\phi}) + (\nabla \tilde{p}_h(t), \tilde{\phi}) &= (f, \tilde{\phi}) - b(u_h(t), u_h(t), \tilde{\phi}) \\ &\quad - (\dot{u}_h(t), \tilde{\phi}) \quad \forall \tilde{\phi} \in \tilde{X}, \end{aligned}$$

$$(2.29) \quad (\nabla \cdot \tilde{u}_h(t), \tilde{\psi}) = 0 \quad \forall \tilde{\psi} \in \tilde{Q},$$

where (\tilde{X}, \tilde{Q}) is either:

- (a) The same-order MFE over a finer grid. That is, for $h' < h$, we choose $(\tilde{X}, \tilde{Q}) = (X_{h',r}, Q_{h',r-1})$.
- (b) A higher-order MFE over the same grid. In this case we choose $(\tilde{X}, \tilde{Q}) = (X_{h,r+1}, Q_{h,r})$.

In both cases, we will denote by \tilde{V} the corresponding discretely divergence-free space that can be either $\tilde{V} = V_{h',r}$ or $\tilde{V} = V_{h,r+1}$ depending on the selection of the postprocessing space. The discrete orthogonal projection into \tilde{V} will be denoted by $\tilde{\Pi}_h$, and we will represent by \tilde{A}_h the discrete Stokes operator acting on functions in \tilde{V} . Notice then that from (2.28) it follows that $\tilde{u}_h(t) \in \tilde{V}$ and it satisfies

$$(2.30) \quad \tilde{A}_h \tilde{u}_h(t) = \tilde{\Pi}_h (f - F(u_h(t), u_h(t)) - \dot{u}_h(t)).$$

In [18] the following result is proved.

THEOREM 2.2. *Let (u, p) be the solution of (1.1)–(1.3) and for $r = 3, 4$, let (2.4)–(2.5) hold with $k = r + 2$ and let (2.11) hold for $2 \leq l \leq r$. Then, there exists a*

positive constant C such that the postprocessed MFE approximation to u , \tilde{u}_h satisfies the following bounds for $r = 3, 4$ and $t \in (0, T]$:

(i) if the postprocessing element is $(\tilde{X}, \tilde{Q}) = (X_{h',r}, Q_{h',r-1})$, then

$$(2.31) \quad \|u(t) - \tilde{u}_h(t)\|_j \leq \frac{C}{t^{(r-2)/2}}(h')^{r-j} + \frac{C}{t^{(r-1)/2}}h^{r+1-j}|\log(h)|, \quad j = 0, 1,$$

$$(2.32) \quad \|p(t) - \tilde{p}_h(t)\|_{L^2/\mathbb{R}} \leq \frac{C}{t^{(r-2)/2}}(h')^{r-1} + \frac{C}{t^{(r-1)/2}}h^r|\log(h)|,$$

(ii) if the postprocessing element is $(\tilde{X}, \tilde{Q}) = (X_{h,r+1}, Q_{h,r})$, then

$$(2.33) \quad \|u(t) - \tilde{u}_h(t)\|_j \leq \frac{C}{t^{(r-1)/2}}h^{r+1-j}|\log(h)|, \quad j = 0, 1,$$

$$(2.34) \quad \|p(t) - \tilde{p}_h(t)\|_{L^2/\mathbb{R}} \leq \frac{C}{t^{(r-1)/2}}h^r|\log(h)|.$$

Since the constant C depends on the type of element used, the result is stated for a particular kind of MFE methods, but it applies to any kind of MFE method satisfying the *LBB* condition (2.7), the approximation properties (2.11)–(2.13), as well as negative norm estimates, that is,

$$\|u - s_h\|_{-m} \leq Ch^{l+\min(m,r-2)}(\|u\|_l + \|p\|_{H^{l-1}/\mathbb{R}})$$

for $m = 1, 2$ and $1 \leq l \leq r$. For these negative norm estimates to hold, it is necessary on the one hand that Ω is of class C^{2+m} , and, on the other hand, that $X_{h,r} \subset H_0^1(\Omega)^d$, so that $X_{h,r}$ consists of continuous functions vanishing on $\partial\Omega$ (i.e., discontinuous elements are excluded).

As pointed out in [18, Remark 4.2], with a much simpler analysis than that needed to prove Theorem 2.2, together with results in [2], the previous result applies to the so-called mini element ($r = 2$) but excluding the case $j = 0$ (L^2 errors) in (2.31) and (2.33).

3. Analysis of fully discrete postprocessed methods.

3.1. The general case. As mentioned in the Introduction, in practice, it is hardly ever possible to compute the MFE approximation exactly, and, instead, some time-stepping procedure must be used to approximate the solution of (2.18)–(2.19). Hence, for some time levels $0 = t_0 < t_1 < \dots < t_N = T$, approximations $U_h^{(n)} \approx u_h(t_n)$ and $P_h^{(n)} \approx p_h(t_n)$ are obtained. Then, given an approximation $d_t^*U_h^{(n)}$ to $\dot{u}_h(t_n)$, the fully discrete postprocessed approximation $(\tilde{U}_h^{(n)}, \tilde{P}_h^{(n)})$ is obtained as the solution of the following Stokes problem:

$$(3.1) \quad \left(\nabla \tilde{U}_h^{(n)}, \nabla \tilde{\phi} \right) + \left(\nabla \tilde{P}_h^{(n)}, \tilde{\phi} \right) = \left(f, \tilde{\phi} \right) - b \left(U_h^{(n)}, U_h^{(n)}, \tilde{\phi} \right) - \left(d_t^* U_h^{(n)}, \tilde{\phi} \right) \quad \forall \tilde{\phi} \in \tilde{X},$$

$$(3.2) \quad \left(\nabla \cdot \tilde{U}_h^{(n)}, \tilde{\psi} \right) = 0 \quad \forall \tilde{\psi} \in \tilde{Q},$$

where (\tilde{X}, \tilde{Q}) is as in (2.28)–(2.29). Notice then that $\tilde{U}_h^{(n)} \in \tilde{V}$ and it satisfies

$$(3.3) \quad \tilde{A}_h \tilde{U}_h^{(n)} = \tilde{\Pi}_h \left(f - F \left(U_h^{(n)}, U_h^{(n)} \right) - d_t^* U_h^{(n)} \right).$$

For reasons already analyzed in [17] and confirmed in the arguments that follow, we propose

$$(3.4) \quad d_t^* U_h^{(n)} = \Pi_h f - A_h U_h^{(n)} - B_h \left(U_h^{(n)}, U_h^{(n)} \right)$$

as an adequate approximation to the time derivative $\dot{u}_h(t_n)$, which is very convenient from the practical point of view.

We decompose the errors $u(t) - \tilde{U}_h^{(n)}$ and $p(t) - \tilde{P}_h^{(n)}$ as follows,

$$(3.5) \quad u(t) - \tilde{U}_h^{(n)} = (u(t) - \tilde{u}_h(t_n)) + \tilde{e}_n,$$

$$(3.6) \quad p(t_n) - \tilde{P}_h^{(n)} = (p(t_n) - \tilde{p}_h(t_n)) + \tilde{\pi}_n,$$

where $\tilde{e}_n = \tilde{u}_h(t_n) - \tilde{U}_h^{(n)}$ and $\tilde{\pi}_n = \tilde{p}_h(t_n) - \tilde{P}_h^{(n)}$ are the temporal errors of the fully discrete postprocessed approximation $(\tilde{U}_h^{(n)}, \tilde{P}_h^{(n)})$. The first terms on the right-hand sides of (3.5)–(3.6) are the errors of the (semidiscrete) postprocessed approximation whose size is estimated in Theorem 2.2. In the present section we analyze the time discretization errors \tilde{e}_n and $\tilde{\pi}_n$.

To estimate the size of \tilde{e}_n and $\tilde{\pi}_n$, we bound them in terms of

$$e_n = u_h(t_n) - U_h^{(n)},$$

the temporal error of the MFE approximation. We do this for any time-stepping procedure satisfying the following assumption

$$(3.7) \quad \lim_{k \rightarrow 0} \max_{0 \leq n \leq N} \|e_n\|_0 = 0, \quad \text{and} \quad \limsup_{k \rightarrow 0} \max_{0 \leq n \leq N} \|e_n\|_1 = O(1),$$

where $k = \max\{t_n - t_{n-1} \mid 1 \leq n \leq N\}$. Bounds for $\|e_n\|_0$ and $\|e_n\|_1$ of size $O(k^2/t_n)$ and $O(k/t_n^{1/2})$, respectively, have been proven for the Crank–Nicolson method in [34] (see also [41]). The arguments in [10] can be adapted to show that, for the two-step BDF, $\|e_n\|_j \leq Ck^{2-j}/t_n$, for $2 \leq n \leq N$, $j = 0, 1$ (although in section 5 we shall obtain sharper bounds of $\|e_n\|_1$). For problems in two spatial dimensions, bounds for a variety of methods can be found in the literature (see a summary in [31]).

In the arguments in the present section we use the following inequalities [34, (3.7)] which hold for all $v_h, w_h \in V_h$ and $\phi \in H_0^1(\Omega)^d$:

$$(3.8) \quad |b(v_h, v_h, \phi)| \leq c \|v_h\|_1^{3/2} \|A_h v_h\|_0^{1/2} \|\phi\|_0,$$

$$(3.9) \quad |b(v_h, w_h, \phi)| + |b(w_h, v_h, \phi)| \leq c \|v_h\|_1 \|A_h w_h\|_0 \|\phi\|_0,$$

$$(3.10) \quad |b(v_h, w_h, \phi)| + |b(w_h, v_h, \phi)| \leq \|v_h\|_1 \|w_h\|_1 \|\phi\|_1.$$

PROPOSITION 3.1. *Let (2.11) hold for $l = 2$. Then, there exists a positive constant $C = C(\max_{0 \leq t \leq T} \|A_h u_h(t)\|_0)$ such that*

$$(3.11) \quad \|\tilde{e}_n - e_n\|_j \leq Ch^{2-j} (\|e_n\|_1 + \|e_n\|_1^3 + \|A_h e_n\|_0), \quad j = 0, 1, \quad 1 \leq n \leq N,$$

$$(3.12) \quad \|\tilde{\pi}_n\|_{L^2(\Omega)/\mathbb{R}} \leq C(\|\tilde{e}_n\|_1 + \|e_n\|_1 + \|e_n\|_1^2), \quad 1 \leq n \leq N.$$

Proof. From (3.4) and (2.20) it follows that

$$(3.13) \quad \dot{u}_h(t_n) - d_t^* U_h^{(n)} = -A_h e_n + \Pi_h \left(F \left(U_h^{(n)}, U_h^{(n)} \right) - F(u_h(t_n), u_h(t_n)) \right),$$

so that subtracting (3.3) from (2.30) and multiplying by \tilde{A}_h^{-1} we get

$$(3.14) \quad \tilde{e}_n = -\tilde{A}_h^{-1}\tilde{\Pi}_h(I - \Pi_h)g + \tilde{A}_h^{-1}\tilde{\Pi}_h A_h e_n,$$

where $g = F(u_h(t_n), u_h(t_n)) - F(U_h^{(n)}, U_h^{(n)})$. By writing

$$\tilde{A}_h^{-1}\tilde{\Pi}_h A_h e_n = e_n + \left(\tilde{A}_h^{-1}\tilde{\Pi}_h - A_h^{-1}\right) A_h e_n,$$

and applying (2.14) we get

$$(3.15) \quad \|\tilde{e}_n - e_n\|_j \leq \left\| \tilde{A}_h^{-1}\tilde{\Pi}_h(I - \Pi_h)g \right\|_j + Ch^{2-j}\|A_h e_n\|_0, \quad j = 0, 1.$$

Similarly, for g we write

$$(3.16) \quad \tilde{A}_h^{-1}\tilde{\Pi}_h(I - \Pi_h)g = \left(\tilde{A}_h^{-1}\tilde{\Pi}_h - A^{-1}\Pi\right)(I - \Pi_h)g + A^{-1}\Pi(I - \Pi_h)g.$$

In order to bound the first term on the right-hand side above we first apply (2.14), and then we observe that $\|(I - \Pi_h)g\|_0 \leq \|g\|_0$. For the second term on the right-hand side of (3.16), we may use a simple duality argument and (2.8), so that we have $\|\tilde{A}_h^{-1}\tilde{\Pi}_h(I - \Pi_h)g\|_j \leq Ch^{2-j}\|g\|_0$. Now, by writing g as

$$(3.17) \quad g = F(e_n, u_h(t_n)) + F(u_h(t_n), e_n) - F(e_n, e_n),$$

a duality argument and (3.8)–(3.9) show that

$$\left\| \tilde{A}_h^{-1}\tilde{\Pi}_h(I - \Pi_h)g \right\|_j \leq Ch^{2-j} \left(\|A_h u_h(t_n)\|_0 \|e_n\|_1 + \|e_n\|_1^{3/2} \|A_h e_n\|_0^{1/2} \right).$$

Applying Hölder’s inequality to the last term on the right-hand side above, the bound (3.11) follows from (3.14) and (3.15).

For the pressure, subtracting (3.1) from (2.28) and recalling (3.13) we have

$$\left(\tilde{\pi}_n, \nabla \cdot \tilde{\phi}\right) = \left(\nabla \tilde{e}_n, \nabla \tilde{\phi}\right) + \left(g, \tilde{\phi}\right) + \left(\dot{u}_h - d_t^* U_h^{(n)}, \tilde{\phi}\right),$$

for all $\tilde{\phi} \in \tilde{X}$, where g is as in (3.17). Then, thanks to the inf-sup condition (2.7), we have

$$\|\tilde{\pi}_n\|_{L^2(\Omega)/\mathbb{R}} \leq C \left(\|\tilde{e}_n\|_1 + \sup_{\tilde{\phi} \in \tilde{X}} \frac{|(g, \tilde{\phi})|}{\|\tilde{\phi}\|_1} + \left\| \dot{u}_h(t_n) - d_t^* U_h^{(n)} \right\|_{-1} \right).$$

Taking into account the expression of g in (3.17) and applying (3.10) it follows that

$$\|\tilde{\pi}_n\|_{L^2(\Omega)/\mathbb{R}} \leq C \left(\|\tilde{e}_n\|_1 + \|e_n\|_1 (\|u_h(t_n)\|_1 + \|e_n\|_1) + \left\| \dot{u}_h - d_t^* U_h^{(n)} \right\|_{-1} \right),$$

so that (3.12) follows by applying Lemma 3.2 below, (2.15), and using the fact that $\|u_h(t_n)\|_1 \leq C\|A_h u_h(t_n)\|_0$. \square

LEMMA 3.2. *Under the hypotheses of Proposition 3.1, there exists a constant $C = C(\max_{0 \leq t \leq T} \|u_h(t)\|_1) > 0$ such that the following bound holds for $1 \leq n \leq N$:*

$$(3.18) \quad \left\| \dot{u}_h - d_t^* U_h^{(n)} \right\|_{-1} \leq C \left\| A_h^{1/2} e_n \right\|_0 \left(1 + \|A_h^{1/2} e_n\|_0 \right).$$

Proof. Since $\dot{u}_h - d_t^* U_h^{(n)} \in V_h$, we have

$$\left\| \dot{u}_h - d_t^* U_h^{(n)} \right\|_{-1} \leq C \left\| A_h^{-1/2} \left(\dot{u}_h - d_t^* U_h^{(n)} \right) \right\|_0,$$

due to (2.15). Thus, in view of (3.13), the lemma is proved if for

$$g = F(u_h(t_n), u_h(t_n)) - F \left(U_h^{(n)}, U_h^{(n)} \right),$$

we show that $\|A_h^{-1/2} \Pi_h g\|_0$ can be bounded by the right-hand side of (3.18). If we write g as in (3.17), a simple duality argument, (3.10), and the equivalence (2.15) show that, indeed,

$$\left\| A_h^{-1/2} \Pi_h g \right\|_0 \leq C \|e_n\|_1 (\|u_h(t_n)\|_1 + \|e_n\|_1).$$

Since according to (2.15), $\|e_n\|_1$ and $\|A_h^{1/2} e_n\|_0$ are equivalent, then the result follows. \square

Since we are assuming that the meshes are quasiuniform and, hence, both (2.6) and (2.15) hold, we have $Ch^{2-j} \|A_h e_n\|_0 \leq C \|e_n\|_j$ and $Ch \|e_n\|_1 \leq C \|e_n\|_0$. Thus, from (3.11)–(3.12) and (3.7) the following result follows.

THEOREM 3.3. *Let (2.11) hold for $l = 2$ and let (2.3) hold. Then there exists a positive constant C depending on $\max\{F_{2,0}(t) \mid 0 \leq t \leq T\}$, such that if the errors $e_n = u_h(t_n) - U_h^{(n)}$, $1 \leq n \leq N$ of any approximation $U_h^{(n)} \approx u_h(t_n)$ for $0 = t_0 < \dots < t_N$ satisfy (3.7), then the (fully discrete) postprocessed approximations $(\tilde{U}_h^{(n)}, \tilde{P}_h^{(n)})$ solution of (3.1)–(3.2) satisfy*

$$(3.19) \quad \left\| \tilde{u}_h(t_n) - \tilde{U}_h^{(n)} \right\|_j \leq C \left\| u_h(t_n) - U_h^{(n)} \right\|_j, \quad 1 \leq n \leq N, \quad j = 0, 1,$$

$$(3.20) \quad \left\| \tilde{p}_h(t_n) - \tilde{P}_h^{(n)} \right\|_{L^2(\Omega)/\mathbb{R}} \leq C \left\| u_h(t_n) - U_h^{(n)} \right\|_1, \quad 1 \leq n \leq N,$$

for k sufficiently small, where $(\tilde{u}_h(t_n), \tilde{p}_h(t_n))$ is the (semidiscrete) postprocessed approximation defined in (2.28)–(2.29).

3.2. The case of the BDF. Better estimates than (3.19) can be obtained when $\|A_h e_n\|_0$ can be shown to decay with k at the same rate as $\|e_n\|_0$. As mentioned in the introduction this will be shown to be the case of two (fixed time-step) time integration procedures in section 5: the backward Euler method and the two-step BDF [9] (see also [25, section III.1]). We describe them now.

For $N \geq 2$ integer, we fix $k = \Delta t = T/N$, and we denote $t_n = nk$, $n = 0, 1, \dots, N$. For a sequence $(y_n)_{n=0}^N$ we denote

$$Dy_n = y_n - y_{n-1}, \quad n = 1, 2, \dots, N.$$

Given $U_h^{(0)} = u_h(0)$, a sequence $(U_h^{(n)}, P_h^{(n)})$ of approximations to $(u_h(t_n), p_h(t_n))$, $n = 1, \dots, N$, is obtained by means of the following recurrence relation:

$$(3.21) \quad \left(d_t U_h^{(n)}, \phi_h \right) + \left(\nabla U_h^{(n)}, \nabla \phi_h \right) + b \left(U_h^{(n)}, U_h^{(n)}, \phi_h \right) - \left(P_h^{(n)}, \nabla \cdot \phi_h \right) = (f, \phi_h) \quad \forall \phi_h \in X_{h,r},$$

$$(3.22) \quad \left(\nabla \cdot U_h^{(n)}, \psi_h \right) = 0, \quad \forall \psi_h \in Q_{h,r-1},$$

where $d_t = k^{-1}D$ in the case of the backward Euler method and $d_t = k^{-1}(D + \frac{1}{2}D^2)$ for the two-step BDF. In this last case, a second starting value $U_h^{(1)}$ is needed. In the present paper, we will always assume that $U_h^{(1)}$ is obtained by one step of the backward Euler method. Also, for both the backward Euler and the two-step BDF, we assume that $U_h^{(0)} = u_h(0)$, which is usually the case in practical situations.

In order to cope for the minor differences between the two methods, we set

$$(3.23) \quad l_0 = \begin{cases} 1, & \text{for the backward Euler method,} \\ 2, & \text{for the two-step BDF.} \end{cases}$$

Under these conditions, we show in Lemma 5.2 and Theorems 5.4 and 5.7 in section 5 that the errors e_n of these two time integration procedures satisfy that

$$(3.24) \quad \|e_n\|_0 + t_n \|A_h e_n\|_0 \leq C_{l_0} \frac{k^{l_0}}{t_n^{l_0-1}}, \quad 1 \leq n \leq N,$$

for a certain constants C_1 and C_2 . These are, respectively, the terms between parentheses in (5.23) and (5.33) below, which as Proposition 2.1 above and Lemma 4.3 below show, can be bounded for $T > 0$ fixed. Thus, from Proposition 3.1 and (3.24) the following result follows readily.

THEOREM 3.4. *Under the hypotheses of Proposition 3.1, let the approximations $U_h^{(n)}$, $n = 1, \dots, N$ be obtained by either the backward Euler method or the two-step BDF under the conditions stated above. Then, there exist positive constants $C'_l = C(C_l)$, for $l = 1, 2$, and k' , such that for $k < k'$ the temporal errors \tilde{e}_n of the fully discrete postprocessed approximation satisfy that $\tilde{e}_n = e_n + r_n$, and*

$$\|r_n\|_j \leq C'_{l_0} h^{2-j} \frac{k^{l_0}}{t_n^{l_0}}, \quad j = 0, 1, \quad 1 \leq n \leq N.$$

We remark that a consequence of the above result is that for these two methods the temporal errors of the postprocessed and MFE approximations are asymptotically the same as $h \rightarrow 0$. This allows to use the difference $\gamma_h^{(n)} = \tilde{U}_h^{(n)} - U_h^{(n)}$ as an a posteriori error estimator of the spatial error of the MFE approximation, since, as shown in [15, 16, 17], its size is that of $u(t) - u_h(t)$ so long as the spatial and temporal errors are not too unbalanced.

We also remark that at a price of lengthening the already long and elaborate analysis in the present paper, variable stepsizes could have been considered following ideas in [4], but, for the sake of simplicity we consider only fixed stepsize in the analysis that follows.

4. Technical results.

4.1. Inequalities for the nonlinear term. We now obtain several estimates for the quadratic form $B_h(v, w) = \Pi_h F(u, v)$ that will be frequently used in our analysis. We start by proving an auxiliary result.

LEMMA 4.1. *Let (2.11) hold for $l = 2$. Then, the following bound holds for any f_h, g_h , and ψ_h in V_h :*

$$(4.1) \quad |b(f_h, g_h, \psi_h)| + |b(g_h, f_h, \psi_h)| \leq C \|A_h f_h\|_0 \|A_h^{-1/2} g_h\|_0 \|A_h \psi_h\|_0.$$

Proof. To prove this bound we will use the following identity,

$$I = A^{-1} \Pi A_h + (A_h^{-1} - A^{-1} \Pi) A_h.$$

It will be applied to either f_h or ψ_h whenever any of their derivatives appears in the expressions of $b(f_h, g_h, \psi_h)$ and $b(g_h, f_h, \psi_h)$. We deal first with the second term on the left-hand side of (4.1). Integrating by parts we may write

$$\begin{aligned} b(g_h, f_h, \psi_h) &= \frac{1}{2}((g_h \cdot \nabla)f_h, \psi_h) - \frac{1}{2}((g_h \cdot \nabla)\psi_h, f_h) \\ &= \frac{1}{2}((g_h \cdot \nabla)A^{-1}\Pi A_h f_h, \psi_h) - \frac{1}{2}((g_h \cdot \nabla)A^{-1}\Pi A_h \psi_h, f_h) \\ &\quad + \frac{1}{2}((g_h \cdot \nabla)(A_h^{-1} - A^{-1}\Pi)A_h f_h, \psi_h) \\ &\quad - \frac{1}{2}((g_h \cdot \nabla)(A_h^{-1} - A^{-1}\Pi)A_h \psi_h, f_h). \end{aligned}$$

Using (2.14) with $j = 1$ and (2.16), the last two terms on the right-hand side above can be bounded by

$$Ch(\|A_h f_h\|_0 \|\psi_h\|_\infty + \|A_h \psi_h\|_0 \|f_h\|_\infty) \|g_h\|_0 \leq Ch \|A_h f_h\|_0 \|A_h \psi_h\|_0 \|g_h\|_0.$$

By writing $\|g_h\|_0 \leq \|A_h^{1/2}\|_0 \|A_h^{-1/2} g_h\|_0 \leq Ch^{-1} \|A_h^{-1/2} g_h\|_0$, we thus have

$$(4.2) \quad \begin{aligned} |b_h(g_h, f_h, \psi_h)| &\leq \frac{1}{2} |((g_h \cdot \nabla)A^{-1}\Pi A_h f_h, \psi_h)| + \frac{1}{2} |((g_h \cdot \nabla)A^{-1}\Pi A_h \psi_h, f_h)| \\ &\quad + C \|A_h f_h\|_0 \|A_h \psi_h\|_0 \left\| A_h^{-1/2} g_h \right\|_0. \end{aligned}$$

Now, applying Hölder's inequality it easily follows that

$$\begin{aligned} |((g_h \cdot \nabla)A^{-1}\Pi f_h, \psi_h)| &\leq \\ &C \|g_h\|_{-1} (\|A^{-1}\Pi A_h f_h\|_2 \|\psi_h\|_\infty + \|\nabla A^{-1}\Pi A_h f_h\|_{L^6} \|\nabla \psi_h\|_{L^3}), \end{aligned}$$

so that, applying (2.16)–(2.17) and regularity estimates for the Stokes problem, and standard Sobolev's inequalities we have

$$(4.3) \quad |((g_h \cdot \nabla)A^{-1}\Pi f_h, \psi_h)| \leq C \|g_h\|_{-1} \|A_h f_h\|_0 \|A_h \psi_h\|_0.$$

Also, arguing similarly, $|((g_h \cdot \nabla)A^{-1}\Pi A_h \psi_h, f_h)| \leq C \|g_h\|_{-1} \|A_h f_h\|_0 \|A_h \psi_h\|_0$, so that from (4.2) and (4.3) it follows that

$$|b(g_h, f_h, \psi_h)| \leq C \left(\|g_h\|_{-1} + \left\| A_h^{-1/2} g_h \right\|_0 \right) \|A_h f_h\|_0 \|A_h \psi_h\|_0.$$

Now, recalling the equivalence (2.15) we have

$$(4.4) \quad |b(g_h, f_h, \psi_h)| \leq C \left\| A_h^{-1/2} g_h \right\|_0 \|A_h f_h\|_0 \|A_h \psi_h\|_0.$$

For the second term on the left-hand side of (4.1), thanks to (2.1) we may write

$$\begin{aligned} |b(f_h, g_h, \psi_h)| &\leq |((f_h \cdot \nabla)A^{-1}\Pi A_h \psi_h, g_h)| + |((\nabla \cdot A^{-1}\Pi A_h f_h) \psi_h, g_h)| \\ &\quad + |((f_h \cdot \nabla)(A_h^{-1} - A^{-1}\Pi)A_h \psi_h, g_h)| \\ &\quad + |((\nabla \cdot (A_h^{-1} - A^{-1}\Pi)A_h f_h) \psi_h, g_h)|, \end{aligned}$$

so that

$$\begin{aligned} |b(f_h, g_h, \psi_h)| &\leq (\|(f_h \cdot \nabla)A^{-1}\Pi A_h \psi_h\|_1 + \|(\nabla \cdot A^{-1}\Pi A_h f_h) \psi_h\|_1) \|g_h\|_{-1} \\ &\quad + Ch \|A_h f_h\|_0 \|A_h \psi_h\|_0 \|g_h\|_0. \end{aligned}$$

Then, recalling that $\|g_h\|_0 \leq Ch^{-1}\|A_h^{-1/2}g_h\|_0$ and (2.15), arguments like those used from (4.2) to (4.4) also show that

$$|b_h(f_h, g_h, \psi_h)| \leq C\|A_h^{-1/2}g_h\|_0\|A_h f_h\|_0\|A_h \psi_h\|_0,$$

so that, in view of (4.4), the proof of (4.1) is finished. \square

LEMMA 4.2. *Under the conditions of Lemma 4.1, there exists a constant $C > 0$ such that the following bounds hold for $v_h, w_h \in V_h$*

$$(4.5) \quad \left\|A_h^{j/2}B_h(v_h, w_h)\right\|_0 + \left\|A_h^{j/2}B_h(w_h, v_h)\right\|_0 \leq C\left\|A_h^{(j+1)/2}v_h\right\|_0\|A_h w_h\|_0,$$

for $j = -2, -1, 0, 1$, and

$$(4.6) \quad \|B_h(v_h, v_h)\|_0 \leq C\|v_h\|_1^{3/2}\|A_h v_h\|_0^{1/2},$$

$$(4.7) \quad \|A_h^{-1}B_h(v_h, v_h)\|_0 \leq C\|v_h\|_0\|v_h\|_1.$$

Proof. The cases $j = -1, 0$ in (4.5) as well as (4.6) are easily deduced from the fact that for every $v_h \in V_h$, $\|A_h^{1/2}v_h\|_0 = \|\nabla v_h\|_0$, (2.16), and from standard bounds (e.g., (3.8), [34, (3.7)]).

If we denote $f_h = w_h$, $g_h = v_h$, and, for $\phi_h \in V_h$ $\psi_h = A_h^{-1}\phi_h$, case $j = -2$ in (4.5) is a direct consequence of standard duality arguments and (4.1). Also, arguing by duality the bound (4.7) is a straightforward consequence of well-known bounds for the trilinear form b (e.g., [34, (3.7)]).

Finally, for the case $j = 1$ in (4.5), we argue by duality. For $\phi_h \in V_h$, thanks to (2.1), we have

$$b(v_h, w_h, A_h^{1/2}\phi_h) + b(w_h, v_h, A_h^{1/2}\phi_h) = -b(v_h, A_h^{1/2}\phi_h, w_h) - b(w_h, A_h^{1/2}\phi_h, v_h),$$

so that, by denoting $g_h = A_h^{1/2}\phi_h$, the case $j = 1$ in (4.5) is a direct consequence of (4.1). \square

4.2. Further a priori estimates for the finite-element solution. We maintain the notation tacitly introduced in Proposition 2.1,

$$F_{l,r} = t^{r+2(l-1)}\left\|A_h^{r/2}\frac{d^l}{ds^l}u_h(t)\right\|_0^2, \quad \text{and} \quad I_{l,r} = \int_0^t s^{r+2l-3}\left\|A_h^{r/2}\frac{d^l}{ds^l}u_h(t)\right\|_0^2 ds.$$

LEMMA 4.3. *Under the conditions of Proposition 2.1, there exists a positive constant \tilde{M}_4 such that for $0 \leq t \leq T$ the following bounds hold:*

$$(4.8) \quad F_{2,-2}(t) = \|A_h^{-1}\ddot{u}_h(t)\|_0 \leq \tilde{M}_4,$$

$$(4.9) \quad I_{3,-3}(t) = \int_0^t \left\|A_h^{-3/2}\ddot{\ddot{u}}_h(s)\right\|_0^2 ds \leq \tilde{M}_4,$$

$$(4.10) \quad I_{2,2}(t) = \int_0^t s^3 \|A_h \ddot{u}_h(s)\|_0^2 ds \leq \tilde{M}_4,$$

$$(4.11) \quad I_{3,-1}(t) = \int_0^t s^2 \left\|A_h^{-1/2}\ddot{\ddot{u}}_h(s)\right\|_0^2 ds \leq \tilde{M}_4,$$

$$(4.12) \quad I_{3,1}(t) + F_{2,2}(t) = \int_0^t s^4 \left\|A_h^{1/2}\ddot{\ddot{u}}_h(s)\right\|_0^2 ds + t^4 \|A_h \ddot{u}_h(t)\|_0^2 \leq \tilde{M}_4.$$

Proof. Taking derivatives with respect to t in (2.20) and multiplying by A_h^{-1} we have

$$A_h^{-1}\ddot{u}_h = A_h^{-1}\Pi_h f_t - \dot{u}_h - A_h^{-1}(B_h(\dot{u}_h, u_h) + B_h(u_h, \dot{u}_h)).$$

Applying Lemma 4.2 we have

$$\|A_h^{-1}(B_h(\dot{u}_h, u_h) + B_h(u_h, \dot{u}_h))\|_0 \leq C\|A_h u_h\|_0 \|A_h^{-1/2}\dot{u}_h\|_0 \leq C\|A_h u_h\|_0 \|\dot{u}_h\|_0,$$

so that (4.8) follows from (2.3), (2.23), and (2.24) with $r = 0$.

We now prove (4.9). Taking derivatives twice with respect to t in (2.20) and multiplying by $A_h^{-3/2}$ we have

$$(4.13) \quad A_h^{-3/2}\ddot{u}_h = A_h^{-3/2}\Pi_h f_{tt} - A_h^{-1/2}\ddot{u}_h - A_h^{-3/2}\frac{d^2}{dt^2}B_h(u_h, u_h).$$

Taking into account that for $v_h \in V_h$, we have $\|A_h^{-3/2}v_h\|_0 \leq C\|A_h^{-1}v_h\|_0$, and that

$$(4.14) \quad \frac{d^2}{dt^2}B_h(u_h, u_h) = B_h(\ddot{u}_h, u_h) + 2B_h(\dot{u}_h, \dot{u}_h) + B_h(u_h, \ddot{u}_h),$$

then, applying the bound (4.5) for $j = -2$ with $v_h = \ddot{u}_h$, and $w_h = u_h$, on the one hand, and, on the other, (4.7) with $v_h = \dot{u}_h$, it follows

$$\left\|A_h^{-3/2}\frac{d^2}{dt^2}B_h(u_h, u_h)\right\|_0 \leq C\|A_h u_h\|_0 \left\|A_h^{-1/2}\ddot{u}_h\right\|_0 + \|\dot{u}_h\|_0 \|\dot{u}_h\|_1,$$

so that the bound (4.9) follows from (4.13), (2.3), and the fact that $A_h^{-3/2}$ is bounded independently of h , together with (2.23), (2.24) with $r = 0$, (2.26) with $r = 1$, and (2.27) with $r = -1$.

We now prove (4.10). Taking derivatives twice with respect to t in (2.18) and then setting $\phi = t^3 A_h \ddot{u}_h$, we have

$$\frac{1}{2}t^3 \frac{d}{dt} \left\|A_h^{1/2}\ddot{u}_h\right\|_0^2 + t^3 \|A_h \ddot{u}_h\|_0^2 = t^3 \left(f_{tt} - \frac{d^2}{dt^2}B_h(u_h, u_h), A_h \ddot{u}_h \right).$$

Since $|(f_{tt} - \frac{d^2}{dt^2}B_h(u_h, u_h), A_h \ddot{u}_h)| \leq \|f_{tt}\|_0^2 + \|\frac{d^2}{dt^2}B_h(u_h, u_h)\|_0^2 + \frac{1}{2}\|A_h \ddot{u}_h\|_0^2$, it follows that

$$(4.15) \quad \frac{d}{dt} \left(t^3 \|A_h^{1/2}\ddot{u}_h\|_0^2 \right) + t^3 \|A_h \ddot{u}_h\|_0^2 \leq 2t^3 \|f_{tt}\|_0^2 \left\| \frac{d^2}{dt^2}B_h(u_h, u_h) \right\|_0^2 + 3t^2 \left\| A_h^{1/2}\ddot{u}_h \right\|_0^2.$$

Now recall (4.14) and apply (4.5) with $v_h = \ddot{u}_h$, and $w_h = u_h$, on the one hand, and, on the other, (4.6) with $v_h = \dot{u}_h$ to get

$$\left\| \frac{d^2}{dt^2}B_h(u_h, u_h) \right\|_0 \leq C \left(\|A_h u_h\|_0 \|A_h^{1/2}\ddot{u}_h\|_0 + \|A_h^{1/2}\dot{u}_h\|_0^{3/2} \|A_h \dot{u}_h\|_0^{1/2} \right),$$

so that, in view of (2.23)–(2.25) it follows that

$$(4.16) \quad t^3 \left\| \frac{d^2}{dt^2}B_h(u_h, u_h) \right\|_0^2 \leq C \left(1 + t^{1/2} \right) \tilde{M}_3^4.$$

Integrating with respect to t in (4.15) and taking into account (2.3), (2.25) with $r = 1$, and (4.16), the bound (4.10) follows.

To prove (4.12), we take derivatives twice with respect to t in (2.18) and then we set $\phi = t^4 A_h \ddot{u}_h$, so that

$$t^4 \left\| A_h^{1/2} \ddot{u}_h \right\|_0^2 + \frac{1}{2} t^4 \frac{d}{dt} \|A_h \dot{u}_h\|_0^2 = t^4 \left(f_{tt} - \frac{d^2}{dt^2} B_h(u_h, u_h), A_h \ddot{u}_h \right),$$

and, since $\|A_h^{1/2} \Pi_h f_{tt}\|_0 \leq C \|f_{tt}\|_1$,

$$(4.17) \quad t^4 \left\| A_h^{1/2} \ddot{u}_h \right\|_0^2 + \frac{d}{dt} (t^4 \|A_h \dot{u}_h\|_0^2) \leq 4t^3 \|A_h \dot{u}_h\|_0^2 + 2t^4 \left(C^2 \|f_{tt}\|_1^2 + \left\| A_h^{1/2} \frac{d^2}{dt^2} B_h(u_h, u_h) \right\|_0^2 \right).$$

Applying (4.5) to bound the third term on the right-hand side above we have

$$(4.18) \quad t^4 \left\| A_h^{1/2} \frac{d^2}{dt^2} B_h(u_h, u_h) \right\|_0^2 \leq C t^4 (\|A_h u_h\|_0^2 \|A_h \dot{u}_h\|_0^2 + \|A_h \dot{u}_h\|_0^4) \leq C t \tilde{K} (1 + \tilde{M}_4^2) (t^3 \|A_h \dot{u}_h\|_0^2 + t \|A_h \dot{u}_h\|_0^2),$$

the last inequality being a consequence of (2.23) and (2.24) with $r = 2$. Thus, integrating with respect to t in (4.17) and applying (2.3), (2.26) with $r = 2$, (4.18), and (4.10), the bound (4.12) follows.

Finally, since standard spectral theory of positive self-adjoint operators shows that $\|A_h^{-1/2} \ddot{u}_h\|_0^2 \leq \|A_h^{-3/2} \ddot{u}_h\|_0 \|A_h^{1/2} \ddot{u}_h\|_0$, by applying Hölder's inequality the bound (4.11) follows from (4.9) and (4.12). \square

5. Error estimates. In this section we obtain error estimates for the temporal errors e_n of the two BDF described in section 3.2, the backward Euler method and the two-step formula (3.21)–(3.22), for which, an equivalent formulation is

$$(5.1) \quad d_t U_h^{(n)} = -A_h U_h^{(n)} - B_h(U_h^{(n)}, U_h^{(n)}) + \Pi_h f(t_n), \quad l_0 \leq n \leq N.$$

We remark that although higher regularity was required in section 2.3, in what follows it is only required that Ω is of class C^2 and that (2.11)–(2.13) hold for $l = 2$.

A simple calculation shows that for a sequence $(y_n)_{n=0}^N$ in V_h ,

$$(5.2) \quad \sum_{j=l}^n (y_j, D y_j) = \frac{1}{2} \|y_n\|_0^2 - \frac{1}{2} \|y_{l-1}\|_0^2 + \frac{1}{2} \sum_{j=l}^n \|D y_j\|_0^2, \quad 1 \leq l \leq n \leq N,$$

and, for $2 \leq l \leq n \leq N$, (see, e.g., [10, (2.4b)])

$$(5.3) \quad \sum_{j=l}^n \left(y_j, \left(D + \frac{1}{2} D^2 \right) y_j \right) = \frac{1}{4} \|y_n\|_0^2 + \frac{1}{4} \|y_n + D y_n\|_0^2 + \frac{1}{4} \sum_{j=l}^n \|D^2 y_j\|_0^2 - \frac{1}{4} \|y_{l-1}\|_0^2 - \frac{1}{4} \|y_{l-1} + D y_{l-1}\|_0^2.$$

As mentioned in section 3.2, we shall assume that $e_0 = 0$ although in some of the previous lemmas this condition will not be required. It must be noticed that $e_0 = 0$ is not a serious restriction, since, on the one hand, it is usually satisfied in practice, and, on the other hand, were it not satisfied, there are standard ways to show that the effect of $e_0 \neq 0$ decays exponentially with time.

The finite-element approximation u_h to the velocity satisfies

$$(5.4) \quad d_t u_h(t_n) + A_h u_h(t_n) + B_h(u_h(t_n), u_h(t_n)) - \Pi_h f(t_n) = \tau_n,$$

where

$$(5.5) \quad \tau_n = d_t u_h(t_n) - \dot{u}_h(t_n) = \frac{1}{k} \int_{t_{n-1}}^{t_n} (t - t_{n-1}) \ddot{u}_h(t) dt, \quad n = 1, 2, \dots, N,$$

for the backward Euler method, and, for the two-step BDF,

$$(5.6) \quad \tau_n = \frac{1}{k} \int_{t_{n-2}}^{t_n} \left(2(t - t_{n-1})_+ - \frac{1}{2}(t - t_{n-2}) \right) \ddot{u}_h(t) dt,$$

where for $x \in \mathbb{R}$, $x_+ = \max(x, 0)$, and, also,

$$(5.7) \quad \tau_n = \frac{1}{2k} \int_{t_{n-2}}^{t_n} \left(2(t - t_{n-1})_+^2 - \frac{1}{2}(t - t_{n-2})^2 \right) \frac{d^3}{dt^3} u_h(t) dt.$$

Subtracting (5.1) from (5.4), we obtain that the temporal error e_n satisfies

$$(5.8) \quad d_t e_n + A_h e_n + B_h(e_n, u_h(t_n)) + B(U_h^{(n)}, e_n) = \tau_n, \quad n = 2, 3, \dots, N.$$

We shall now prove a result valid for both the backward Euler method and the two-step BDF.

LEMMA 5.1. *Fix $T > 0$ and $M > 0$. Then, there exist positive constants k_0 and C , such that for any $k \leq k_0$ with $Nk = T$, and any four sequences $(Y_n)_{n=0}^N$, $(V_n)_{n=0}^N$, $(W_n)_{n=0}^N$, and $(g_n)_{n=0}^N$ in V_h satisfying*

$$(5.9) \quad \max(\|A_h V_n\|, \|A_h W_n\|) \leq M, \quad n = 0, 1, \dots, N,$$

and

$$(5.10) \quad d_t Y_i + A_h Y_i + B_h(Y_i, V_i) + B_h(W_i, Y_i) = g_i, \quad i = l_0, \dots, N,$$

where l_0 is the value defined in (3.23), the following bound holds for $n = l_0, \dots, N$, and $j = -2, -1, 0, 1, 2$.

$$(5.11) \quad \begin{aligned} & \left\| A_h^{j/2} Y_n \right\|_0^2 + k \sum_{i=l_0}^n \left\| A_h^{(j+1)/2} Y_i \right\|_0^2 \\ & \leq C^2 \left(\left\| A_h^{j/2} Y_0 \right\|_0^2 + (l_0 - 1) \left\| A_h^{j/2} Y_1 \right\|_0^2 + k \sum_{i=l_0}^n \left\| A_h^{(j-1)/2} g_i \right\|_0^2 \right). \end{aligned}$$

When $j = 0$, condition (5.9) can be relaxed to $\|A_h V_n\| \leq M$, for $n = 0, 1, \dots, N$.

Proof. Take inner product with $A^j Y_i$ in (5.10) so that we have

$$(5.12) \quad \left(d_t A_h^{j/2} Y_i, A_h^{j/2} Y_i \right) + \left\| A_h^{(j+1)/2} Y_i \right\|_0^2 \leq \left| \left(Z_i, A_h^j Y_i \right) \right| + \left| \left(A_h^{(j-1)/2} g_i, A_h^{(j+1)/2} Y_i \right) \right|,$$

where

$$(5.13) \quad Z_i = B(Y_i, V_i) + B(W_i, Y_i).$$

Applying Hölder's inequality to the last term on the right-hand side of (5.12) and rearranging terms we have

$$(5.14) \quad \left(d_t A_h^{j/2} Y_i, A_h^{j/2} Y_i \right) + \frac{1}{2} \left\| A_h^{(j+1)/2} Y_i \right\|_0^2 \leq \left| \left(Z_i, A_h^j Y_i \right) \right| + \frac{1}{2} \left\| A_h^{(j-1)/2} g_i \right\|_0^2.$$

For $j > -2$, we write $(Z_i, A_h^j Y_i) = (A_h^{(j-1)/2} Z_i, A_h^{(j+1)/2} Y_i)$, so that applying Hölder's inequality and Lemma 4.2 with V_i and W_i taking the role of v_h and w_h in (4.5), and recalling (5.9) we have

$$(5.15) \quad \begin{aligned} \left| \left(Z_i, A_h^j Y_i \right) \right| &\leq \left\| A_h^{(j-1)/2} Z_i \right\|_0^2 + \frac{1}{4} \left\| A_h^{(j+1)/2} Y_i \right\|_0^2 \\ &\leq C^2 M^2 \left\| A_h^{j/2} Y_i \right\|_0^2 + \frac{1}{4} \left\| A_h^{(j+1)/2} Y_i \right\|_0^2. \end{aligned}$$

Notice that when $j = 0$, due to the skew-symmetry property (2.1) of the trilinear form b we have $|(Z_i, Y_i)| = |b(Y_i, V_i, Y_i)|$, so that only $\|A_h V_i\|_0 \leq M$ is necessary for (5.15) to hold; that is, no condition on $A_h W_i$ is required. For $j = 2$, on the other hand, we write $(Z_i, A_h^2 Y_i) = (A_h^{j/2} Z_i, A_h^{j/2} Y_i)$, so that arguing similarly we have

$$(5.16) \quad \begin{aligned} \left| \left(Z_i, A_h^{-2} Y_i \right) \right| &\leq \left\| A_h^{-1} Z_i \right\|_0 \left\| A_h^{-1} Y_i \right\|_0 \leq C M \left\| A_h^{-1/2} Y_i \right\|_0 \left\| A_h^{-1} Y_i \right\|_0 \\ &\leq C^2 M^2 \left\| A_h^{-1} Y_i \right\|_0^2 + \frac{1}{4} \left\| A_h^{-1/2} Y_i \right\|_0^2. \end{aligned}$$

In all cases, then, from (5.14) it follows that for an appropriate constant $C_0 > 0$

$$\left(d_t A_h^{j/2} Y_i, A_h^{j/2} Y_i \right) + \frac{1}{4} \left\| A_h^{(j+1)/2} Y_i \right\|_0^2 \leq C_0^2 M^2 \left\| A_h^{j/2} Y_i \right\|_0^2 + \frac{1}{2} \left\| A_h^{(j-1)/2} g_i \right\|_0^2,$$

so that multiplying this inequality by k and summing from l_0 to n , and recalling (5.2–5.3), after some rearrangements we can write

$$(5.17) \quad \begin{aligned} \frac{1}{2l_0} \left\| A_h^{j/2} Y_n \right\|_0^2 + \frac{k}{4} \sum_{i=l_0}^n \left\| A_h^{(j+1)/2} Y_i \right\|_0^2 \\ \leq \frac{1}{2l_0} \left(\left\| A_h^{j/2} Y_{l_0-1} \right\|_0^2 + (l_0 - 1) \left\| A_h^{j/2} (Y_{l_0-1} + DY_{l_0-1}) \right\|_0^2 \right) \\ + C_0^2 M^2 k \sum_{i=l_0}^n \left\| A_h^{j/2} Y_i \right\|_0^2 + \frac{k}{2} \sum_{i=l_0}^n \left\| A_h^{(j-1)/2} g_i \right\|_0^2. \end{aligned}$$

A simple calculation shows that

$$\left\| A_h^{j/2} Y_1 \right\|_0^2 + \left\| A_h^{j/2} (Y_1 + DY_1) \right\|_0^2 \leq 7 \left\| A_h^{j/2} Y_1 \right\|_0^2 + 3 \left\| A_h^{j/2} Y_0 \right\|_0^2,$$

so that multiplying by $2l_0$ in (5.17) and taking into account that $2l_0/4 \leq 1$, for an appropriate constant C' we may write

$$\begin{aligned} \left\| A_h^{j/2} Y_n \right\|_0^2 + k \sum_{i=l_0}^n \left\| A_h^{(j+1)/2} Y_i \right\|_0^2 &\leq 2l_0 C_0^2 M^2 k \sum_{i=l_0}^n \left\| A_h^{j/2} Y_i \right\|_0^2 \\ &+ C' \left(\left\| A_h^{j/2} Y_0 \right\|_0^2 + (l_0 - 1) \left\| A_h^{j/2} Y_1 \right\|_0^2 + k \sum_{i=l_0}^n \left\| A_h^{(j-1)/2} g_i \right\|_0^2 \right). \end{aligned}$$

Now, for k sufficiently small so that $2l_0 C_0^2 M^2 k < 1/2$, applying a standard discrete Gronwall lemma (e.g., [34, Lemma 5.1]) we have that (5.11) holds, with C^2 being $C' \exp(4l_0 C_0^2 M^2 T)$. \square

LEMMA 5.2. *Let (2.11) hold for $l = 2$. Then, there exist positive constants k_0 and c_1 such that the errors e_n satisfy the following bound for $1 \leq n \leq N$, $k \leq k_0$:*

$$(5.18) \quad E_n \equiv \|e_n\|_0^2 + k \sum_{i=l_0}^n \left\| A_h^{1/2} e_i \right\|_0^2 \leq c_1^2 (\|e_0\|_0^2 + (l_0 - 1)\|e_1\|_0^2 + k^2 I_{2,-1}(t_n)).$$

Proof. We apply Lemma 5.1 to (5.8) in the case where $j = 0$ and $Y_i = e_i$, $V_i = u_h(t_i)$, and $W_i = U_h^{(i)}$. Observe that since we are in the case $j = 0$, only one of the two sequences $(A_h u_h(t_i))_{i=0}^N$, $(A_h U_h^{(i)})_{i=0}^N$ has to be bounded, and, in the present case, the first one is bounded according to (2.23). Thus, we have

$$(5.19) \quad \|e_n\|_0^2 + k \sum_{i=l_0}^n \left\| A_h^{1/2} e_i \right\|_0^2 \leq C^2 \left(\|e_0\|_0^2 + (l_0 - 1)\|e_1\|_0^2 + k \sum_{i=1}^n \left\| A_h^{-1/2} \tau_j \right\|_0^2 \right).$$

Now, applying Hölder’s inequality to the right-hand side of (5.5) we have

$$\begin{aligned} k \left\| A_h^{-1/2} \tau_j \right\|_0^2 &\leq \frac{1}{k} \int_{t_{j-1}}^{t_j} (t - t_{j-1})^2 dt \int_{t_{j-1}}^{t_j} \left\| A_h^{-1/2} \ddot{u}_h(t) \right\|_0^2 dt \\ &= \frac{k^2}{3} \int_{t_{j-1}}^{t_j} \left\| A_h^{-1/2} \ddot{u}_h(t) \right\|_0^2 dt. \end{aligned}$$

Similarly, for the two-step BDF, applying Hölder’s inequality to the right-hand side of (5.6) we have

$$\begin{aligned} k \left\| A_h^{-1/2} \tau_n \right\|_0^2 &\leq \frac{1}{4k} \int_{t_{n-2}}^{t_n} (4(t - t_{n-1})_+ - (t - t_{n-2}))^2 dt \int_{t_{n-2}}^{t_n} \left\| A_h^{-1/2} \ddot{u}_h(t) \right\|_0^2 dt \\ &\leq \frac{5}{3} k^2 \int_{t_{n-2}}^{t_n} \left\| A_h^{-1/2} \ddot{u}_h(t) \right\|_0^2 dt. \end{aligned}$$

Thus, the statement of the lemma follows from (5.19). \square

The previous result allows us to deduce a bound for $\|A_h e_n\|_0$ in the following lemma. The values of $I_{2,-1}$, $I_{2,0}$ are those of Proposition 2.1.

LEMMA 5.3. *Under the conditions of Lemma 5.2, there exist positive constants \tilde{k}_0 and \tilde{c}_0 such that if $e_0 = 0$ and, in the case of the two-step BDF, also $U_h^{(1)}$ is given by the backward Euler method, the following bound holds for $k \leq \tilde{k}_0$ and $n = 1, 2, \dots, N$:*

$$(5.20) \quad \|A_h e_n\|_0 \leq \tilde{c}_0 J_n,$$

where $J_n = (I_{2,-1}(t_n) + I_{2,0}(t_n))^{1/2}$.

Proof. If $e_0 = 0$, from (5.18) for the Euler method (for which $l_0 = 1$) it follows that

$$(5.21) \quad \|e_n\|_0 \leq c'k(I_{2,-1}(t_n))^{1/2} \leq c'kJ_n, \quad n = 1, 2, \dots, N$$

for $c' = c_1$. In the case of the two-step BDF, if $U_h^{(1)}$ is obtained by the backward Euler method, if we allow for a larger value of c' , it is clear that (5.21) also holds.

Furthermore, it is immediate to check that $\|d_t e_n\|_0 \leq 2l_0 k^{-1} \max_{0 \leq i \leq l_0} \|e_n\|_0$, so that, in view of (5.21), from (5.8) it follows that

$$\|A_h e_n\|_0 \leq c''J_n + \left\| B_h(e_n, u_h(t_n)) + B(U_h^{(n)}, e_n) \right\|_0 + \|\tau_n\|_0,$$

for some constant $c'' > 0$. For the Euler method, recalling the expression of τ_n in (5.5) we can write

$$\|\tau_n\|_0^2 = \left\| \frac{1}{k} \int_{t_{n-1}}^{t_n} (t - t_{n-1}) \ddot{u}_h \, dt \right\|_0^2 \leq \frac{1}{k^2} \int_{t_{n-1}}^{t_n} \frac{(t - t_{n-1})^2}{t} \, dt \int_{t_{n-1}}^{t_n} t \|\ddot{u}_h\|_0^2 \, dt.$$

A simple calculation shows that the first factor on the right-hand side above can be bounded by $k/t_n \leq 1$ for $n = 1, 2, \dots, N$. Furthermore, a similar bound can be also obtained in the case of the two-step BDF. Thus, we have

$$\|A_h e_n\|_0 \leq c'''J_n + \left\| B_h(e_n, u_h(t_n)) + B(U_h^{(n)}, e_n) \right\|_0,$$

for an appropriate constant $c''' > 0$. Finally,

$$\left\| B_h(e_n, u_h(t_n)) + B(U_h^{(n)}, e_n) \right\|_0 = \|B_h(e_n, u_h(t_n)) + B(u_h(t_n), e_n) - B_h(e_n, e_n)\|_0.$$

Applying (4.5) and (4.6) we get

$$\|A_h e_n\|_0 \leq c'''J_n + C \left\| A_h^{1/2} e_n \right\|_0 \|A_h u_h\|_0 + C \left\| A_h^{1/2} e_n \right\|_0^{3/2} \|A_h e_n\|_0^{1/2},$$

and, thus,

$$\frac{1}{2} \|A_h e_n\|_0 \leq c'''J_n + C \left\| A_h^{1/2} e_n \right\|_0 \|A_h u_h\|_0 + \frac{1}{2} C^2 \left\| A_h^{1/2} e_n \right\|_0^3.$$

Since $\|A_h u_h\|_0$ is bounded (recall Proposition 2.1) and, arguing as in (5.21), we have $\|A_h^{1/2} e_n\|_0 \leq c_1(kI_{2,-1}(t_n))^{1/2}$, the bound (5.20) follows for k sufficiently small. \square

Remark 5.1. Observe that from the previous lemma and Proposition 2.1 it follows that $\|A_h U_h^{(n)}\|_0 \leq c\tilde{M}_3$ where $c = 1 + \tilde{c}_0\sqrt{2}$. Thus, as long as $e_0 = 0$ and, in case of the two-step BDF, also $U_h^{(1)}$ is given by the Euler method, we may apply Lemma 5.1 for $j \neq 0$, with V_n and W_n replaced by $u_h(t_n)$ and $U_h^{(n)}$, respectively.

We now study the errors $A_h t_n e_n$. We deal first with the backward Euler method. Observe that $D(t_n e_n) = t_n D e_n + k e_{n-1}$, so that multiplying by t_n in (5.8), after some rearrangements we get

$$(5.22) \quad d_t(t_n e_n) + A_h(t_n e_n) + B_h(t_n e_n, u_h) + B_h(U_h^{(n)}, t_n e_n) = e_{n-1} + t_n \tau_n.$$

We have the following result.

THEOREM 5.4. *Let (2.11) hold for $l = 2$. Then, there exist positive constants k_2 and c_2 such that for $k \leq k_2$, if $e_0 = 0$ the errors $\epsilon_n = t_n e_n$ satisfy*

$$(5.23) \quad \left(\|A_h \epsilon_n\|_0^2 + k \sum_{i=1}^n \left\| A_h^{3/2} \epsilon_i \right\|_0^2 \right)^{1/2} \leq c_2 k (I_{2,-1}(t_n) + I_{2,1}(t_n))^{1/2}, \quad 1 \leq n \leq N.$$

Proof. Applying Lemma 5.1 with $j = 2$ to (5.22) we have

$$\|A_h \epsilon_n\|_0^2 + k \sum_{i=1}^n \left\| A_h^{3/2} \epsilon_i \right\|_0^2 \leq c \left(k \sum_{i=1}^{n-1} \left\| A_h^{1/2} e_i \right\|_0^2 + k \sum_{i=1}^n \left\| t_i A_h^{1/2} \tau_i \right\|_0^2 \right).$$

The first term on the right-hand side above is bounded Lemma 5.2 by $c_1^2 k^2 I_{2,-1}(t_n)$. For the second one we notice that

$$k \left\| t_i A_h^{1/2} \tau_i \right\|_0^2 = k \left\| \frac{t_i}{k} \int_{t_{i-1}}^{t_i} \frac{t - t_{i-1}}{t} t A_h^{1/2} \ddot{u}_h(t) dt \right\|_0^2,$$

so that, for $i \geq 2$ since $t_i/t \leq t_i/t_{i-1} \leq 2$, if $t \in (t_{i-1}, t_i)$, we may bound $k \|t_i A_h^{1/2} \tau_i\|_0^2$ by

$$\frac{4}{k} \int_{t_{i-1}}^{t_i} (t - t_{i-1})^2 dt \int_{t_{i-1}}^{t_i} t^2 \left\| A_h^{1/2} \ddot{u}_h(t) \right\|_0^2 dt \leq \frac{4k^2}{3} \int_{t_{i-1}}^{t_i} t^2 \left\| A_h^{1/2} \ddot{u}_h(t) \right\|_0^2 dt,$$

and, for $i = 1$, since $t_1/k = 1$, and $(t - t_0)/t = 1$, we may bound $k \|t_i A_h^{1/2} \tau_i\|_0^2$ by

$$k \int_{t_{i-1}}^{t_i} dt \int_{t_{i-1}}^{t_i} t^2 \left\| A_h^{1/2} \ddot{u}_h(t) \right\|_0^2 dt \leq k^2 \int_{t_{i-1}}^{t_i} t^2 \left\| A_h^{1/2} \ddot{u}_h(t) \right\|_0^2 dt.$$

Thus, (5.23) follows. \square

LEMMA 5.5. *Under the conditions of Lemma 5.2, there exist positive constants k'_0 and c'_1 such that for $k \leq k'_0$, if $e_0 = 0$, in the backward Euler method e_1 satisfies*

$$(5.24) \quad \left\| A_h^{-1} e_1 \right\|_0^2 + k \left\| A_h^{-1/2} e_1 \right\|_0^2 \leq c'_1 k^4 G_1,$$

where $G_1 = \max_{0 \leq s \leq k} F_{2,-2}(s)$.

Proof. We take inner product with $2k A_h^{-2} e_1$ in (5.8) for $n = 1$, and recalling (5.2) and taking into account that $e_0 = 0$, after some rearrangements we have

$$(5.25) \quad \left\| A_h^{-1} e_1 \right\|_0^2 + 2k \left\| A_h^{-1/2} e_1 \right\|_0^2 \leq 2k |(Z_1, A_h^{-2} e_1)| + 2k |(A_h^{-1} \tau_1, A_h^{-1} e_1)|,$$

where Z_1 is as in (5.13) but with Y_1 , V_1 , and W_1 replaced by e_1 , $u_h(t_1)$, and $U_h^{(1)}$, respectively. Thus, arguing as in (5.16)

$$\left(1 - 2k C^2 \tilde{M}_3^2 \right) \left\| A_h^{-1} e_1 \right\|_0^2 + \frac{3}{2} k \left\| A_h^{-1/2} e_1 \right\|_0^2 \leq 2k |(A_h^{-1} \tau_1, A_h^{-1} e_1)|,$$

so that taking into account that

$$2k |(A_h^{-1} e_1, A_h^{-1} \tau_1)| \leq \left\| A_h^{-1} e_1 \right\|_0^2 / 2 + k^2 \left\| A_h^{-1} \tau_1 \right\|_0^2,$$

from (5.25) it follows that

$$(5.26) \quad \left(\frac{1}{2} - 2kC^2\tilde{M}_3^2\right) \|A_h^{-1}e_1\|^2 + \frac{3}{2}k \|A_h^{-1/2}e_1\|_0^2 \leq k^2 \|A_h^{-1}\tau_1\|_0^2.$$

Recalling the expression of τ_n in (5.5) we can write

$$\|A_h^{-1}\tau_1\|_0^2 \leq \max_{0 \leq t \leq t_1} \|A_h^{-1}\ddot{u}_h(t)\|_0^2 \left(\frac{1}{k} \int_0^k t \, dt\right)^2 = \frac{k^2}{4} \max_{0 \leq t \leq t_1} \|A_h^{-1}\ddot{u}_h(t)\|_0^2,$$

we then have that (5.24) follows from (5.26) provided that k is sufficiently small. \square

LEMMA 5.6. *Under the conditions of Lemma 5.2, let $e_0 = 0$ and let $U_h^{(1)}$ be given by the backward Euler method. Then, there exist positive constants k_0 and c_1 such that the errors e_n of the two-step BDF satisfy*

$$(5.27) \quad E'_n \equiv \|A_h^{-1}e_n\|_0^2 + k \sum_{j=2}^n \|A_h^{-1/2}e_j\|_0^2 \leq c_1k^4(G_1 + I_{3,-3}(t_n)), \quad 2 \leq n \leq N,$$

where G_1 is given after (5.24).

Proof. In view of the comments in Remark 5.1, we can apply Lemma 5.1 with $j = -2$ to (5.8), so that, recalling that $e_0 = 0$, we have

$$(5.28) \quad \|A_h^{-1}e_n\|^2 + k \sum_{i=2}^n \|A_h^{-1/2}e_i\|_0^2 \leq c \left(\|A_h^{-1}e_1\|_0^2 + k \sum_{j=2}^n \|A_h^{-3/2}\tau_j\|_0^2 \right).$$

Notice that, as we showed in Lemma 5.5, the first term on the right-hand side above is bounded by $c_1k^4G_1$. For the second term on the right-hand side of (5.28), in view of (5.7) a simple calculation shows that

$$k \|A_h^{-3/2}\tau_j\|_0^2 \leq Ck^4 \int_{t_{n-2}}^{t_n} \left\| A_h^{-3/2} \frac{d^3 u_h(s)}{ds^3} \right\|_0^2 ds.$$

Thus, (5.27) follows. \square

For any two sequences $(y_n)_{n=0}^\infty$ and $(z_n)_{n=0}^\infty$, it is easy to check that $D(y_n z_n) = y_n D z_n + z_{n-1} D y_n$, for $n = 1, 2, \dots$, and, also,

$$D^2(y_n z_n) = y_n D^2 z_n + 2 D y_n D z_{n-1} + z_{n-2} D^2 y_n.$$

Thus, for the two-step BDF, multiplying (5.8) by t_n and t_n^2 and rearranging terms, for $j = 2, 3, \dots, N$, we have

$$(5.29) \quad \begin{aligned} d_t(t_n e_n) + A_h t_n e_n + B_h(u_h(t_n), t_n e_n) \\ - t_n B(U_h^{(n)}, t_n e_n) = t_n \tau_n + (e_{n-1} + D e_{n-1}), \end{aligned}$$

and

$$(5.30) \quad d_t(t_n^2 e_n) + A_h t_n^2 e_n + B_h(t_n^2 e_n, u_h(t_n)) - B(U_h^{(n)}, t_n^2 e_n) = t_n^2 \tau_j + \sigma_{n-1},$$

where

$$(5.31) \quad \sigma_{n-1} = (t_n + t_{n-1})(e_{n-1} + D e_{n-1}) + k e_{n-2}.$$

THEOREM 5.7. *Under the conditions of Theorem 5.4, there exist positive constants k_1 and c_2 such that for $k \leq k_1$ if $e_0 = 0$ and $U_h^{(1)}$ be given by the backward Euler method, the errors $\epsilon_n = t_n e_n$ and $\epsilon'_n = t_n^2 e_n$ of the two-step BDF satisfy the following bounds for $2 \leq n \leq N$:*

$$(5.32) \quad \mathcal{E}_n \equiv \left(\|\epsilon_n\|_0^2 + k \sum_{i=2}^n \|A_h^{1/2} \epsilon_i\|_0^2 \right)^{\frac{1}{2}} \leq c_2 k^2 (H_1 + I(t_n))^{\frac{1}{2}},$$

$$(5.33) \quad \mathcal{E}'_n \equiv \left(\|A_h \epsilon'_n\|_0^2 + k \sum_{i=2}^n \|A_h^{3/2} \epsilon'_i\|_0^2 \right)^{\frac{1}{2}} \leq c_2 k^2 (H_1 + I(t_n) + J_1^2 + I_{3,1}(t_n))^{1/2},$$

where J_1 is given after (5.20), $H_1 = I_{2,-1}(t_1) + G_1$, where G_1 is given after (5.24), and $I(t) = I_{3,-1}(t) + I_{3,-3}(t)$.

Proof. To prove (5.32) we apply Lemma 5.1 with $j = 0$ to (5.29), so that taking into account that $e_0 = 0$ we have

$$(5.34) \quad \|\epsilon_n\|_0^2 + k \sum_{j=2}^n \|A_h^{1/2} \epsilon_j\|_0^2 \leq c \left(\|\epsilon_1\|_0^2 + k \sum_{i=1}^{n-1} \|A_h^{-1/2} (e_i + De_i)\|_0^2 + k \sum_{i=2}^n \|t_i A_h^{-1/2} \tau_i\|_0^2 \right).$$

Notice that since $e_i + De_i = 2e_i - e_{i-1}$, the second term on the right-hand side above can be bounded by $7k \sum_{i=1}^{n-1} \|A_h^{-1/2} e_i\|_0^2$, a quantity that has already been bounded in Lemma 5.6. Also, by writing

$$t_i \tau_i = \frac{t_i}{2k} \int_{t_{i-2}}^{t_i} \frac{1}{t} \left(2(t - t_{i-1})_+^2 - \frac{1}{2}(t - t_{i-2})^2 \right) t \frac{d^3}{dt^3} u_h(t) dt,$$

and noticing that $t_i/t_{i-2} \leq 3$ for $i \geq 3$, and $t_2/k \leq 2$, a straightforward calculation shows that

$$(5.35) \quad k \|A_h^{-1/2} t_i \tau_j\|_0^2 \leq C k^4 \int_{t_{i-2}}^{t_i} t^2 \|A_h^{-1/2} \ddot{u}_h(t)\|_0^2 dt, \quad j = 2, \dots, N.$$

Finally, noticing that $\|\epsilon_1\|_0^2 = k^2 \|e_1\|_0^2$ and recalling Lemma 5.2, we have that (5.32) follows from (5.34) and (5.35).

To prove (5.33) we apply Lemma 5.1 with $j = 2$ to (5.30) to get

$$(5.36) \quad \|A_h \epsilon'_n\|_0^2 + k \sum_{j=2}^n \|A_h^{3/2} \epsilon'_j\|_0^2 \leq c \left(\|A_h \epsilon'_1\|_0^2 + k \sum_{i=1}^{n-1} \|A_h^{1/2} \sigma_i\|_0^2 + k \sum_{i=2}^n \|t_i^2 A_h^{1/2} \tau_i\|_0^2 \right).$$

For the first term on the right-hand side above, in view of Lemma 5.3 we can write

$$(5.37) \quad \|A_h \epsilon'_1\|_0 = k^2 \|A_h e_1\| \leq k^2 \tilde{c}_0 J_1.$$

For the second term on the right-hand side of (5.36), we first recall the expression of σ_i in (5.31) and then we notice that for $i \geq 1$ we have that $k \leq t_i$, $t_{i+2}/t_i \leq 3$ and $t_{i+1}/t_i \leq 2$, so that, recalling that $e_0 = 0$, for an appropriate constant $C > 0$ we may write

$$(5.38) \quad k \sum_{i=1}^{n-1} \|A_h^{1/2} \sigma_i\|_0^2 \leq C \left(k \|A_h^{1/2} \epsilon_1\|_0^2 + k \sum_{i=2}^{n-1} \|A_h^{1/2} \epsilon_i\|_0^2 \right) \leq C \left(k^3 \|A_h^{1/2} e_1\|_0^2 + \mathcal{E}_n^2 \right),$$

\mathcal{E}_n being the quantity in (5.32). Also, by writing

$$t_i^2 \tau_i = \frac{t_i^2}{2k} \int_{t_{i-2}}^{t_i} \frac{1}{t^2} \left(2(t - t_{i-1})_+^2 - \frac{1}{2}(t - t_{i-2})^2 \right) t^2 \frac{d^3}{dt^3} u_h(t) dt,$$

and noticing that $t_i^2/t_{i-2}^2 \leq 9$ for $i \geq 3$, and $t_2^2/k \leq 4k$, we get

$$(5.39) \quad k \left\| A_h^{1/2} t_i^2 \tau_i \right\|_0^2 \leq C k^4 \int_{t_{i-2}}^{t_i} t^4 \left\| A_h^{1/2} \ddot{u}_h(t) \right\|_0^2 dt, \quad j = 2, \dots, N.$$

Thus, (5.33) follows from (5.36), (5.37), (5.38), (5.32), (5.18), and (5.39). \square

Although not strictly necessary for the analysis of the postprocessed approximation, for the sake of completeness we include an error bound for the pressure. We first notice that as a consequence of the LBB condition (2.7) we have that the error $\pi_n = p_h(t_n) - P_h^{(n)}$ satisfies

$$\|\pi_n\|_{L^2(\Omega)/\mathbb{R}} \leq \frac{1}{\beta} \sup_{\phi_h \in X_{h,r}} \frac{|(\pi_n, \nabla \cdot \phi_h)|}{\|\phi_h\|_1}.$$

Furthermore subtracting (3.21) from (2.18), we have

$$(\pi_n, \nabla \cdot \phi_h) = \left(\dot{u}_h - d_t U_h^{(n)}, \phi_h \right) + (\nabla e_n, \nabla \phi_h) + b(e_n, u_h, \phi_h) + b\left(U_h^{(n)}, e_n, \phi_h \right)$$

for all $\phi_h \in X_{h,r}$. Using standard bounds for the trilinear form b (e.g., [34, (3.7)]) we can write

$$(5.40) \quad \|\pi_n\|_{L^2(\Omega)/\mathbb{R}} \leq C \left(\|e_n\|_1 + \left\| \dot{u}_h - d_t U_h^{(n)} \right\|_{-1} \right).$$

Recalling the expression of $d_t^* U_h^{(n)}$ in (3.4), we see that $\dot{u}_h - d_t U_h^{(n)} = \dot{u}_h - d_t^* U_h^{(n)}$, so that applying Lemma 3.2 and taking into account the equivalence (2.15) between $\|e_n\|_1$ and $\|A_h^{1/2} e_n\|_0$, we have $\|\pi_n\|_{L^2(\Omega)/\mathbb{R}} \leq C \|A_h^{1/2} e_n\|_0$. Since using standard spectral theory of positive self-adjoint operators it is straightforward to show that $\|A_h^{1/2} e_n\|_0 \leq C \|e_n\|_0^{1/2} \|A_h e_n\|_0^{1/2}$, applying Lemma 5.2 and Theorem 5.4 in the case of the backward Euler method, and Theorem 5.7 in the case of the two-step BDF, we conclude the following result.

THEOREM 5.8. *Under the conditions of Theorem 5.4, there exist positive constants k_3 and c_4 such that if $e_0 = 0$ and $U_h^{(1)}$ is obtained by the backward Euler method, the following bound holds for $k < k_3$ and for $n = l_0, \dots, N$: For the backward Euler method,*

$$\left\| p_h(t_n) - P_h^{(n)} \right\|_{L^2(\Omega)/\mathbb{R}} \leq c_4 C_1^{1/2} \frac{k}{t_n^{1/2}},$$

where $C_1 = (I_{2,-1}(I_{2,-1} + I_{2,1}))^{1/2}$, and, for the two-step BDF,

$$\left\| p_h(t_n) - P_h^{(n)} \right\|_{L^2(\Omega)/\mathbb{R}} \leq c_4 C_2^{1/2} \frac{k^2}{t_n^{3/2}},$$

where C_2 is the product of the quantities between parentheses on the right-hand sides of (5.32) and (5.33).

REFERENCES

- [1] B. AYUSO AND B. GARCÍA-ARCHILLA, *Regularity constants of the Stokes problem. Application to finite-element methods on curved domains*, Math. Models Methods Appl. Sci., 15 (2005), pp. 437–470.
- [2] B. AYUSO, J. DE FRUTOS, AND J. NOVO, *Improving the accuracy of the mini-element approximation to Navier–Stokes equations*, IMA J. Numer. Anal., 27 (2007), pp. 198–218.
- [3] B. AYUSO, B. GARCÍA-ARCHILLA, AND J. NOVO, *The postprocessed mixed finite-element method for the Navier–Stokes equations*, SIAM J. Numer. Anal., 43 (2005), pp. 1091–1111.
- [4] J. BECKER, *A second order backward difference method with variable steps for a parabolic problem*, BIT, 38 (1998), pp. 644–662.
- [5] F. BREZZI AND R. S. FALK, *Stability of higher-order Hood–Taylor methods*, SIAM J. Numer. Anal., 28 (1991), pp. 581–590.
- [6] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer, New York, 1991.
- [7] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
- [8] M. CROUZEIX AND P.-A. RAVIART, *Conforming and nonconforming finite element methods for solving the stationary Stokes equations*, RAIRO, 7 (1973), pp. 33–75.
- [9] C. F. CURTISS AND J. O. HIRSHFELDER, *Integration of stiff equations*, Proc. Natl. Acad. Sci., 38 (1952), pp. 235–243.
- [10] E. EMMRICH, *Error of the two-step BDF for the incompressible Navier–Stokes problem*, M2AN Math. Model. Numer. Anal., 38 (2004), pp. 757–764.
- [11] J. DE FRUTOS, B. GARCÍA-ARCHILLA, AND J. NOVO, *A postprocessed Galerkin method with Chebyshev or Legendre polynomials*, Numer. Math., 86 (2000), pp. 419–442.
- [12] J. DE FRUTOS AND J. NOVO, *A spectral element method for the Navier–Stokes equations with improved accuracy*, SIAM J. Numer. Anal., 38 (2000), pp. 799–819.
- [13] J. DE FRUTOS AND J. NOVO, *A postprocess based improvement of the spectral element method*, Appl. Numer. Math., 33 (2000), pp. 217–223.
- [14] J. DE FRUTOS AND J. NOVO, *Postprocessing the linear-finite-element method*, SIAM J. Numer. Anal., 40 (2002), pp. 805–819.
- [15] J. DE FRUTOS AND J. NOVO, *A posteriori error estimation with the p version of the finite element method for nonlinear parabolic differential equations*, Comput. Meth. Appl. Mech. Engrg., 191 (2002), pp. 4893–4904.
- [16] J. DE FRUTOS AND J. NOVO, *Element-wise a posteriori estimates based on hierarchical bases for nonlinear parabolic problems*, Int. J. Numer. Meth. Engrg., 63 (2005), pp. 1146–1173.
- [17] J. DE FRUTOS, B. GARCÍA-ARCHILLA, AND J. NOVO, *A posteriori error estimates for fully discrete nonlinear parabolic problems*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 3462–3474.
- [18] J. DE FRUTOS, B. GARCÍA-ARCHILLA, AND J. NOVO, *The postprocessed mixed finite-element method for the Navier–Stokes equations: Improved error bounds*, SIAM J. Numer. Anal., 46 (2007–2008), pp. 201–230.
- [19] G. P. GALDI, *An Introduction to the Mathematical Theory of the Navier–Stokes Equations. Vol. I. Linearized Steady Problems*, Springer-Verlag, New York, 1994.
- [20] B. GARCÍA-ARCHILLA, J. NOVO, AND E. S. TITI, *Postprocessing the Galerkin method: A novel approach to approximate inertial manifolds*, SIAM J. Numer. Anal., 35 (1998), pp. 941–972.
- [21] B. GARCÍA-ARCHILLA, J. NOVO, AND E. S. TITI, *An approximate inertial manifold approach to postprocessing Galerkin methods for the Navier–Stokes equations*, Math. Comput., 68 (1999), pp. 893–911.
- [22] B. GARCÍA-ARCHILLA AND E. S. TITI, *Postprocessing the Galerkin method: The finite-element case*, SIAM J. Numer. Anal., 37 (2000), pp. 470–499.
- [23] V. GIRAULT AND J. L. LIONS, *Two-grid finite-element schemes for the transient Navier–Stokes problem*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 945–980.
- [24] V. GIRAULT AND P. A. RAVIART, *Finite Element Methods for Navier–Stokes Equations*, Springer-Verlag, Berlin, 1986.
- [25] E. HAIRER, S. P. NØRSETT, AND G. WANNER, *Solving Ordinary Differential Equations I*, 2nd ed., Springer-Verlag, Berlin, 1993.
- [26] E. HAIRER AND G. WANNER, *Solving Ordinary Differential Equations II*, 2nd. ed., Springer-Verlag, Berlin, 1997.
- [27] Y. HE, *Two-level method based on finite element and Crank–Nicolson extrapolation for the time-dependent Navier–Stokes equations*, SIAM J. Numer. Anal., 41 (2003), pp. 1263–1285.
- [28] Y. HE AND K. M. LIU, *A multilevel finite element method for the Navier–Stokes problem*, Numer. Methods Partial Differential Equations, 21 (2005), pp. 1052–1078.

- [29] Y. HE AND K. M. LIU, *A multilevel spectral Galerkin method for the Navier-Stokes equations, II: Time discretization*, Adv. Comput. Math., 25 (2006), pp. 403–433.
- [30] Y. HE, K. M. LIU, AND W. SUN, *Multi-level spectral Galerkin method for the Navier-Stokes problem I: Spatial discretization*, Numer. Math., 101 (2005), pp. 501–522.
- [31] Y. HE AND W. SUN, *Stability and convergence of the Crank-Nicolson/Adams-Bashforth scheme for the time-dependent Navier-Stokes equations*, SIAM J. Numer. Anal., 45 (2007), pp. 837–869.
- [32] J. G. HEYWOOD AND R. RANNACHER, *Finite element approximation of the nonstationary Navier-Stokes problem. Part I. Regularity of solutions and second-order error estimates for spatial discretization*, SIAM J. Numer. Anal., 19 (1982), pp. 275–311.
- [33] J. G. HEYWOOD AND R. RANNACHER, *Finite element approximation of the nonstationary Navier-Stokes problem. Part III: Smoothing property and higher order error estimates for spatial discretization*, SIAM J. Numer. Anal., 25 (1988), pp. 489–512.
- [34] J. G. HEYWOOD AND R. RANNACHER, *Finite-element approximation of the nonstationary Navier-Stokes problem. Part IV: Error analysis for second-order time discretization*, SIAM J. Numer. Anal., 27 (1990), pp. 353–384.
- [35] P. HOOD AND C. TAYLOR, *A numerical solution of the Navier-Stokes equations using the finite element technique*, Comput. Fluids, 1 (1973), pp. 73–100.
- [36] Y. HOU AND K. LIU, *A small eddy correction method for nonlinear dissipative evolutionary equations*, SIAM J. Numer. Anal., 41 (2003), pp. 1101–1130.
- [37] C. R. LAING, A. MCROBIE, AND J. M. T. THOMPSON, *The post-processed Galerkin method applied to non-linear shell vibrations*, Dynam. Stability Systems, 14 (1999), pp. 163–181.
- [38] G. J. LORD AND T. SHARDLOW, *Postprocessing for stochastic parabolic partial differential equations*, SIAM J. Numer. Anal., 45 (2007), pp. 870–889.
- [39] L. G. MARGOLIN, E. S. TITI, AND S. WYNNE, *The postprocessing Galerkin and nonlinear Galerkin methods—A truncation analysis point of view*, SIAM J. Numer. Anal., 41 (2003), pp. 695–714.
- [40] M. A. OLSHANSKII, *Two-level method and some a priori estimates in unsteady Navier-Stokes calculations*, J. Comp. Appl. Math., 104 (1999), pp. 173–191.
- [41] J. RODENKIRCHEN, *Maximum L^2 -Convergence Rates of the Crank-Nicolson Scheme to the Stokes Initial Value Problem*, SIAM J. Numer. Anal., 45 (2007), pp. 484–499.
- [42] S. C. SINHA, S. REDKAR, V. DESHMUKH, AND E. A. BUTCHER, *Order reduction of parametrically excited nonlinear systems: Techniques and applications*, Nonlinear Dynamics, 41 (2005), pp. 237–273.
- [43] C. SANSOUR, P. WRIGGERS, AND J. SANSOUR, *A finite element post-processed Galerkin method for dimensional reduction in the non-linear dynamics of solids. Applications to shells*, Comput. Mech., 32 (2003), pp. 104–114.
- [44] Y. YAN, *Postprocessing the finite element-method for semilinear parabolic problems*, SIAM J. Numer. Anal., 44 (2006), pp. 1681–1702.

A BOUNDED ARTIFICIAL VISCOSITY LARGE EDDY SIMULATION MODEL*

JEFF BORGGAARD[†], TRAIAN ILIESCU[‡], AND JOHN PAUL ROOP[§]

Abstract. In this paper, we present a rigorous numerical analysis for a bounded artificial viscosity model ($\tau = \mu\delta^\sigma a(\delta\|\nabla^s \mathbf{u}\|_F)\nabla^s \mathbf{u}$) for the numerical simulation of turbulent flows. In practice, the commonly used Smagorinsky model ($\tau = (c_s\delta)^2\|\nabla^s \mathbf{u}\|_F \nabla^s \mathbf{u}$) is overly dissipative and yields unphysical results. To date, several methods for “clipping” the Smagorinsky viscosity have proven useful in improving the physical characteristics of the simulated flow. However, such heuristic strategies strongly rely upon a priori knowledge of the flow regime. The bounded artificial viscosity model relies on a highly nonlinear, but monotone and smooth, semilinear elliptic form for the artificial viscosity. For this model, we have introduced a variational computational strategy, provided finite element error convergence estimates, and included several computational examples indicating its improvement on the overly diffusive Smagorinsky model.

Key words. large eddy simulation, turbulence, artificial viscosity, Smagorinsky model

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

DOI. 10.1137/060656164

1. Introduction. Turbulence is central to many important applications. Direct numerical simulation is not feasible for the foreseeable future in many of these applications. Indeed, Kolmogorov’s 1941 theory (K-41) of homogeneous, isotropic turbulence predicts that small scales exist down to $O(Re^{-3/4})$, where $Re > 0$ is the Reynolds number. Thus, in order to capture all scales on a mesh, we need a mesh-size $h \sim Re^{-3/4}$ and consequently (in three-dimensional (3D) case) $N \sim Re^{9/4}$ mesh points.

Large eddy simulation (LES) is one of the most successful approaches in the numerical simulation of turbulent flows. LES seeks to calculate the large, energetic structures (the large eddies) in a turbulent flow. The large structures are defined by convolving the flow variables with a rapidly decaying spatial filter g_δ . To derive equations for $\bar{\mathbf{u}}$, the large eddy flow structure, we convolve the Navier–Stokes equations (NSE) with $g_\delta(\mathbf{x})$. The resulting system is not closed, since it involves both \mathbf{u} and $\bar{\mathbf{u}}$. The tensor $\boldsymbol{\tau}(\mathbf{u}, \mathbf{u}) = \overline{\mathbf{u}\mathbf{u}^T} - \bar{\mathbf{u}}\bar{\mathbf{u}}^T$ is often called the subgrid-scale stress (SGS) tensor. Thus, the closure problem in LES is to model the SGS tensor $\boldsymbol{\tau}(\mathbf{u}, \mathbf{u})$.

The simplest and most commonly used approach to the closure problem is the eddy viscosity (EV) model. EV models are motivated by the idea that the global effect of the SGS stress tensor $\boldsymbol{\tau}(\mathbf{u}, \mathbf{u})$, in the mean, is to transfer energy from resolved to

*Received by the editors April 4, 2006; accepted for publication (in revised form) September 10, 2008; published electronically January 16, 2009.

<http://www.siam.org/journals/sinum/47-1/65616.html>

[†]Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061 (jborggaard@vt.edu). This author’s research was partially supported by AFOSR grants F49620-00-1-0299, F49620-03-1-0243, and FA9550-05-1-0449 and NSF grants DMS-0322852 and DMS-0513542.

[‡]Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061 (iliescu@vt.edu). This author’s research was partially supported by AFOSR grants F49620-03-1-0243 and FA9550-05-1-0449 and NSF grants DMS-0209309, DMS-0322852, and DMS-0513542.

[§]Department of Mathematics, North Carolina A & T University, Greensboro, NC 27411 (jproop@ncat.edu). This author’s research was partially supported by AFOSR grant F49620-00-1-0299.

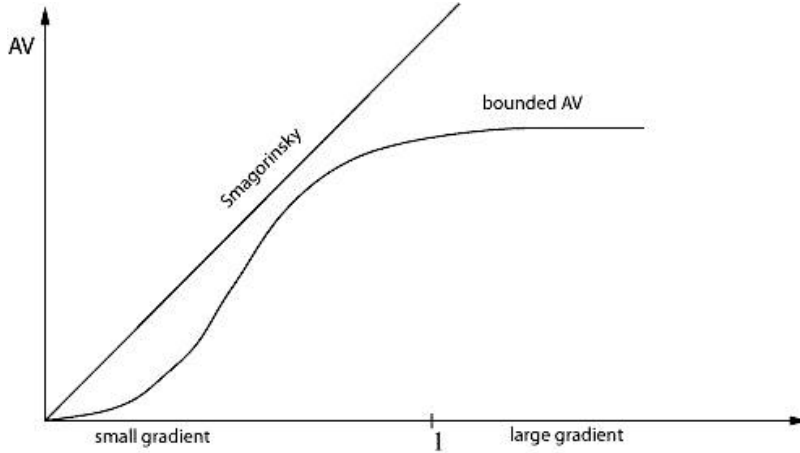


FIG. 1.1. The amount of artificial viscosity introduced by the Smagorinsky and the bounded AV models against $\|\delta \nabla^s \bar{\mathbf{u}}\|_F$.

unresolved scales through inertial interactions [4, 45]:

$$\nabla \cdot \boldsymbol{\tau}(\mathbf{u}, \mathbf{u}) \approx -\nabla \cdot (\nu_T \nabla^s \bar{\mathbf{u}}) + \text{terms incorporated into } \bar{p},$$

where $\nabla^s \bar{\mathbf{u}}$ is the deformation tensor and $\nu_T \geq 0$ is the “turbulent viscosity coefficient.” The most common EV model is known in LES as the *Smagorinsky model* [32, 33, 36, 47, 51] in which

$$(1.1) \quad \nu_T = \nu_{\text{Smag}}(\bar{\mathbf{u}}, \delta) := (c_s \delta)^2 \|\nabla^s \bar{\mathbf{u}}\|_F.$$

Although the Smagorinsky model is easy to implement, is stable, and replicates energy dissipation rates, it is quite inaccurate for many problems. Probably the most common complaint for the Smagorinsky model (1.1) is that it is *too dissipative*. The reason is clearly illustrated in Figure 1.1: for large values of the deformation tensor $\nabla^s \bar{\mathbf{u}}$, the Smagorinsky model introduces an unbounded amount of artificial viscosity (AV). This behavior is manifest in practical computations of flows displaying large velocity deformation tensors, such as wall-bounded flows. For example, in turbulent channel flows and pipe flows [4], the Smagorinsky model yields unphysical results.

Different approaches have been devised to cope with this limitation: the “clipping procedure” [2, 8, 20, 31, 52], the van Driest damping [4, 26, 27, 50], the *Ri*-dependent Smagorinsky model [11, 15, 38, 44, 46, 48] (where *Ri* is the Richardson number, the square of the ratio of the buoyancy frequency, and the vertical shear), the dynamic SGS model [19], and the Lagrangian dynamic SGS model [39, 41]. All of these approaches target the same deficiency of the Smagorinsky model—its *overly diffusive character*.

In this paper, we consider a *bounded AV model* for the numerical simulation of turbulent flows with high velocity deformation tensors. The bounded AV model has a *general* form: it can be used to reduce the overly dissipative nature of the Smagorinsky model without massive a priori knowledge of the flow regime. The bounded AV model reads

$$(1.2) \quad \nu_T = \mu \delta^\sigma a(\delta \|\nabla^s \bar{\mathbf{u}}\|_F) \nabla^s \bar{\mathbf{u}},$$

where $a(\cdot)$ is a general function whose graph resembles that in Figure 1.1.

The bounded AV model was proposed in [28] as an alternative to the Smagorinsky model and yielded improved results for convection-dominated convection-diffusion problems. In this paper, we analyze and test the bounded AV model (1.2) in the numerical simulation of incompressible fluid flows.

The paper is organized as follows: In section 2, we discuss the commonly used EV models. We note the benefits and limitations of heuristic procedures in which a “clipping” of the Smagorinsky AV is performed and present (1.2) as a viable alternative to such strategies. In section 3, we provide the variational setting for which the NSE with (1.2) is solved and introduce the necessary notation. In section 4, we present some stability results for the variational solution to NSE with (1.2), which are generalizations of Leray’s inequality for the usual Navier–Stokes system. In section 5, we prove an error estimate for the semidiscrete finite element approximation of the NSE with (1.2). In section 6, we discuss the Newton approximation scheme as applied to the NSE with the bounded AV term (1.2). Finally, in section 7, we include finite element calculations for NSE with (1.2), which both support the theoretical error estimate of section 5 and show that the bounded AV model (1.2) yields better results than the Smagorinsky model (1.1) in the numerical simulation of turbulent flow in a 3D square duct. We provide both sequential computations for an academic vortex decay problem and parallel computations for a 3D square duct flow, using the Virginia Tech large eddy simulator (ViTLES).

2. Large eddy simulation. LES is a natural computational idea: when a numerical mesh is so coarse that the problem data and solution fluctuate significantly inside each mesh cell, it is reasonable to replace the problem data by mesh cell averages of that data and define an approximate solution that represents a mesh cell average of the true solution. Thus, if δ is the mesh cell width, then we should not seek to approximate the pointwise fluid velocity $\mathbf{u}(\mathbf{x}, t)$ but rather some mesh cell average $\bar{\mathbf{u}}(\mathbf{x}, t)$. The simplest such average is given by the convolution of the velocity \mathbf{u} with a rapidly decaying spatial filter $g_\delta(\mathbf{x})$ such as the sharp cut-off, box (top hat), Gaussian, or differential filters.

The essential idea of LES is the following: Pick a spatial filter $g_\delta(\mathbf{x})$ and define $\bar{\mathbf{u}}(\mathbf{x}, t) := (g_\delta * \mathbf{u})(\mathbf{x}, t)$. Derive appropriate equations for $\bar{\mathbf{u}}$ by convolving the NSE with the spatial filter. Solve the closure problem. Impose accurate boundary conditions for $\bar{\mathbf{u}}$. Then discretize the resulting continuum model and solve it. Generally, such an averaging suppresses any fluctuations in \mathbf{u} below $O(\delta)$ and preserves those on scales larger than $O(\delta)$. In many flows, the portion of the flow that must be modeled $\mathbf{u}' := \mathbf{u} - \bar{\mathbf{u}}$ is small relative to the portion that is calculated $\bar{\mathbf{u}}$. Models in LES tend to be both simple and accurate, and the overall computational cost is comparable to that of computing an (unreliable, underrefined) solution of the NSE on the same mesh.

For an incompressible fluid, the nondimensionalized form of the NSE is

$$(2.1) \quad \mathbf{u}_t - Re^{-1} \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla) \mathbf{u} + \nabla p = \mathbf{f}, \quad \text{in } \Omega \times (0, T),$$

$$(2.2) \quad \nabla \cdot \mathbf{u} = 0, \quad \text{in } \Omega \times (0, T),$$

$$(2.3) \quad \mathbf{u} = 0, \quad \text{on } \partial\Omega \times (0, T),$$

where $\Omega \subset \mathbb{R}^d$, $d = 2, 3$ is the flow domain, T is the final time, \mathbf{u} is the velocity, p is the pressure, and $Re := UL/\nu$ is the Reynolds number, defined as the ratio between the product of a characteristic length-scale L and a characteristic velocity U , and the kinematic viscosity ν .

To derive equations for $\bar{\mathbf{u}}$, we convolve the NSE with the chosen filter function $g_\delta(\mathbf{x})$. Using the fact that (for constant $\delta > 0$ and in the absence of boundaries) filtering commutes with differentiation gives the space-filtered NSE:

$$(2.4) \quad \bar{\mathbf{u}}_t - Re^{-1}\Delta\bar{\mathbf{u}} + \nabla \cdot (\bar{\mathbf{u}}\bar{\mathbf{u}}^T) + \nabla\bar{p} = -\nabla \cdot \boldsymbol{\tau}(\mathbf{u}, \mathbf{u}) + \bar{\mathbf{f}} \quad \text{in } \Omega \times (0, T),$$

$$(2.5) \quad \nabla \cdot \bar{\mathbf{u}} = 0 \quad \text{in } \Omega \times (0, T).$$

This system is not closed, since it involves both \mathbf{u} and $\bar{\mathbf{u}}$. The tensor $\boldsymbol{\tau}(\mathbf{u}, \mathbf{u}) = \overline{\mathbf{u}\mathbf{u}^T} - \bar{\mathbf{u}}\bar{\mathbf{u}}^T$ or $\tau_{ij}(\mathbf{u}, \mathbf{u}) = \overline{u_i u_j} - \bar{u}_i \bar{u}_j$ is often called the SGS tensor. Thus, the closure problem in LES is to model the SGS tensor $\boldsymbol{\tau}(\mathbf{u}, \mathbf{u})$, i.e., to specify a tensor $\mathcal{S} = \mathcal{S}(\bar{\mathbf{u}}, \bar{\mathbf{u}})$ to replace $\boldsymbol{\tau}(\mathbf{u}, \mathbf{u})$ in (2.4).

2.1. Eddy viscosity models. The most popular approach to the closure problem is the EV model, which is motivated by the idea that the global effect of $\boldsymbol{\tau}(\mathbf{u}, \mathbf{u})$, in the mean, is to transfer energy from resolved to unresolved scales through inertial interactions. EV models are motivated by Kolmogorov’s K-41 theory ([4, 45]), and, in particular, by the *energy cascade*. The essence of the energy cascade (see [43]) is that kinetic energy enters the turbulent flow at the largest scales of motion, and is then transferred by inviscid processes to smaller and smaller scales, until it is eventually dissipated through viscous effects. Thus, the action of $\boldsymbol{\tau}(\mathbf{u}, \mathbf{u})$ is thought of as having a dissipative effect on the mean flow: the scales uncaptured on the numerical mesh (above the cutoff wavenumber k_c) should dissipate energy from the large scales (below the cutoff wavenumber k_c).

Boussinesq [5] first formulated the *EV/Boussinesq hypothesis* based upon an analogy between the interaction of small eddies and the perfectly elastic collision of molecules (e.g., molecular viscosity or heat): “*Turbulent fluctuations are dissipative in the mean.*” The mathematical realization is the model

$$\nabla \cdot \boldsymbol{\tau}(\mathbf{u}, \mathbf{u}) \approx -\nabla \cdot (\nu_T \nabla^s \bar{\mathbf{u}}) + \text{terms incorporated into } \bar{p},$$

where $\nabla^s \bar{\mathbf{u}} := (\nabla \bar{\mathbf{u}} + \nabla \bar{\mathbf{u}}^T)/2$ is the deformation tensor of $\bar{\mathbf{u}}$ and $\nu_T \geq 0$ is the “turbulent viscosity coefficient.” The modeling problem then reduces to determining one parameter: the turbulent viscosity coefficient $\nu_T(\bar{\mathbf{u}}, \delta)$.

2.2. The Smagorinsky model. The most common EV model is known in LES as the Smagorinsky model, in which

$$(2.6) \quad \nu_T = \nu_{\text{Smag}}(\bar{\mathbf{u}}, \delta) := (c_s \delta)^2 \|\nabla^s \bar{\mathbf{u}}\|_F,$$

where δ is the filter radius, c_s is the Smagorinsky constant, and $\|\sigma\|_F := \sqrt{\sum_{i,j=1}^d |\sigma_{ij}|^2}$ is the Frobenius norm of the tensor σ . This model was studied in [51] as a nonlinear AV in gas dynamics and in [47] for geophysical flow calculations. A complete mathematical theory for partial differential equations involving this term was constructed by Ladyžhenskaya [32, 33].

The Smagorinsky model (1.1) where $c_s \sim 0.17$ [36] seems to be a universal answer in LES. It is easy to implement, stable, and (under “optimistic” assumptions) it replicates energy dissipation rates. Unfortunately, it can be also quite inaccurate for many problems.

The most successful form of the Smagorinsky model is the *dynamic* SGS model of [19], in which c_s is chosen locally in space and time, $c_s = c_s(\mathbf{x}, t)$. An essential improvement is that the dynamic SGS model introduces *backscatter*, the inverse transfer of energy from small scales to large scales [27, 4]. A yet improved version of the dynamic SGS model is the *Lagrangian dynamic* SGS model of [39, 41].

2.3. The overly diffusive character of the Smagorinsky model. Whether simplistic or more involved, all these approaches target the same deficiency of the Smagorinsky model—its overly diffusive character. This negative feature of the Smagorinsky model is clearly illustrated by the schematic in Figure 1.1. Plotting the amount of AV introduced by the Smagorinsky model against $\|\delta \nabla^s \bar{\mathbf{u}}\|_F$, we obtain a linear profile: Indeed, (1.1) can be rewritten as

$$(2.7) \quad \nu_T = \nu_{\text{Smag}}(\bar{\mathbf{u}}, \delta) = c_s^2 \delta \|\delta \nabla^s \bar{\mathbf{u}}\|_F,$$

which yields a linear profile for ν_T (if δ is held constant). In smooth regions of the flow, where the deformation tensor is relatively small ($\|\nabla^s \bar{\mathbf{u}}\|_F \leq O(1/\delta)$), the Smagorinsky model will introduce a moderate amount of AV ($\nu_T \leq O(\delta)$). In those regions of the flow where the deformation tensor is large ($\|\nabla^s \bar{\mathbf{u}}\|_F \geq O(1/\delta^2)$, for example), the Smagorinsky model will introduce an *unphysical* amount of AV ($\nu_T \sim O(1)$).

The overly diffusive feature of the Smagorinsky model is manifested in practical computations of flows displaying a large deformation tensor, such as wall-bounded flows. Indeed, for turbulent channel flows and pipe flows, because the velocity deformation tensor is very large near the solid wall, the Smagorinsky model introduces an unphysical amount of AV. Similarly, in stratified flows with large shear (and thus large deformation tensors), the Smagorinsky model introduces an unphysical amount of AV in the vertical direction.

There have been numerous modifications of the Smagorinsky model, all trying to attenuate its overly diffusive character. The simplest such approach is the “clipping procedure”

$$(2.8) \quad \nu_T = \nu_{\text{Smag}}^{\text{clipping}}(\bar{\mathbf{u}}, \delta) := \min\{\nu_{\text{Smag}}(\bar{\mathbf{u}}, \delta), C\},$$

where C is a user-defined constant [2, 8, 20, 31, 52].

A more involved approach for wall-bounded flows (such as channel and pipe flows) is the van Driest damping [4, 26, 27, 50], in which

$$(2.9) \quad \nu_T = \nu_{\text{Smag}}^{VD}(\bar{\mathbf{u}}, \delta) := \left[\left(1 - e^{-\frac{y^+}{25}} \right) \right] \nu_{\text{Smag}}(\bar{\mathbf{u}}, \delta),$$

where y^+ is the nondimensionalized distance to the wall (see Chapter 12 in [4] for more details). The main improvement over the ad hoc clipping procedure (2.8) is that the damping function in (2.9) is chosen so that the resulting flow satisfies the turbulent boundary layer theory [4].

In stratified flows, the Smagorinsky model is used with a damping function in the vertical direction [11, 15, 38, 44, 46, 48]:

$$(2.10) \quad \nu_T^z = \nu_{\text{Smag}}^{Ri}(\bar{\mathbf{u}}, \delta) := \sqrt{1 - \frac{Ri}{Ri_c}} \nu_{\text{Smag}}(\bar{\mathbf{u}}, \delta),$$

where Ri is the Richardson number, the square of the ratio of the buoyancy frequency and the vertical shear, and Ri_c is a critical Richardson number (a popular choice is $Ri_c \sim 0.25$) [40, 44].

2.4. The bounded AV model. In this paper, we consider the bounded AV model, a general, mathematically sound alternative to the Smagorinsky model. The bounded AV model reads

$$(2.11) \quad \nu_T = \mu \delta^\sigma a(\delta \|\nabla^s \bar{\mathbf{u}}\|_F) \nabla^s \bar{\mathbf{u}},$$

where $a(\cdot)$ is a general function whose graph resembles that in Figure 1.1. This new model, proposed in [28] for convection-diffusion problems, is a clear improvement over the Smagorinsky model. Indeed, in the flow regions with large velocity deformation tensors, the bounded AV model introduces a *bounded* amount of AV, just enough to spread the solution onto the computational mesh. This is in clear contrast with the Smagorinsky model, which introduces an unbounded amount of AV, thus being overly dissipative. The improvement of the bounded AV model over the Smagorinsky model is clearly supported by the numerical simulation of a turbulent flow in a 3D square duct in section 7.

Another distinct advantage of the bounded AV model over other modifications of the Smagorinsky model is that, when appropriately chosen, the bounded AV term represents a monotonic semilinear operator. This property allows for existence and uniqueness results for the finite element approximation as well as the error estimates presented herein. It is important to note that the results of this paper are more straightforward for the bounded AV model than for the previously mentioned heuristic AV bounding techniques, as the bounded AV model translates more readily into the mathematical framework established for the NSE.

The bounded AV model is general. Indeed, the function $a(\cdot)$ in (1.2) is just required to be bounded and monotonically increasing (see Figure 1.1). Thus, the bounded AV model clearly includes the ad hoc “clipped” Smagorinsky model (2.8) as a particular case. Although the bounded AV model does not directly include the Smagorinsky model with van Driest damping (2.9) ($a(\cdot)$ must be monotonically increasing) or the *Ri*-dependent Smagorinsky model (2.10) ($a(\cdot)$ depends on $\nabla^s \mathbf{u}$, whereas (2.10) depends on $\frac{\partial \mathbf{u}}{\partial z}$), it is certainly related to these two models, targeting the overly diffusive character of the Smagorinsky model. Note that while models (2.9) and (2.10) are tailored for specific flows (wall-bounded and stratified, respectively), the bounded AV model is not restricted to any particular type of flow. Of course, the function $a(\cdot)$ should be optimized for each particular flow setting in which the bounded AV model is used. To this end, extensive a priori and a posteriori testing [4] should be carried out for each such flow setting.

It was shown in [28] that the bounded AV model yields a clear improvement over the Smagorinsky model in the numerical simulation of convection-dominated convection-diffusion problems with sharp transition layers. In this paper, we show that the bounded AV model is a dramatic improvement over the Smagorinsky model in the numerical simulation of a turbulent flow in a 3D square duct.

There are numerous challenges in the numerical analysis of LES, where the study of classic topics such as consistency, stability, and convergence of the LES discretization are still at an initial stage. Only the first few steps along these lines have been made, some of which are presented in the exquisite monograph of John [30]. A thorough numerical analysis for the finite element implementation of the Smagorinsky model has been presented in [12, 13]. Further studies have been presented in [29, 34].

In this paper, we present a rigorous numerical analysis for the finite element implementation of the bounded AV model:

$$\begin{aligned}
 (2.12) \quad & \mathbf{w}_t - Re^{-1} \Delta \mathbf{w} - \nabla \cdot (\mu \delta^\sigma a(\delta \|\nabla^s \mathbf{w}\|_F) \nabla^s \mathbf{w}) \\
 & + (\mathbf{w} \cdot \nabla) \mathbf{w} - \nabla q = \bar{\mathbf{f}} \quad \text{in } \Omega \times (0, T), \\
 (2.13) \quad & \nabla \cdot \mathbf{w} = 0 \quad \text{in } \Omega \times (0, T), \\
 (2.14) \quad & \mathbf{w} = 0 \quad \text{on } \partial\Omega \times (0, T).
 \end{aligned}$$

We also illustrate our error estimates with numerical simulations, using the bounded

AV model for the 3D square duct turbulent flow and the two-dimensional (2D) vortex decay problem.

Remark 2.1. The numerical analysis presented herein is concerned with the *numerical error* (i.e., $\mathbf{w} - \mathbf{w}_h$) associated with the bounded AV model, and not the *modeling error* (i.e., $\bar{\mathbf{u}} - \mathbf{w}$) associated with the proposed EV model.

3. The variational formulation. In this section, we develop the variational formulation for (2.12)–(2.14). We will denote the usual Sobolev spaces [1] by $W^{m,p}(\Omega)$, with norms $\|\cdot\|_{W^{m,p}}$ and seminorms $|\cdot|_{W^{m,p}}$, and set $H^m(\Omega) := W^{m,2}(\Omega)$ and $L^p(\Omega) := W^{0,p}(\Omega)$. In what follows, we will denote $\|\cdot\|$ and (\cdot, \cdot) the norm and inner product for $L^2(\Omega)$, and $\|\cdot\|_m$ the norm for $H^m(\Omega)$. The vector spaces and vector functions will be indicated by boldface type letters.

Specifically, we use the following function spaces for the variational formulation:

$$\begin{aligned} \text{Velocity space : } \mathbf{X} &:= \mathbf{H}_0^1(\Omega) := \{ \mathbf{v} \in \mathbf{H}^1(\Omega) : \mathbf{v} = 0 \text{ on } \partial\Omega \}, \\ \text{Pressure space : } Q &:= L_0^2(\Omega) := \left\{ q \in L^2(\Omega) : \int_{\Omega} q \, dx = 0 \right\}. \end{aligned}$$

The velocity and pressure spaces \mathbf{X} and Q satisfy the inf-sup condition [21]

$$(3.1) \quad \inf_{\lambda \in Q} \sup_{\mathbf{v} \in \mathbf{X}} \frac{(\lambda, \nabla \cdot \mathbf{v})}{\|\lambda\| \|\mathbf{v}\|_1} \geq \beta > 0.$$

The inf-sup condition (3.1), in turn, implies that the space of weakly divergence-free functions \mathbf{V} :

$$(3.2) \quad \mathbf{V} := \{ \mathbf{v} \in \mathbf{X} : (\lambda, \nabla \cdot \mathbf{v}) = 0, \forall \lambda \in Q \}$$

is a well-defined, nontrivial, closed subspace of \mathbf{X} [21].

The variational formulation of (2.12)–(2.14) proceeds in the usual manner. Multiplying (2.12) and (2.13) by a velocity (\mathbf{v}) and pressure (λ) test function, respectively, integrating over Ω , and integrating by parts (using the fact that $\mathbf{v} = 0$ on $\partial\Omega$), we obtain

$$(3.3) \quad (\mathbf{w}_t, \mathbf{v}) + A(\mathbf{w}, \mathbf{v}) + B(\mathbf{w}, \mathbf{w}, \mathbf{v}) + C(\mathbf{w}, \mathbf{w}, \mathbf{v}) - (q, \nabla \cdot \mathbf{v}) = (\bar{\mathbf{f}}, \mathbf{v}), \quad \forall \mathbf{v} \in \mathbf{X}, \quad t \in [0, T],$$

$$(3.4) \quad (\nabla \cdot \mathbf{w}, \lambda) = 0, \quad \forall \lambda \in Q, \quad t \in [0, T],$$

where the bilinear form $A(\cdot, \cdot)$ is defined by

$$A(\mathbf{w}, \mathbf{v}) := Re^{-1} (\nabla \mathbf{w}, \nabla \mathbf{v})$$

and the trilinear forms $B(\cdot, \cdot, \cdot)$ and $C(\cdot, \cdot, \cdot)$ are defined by

$$\begin{aligned} B(\mathbf{u}, \mathbf{v}, \mathbf{w}) &:= \mu \delta^\sigma (a(\delta \|\nabla \mathbf{u}\|_F) \nabla \mathbf{v}, \nabla \mathbf{w}), \\ C(\mathbf{u}, \mathbf{v}, \mathbf{w}) &:= \frac{1}{2} (\mathbf{u} \cdot \nabla \mathbf{v}, \mathbf{w}) - \frac{1}{2} (\mathbf{u} \cdot \nabla \mathbf{w}, \mathbf{v}). \end{aligned}$$

It is a simple index calculation to check that, for $\mathbf{v} \in \mathbf{X}$, $\mathbf{w} \in \mathbf{V}$, $(\mathbf{w} \cdot \nabla \mathbf{w}, \mathbf{v}) = C(\mathbf{w}, \mathbf{w}, \mathbf{v})$.

Remark 3.1. Although the bounded AV model (2.12)–(2.14) depends on $\nabla^s \mathbf{w}$, for clarity we will replace $\nabla^s \mathbf{w}$ by $\nabla \mathbf{w}$. The same numerical analysis can be carried out with the $\nabla^s \mathbf{w}$ by using Korn’s inequalities, which relate the L^p -norms of the deformation tensor $\nabla^s \mathbf{w}$ to the same norms of the gradient $\nabla \mathbf{w}$ for $1 < p < \infty$ [16, 29].

3.1. Finite element spaces. Let $\Omega \subset \mathbb{R}^d$, ($d = 2, 3$) be a polygonal domain, and let T_h denote a triangulation of Ω made up of triangles (in \mathbb{R}^2) or tetrahedra (in \mathbb{R}^3). Thus, the computational domain is defined by

$$\Omega = \bigcup K, \quad K \in T_h.$$

We also assume that for a particular triangulation T_h of Ω , there exist positive constants c_1, c_2 such that

$$c_1 h \leq h_K \leq c_2 \rho_K,$$

where h_K is the diameter of K , ρ_K is the diameter of the greatest ball (sphere) included in K , and $h = \max_{K \in T_h} h_K$. Let $P_k(A)$ denote the space of polynomials on a subdomain A of degree no greater than k . We define the conforming finite element spaces associated with the velocity and pressure spaces as follows:

$$(3.5) \quad \mathbf{X}_h := \{ \mathbf{v}_h \in \mathbf{X} \cap C(\overline{\Omega})^d : \mathbf{v}_h|_K \in P_k(K), \forall K \in T_h \},$$

$$(3.6) \quad Q_h := \{ \lambda_h \in Q \cap C(\overline{\Omega}) : \lambda_h|_K \in P_l(K), \forall K \in T_h \},$$

where $C(\overline{\Omega})$ denotes the set of continuous functions on the closure of Ω . Analogous to the continuous inf-sup condition, the spaces \mathbf{X}_h, Q_h satisfy the *discrete* inf-sup condition [14, 21]

$$(3.7) \quad \inf_{\lambda_h \in Q_h} \sup_{\mathbf{v}_h \in \mathbf{X}_h} \frac{(\lambda_h, \nabla \cdot \mathbf{v}_h)}{\|\lambda_h\| \|\mathbf{v}_h\|_1} \geq \beta > 0.$$

The *discrete* inf-sup condition (3.7), in turn, implies that the space of weakly divergence-free functions \mathbf{V}_h

$$(3.8) \quad \mathbf{V}_h := \{ \mathbf{v}_h \in \mathbf{X}_h : (\lambda_h, \nabla \cdot \mathbf{v}_h) = 0, \forall \lambda_h \in Q_h \}$$

is a well-defined, nontrivial, closed subspace of \mathbf{X}_h [14, 21, 22].

We assume that the finite element spaces \mathbf{X}_h, Q_h satisfy the usual approximation properties [6, 21]: For $(\mathbf{w}, \lambda) \in \mathbf{H}^{k+1}(\Omega) \times H^{l+1}(\Omega)$, there exist interpolants $(I_h \mathbf{w}, I_h \lambda) \in \mathbf{X}_h \times Q_h$ satisfying [6, 21]

$$(3.9) \quad \|\mathbf{w} - I_h \mathbf{w}\| \leq C_I h^{k+1} |\mathbf{w}|_{\mathbf{H}^{k+1}},$$

$$(3.10) \quad \|\mathbf{w} - I_h \mathbf{w}\|_1 \leq C_I h^k |\mathbf{w}|_{\mathbf{H}^{k+1}},$$

$$(3.11) \quad \|\lambda - I_h \lambda\| \leq C_I h^{l+1} |\lambda|_{H^{l+1}}.$$

From [6], we have the following useful results concerning interpolation.

LEMMA 3.1. *Let $\{T_h\}$ ($0 < h \leq 1$) denote a quasi-uniform family of subdivisions of a polyhedral domain $\Omega \subset \mathbb{R}^d$. Let (\hat{K}, P, N) be a reference finite element such that $P \subset W^{l,p}(\hat{K}) \cap W^{m,q}(\hat{K})$, where $1 \leq p, q \leq \infty$, and $0 \leq m \leq l$. For $K \in T_h$, let (K, P_K, N_K) be the affine equivalent element and*

$$V_h := \left\{ v : v \text{ is measurable and } v|_K \in P_K, \forall K \in T_h \right\}.$$

Then there exists $C = C(l, p, q)$ such that

$$(3.12) \quad \left[\sum_{K \in T_h} \|v\|_{W^{l,p}(K)}^p \right]^{1/p} \leq C h^{m-l+\min(0, \frac{d}{p}-\frac{d}{q})} \left[\sum_{K \in T_h} \|v\|_{W^{m,q}(K)}^q \right]^{1/q}.$$

4. Stability results. In this section, we prove some stability results concerning the variational problem (3.3)–(3.4), as well as its semidiscrete finite element approximation. Useful in the following analysis are the following three lemmas.

LEMMA 4.1 (monotonicity of $B(\cdot, \cdot, \cdot)$ [28]). *For $\mathbf{u}, \mathbf{v} \in \mathbf{X}$, and the function $a(\cdot)$ satisfying*

$$0 \leq a(x) \leq 1, \quad a'(x) \geq 0, \quad \forall x \in [0, \infty),$$

we have

$$B(\mathbf{u}, \mathbf{u}, \mathbf{u} - \mathbf{v}) - B(\mathbf{v}, \mathbf{v}, \mathbf{u} - \mathbf{v}) \geq 0.$$

Proof. Consider the functional $I : \mathbf{X} \rightarrow \mathbb{R}$, defined by

$$I(\mathbf{U}) := \int_{\Omega} A(\|\nabla \mathbf{U}\|_F) \, d\mathbf{x},$$

where the function $A : [0, \infty) \rightarrow \mathbb{R}$ is defined by

$$A(x) := \int_0^x t a(t) dt.$$

First, note that

$$dI(\mathbf{U}, \mathbf{V}) = \int_{\Omega} A'(\|\nabla \mathbf{U}\|_F) \frac{\nabla \mathbf{U}}{\|\nabla \mathbf{U}\|_F} \nabla \mathbf{V} \, d\mathbf{x} = \int_{\Omega} a(\|\nabla \mathbf{U}\|_F) \nabla \mathbf{U} \nabla \mathbf{V} \, d\mathbf{x},$$

where $dI(\mathbf{U}, \mathbf{V})$ is the Gâteaux derivative of I at \mathbf{U} in the direction of \mathbf{V} .

Therefore, setting $\mathbf{U}_1 := \delta \mathbf{u}$, $\mathbf{U}_2 := \delta \mathbf{v}$, and $\mathbf{V} := \mathbf{U}_1 - \mathbf{U}_2$, we have

$$(4.1) \quad B(\mathbf{u}, \mathbf{u}, \mathbf{u} - \mathbf{v}) - B(\mathbf{v}, \mathbf{v}, \mathbf{u} - \mathbf{v}) = \frac{\mu \delta^\sigma}{\delta^2} (dI(\mathbf{U}_1, \mathbf{V}) - dI(\mathbf{U}_2, \mathbf{V})).$$

However, we can rewrite this expression as

$$\begin{aligned} dI(\mathbf{U}_1, \mathbf{V}) - dI(\mathbf{U}_2, \mathbf{V}) &= \int_0^1 \frac{d}{dt} dI(\mathbf{U}_2 + t(\mathbf{U}_1 - \mathbf{U}_2), \mathbf{V}) \, dt \\ &= \int_0^1 \frac{d}{dt} \int_{\Omega} a(\|\nabla(\mathbf{U}_2 + t(\mathbf{U}_1 - \mathbf{U}_2))\|_F) \nabla(\mathbf{U}_2 + t(\mathbf{U}_1 - \mathbf{U}_2)) \nabla \mathbf{V} \, d\mathbf{x} \, dt \\ &= \int_0^1 \int_{\Omega} a'(\|\nabla(\mathbf{U}_2 + t(\mathbf{U}_1 - \mathbf{U}_2))\|_F) \frac{\nabla(\mathbf{U}_2 + t(\mathbf{U}_1 - \mathbf{U}_2)) \nabla \mathbf{V}}{\|\nabla(\mathbf{U}_2 + t(\mathbf{U}_1 - \mathbf{U}_2)) \nabla \mathbf{V}\|_F} \\ &\quad \nabla(\mathbf{U}_2 + t(\mathbf{U}_1 - \mathbf{U}_2)) \nabla \mathbf{V} \, d\mathbf{x} \, dt \\ (4.2) \quad &+ \int_0^1 \int_{\Omega} a(\|\nabla(\mathbf{U}_2 + t(\mathbf{U}_1 - \mathbf{U}_2))\|_F) \|\nabla \mathbf{V}\|_F^2 \, d\mathbf{x} \, dt. \end{aligned}$$

As $a(x), a'(x) \geq 0$, it is clear that the expression in (4.2) is nonnegative. Finally, using (4.1), we obtain the stated result. \square

Remark 4.1. Notice that Lemma 4.1 states that the bounded AV operator $B(\cdot, \cdot, \cdot)$ is monotone but not strongly monotone. The Smagorinsky AV operator

$$B^{Smag}(\mathbf{u}, \mathbf{v}, \mathbf{w}) := ((c_s \delta)^2 \|\nabla^s \mathbf{u}\|_F \nabla^s \mathbf{v}, \nabla^s \mathbf{w}),$$

on the other hand, is strongly monotone [13, 29, 34]. Indeed, $\forall \mathbf{u}, \mathbf{v} \in \mathbf{W}^{1,3}(\Omega)$,

$$B^{Smag}(\mathbf{u}, \mathbf{u}, \mathbf{u} - \mathbf{v}) - B^{Smag}(\mathbf{v}, \mathbf{v}, \mathbf{u} - \mathbf{v}) \geq C \delta^2 \|\nabla^s(\mathbf{u} - \mathbf{v})\|_{L^3}^3.$$

Therefore, the error estimate we prove in Theorem 5.1 for the bounded AV model assumes higher regularity for the solution \mathbf{w} ($\mathbf{w} \in L^4(0, T; \mathbf{W}^{1,\infty}(\Omega))$) than the regularity for \mathbf{w} assumed for the Smagorinsky model [29] ($\mathbf{w} \in L^2(0, T; \mathbf{W}^{1,\infty}(\Omega))$).

LEMMA 4.2. For $\mathbf{u}_1, \mathbf{u}_2, \mathbf{v}, \mathbf{w} \in \mathbf{X}$, and the function $a(\cdot)$ satisfying

$$0 \leq a(x) \leq 1, \quad 0 \leq a'(x) \leq M_a, \quad \forall x \in [0, \infty),$$

we have

$$(4.3) \quad |B(\mathbf{u}_1, \mathbf{v}, \mathbf{w}) - B(\mathbf{u}_2, \mathbf{v}, \mathbf{w})| \leq M_a \mu \delta^{\sigma+1} (\|\nabla \mathbf{u}_1 - \nabla \mathbf{u}_2\|_F \|\nabla \mathbf{v}\|_F, \|\nabla \mathbf{w}\|_F)$$

and

$$(4.4) \quad |B(\mathbf{u}_1, \mathbf{v}, \mathbf{w})| \leq M_a \mu \delta^{\sigma+1} (\|\nabla \mathbf{u}_1\|_F \|\nabla \mathbf{v}\|_F, \|\nabla \mathbf{w}\|_F).$$

Proof. Without loss of generality, we assume that $\|\nabla \mathbf{u}_1\|_F \geq \|\nabla \mathbf{u}_2\|_F$. Immediately, we have

$$(4.5) \quad |B(\mathbf{u}_1, \mathbf{v}, \mathbf{w}) - B(\mathbf{u}_2, \mathbf{v}, \mathbf{w})| = \mu \delta^\sigma |(a(\delta \|\nabla \mathbf{u}_1\|_F) - a(\delta \|\nabla \mathbf{u}_2\|_F)) \nabla \mathbf{v}, \nabla \mathbf{w}|.$$

Now, by the mean value theorem, there exists $c_a \in [\delta \|\nabla \mathbf{u}_2\|_F, \delta \|\nabla \mathbf{u}_1\|_F]$ such that

$$a(\delta \|\nabla \mathbf{u}_1\|_F) - a(\delta \|\nabla \mathbf{u}_2\|_F) = a'(c_a) \delta (\|\nabla \mathbf{u}_1\|_F - \|\nabla \mathbf{u}_2\|_F).$$

Combining this with the reverse triangle inequality $||x| - |y|| \leq |x - y|$, we have

$$(4.6) \quad |a(\delta \|\nabla \mathbf{u}_1\|_F) - a(\delta \|\nabla \mathbf{u}_2\|_F)| \leq a'(c_a) \delta \|\nabla \mathbf{u}_1 - \nabla \mathbf{u}_2\|_F.$$

Finally, substituting (4.6) into (4.5) and noting that $a'(c_a) \leq M_a$, we obtain (4.3). The result (4.4) follows directly. \square

Remark 4.2. Again, the bounded AV operator $B(\cdot, \cdot, \cdot)$ satisfies a weaker inequality than the Smagorinsky AV operator $B^{Smag}(\cdot, \cdot, \cdot)$ [13, 29, 34].

LEMMA 4.3 (Leray’s inequality for the bounded AV model). A solution of (3.3)–(3.4) satisfies

$$\frac{1}{2} \|\mathbf{w}(t)\|^2 + \int_0^t Re^{-1} \|\nabla \mathbf{w}\|^2 ds \leq \frac{1}{2} \|\mathbf{w}(0)\|^2 + \int_0^t (\mathbf{f}, \mathbf{w}) ds \quad \forall t \in [0, T].$$

Proof. The stated result follows by setting $\mathbf{v} = \mathbf{w}$ and $\lambda = q$ in (3.3) and (3.4), noting that

$$B(\mathbf{w}, \mathbf{w}, \mathbf{w}) \geq 0, \quad C(\mathbf{w}, \mathbf{w}, \mathbf{w}) = 0,$$

and integrating from 0 to t . \square

We now define the semidiscrete approximation as the solution of (3.3)–(3.4) restricted to the finite element spaces \mathbf{X}_h, Q_h .

DEFINITION 4.4 (The semidiscrete approximation). The semidiscrete approximation is defined to be an element $(\mathbf{w}_h, q_h) \in C(0, T; \mathbf{X}_h) \cap C(0, T; Q_h)$ such that

$$(4.7) \quad (\mathbf{w}_{h,t}, \mathbf{v}_h) + A(\mathbf{w}_h, \mathbf{v}_h) + B(\mathbf{w}_h, \mathbf{w}_h, \mathbf{v}_h) + C(\mathbf{w}_h, \mathbf{w}_h, \mathbf{v}_h) - (q_h, \nabla \cdot \mathbf{v}_h) = (\bar{\mathbf{f}}, \mathbf{v}_h), \quad \forall \mathbf{v}_h \in \mathbf{X}_h, \quad t \in [0, T],$$

$$(4.8) \quad (\nabla \cdot \mathbf{w}_h, \lambda_h) = 0, \quad \forall \lambda_h \in Q_h, \quad t \in [0, T].$$

We immediately obtain the following two lemmas.

LEMMA 4.5 (Leray's inequality for \mathbf{w}_h). *A solution of (4.7)–(4.8) satisfies*

$$\frac{1}{2} \|\mathbf{w}_h(t)\|^2 + \int_0^t Re^{-1} \|\nabla \mathbf{w}_h\|^2 ds \leq \frac{1}{2} \|\mathbf{w}_h(0)\|^2 + \int_0^t (\overline{\mathbf{f}}, \mathbf{w}_h) ds.$$

Proof. Setting $\mathbf{v}_h = \mathbf{w}_h$ and $\lambda_h = q_h$ in (4.7)–(4.8), we immediately have

$$(4.9) \quad \frac{1}{2} \frac{d}{dt} \|\mathbf{w}_h(t)\|^2 + Re^{-1} \|\nabla \mathbf{w}_h\|^2 \leq (\overline{\mathbf{f}}, \mathbf{w}_h).$$

Integrating from 0 to t thus yields the stated result. \square

LEMMA 4.6 (Stability of \mathbf{w}_h). *A solution \mathbf{w}_h of (4.7)–(4.8) satisfies*

$$(4.10) \quad \begin{aligned} \|\mathbf{w}_h(t)\|^2 + Re^{-1} C(\Omega) \int_0^t \|\mathbf{w}_h\|_{\mathbf{H}^1}^2 ds \\ \leq \|\mathbf{w}_h(0)\|^2 + \frac{Re}{C(\Omega)} \int_0^t \|\overline{\mathbf{f}}\|_{\mathbf{H}^{-1}}^2 ds \end{aligned}$$

and

$$(4.11) \quad \begin{aligned} \|\mathbf{w}_h(t)\|^2 + 2Re^{-1} \int_0^t e^{t-s} \|\nabla \mathbf{w}_h\|^2 ds \\ \leq e^t \|\mathbf{w}_h(0)\|^2 + \int_0^t e^{t-s} \|\overline{\mathbf{f}}\|^2 ds, \end{aligned}$$

where $C(\Omega)$ denotes a generic constant depending on Ω .

Proof. By using the Cauchy–Schwarz and Young's inequalities, we have

$$(4.12) \quad (\overline{\mathbf{f}}, \mathbf{w}_h) \leq \frac{\varepsilon}{2} \|\mathbf{w}_h\|_{\mathbf{H}^1}^2 + \frac{1}{2\varepsilon} \|\overline{\mathbf{f}}\|_{\mathbf{H}^{-1}}^2.$$

By using Poincaré's inequality [17], we get

$$(4.13) \quad Re^{-1} C(\Omega) \|\mathbf{w}_h\|_{\mathbf{H}^1}^2 \leq Re^{-1} \|\nabla \mathbf{w}_h\|^2.$$

Inserting (4.12) with $\varepsilon := Re^{-1} C(\Omega)$ and (4.13) in (4.9), we obtain

$$(4.14) \quad \frac{1}{2} \frac{d}{dt} \|\mathbf{w}_h\|^2 + \frac{Re^{-1} C(\Omega)}{2} \|\mathbf{w}_h\|_{\mathbf{H}^1}^2 \leq \frac{Re}{2C(\Omega)} \|\overline{\mathbf{f}}\|_{\mathbf{H}^{-1}}^2.$$

By integrating (4.14) from 0 to t , we get (4.10).

By using the Cauchy–Schwarz and Young's inequalities, we have

$$(4.15) \quad (\overline{\mathbf{f}}, \mathbf{w}_h) \leq \frac{1}{2} \|\mathbf{w}_h\|^2 + \frac{1}{2} \|\overline{\mathbf{f}}\|^2.$$

By using (4.15) in (4.9), we obtain

$$\frac{1}{2} \frac{d}{dt} \|\mathbf{w}_h\|^2 - \frac{1}{2} \|\mathbf{w}_h\|^2 + Re^{-1} \|\nabla \mathbf{w}_h\|^2 \leq \frac{1}{2} \|\overline{\mathbf{f}}\|^2.$$

The positivity of the exponential implies

$$e^{-s} \left(\frac{d}{ds} \|\mathbf{w}_h\|^2 - \|\mathbf{w}_h\|^2 + 2Re^{-1} \|\nabla \mathbf{w}_h\|^2 \right) \leq e^{-s} \|\overline{\mathbf{f}}\|^2.$$

By integrating from 0 to $t < T$ and then multiplying by e^t , we get (4.11). \square

LEMMA 4.7 (Existence of (\mathbf{w}_h, q_h)). *There exists a solution (\mathbf{w}_h, q_h) of the semidiscrete approximation (4.7)–(4.8).*

Proof. Since $\dim(\mathbf{X}_h) < \infty$ and any possible solution (\mathbf{w}_h, q_h) satisfies the a priori stability estimates in Lemma 4.6, Schauder’s fixed point theorem [35] implies existence of a solution (\mathbf{w}_h, q_h) of the semidiscrete approximation (4.7)–(4.8). \square

REMARK 4.3 (Uniqueness of (\mathbf{w}_h, q_h)). The uniqueness of the solution (\mathbf{w}_h, q_h) of the semidiscrete approximation (4.7)–(4.8) would follow by using a general argument: Assume that there exist two distinct solutions $(\mathbf{w}_{1h}, q_{1h})$ and $(\mathbf{w}_{2h}, q_{2h})$ of (4.7)–(4.8); subtract (4.7)–(4.8) corresponding to the two solutions; use the coercivity of the operators in the error equation to obtain inequalities of the form $\|\mathbf{w}_{1h} - \mathbf{w}_{2h}\| \leq 0$ and $\|q_{1h} - q_{2h}\| \leq 0$.

This approach, however, fails for (4.7)–(4.8) because of the nonlinear term $C(\cdot, \cdot, \cdot)$ and the bounded AV operator $B(\cdot, \cdot, \cdot)$ is just monotone and *not* strongly monotone (see Lemma 4.1).

Note that when $\delta \rightarrow 0$, the bounded AV operator $B(\cdot, \cdot, \cdot)$ becomes strongly monotone (see Remark 4.1), which could, in turn, allow us to prove the uniqueness of (\mathbf{w}_h, q_h) .

5. An a priori error estimate. In order to prove an a priori error estimate for the semidiscrete approximation (\mathbf{w}_h, q_h) , we will assume that the solution to the continuous problem satisfies $\mathbf{w} \in L^4(0, T; \mathbf{W}^{1,\infty}(\Omega))$.

THEOREM 5.1. *Assume that the system (3.3)–(3.4) has a solution $(\mathbf{w}, q) \in \mathbf{X} \times Q$ which satisfies*

$$(5.1) \quad \mathbf{w} \in L^4(0, T; \mathbf{W}^{1,\infty}(\Omega)).$$

Then, there exist generic constants C and $C_1(\mathbf{w})$ independent of Re , such that the error $\mathbf{w} - \mathbf{w}_h$ satisfies for $T > 0$

$$(5.2) \quad \begin{aligned} & \|\mathbf{w} - \mathbf{w}_h\|_{L^\infty(0,T;L^2)}^2 + Re^{-1} \|\nabla(\mathbf{w} - \mathbf{w}_h)\|_{L^2(0,T;L^2)}^2 \\ & \leq C \exp(C_1(\mathbf{w})) \|(\mathbf{w} - \mathbf{w}_h)(\mathbf{x}, 0)\|^2 \\ & \quad + C \inf_{\tilde{\mathbf{w}} \in \mathbf{V}_h, \lambda_h \in Q_h} \mathcal{F}(\mathbf{w} - \tilde{\mathbf{w}}, q - \lambda_h, \delta, Re), \end{aligned}$$

where

$$\begin{aligned} & \mathcal{F}(\mathbf{w} - \tilde{\mathbf{w}}, q - \lambda_h, \delta, Re) \\ := & \|\mathbf{w} - \tilde{\mathbf{w}}\|_{L^\infty(0,T;L^2)}^2 + Re^{-1} \|\nabla(\mathbf{w} - \tilde{\mathbf{w}})\|_{L^2(0,T;L^2)}^2 \\ & + \exp(C_1(\mathbf{w})) \left[\|(\mathbf{w} - \tilde{\mathbf{w}})_t\|_{L^2(0,T;L^2)}^2 \right. \\ & + \|\mathbf{w} - \tilde{\mathbf{w}}\|_{L^2(0,T;L^2)}^2 + \|\nabla(\mathbf{w} - \tilde{\mathbf{w}})\|_{L^2(0,T;L^2)}^2 \\ & \left. + \|\nabla(\mathbf{w} - \tilde{\mathbf{w}})\|_{L^4(0,T;L^2)}^2 + \|q - \lambda_h\|_{L^2(0,T;L^2)}^2 \right]. \end{aligned}$$

Proof. First note that, for standard piecewise polynomial finite element spaces, it is known that the L^p -projection of a function in $L^p, p \geq 2$, is in L^p itself, and the L^2 -projection operator is stable in $L^p, 2 \leq p \leq \infty$ [9].

Let the error in \mathbf{w} be denoted by $\mathbf{e} := \mathbf{w} - \mathbf{w}_h$, and $\tilde{\mathbf{w}}$ denote a stable approximation of \mathbf{w} in \mathbf{V}_h , for example, the L^2 -projection under the conditions of [9].

The error \mathbf{e} is decomposed as

$$(5.3) \quad \mathbf{e} = (\mathbf{w} - \tilde{\mathbf{w}}) + (\tilde{\mathbf{w}} - \mathbf{w}_h) := \boldsymbol{\eta} + \boldsymbol{\phi}_h,$$

where $\boldsymbol{\eta} := \mathbf{w} - \tilde{\mathbf{w}}$ and $\boldsymbol{\phi}_h := \tilde{\mathbf{w}} - \mathbf{w}_h \in \mathbf{V}_h$. By subtracting (4.7) from (3.3) and using that $\mathbf{w} \in \mathbf{V}$, we obtain an error equation

$$(5.4) \quad (\mathbf{e}_t, \mathbf{v}_h) + A(\mathbf{e}, \mathbf{v}_h) + [B(\mathbf{w}, \mathbf{w}, \mathbf{v}_h) - B(\mathbf{w}_h, \mathbf{w}_h, \mathbf{v}_h)] \\ + [C(\mathbf{w}, \mathbf{w}, \mathbf{v}_h) - C(\mathbf{w}_h, \mathbf{w}_h, \mathbf{v}_h)] - (q - \lambda_h, \nabla \cdot \mathbf{v}_h) = 0 \quad \forall (\mathbf{v}_h, \lambda_h) \in \mathbf{V}_h \times Q_h.$$

By adding and subtracting terms and setting $\mathbf{v}_h := \boldsymbol{\phi}_h$, (5.4) becomes

$$(5.5) \quad (\boldsymbol{\phi}_{h,t}, \boldsymbol{\phi}_h) + A(\boldsymbol{\phi}_h, \boldsymbol{\phi}_h) + [B(\tilde{\mathbf{w}}, \tilde{\mathbf{w}}, \boldsymbol{\phi}_h) - B(\mathbf{w}_h, \mathbf{w}_h, \boldsymbol{\phi}_h)] \\ = -(\boldsymbol{\eta}_t, \boldsymbol{\phi}_h) - A(\boldsymbol{\eta}, \boldsymbol{\phi}_h) - [B(\mathbf{w}, \mathbf{w}, \boldsymbol{\phi}_h) - B(\tilde{\mathbf{w}}, \tilde{\mathbf{w}}, \boldsymbol{\phi}_h)] \\ - [C(\mathbf{w}, \mathbf{w}, \boldsymbol{\phi}_h) - C(\mathbf{w}_h, \mathbf{w}_h, \boldsymbol{\phi}_h)] + (q - \lambda_h, \nabla \cdot \boldsymbol{\phi}_h).$$

By using the monotonicity of $B(\cdot, \cdot, \cdot)$ (Lemma 4.1), we have

$$B(\tilde{\mathbf{w}}, \tilde{\mathbf{w}}, \boldsymbol{\phi}_h) - B(\mathbf{w}_h, \mathbf{w}_h, \boldsymbol{\phi}_h) \geq 0.$$

Thus, the left-hand side of (5.5) can be bounded from below as follows:

$$(5.6) \quad (\boldsymbol{\phi}_{h,t}, \boldsymbol{\phi}_h) + A(\boldsymbol{\phi}_h, \boldsymbol{\phi}_h) + [B(\tilde{\mathbf{w}}, \tilde{\mathbf{w}}, \boldsymbol{\phi}_h) - B(\mathbf{w}_h, \mathbf{w}_h, \boldsymbol{\phi}_h)] \geq \frac{1}{2} \frac{d}{dt} \|\boldsymbol{\phi}_h\|^2 + Re^{-1} \|\nabla \boldsymbol{\phi}_h\|^2.$$

We now start bounding from above the five terms on the right-hand side of (5.5). By using the Cauchy–Schwarz and Young’s inequalities, we obtain

$$(5.7) \quad -(\boldsymbol{\eta}_t, \boldsymbol{\phi}_h) \leq |(\boldsymbol{\eta}_t, \boldsymbol{\phi}_h)| \leq \frac{1}{2} \|\boldsymbol{\phi}_h\|^2 + \frac{1}{2} \|\boldsymbol{\eta}_t\|^2,$$

$$(5.8) \quad -A(\boldsymbol{\eta}, \boldsymbol{\phi}_h) \leq |A(\boldsymbol{\eta}, \boldsymbol{\phi}_h)| \leq Re^{-1} \varepsilon_1 \|\nabla \boldsymbol{\phi}_h\|^2 + \frac{Re^{-1}}{4\varepsilon_1} \|\nabla \boldsymbol{\eta}\|^2.$$

By adding and subtracting terms and using Lemma 4.2, we get

$$(5.9) \quad B(\mathbf{w}, \mathbf{w}, \boldsymbol{\phi}_h) - B(\tilde{\mathbf{w}}, \tilde{\mathbf{w}}, \boldsymbol{\phi}_h) \leq |B(\mathbf{w}, \mathbf{w}, \boldsymbol{\phi}_h) - B(\tilde{\mathbf{w}}, \tilde{\mathbf{w}}, \boldsymbol{\phi}_h)| \\ \leq |B(\mathbf{w}, \mathbf{w}, \boldsymbol{\phi}_h) - B(\tilde{\mathbf{w}}, \mathbf{w}, \boldsymbol{\phi}_h)| + |B(\tilde{\mathbf{w}}, \mathbf{w}, \boldsymbol{\phi}_h) - B(\tilde{\mathbf{w}}, \tilde{\mathbf{w}}, \boldsymbol{\phi}_h)| \\ \leq M_a \mu \delta^{\sigma+1} (\|\nabla \boldsymbol{\eta}\|_F \|\nabla \mathbf{w}\|_F, \|\nabla \boldsymbol{\phi}_h\|_F) \\ + M_a \mu \delta^{\sigma+1} (\|\nabla \tilde{\mathbf{w}}\|_F \|\nabla \boldsymbol{\eta}\|_F, \|\nabla \boldsymbol{\phi}_h\|_F).$$

Note that the stability estimates in [9] imply

$$(5.10) \quad \|\nabla \tilde{\mathbf{w}}\|_{L^\infty} \leq \tilde{C} \|\nabla \mathbf{w}\|_{L^\infty}.$$

Thus,

$$(5.11) \quad B(\mathbf{w}, \mathbf{w}, \boldsymbol{\phi}_h) - B(\tilde{\mathbf{w}}, \tilde{\mathbf{w}}, \boldsymbol{\phi}_h) \\ \leq M_a \mu \delta^{\sigma+1} \|\nabla \mathbf{w}\|_{L^\infty} \|\nabla \boldsymbol{\eta}\| \|\nabla \boldsymbol{\phi}_h\| + M_a \mu \delta^{\sigma+1} \tilde{C} \|\nabla \mathbf{w}\|_{L^\infty} \|\nabla \boldsymbol{\eta}\| \|\nabla \boldsymbol{\phi}_h\| \\ \leq C \left(M_a, \mu, \delta, \sigma, \tilde{C} \right) \frac{1}{4\varepsilon_2} \|\nabla \mathbf{w}\|_{L^\infty}^2 \|\nabla \boldsymbol{\eta}\|^2 + \varepsilon_2 \|\nabla \boldsymbol{\phi}_h\|^2.$$

By adding and subtracting terms, we obtain

$$\begin{aligned}
 & C(\mathbf{w}, \mathbf{w}, \phi_h) - C(\mathbf{w}_h, \mathbf{w}_h, \phi_h) \\
 & \leq |C(\mathbf{w}, \mathbf{w}, \phi_h) - C(\mathbf{w}, \mathbf{w}_h, \phi_h)| + |C(\mathbf{w}, \mathbf{w}_h, \phi_h) - C(\mathbf{w}_h, \mathbf{w}_h, \phi_h)| \\
 & = |C(\mathbf{w}, \boldsymbol{\eta} + \phi_h, \phi_h)| + |C(\boldsymbol{\eta} + \phi_h, \mathbf{w}_h, \phi_h)| \\
 & \quad \left(\text{since } C(\boldsymbol{\eta} + \phi_h, \tilde{\mathbf{w}} - \mathbf{w}_h, \phi_h) = C(\boldsymbol{\eta} + \phi_h, \phi_h, \phi_h) = 0 \right) \\
 & = |C(\mathbf{w}, \boldsymbol{\eta}, \phi_h)| + |C(\boldsymbol{\eta} + \phi_h, \tilde{\mathbf{w}}, \phi_h)| \\
 & = |(\mathbf{w} \cdot \nabla \boldsymbol{\eta}, \phi_h)| + |(\boldsymbol{\eta} \cdot \nabla \tilde{\mathbf{w}}, \phi_h)| + |(\phi_h \cdot \nabla \tilde{\mathbf{w}}, \phi_h)| \\
 & \leq \left(\frac{1}{2} \|\nabla \boldsymbol{\eta}\|^2 + \frac{1}{2} \|\mathbf{w}\|_{L^\infty}^2 \|\phi_h\|^2 \right) \\
 (5.12) \quad & + \left(\frac{1}{2} \|\boldsymbol{\eta}\|^2 + \frac{1}{2} \|\nabla \tilde{\mathbf{w}}\|_{L^\infty}^2 \|\phi_h\|^2 \right) + (\|\nabla \tilde{\mathbf{w}}\|_{L^\infty} \|\phi_h\|^2).
 \end{aligned}$$

By using the Cauchy–Schwarz and Young’s inequalities, we obtain

$$(5.13) \quad (q - \lambda_h, \nabla \cdot \phi_h) \leq |(q - \lambda_h, \nabla \cdot \phi_h)| \leq \varepsilon_3 \|\nabla \phi_h\|^2 + \frac{1}{4\varepsilon_3} \|q - \lambda_h\|^2.$$

Inserting estimates (5.6)–(5.13) into (5.5) and picking $\varepsilon_1 := 1/6$, $\varepsilon_2 := Re^{-1}/6$, $\varepsilon_3 := Re^{-1}/6$, we get

$$\begin{aligned}
 & \frac{1}{2} \frac{d}{dt} \|\phi_h\|^2 + \frac{Re^{-1}}{2} \|\nabla \phi_h\|^2 \\
 & \leq \left(\frac{1}{2} + \frac{6}{4} Re^{-1} + \frac{6}{4} Re C(M_a, \mu, \delta, \sigma, \tilde{C}) \|\nabla \mathbf{w}\|_{L^\infty}^2 \right) \|\nabla \boldsymbol{\eta}\|^2 \\
 & \quad + \frac{1}{2} \|\boldsymbol{\eta}_t\|^2 + \frac{1}{2} \|\boldsymbol{\eta}\|^2 + \frac{6}{4} Re \|q - \lambda_h\|^2 \\
 (5.14) \quad & + \left(\frac{1}{2} + \frac{1}{2} \|\mathbf{w}\|_{L^\infty}^2 + \frac{\tilde{C}^2}{2} \|\nabla \mathbf{w}\|_{L^\infty}^2 + \tilde{C} \|\nabla \mathbf{w}\|_{L^\infty} \right) \|\phi_h\|^2.
 \end{aligned}$$

In order to apply Gronwall’s lemma, we need

$$\begin{aligned}
 b(t) & := \left(\frac{1}{2} + \frac{1}{2} \|\mathbf{w}\|_{L^\infty}^2 + \frac{\tilde{C}^2}{2} \|\nabla \mathbf{w}\|_{L^\infty}^2 + \tilde{C} \|\nabla \mathbf{w}\|_{L^\infty} \right) \\
 (5.15) \quad & \in L^1(0, T).
 \end{aligned}$$

This follows immediately from the hypothesis ($\mathbf{w} \in L^4(0, T; \mathbf{W}^{1,\infty}(\Omega))$).

Hiding all constants in generic C ’s, Gronwall’s lemma now implies, for almost all $t \in [0, T]$, that

$$\begin{aligned}
 & \|\phi_h(\mathbf{x}, t)\|^2 + Re^{-1} \|\nabla \phi_h\|_{L^2(0,t;L^2)}^2 \\
 & \leq C \exp(\|b(t)\|_{L^1(0,t)}) \|\phi_h(\mathbf{x}, 0)\|^2 \\
 & \quad + C \exp(\|b(t)\|_{L^1(0,t)}) \left[\|\boldsymbol{\eta}_t\|_{L^2(0,t;L^2)}^2 + \|\boldsymbol{\eta}\|_{L^2(0,t;L^2)}^2 \right. \\
 & \quad + (1 + 3Re^{-1}) \|\nabla \boldsymbol{\eta}\|_{L^2(0,t;L^2)}^2 + C(M_a, \mu, \delta, \sigma, \tilde{C}) \int_0^t 3Re \|\nabla \mathbf{w}\|_{L^\infty}^2 \|\nabla \boldsymbol{\eta}\|^2 ds \\
 (5.16) \quad & \left. + 3Re \|q - \lambda_h\|^2 \right].
 \end{aligned}$$

By using the Cauchy–Schwarz inequality in $L^2(0, t), t \in [0, T]$, and the hypothesis ($\mathbf{w} \in L^4(0, T; \mathbf{W}^{1,\infty}(\Omega))$), we get

$$(5.17) \quad \int_0^t \|\nabla \mathbf{w}\|_{L^\infty}^2 \|\nabla \boldsymbol{\eta}\|^2 ds \leq \|\nabla \mathbf{w}\|_{L^4(0,t;L^\infty)}^2 \|\nabla \boldsymbol{\eta}\|_{L^4(0,t;L^2)}^2.$$

We apply now the essential supremum over $t \in [0, T]$ on both sides of inequality (5.16). By using the triangle inequality, the error estimate in the theorem now follows. \square

Remark 5.1. The error estimate in Theorem 5.1 is not uniform in Re , as the error estimate for the Smagorinsky model (Theorem 4.2 in [29]). The reason is that our bounded AV operator $B(\cdot, \cdot, \cdot)$ is just monotone (Lemma 4.1) and not strongly monotone as the Smagorinsky AV operator in [29].

Remark 5.2. The regularity of \mathbf{w} ($\mathbf{w} \in L^4(0, T; \mathbf{W}^{1,\infty}(\Omega))$) assumed in Theorem 5.1 is higher than the regularity of \mathbf{w} assumed in the error estimate for the Smagorinsky model in [29] ($\mathbf{w} \in L^2(0, T; \mathbf{W}^{1,\infty}(\Omega))$). Again, the reason is that our bounded AV operator $B(\cdot, \cdot, \cdot)$ is just monotone (Lemma 4.1), whereas the Smagorinsky AV operator is strongly monotone.

Remark 5.3. The multiplicative constant $C_1(\mathbf{w}) := \exp(\|b(t)\|_{L^1(0,T)})$ in the error estimate in Theorem 5.1 does not depend on Re .

COROLLARY 5.2. *Assuming $\mathbb{P}_k/\mathbb{P}_{k-1}$ velocity pressure discretization (i.e., choosing $l := k - 1$ in (3.6)), the order of convergence of*

$$(5.18) \quad \|\mathbf{w} - \mathbf{w}_h\|_{L^\infty(0,T;L^2)} \text{ is } O(h^k),$$

$$(5.19) \quad \|\mathbf{w} - \mathbf{w}_h\|_{L^2(0,T;H^1)} \text{ is } O(h^k).$$

Proof. The proof is an immediate consequence of Theorem 5.1 and the approximation properties of the interpolation (3.9)–(3.11). \square

6. Newton approximation scheme for the bounded AV model. In this section, we discuss the Newton approximation scheme as applied to the NSE with the bounded AV term (1.2). The analysis in this section is especially relevant to the numerical discretization of the vortex decay problem used in section 7.2.1. Note that approximate solutions $(\mathbf{u}_h, p_h) \in \mathbf{X}_h \times Q_h$ for the NSE with the bounded AV term (1.2) must satisfy a nonlinear system of ordinary differential equations. In this section, we derive the Newton approximation scheme for the semidiscrete variational problem, and note that, in practice, one would apply a Newton iteration at each time step for a fully discrete approximation.

It is a straightforward calculation to show that the Gâteaux derivative for the bounded AV term considered in this paper satisfies the following.

THEOREM 6.1. *Suppose that the function $a(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ is analytic, and define the (continuous) map $\mathcal{G}_1 : \mathbf{X}_h \rightarrow \mathbf{X}_h$ as*

$$\langle \mathcal{G}_1[\mathbf{u}], \mathbf{v} \rangle := (a(\|\nabla \mathbf{u}\|_F) \nabla \mathbf{u}, \nabla \mathbf{v}).$$

Then the Gâteaux derivative in the direction of \mathbf{u} evaluated at \mathbf{w} , denoted as $\mathcal{J}_\mathbf{u} \mathcal{G}_1[\mathbf{w}]$, is equal to

$$\langle \mathcal{J}_\mathbf{u} \mathcal{G}_1[\mathbf{w}], \mathbf{v} \rangle = (a(\|\nabla \mathbf{w}\|_F) \nabla \mathbf{w}, \nabla \mathbf{v}) + \left(\frac{a'(\|\nabla \mathbf{w}\|_F)}{\|\nabla \mathbf{w}\|_F} [\nabla \mathbf{u} : \nabla \mathbf{w}] \nabla \mathbf{u}, \nabla \mathbf{v} \right).$$

COROLLARY 6.2. *Suppose that the function $a(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ is analytic, and define the (continuous) map $\mathcal{G}_2 : \mathbf{X}_h \rightarrow \mathbf{X}_h$ as*

$$\langle \mathcal{G}_2[\mathbf{u}], \mathbf{v} \rangle := (a(\delta \|\nabla \mathbf{u}\|_F) \nabla \mathbf{u}, \nabla \mathbf{v}).$$

Then the Gâteaux derivative in the direction of \mathbf{u} evaluated at \mathbf{w} , denoted as $\mathcal{J}_{\mathbf{u}} \mathcal{G}_2[\mathbf{w}]$, is equal to

$$\langle \mathcal{J}_{\mathbf{u}} \mathcal{G}_2[\mathbf{w}], \mathbf{v} \rangle = (a(\|\delta \nabla \mathbf{u}\|_F) \nabla \mathbf{w}, \nabla \mathbf{v}) + \delta \left(\frac{a'(\|\delta \nabla \mathbf{u}\|_F)}{\|\nabla \mathbf{u}\|_F} [\nabla \mathbf{u} : \nabla \mathbf{w}] \nabla \mathbf{u}, \nabla \mathbf{v} \right).$$

We now consider the semidiscrete approximation $(\mathbf{u}_h, p_h) \in \mathbf{X}_h \times Q_h$, which solves the NSE with the bounded AV term (1.2). For notational simplicity, we drop the “ h ” subscripts from \mathbf{u} and p . Define the (continuous) map $\mathcal{G} : \mathbf{X}_h \times Q_h \rightarrow \mathbf{X}_h \times Q_h$ as

$$\begin{aligned} \langle \mathcal{G}[\mathbf{u}, p], (\mathbf{v}, q) \rangle &:= (\mathbf{u}_t, \mathbf{v}) + Re^{-1} (\nabla \mathbf{u}, \nabla \mathbf{v}) + \mu \delta^\sigma (a(\delta \|\nabla \mathbf{u}\|_F) \nabla \mathbf{u}, \nabla \mathbf{v}) \\ &\quad + (\mathbf{u} \cdot \nabla \mathbf{u}, \mathbf{v}) - (p, \nabla \cdot \mathbf{v}) + (q, \nabla \cdot \mathbf{u}) - (\mathbf{f}, \mathbf{v}). \end{aligned}$$

Therefore, we immediately have that the Gâteaux derivative of \mathcal{G} in the direction of (\mathbf{u}, p) , evaluated at (\mathbf{w}, r) , is given by

$$\begin{aligned} \langle \mathcal{J}_{(\mathbf{u}, p)} \mathcal{G}[\mathbf{w}, r], (\mathbf{v}, q) \rangle &= (\mathbf{w}_t, \mathbf{v}) + Re^{-1} (\nabla \mathbf{w}, \nabla \mathbf{v}) + \mu \delta^\sigma (a(\delta \|\nabla \mathbf{u}\|_F) \nabla \mathbf{w}, \nabla \mathbf{v}) \\ &\quad + \mu \delta^{\sigma+1} \left(\frac{a'(\delta \|\nabla \mathbf{u}\|_F)}{\|\nabla \mathbf{u}\|_F} [\nabla \mathbf{u} : \nabla \mathbf{w}] \nabla \mathbf{u}, \nabla \mathbf{v} \right) \\ &\quad + (\mathbf{u} \cdot \nabla \mathbf{w}, \mathbf{v}) + (\mathbf{w} \cdot \nabla \mathbf{u}, \mathbf{v}) \\ &\quad - (r, \nabla \cdot \mathbf{v}) + (q, \nabla \cdot \mathbf{w}) - (\mathbf{f}, \mathbf{v}). \end{aligned}$$

Now, substituting this formula for $\langle \mathcal{J}_{(\mathbf{u}, p)} \mathcal{G}[\mathbf{w}, r], (\mathbf{v}, q) \rangle$ into the Newton iteration system

$$\langle \mathcal{J}_{(\mathbf{u}^{(n-1)}, p^{(n-1)})} \mathcal{G}[\mathbf{u}^{(n)} - \mathbf{u}^{(n-1)}, p^{(n)} - p^{(n-1)}], (\mathbf{v}, q) \rangle = - \langle \mathcal{G}[\mathbf{u}^{(n-1)}, p^{(n-1)}], (\mathbf{v}, q) \rangle,$$

we obtain the Newton iteration scheme

$$\begin{aligned} &(\mathbf{u}_t^{(n)}, \mathbf{v}) + Re^{-1} (\nabla \mathbf{u}^{(n)}, \nabla \mathbf{v}) + \mu \delta^\sigma (a(\delta \|\nabla \mathbf{u}^{(n-1)}\|_F) \nabla \mathbf{u}^{(n)}, \nabla \mathbf{v}) \\ (6.1) \quad &+ \mu \delta^{\sigma+1} \left(\frac{a'(\delta \|\nabla \mathbf{u}^{(n-1)}\|_F)}{\|\nabla \mathbf{u}^{(n-1)}\|_F} [\nabla \mathbf{u}^{(n-1)} : \nabla \mathbf{u}^{(n)}] \nabla \mathbf{u}^{(n-1)}, \nabla \mathbf{v} \right) \\ &+ (\mathbf{u}^{(n-1)} \cdot \nabla \mathbf{u}^{(n)}, \mathbf{v}) + (\mathbf{u}^{(n)} \cdot \nabla \mathbf{u}^{(n-1)}, \mathbf{v}) - (p^{(n)}, \nabla \cdot \mathbf{v}) + (q, \nabla \cdot \mathbf{u}^{(n)}) \\ &= (\mathbf{f}, \mathbf{v}) + \mu \delta^{\sigma+1} (a'(\delta \|\nabla \mathbf{u}^{(n-1)}\|_F) \|\nabla \mathbf{u}^{(n-1)}\|_F \nabla \mathbf{u}^{(n-1)}, \nabla \mathbf{v}) \\ &+ (\mathbf{u}^{(n-1)} \cdot \nabla \mathbf{u}^{(n-1)}, \mathbf{v}) \end{aligned}$$

$\forall (\mathbf{v}, q) \in \mathbf{X}_h \times Q_h$.

Remark 6.1. Given an initial $(\mathbf{u}^{(0)}, p^{(0)}) \in \mathbf{X}_h \times Q_h$, the existence of a solution $(\mathbf{u}^{(n)}, p^{(n)}) \in \mathbf{X}_h \times Q_h \forall n = 1, 2, \dots$ to (6.1) follows by using the same arguments as in section 4 (Lemmas 4.6 and 4.7). Since the system (6.1), however, is linear, we can easily prove (by using the properties of $a(\cdot)$) that the solution $(\mathbf{u}^{(n)}, p^{(n)})$ is also unique.

7. Numerical results. In this section, we present numerical results that illustrate the benefits of the bounded AV model (2.12)–(2.14). We begin with a numerical simulation of turbulent flow in a 3D square duct, comparing the bounded AV model and Smagorinsky model with a direct numerical simulation (DNS) of the flow (section 7.1). This is followed with a careful mesh-refinement study that supports the theoretical error estimates shown above (section 7.2).

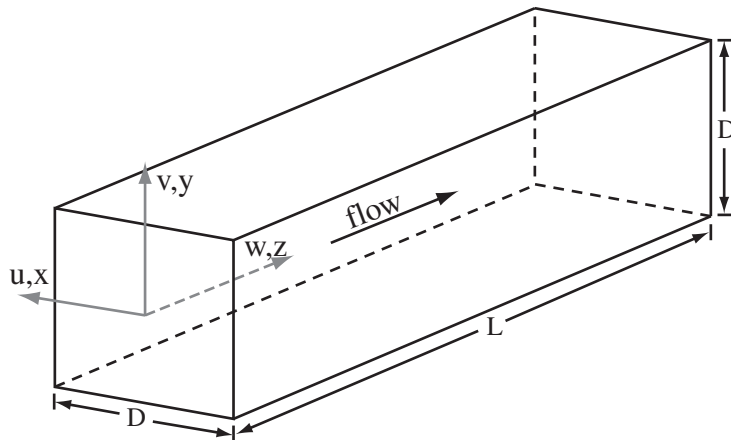


FIG. 7.1. Computational domain for the turbulent flow in a 3D square duct.

7.1. Turbulent flow in a 3D square duct. To demonstrate the improvement of the bounded AV model over the Smagorinsky model for a wall-bounded flow, we consider the numerical simulation of a turbulent flow in a 3D square duct. This test problem is often employed in the validation of LES models, e.g., [18, 24, 25, 37]. For our demonstration, we use the same parameters used in [24, 25] for their DNS simulation.

We performed a posteriori testing of the LES models, comparing coarser discretizations of the bounded AV model and the Smagorinsky model to a DNS in the *same geometry*, with the *same level of discretization*, and the *same Reynolds number* (based on the bulk velocity) as in [24]. In particular, we used the same geometry, illustrated in Figure 7.1, where $\Omega = (-D/2, D/2) \times (-D/2, D/2) \times (0, L) \subset \mathbb{R}^3$, with $D = 1$ and $L = 6.4$. The grid point distributions in the x - and y -directions also follow [24]:

$$(7.1) \quad x_j = -\frac{1}{2} + \frac{1}{2}(b-1) \left[\frac{a^{2j/N_x} - 1}{a^{(2j/N_x - 1)} + 1} \right], \quad j = 0, 1, \dots, N_x - 1,$$

where $a = \frac{b+1}{b-1}$, N_x is the number of grid points in the x -direction, and $b = 1.1$ is a stretching parameter. We use a similar distribution for the y -direction. In the z -direction, we used a uniform grid point distribution

$$(7.2) \quad z_j = j \frac{L}{N_z - 1}, \quad j = 0, 1, \dots, N_z,$$

where N_z is number of grid points in the z -direction.

- For the benchmark DNS, we used similar discretization levels as used for the DNS in [24]:

$$N_x \times N_y \times N_z = 97 \times 97 \times 145$$

yielding 3.6 million degrees of freedom. This is about 50% finer in the z -direction.

- For the two LES simulations (bounded AV and Smagorinsky), we used the coarser discretization:

$$N_x \times N_y \times N_z = 25 \times 25 \times 97$$

yielding 0.14 million degrees of freedom.

Note that the LES runs were performed on meshes that employed roughly 4 times fewer nodes in the wall normal directions. This is typical in the a posteriori testing of LES models (e.g., LES of a turbulent channel flow [26]).

For all three numerical simulations (DNS, Smagorinsky, and bounded AV models), we used the following boundary conditions: no-slip ($\mathbf{u} = 0$) on the walls, stress-free, “do-nothing” boundary conditions at the outlet ($z = 6.4$)

$$(7.3) \quad (-p\mathbf{I} + 2 Re^{-1} \nabla^s \mathbf{u}) \cdot \mathbf{n} = 0,$$

and Dirichlet boundary conditions at the inlet. The inflow profile is

$$(7.4) \quad u(x, y, z = 0, t) = 0,$$

$$(7.5) \quad v(x, y, z = 0, t) = 0,$$

$$(7.6) \quad w(x, y, z = 0, t) = C_1 \frac{2}{\pi} \arctan \left[C_2 \left(\frac{1}{2} - \max\{|x|, |y|\} \right) \right],$$

where $C_2 = 100$ and $C_1 = \frac{\pi}{2 \arctan(C_2/2)} \approx 1.01$. The constant C_2 controls the steepness of the inflow profile near the walls, whereas the constant C_1 was chosen to yield a centerline streamwise velocity of 1.

The initial conditions were obtained from the inflow conditions $\mathbf{u}(x, y, z, 0) = \mathbf{u}(x, y, z = 0, t)$ for $(x, y, z) \subset \Omega$. The initial pressure was obtained by substituting the initial conditions into the momentum equations, integrating in z , and using (7.3) at $(x = 0, y = 0)$ to determine the integration constant.

The inflow velocity (7.4)–(7.6) was used to compute the bulk velocity u_B :

$$(7.7) \quad u_B = \frac{1}{\text{area}(\Omega_{cs})} \int_{\Omega_{cs}} u(x, y, z, t) dV,$$

where Ω_{cs} is the cross-section of the computational domain at the inlet. Formulas (7.4)–(7.6) and the values of C_1 and C_2 immediately yield the bulk velocity

$$(7.8) \quad u_B \approx 0.9.$$

We define the Reynolds number based on the bulk velocity in the usual way [24, 25]:

$$(7.9) \quad Re_B := \frac{u_B D}{\nu},$$

where ν is the kinematic viscosity. Thus, in order to achieve the Re_B used in [24]

$$(7.10) \quad Re_B = 10,320,$$

we used the following value for the kinematic viscosity ν :

$$(7.11) \quad \nu = \frac{u_B D}{Re_B} \approx 8.7207 \times 10^{-5}.$$

A time dependent forcing function in the z -direction

$$f_z(x, y, z = 0, t) = \begin{cases} 0.1 \sin(20t) \cos(10\pi x) \cos(10\pi y), & 0 < t < 1, \\ 0, & t \geq 1, \end{cases}$$

was used to induce the flow transition to turbulence.

The flow was integrated in time up to the final time $T = 14.5$ with a time step $\Delta T = 0.02$, which is similar to that used in [24]. Since the centerline streamwise velocity has magnitude 1, this allows a particle in the center to travel the entire length of the 3D square duct 2.25 times. Each DNS run was performed on 48 nodes of SystemX (www.arc.vt.edu) for two CPU weeks. The high computational cost associated with the DNS runs has restricted the final time to the value presented above. The numerical results presented below represent instantaneous results and not time and spatial averages. We emphasize, however, that the qualitative behavior of the numerical simulations is representative for the entire duration of the numerical simulation and not only for the particular instantaneous time frame that we present.

The numerical simulations in this section have been carried out with the ViTLES (see Appendix A for a detailed description of the algorithm and computational implementation). We used a small penalty parameter $\varepsilon = 10^{-4}$ in the penalty method to compute the pressure. The nonlinear system at each time step was solved with a Newton iteration up to a Euclidean norm of the residual vector less than 10^{-8} .

For the Smagorinsky model, we chose a filter radius $\delta = 0.01$ and the Smagorinsky constant $c_s = 0.17$, which is a popular choice in the literature. For the bounded AV model, we chose the parameters $\mu \sim 0.4$ and $\sigma = 1$. The function $a(\cdot)$ was chosen to resemble that in Figure 1.1:

$$(7.12) \quad a(\delta \|\nabla^s \mathbf{u}\|_F) := -0.02 + \frac{1}{1 + 49 e^{-5.7\delta \|\nabla^s \mathbf{u}\|_F}}.$$

Remark 7.1. This is exactly the function that was chosen in [28] for a *completely different setting* (a rotating pulse for the convection-diffusion equation). Therefore, this choice is *not optimal* for our present setting (the 3D square duct turbulent flow). Further testing is needed to find the optimal values. We thus emphasize that we compare the Smagorinsky model with tuned parameters with the bounded AV model with nonoptimized parameters.

Figure 7.2 presents numerical approximations for the pressure (p) for the following cases (with the given pressure ranges): (i) the DNS ($p \in [-.003, .124]$); (ii) the Smagorinsky model (1.1) ($p \in [-.018, .710]$); and (iii) the bounded AV model (1.2) ($p \in [-.005, .160]$). Note that the pressure loss for the bounded AV model is more consistent with the DNS. The larger pressure loss observed in the Smagorinsky model is indicative of the overly diffusive nature of this model. A larger pressure drop is required to overcome the diffusion and maintain the inflow velocity.

Remark 7.2. The results obtained in the numerical simulation of turbulent flow in a 3D square duct are encouraging. They do, however, represent just a first step in a thorough validation of the bounded AV model. Testing in other settings, such as turbulent channel flow simulations [4], is needed. Furthermore, improved versions of the Smagorinsky model (such as the use of the van Driest damping (2.9)) should be used in the comparison with the bounded AV model. In this case, however, optimized parameters in the bounded AV model (such as the function $a(\cdot)$) should be employed to allow for a fair comparison with the optimized version of the Smagorinsky model.

7.2. Mesh-refinement study. In this section, we present a careful mesh-refinement study supporting the error estimates in Theorem 5.1 for the bounded AV model (1.2).

7.2.1. The 2D vortex decay problem. In this section, we present numerical results for the bounded AV model applied to the vortex decay problem of Chorin [7, 49]. A similar study for the vortex decay problem using the Smagorinsky model was

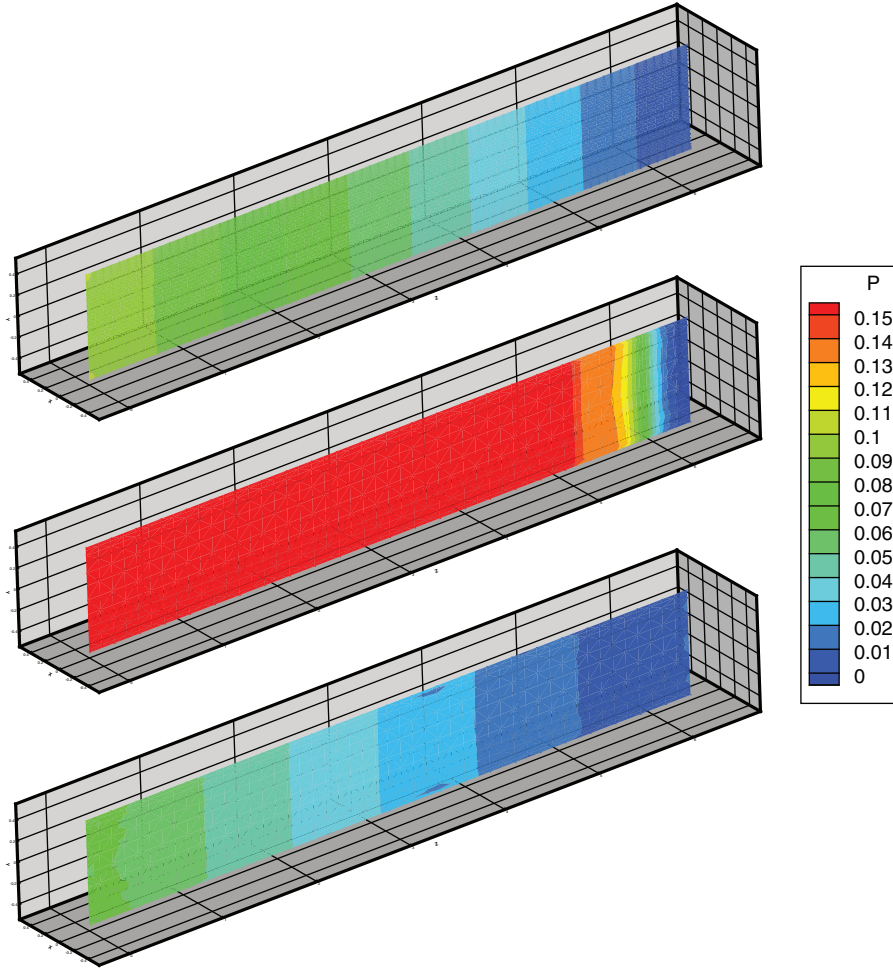


FIG. 7.2. 3D square duct flow, vertical cross-section. The pressure distribution (top to bottom): DNS (first), Smagorinsky (second), and bounded AV (third).

presented in [29]. For the vortex decay problem, we define the domain $\Omega = (0, 1)^2 \subset \mathbb{R}^2$ and specify

$$\begin{aligned}
 (7.13) \quad u_1 &:= -\cos(n\pi x) \sin(n\pi y) \exp(-2n^2\pi^2 t/\tau), \\
 u_2 &:= \sin(n\pi x) \cos(n\pi y) \exp(-2n^2\pi^2 t/\tau), \\
 p &:= -\frac{1}{4}(\cos(2n\pi x) + \cos(2n\pi y)) \exp(-4n^2\pi^2 t/\tau).
 \end{aligned}$$

Note that for $\tau := Re^{-1}$, the set (u_1, u_2, p) solves the time-dependent NSE with the appropriate (time-dependent, Dirichlet) boundary conditions. For our purposes, we take (7.13) as the solution to (2.12)–(2.14) and illustrate two points: that the spatial semidiscretization error estimates (5.18)–(5.19) in Corollary 5.2 are satisfied and that the estimates are bounded uniformly with respect to the Reynolds number.

TABLE 7.1
Finite element convergence estimates for the vortex decay problem.

h	$\ \mathbf{w} - \mathbf{w}_h\ _{L^\infty(0,T;L^2)}$	rate	$\ \mathbf{w} - \mathbf{w}_h\ _{L^2(0,T;L^2)}$	rate	$\ \mathbf{w} - \mathbf{w}_h\ _{L^2(0,T;H^1)}$	rate
1/8	$4.020927 \cdot 10^{-1}$		$4.936365 \cdot 10^{-1}$		$1.560773 \cdot 10^1$	
1/16	$3.103567 \cdot 10^{-2}$	3.70	$3.952673 \cdot 10^{-2}$	3.64	$2.886625 \cdot 10^0$	2.43
1/24	$5.534371 \cdot 10^{-3}$	4.25	$7.594030 \cdot 10^{-3}$	4.07	$1.096755 \cdot 10^0$	2.39
1/32	$1.822532 \cdot 10^{-3}$	3.86	$2.418206 \cdot 10^{-3}$	3.98	$5.182457 \cdot 10^{-1}$	2.61
1/40	$7.778230 \cdot 10^{-4}$	3.82	$1.018835 \cdot 10^{-3}$	3.87	$2.870239 \cdot 10^{-1}$	2.65
1/48	$4.227375 \cdot 10^{-4}$	3.34	$5.138958 \cdot 10^{-4}$	3.75	$1.763596 \cdot 10^{-1}$	2.68
1/56	$2.567581 \cdot 10^{-4}$	3.26	$2.907396 \cdot 10^{-4}$	3.70	$1.166515 \cdot 10^{-1}$	2.68
1/64	$1.645877 \cdot 10^{-4}$	3.33	$1.779283 \cdot 10^{-4}$	3.68	$8.151033 \cdot 10^{-2}$	2.68
1/72	$1.102856 \cdot 10^{-4}$	3.40	$1.154125 \cdot 10^{-4}$	3.68	$5.940966 \cdot 10^{-2}$	2.69

Accordingly, we specify the following parameters:

$$\begin{aligned}
 Re &:= 10^{10}, \\
 \tau &:= 1000, \\
 \text{final time } T &:= 2, \\
 \text{filter radius } \delta &:= 0.1, \\
 \mu = c_s^2 &:= 0.17^2, \\
 \Delta t &:= 0.01.
 \end{aligned}$$

For our calculations, we assume $n = 3$, i.e., a 3×3 array of vortices and study the finite element convergence rates for fixed $\delta := 0.1$ as $h \rightarrow 0$. For the spatial discretization, we take the Taylor–Hood finite element pair and implement the Newton iteration scheme as described in section 6. For the temporal discretization, we use the Crank–Nicolson scheme. As indicated in Table 7.1, the spatial semidiscretization errors are of order 3 in the spatial L^2 norm and 2 in the spatial H^1 norm. Thus, for this test problem, we obtain superconvergence in both norms (see Corollary 5.2). Also, note that these estimates are *independent* of the selected Reynolds number, as we have taken a relatively high value for Re and a relatively high value for Δt .

Appendix A. The Virginia Tech large eddy simulator (ViTLES). In this paper, we have used *ViTLES*, a *parallel, finite element* computational platform for the numerical validation of CFD and LES models. In what follows, we briefly describe the algorithm around which ViTLES is developed. More details can be found at <http://icam.vt.edu/ViTLES>.

For spatial discretization, the computational domain is decomposed in a collection of nonoverlapping triangles (in 2D) or tetrahedra (in 3D). We employ the traditional Taylor–Hood finite element pair (continuous quadratic velocities and continuous linear pressures), which satisfies the discrete inf-sup (LBB_h) condition [6].

For time discretization, we employ the second-order accurate, unconditionally stable Crank–Nicolson scheme [10].

We also employ the penalty method [10, 22, 42], in which the incompressibility constraint $\nabla \cdot \mathbf{u} = 0$ in the NSE (2.1)–(2.3) is relaxed by setting

$$(A.1) \quad \varepsilon p_\varepsilon + \nabla \cdot \mathbf{u}_\varepsilon = 0,$$

where ε is a small parameter. In our computations, we used $\varepsilon = 0.0001$. As $\varepsilon \rightarrow 0$, the solution of the penalized problem converges to that of the unpenalized problem [22].

We employ a Newton iteration scheme for solving the nonlinear system at each time step. The scheme implemented in ViTLES explicitly constructs finite difference approximations of the Jacobians rather than calculating the actual Jacobian matrix.

ViTLES is written on top of PETSc (the portable, extensible toolkit for scientific computing) developed at Argonne National Laboratory [3]. ViTLES makes use of the message passing interface (MPI), the LINPACK library, and the basic linear algebra subprograms (BLAS) library. We have also used ADIC, the automatic differentiation tool [23], to compute Jacobians. Tecplot (<http://www.tecplot.com/>) was used to visualize the numerical results.

Acknowledgments. We thank the two anonymous referees whose comments and suggestions have greatly improved the paper. We also acknowledge computer time on System X (www.arc.vt.edu/arc/SystemX) and a cluster obtained through the NSF scientific computing research environments for the mathematical sciences (SCREMS) program (NSF grant DMS-0322852).

REFERENCES

- [1] R.A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
- [2] J.C. ANDRÉ, G.D. MOOR, P. LACARRERE, AND R.D. VACHAT, *Turbulence approximations for inhomogeneous flows. Part I: The clipping approximation*, J. Atmospheric Sci., 33 (1976), pp. 476–481.
- [3] S. BALAY, K. BUSCHELMAN, V. EIJKHOUT, W.D. GROPP, D. KAUSHIK, M.G. KNEPLEY, L. CURFMAN MCINNES, B.F. SMITH, AND H. ZHANG, *PETSc Users Manual*, Technical report ANL-95/11 - Revision 2.1.5, Argonne National Laboratory, Argonne, IL, 2004.
- [4] L.C. BERSELLI, T. ILIESCU, AND W.J. LAYTON, *Mathematics of Large Eddy Simulation of Turbulent Flows*, Springer-Verlag, New York, 2005.
- [5] J. BOUSSINESQ, *Essai sur la théorie des eaux courantes*, Mém. prés par div. savants à la Acad. Sci., 23 (1877), pp. 1–680.
- [6] S.C. BRENNER AND L.R. SCOTT, *The Mathematical Theory of Finite Element Methods*, Texts Appl. Math. 15, Springer-Verlag, New York, 1994.
- [7] A.J. CHORIN, *Numerical solution of the Navier-Stokes equations*, Math. Comp., 22 (1968), pp. 745–762.
- [8] G.-H. COTTET AND A.A. WRAY, *Anisotropic grid-based formulas for subgrid-scale models*, Annual Research Briefs, Center for Turbulence Research, Stanford University and NASA Ames, Stanford, CA, 1997, pp. 113–122.
- [9] M. CROUZEIX AND V. THOMÉE, *The stability in L_p and W_p^1 of the L_2 -projection onto finite element function spaces*, Math. Comp., 48 (1987), pp. 521–532.
- [10] C. CUVELIER, A. SEGAL, AND A.A. VAN STEENHOVEN, *Finite Element Methods and Navier–Stokes Equations*, Math. Appl. 22, D. Reidel, Dordrecht, The Netherlands, 1986.
- [11] A. DÖRNBRACK, *Turbulent mixing by breaking gravity waves*, J. Fluid Mech., 375 (1998), pp. 113–141.
- [12] Q. DU AND M.D. GUNZBURGER, *Finite-element approximations of a Ladyzhenskaya model for stationary incompressible viscous flow*, SIAM J. Numer. Anal., 27 (1990), pp. 1–19.
- [13] Q. DU AND M.D. GUNZBURGER, *Analysis of a Ladyzhenskaya model for incompressible viscous flow*, J. Math. Anal. Appl., 155 (1991), pp. 21–45.
- [14] A. ERN AND J.-L. GUERMOND, *Theory and Practice of Finite Elements*, Appl. Math. Sci. 159, Springer-Verlag, New York, 2004.
- [15] H.J.S. FERNANDO, *Aspects of stratified turbulence*, in Developments in Geophysical Turbulence, R.M. Kerr and Y. Kimura, eds., Kluwer, Norwell, MA, 2000, pp. 81–92.
- [16] G.P. GALDI, J.G. HEYWOOD, AND R. RANNACHER (EDS.), *Fundamental directions in mathematical fluid mechanics*, in Advances in Mathematical Fluid Mechanics, Birkhäuser-Verlag, Basel, 2000.
- [17] G.P. GALDI, *An Introduction to the Mathematical Theory of the Navier–Stokes Equations*, Vol. I: Linearized Steady Problems, Springer Tracts in Natural Philosophy 38, Springer-Verlag, New York, 1994.
- [18] S. GAVRILAKIS, *Numerical simulation of low-Reynolds-number turbulent flow through a straight square duct*, J. Fluid Mech., 244 (1992), pp. 101–129.

- [19] M. GERMANO, U. PIOMELLI, P. MOIN, AND W.H. CABOT, *A dynamic subgrid-scale eddy viscosity model*, Phys. Fluids A, 3 (1991), pp. 1760–1765.
- [20] S. GHOSAL, T.S. LUND, P. MOIN, AND K. AKSELVOLL, *A dynamic localization model for large-eddy simulation of turbulent flows*, J. Fluid Mech., 286 (1995), pp. 229–255.
- [21] V. GIRAULT AND P.-A. RAVIART, *Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms*, Springer Ser. Comput. Math. 5, Springer-Verlag, Berlin, 1986.
- [22] M.D. GUNZBURGER, *Finite Element Methods for Viscous Incompressible Flows: A Guide to Theory, Practice, and Algorithms*, Computer Science and Scientific Computing, Academic Press, Boston, MA, 1989.
- [23] P. HOVLAND, B. NORRIS, AND C. BISCHOF, *ADIC Web Page*, <http://www-fp.mcs.anl.gov/adic/>.
- [24] A. HUSER AND S. BIRINGEN, *Direct numerical simulation of turbulent flow in a square duct*, J. Fluid Mech., 257 (1993), pp. 65–95.
- [25] A. HUSER, S. BIRINGEN, AND F.F. HATAY, *Direct simulation of turbulent flow in a square duct: Reynolds-stress budgets*, Phys. Fluids, 6 (1994), pp. 3144–3152.
- [26] T. ILIESCU AND P.F. FISCHER, *Large eddy simulation of turbulent channel flows by the rational LES model*, Phys. Fluids, 15 (2003), pp. 3036–3047.
- [27] T. ILIESCU AND P.F. FISCHER, *Backscatter in the Rational LES model*, Comput. & Fluids, 35 (2004), pp. 783–790.
- [28] T. ILIESCU, *Genuinely nonlinear models for convection-dominated problems*, Comput. Math. Appl., 48 (2004), pp. 1677–1692.
- [29] V. JOHN AND W.J. LAYTON, *Analysis of numerical errors in large eddy simulation*, SIAM J. Numer. Anal., 40 (2002), pp. 995–1020.
- [30] V. JOHN, *Large Eddy Simulation of Turbulent Incompressible Flows: Analytical and Numerical Results for a Class of LES Models*, Lect. Notes Comput. Sci. Eng. 34, Springer-Verlag, Berlin, 2004.
- [31] N.V. KORNEV, I.V. TKATCHENKO, AND E. HASSEL, *A simple clipping procedure for the dynamic mixed model based on Taylor series approximation*, Comm. Numer. Methods Engrg., 22 (2006), pp. 55–61.
- [32] O.A. LADYŽHENSKAYA, *New equations for the description of motion of viscous incompressible fluids and solvability in the large of boundary value problems for them*, Proc. Steklov Inst. Math., 102 (1967), pp. 95–118.
- [33] O.A. LADYŽHENSKAYA, *Modifications of the Navier-Stokes equations for large gradients of the velocities*, Zap. Naučn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI), 7 (1968), pp. 126–154.
- [34] W.J. LAYTON, *A nonlinear, subgridscale model for incompressible viscous flow problems*, SIAM J. Sci. Comput., 17 (1996), pp. 347–357.
- [35] J. LERAY AND J. SCHAUDER, *Topologie et équations fonctionnelles*, Ann. Sci. École Norm. Sup., 51 (1934), pp. 45–78.
- [36] D.K. LILLY, *The representation of small scale turbulence in numerical simulation experiments*, in Proceedings of the IBM Scientific Computing Symposium on Environmental Sciences, H.H. Goldstine, ed., Yorktown Heights, NY, 1967, pp. 195–210.
- [37] R.K. MADABHUSHI AND S.P. VANKA, *Large eddy simulation of turbulence-driven secondary flow in a square duct*, Phys. Fluids, 3 (1991), pp. 2734–2745.
- [38] P.J. MASON, *Large-eddy simulation of the convective atmospheric boundary layer*, J. Atmospheric Sci., 46 (1989), pp. 1492–1516.
- [39] C. MENEVEAU, T. LUND, AND W. CABOT, *A Lagrangian dynamic subgrid-scale model of turbulence*, J. Fluid Mech., 319 (1996), pp. 353–385.
- [40] J.W. MILES, *On the stability of heterogeneous shear flows*, J. Fluid Mech., 10 (1961), pp. 496–508.
- [41] F. PORTÉ-AGEL, C. MENEVEAU, AND M.B. PARLANGE, *A scale-dependent dynamic model for large-eddy simulation: Application to a neutral atmospheric boundary layer*, J. Fluid Mech., 415 (2000), pp. 261–284.
- [42] A. QUARTERONI AND A. VALLI, *Numerical Approximation of Partial Differential Equations*, Springer Ser. Comput. Math. 23, Springer-Verlag, Berlin, 1994.
- [43] L.F. RICHARDSON, *Weather Prediction by Numerical Process*, Cambridge University Press, Cambridge, 1922.
- [44] J.J. ROHR, E.C. ITSWEIRE, K.N. HELLAND, AND C.W. VAN ATTA, *Growth and decay of turbulence in stably stratified shear flow*, J. Fluid Mech., 195 (1988), pp. 77–111.
- [45] P. SAGAUT, *Large eddy simulation for incompressible flows: An introduction*, third ed., Scientific Computation, Springer-Verlag, Berlin, 2006 (in French).
- [46] U. SCHUMANN, *Subgrid scale model for finite difference simulations of turbulent flows in plane channels and annuli*, J. Comput. Phys., 18 (1975), pp. 376–404.

- [47] J.S. SMAGORINSKY, *General circulation experiments with the primitive equations*, Monthly Weather Rev., 91 (1963), pp. 99–164.
- [48] B. STEVENS, C.-H. MOENG, AND P.P. SULLIVAN, *Large-eddy simulations of radiatively driven convection: Sensitivities to the representation of small scales*, J. Atmospheric Sci., 56 (1999), pp. 3963–3984.
- [49] D. TAFTI, *Comparison of some upwind-biased high-order formulations with a second-order central-difference scheme for time integration of the incompressible Navier-Stokes equations*, Comput. & Fluids, 25 (1996), pp. 647–665.
- [50] E.R. VAN DRIEST, *On turbulent flow near a wall*, J. Aerosp. Sci., 23 (1956), pp. 1007–1011.
- [51] J. VON NEUMANN AND R.D. RICHTMYER, *A method for the numerical calculation of hydrodynamic shocks*, J. Appl. Phys., 21 (1950), pp. 232–237.
- [52] Y. ZANG, R.L. STREET, AND J.R. KOSEFF, *A dynamic mixed subgrid-scale model and its application to turbulent recirculating flows*, Phys. Fluids A, 5 (1993), pp. 3186–3196.

SUPERCONVERGENCE OF SOME PROJECTION APPROXIMATIONS FOR WEAKLY SINGULAR INTEGRAL EQUATIONS USING GENERAL GRIDS*

ANDREY AMOSOV[†], MARIO AHUES[‡], AND ALAIN LARGILLIER[‡]

Abstract. This paper deals with superconvergence phenomena in general grids when projection-based approximations are used for solving Fredholm integral equations of the second kind with weakly singular kernels. Four variants of the Galerkin method are considered. They are the classical Galerkin method, the iterated Galerkin method, the Kantorovich method, and the iterated Kantorovich method. It is proved that the iterated Kantorovich approximation exhibits the best superconvergence rate if the right-hand side of the integral equation is nonsmooth. All error estimates are derived for an arbitrary grid without any uniformity or quasi-uniformity condition on it, and are formulated in terms of the data without any additional assumption on the solution. Numerical examples concern the equation governing transfer of photons in stellar atmospheres. The numerical results illustrate the fact that the error estimates proposed in the different theorems are quite sharp, and confirm the superiority of the iterated Kantorovich scheme.

Key words. projections, grids, superconvergence, weak singularity

AMS subject classifications. 65R20, 65B99, 65J10

DOI. 10.1137/070685464

1. Introduction. It is well known that some variants of the Galerkin approximation can be superconvergent in the sense that convergence to the solution is faster than the convergence to it of its own projection.

We recall that in a Hilbert space H , if π^h is a family of orthogonal projections pointwise convergent to the identity operator I and if T is a compact operator for which 1 is not an eigenvalue, then the operator $T^h := \pi^h T$ gives the Galerkin approximation φ^h to $\varphi = T\varphi + f$ as $\varphi^h = (I - T^h)^{-1} \pi^h f$ for all $f \in H$, where the inverses $(I - T^h)^{-1}$ exist and are uniformly bounded in the operator norm; i.e., $\|(I - T^h)^{-1}\| \leq c$ for some constant c independent of h , for sufficiently small h . The identity

$$\varphi^h - \varphi = (I - T^h)^{-1} \pi^h (I - T)\varphi - (I - T^h)^{-1} (I - T^h)\varphi = (I - T^h)^{-1} (\pi^h - I)\varphi$$

shows that

$$\|(I - \pi^h)\varphi\| \leq \|\varphi - \varphi^h\| \leq c \|(I - \pi^h)\varphi\|,$$

so the Galerkin approximation φ^h converges to the solution φ as slow as the projection of φ onto the range of π^h converges to φ .

For integral equations the phenomenon of superconvergence of some variants of the Galerkin approximation has been studied since 1976 (cf. [16], [17], [18], [11], [12], [4], [19], [6], [23], [20], [15], [24], [9], and references therein).

*Received by the editors March 16, 2007; accepted for publication (in revised form) September 12, 2008; published electronically January 16, 2009. This work was also partially supported by project 04-01-00539 of the RFBR.

<http://www.siam.org/journals/sinum/47-1/68546.html>

[†]Department of Mathematical Modelling, Moscow Power Engineering Institute (Technical University), Krasnokazarmennaja str. 14, 111250 Moscow, Russia (AmosovAA@mpei.ru).

[‡]Laboratoire de Mathématiques de l'Université de Saint-Etienne (LaMUSE), EA 3989, 23 rue du Dr. Paul Michelon, F-42023 Saint-Étienne, France (mario.ahues@univ-st-etienne.fr, largillier@univ-st-etienne.fr).

In this paper we study superconvergence results on three other projection methods for solving weakly singular integral equations of the form

$$(1) \quad \varphi(\tau) = \varpi_0 \int_0^{\tau_*} \mathcal{E}(|\tau - \tau'|)\varphi(\tau') d\tau' + f(\tau), \quad \tau \in J :=]0, \tau_*[,$$

where φ is the unknown.

The phenomenon of superconvergence is studied here following the properties of the Banach space X , which may be $L^p(J)$, $C(\overline{J})$, $BV(\overline{J})$, or $W^{1,p}(J)$, in which the equation is settled. In all the spaces, the discretization procedure is the weakest one we can conceive in the Lebesgue space L^1 , namely, projection techniques based on piecewise constant functions. We are particularly interested in the case in which ϖ_0 is close to 1 from below, and τ_* is much bigger than 1. For instance, in Astrophysics, typical values may be $\varpi_0 = 0.9999$ and $\tau_* = 10^6$.

2. Description of the methods and formulation of some results. We assume that the kernel \mathcal{E} is a positive decreasing function defined on $\mathbb{R}^+ :=]0, +\infty[$, such that $\mathcal{E}(0^+) = +\infty$ (that is why (1) is called singular), $\mathcal{E} \in L^r(\mathbb{R})$ for all $r \in [1, \infty[$ (that is why (1) is only weakly singular), and

$$(2) \quad \|\mathcal{E}\|_{L^1(\mathbb{R}^+)} \leq 1/2.$$

We suppose that ϖ_0 is a fixed real number in $]0, 1[$, and that $f \in L^p(J)$ for some $p \in [1, \infty]$.

The heat transfer integral equation

$$(3) \quad \varphi(\tau) = \frac{\varpi_0}{2} \int_0^{\tau_*} E_1(|\tau - \tau'|)\varphi(\tau') d\tau' + f(\tau), \quad \tau \in J,$$

widely used for mathematical modelling in Astrophysics (cf. [3], [5], [10], and [22]), is an important example of (1), where

$$E_1(\tau) := \int_0^1 \mu^{-1} e^{-\tau/\mu} d\mu, \quad \tau > 0,$$

is the exponential-integral function of order 1. Usually, in Astrophysics, τ_* , called the optical depth, is a very large number, and ϖ_0 , called the albedo, may be very close to 1. So it is important to find accurate estimates of the error which are independent of τ_* but depend on ϖ_0 in an explicit form.

Let us define the integral operator

$$(\Lambda\varphi)(\tau) := \int_0^{\tau_*} \mathcal{E}(|\tau - \tau'|)\varphi(\tau') d\tau', \quad \varphi \in X, \quad \tau \in [0, \tau_*],$$

where $X \subseteq L^1(J)$ is a suitable Banach space, and rewrite (1) as follows.

For a given $f \in X$, find $\varphi \in X$ such that

$$(4) \quad \varphi = \varpi_0 \Lambda\varphi + f.$$

Different choices of X will be made precise throughout the paper.

For integers p and q such that $p \leq q$, we denote by $\llbracket p, q \rrbracket$ the set of all integers j such that $p \leq j \leq q$.

In order to compute approximate solutions to (4), we consider the following numerical process:

Let $J^h := \{\tau_j : j \in \llbracket 0, n \rrbracket\}$ be a finite set of $n + 1$ points in $[0, \tau_*]$ such that

$$0 = \tau_0 < \tau_1 < \dots < \tau_{n-1} < \tau_n = \tau_*.$$

Set

$$(5) \quad h_i := \tau_i - \tau_{i-1}, \quad i \in \llbracket 1, n \rrbracket, \quad h_{\max} := \max_{i \in \llbracket 1, n \rrbracket} h_i, \quad h := (h_1, h_2, \dots, h_n),$$

$$(6) \quad J_{i-1/2} := [\tau_{i-1}, \tau_i[, \quad i \in \llbracket 1, n \rrbracket.$$

Let $S_{1/2}^h(J)$ be the following space of piecewise constant functions:

$$S_{1/2}^h(J) := \{f : f(\tau) = f_{i-1/2} \quad \text{for all } i \in \llbracket 1, n \rrbracket, \text{ and all } \tau \in J_{i-1/2}\}.$$

Let $\pi^h : L^p(J) \rightarrow L^p(J)$ be the bounded projection defined, for all $\varphi \in L^p(J)$, as follows:

For each $i \in \llbracket 1, n \rrbracket$ and all $\tau \in J_{i-1/2}$,

$$(7) \quad (\pi^h \varphi)(\tau) := \frac{1}{h_i} \int_{\tau_{i-1}}^{\tau_i} \varphi(\tau') d\tau'.$$

It is clear that the range of π^h is equal to $S_{1/2}^h(J)$.

We shall deal with four different projection approximations based on π^h :

The first one is the well-known classical Galerkin approximation φ^h which solves

$$(8) \quad \varphi^h = \varpi_0 \pi^h \Lambda \varphi^h + \pi^h f.$$

We observe that $\varphi^h \in S_{1/2}^h(J)$. Since $\pi^h \Lambda$ is a finite rank operator, φ^h may be computed with the help of the solution of an auxiliary system of linear algebraic equations (cf. [2]). Moreover, in this particular case, the coefficient matrix of the linear system is self-adjoint and positive definite with respect to the inner product

$$\langle \varphi^h, \psi^h \rangle = \sum_{i=1}^n \varphi_{i-1/2}^h \psi_{i-1/2}^h h_i$$

defined in $S_{1/2}^h(J) \times S_{1/2}^h(J)$.

The second approximation is the iterated Galerkin approximation $\overline{\varphi}^h$ [21], also called Sloan approximation, which solves

$$(9) \quad \overline{\varphi}^h = \varpi_0 \Lambda \pi^h \overline{\varphi}^h + f.$$

$\overline{\varphi}^h$ may be computed in the following way: Apply π^h to (9) and note that $\pi^h \overline{\varphi}^h = \varphi^h$, the Galerkin approximation. Hence

$$\overline{\varphi}^h = \varpi_0 \Lambda \varphi^h + f.$$

It is well known (cf. [16], [19], [4], [6], [7], [23]) that the iterated Galerkin method has a higher order of convergence than the Galerkin method itself.

The third one, called Kantorovich approximation, is based on *Kantorovich regularization* (cf. [8]) of (4):

$$(10) \quad y = \varpi_0 \Lambda y + \Lambda f.$$

The Kantorovich approximation is defined to be the solution of

$$(11) \quad \tilde{\varphi}^h = \varpi_0 \pi^h \Lambda \tilde{\varphi}^h + f.$$

To solve (11), set $y := \Lambda \varphi$, $y^h := \pi^h \Lambda \tilde{\varphi}^h$ and note that

$$\varphi = \varpi_0 y + f, \quad \tilde{\varphi}^h = \varpi_0 y^h + f,$$

and that y solves (10) while y^h solves

$$y^h = \varpi_0 \pi^h \Lambda y^h + \pi^h \Lambda f.$$

So $y^h \in S_{1/2}^h(J)$ is the Galerkin approximation to (10).

This method was considered in [11], [12], [13], [15], [20]. Since it approximates $y = \Lambda \varphi$, this method is appropriate if $\Lambda \varphi$ is more smooth than φ .

The fourth approximation is the iterated Kantorovich approximation (cf. [20], [14]) defined to be the solution of

$$(12) \quad \hat{\varphi}^h = \varpi_0 \Lambda \pi^h \hat{\varphi}^h + f + \varpi_0 \Lambda (I - \pi^h) f.$$

As it is quoted in [20], to compute $\hat{\varphi}^h$ we may find the Sloan approximation \bar{y}^h to (10),

$$\bar{y}^h = \varpi_0 \Lambda \pi^h \bar{y}^h + \Lambda f,$$

and then use the fact that $\hat{\varphi}^h = \varpi_0 \bar{y}^h + f$.

Remark 1. Note that $\hat{\varphi}^h - \varphi = \varpi_0 (y^h - y)$ and $\tilde{\varphi}^h - \varphi = \varpi_0 (\bar{y}^h - y)$. So for estimating the errors of Kantorovich and iterated Kantorovich approximations, we may apply the estimates we have obtained for the error of Galerkin and Sloan approximations but to (10) with right-hand side Λf instead of f and with solution $y = \Lambda \varphi$ instead of φ .

Remark 2. We recall that the classical Galerkin approximation $\varphi^h \in S_{1/2}^h(J)$ is defined by reducing to zero the projection of the residual onto the finite dimensional approximating subspace $S_{1/2}^h(J)$, the range of π^h :

$$\pi^h (\varphi^h - \varpi_0 \pi^h \Lambda \varphi^h - f) = 0.$$

Another Galerkin-type approximation ψ^h can be built as the solution of

$$\psi^h = \varpi_0 \pi^h \Lambda \pi^h \psi^h + f.$$

In this case, f is kept as it is and no projection acts on it. Obviously, $\varphi^h = \pi^h \psi^h$, but in general, φ^h and ψ^h do not coincide. However, if f belongs to the range of π^h , i.e., if $\pi^h f = f$, then not only $\psi^h = \varphi^h$ but $\tilde{\varphi}^h = \varphi^h$ too, and hence the iterated Galerkin approximation and the iterated Kantorovich approximation coincide: $\hat{\varphi}^h = \tilde{\varphi}^h$.

Throughout this paper, p , q , and s are real numbers in $[1, \infty]$ such that $p \leq q$, and

$$(13) \quad 1/s = 1 - 1/p + 1/q.$$

Note that $s \in [1, \infty]$, $s = 1$ if $p = q$ and that $s = \infty$ means that $p = 1$ and $q = \infty$.

Let

$$\gamma_0 := \frac{1}{1 - \varpi_0}, \quad \gamma_1 := \frac{\varpi_0}{1 - \varpi_0}, \quad M_p := 2^{1/p} \|\mathcal{E}\|_{L^p(\mathbb{R}^+)}.$$

Let $C^0(\overline{J})$ denote the space of all continuous functions on \overline{J} , $BV(\overline{J})$ the space of functions with bounded variation, and $W^{1,p}(J)$ the classical Sobolev space, with respective norms

$$\begin{aligned} \|f\|_{C^0} &:= \max_{\tau \in \overline{J}} |f(\tau)|, & \|f\|_{BV} &:= \operatorname{var} f + \gamma_1 \sup_{\tau \in \overline{J}} |f(\tau)|, \\ \|f\|_{W^{1,p}} &:= \|Df\|_{L^p} + 2^{1-1/p} M_p \gamma_1 \|f\|_{C^0}. \end{aligned}$$

Here and later $Df = f'$.

Let us formulate the error estimates which are proved in this paper in the case of the heat transfer equation (3).

These errors will be denoted by

$$\varepsilon^h := \varphi^h - \varphi, \quad \overline{\varepsilon}^h := \overline{\varphi}^h - \varphi, \quad \tilde{\varepsilon}^h := \tilde{\varphi}^h - \varphi, \quad \widehat{\varepsilon}^h := \widehat{\varphi}^h - \varphi.$$

Remark 3. All error estimates are derived for an arbitrary grid without any uniformity or quasi-uniformity condition on it, and are formulated in the terms of the data without any additional assumption on the solution. All the error bounds are independent of τ_* and depend on ϖ_0 through γ_0 and γ_1 .

THEOREM 1. *Assume that $\mathcal{E} = \frac{1}{2}E_1$, $f \in L^p(J)$, and $1 \leq p \leq q \leq \infty$. The following estimates hold as $h_{\max} \rightarrow 0$:*

$$\begin{aligned} \|\pi^h \varepsilon^h\|_{L^q} &\leq \gamma_0 \gamma_1 h_{\max}^{1-1/p+1/q} \ln \frac{1}{h_{\max}} (1 + o(1)) \|f\|_{L^p} \quad \text{for } s < \infty, \\ \|\overline{\varepsilon}^h\|_{L^q} &\leq \gamma_0 \gamma_1 h_{\max}^{1-1/p+1/q} \ln \frac{1}{h_{\max}} (1 + o(1)) \|f\|_{L^p} \quad \text{for } s < \infty, \\ \|\pi^h \tilde{\varepsilon}^h\|_{L^q} &\leq \gamma_1^2 h_{\max}^{2-1/p+1/q} \ln^2 \frac{1}{h_{\max}} (1 + o(1)) \|f\|_{L^p}, \\ \|\widehat{\varepsilon}^h\|_{L^q} &\leq \gamma_0 \gamma_1 h_{\max}^{1-1/p+1/q} \ln \frac{1}{h_{\max}} (1 + o(1)) \|f\|_{L^p} \quad \text{for } s < \infty, \\ \|\widehat{\varepsilon}^h\|_{L^q} &\leq \gamma_1^2 h_{\max}^{2-1/p+1/q} \ln^2 \frac{1}{h_{\max}} (1 + o(1)) \|f\|_{L^p}. \end{aligned}$$

As a corollary

$$\begin{aligned} \|\pi^h \varepsilon^h\|_{L^p} &\leq \gamma_0 \gamma_1 h_{\max} \ln \frac{1}{h_{\max}} (1 + o(1)) \|f\|_{L^p}, \\ \|\overline{\varepsilon}^h\|_{L^p} &\leq \gamma_0 \gamma_1 h_{\max} \ln \frac{1}{h_{\max}} (1 + o(1)) \|f\|_{L^p}, \\ \|\pi^h \tilde{\varepsilon}^h\|_{L^p} &\leq \gamma_1^2 h_{\max}^2 \ln^2 \frac{1}{h_{\max}} (1 + o(1)) \|f\|_{L^p}, \\ \|\widehat{\varepsilon}^h\|_{L^p} &\leq \gamma_0 \gamma_1 h_{\max} \ln \frac{1}{h_{\max}} (1 + o(1)) \|f\|_{L^p}, \\ \|\widehat{\varepsilon}^h\|_{L^p} &\leq \gamma_1^2 h_{\max}^2 \ln^2 \frac{1}{h_{\max}} (1 + o(1)) \|f\|_{L^p}. \end{aligned}$$

As we see, when $f \in L^p(J)$, then the iterated Galerkin method and the Kantorovich method have the same accuracy. The iterated Kantorovich method is appreciably more accurate.

THEOREM 2. Assume that $\mathcal{E} = \frac{1}{2}E_1$, $f \in BV(\overline{J})$, and $1 \leq q \leq \infty$. The following estimates hold as $h_{\max} \rightarrow 0$:

$$\begin{aligned} \|\pi^h \varepsilon^h\|_{L^q} &\leq \gamma_0 \gamma_1 h_{\max}^{1+1/q} \ln \frac{1}{h_{\max}} (1 + o(1)) \|f\|_{BV}, \\ \|\varepsilon^h\|_{L^q} &\leq \gamma_0^2 h_{\max}^{1/q} \|f\|_{BV}, \\ \|\overline{\varepsilon}^h\|_{L^q} &\leq \gamma_0 \gamma_1 h_{\max}^{1+1/q} \ln \frac{1}{h_{\max}} (1 + o(1)) \|f\|_{BV}, \\ \|\pi^h \widehat{\varepsilon}^h\|_{L^q} &\leq \gamma_1^2 h_{\max}^2 \ln \frac{1}{h_{\max}} (1 + o(1)) \|\Lambda f\|_{W^{1,q}} \quad \text{for } q < \infty, \\ \|\widehat{\varepsilon}^h\|_{L^q} &\leq \gamma_0 \gamma_1 h_{\max} \|\Lambda f\|_{W^{1,q}} \quad \text{for } q < \infty, \\ \|\widehat{\varepsilon}^h\|_{L^q} &\leq \gamma_1^2 h_{\max}^2 \ln \frac{1}{h_{\max}} (1 + o(1)) \|\Lambda f\|_{W^{1,q}} \quad \text{for } q < \infty, \end{aligned}$$

The following estimates hold too:

$$\begin{aligned} \|\pi^h \varepsilon^h\|_{BV} &\leq 3\gamma_0 \gamma_1 h_{\max} \ln \frac{1}{h_{\max}} (1 + o(1)) \|f\|_{BV}, \\ \|\overline{\varepsilon}^h\|_{W^{1,1}} &\leq 3\gamma_0 \gamma_1 h_{\max} \ln \frac{1}{h_{\max}} (1 + o(1)) \|f\|_{BV}, \\ \|\pi^h \widehat{\varepsilon}^h\|_{BV} &\leq 3\gamma_0 \gamma_1 h_{\max} \ln \frac{1}{h_{\max}} (1 + o(1)) \|f\|_{BV}, \\ \|\widehat{\varepsilon}^h\|_{W^{1,1}} &\leq 3\gamma_0 \gamma_1 h_{\max} \ln \frac{1}{h_{\max}} (1 + o(1)) \|f\|_{BV}. \end{aligned}$$

Theorem 2 shows that, when $f \in BV(\overline{J})$, then the iterated Galerkin method is more accurate than the Kantorovich method. The iterated Kantorovich method keeps an advantage.

THEOREM 3. Assume that $\mathcal{E} = \frac{1}{2}E_1$, $f \in W^{1,p}(J)$ for some $p \in [1, \infty[$, and $p \leq q \leq \infty$. The following estimates hold as $h_{\max} \rightarrow 0$:

$$\begin{aligned} \|\pi^h \varepsilon^h\|_{L^q} &\leq \gamma_0 \gamma_1 h_{\max}^{2-1/p+1/q} \ln \frac{1}{h_{\max}} (1 + o(1)) \|f\|_{W^{1,p}}, \\ \|\varepsilon^h\|_{L^q} &\leq \gamma_0^2 h_{\max}^{1-1/p+1/q} \|f\|_{W^{1,p}}, \\ \|\overline{\varepsilon}^h\|_{L^q} &\leq \gamma_0 \gamma_1 h_{\max}^{2-1/p+1/q} \ln \frac{1}{h_{\max}} (1 + o(1)) \|f\|_{W^{1,p}}. \end{aligned}$$

As a corollary

$$\begin{aligned} \|\pi^h \varepsilon^h\|_{L^p} &\leq \gamma_0 \gamma_1 h_{\max} \ln \frac{1}{h_{\max}} (1 + o(1)) \|f\|_{W^{1,p}}, \\ \|\varepsilon^h\|_{L^p} &\leq \gamma_0^2 h_{\max} \|f\|_{W^{1,p}}, \\ \|\overline{\varepsilon}^h\|_{L^p} &\leq \gamma_0 \gamma_1 h_{\max}^2 \ln \frac{1}{h_{\max}} (1 + o(1)) \|f\|_{W^{1,p}}. \end{aligned}$$

Theorems 2 and 3 show us that, when $f \in W^{1,p}(J)$, $p \in [1, \infty[$, then the classical Galerkin method has almost the same accuracy as the Kantorovich method, and the iterated Galerkin method has almost the same accuracy as the iterated Kantorovich method.

Theorems 1–3 follow from Theorem 11 and Theorem 14 because E_1 satisfies the following properties:

1. $E_1 \in W^{1,r}(\delta, \infty)$ for all $\delta > 0$ and $r \in [1, \infty[$,
2. $DE_1(t) = o(t^{-1}E_1(t))$ as $t \rightarrow 0^+$,
3. $E_1(t) = \ln \frac{1}{t}(1 + o(1))$ as $t \rightarrow 0^+$ (cf. [1]).

Remark 4. It is proved in section 3 that, if $f \in BV(\overline{J})$, then $\Lambda f \in BV(\overline{J})$, $\Lambda f \in W^{1,q}(J)$ for all $q \in [1, \infty[$, and

$$\begin{aligned} \|\Lambda f\|_{BV(\overline{J})} &\leq \operatorname{var}_{\overline{J}} f + \gamma_0 \sup_{\tau \in \overline{J}} |f(\tau)|, \\ \|\Lambda f\|_{W^{1,q}} &\leq M_q \left(\operatorname{var}_{\overline{J}} f + 2^{1-1/q} \gamma_0 \sup_{\tau \in \overline{J}} |f(\tau)| \right). \end{aligned}$$

It is proved also that, if $f \in W^{1,p}(J)$ for some $p \in [1, \infty[$, then $\Lambda f \in W^{1,q}(J)$ for all $q \in [p, \infty[$, and

$$\|\Lambda f\|_{W^{1,q}} \leq M_s \|Df\|_{L^p} + 2^{1-1/q} \gamma_0 M_q \|f\|_{C^0(\overline{J})}.$$

Remark 5. The Kantorovich approach has no sense if $f \in W^{1,\infty}(J)$ is such that $f(0) \neq 0$ or $f(\tau_*) \neq 0$. It follows from formula (21) that in that case $\Lambda f \notin W^{1,\infty}(J)$, and, as a corollary, $\Lambda \varphi \notin W^{1,\infty}(J)$. So the iterated Galerkin method should be preferred if f belongs to $W^{1,\infty}(J)$ or if it is smoother than that.

3. Some properties of the solution. We use the notations

$$\begin{aligned} (\varphi, \psi) &:= \int_0^{\tau_*} \varphi(\tau) \psi(\tau) d\tau, \\ \mathcal{E}_*(\tau) &:= \mathcal{E}(\tau_* - \tau). \end{aligned}$$

Let $f \in L^p(J)$, and $\mathcal{E} \in L^r(\mathbb{R}^+)$ for all $r \in [1, \infty[$.

In this section, h denotes a strictly positive real number. We introduce differences

$$\Delta_h f(\tau) := f(\tau + h) - f(\tau), \quad \Delta_{-h} f(\tau) := f(\tau - h) - f(\tau),$$

and the modulus of continuity of the kernel \mathcal{E} :

$$\omega_r(\mathcal{E}, h) := \sup_{0 < \delta \leq h} \|\Delta_\delta \mathcal{E}(\cdot)\|_{L^r(\mathbb{R})},$$

that is to say

$$\omega_r(\mathcal{E}, h) = \sup_{0 < \delta \leq h} \left(\int_{\mathbb{R}} |\mathcal{E}(|\tau + \delta|) - \mathcal{E}(|\tau|)|^r d\tau \right)^{1/r}.$$

LEMMA 1. *Let $s \neq \infty$. Then Λ is a bounded linear operator from $L^p(J)$ into $L^q(J)$, and the following estimates hold:*

(14) $\|\Lambda\|_{L^p \rightarrow L^q} \leq M_s,$

(15) $\|\Delta_h \Lambda\|_{L^p \rightarrow L^q} \leq \omega_s(\mathcal{E}, h).$

Proof. For $p \in [1, \infty[$ and by Hölder inequality,

$$\begin{aligned} |\Lambda f(\tau)|^p &\leq \left(\int_0^{\tau_*} \mathcal{E}(|\tau - \tau'|) |f(\tau')| d\tau' \right)^p \\ &= \left(\int_0^{\tau_*} \mathcal{E}(|\tau - \tau'|)^{s(1-1/p)} \mathcal{E}(|\tau - \tau'|)^{s/q} |f(\tau')| d\tau' \right)^p \\ &\leq \left[\int_0^{\tau_*} \mathcal{E}(|\tau - \tau'|)^s d\tau' \right]^{p-1} \int_0^{\tau_*} \mathcal{E}(|\tau - \tau'|)^{sp/q} |f(\tau')|^p d\tau' \\ &\leq M_s^{s(p-1)} \int_0^{\tau_*} \mathcal{E}(|\tau - \tau'|)^{sp/q} |f(\tau')|^p d\tau', \end{aligned}$$

and similarly

$$\begin{aligned} |\Delta_h \Lambda f(\tau)|^p &= \left| \int_0^{\tau_*} \Delta_h \mathcal{E}(|\tau - \tau'|) f(\tau') d\tau' \right|^p \\ &\leq \omega_s(\mathcal{E}, h)^{s(p-1)} \int_0^{\tau_*} |\Delta_h \mathcal{E}(|\tau - \tau'|)|^{sp/q} |f(\tau')|^p d\tau'. \end{aligned}$$

If $q = \infty$ these estimates immediately imply inequalities (14), (15).

If $q < \infty$, then by generalized Minkowski inequality we get

$$\begin{aligned} \|\Lambda f\|_{L^q}^p &\leq M_s^{s(p-1)} \left\| \int_0^{\tau_*} \mathcal{E}(|\cdot - \tau'|)^{sp/q} |f(\tau')|^p d\tau' \right\|_{L^{q/p}} \\ &\leq M_s^{s(p-1)} \int_0^{\tau_*} \|\mathcal{E}(|\cdot - \tau'|)^{sp/q}\|_{L^{q/p}} |f(\tau')|^p d\tau' \\ &= M_s^{s(p-1)} \int_0^{\tau_*} \|\mathcal{E}(|\cdot - \tau'|)\|_{L^s}^{sp/q} |f(\tau')|^p d\tau' \leq M_s^p \|f\|_{L^p}^p, \end{aligned}$$

and

$$\begin{aligned} \|\Delta_h \Lambda f\|_{L^q}^p &\leq \omega_s(\mathcal{E}, h)^{s(p-1)} \left\| \int_0^{\tau_*} |\Delta_h \mathcal{E}(|\cdot - \tau'|)|^{sp/q} |f(\tau')|^p d\tau' \right\|_{L^{q/p}} \\ &\leq \omega_s(\mathcal{E}, h)^{s(p-1)} \int_0^{\tau_*} \|\Delta_h \mathcal{E}(|\cdot - \tau'|)^{sp/q}\|_{L^{q/p}} |f(\tau')|^p d\tau' \\ &= \omega_s(\mathcal{E}, h)^{s(p-1)} \int_0^{\tau_*} \|\Delta_h \mathcal{E}(|\cdot - \tau'|)\|_{L^s}^{sp/q} |f(\tau')|^p d\tau' \leq \omega_s(\mathcal{E}, h)^p \|f\|_{L^p}^p. \end{aligned}$$

If $p = q = \infty$, then

$$|\Lambda f(\tau)| \leq \int_0^{\tau_*} \mathcal{E}(|\tau - \tau'|) d\tau' \|f\|_{L^\infty} \leq M_1 \|f\|_{L^\infty}, \quad |\Delta_h \Lambda f(\tau)| \leq \omega_1(\mathcal{E}, h) \|f\|_{L^\infty}.$$

The proof is complete. \square

COROLLARY 1. *The following estimate holds for all $p \in [1, \infty[$:*

$$(16) \quad \|\Lambda\|_{L^p \rightarrow L^p} \leq M_1 = 1.$$

COROLLARY 2. *If $1 < p \leq \infty$, then Λ is a bounded linear operator from $L^p(J)$ into $C^0(\overline{J})$.*

THEOREM 4. *If $f \in L^p(J)$ for some $p \in [1, \infty]$, then there exists a unique solution $\varphi \in L^p(J)$ of (1) and*

$$(17) \quad \|\varphi\|_{L^p} \leq \gamma_0 \|f\|_{L^p}.$$

If $f \in C^0(\bar{J})$, then $\varphi \in C^0(\bar{J})$ and

$$(18) \quad \|\varphi\|_{C^0} \leq \gamma_0 \|f\|_{C^0}.$$

Proof. It follows from (16) and the Banach Fixed Point Theorem. \square

LEMMA 2. *Assume that $\varphi, y \in L^1(J)$ and $h^{-1}\Delta_h\varphi \rightarrow y$ in $L^1_{loc}(J)$ as $h \rightarrow 0$. Then $\varphi \in W^{1,1}(J)$ and $D\varphi = y$.*

Proof. It is easy to see that, if h is a sufficiently small number, then

$$(h^{-1}\Delta_h\varphi, \psi) = (\varphi, h^{-1}\Delta_{-h}\psi) \quad \text{for all } \psi \in C_0^\infty(\bar{J}).$$

After a limit transition we get

$$(y, \psi) = -(\varphi, D\psi) \quad \text{for all } \psi \in C_0^\infty(\bar{J}).$$

This means that $y = D\varphi$. \square

THEOREM 5. *If $f \in W^{1,p}(J)$ for some $p \in [1, \infty[$, then $\varphi \in W^{1,p}(J)$ and*

$$(19) \quad D\varphi = \varpi_0 \Lambda D\varphi + \varpi_0 \varphi(0) \mathcal{E} - \varpi_0 \varphi(\tau_*) \mathcal{E}_* + Df \text{ in } J,$$

$$(20) \quad \|D\varphi\|_{L^p} \leq \gamma_0 \|f\|_{W^{1,p}}.$$

Proof. Let $h \in]0, \tau_*[$ and $\tau \in J_h :=]0, \tau_* - h[$. It follows from (1) that

$$\begin{aligned} \Delta_h\varphi(\tau) &= \varpi_0 \int_0^{\tau_*} \Delta_h \mathcal{E}(|\tau - \tau'|) \varphi(\tau') d\tau' + \Delta_h f(\tau) \\ &= \varpi_0 \int_0^{\tau_*-h} \mathcal{E}(|\tau - \tau'|) \Delta_h \varphi(\tau') d\tau' + \varpi_0 \int_0^h \mathcal{E}(|\tau + h - \tau'|) \varphi(\tau') d\tau' \\ &\quad - \varpi_0 \int_{\tau_*-h}^{\tau_*} \mathcal{E}(|\tau - \tau'|) \varphi(\tau') d\tau' + \Delta_h f(\tau). \end{aligned}$$

Hence the function $y^h := h^{-1}\Delta_h\varphi$ solves

$$y^h(\tau) = \varpi_0 \int_0^{\tau_*-h} \mathcal{E}(|\tau - \tau'|) y^h(\tau') d\tau' + \psi^h(\tau) \quad \text{for all } \tau \in J_h,$$

where

$$\begin{aligned} \psi^h(\tau) &:= \varpi_0 h^{-1} \int_0^h \mathcal{E}(|\tau + h - \tau'|) \varphi(\tau') d\tau' - \varpi_0 h^{-1} \int_{\tau_*-h}^{\tau_*} \mathcal{E}(|\tau - \tau'|) \varphi(\tau') d\tau' + h^{-1} \Delta_h f(\tau). \end{aligned}$$

Note that

$$\begin{aligned} &\|h^{-1} \int_0^h \mathcal{E}(|\cdot + h - \tau'|) \varphi(\tau') d\tau' - \varphi(0) \mathcal{E}(\cdot)\|_{L^1} \\ &\leq h^{-1} \int_0^h \|\mathcal{E}(|\cdot + h - \tau'|)\|_{L^1} |\varphi(\tau') - \varphi(0)| d\tau' \\ &\quad + h^{-1} \int_0^h \|\mathcal{E}(|\cdot + h - \tau'|) - \mathcal{E}(|\cdot|)\|_{L^1} d\tau' |\varphi(0)| \\ &\leq h^{-1} \int_0^h |\varphi(\tau') - \varphi(0)| d\tau' + \omega_1(\mathcal{E}, h) |\varphi(0)|, \end{aligned}$$

which tends to 0 as $h \rightarrow 0$.

Similarly,

$$\left\| h^{-1} \int_{\tau_*-h}^{\tau_*} \mathcal{E}(|\cdot - \tau'|) \varphi(\tau') d\tau' - \varphi(\tau_*) \mathcal{E}_*(\cdot) \right\|_{L^1} \rightarrow 0 \quad \text{as } h \rightarrow 0.$$

But $Df \in L^p(J)$ and hence $\|h^{-1} \Delta_h f - Df\|_{L^p(J_h)} \rightarrow 0$ as $h \rightarrow 0$. Thus, $\|\psi^h - \psi\|_{L^1(J_h)} \rightarrow 0$, where

$$\psi := \varpi_0 \varphi(0) \mathcal{E} - \varpi_0 \varphi(\tau_*) \mathcal{E}_* + Df \in L^p(J).$$

Let $y \in L^p(J)$ be a solution of the equation

$$y(\tau) = \varpi_0 \int_0^{\tau_*} \mathcal{E}(|\tau - \tau'|) y(\tau') d\tau' + \psi(\tau), \quad \tau \in J.$$

Applying Theorem 3 to this equation we get

$$\|y\|_{L^p} \leq \gamma_0 \|\psi\|_{L^p} \leq 2\gamma_1 \|\mathcal{E}\|_{L^p} \|\varphi\|_{C^0} + \gamma_0 \|Df\|_{L^p} \leq \gamma_0 \|f\|_{W^{1,p}}.$$

Considering $z^h := y^h - y$ as a solution of equation

$$z^h(\tau) = \varpi_0 \int_0^{\tau_*-h} \mathcal{E}(|\tau - \tau'|) z^h(\tau') d\tau' + g^h(\tau), \quad \tau \in J_h,$$

where

$$g^h(\tau) := \psi^h(\tau) - \psi(\tau) + \varpi_0 \int_{\tau_*-h}^{\tau_*} \mathcal{E}(|\tau - \tau'|) y(\tau') d\tau',$$

we get the estimate

$$\|y^h - y\|_{L^1(J_h)} \leq \gamma_0 \left[\|\psi^h - \psi\|_{L^1(J_h)} + \varpi_0 \int_{\tau_*-h}^{\tau_*} \|\mathcal{E}(|\cdot - \tau'|)\|_{L^1} |y(\tau')| d\tau' \right],$$

which tends to 0 as $h \rightarrow 0$. So $h^{-1} \varphi_h \rightarrow y$ in $L^1_{\text{loc}}(J)$ and, from Lemma 2, $y = D\varphi$. \square

LEMMA 3. If $f \in W^{1,p}(J)$ for some $p \in [1, \infty[$, then $\Lambda f \in W^{1,q}(J)$ for all $q \in [p, \infty[$, and

$$(21) \quad D\Lambda f = \Lambda Df + f(0) \mathcal{E} - f(\tau_*) \mathcal{E}_*,$$

$$(22) \quad \|D\Lambda f\|_{L^q} \leq M_s \|Df\|_{L^p} + 2^{1-1/q} M_q \|f\|_{C^0}.$$

Proof. As in the proof of Theorem 5

$$\begin{aligned} h^{-1} \Delta_h \Lambda f(\tau) &= \int_0^{\tau_*-h} \mathcal{E}(|\tau - \tau'|) h^{-1} \Delta_h f(\tau') d\tau' + h^{-1} \int_0^h \mathcal{E}(|\tau + h - \tau'|) f(\tau') d\tau' \\ &- h^{-1} \int_{\tau_*-h}^{\tau_*} \mathcal{E}(|\tau - \tau'|) f(\tau') d\tau' \rightarrow \int_0^{\tau_*} \mathcal{E}(|\tau - \tau'|) Df(\tau') d\tau' + f(0) \mathcal{E}(\tau) - f(\tau_*) \mathcal{E}_*(\tau) \end{aligned}$$

in $L^1_{\text{loc}}(J)$. So (21) is true. Inequality (22) follows from (21). \square

LEMMA 4. If $f \in BV(\overline{J})$, then $\Lambda f \in W^{1,p}(J)$ for all $p \in [1, \infty[$ and

$$(23) \quad \|D\Lambda f\|_{L^p} \leq M_p \left(\text{var}_{\overline{J}} f + 2^{1-1/p} \sup_{\tau \in \overline{J}} |f(\tau)| \right).$$

Proof. Put $f(\tau) := f(\tau_*)$ for $\tau > \tau_*$. Consider the average function

$$(24) \quad f_\delta(\tau) := \delta^{-1} \int_0^\delta f(\tau + \tau') d\tau'.$$

It is well known that $f_\delta \in W^{1,\infty}(J)$ and that, for all $p \in [1, \infty[$, $f_\delta \rightarrow f$ in $L^p(J)$ as $\delta \rightarrow 0$. Also,

$$(25) \quad \|Df_\delta\|_{L^1} \leq \text{var}_{\bar{J}} f, \quad \|f_\delta\|_{C^0} \leq \sup_{\tau \in \bar{J}} |f(\tau)|.$$

It follows from Lemma 3 that

$$\|D\Lambda f_\delta\|_{L^p} \leq M_p \left(\text{var}_{\bar{J}} f + 2^{1-1/p} \sup_{\tau \in \bar{J}} |f(\tau)| \right) \quad \text{for all } p \in [1, \infty[.$$

If $p \in]1, \infty[$ this inequality implies that there exists $D\Lambda f \in L^p(J)$ and inequality (23) holds. For $p = 1$ the inequality may be justified by a limiting process as $p \rightarrow 1$. \square

THEOREM 6. *Let $f \in BV(\bar{J})$. Then there exists a function $\varphi \in BV(\bar{J})$ satisfying (4) for all $\tau \in \bar{J}$ and*

$$(26) \quad \text{var}_{\bar{J}} \varphi \leq \gamma_0 \|f\|_{BV},$$

$$(27) \quad \sup_{\tau \in \bar{J}} |\varphi(\tau)| \leq \gamma_0 \sup_{\tau \in \bar{J}} |f(\tau)|.$$

Proof. As $f \in BV(\bar{J}) \subset L^\infty(J)$, there exists a solution $\varphi \in L^\infty(J)$ for which $\Lambda\varphi \in C^0(\bar{J})$ and (4) is satisfied almost everywhere. Putting $\varphi_* := \varpi_0 \Lambda\varphi + f$, we note that this function is equivalent to φ and $\varphi_* = \varpi_0 \Lambda\varphi_* + f$ for all $\tau \in \bar{J}$. We will consider this special solution of (4) and will denote it by φ again.

It is evident that function φ is bounded. It follows from (4) that

$$\sup_{\tau \in \bar{J}} |\varphi(\tau)| \leq \varpi_0 \sup_{\tau \in \bar{J}} |\varphi(\tau)| + \sup_{\tau \in \bar{J}} |f(\tau)|.$$

So (27) holds.

Now put $\varphi(\tau) := \varphi(0)$ for $\tau < 0$ and $\varphi(\tau) := \varphi(\tau_*)$ for $\tau > \tau_*$. Let

$$V_N := \max_{M \in \llbracket 1, N \rrbracket} \left[\sup_{0 = \tau_0 < \tau_1 < \dots < \tau_M = \tau_*} \sum_{i=0}^{M-1} |\varphi(\tau_{i+1}) - \varphi(\tau_i)| \right].$$

Note that $V_N < \infty$ and that

$$V_N = \max_{M \in \llbracket 1, N \rrbracket} \left[\sup_{-\infty < \tau_0 < \tau_1 < \dots < \tau_M < \infty} \sum_{i=0}^{M-1} |\varphi(\tau_{i+1}) - \varphi(\tau_i)| \right].$$

It follows from (4) that

$$\begin{aligned} \varphi(\tau_{i+1}) - \varphi(\tau_i) &= \varpi_0 \int_{\mathbb{R}} \mathcal{E}(|\tau|) (\varphi(\tau + \tau_{i+1}) - \varphi(\tau + \tau_i)) d\tau \\ &\quad + \varpi_0 \int_{-\tau_{i+1}}^{-\tau_i} \mathcal{E}(|\tau|) d\tau \varphi(0) - \varpi_0 \int_{\tau_* - \tau_{i+1}}^{\tau_* - \tau_i} \mathcal{E}(|\tau|) d\tau \varphi(\tau_*) \\ &\quad + f(\tau_{i+1}) - f(\tau_i). \end{aligned}$$

So

$$\begin{aligned} \sum_{i=0}^{M-1} |\varphi(\tau_{i+1}) - \varphi(\tau_i)| &\leq \varpi_0 \int_{\mathbb{R}} \mathcal{E}(|\tau|) \sum_{i=0}^{M-1} |\varphi(t + \tau_{i+1}) - \varphi(t + \tau_i)| \, dt \\ &+ \varpi_0 \int_{-\tau_*}^0 \mathcal{E}(|\tau|) \, d\tau |\varphi(0)| + \varpi_0 \int_0^{\tau_*} \mathcal{E}(\tau) \, d\tau |\varphi(\tau_*)| + \sum_{i=0}^{M-1} |f(\tau_{i+1}) - f(\tau_i)| \\ &\leq \varpi_0 V_N + \varpi_0 \sup_{\tau \in \bar{J}} |\varphi(\tau)| + \operatorname{var}_{\bar{J}} f \leq \varpi_0 V_N + \gamma_1 \sup_{\tau \in \bar{J}} |f(\tau)| + \operatorname{var}_{\bar{J}} f. \end{aligned}$$

As a consequence $V_N \leq \varpi_0 V_N + \|f\|_{BV}$ and $V_N \leq \gamma_0 \|f\|_{BV}$. The last inequality implies (26). \square

COROLLARY 3. *If $\Lambda f \in BV(\bar{J})$, then $\Lambda\varphi \in BV(\bar{J})$ and*

$$(28) \quad \operatorname{var}_{\bar{J}} \Lambda\varphi \leq \gamma_0 \|\Lambda f\|_{BV},$$

$$(29) \quad \sup_{\tau \in \bar{J}} |\Lambda\varphi(\tau)| \leq \gamma_0 \sup_{\tau \in \bar{J}} |\Lambda f(\tau)|.$$

To prove Corollary 3 we remind that $y = \Lambda\varphi$ solves (10).

4. Some subsidiary results. In this section we will consider some properties of operators π^h and $I - \pi^h$. We come back to the general notations of the paper, introduced in section 2. In addition, $p' \in [1, \infty]$ will denote the conjugate of $p \in [1, \infty]$ in the sense that $1/p + 1/p' = 1$.

It is evident that $(\pi^h)^2 = \pi^h$. So π^h and $I - \pi^h$ are projections from $L^p(J)$ into $L^p(J)$. It is easy to see that for all $p \in [1, \infty]$ we have

$$\|\pi^h\|_{L^p \rightarrow L^p} = 1.$$

Moreover, π^h has this self-adjoint-like property:

$$(\pi^h \varphi, \psi) = (\varphi, \pi^h \psi) \quad \text{for all } \varphi \in L^p(J), \psi \in L^{p'}(J).$$

Let $f \in L^p(J)$, $\mathcal{E} \in L^r(\mathbb{R}^+)$ for all $r \in [1, \infty[$, and $\alpha \in]0, 1[$. We introduce the following difference operator Δ_α^h and two moduli of continuity, all three associated with the grid J^h :

$$\begin{aligned} \Delta_\alpha^h f(\tau) &:= \begin{cases} \Delta_{\alpha h_i} f(\tau) & \text{for all } \tau \in [\tau_{i-1}, \tau_i - \alpha h_i], \\ 0 & \text{for all } \tau \in]\tau_i - \alpha h_i, \tau_i[, \end{cases} \quad i \in \llbracket 1, n \rrbracket, \\ \omega_p(f, J^h) &:= \begin{cases} 2^{1/p} \left[\int_0^1 \|\Delta_\alpha^h f\|_{L^p(J)}^p \, d\alpha \right]^{1/p} & \text{for all } p \in [1, \infty[, \\ \max_{i \in \llbracket 1, n \rrbracket} \operatorname{esssup}_{(\tau, \tau') \in [\tau_{i-1}, \tau_i]^2} |f(\tau) - f(\tau')| & \text{for } p = \infty, \end{cases} \\ \bar{\omega}_r(\mathcal{E}, J^h) &:= \sup_{0 < \tau' < \tau_*} \omega_r(\mathcal{E}(|\cdot - \tau'|), J^h) \\ &= 2^{1/r} \sup_{0 < \tau' < \tau_*} \left[\int_0^1 \|\Delta_\alpha^h \mathcal{E}(|\cdot - \tau'|)\|_{L^r}^r \, d\alpha \right]^{1/r} \quad \text{for all } r \in [1, \infty[. \end{aligned}$$

LEMMA 5. *The following estimates hold:*

$$(30) \quad \|(I - \pi^h) f\|_{L^p} \leq \omega_p(f, J^h) \quad \text{for all } f \in L^p(J),$$

$$(31) \quad \|(I - \pi^h) f\|_{L^q} \leq h_{\max}^{1/q} \operatorname{var}_{\bar{J}} f \quad \text{for all } f \in BV(\bar{J}),$$

$$(32) \quad \|(I - \pi^h) f\|_{L^q} \leq h_{\max}^{1/s} \|Df\|_{L^p} \quad \text{for all } f \in W^{1,p}(J).$$

Proof. Let $f \in L^p(J)$. It follows from (7) that, for all $i \in \llbracket 1, n \rrbracket$ and $\tau \in J_{i-1/2}$,

$$(33) \quad (I - \pi^h) f(\tau) = \frac{1}{h_i} \int_{\tau_{i-1}}^{\tau_i} (f(\tau) - f(\tau')) d\tau'.$$

If $p < \infty$, then

$$\begin{aligned} & \int_{\tau_{i-1}}^{\tau_i} |(I - \pi^h) f(\tau)|^p d\tau \leq \int_{\tau_{i-1}}^{\tau_i} \left[\frac{1}{h_i} \int_{\tau_{i-1}}^{\tau_i} |f(\tau') - f(\tau)| d\tau' \right]^p d\tau \\ & \leq \frac{1}{h_i} \int_{\tau_{i-1}}^{\tau_i} \left[\int_{\tau_{i-1}}^{\tau_i} |f(\tau') - f(\tau)|^p d\tau' \right] d\tau = \frac{2}{h_i} \int_{\tau_{i-1}}^{\tau_i} \left[\int_{\tau}^{\tau_i} |f(\tau') - f(\tau)|^p d\tau' \right] d\tau \\ & = 2 \int_{\tau_{i-1}}^{\tau_i} \left[\int_0^{(\tau_i - \tau)/h_i} |f(\tau + \alpha h_i) - f(\tau)|^p d\alpha \right] d\tau = 2 \int_0^1 \left[\int_{\tau_{i-1}}^{\tau_i - \alpha h_i} |\Delta_{\alpha h_i} f(\tau)|^p d\tau \right] d\alpha. \end{aligned}$$

As a consequence

$$\|(I - \pi^h) f\|_{L^p}^p \leq 2 \sum_{i=1}^n \int_0^1 \left[\int_{\tau_{i-1}}^{\tau_i - \alpha h_i} |\Delta_{\alpha h_i} f(\tau)|^p d\tau \right] d\alpha = \omega_p(f, J^h)^p.$$

In the case $p = \infty$ the estimate is evident.

Let $f \in BV(J)$. It follows from (33) that, for all $i \in \llbracket 1, n \rrbracket$,

$$\sup_{\tau \in [\tau_{i-1}, \tau_i[} |(I - \pi^h) f(\tau)| \leq \operatorname{var}_{[\tau_{i-1}, \tau_i]} f,$$

so

$$\|(I - \pi^h) f\|_{L^\infty} \leq \frac{\operatorname{var} f}{J},$$

and if $q < \infty$, then

$$\|(I - \pi^h) f\|_{L^q}^q = \sum_{i=1}^n \|(I - \pi^h) f\|_{L^q(\tau_{i-1}, \tau_i)}^q \leq \sum_{i=1}^n h_i \left(\operatorname{var}_{[\tau_{i-1}, \tau_i]} f \right)^q \leq h_{\max} \left(\frac{\operatorname{var} f}{J} \right)^q.$$

Let $f \in W^{1,p}(J)$. It follows from (33) that, for all $i \in \llbracket 1, n \rrbracket$,

$$\sup_{\tau \in [\tau_{i-1}, \tau_i]} |(I - \pi^h) f(\tau)| \leq \|Df\|_{L^1(\tau_{i-1}, \tau_i)} \leq h_i^{1-1/p} \|Df\|_{L^p(\tau_{i-1}, \tau_i)}.$$

If $q = \infty$, then (32) is evident. If $q < \infty$, then

$$\begin{aligned} \|(I - \pi^h) f\|_{L^q(J)}^q &= \sum_{i=1}^n \|(I - \pi^h) f\|_{L^q(\tau_{i-1}, \tau_i)}^q \\ &\leq \sum_{i=1}^n h_i^{q-q/p+1} \|Df\|_{L^p(\tau_{i-1}, \tau_i)}^q \\ &\leq h_{\max}^{q-q/p+1} \|Df\|_{L^p}^q. \end{aligned}$$

The proof is complete. \square

LEMMA 6. Let $g \in L^1(0, 1)$ be a positive nonincreasing function. Then $g(t) = o(t^{-1})$ as $t \rightarrow 0^+$.

Proof. Otherwise there exist a number $c_0 > 0$ and a sequence $(t_k)_{k=1}^\infty$ in $]0, 1[$ such that $\lim_{k \rightarrow \infty} t_k = 0$, $t_{k+1} < t_k/2$, and $g(t_k) \geq c_0 t_k^{-1}$ for all $k \geq 1$. So

$$\int_0^1 g(t) dt \geq \sum_{k=1}^\infty g(t_k)t_k(1 - t_{k+1}/t_k) \geq c_0 \sum_{k=1}^\infty (1 - t_{k+1}/t_k) \geq \sum_{k=1}^\infty 1/2 = +\infty.$$

This contradiction proves the lemma. \square

COROLLARY 4. $\mathcal{E}^r(t) = o(t^{-1})$ as $t \rightarrow 0^+$ for all $r \in [1, \infty[$.

LEMMA 7. If

$$(34) \quad \mathcal{E} \in W^{1,r}(\delta, \infty) \quad \text{for all } \delta > 0 \text{ and all } r \in [1, \infty[,$$

then

$$\omega_r(\mathcal{E}, \eta) \leq \sup_{0 < \delta \leq \eta} 4^{1/r} \left(\int_0^{\delta/2} \mathcal{E}(\tau)^r d\tau + \delta^r \int_{\delta/2}^\infty |D\mathcal{E}(\tau)|^r d\tau \right)^{1/r} \quad \text{for all } r \in]1, \infty[.$$

(35)

If in addition

$$(36) \quad D\mathcal{E}(\tau) = o(\tau^{-1}\mathcal{E}(\tau)) \quad \text{as } \tau \rightarrow 0^+,$$

then, for all $r \in [1, \infty[$,

$$(37) \quad \omega_r(\mathcal{E}, \eta) \leq 2^{1/r} \eta^{1/r} \mathcal{E}(\eta/2)(1 + o(1)) \quad \text{as } \eta \rightarrow 0^+.$$

Proof. Note that

$$\begin{aligned} & \int_{-\infty}^\infty |\mathcal{E}(|\tau + \delta|) - \mathcal{E}(|\tau|)|^r d\tau = 2 \int_0^\infty |\mathcal{E}(\tau + \delta/2) - \mathcal{E}(\tau - \delta/2)|^r d\tau \\ & \leq 2 \int_0^{\delta/2} \mathcal{E}(|\tau - \delta/2|)^r d\tau + 2 \int_{\delta/2}^\delta \mathcal{E}(|\tau - \delta/2|)^r d\tau + 2 \int_\delta^\infty (\mathcal{E}(\tau - \delta/2) - \mathcal{E}(\tau + \delta/2))^r d\tau \\ & \leq 4 \int_0^{\delta/2} \mathcal{E}(\tau)^r d\tau + 2 \int_{\delta/2}^\infty (\mathcal{E}(\tau) - \mathcal{E}(\tau + \delta))^r d\tau \\ & \leq 4 \int_0^{\delta/2} \mathcal{E}(\tau)^r d\tau + 2 \int_{\delta/2}^\infty \left[\int_\tau^{\tau+\delta} |D\mathcal{E}(\tau')| d\tau' \right]^r d\tau \\ & \leq 4 \int_0^{\delta/2} \mathcal{E}(\tau)^r d\tau + 2\delta^{r-1} \int_{\delta/2}^\infty \left[\int_\tau^{\tau+\delta} |D\mathcal{E}(\tau')|^r d\tau' \right] d\tau \\ & \leq 4 \int_0^{\delta/2} \mathcal{E}(\tau)^r d\tau + 2\delta^r \int_{\delta/2}^\infty |D\mathcal{E}(\tau)|^r d\tau. \end{aligned}$$

So (35) holds.

To prove (37) we note that $\delta \mathcal{E}^r(\delta/2) \rightarrow 0$ as $\delta \rightarrow 0^+$. (See Corollary 4.) By our assumptions and the L'Hospital rule,

$$\begin{aligned} \lim_{\delta \rightarrow 0^+} \frac{2 \int_0^{\delta/2} \mathcal{E}^r(\tau) d\tau}{\delta \mathcal{E}^r(\delta/2)} &= \lim_{\delta \rightarrow 0^+} \frac{\mathcal{E}^r(\delta/2)}{\mathcal{E}^r(\delta/2) + (r/2)\delta \mathcal{E}^{r-1}(\delta/2) D\mathcal{E}(\delta/2)} = 1, \\ \lim_{\delta \rightarrow 0^+} \frac{\delta^r \int_{\delta/2}^{\infty} |D\mathcal{E}(\tau)|^r d\tau}{\delta \mathcal{E}^r(\delta/2)} &= \lim_{\delta \rightarrow 0^+} \frac{\int_{\delta/2}^{\infty} |D\mathcal{E}(\tau)|^r d\tau}{\delta^{1-r} \mathcal{E}^r(\delta/2)} \\ &= \lim_{\delta \rightarrow 0^+} \frac{-|D\mathcal{E}(\delta/2)|^r/2}{(1-r)\delta^{-r} \mathcal{E}^r(\delta/2) + (r/2)\delta^{1-r} \mathcal{E}^{r-1}(\delta/2) D\mathcal{E}(\delta/2)} = 0. \end{aligned}$$

So (37) follows from (35). \square

Remark 6. We do not need the assumption (34) to prove that

$$\omega_1(\mathcal{E}, \eta) \leq 4 \int_0^{\eta/2} \mathcal{E}(\tau) d\tau.$$

Proof. In fact,

$$\begin{aligned} &\int_{-\infty}^{\infty} |\mathcal{E}(|\tau + \delta|) - \mathcal{E}(|\tau|)| d\tau = \int_{-\infty}^{\infty} |\mathcal{E}(|\tau + \delta/2|) - \mathcal{E}(|\tau - \delta/2|)| d\tau \\ &= 2 \int_0^{\infty} |\mathcal{E}(|\tau + \delta/2|) - \mathcal{E}(|\tau - \delta/2|)| d\tau = 2 \int_0^{\infty} (\mathcal{E}(|\tau - \delta/2|) - \mathcal{E}(|\tau + \delta/2|)) d\tau \\ &= 2 \int_{-\delta/2}^{\infty} \mathcal{E}(|\tau|) d\tau - 2 \int_{\delta/2}^{\infty} \mathcal{E}(|t|) dt = 4 \int_0^{\delta/2} \mathcal{E}(\tau) d\tau. \end{aligned}$$

The proof is complete. \square

LEMMA 8. *If \mathcal{E} satisfies (34), then*

$$(38) \quad \bar{\omega}_r(\mathcal{E}, J^h) \leq 4^{1/r} \left[\int_0^{h_{\max}/2} \mathcal{E}(\tau)^r d\tau + h_{\max}^r \int_{h_{\max}/2}^{\infty} |D\mathcal{E}(\tau)|^r d\tau \right]^{1/r}.$$

If in addition \mathcal{E} satisfies (36), then

$$(39) \quad \bar{\omega}_r(\mathcal{E}, J^h) \leq 2^{1/r} h_{\max}^{1/r} \mathcal{E}(h_{\max}/2) (1 + o(1)) \quad \text{as } h_{\max} \rightarrow 0^+.$$

Proof. Note that

$$\bar{\omega}_r(\mathcal{E}, J^h)^r = 2 \sup_{0 < \tau' < \tau_*} \sum_{i=1}^n \int_0^1 \left[\int_{t_{i-1}}^{t_i - \alpha h_i} |\mathcal{E}(|\tau + \alpha h_i|) - \mathcal{E}(|\tau|)|^r d\tau \right] d\alpha,$$

where $t_i := \tau_i - \tau'$ for all $i \in \llbracket 0, n \rrbracket$.

Let us estimate

$$\begin{aligned} I_i &:= \int_0^1 \left[\int_{t_{i-1}}^{t_i - \alpha h_i} |\mathcal{E}(|\tau + \alpha h_i|) - \mathcal{E}(|\tau|)|^r d\tau \right] d\alpha \\ &= \int_{t_{i-1}}^{t_i} \left[\int_0^{(t_i - \tau)/h_i} |\mathcal{E}(|\tau + \alpha h_i|) - \mathcal{E}(|\tau|)|^r d\alpha \right] d\tau. \end{aligned}$$

If $-h_{\max}/2 \leq t_{i-1} < t_i \leq h_{\max}/2$, then

$$\begin{aligned} I_i &\leq \int_0^1 \left[\int_{t_{i-1}}^{t_i - \alpha h_i} (\mathcal{E}(|\tau + \alpha h_i|)^r + \mathcal{E}(|\tau|)^r) d\tau \right] d\alpha \\ &= \int_0^1 \left[\int_{t_{i-1} + \alpha h_i}^{t_i} \mathcal{E}(|\tau|)^r d\tau \right] d\alpha + \int_0^1 \left[\int_{t_{i-1}}^{t_i - \alpha h_i} \mathcal{E}(|\tau|)^r d\tau \right] d\alpha \\ &= \int_{t_{i-1}}^{t_i} \frac{\tau - t_{i-1}}{h_i} \mathcal{E}(|\tau|)^r d\tau + \int_{t_{i-1}}^{t_i} \frac{t_i - \tau}{h_i} \mathcal{E}(|\tau|)^r d\tau = \int_{t_{i-1}}^{t_i} \mathcal{E}(|\tau|)^r d\tau. \end{aligned}$$

Let $t_{i-1} < 0 < h_{\max}/2 < t_i$. Note that $0 < -t_{i-1} < h_{\max}/2$ and

$$|\mathcal{E}(|\tau + \alpha h_i|) - \mathcal{E}(|\tau|)| \leq \begin{cases} \mathcal{E}(|\tau + \alpha h_i|) & \text{for } \tau \in (t_{i-1}, -\alpha h_i/2), \\ \mathcal{E}(|\tau|) & \text{for } \tau \in [-\alpha h_i/2, h_{\max}/2], \\ \mathcal{E}(h_{\max}/2) - \mathcal{E}(t_i) & \text{for } \tau \in (h_{\max}/2, t_i - \alpha h_i). \end{cases}$$

So

$$\begin{aligned} I_i &\leq \int_{t_{i-1}}^0 \left[\int_0^{-2\tau/h_i} \mathcal{E}(|\tau + \alpha h_i|)^r d\alpha \right] d\tau + \int_{t_{i-1}}^0 \left[\int_{-2\tau/h_i}^{(t_i - \tau)/h_i} \mathcal{E}(|\tau|)^r d\alpha \right] d\tau \\ &\quad + \int_0^{h_{\max}/2} \left[\int_0^{(t_i - \tau)/h_i} \mathcal{E}(|\tau|)^r d\alpha \right] d\tau + h_i \left| \int_{h_{\max}/2}^{t_i} D\mathcal{E}(\tau) d\tau \right|^r \\ &= \int_{t_{i-1}}^0 \left[\int_0^{(\tau - t_{i-1})/h_i} \mathcal{E}(|\tau|)^r d\alpha \right] d\tau + \int_0^{-t_{i-1}} \left[\int_{2\tau/h_i}^{(\tau - t_{i-1})/h_i} \mathcal{E}(|\tau|)^r d\alpha \right] d\tau \\ &\quad + \int_{t_{i-1}}^0 \frac{t_i + \tau}{h_i} \mathcal{E}(|\tau|)^r d\tau + \int_0^{h_{\max}/2} \frac{t_i - \tau}{h_i} \mathcal{E}(|\tau|)^r d\tau + h_i \left| \int_{h_{\max}/2}^{t_i} D\mathcal{E}(\tau) d\tau \right|^r \\ &\leq \int_{t_{i-1}}^0 \left(1 + \frac{2\tau}{h_i} \right) \mathcal{E}(|\tau|)^r d\tau + \int_0^{-t_{i-1}} \mathcal{E}(|\tau|)^r d\tau + \int_{-t_{i-1}}^{h_{\max}/2} \frac{t_i - \tau}{h_i} \mathcal{E}(|\tau|)^r d\tau \\ &\quad + h_{\max} \left| \int_{h_{\max}/2}^{t_i} D\mathcal{E}(\tau) d\tau \right|^r \leq \int_{t_{i-1}}^{h_{\max}/2} \mathcal{E}(|\tau|)^r d\tau + h_{\max}^r \int_{h_{\max}/2}^{t_i} |D\mathcal{E}(\tau)|^r d\tau. \end{aligned}$$

If $t_{i-1} \leq -h_{\max}/2 < 0 < t_i$, then in an analogous manner

$$I_i \leq \int_{-h_{\max}/2}^{t_i} \mathcal{E}(|\tau|)^r d\tau + h_{\max}^r \int_{t_{i-1}}^{-h_{\max}/2} |D\mathcal{E}(\tau)|^r d\tau.$$

If $0 \leq t_{i-1} \leq h_{\max}/2 < t_i$, then

$$\begin{aligned} I_i &\leq \int_{t_{i-1}}^{h_{\max}/2} \mathcal{E}(\tau)^r d\tau + \int_{h_{\max}/2}^{t_i} (\mathcal{E}(\tau) - \mathcal{E}(t_i))^r d\tau \\ &\leq \int_{t_{i-1}}^{h_{\max}/2} \mathcal{E}(\tau)^r d\tau + h_i^r \int_{h_{\max}/2}^{t_i} |D\mathcal{E}(\tau)|^r d\tau. \end{aligned}$$

In an analogous way if $t_{i-1} < -h_{\max}/2 \leq t_i \leq 0$, then

$$I_i \leq h_i^r \int_{t_{i-1}}^{-h_{\max}/2} |D\mathcal{E}(\tau)|^r d\tau + \int_{-h_{\max}/2}^{t_i} \mathcal{E}(|\tau|)^r d\tau.$$

If $h_{\max}/2 \leq t_{i-1}$, then

$$\begin{aligned} I_i &= \int_0^1 \left[\int_{t_{i-1}}^{t_i - \alpha h_i} (\mathcal{E}(\tau) - \mathcal{E}(\tau + \alpha h_i))^r d\tau \right] d\alpha \\ &\leq h_i \left[\int_{t_{i-1}}^{t_i} |D\mathcal{E}(\tau)| d\tau \right]^r \leq h_i^r \int_{t_{i-1}}^{t_i} |D\mathcal{E}(\tau)|^r d\tau. \end{aligned}$$

Analogously, if $t_i \leq -h_{\max}/2$ then

$$I_i \leq h_i^r \int_{t_{i-1}}^{t_i} |D\mathcal{E}(\tau)|^r d\tau.$$

The summation of all the estimates for I_i produced above gives us

$$\begin{aligned} \omega_r(\mathcal{E}(|\cdot - \tau'|), J^h)^r &\leq 2 \left[h_{\max}^r \int_{-\infty}^{-h_{\max}/2} |D\mathcal{E}(\tau)|^r d\tau + \int_{-h_{\max}/2}^{h_{\max}/2} \mathcal{E}(|\tau|)^r d\tau \right. \\ &\quad \left. + h_{\max}^r \int_{h_{\max}/2}^{\infty} |D\mathcal{E}(\tau)|^r d\tau \right] \\ &= 4 \left[\int_0^{h_{\max}/2} \mathcal{E}(\tau)^r d\tau + h_{\max}^r \int_{h_{\max}/2}^{\infty} |D\mathcal{E}(\tau)|^r d\tau \right]. \end{aligned}$$

So estimate (38) holds. As in the proof of Lemma 7, (38) implies (39). \square

LEMMA 9. *If in (13) $s < \infty$, then*

$$(40) \quad \|(I - \pi^h) \Lambda\|_{L^p \rightarrow L^q} \leq \omega_s(\mathcal{E}, h_{\max})^{1-s/q} \overline{\omega}_s(\mathcal{E}, J^h)^{s/q},$$

$$(41) \quad \|\Lambda(I - \pi^h)\|_{L^p \rightarrow L^q} \leq \omega_s(\mathcal{E}, h_{\max})^{s/q} \overline{\omega}_s(\mathcal{E}, J^h)^{1-s/q}.$$

Proof. If $q \neq \infty$, Lemma 5 and generalized Minkowsky inequality we get

$$\begin{aligned} \|(I - \pi^h) \Lambda f\|_{L^q}^q &\leq \omega_q(\Lambda f, J^h)^q = 2 \int_0^1 \|\Delta_\alpha^h \Lambda f\|_{L^q}^q d\alpha \\ &\leq 2 \int_0^1 \left\| \left[\int_0^{\tau_*} |\Delta_\alpha^h \mathcal{E}(|\cdot - \tau'|)|^s d\tau' \right]^{1/s-1/q} \left[\int_0^{\tau_*} |\Delta_\alpha^h \mathcal{E}(|\cdot - \tau'|)|^{sp/q} |f(\tau')|^p d\tau' \right]^{1/p} \right\|_{L^q}^q d\alpha \\ &\leq 2\omega_s(\mathcal{E}, h_{\max})^{q-s} \int_0^1 \left\| \int_0^{\tau_*} |\Delta_\alpha^h \mathcal{E}(|\cdot - \tau'|)|^{sp/q} |f(\tau')|^p d\tau' \right\|_{L^{q/p}}^{q/p} d\alpha \\ &\leq 2\omega_s(\mathcal{E}, h_{\max})^{q-s} \left[\int_0^{\tau_*} \left(\int_0^1 \|\Delta_\alpha^h \mathcal{E}(|\cdot - \tau'|)\|_{L^s}^s d\alpha \right)^{p/q} |f(\tau')|^p d\tau' \right]^{q/p} \\ &\leq \omega_s(\mathcal{E}, h_{\max})^{q-s} \overline{\omega}_s(\mathcal{E}, J^h)^s \|f\|_{L^p}^q. \end{aligned}$$

If $q = \infty$, then $s = p'$, and it follows from (33) that

$$\begin{aligned} |(I - \pi^h) \Lambda f(\tau)| &\leq \frac{1}{h_i} \int_{\tau_{i-1}}^{\tau_i} |\Lambda f(\tau) - \Lambda f(\tau')| d\tau' \\ &\leq \frac{1}{h_i} \int_{\tau_{i-1}}^{\tau_i} \left[\int_0^{\tau_*} |\mathcal{E}(|\tau - \tilde{\tau}|) - \mathcal{E}(|\tau' - \tilde{\tau}|)| |f(\tilde{\tau})| d\tilde{\tau} \right] d\tau' \\ &\leq \frac{1}{h_i} \int_{\tau_{i-1}}^{\tau_i} \|\mathcal{E}(|\tau - \cdot|) - \mathcal{E}(|\tau' - \cdot|)\|_{L^s} d\tau' \|f\|_{L^p} \leq \omega_s(\mathcal{E}, h_{\max}) \|f\|_{L^p}. \end{aligned}$$

So

$$\|(I - \pi^h) \Lambda f\|_{L^\infty} \leq \omega_s(\mathcal{E}, h_{\max}) \|f\|_{L^p}.$$

The inequality (40) is proved.

Note that

$$(\Lambda (I - \pi^h) \varphi, \psi) = (\varphi, (I - \pi^h) \Lambda \psi) \quad \text{for all } \varphi \in L^p(J), \psi \in L^{q'}(J),$$

and hence

$$\|\Lambda (I - \pi^h)\|_{L^p \rightarrow L^q} = \|(I - \pi^h) \Lambda\|_{L^{q'} \rightarrow L^{p'}}.$$

So (41) holds. \square

LEMMA 10. *The following bound holds with $2/r = 1 + 1/s$:*

$$(42) \quad \|\Lambda (1 - \pi^h) \Lambda\|_{L^p \rightarrow L^q} \leq \omega_r(\mathcal{E}, h_{\max})^{2-r} \bar{\omega}_r(\mathcal{E}, J^h)^r.$$

Proof. Let $2/\mu = 1/p + 1/q$. Using the formula

$$\Lambda (I - \pi^h) \Lambda = \Lambda (1 - \pi^h) (1 - \pi^h) \Lambda$$

and Lemma 9, we get

$$\begin{aligned} \|\Lambda (1 - \pi^h) \Lambda\|_{L^p \rightarrow L^q} &\leq \|\Lambda (1 - \pi^h)\|_{L^\mu \rightarrow L^q} \|(1 - \pi^h) \Lambda\|_{L^p \rightarrow L^\mu} \\ &\leq \omega_r(\mathcal{E}, h_{\max})^{r/q} \bar{\omega}_r(\mathcal{E}, J^h)^{1-r/q} \omega_r(\mathcal{E}, h_{\max})^{1-r/\mu} \bar{\omega}_r(\mathcal{E}, J^h)^{r/\mu} \\ &= \omega_r(\mathcal{E}, h_{\max})^{2-r} \bar{\omega}_r(\mathcal{E}, J^h)^r. \end{aligned}$$

And the proof is complete. \square

5. Error estimates in $L^q(J)$.

THEOREM 7. *For the Galerkin approximation, the following error estimates hold:*

$$(43) \quad \|\pi^h \varepsilon^h\|_{L^q} \leq \gamma_1 \|\Lambda (I - \pi^h) \varphi\|_{L^q},$$

$$(44) \quad \|\varepsilon^h\|_{L^q} \leq \gamma_0 \|(I - \pi^h) \varphi\|_{L^q}.$$

Proof. Applying π^h to (4),

$$\pi^h \varphi = \varpi_0 \pi^h \Lambda \pi^h \varphi + \pi^h f + \varpi_0 \pi^h \Lambda (I - \pi^h) \varphi.$$

Subtracting this equation from (8),

$$\pi^h \varepsilon^h = \varpi_0 \pi^h \Lambda \pi^h \varepsilon^h - \varpi_0 \pi^h \Lambda (I - \pi^h) \varphi.$$

As $\|\pi^h\|_{L^q \rightarrow L^q} = 1$ and $\|\Lambda\|_{L^q \rightarrow L^q} \leq 1$,

$$\|\pi^h \varepsilon^h\|_{L^q} \leq \varpi_0 \|\pi^h \varepsilon^h\|_{L^q} + \varpi_0 \|\Lambda (I - \pi^h) \varphi\|_{L^q}.$$

Hence (43) holds. The estimate (44) follows from (43) and the inequality

$$\|\varepsilon^h\|_{L^q} \leq \|\pi^h \varepsilon^h\|_{L^q} + \|(I - \pi^h) \varphi\|_{L^q}.$$

This ends the proof. \square

COROLLARY 5. *It follows from Theorem 4 and inequality (44), that if $f \in L^p(J)$ for some $p < \infty$, then $\varphi \in L^p(J)$ and $\varepsilon^h \rightarrow 0$ in $L^p(J)$ as $h_{\max} \rightarrow 0$.*

COROLLARY 6. *If $f \in L^p(J)$ and $s < \infty$, then*

$$(45) \quad \|\pi^h \varepsilon^h\|_{L^q} \leq \gamma_0 \gamma_1 \omega_s(\mathcal{E}, h_{\max})^{s/q} \overline{\omega}_s(\mathcal{E}, J^h)^{1-s/q} \|f\|_{L^p}.$$

If $f \in BV(J)$, then

$$(46) \quad \|\pi^h \varepsilon^h\|_{L^q} \leq \gamma_0 \gamma_1 \omega_1(\mathcal{E}, h_{\max})^{1/q} \overline{\omega}_1(\mathcal{E}, J^h)^{1-1/q} h_{\max}^{1/q} \|f\|_{BV},$$

$$(47) \quad \|\varepsilon^h\|_{L^q} \leq \gamma_0^2 h_{\max}^{1/q} \|f\|_{BV}.$$

If $f \in W^{1,p}(J)$ for some $p < \infty$, then

$$(48) \quad \|\pi^h \varepsilon^h\|_{L^q} \leq \gamma_0 \gamma_1 \omega_1(\mathcal{E}, h_{\max})^{1/q} \overline{\omega}_1(\mathcal{E}, J^h)^{1-1/q} h_{\max}^{1/s} \|f\|_{W^{1,p}},$$

$$(49) \quad \|\varepsilon^h\|_{L^q} \leq \gamma_0^2 h_{\max}^{1/s} \|f\|_{W^{1,p}}.$$

Proof. Inequality (45) follows from (43), (41), and (17). To prove (46)–(49) we use estimates (31), (26), and (32), (20), and for (46), (48), in addition, the formula

$$\Lambda(I - \pi^h)\varphi = \Lambda(I - \pi^h)(I - \pi^h)\varphi$$

and the inequality

$$\|\Lambda(I - \pi^h)\varphi\|_{L^q} \leq \|\Lambda(I - \pi^h)\|_{L^q \rightarrow L^q} \|(I - \pi^h)\varphi\|_{L^q}.$$

This completes the proof. \square

THEOREM 8. *For the Sloan approximation, the following error estimate holds:*

$$(50) \quad \|\overline{\varepsilon}^h\|_{L^q} \leq \gamma_1 \|\Lambda(I - \pi^h)\varphi\|_{L^q}.$$

Proof. Subtracting (4) from (9),

$$\overline{\varepsilon}^h = \varpi_0 \Lambda \pi^h \overline{\varepsilon}^h - \varpi_0 \Lambda (I - \pi^h) \varphi.$$

It is clear that (50) holds. \square

COROLLARY 7. *If $f \in L^p(J)$ and $s < \infty$, then*

$$(51) \quad \|\overline{\varepsilon}^h\|_{L^q} \leq \gamma_0 \gamma_1 \omega_s(\mathcal{E}, h_{\max})^{s/q} \overline{\omega}_s(\mathcal{E}, J^h)^{1-s/q} \|f\|_{L^p}.$$

If $f \in BV(\overline{J})$, then

$$(52) \quad \|\overline{\varepsilon}^h\|_{L^q} \leq \gamma_0 \gamma_1 \omega_1(\mathcal{E}, h_{\max})^{1/q} \overline{\omega}_1(\mathcal{E}, J^h)^{1-1/q} h_{\max}^{1/q} \|f\|_{BV}.$$

If $f \in W^{1,p}$ for some $p < \infty$, then

$$(53) \quad \|\overline{\varepsilon}^h\|_{L^q} \leq \gamma_0 \gamma_1 \omega_1(\mathcal{E}, h_{\max})^{1/q} \overline{\omega}_1(\mathcal{E}, J^h)^{1-1/q} h_{\max}^{1/s} \|f\|_{W^{1,p}}.$$

THEOREM 9. *For Kantorovich approximation, the following error estimates hold:*

$$(54) \quad \|\pi^h \overline{\varepsilon}^h\|_{L^q} \leq \varpi_0 \gamma_1 \|\Lambda(I - \pi^h)\Lambda\varphi\|_{L^q},$$

$$(55) \quad \|\overline{\varepsilon}^h\|_{L^q} \leq \gamma_1 \|(I - \pi^h)\Lambda\varphi\|_{L^q}.$$

Proof. The estimates (54) and (55) follow from Theorem 7 and Remark 1. \square
 COROLLARY 8. *If $f \in L^p(J)$ and $2/r = 1 + 1/s$, then*

$$(56) \quad \|\pi^h \tilde{\varepsilon}^h\|_{L^q} \leq \gamma_1^2 \omega_r(\mathcal{E}, h_{\max})^{2-r} \bar{\omega}_r(\mathcal{E}, J^h)^r \|f\|_{L^p},$$

$$(57) \quad \|\tilde{\varepsilon}^h\|_{L^q} \leq \gamma_0 \gamma_1 \omega_s(\mathcal{E}, h_{\max})^{1-s/q} \bar{\omega}_s(\mathcal{E}, J^h)^{s/q} \|f\|_{L^p}, \quad \text{if } s < \infty.$$

If $\Lambda f \in BV(J)$, then, for all $q \in [1, \infty[$,

$$(58) \quad \|\pi^h \tilde{\varepsilon}^h\|_{L^q} \leq \gamma_1^2 \omega_1(\mathcal{E}, h_{\max})^{1/q} \bar{\omega}_1(\mathcal{E}, J^h)^{1-1/q} h_{\max}^{1/q} \|\Lambda f\|_{BV},$$

$$(59) \quad \|\tilde{\varepsilon}^h\|_{L^q} \leq \gamma_0 \gamma_1 h_{\max}^{1/q} \|\Lambda f\|_{BV}.$$

If $\Lambda f \in W^{1,p}$ for some $p \in [1, \infty[$, then, for all $q \in [1, \infty[$,

$$(60) \quad \|\pi^h \tilde{\varepsilon}^h\|_{L^q} \leq \gamma_1^2 \omega_1(\mathcal{E}, h_{\max})^{1/q} \bar{\omega}_1(\mathcal{E}, J^h)^{1-1/q} h_{\max}^{1/s} \|\Lambda f\|_{W^{1,p}},$$

$$(61) \quad \|\tilde{\varepsilon}^h\|_{L^q} \leq \gamma_0 \gamma_1 h_{\max}^{1/s} \|\Lambda f\|_{W^{1,p}}.$$

Proof. To prove (56) and (57) we use estimates (42), (40), and (17). To prove (58) and (59) we use estimates

$$(62) \quad \|(I - \pi^h) \Lambda \varphi\|_{L^q} \leq h_{\max}^{1/q} \text{var}_{\bar{J}} \Lambda \varphi,$$

$$(63) \quad \|\Lambda(I - \pi^h) \Lambda \varphi\|_{L^q} \leq \|\Lambda(I - \pi^h)\|_{L^q \rightarrow L^q} \|(I - \pi^h) \Lambda \varphi\|_{L^q}$$

and (41), (28). To prove (60), (61) we use the following inequality

$$\|(I - \pi^h) \Lambda \varphi\|_{L^q} \leq h_{\max}^{1/s} \|D\Lambda \varphi\|_{L^p},$$

as well as (63), (41), and (20), applied to $y = \Lambda \varphi$ which solves (10). \square

Remark 7. This is a reminder that if $f \in BV(\bar{J})$, then $\Lambda f \in BV(\bar{J})$ and $\Lambda f \in W^{1,p}(J)$ for all $p \in [1, \infty[$. (See Lemma 3 and Lemma 4).

THEOREM 10. *For the iterated Kantorovich approximation, the following error estimate holds:*

$$(64) \quad \|\tilde{\varepsilon}^h\|_{L^q} \leq \varpi_0 \gamma_1 \|\Lambda(I - \pi^h) \Lambda \varphi\|_{L^q}.$$

Proof. The estimate follows from Theorem 8 and Remark 1. \square

COROLLARY 9. *If $f \in L^p(J)$ and $2/r = 1 + 1/s$, then*

$$(65) \quad \|\tilde{\varepsilon}^h\|_{L^q} \leq \gamma_1^2 \omega_r(\mathcal{E}, h_{\max})^{2-r} \bar{\omega}_r(\mathcal{E}, J^h)^r \|f\|_{L^p}.$$

If $\Lambda f \in BV(\bar{J})$, then, for all $q \in [1, \infty[$,

$$(66) \quad \|\tilde{\varepsilon}^h\|_{L^q} \leq \gamma_1^2 \omega_1(\mathcal{E}, h_{\max})^{1/q} \bar{\omega}_1(\mathcal{E}, J^h)^{1-1/q} h_{\max}^{1/q} \|\Lambda f\|_{BV}.$$

If $\Lambda f \in W^{1,p}(J)$ for some $p < \infty$, then

$$(67) \quad \|\tilde{\varepsilon}^h\|_{L^q} \leq \gamma_1^2 \omega_1(\mathcal{E}, h_{\max})^{1/q} \bar{\omega}_1(\mathcal{E}, J^h)^{1-1/q} h_{\max}^{1/s} \|\Lambda f\|_{W^{1,p}}.$$

If \mathcal{E} satisfies (34) and (36), then the bounds in corollaries 6, 7, 8, and 9 may be specified using (37) and (39).

THEOREM 11. Assume that \mathcal{E} satisfies (34) and (36). The following error estimates hold as $h_{\max} \rightarrow 0$:

If $f \in L^p(J)$, then

$$\begin{aligned} \|\pi^h \varepsilon^h\|_{L^q} &\leq 2^{1/s} \gamma_0 \gamma_1 h_{\max}^{1/s} \mathcal{E}(h_{\max}/2)(1 + o(1)) \|f\|_{L^p} \quad \text{for } s < \infty, \\ \|\bar{\varepsilon}^h\|_{L^q} &\leq 2^{1/s} \gamma_0 \gamma_1 h_{\max}^{1/s} \mathcal{E}(h_{\max}/2)(1 + o(1)) \|f\|_{L^p} \quad \text{for } s < \infty, \\ \|\pi^h \hat{\varepsilon}^h\|_{L^q} &\leq 2^{1+1/s} \gamma_1^2 h_{\max}^{1+1/s} \mathcal{E}^2(h_{\max}/2)(1 + o(1)) \|f\|_{L^p}, \\ \|\bar{\varepsilon}^h\|_{L^q} &\leq 2^{1/s} \gamma_0 \gamma_1 h_{\max}^{1/s} \mathcal{E}(h_{\max}/2)(1 + o(1)) \|f\|_{L^p} \quad \text{for } s < \infty, \\ \|\hat{\varepsilon}^h\|_{L^q} &\leq 2^{1+1/s} \gamma_1^2 h_{\max}^{1+1/s} \mathcal{E}^2(h_{\max}/2)(1 + o(1)) \|f\|_{L^p}. \end{aligned}$$

If $\Lambda f \in BV(\bar{J})$, then

$$\begin{aligned} \|\pi^h \bar{\varepsilon}^h\|_{L^q} &\leq 2\gamma_1^2 h_{\max}^{1+1/q} \mathcal{E}(h_{\max}/2)(1 + o(1)) \|\Lambda f\|_{BV}, \\ \|\hat{\varepsilon}^h\|_{L^q} &\leq \gamma_0 \gamma_1 h_{\max}^{1/q} \|\Lambda f\|_{BV}, \\ \|\bar{\varepsilon}^h\|_{L^q} &\leq 2\gamma_1^2 h_{\max}^{1+1/q} \mathcal{E}(h_{\max}/2)(1 + o(1)) \|\Lambda f\|_{BV}. \end{aligned}$$

If $\Lambda f \in W^{1,p}(J)$ for some $p < \infty$, then

$$\begin{aligned} \|\pi^h \bar{\varepsilon}^h\|_{L^q} &\leq 2\gamma_1^2 h_{\max}^{1+1/s} \mathcal{E}(h_{\max}/2)(1 + o(1)) \|\Lambda f\|_{W^{1,p}}, \\ \|\hat{\varepsilon}^h\|_{L^q} &\leq \gamma_0 \gamma_1 h_{\max}^{1/s} \|\Lambda f\|_{W^{1,p}}, \\ \|\bar{\varepsilon}^h\|_{L^q} &\leq 2\gamma_1^2 h_{\max}^{1+1/s} \mathcal{E}(h_{\max}/2)(1 + o(1)) \|\Lambda f\|_{W^{1,p}}. \end{aligned}$$

If $f \in BV(\bar{J})$, then

$$\begin{aligned} \|\pi^h \varepsilon^h\|_{L^q} &\leq 2\gamma_0 \gamma_1 h_{\max}^{1+1/q} \mathcal{E}(h_{\max}/2)(1 + o(1)) \|f\|_{BV}, \\ \|\varepsilon^h\|_{L^q} &\leq \gamma_0^2 h_{\max}^{1/q} \|f\|_{BV}, \\ \|\bar{\varepsilon}^h\|_{L^q} &\leq 2\gamma_0 \gamma_1 h_{\max}^{1+1/q} \mathcal{E}(h_{\max}/2)(1 + o(1)) \|f\|_{BV}. \end{aligned}$$

If $f \in W^{1,p}(J)$ for some $p < \infty$, then

$$\begin{aligned} \|\pi^h \varepsilon^h\|_{L^q} &\leq 2\gamma_0 \gamma_1 h_{\max}^{1+1/s} \mathcal{E}(h_{\max}/2)(1 + o(1)) \|f\|_{W^{1,p}}, \\ \|\varepsilon^h\|_{L^q} &\leq \gamma_0^2 h_{\max}^{1/s} \|f\|_{W^{1,p}}, \\ \|\bar{\varepsilon}^h\|_{L^q} &\leq 2\gamma_0 \gamma_1 h_{\max}^{1+1/s} \mathcal{E}(h_{\max}/2)(1 + o(1)) \|f\|_{W^{1,p}}. \end{aligned}$$

6. Error estimates in $W^{1,1}(J)$ and $BV(\bar{J})$. In this section we will estimate projection approximation errors in the norm of the space $BV(\bar{J})$. To notations (5), (6) we add

$$\begin{aligned} \tau_{i-1/2} &:= (\tau_{i-1} + \tau_i)/2, \quad i \in \llbracket 1, n \rrbracket, \\ h_{i+1/2} &:= \tau_{i+1/2} - \tau_{i-1/2} \quad \text{for all } i \in \llbracket 1, n-1 \rrbracket, \quad h_{1/2} := h_1/2, \quad h_{n+1/2} := h_n/2, \\ J_i &:=]\tau_{i-1/2}, \tau_{i+1/2}] \quad \text{for all } i \in \llbracket 1, n-1 \rrbracket, \quad J_0 := [0, \tau_{1/2}], \quad J_n :=]\tau_{n-1/2}, \tau_*]. \end{aligned}$$

Let $S^h(J)$ be the following space of piecewise constant functions:

$$S^h(J) := \{\psi : \psi(\tau) = \psi_i \quad \text{for all } \tau \in J_i, \text{ and all } i \in \llbracket 0, n \rrbracket\},$$

We consider the following family of standard hat functions: For $i \in \llbracket 0, n \rrbracket$,

$$e_i(\tau) := \begin{cases} (\tau - \tau_{i-1})/h_i & \text{for all } \tau \in [\tau_{i-1}, \tau_i], \\ (\tau_{i+1} - \tau)/h_{i+1} & \text{for all } \tau \in [\tau_i, \tau_{i+1}], \\ 0 & \text{for all } \tau \notin [\tau_{i-1}, \tau_{i+1}], \end{cases}$$

$$e_0(\tau) := \begin{cases} (\tau_1 - \tau)/h_1 & \text{for all } \tau \in [\tau_0, \tau_1], \\ 0 & \text{for all } \tau \notin [\tau_0, \tau_1], \end{cases}$$

$$e_n(\tau) := \begin{cases} (\tau - \tau_{n-1})/h_n & \text{for all } \tau \in [\tau_{n-1}, \tau_n], \\ 0 & \text{for all } \tau \notin [\tau_{n-1}, \tau_n]. \end{cases}$$

Note that

$$(1, e_i) = h_{i+1/2} \quad \text{for all } i \in \llbracket 0, n \rrbracket,$$

$$\sum_{i=0}^n e_i = 1 \quad \text{in } J.$$

Let us introduce the operators $\sigma^h : L^1(J) \rightarrow S^h(J)$, $\partial^h : L^1(J) \rightarrow S^h(J)$ such that

$$(\sigma^h \psi)(\tau) := \sigma^h \psi_i = h_{i+1/2}^{-1}(\psi, e_i) \quad \text{for all } \tau \in J_i, \text{ and all } i \in \llbracket 0, n \rrbracket,$$

$$(\partial^h \psi)(\tau) := \partial^h \psi_i = \begin{cases} \sigma^h \psi_i & \text{for all } \tau \in J_i, \text{ and all } i \in \llbracket 1, n-1 \rrbracket, \\ 0 & \text{for all } \tau \in J_i, \text{ if } i = 0 \text{ or } i = n. \end{cases}$$

LEMMA 11. *We have*

$$(68) \quad \|\sigma^h\|_{L^p(J) \rightarrow L^p(J)} = 1 \quad \text{for all } p \in [1, \infty].$$

Proof. If $p \in [1, \infty[$ and $\psi \in L^p(J)$, then, by Hölder inequality,

$$\begin{aligned} \|\sigma^h \psi\|_{L^p}^p &= \sum_{i=0}^n |\sigma^h \psi_i|^p h_{i+1/2} \leq \sum_{i=0}^n h_{i+1/2}^{1-p} (|\psi|, e_i)^p \\ &\leq \sum_{i=0}^n h_{i+1/2}^{1-p} (1, e_i)^{p-1} (|\psi|^p, e_i) = \left(|\psi|^p, \sum_{i=0}^n e_i \right) = \|\psi\|_{L^p}^p. \end{aligned}$$

For $p = \infty$ we have

$$\|\sigma^h \psi\|_{L^\infty} = \max_{i \in \llbracket 0, n \rrbracket} |\sigma^h \psi_i| \leq \max_{i \in \llbracket 0, n \rrbracket} h_{i+1/2}^{-1} (1, e_i) \|\psi\|_{L^\infty} = \|\psi\|_{L^\infty}.$$

Finally taking into account that $\sigma^h 1 = 1$ we get (68).

Let $D^h : S_{1/2}^h(J) \rightarrow S^h(J)$ be a finite-difference differentiation operator such that

$$D^h \varphi(\tau) = D^h \varphi_i = \begin{cases} \frac{\varphi_{i+1/2} - \varphi_{i-1/2}}{h_{i+1/2}} & \text{for all } \tau \in J_i \text{ and } i \in \llbracket 1, n-1 \rrbracket, \\ 0 & \text{for all } \tau \in J_i \text{ and } i = 0 \text{ or } i = n. \end{cases}$$

The proof is complete. \square

LEMMA 12. *For all $\psi \in W^{1,1}(J)$, the following formulas hold:*

$$(69) \quad \sigma^h D \psi_i = D^h \pi^h \psi_i = \frac{\pi^h \psi_{i+1/2} - \pi^h \psi_{i-1/2}}{h_{i+1/2}} \quad \text{for all } i \in \llbracket 1, n-1 \rrbracket,$$

$$(70) \quad \sigma^h D \psi_0 = \frac{\pi^h \psi_{1/2} - \psi(0)}{h_{1/2}}, \quad \sigma^h D \psi_n = \frac{\psi(\tau_*) - \pi^h \psi_{n-1/2}}{h_{n+1/2}}.$$

Proof. Integration by parts gives

$$\begin{aligned} \sigma^h D\psi_i &= h_{i+1/2}^{-1}(D\psi, e_i) = -h_{i+1/2}^{-1}(\psi, De_i) \\ &= h_{i+1/2}^{-1}(\pi^h \psi_{i+1/2} - \pi^h \psi_{i-1/2}) \quad \text{for all } i \in \llbracket 1, n-1 \rrbracket, \\ \sigma^h D\psi_0 &= h_{1/2}^{-1}(D\psi, e_0) = -h_{1/2}^{-1}(\psi(0) + (\psi, De_0)) = h_{1/2}^{-1}(\pi^h \psi_{1/2} - \psi(0)), \\ \sigma^h D\psi_n &= h_{n+1/2}^{-1}(D\psi, e_n) = h_{n+1/2}^{-1}(\psi(\tau_*) - (\psi, De_n)) = h_{n+1/2}^{-1}(\psi(\tau_*) - \pi^h \psi_{n-1/2}), \end{aligned}$$

and the proof is complete. \square

Let us introduce an operator $\Lambda^h : S^h(J) \rightarrow L^q(J)$, $q \in [1, \infty[$ by formula

$$(\Lambda^h \psi^h)(\tau) := \sum_{k=0}^n \mathcal{E}(|\tau - \tau_k|) \psi_k^h h_{k+1/2}.$$

LEMMA 13. *The following inequality holds:*

$$(71) \quad \|\Lambda^h \psi^h\|_{L^1} \leq \|\psi^h\|_{L^1} \quad \text{for all } \psi^h \in S^h(J).$$

Proof. In fact,

$$\|\Lambda^h \psi^h\|_{L^1} \leq \sum_{k=0}^n \int_0^{\tau_*} \mathcal{E}(|\tau - \tau_k|) d\tau |\psi_k^h| h_{k+1/2} \leq \sum_{k=0}^n |\psi_k^h| h_{k+1/2} = \|\psi^h\|_{L^1},$$

which proves the inequality. \square

LEMMA 14. *The following estimate holds:*

$$(72) \quad \|\Lambda^h \sigma^h - \Lambda\|_{L^1 \rightarrow L^1} \leq \omega_1(\mathcal{E}, h_{\max}).$$

Proof. Since

$$(\Lambda^h \sigma^h - \Lambda) \psi(\tau) = \sum_{k=0}^n \int_0^{\tau_*} (\mathcal{E}(|\tau - \tau_k|) - \mathcal{E}(|\tau - \tau'|)) e_k(\tau') \psi(\tau') d\tau',$$

then

$$\begin{aligned} \|(\Lambda^h \sigma^h - \Lambda) \psi\|_{L^1} &\leq \sum_{k=0}^n \int_0^{\tau_*} \|\mathcal{E}(|\cdot - \tau_k|) - \mathcal{E}(|\cdot - \tau'|)\|_{L^1} e_k(\tau') |\psi(\tau')| d\tau' \\ &\leq \omega_1(\mathcal{E}, h_{\max}) \int_0^{\tau_*} \sum_{k=0}^n e_k(\tau') |\psi(\tau')| d\tau' = \omega_1(\mathcal{E}, h_{\max}) \|\psi\|_{L^1}, \end{aligned}$$

and the estimate is proved. \square

LEMMA 15. *The following formulas hold:*

$$(73) \quad D\Lambda \psi^h = \Lambda^h D^h \psi^h + \psi_{1/2}^h \mathcal{E} - \psi_{n-1/2}^h \mathcal{E}_* \quad \text{for all } \psi^h \in S_{1/2}^h(J),$$

$$(74) \quad D\Lambda \pi^h \psi = \Lambda^h \sigma^h D\psi + \psi(0) \mathcal{E} - \psi(\tau_*) \mathcal{E}_* \quad \text{for all } \psi \in W^{1,1}(J).$$

Proof. Let $\psi^h \in S_{1/2}^h(J)$. Then

$$(\Lambda \psi^h)(\tau) = \sum_{k=1}^n \int_{\tau_{k-1}}^{\tau_k} \mathcal{E}(|\tau - \tau'|) d\tau' \psi_{k-1/2}^h = \sum_{k=1}^n \int_{\tau_{k-1}-\tau}^{\tau_k-\tau} \mathcal{E}(|s|) ds \psi_{k-1/2}^h.$$

Hence

$$\begin{aligned} (D\Lambda\psi^h)(\tau) &= \sum_{k=1}^n (\mathcal{E}(|\tau_{k-1} - \tau|) - \mathcal{E}(|\tau_k - \tau|))\psi_{k-1/2}^h \\ &= \sum_{k=1}^{n-1} \mathcal{E}(|\tau - \tau_k|) \frac{\psi_{k+1/2}^h - \psi_{k-1/2}^h}{h_{k+1/2}} h_{k+1/2} + \mathcal{E}(\tau)\psi_{1/2}^h - \mathcal{E}(\tau_n - \tau)\psi_{n-1/2}^h \\ &= (\Lambda^h D^h \psi^h)(\tau) + \psi_{1/2}^h \mathcal{E}(\tau) - \psi_{n-1/2}^h \mathcal{E}_*(\tau). \end{aligned}$$

Let $\psi \in W^{1,1}(J)$. Applying formula (73) to $\psi^h = \pi^h \psi$ and taking into account (69) and (70),

$$\begin{aligned} D\Lambda\pi^h\psi &= \Lambda^h\sigma^h D\psi - (\sigma^h D\psi)_0 \mathcal{E}h_{1/2} - (\sigma^h D\psi)_n \mathcal{E}_*h_{n+1/2} + \pi^h\psi_{1/2}\mathcal{E} - \pi^h\psi_{n-1/2}\mathcal{E}_* \\ &= \Lambda^h\sigma^h D\psi + \psi(0)\mathcal{E} - \psi(\tau_n)\mathcal{E}_*. \end{aligned}$$

This ends the proof. \square

THEOREM 12. *If $f \in W^{1,1}(J)$, then*

$$(75) \quad \text{var}_{\mathcal{J}} \pi^h \varepsilon^h \leq \gamma_0 \gamma_1 (\omega_1(\mathcal{E}, h_{\max}) + \bar{\omega}_1(\mathcal{E}, J^h)) \|f\|_{W^{1,1}},$$

$$(76) \quad \|D\bar{\varepsilon}^h\|_{L^1} \leq \gamma_0 \gamma_1 (\omega_1(\mathcal{E}, h_{\max}) + \bar{\omega}_1(\mathcal{E}, J^h)) \|f\|_{W^{1,1}}.$$

If $\Lambda f \in W^{1,1}$, then

$$(77) \quad \text{var}_{\mathcal{J}} \pi^h \bar{\varepsilon}^h \leq \gamma_1^2 (\omega_1(\mathcal{E}, h_{\max}) + \bar{\omega}_1(\mathcal{E}, J^h)) \|\Lambda f\|_{W^{1,1}},$$

$$(78) \quad \|D\hat{\varepsilon}^h\|_{L^1} \leq \gamma_1^2 (\omega_1(\mathcal{E}, h_{\max}) + \bar{\omega}_1(\mathcal{E}, J^h)) \|\Lambda f\|_{W^{1,1}}.$$

Proof. Applying operator D^h to (8) and taking into account (69) and (73), we obtain

$$\begin{aligned} (79) \quad D^h \varphi^h &= \varpi_0 \overset{\circ}{\partial}^h D\Lambda\varphi^h + D^h \pi^h f \\ &= \varpi_0 \overset{\circ}{\partial}^h [\Lambda^h D^h \varphi^h + \varphi_{1/2}^h \mathcal{E} - \varphi_{n-1/2}^h \mathcal{E}_*] + D^h \pi^h f. \end{aligned}$$

Applying operator $\overset{\circ}{\partial}^h$ to (21) and taking into account (69) and (70),

$$\begin{aligned} (80) \quad D^h \pi^h \varphi &= \varpi_0 \overset{\circ}{\partial}^h [\Lambda D\varphi + \varphi(0)\mathcal{E} - \varphi(\tau_n)\mathcal{E}_*] + D^h \pi^h f \\ &= \varpi_0 \overset{\circ}{\partial}^h [\Lambda^h D^h \pi^h \varphi + (\Lambda - \Lambda^h \sigma^h) D\varphi + \pi^h \varphi_{1/2} \mathcal{E} - \pi^h \varphi_{n-1/2} \mathcal{E}_*] \\ &\quad + D^h \pi^h f. \end{aligned}$$

Subtracting (81) from (80) we get

$$D^h \pi^h \varepsilon^h = \varpi_0 \overset{\circ}{\partial}^h \Lambda^h D^h \pi^h \varepsilon^h + \varpi_0 \overset{\circ}{\partial}^h [(\Lambda^h \sigma^h - \Lambda) D\varphi + \pi^h \varepsilon_{1/2}^h \mathcal{E} - \pi^h \varepsilon_{n-1/2}^h \mathcal{E}_*].$$

Using inequalities (68) and (71) we get

$$\|D^h \pi^h \varepsilon^h\|_{L^1} \leq \varpi_0 \|D^h \pi^h \varepsilon^h\|_{L^1} + \varpi_0 \|\Lambda^h \sigma^h - \Lambda\|_{L^1 \rightarrow L^1} \|D\varphi\|_{L^1} + \varpi_0 \|\pi^h \varepsilon^h\|_{L^\infty}.$$

From the equality

$$\|D^h \pi^h \varepsilon^h\|_{L^1} = \text{var}_{\mathcal{J}} \pi^h \varepsilon^h$$

and from the inequality (72), it follows that

$$\operatorname{var}_{\overline{J}} \pi^h \varepsilon^h \leq \gamma_1 (\omega_1(\mathcal{E}, h_{\max}) \|D\varphi\|_{L^1} + \|\pi^h \varepsilon^h\|_{L^\infty}).$$

From (45), it follows that

$$(81) \quad \|\pi^h \varepsilon^h\|_{L^\infty} \leq \gamma_0 \gamma_1 \overline{\omega}_1(\mathcal{E}, J^h) \|f\|_{L^\infty} \leq \gamma_0 \overline{\omega}_1(\mathcal{E}, J^h) \|f\|_{W^{1,1}}.$$

Now, using the estimate (20) with $p = 1$ we get (75).

Applying (74) to the Sloan approximation $\overline{\varphi}^h$, which solves $\overline{\varphi}^h = \varpi_0 \Lambda \pi^h \overline{\varphi}^h + f$, we get

$$D\overline{\varphi}^h = \varpi_0 \Lambda^h \sigma^h D\overline{\varphi}^h + \varpi_0 \overline{\varphi}^h(0) \mathcal{E} - \varpi_0 \overline{\varphi}^h(\tau_*) \mathcal{E}_* + Df.$$

Subtracting this equality from (19) we get

$$D\overline{\varepsilon}^h = \varpi_0 \Lambda^h \sigma^h D\overline{\varepsilon}^h + \varpi_0 (\sigma^h \Lambda^h - \Lambda) D\varphi + \varpi_0 \overline{\varepsilon}^h(0) \mathcal{E} - \varpi_0 \overline{\varepsilon}^h(\tau_*) \mathcal{E}_*,$$

and hence

$$\|D\overline{\varepsilon}^h\|_{L^1} \leq \gamma_1 (\omega_1(\mathcal{E}, h_{\max}) \|D\varphi\|_{L^1} + \|\overline{\varepsilon}^h\|_{L^\infty}).$$

It follows from (51) that

$$(82) \quad \|\overline{\varepsilon}^h\|_{C^0} \leq \gamma_0 \gamma_1 (\mathcal{E}, J^h) \|f\|_{C^0} \leq \gamma_1 \overline{\omega}_1(\mathcal{E}, J^h) \|f\|_{W^{1,1}}.$$

Now, using the estimate (20) with $p = 1$ we get (76).

Inequalities (77) and (78) follow from (75), (76) and Remark 1. \square

THEOREM 13. *If $f \in BV(\overline{J})$, then*

$$(83) \quad \|\pi^h \varepsilon^h\|_{BV} \leq \gamma_0 \gamma_1 (\omega_1(\mathcal{E}, h_{\max}) + 2\overline{\omega}_1(\mathcal{E}, J^h)) \|f\|_{BV},$$

$$(84) \quad \|\overline{\varepsilon}^h\|_{W^{1,1}} \leq \gamma_0 \gamma_1 (\omega_1(\mathcal{E}, h_{\max}) + 2\overline{\omega}_1(\mathcal{E}, J^h)) \|f\|_{BV}.$$

If $\Lambda f \in BV(\overline{J})$, then

$$(85) \quad \|\pi^h \overline{\varepsilon}^h\|_{BV} \leq \gamma_1^2 (\omega_1(\mathcal{E}, h_{\max}) + 2\overline{\omega}_1(\mathcal{E}, J^h)) \|\Lambda f\|_{BV},$$

$$(86) \quad \|\widehat{\varepsilon}^h\|_{W^{1,1}} \leq \gamma_1^2 (\omega_1(\mathcal{E}, h_{\max}) + 2\overline{\omega}_1(\mathcal{E}, J^h)) \|\Lambda f\|_{BV}.$$

Proof. Let f_δ be the average function defined in (24), and φ_δ solve

$$\varphi_\delta = \varpi_0 \Lambda \varphi_\delta + f_\delta.$$

Let φ_δ^h and $\overline{\varphi}_\delta^h$ be corresponding Galerkin and Sloan approximations to φ_δ :

$$\varphi_\delta^h = \varpi_0 \pi^h \Lambda \varphi_\delta^h + \pi^h f_\delta, \quad \overline{\varphi}_\delta^h = \varpi_0 \Lambda \pi^h \varphi_\delta^h + f_\delta.$$

Let $\varepsilon_\delta^h := \varphi_\delta^h - \varphi_\delta$ and $\overline{\varepsilon}_\delta^h := \overline{\varphi}_\delta^h - \overline{\varphi}_\delta$. As $\|f_\delta\|_{W^{1,1}} \leq \|f\|_{BV}$ (see (25)), it follows, from (75), (46) and (76), (52), that

$$(87) \quad \operatorname{var}_{\overline{J}} \pi^h \varepsilon_\delta^h \leq \gamma_0 \gamma_1 (\omega_1(\mathcal{E}, h_{\max}) + \overline{\omega}_1(\mathcal{E}, J^h)) \|f\|_{BV},$$

$$(88) \quad \|\pi^h \varepsilon_\delta^h\|_{L^\infty} \leq \gamma_0 \gamma_1 \overline{\omega}_1(\mathcal{E}, J^h) \|f\|_{BV},$$

$$(89) \quad \operatorname{var}_{\overline{J}} \overline{\varepsilon}_\delta^h \leq \gamma_0 \gamma_1 (\omega_1(\mathcal{E}, h_{\max}) + \overline{\omega}_1(\mathcal{E}, J^h)) \|f\|_{BV},$$

$$(90) \quad \|\overline{\varepsilon}_\delta^h\|_{L^\infty} \leq \gamma_0 \gamma_1 \overline{\omega}_1(\mathcal{E}, J^h) \|f\|_{BV}.$$

It is easy to see that $\varepsilon_\delta^h \rightarrow \varepsilon^h$ in $L^1(J)$, and that $\pi^h \varepsilon_\delta^h \rightarrow \pi^h \varepsilon^h$ in the finite dimensional space $S_{1/2}^h(J)$, as $\delta \rightarrow 0$. As a consequence, $\|\pi^h \varepsilon_\delta^h\|_{L^\infty} \rightarrow \|\pi^h \varepsilon^h\|_{L^\infty}$ and $\text{var}_{\mathcal{J}} \pi^h \varepsilon_\delta^h \rightarrow \text{var}_{\mathcal{J}} \pi^h \varepsilon^h$ and a limit transition in (87) and (88) gives (83).

It follows, from the formula $\bar{\varepsilon}_\delta^h = \varpi_0 \Lambda \varepsilon_\delta^h$, that $\bar{\varepsilon}_\delta^h \rightarrow \varpi_0 \Lambda \varepsilon^h$ in $C^0(\bar{J})$ as $\delta \rightarrow 0$. So, by the First Helli Theorem and the property $\bar{\varepsilon}^h \in W^{1,1}(J)$, (84) follows from (89) and (90).

Estimates (85) and (86) follow from (83), (84), and Remark 1. \square

As an immediate result, we have the following.

THEOREM 14. *Assume that \mathcal{E} satisfies (34) and (36).*

If $f \in BV(\bar{J})$, then

$$\|\pi^h \varepsilon^h\|_{BV} \leq 6\gamma_0 \gamma_1 h_{\max} \mathcal{E}(h_{\max}/2)(1 + o(1)) \|f\|_{BV},$$

$$\|\bar{\varepsilon}^h\|_{W^{1,1}} \leq 6\gamma_0 \gamma_1 h_{\max} \mathcal{E}(h_{\max}/2)(1 + o(1)) \|f\|_{BV}.$$

If $\Lambda f \in BV(\bar{J})$, then

$$\|\pi^h \tilde{\varepsilon}^h\|_{BV} \leq 6\gamma_1^2 h_{\max} \mathcal{E}(h_{\max}/2)(1 + o(1)) \|\Lambda f\|_{BV},$$

$$\|\hat{\varepsilon}^h\|_{W^{1,1}} \leq 6\gamma_1^2 h_{\max} \mathcal{E}(h_{\max}/2)(1 + o(1)) \|\Lambda f\|_{BV}.$$

7. Numerical evidence. The results exhibited in the Figures 1–4 are obtained for the heat transfer equation (3) with the following data:

Albedo: $\varpi_0 = 0.99$,

Optical depth: $\tau_* = 50$,

Source function: $f(\tau) = E_1(\tau)$,

Number of grid points: $n = 680$.

All three, φ , τ_* , and f , have physical meanings or interpretations. The albedo φ represents the reflectiveness of the stellar atmosphere, the optical depth τ_* is an

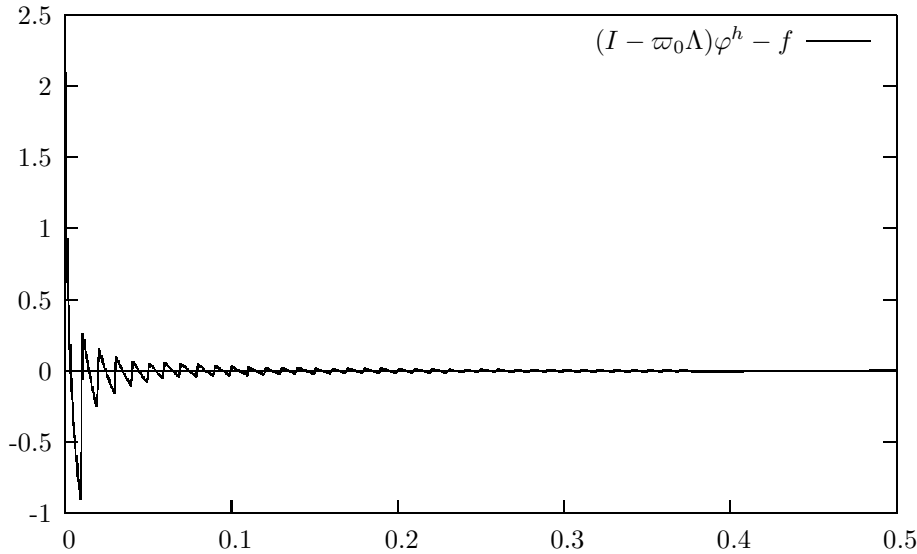
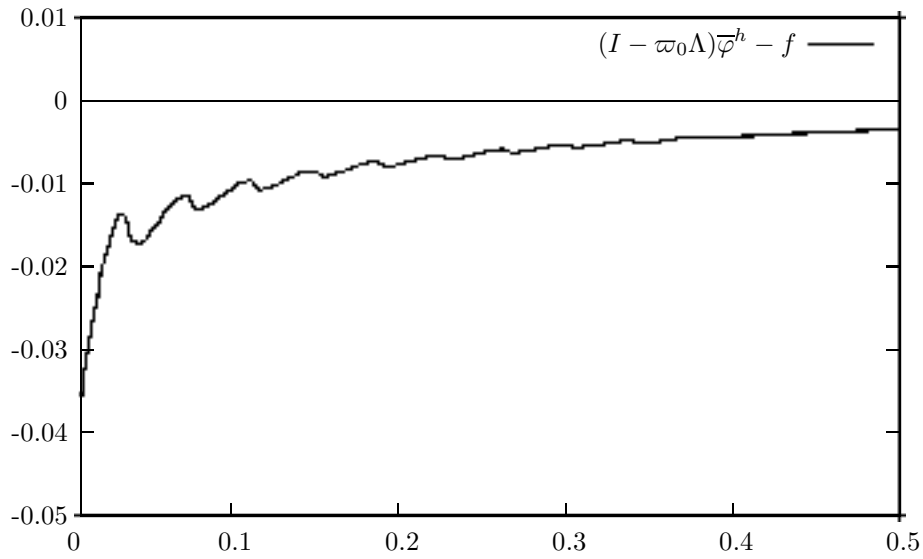
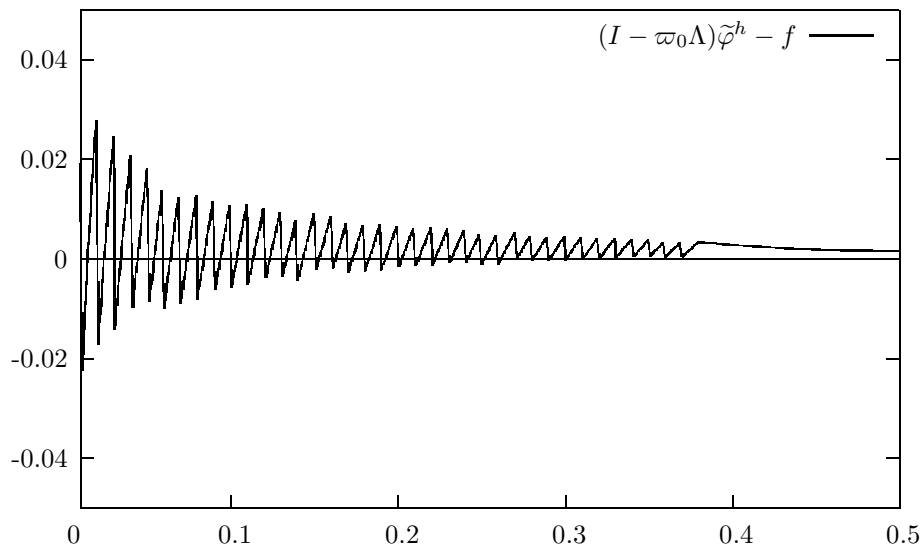


FIG. 1. Galerkin residual function.

FIG. 2. *Sloan residual function.*FIG. 3. *Kantorovich residual function.*

increasing function of the geometrical thickness of the atmosphere and the source function f represents an intense source of photons located on the surface of the atmosphere.

The grid is built with first 500 equally spaced points separated by 0.01 one from the next and last 180 equally spaced points separated by 0.25 one from the next.

Figures 1–4 show the residual functions corresponding to the Galerkin, Sloan, Kantorovich, and iterated Kantorovich approximations in the interval $[0, 0.5]$.

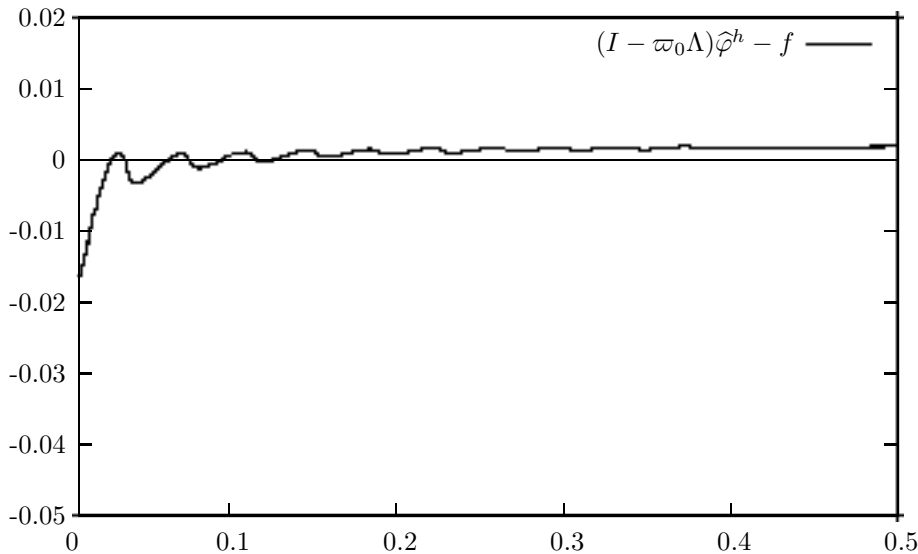


FIG. 4. Iterated Kantorovich residual function.

Acknowledgments. The authors are thankful to V.V. Dmitriev for his assistance in the preparation of Section 7. The author Andrey Amosov is thankful to the Laboratoire de Mathématiques de l'Université de Saint-Étienne (LaMUSE), France, for its hospitality as the paper was written when he was on leave at this team.

REFERENCES

- [1] M. ABRAMOVITZ AND I.A. STEGUN, *Handbook of Mathematical Functions*, Dover, New York, 1964.
- [2] M. AHUES, A. LARGILLIER, AND B.V. LIMAYE, *Spectral Computations with Bounded Operators*, CRC, Boca Raton, FL, 2001.
- [3] I.W. BUSBRIDGE, *The Mathematics of Radiative Transfer*, Cambridge University Press, Cambridge, UK, 1960.
- [4] G.A. CHANDLER, *Superconvergence for second kind integral equations*, in *The Application and Numerical Solution of Integral Equations*, R.S. Anderssen, F.R. de Hoog, and M.A. Lukas, eds., Alphen aan den Rijn: Sijthoff and Noordhoff, 1980, pp. 103–107.
- [5] S. CHANDRASEKAR, *Radiative Transfer*, Oxford Calderon Press, 1950.
- [6] I.G. GRAHAM, *Galerkin method for second kind integral equations with singularities*, *Math. Comput.*, 39 (1982), pp. 519–533.
- [7] G.C. HSIAO AND W.L. WENDLAND, *The Aubin-Nitsche lemma for integral equations*, *J. Integral Equations*, 3 (1981), pp. 299–315.
- [8] L.V. KANTOROVICH, *Functional analysis and applied mathematics*, *Usp. Mat. Nauk.*, 3 (1948), pp. 89–185 (in Russian). English translation: N.B.S. Report 1509 (1952).
- [9] H. KANEKO AND Y. XU, *Superconvergence of the iterated Galerkin methods for Hammerstein equations*, *SIAM J. Numer. Anal.*, 33 (1996), pp. 1048–1064.
- [10] V. KOURGANOFF, *Basic Methods in Transfer Problems*, Dover Publications, New York, 1963.
- [11] E. SCHOCK, *Über die Konvergenzgeschwindigkeit projektiver Verfahren*, *Math. Z.*, 120 (1971), pp. 148–156.
- [12] E. SCHOCK, *Über die Konvergenzgeschwindigkeit projektiver Verfahren II*, *Math. Z.*, 127 (1972), pp. 191–198.
- [13] E. SCHOCK, *Galerkin-like methods for equations of the second kind*, *J. Integral Equations*, 4 (1982), pp. 361–364.
- [14] E. SCHOCK, *Numerische Lösung Fredholmscher Integralgleichungen*, Lecture Notes, University of Kaiserslautern, 1982.

- [15] E. SCHOCK, *Arbitrarily slow convergence, uniform convergence and superconvergence of Galerkin-like methods*, IMA J. Numer. Anal., 5 (1985), pp. 153–160.
- [16] I.H. SLOAN, *Error analysis for a class of degenerate-kernel methods*, Numer. Math., 25 (1976), pp. 231–238.
- [17] I.H. SLOAN, *Improvement by iteration for compact operator equations*, Math. Comput., 30 (1976), pp. 758–764.
- [18] I.H. SLOAN, *Iterated Galerkin method for eigenvalue problems*, SIAM J. Numer. Anal., 13 (1976), pp. 753–764.
- [19] I.H. SLOAN, *Superconvergence and the Galerkin method for integral equations of the second kind*, in Treatment of Integral Equations by Numerical Methods, T.N. Christopher, Baker and G.F. Miller, eds., Academic Press, New York, 1982, pp. 197–206.
- [20] I.H. SLOAN, *Four variants of the Galerkin method for integral equations of the second kind*, IMA J. Numer. Anal., 4 (1984), pp. 9–17.
- [21] I.H. SLOAN, B.J. BURN, AND N. DATINER, *A new approach to the numerical solution of integral equations*, J. Comp. Phys., 18 (1975), pp. 92–105.
- [22] V.V. SOBOLEV, *A Treatise on Radiative Transfer*, D. Van Nostrand, Princeton, NJ, 1963.
- [23] A. SPENCE AND K.S. THOMAS, *On superconvergence properties of Galerkin's method for compact operator equations*, IMA J. Num. Analysis, 3 (1983), pp. 253–271.
- [24] M. THAMBAN NAIR AND R.S. ANDERSSSEN, *Superconvergence of modified projection method for integral equations of the second kind*, J. Integral Equations and Applications, 3 (1991), pp. 255–269.

THE DIRECT DISCONTINUOUS GALERKIN (DDG) METHODS FOR DIFFUSION PROBLEMS*

HAILIANG LIU[†] AND JUE YAN[†]

Abstract. A new discontinuous Galerkin finite element method for solving diffusion problems is introduced. Unlike the traditional local discontinuous Galerkin method, the scheme called the direct discontinuous Galerkin (DDG) method is based on the direct weak formulation for solutions of parabolic equations in each computational cell and lets cells communicate via the numerical flux \widehat{u}_x only. We propose a general numerical flux formula for the solution derivative, which is consistent and conservative; and we then introduce a concept of admissibility to identify a class of numerical fluxes so that the nonlinear stability for both one-dimensional (1D) and multidimensional problems are ensured. Furthermore, when applying the DDG scheme with admissible numerical flux to the 1D linear case, k th order accuracy in an energy norm is proven when using k th degree polynomials. The DDG method has the advantage of easier formulation and implementation and efficient computation of the solution. A series of numerical examples are presented to demonstrate the high order accuracy of the method. In particular, we study the numerical performance of the scheme with different admissible numerical fluxes.

Key words. diffusion, discontinuous Galerkin methods, stability, convergence rate, numerical trace

AMS subject classifications. 65M12, 65M60

DOI. 10.1137/080720255

1. Introduction. In this paper, we introduce a new discontinuous Galerkin (DG) method for solving nonlinear diffusion equations of the form

$$(1.1) \quad \partial_t U - \nabla \cdot (A(U)\nabla U) = 0, \quad \Omega \times (0, T),$$

where $\Omega \subset \mathbb{R}^d$, the matrix $A(U) = (a_{ij}(U))$ is symmetric and positive definite, and U is an unknown function of (x, t) .

The novelty of our method is to use the direct weak formulation for solutions of (1.1) in each computational cell and let cells communicate through a numerical trace of $A(U)\nabla U$ only. It is from this feature that the method proposed here derives its name: the direct DG (DDG) method. Here we carefully design a class of numerical fluxes in such a way that a stable and high order accurate DG method for the nonlinear diffusion equation (1.1) is achieved.

The DG method is a finite element method using a completely discontinuous piecewise polynomial space for the numerical solution and the test functions. A key ingredient of this method is the suitable design of the interelement boundary treatments (the so-called numerical fluxes) to obtain high order accurate and stable schemes. The DG method has been vigorously developed for hyperbolic problems since it was first introduced in 1973 by Reed and Hill [25] for neutron transport equations. A major development of the DG method is carried out by Cockburn, Shu, and collaborators in a series of papers [16, 15, 14, 11, 18] for nonlinear hyperbolic

*Received by the editors April 4, 2008; accepted for publication (in revised form) September 9, 2008; published electronically January 16, 2009.

<http://www.siam.org/journals/sinum/47-1/72025.html>

[†]Mathematics Department, Iowa State University, Ames, IA 50011 (hliu@iastate.edu, jyan@iastate.edu). The first author's research was supported by the National Science Foundation under grant DMS05-05975.

conservation laws. While it is being actively developed, the DG method has found rapid applications in many areas; we refer to [20, 13, 19] for further references.

However, the DG method when applied to diffusion problems encounters subtle difficulties, which can be illustrated by the simple one-dimensional (1D) heat equation

$$u_t - u_{xx} = 0.$$

Indeed, using this equation in [28], Shu illustrated some typical “pitfalls” in using the DG method for viscous terms. The DG method when applied to the heat equation formally leads to

$$(1.2) \quad \int_{I_j} u_t v + \int_{I_j} u_x v_x dx - \widehat{(u_x)}_{j+1/2} v_{j+1/2}^- + \widehat{(u_x)}_{j-1/2} v_{j-1/2}^+ = 0,$$

where both u and v are piecewise polynomials on each computational cell $I_j = (x_{j-1/2}, x_{j+1/2})$. Notice that u itself is discontinuous at cell interfaces; the formulation (1.2) even requires approximations of u_x at cell interfaces, which we call the numerical flux $\widehat{(u_x)}$!

A primary choice is the slope average $\widehat{(u_x)}_{j+1/2} = ((u_x)_{j+1/2}^- + (u_x)_{j+1/2}^+)/2$. But the scheme produces a completely incorrect, therefore inconsistent, solution; see Figure 1 (left). This is called “subtle inconsistency” by Shu in [28].

There are two ways to remedy this problem which were suggested in the literature. One is to rewrite the heat equation into a 1st order system and solve it with the DG method

$$u_t - q_x = 0, \quad q - u_x = 0.$$

Here both u and the auxiliary variable q are evolved in each computational cell. This method was originally proposed for the compressible Navier–Stokes equation by Bassi and Rebay [4]. Subsequently, a generalization called the local discontinuous Galerkin (LDG) method was introduced in [17] by Cockburn and Shu and further studied in [10, 7, 12, 8]. More recently, the LDG methods have been successfully extended to higher order partial differential equations; see, e.g., [32, 22, 31, 23].

Another one is to add extra cell boundary terms so that a weak stability property is ensured. The scheme thus takes the following form:

$$\begin{aligned} & \int_{I_j} u_t v + \int_{I_j} u_x v_x dx - \widehat{(u_x)}_{j+1/2} v_{j+1/2}^- + \widehat{(u_x)}_{j+1/2} v_{j-1/2}^+ \\ & - \frac{1}{2} (v_x)_{j+1/2}^- (u_{j+1/2}^+ - u_{j+1/2}^-) - \frac{1}{2} (v_x)_{j-1/2}^+ (u_{j-1/2}^+ - u_{j-1/2}^-) = 0, \end{aligned}$$

where again the slope average was chosen as the numerical flux. Such a method was introduced by Baumann and Oden [5]; see also Oden, Babuska, and Baumann [24]. This later scheme, once written into a primal formulation, is similar to a class of *interior penalty* methods, independently proposed and studied for elliptic and parabolic problems in the 1970s; see, e.g., [2, 3, 30]. Considering the similarities among the recently introduced DG methods, Arnold et al. [1] have set the existing DG methods into a unified framework with a systematic analysis of these methods via linear elliptic problems. Another framework using both the equation in each element and continuity relations across interfaces was recently analyzed in [6].

Notice that the above two ways suggest modifications mainly on the scheme formulation but not on the numerical flux $\widehat{u_x}$. The main goal of this work is to propose

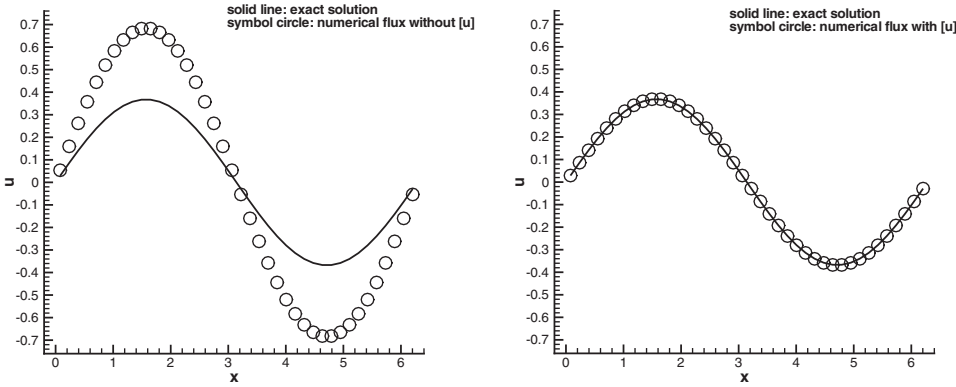


FIG. 1. On the left $\widehat{u}_x = \overline{u_x}$ and on the right $\widehat{u}_x = \frac{[u]}{\Delta x} + \overline{u_x}$ at $t = 1$, mesh size $N = 40$. p^1 polynomial approximation.

a path which sticks to the direct weak formulation (1.2) but with new choices of numerical flux \widehat{u}_x to obtain a stable and accurate DG scheme. More precisely, the heart of the DDG method is to use the direct weak formulation for parabolic equations and let cells communicate via the numerical flux \widehat{u}_x . A key observation is that the jump of the function itself relative to the mesh size, when numerically measuring slopes of a discontinuous function, plays an essential role. For example, for the piecewise constant approximation ($k = 0$), the choice of

$$\widehat{u}_x = \frac{u^+ - u^-}{\Delta x}$$

leads to the standard central finite difference scheme. When we use the numerical flux

$$\widehat{u}_x = \frac{u^+ - u^-}{\Delta x} + \frac{1}{2} (u_x^+ + u_x^-),$$

the resulting scheme with piecewise linear approximation is found of 2nd order accurate and of course gives the correct solution; see Figure 1 (right).

However, the trace of the solution derivative under a diffusion process is rather subtle. From the PDE point of view, jumps of all even order derivatives as well as the average of odd order derivatives all contribute to the trace of the solution derivative. We propose a general numerical flux formula, which is consistent with the solution gradient and conservative. The form of the numerical flux is motivated by an exact trace formulation derived from solving the heat equation with smooth initial data having only one discontinuous point.

We then introduce a concept of admissibility for numerical fluxes. The admissibility condition serves as a criterion for selecting suitable numerical fluxes to guarantee nonlinear stability of the DDG method and corresponding error estimates. Indeed in the linear case, the convergence rate of order $(\Delta x)^k$ for the error in a parabolic energy norm $L^\infty(0, T; L^2) \cap L^2(0, T; H^1)$ is obtained when p^k polynomials are used.

In this paper, we restrict ourselves to diffusion problems with periodic boundary conditions. We shall display the most distinctive features of the DDG method using as simple a setting as possible. This paper is organized as follows. In section 2, we introduce the DDG methods for the 1D problems. For this model problem, the main

idea of how to devise the method is presented. The nonlinear stability and error estimate for the linear case are discussed in section 3. In section 4, we extend the DDG methods to multidimensional problems in which U is a scalar and $A = (a_{ij})_{d \times d}$ is a positive and semidefinite matrix. The nonlinear stability is established. Finally in section 5, we present a series of numerical results to validate our DDG methods. For completeness some projection properties and a trace formula for the heat equation are presented in the appendix.

Finally, we note that formulating a DG method without rewriting the equation into a 1st order system as in the LDG method was also explored in three more recent works [29, 21, 9]. But they all rely on repeated integration by parts for the diffusion term so that the interface values can be imposed for both the solution and its derivatives. In contrast, we use the standard weak formulation for parabolic equations with integration by parts only once, and the interface continuity is enforced by defining suitable interface values of the solution derivative only.

2. 1D diffusion process. In this section, we introduce the formulation of the DDG method for the simple 1D case

$$(2.1) \quad U_t - (a(U)U_x)_x = 0 \quad \text{in } (0, 1) \times (0, T)$$

subject to initial data

$$(2.2) \quad U(x, 0) = U_0(x) \quad \text{on } (0, 1)$$

and periodic boundary conditions.

The unknown function U is a scalar, and we assume the diffusion coefficient a to be a nonnegative function of U . The DDG method is constructed upon the direct weak formulation of parabolic equations.

First, we partition the domain $(0, 1)$ by grid points $0 = x_{1/2} < x_{3/2} < \dots < x_{N+1/2} = 1$; we define the mesh $\{I_j = (x_{j-1/2}, x_{j+1/2}), \quad j = 1 \dots N\}$ and set the mesh size $\Delta x_j = x_{j+1/2} - x_{j-1/2}$. Furthermore, we denote $\Delta x = \max_{1 \leq j \leq N} \Delta x_j$. We seek an approximation u to U such that for any time $t \in [0, T]$, $u \in \mathbb{V}_{\Delta x}$,

$$\mathbb{V}_{\Delta x} := \{v \in L^2(0, 1) : v|_{I_j} \in P^k(I_j), \quad j = 1, \dots, N\},$$

where $P^k(I_j)$ denotes the space of polynomials in I_j with degree at most k . We now formulate our scheme for (2.1) and describe guidelines for defining numerical fluxes.

2.1. Formulation of the scheme. Denote the flux $h := h(U, U_x) = a(U)U_x$. Let U be the exact solution of the underlying problem. Multiply (2.1) by any smooth function $V \in H^1(0, 1)$, integrate on I_j , and have integration by parts to obtain the following equations:

$$(2.3) \quad \int_{I_j} U_t V dx - h_{j+1/2} V_{j+1/2} + h_{j-1/2} V_{j-1/2} + \int_{I_j} a(U)U_x V_x dx = 0,$$

$$(2.4) \quad \int_{I_j} U(x, 0) V(x) dx = \int_{I_j} U_0(x) V(x) dx.$$

Here the time derivative is to be understood in the weak sense, and $h_{j\pm 1/2}$ and $V_{j\pm 1/2}$ denote values of h and V at $x = x_{j\pm 1/2}$, respectively.

Next we replace the smooth function V by any test function $v \in \mathbb{V}_{\Delta x}$ and the exact solution U by the numerical approximate solution u . The flux $h(U, U_x)$ is replaced by the numerical flux \hat{h} that will be defined later.

Thus the approximate solution given by the DDG method is defined as

$$(2.5) \quad \int_{I_j} u_t v dx - \widehat{h}_{j+1/2} v_{j+1/2}^- + \widehat{h}_{j-1/2} v_{j-1/2}^+ + \int_{I_j} a(u) u_x v_x dx = 0,$$

$$(2.6) \quad \int_{I_j} u(x, 0) v(x) dx = \int_{I_j} U_0(x) v(x) dx.$$

Note that u is a well defined function since there are as many equations per element as unknowns. The integral $\int_{I_j} a(u) u_x v_x dx$ could be either computed exactly or approximated by using suitable numerical quadratures. Thus, to complete the DG space discretization, we only have to define the numerical flux \widehat{h} .

2.2. The numerical flux. Crucial for the stability as well as for the accuracy of the DDG method is the choice of the numerical flux \widehat{h} . To define it, we adopt the following notations:

$$u^\pm(t) = u(x_{j+1/2}^\pm, t), \quad [u] = u^+ - u^-, \quad \overline{u} = \frac{u^+ + u^-}{2}.$$

The numerical flux \widehat{h} defined at the cell interface $x_{j+1/2}$ is chosen in such a way that it is a function depending only on the left and right polynomials and that it (i) is consistent with $h = b(u)_x = a(u)u_x$, where $b(u) = \int^u a(s) ds$ when u is smooth; (ii) is conservative in the sense of \widehat{h} being single valued on $x_{j+1/2}$ and

$$\frac{d}{dt} \int_{I_j} u dx = \widehat{h}_{j+1/2} - \widehat{h}_{j-1/2};$$

(iii) ensures the L^2 -stability; and (iv) enforces the high order accuracy of the method.

Motivated by the trace formula of the solution derivative of the heat equation, see (7.3) in the appendix, we propose the following general format of the numerical flux:

$$(2.7) \quad \widehat{h} = D_x b(u) = \beta_0 \frac{[b(u)]}{\Delta x} + \overline{b(u)_x} + \sum_{m=1}^{\lfloor k/2 \rfloor} \beta_m (\Delta x)^{2m-1} [\partial_x^{2m} b(u)],$$

where k is the highest degree of polynomials in two adjacent computational cells and $\lfloor \cdot \rfloor$ is the floor function. Note here in (2.7) and in what follows that for nonuniform mesh Δx should be replaced by $(\Delta x_j + \Delta x_{j+1})/2$ and for uniform mesh $\Delta x = 1/N$.

The numerical flux \widehat{h} , which is an approximation of $b(U)_x$ at the cell interface, involves the average $\overline{b(u)_x}$ and the jumps of even order derivatives of $b(u)$, $[\partial_x^{2m} b(u)]$, up to $m = \lfloor k/2 \rfloor$. For example, with the p^3 polynomial approximation we need to determine suitable β_0 and β_1 to define the numerical flux

$$\widehat{h} = D_x b(u) = \beta_0 \frac{[b(u)]}{\Delta x} + \overline{b(u)_x} + \beta_1 \Delta x [b(u)_{xx}].$$

It is clear for any choice of β_i 's that the numerical flux defined in (2.7) is consistent and conservative. As is known, the underlying solution for the heat equation is smooth, and thus jumps of discrete solutions across cell interfaces have to be properly controlled so that continuities can be enforced at least in a weak sense.

To ensure stability and enhance accuracy and more importantly to measure the goodness of the choice of β_i 's, we introduce a notion of admissibility for numerical fluxes as follows.

DEFINITION 2.1 (admissibility). *We call a numerical flux \widehat{h} of the form (2.7) admissible if there exists a $\gamma \in (0, 1)$ and $\alpha > 0$ such that*

$$(2.8) \quad \gamma \sum_{j=1}^N \int_{I_j} a(u)u_x^2(x, t)dx + \sum_{j=1}^N \widehat{h}_{j+1/2}[u]_{j+1/2} \geq \alpha \sum_{j=1}^N \frac{([b(u)][u])_{j+1/2}}{\Delta x}$$

holds for any piecewise polynomials of degree k , i.e., $u \in \mathbb{V}_{\Delta x}$.

It is shown in the next section that for any admissible flux the DDG scheme is nonlinear stable and has k th order accuracy in an energy norm when using p^k polynomials for linear problems. We note that for error analysis $\alpha > 0$ plays an essential role in controlling the total jumps across cell interfaces.

We now discuss some principles for finding β_i 's. To simplify the presentation we restrict our discussions to the linear case with $\widehat{h} = D_x u$.

For the piecewise constant approximation, $k = 0$, the numerical flux (2.7) reduces to

$$\widehat{u}_x = D_x u = \beta_0 \frac{[u]}{\Delta x}.$$

Clearly we should take $\beta_0 = 1$, for which the DDG scheme is consistent with the central finite difference scheme. Note that $\beta_0 \neq 1$ is admissible but gives $O(1)$ error.

For the piecewise linear approximation, $k = 1$, the numerical flux (2.7) with $\beta_0 = 1$ becomes

$$(2.9) \quad D_x u = \frac{[u]}{\Delta x} + \overline{u}_x.$$

This can be easily verified to be admissible with $\alpha = 1/2$ and $\gamma = 1/2$. The corresponding DDG scheme is of 2nd order as observed numerically in section 5.

We can now prove that (2.9), with possibly an additional amount of $[u]/\Delta x$, is admissible for polynomial approximations of any degree, even for nonlinear diffusion.

THEOREM 2.1. *Consider the 1D diffusion with $a(u) \geq \delta > 0$. The numerical flux*

$$(2.10) \quad D_x u = \beta_0 \frac{[b(u)]}{\Delta x} + \overline{u}_x$$

is admissible for any piecewise polynomial of degree $k \geq 0$ provided β_0 is suitably large.

Proof. It is sufficient to select β_0 so that the underlying flux is admissible locally around each cell, i.e.,

$$\gamma \int_{I_j} a(u)u_x^2 dx + D_x u[u] \geq \alpha [b(u)][u]/\Delta x,$$

which, when combined with (2.10), can be rewritten as

$$\gamma \Delta x \int_{I_j} a(u)u_x^2 dx + \overline{u}_x [u] \Delta x + (\beta_0 - \alpha) \frac{[b(u)]}{[u]} [u]^2 \geq 0.$$

Note for $k = 0$, $\beta(0) = 1$ is admissible for $\alpha \leq 1$. From $a(u) \geq \delta$ we have $\frac{[b(u)]}{[u]} \geq \delta$. Thus the above inequality is ensured to hold for all $u|_{I_j} \in P^k(I_j)$ and therefore for all $[u]$, provided

$$(\bar{u}_x \Delta x)^2 - 4(\beta_0 - \alpha)\gamma \Delta x \delta \int_{I_j} u_x^2 dx \leq 0.$$

Summing this inequality over all indexes $j \in N$ we have

$$\beta_0 \geq \alpha + \frac{1}{4\gamma\delta} \frac{\Delta x \sum_j \bar{u}_x^2}{\sum_j \int_{I_j} a(u) u_x^2 dx}.$$

Maximizing the right side over all $u|_{I_j} \in P^k(I_j)$ we obtain

$$\beta_0 \geq \alpha + \frac{M_k}{4\gamma\delta^2},$$

where

$$M_k = \max_{u \in P^k(I_j)} \frac{\Delta x \sum_j \bar{u}_x^2}{\sum_j \int_{I_j} u_x^2 dx}.$$

For example, when $a(u) = 1$, $M_0 = 0$, $M_1 = 1$, $M_2 = 3$, etc. □

Numerical experiments show that the scheme with numerical flux (2.10) achieves $(k + 1)$ th order accuracy if k is odd but k th order accuracy if k is even, as long as β_0 is chosen above a critical value $\beta^* \sim M_k$ (to guarantee the scheme stability). The scheme accuracy is not sensitive to the choice of β_0 , though the critical value β^* needs to be larger as k increases.

In order to gain the $(k + 1)$ th order accuracy when k is even it is necessary to use higher order derivatives within our DDG framework. We consider exploring higher order approximations. The idea is to construct a higher order polynomial $\tilde{p}(x) \in P^{k+1}(I_j \cup I_{j+1})$ across the interface by interpolating at sample points in two neighboring cells. There are $[k/2] + 1$ pairs of points symmetrically sampled on each side of the underlying interface. Then the numerical flux can be defined as

$$(2.11) \quad D_x u = \partial_x \tilde{p}(x)|_{x_{j+1/2}}.$$

For $k = 2, 3$ we explore the Stirling interpolation formula based on four symmetric points

$$x_{j+1/2} \pm \frac{1}{2}h, \quad x_{j+1/2} \pm h, \quad 0 < h \leq \Delta x,$$

leading to a unique 3rd order polynomial, whose derivative when evaluated at the cell interface $x_{j+1/2}$ gives

$$(2.12) \quad D_x u = \frac{7}{6} \frac{[u]}{h} + \bar{u}_x + \frac{h}{12} [u_{xx}].$$

For p^2 and p^3 polynomials, the numerical flux (2.12) with $h = \Delta x$ enables us to obtain the optimal 3rd and 4th orders of accuracy, respectively. This suggests that the step used in the Stirling interpolation spans exactly the full computational cell on each

side, no more and no less; it is also unbiased. Of course, for nonuniform mesh, Δx needs to be understood as $(\Delta x_j + \Delta x_{j+1})/2$.

Here we note yet another way to select β_1 for $k \leq 2$ based on the exact trace formula (7.3), i.e.,

$$u_x(0, t) = \frac{1}{\sqrt{4\pi t}}[u] + \bar{u}_x + \sqrt{\frac{t}{\pi}}[u_{xx}], \quad 0 < t < \Delta t.$$

Considering the parabolic scaling, the correct mesh ratio should be $t \sim (\Delta x)^2$. Therefore setting $t = (\eta\Delta x)^2$ we obtain the following numerical flux:

$$D_x u = \frac{1}{\sqrt{4\pi} \eta \Delta x} [u] + \bar{u}_x + \frac{\eta \Delta x}{\sqrt{\pi}} [u_{xx}].$$

In section 5, we carry out numerical experiments for these η -schemes, and the choice $\eta = \sqrt{\pi}/12$, i.e., again $\beta_1 = 1/12$, gives the best performance, both in the absolute error and the order of the scheme.

In summary for $p^k, k = 0, \dots, 3$, we advocate the DDG scheme with the following numerical flux:

$$(2.13) \quad D_x u = \frac{[u]}{\Delta x} + \bar{u}_x + \frac{\Delta x}{12} [u_{xx}].$$

For p^k with $k \geq 4$ we employ the simple flux (2.10). It is interesting to note that for the p^2 case the coefficient $\beta_1 = 1/12$ is indeed important, but the β_0 is less important in the sense that with other choices of β_0 3rd order accuracy can also be achieved. In comparison, for the p_0 case $\beta_0 = 1$ is important.

Remark 2.1. The recipe given in (2.11) leads to a class of admissible numerical fluxes (2.12). But numerically only the flux with $h = \Delta x$ delivers the optimal L^2 accuracy for P^2 element, which is also the case for both nonuniform meshes in the 1D setting and hypercube partitions in multidimensions; for the latter see (5.10) in Example 5.5. This fact is further illustrated in Example 5.4 when the equation is nonlinear.

2.3. Time discretization. Up to now, we have taken the method of lines approach and have left t continuous. For time discretization we can use total variation diminishing (TVD) high order Runge–Kutta methods [27, 26] to solve the method of lines ODE

$$(2.14) \quad u_t = L(u).$$

The 3rd order TVD Runge–Kutta method that we use in this paper is given by

$$\begin{aligned} u^{(1)} &= u^n + \Delta t L(u^n), \\ u^{(2)} &= \frac{3}{4}u^n + \frac{1}{4}u^{(1)} + \frac{1}{4}\Delta t L(u^{(1)}), \\ u^{n+1} &= \frac{1}{3}u^n + \frac{2}{3}u^{(2)} + \frac{2}{3}\Delta t L(u^{(2)}). \end{aligned}$$

3. The nonlinear stability and error estimates.

3.1. The nonlinear stability. We first review the stability property for the continuous problem. We let $U \in L^2$ be a smooth solution to the initial value problem (2.1)–(2.2). Set $V = U$ in the weak formulation, and integrate over $[0, T]$; we have the following energy identity:

$$\frac{1}{2} \int_0^1 U^2(x, T) dx + \int_0^T \int_0^1 a(U) U_x^2 dx d\tau = \frac{1}{2} \int_0^1 U_0^2 dx \quad \forall T > 0.$$

We say the DDG scheme is L^2 stable if the numerical solution $u(x, t)$ satisfies

$$\int_0^1 u^2(x, T) dx \leq \int_0^1 U_0^2 dx.$$

In fact, the numerical solution defined by our DDG scheme (2.5) and (2.6) not only satisfies this stability property but also has total control on all jumps crossing cell interfaces $\{x_{j+1/2}\}_{j=1}^N$ due to the admissibility of the numerical flux.

THEOREM 3.1 (energy stability). *Consider the DDG scheme (2.5) and (2.6) with numerical flux (2.7). If the numerical flux is admissible as described in (2.8), then*

$$\begin{aligned} & \frac{1}{2} \int_0^1 u^2(x, T) dx + (1 - \gamma) \int_0^T \sum_{j=1}^N \int_{I_j} a(u) u_x^2(x, t) dx dt \\ (3.1) \quad & + \alpha \int_0^T \sum_{j=1}^N \frac{[b(u)]}{\Delta x} [u] dt \leq \frac{1}{2} \int_0^1 U_0^2(x) dx. \end{aligned}$$

Proof. Setting $v = u$ in (2.5), we have

$$\frac{d}{dt} \int_{I_j} \frac{u^2}{2} dx + \int_{I_j} a(u) u_x^2 dx - \widehat{h}_{j+1/2} u_{j+1/2}^- + \widehat{h}_{j-1/2} u_{j-1/2}^+ = 0.$$

Summation over $j = 1, 2 \dots N$ and integration with respect to t over $[0, T]$ leads to

$$\begin{aligned} & \frac{1}{2} \int_0^1 u^2(x, T) dx + \int_0^T \sum_{j=1}^N \int_{I_j} a(u) u_x^2(x, t) dx dt \\ (3.2) \quad & + \int_0^T \sum_{j=1}^N \widehat{h}_{j+1/2} [u]_{j+1/2} dt = \frac{1}{2} \int_0^1 u^2(x, 0) dx. \end{aligned}$$

From the admissible condition (2.8) of the numerical flux $\widehat{h}_{j+1/2}$ defined in (2.7) it follows that

$$\begin{aligned} & \int_0^T \sum_{j=1}^N \widehat{h}_{j+1/2} [u]_{j+1/2} dt = \int_0^T \sum_{j=1}^N \left(\widehat{b(u)_x} [u] \right)_{j+1/2} dt \\ (3.3) \quad & \geq \alpha \int_0^T \sum_{j=1}^N \frac{[b(u)]}{\Delta x} [u] dt - \gamma \int_0^T \sum_{j=1}^N \int_{I_j} a(u) u_x^2(x, t) dx dt. \end{aligned}$$

Finally, we note that (2.6) with $v(x) = u(x, 0)$ gives

$$(3.4) \quad \int_0^1 u^2(x, 0)dx \leq \int_0^1 U_0^2(x)dx.$$

Insertion of (3.3) and (3.4) into (3.2) leads to the desired stability estimate (3.1). \square

Notice that the usual L^2 -stability follows from such a stability estimate (3.1) since the term $[b(u)][u] = a(u^*)[u]^2$ remains nonnegative for any jumps $[u]$.

3.2. Error estimates. Now we turn to the question of the quality of the approximate solution defined by the DDG method. In the linear case $a(u) = 1$, from the above stability result and from the approximation properties of the finite element space $\mathbb{V}_{\Delta x}$, we can estimate the error $e := u - U$ between the exact solution U and the numerical solution u . Inspired by the stability estimate (3.1) we introduce the following energy norm to measure the solutions and the error:

$$(3.5) \quad |||v(\cdot, t)||| := \left(\int_0^1 v^2 dx + (1 - \gamma) \int_0^t \sum_{j=1}^N \int_{I_j} v_x^2 dx d\tau + \alpha \int_0^t \sum_{j=1}^N \frac{[v]^2}{\Delta x} d\tau \right)^{1/2},$$

with $\gamma \in (0, 1)$ and $\alpha > 0$. From the stability analysis and the smoothness of the exact solution U , we reformulate stability estimates for both the exact solution and the numerical solution in terms of the norm $|||(\cdot, T)|||$:

$$|||U(\cdot, T)||| \leq |||U(\cdot, 0)|||, \quad |||u(\cdot, T)||| \leq |||U(\cdot, 0)|||.$$

This section is devoted to the proof of the following error estimate.

THEOREM 3.2 (error estimate). *Let U be the exact solution and e be the error between the exact solution and the numerical solution by the DDG method with numerical flux (2.7). If the numerical flux is admissible (2.8), then the energy norm of the error satisfies the inequality*

$$(3.6) \quad |||e(\cdot, T)||| \leq C |||\partial_x^{k+1} U(\cdot, T)||| (\Delta x)^k,$$

where $C = C(k, \gamma, \alpha)$ is a constant depending on k, γ , and α but is independent of U and Δx .

Remark 3.1. The error estimates are optimal in Δx for smooth solutions. For initial data in $H^{k+1}(0, 1)$ we can simply replace $|||\partial_x^{k+1} U(\cdot, T)|||$ by $|U_0|_{k+1}$ since for parabolic problem $U_t = U_{xx}$ we have

$$(3.7) \quad \frac{1}{2} \int_0^1 |\partial_x^{k+1} U(x, T)|^2 dx + \int_0^T \int_0^1 |\partial_x^{k+2} U(x, T)|^2 dx \leq \frac{1}{2} \int_0^1 |\partial_x^{k+1} U_0(x)|^2 dx,$$

which holds for solution U with initial data $U_0 \in H^{k+1}(0, 1)$.

Remark 3.2. The k th order energy error (3.6) does not automatically imply a $(k + 1)$ th order L^2 error estimate unless the scheme is adjoint-consistent; see, e.g., [1]. The inclusion of jumps of higher order derivatives in numerical flux in this paper is intended to restore the optimal L^2 error.

Let \mathbb{P} be the L^2 projection operator from $H^1(0, 1)$ to the finite element space $\mathbb{V}_{\Delta x}$, which is defined as the only polynomial $\mathbb{P}(U)(x)$ in $\mathbb{V}_{\Delta x}$ such that

$$\int_{I_j} (\mathbb{P}(U)(x) - U(x))v(x)dx = 0 \quad \forall v \in \mathbb{V}_{\Delta x}.$$

Note by (2.6) and the above L^2 projection definition we have that $u(x, 0) = \mathbb{P}(U_0)$.

To estimate $e = u - U$, we rewrite the error as

$$(3.8) \quad e = u - \mathbb{P}(U) + \mathbb{P}(U) - U = \mathbb{P}(e) - (U - \mathbb{P}(U)).$$

Thus we have

$$(3.9) \quad |||e(\cdot, T)||| \leq |||\mathbb{P}(e)(\cdot, T)||| + |||(U - \mathbb{P}(U))(\cdot, T)|||.$$

It suffices to estimate the two terms on the right. The projection properties are essentially used and summarized in the following auxiliary lemma. The proof of the lemma is based on Bramble–Hilbert lemma 7, and an extended discussion is postponed in the appendix.

LEMMA 3.1 (L^2 projection properties). *Let $U \in H^{s+1}(I_j)$ for $j = 1, \dots, N$ and $s \geq 0$. Then we have the following estimates:*

1. $|\mathbb{P}(U) - U|_{m, I_j} \leq c_k (\Delta x)^{(\min\{k, s\} + 1 - m)} |U|_{s+1, I_j}, \quad m \leq k + 1.$
2. $|\partial_x^m (\mathbb{P}(U) - U)_{x_{j+1/2}}| \leq c_k (\Delta x)^{(\min\{k, s\} + 1/2 - m)} |U|_{s+1, I_{j+1/2}}, \quad m \leq k + 1/2,$
 where $m \geq 0$ is an integer, $I_{j+1/2} = I_j \cup I_{j+1}$, and constant c_k depends on k but is independent of I_j and U ; $|\cdot|_{m, I_j}$ denotes the seminorm of $H^m(I_j)$.

These basic estimates enable us to prove the following lemma.

LEMMA 3.2. *Let U be the smooth exact solution, and for any function v the $D_x v$ at the cell interface $x_{j+1/2}$ is defined by*

$$(3.10) \quad D_x v = \bar{v}_x + \sum_{m=0}^{\lfloor k/2 \rfloor} \beta_m (\Delta x)^{2m-1} [\partial_x^{2m} v].$$

Then we have

- (i) *projection error*

$$|||(U - \mathbb{P}(U))(\cdot, T)||| \leq C |||\partial_x^{k+1} U||| (\Delta x)^k,$$

- (ii) *trace error*

$$\int_0^T \sum_{j=1}^N (D_x(U - \mathbb{P}(U)))_{j+1/2}^2(t) dt \leq C |||\partial_x^{k+1} U|||^2 (\Delta x)^{2k-1}.$$

Proof. (i) Apply the estimates in Lemma 3.1 to $|||(U - \mathbb{P}(U))(\cdot, T)|||^2$ to obtain

$$\begin{aligned} & \sum_{j=1}^N |\mathbb{P}(U) - U|_{0, I_j}^2 + (1 - \gamma) \int_0^T \sum_{j=1}^N |(\mathbb{P}(U) - U)|_{1, I_j}^2 dt + \alpha \int_0^T \sum_{j=1}^N \frac{[\mathbb{P}(U) - U]^2}{\Delta x} dt \\ & \leq C_k \left((\Delta x)^{2k+2} |U|_{k+1, [0,1]}^2 + (\Delta x)^{2k} \int_0^T |U(\cdot, t)|_{k+2, [0,1]}^2 dt \right). \end{aligned}$$

Thus the estimate in (i) is ensured.

(ii) Applying the estimates in Lemma 3.1 to the expression (3.10) with $v = U - \mathbb{P}(U)$ we have

$$\begin{aligned} & \sum_{j=1}^N (D_x(U - \mathbb{P}(U)))_{j+1/2}^2 \\ & \leq C_k \sum_{j=1}^N \left\{ (\Delta x)^{2k-1} |U|_{k+2, I_j}^2 + \sum_{m=0}^{\lfloor k/2 \rfloor} (\Delta x)^{4m-2} (\Delta x)^{2k+1-4m} |U|_{k+2, I_{j+1/2}}^2 \right\} \\ & \leq C (\Delta x)^{2k-1} |U|_{k+2, [0,1]}^2. \end{aligned}$$

This gives the estimate (ii). The proof is thus complete. \square

To finish the estimate of e in (3.9), it remains to estimate $\mathbb{P}(e)$, which favorably lies in $\mathbb{V}_{\Delta x}$.

LEMMA 3.3. *We have*

$$\|\mathbb{P}(e)(\cdot, T)\|^2 \leq \frac{1}{(1-\gamma)^2} \|(U - \mathbb{P}(U))(\cdot, T)\|^2 + \frac{\Delta x}{\alpha} \int_0^T \sum_{j=1}^N (D_x(U - \mathbb{P}(U)))_{j+1/2}^2 dt.$$

A combination of Lemmas 3.2 and 3.3 with the inequality (3.9) yields the desired estimate (3.6), which completes the proof of Theorem 4.1.

We now conclude this section by presenting a detailed proof of Lemma 3.3.

Proof of Lemma 3.3. First, we define a bilinear form $\mathbb{B}(w, v)$ as

$$(3.11) \quad \mathbb{B}(w, v) = \int_0^T \int_0^1 w_t(x, t)v(x, t)dx + \int_0^T \sum_{j=1}^N \int_{I_j} w_x(x, t)v_x(x, t)dxdt + \Theta(T, \widehat{w}_x, v)$$

for any $v \in \mathbb{V}_{\Delta x}$, and

$$(3.12) \quad \Theta(T, \widehat{w}_x, v) = \int_0^T \sum_{j=1}^N \left((\widehat{w}_x)_{j+1/2} [v]_{j+1/2} \right) dt.$$

By the definition of DDG scheme (2.5), we have $\mathbb{B}(u, v) = 0 \forall v \in \mathbb{V}_{\Delta x}$. Exact solution $U(x, t)$ also satisfies $\mathbb{B}(U, v) = 0 \forall v \in \mathbb{V}_{\Delta x}$, and then we have

$$\mathbb{B}(e, v) = \mathbb{B}(u - U, v) = 0.$$

This equality when combined with (3.8) gives

$$\mathbb{B}(\mathbb{P}(e), v) = \mathbb{B}(U - \mathbb{P}(U), v).$$

Taking $v = u - \mathbb{P}(U) = \mathbb{P}(e)$, we have

$$(3.13) \quad \mathbb{B}(\mathbb{P}(e), \mathbb{P}(e)) = \mathbb{B}(U - \mathbb{P}(U), \mathbb{P}(e)).$$

Note the left-hand side of the equality involves the term $\mathbb{P}(e)$ that we want to estimate. The right-hand side of the equality is $\mathbb{B}(U - \mathbb{P}(U), \mathbb{P}(e))$, which is expected to be small because it involves the error between the exact solution and its L^2 projection $U - \mathbb{P}(U)$.

Letting $w = v = \mathbb{P}(e)$ in (3.11) and using $\mathbb{P}(e)(\cdot, 0) = 0$, we have

$$(3.14) \quad \mathbb{B}(\mathbb{P}(e), \mathbb{P}(e)) = \frac{1}{2} \|\mathbb{P}(e)(\cdot, T)\|^2 + \int_0^T \sum_{j=1}^N \|(\mathbb{P}(e))_x(\cdot, t)\|_{I_j}^2 dt + \Theta\left(T, (\widehat{\mathbb{P}(e)})_x, \mathbb{P}(e)\right).$$

Recalling the definition of admissibility for the numerical flux in (2.7) and the interface contribution term Θ defined in (3.12), we obtain

$$\Theta\left(T, (\widehat{\mathbb{P}(e)})_x, \mathbb{P}(e)\right) \geq \alpha \int_0^T \sum_{j=1}^N \frac{[\mathbb{P}(e)]^2}{\Delta x} dt - \gamma \int_0^T \sum_{j=1}^N \|(\mathbb{P}(e))_x(\cdot, t)\|_{I_j}^2 dt.$$

Hence

$$(3.15) \quad \mathbb{B}(\mathbb{P}(e), \mathbb{P}(e)) \geq \|\mathbb{P}(e)(\cdot, T)\|^2 - \frac{1}{2} \|\mathbb{P}(e)(\cdot, T)\|^2.$$

On the other hand,

$$\begin{aligned}
 \mathbb{B}(U - \mathbb{P}(U), \mathbb{P}(e)) &= \int_0^T \int_0^1 (U - \mathbb{P}(U))_t \mathbb{P}(e) dx dt \\
 (3.16) \quad &+ \int_0^T \sum_{j=1}^N \int_{I_j} (U - \mathbb{P}(U))_x (\mathbb{P}(e))_x dx dt + \Theta \left(T, (U - \widehat{\mathbb{P}(U)})_x, \mathbb{P}(e) \right).
 \end{aligned}$$

With $\mathbb{P}(e) \in \mathbb{V}_{\Delta x}$, we have

$$\int_0^T \int_0^1 (U - \mathbb{P}(U))_t \mathbb{P}(e) dx dt = 0.$$

For the second term in (3.16) we obtain

$$\begin{aligned}
 &\int_0^T \sum_{j=1}^N \int_{I_j} (U - \mathbb{P}(U))_x (\mathbb{P}(e))_x dx dt \\
 &\leq \frac{1}{2(1-\gamma)} \int_0^T \sum_{j=1}^N \|(U - \mathbb{P}(U))_x(\cdot, t)\|_{I_j}^2 dt + \frac{(1-\gamma)}{2} \int_0^T \sum_{j=1}^N \|(\mathbb{P}(e))_x(\cdot, t)\|_{I_j}^2 dt.
 \end{aligned}$$

The third term in (3.16) is majored by

$$\begin{aligned}
 &\int_0^T \left[\sum_{j=1}^N (U - \widehat{\mathbb{P}(U)})_x [\mathbb{P}(e)] \right] dt \\
 &\leq \frac{\Delta x}{2\alpha} \int_0^T \sum_{j=1}^N \{D_x(U - \mathbb{P}(U))\}^2 dt + \frac{\alpha}{2} \int_0^T \sum_{j=1}^N \frac{[\mathbb{P}(e)]^2}{\Delta x} dt.
 \end{aligned}$$

The above three estimates when inserted into (3.16) gives

$$\begin{aligned}
 \mathbb{B}(U - \mathbb{P}(U), \mathbb{P}(e)) &\leq \frac{1}{2} \|\mathbb{P}(e)(\cdot, T)\|^2 - \frac{1}{2} \|\mathbb{P}(e)(\cdot, 0)\|^2 \\
 &+ \frac{1}{2(1-\gamma)^2} \|(U - \mathbb{P}(U))(\cdot, T)\|^2 + \frac{\Delta x}{2\alpha} \int_0^T \sum_{j=1}^N \{D_x(U - \mathbb{P}(U))\}^2 dt.
 \end{aligned}$$

This with (3.15) when substituted into (3.16) yields the inequality claimed in Lemma 3.3. \square

4. Multidimensional diffusion process. In this section, we generalize the DDG method discussed in the previous sections to multiple spatial dimensions $x = (x_1, \dots, x_d)$. We solve the following diffusion problem:

$$(4.1) \quad \partial_t U - \sum_{i=1}^d \partial_{x_i} \left(\sum_{j=1}^d a_{ij}(U) \partial_{x_j} U \right) = 0 \quad \text{in } (0, T) \times (0, 1)^d,$$

$$(4.2) \quad U(x, t = 0) = U_0 \quad \text{on } (0, 1)^d,$$

with periodic boundary conditions. The diffusion coefficient matrix (a_{ij}) is assumed to be symmetric, semipositive definite.

Notice that the assumption of a unit box geometry and periodic boundary conditions is for simplicity only and is not essential: the method can be designed for arbitrary domain and for nonperiodic boundary conditions.

Let a partition of the unit box $(0, 1)^d$ be denoted by shape-regular meshes $T_\Delta = \{K\}$, consisting of a nonoverlapping open element covering completely the unit box. We denote by Δ the piecewise constant mesh function with $\Delta(x) \equiv \Delta_K = \text{diam}\{K\}$ when x is in element K . Let each K be a smooth bijective image of a fixed master element: the open hypercube $C = (-1, 1)^d$ through $F_K : C \rightarrow K$. On C we define spaces of polynomials of degree $k \geq 1$ as follows:

$$P^k = \text{span}\{y^\alpha : 0 \leq \alpha_i \leq k, 1 \leq i \leq d\}.$$

We denote the finite element space by

$$(4.3) \quad \mathbb{V}_\Delta = \{v : v|_K \circ F_K \in P^k \quad \forall K \in T_\Delta\}.$$

Note that the master element can also be chosen as the open unit simplex

$$S = \{x \in \mathbb{R}^d : 0 < x_1 + \dots + x_d < 1, x_j > 0, j = 1 \dots d\};$$

then the corresponding polynomial should be changed to $P^k = \text{span}\{y^\alpha : 0 \leq |\alpha| \leq k\}$. The DDG method is obtained by discretizing (4.1) directly with the DG method. This is achieved by multiplying the equation by test functions $v \in \mathbb{V}_\Delta$, integrating over an element $K \in T_\Delta$, and integration by parts. We again need to pay special attention to the boundary terms resulting from the procedure of integration by parts, as in the 1D case. Thus we seek piecewise polynomial solution $u \in \mathbb{V}_\Delta$, where \mathbb{V}_Δ is defined in (4.3) such that for all test functions $v \in \mathbb{V}_\Delta$ we have

$$(4.4) \quad \int_K u_t v dx + \int_K \sum_{i=1}^d \sum_{j=1}^d a_{ij}(u) \partial_{x_j} u \partial_{x_i} v dx - \int_{\partial K} \widehat{h}_{n_K} v^{int_K} ds = 0,$$

where ∂K is the boundary of element K , $n_K = (n_{1,K}, \dots, n_{d,K})$ is the outward unit normal for element K along the element boundary ∂K , and v^{int_K} denotes the value of v evaluated from inside the element K . Correspondingly, we use v^{ext_K} to denote the value of v evaluated from outside the element K (inside the neighboring element). The numerical flux \widehat{h}_{n_K} is defined similarly to the 1D case as

$$(4.5) \quad \widehat{h}_{n_K} = \widehat{h}_{n_K, K}(u^{int_K}, u^{ext_K}) = \sum_{i=1}^d \left(\sum_{j=1}^d \partial_{x_j} \widehat{b_{ij}}(u) \right) n_i,$$

where $b_{ij}(u) = \int^u a_{ij}(s) ds$ and

$$\partial_{x_j} \widehat{b_{ij}}(u) = \beta_0 \frac{[b_{ij}(u)]}{\Delta} n_j + \overline{\partial_{x_j}(b_{ij}(u))},$$

where locally Δ can be defined as the average of diameters of two neighboring elements sharing one common face. Here we have used the following notations:

$$[u] = u^{ext_K} - u^{int_K} \quad \text{and} \quad \overline{\partial_{x_j} u} = \frac{\partial_{x_j} u^{ext_K} + \partial_{x_j} u^{int_K}}{2}.$$

Note that for hyperrectangle meshes we replace Δ by $\overline{\Delta x_j}$, which denotes the average of lengths of two adjacent elements in the x_j direction only. This way the scheme is consistent with the finite difference scheme when $\beta_0 = 1$. In the general case, the stability is ensured by a larger choice of β_0 . The algorithm is now well defined.

We note that the numerical flux defined above enjoys some nice properties similar to those in the 1D case. More precisely, $\widehat{h_{n_K}}(u^{int_K}, u^{ext_K})$ is consistent with $h_{n_K}(u)$ in the sense that $\widehat{h_{n_K}}(u, u) = \sum_{i=1}^d (\sum_{j=1}^d \partial_{x_j}(b_{ij}(u)))n_i$, which is verified for all u smooth enough. It is also conservative (that is, there is only one flux defined at each face shared by two elements), namely,

$$\widehat{h_{n_K, K}}(a, b) = -\widehat{h_{n_{K'}, K'}}(b, a),$$

where K and K' share the same face where the flux is computed and hence $n_{K'} = -n_K$. Moreover, it ensures the L^2 -stability of the method.

THEOREM 4.1 (energy stability). *Assume that for $p \in \mathbb{R}$, $\exists \gamma$ and γ^* such that the eigenvalues of matrix $(a_{ij}(p))$ lie between $[\gamma, \gamma^*]$. Consider the DDG scheme with numerical flux chosen in (4.5). Then the numerical solution satisfies*

$$\begin{aligned} \int_{(0,1)^d} u^2(x, T) dx + \int_0^T \sum_K \int_K \sum_{i=1}^d \sum_{j=1}^d a_{ij}(u) u_{x_i} u_{x_j} dx dt \\ + \gamma \beta_0 \int_0^T \sum_K \int_{\partial K} \frac{[u]^2}{\Delta} ds dt \leq \int_0^1 U_0^2(x) dx, \end{aligned} \tag{4.6}$$

provided $\beta_0 \geq C(k)(\frac{\gamma^*}{\gamma})^2$ for some $C(k)$, depending on the degree k of the approximating polynomial.

Proof. Setting $v = u$ in (4.4) and summing over all elements, we obtain

$$\frac{d}{dt} \int_{\Omega} \frac{u^2}{2} dx + \sum_K \int_K \nabla u \cdot (A(u) \nabla u) dx + \sum_K \int_{\partial K} \hat{h}[u] ds = 0, \quad \Omega := [0, 1]^d. \tag{4.7}$$

The last term involving the flux (4.5) can be bounded from below as follows:

$$\begin{aligned} \sum_K \int_{\partial K} \hat{h}[u] ds &= \sum_K \int_{\partial K} \left(\frac{\beta_0 [u]}{\Delta} \sum_{i,j=1}^d n_i \frac{[b_{ij}(u)]}{[u]} n_j + \sum_{i,j=1}^d \partial_{x_j} u a_{ij}(u) n_i \right) [u] ds \\ &\geq \frac{\beta_0}{\Delta} \sum_K \int_{\partial K} \gamma [u]^2 ds - \gamma^* \sum_K \int_{\partial K} |\nabla u| |n| [u] ds \\ &\geq \gamma \beta_0 \sum_K \int_{\partial K} \frac{[u]^2}{\Delta} ds - \gamma^* \sum_K \|\nabla u\|_{0, \partial K} \| [u] \|_{0, \partial K} \\ &\geq \frac{\beta_0 \gamma}{2} \sum_K \Delta^{-1} \| [u] \|_{0, \partial K}^2 - \frac{(\gamma^*)^2}{2 \beta_0 \gamma} \sum_K \Delta \|\nabla u\|_{0, \partial K}^2, \end{aligned} \tag{4.8}$$

where we used the assumption on matrix $A(u)$, followed by using the inequality $ab \leq \epsilon a^2/2 + b^2/(2\epsilon)$ to achieve the last inequality. Using the trace inequality and the fact that $u \in \mathbb{V}_{\Delta}$ we further obtain

$$\|\nabla u\|_{0, \partial K}^2 \leq C (\Delta^{-1} \|\nabla u\|_{0, K}^2 + \Delta \|\nabla^2 u\|_{0, K}^2) \leq C(k) \Delta^{-1} \|\nabla u\|_{0, K}^2.$$

Hence

$$\sum_K \Delta \|\nabla u\|_{0,\partial K}^2 \leq C(k) \sum_K \|\nabla u\|_{0,\partial K}^2 \leq \frac{C(k)}{\gamma} \sum_K \int_K \nabla u \cdot (A(u)\nabla u) dx.$$

This, together with (4.8) and when inserted into (4.7), gives

$$\frac{d}{dt} \int_{\Omega} \frac{u^2}{2} dx + \left(1 - \frac{C(k)}{2\beta_0} \left(\frac{\gamma^*}{\gamma}\right)^2\right) \sum_K \int_K \nabla u \cdot (A(u)\nabla u) dx + \frac{\gamma\beta_0}{2} \sum_K \int_{\partial K} \frac{[u]^2}{\Delta} ds \leq 0.$$

Thus the asserted inequality follows from time integration of the above over $[0, T]$ and the fact that $\|u_0\|_{0,\Omega} \leq \|U_0\|_{0,\Omega}$, provided $\beta_0 \geq C(k)\left(\frac{\gamma^*}{\gamma}\right)^2$. \square

5. Numerical examples. In this section, we provide a few numerical examples to illustrate the accuracy and capacity of the DDG method. We would like to illustrate the high order accuracy of the method through these numerical examples from 1D to 2D linear and nonlinear problems. In particular, we study the numerical performance of the scheme with different admissible numerical fluxes.

Example 5.1 (1D linear diffusion equation).

$$(5.1) \quad U_t - U_{xx} = 0 \quad \text{in } (0, 2\pi)$$

with initial condition $U(x, 0) = \sin(x)$ and periodic boundary conditions. The exact solution is given by $U(x, t) = e^{-t}\sin(x)$. We compute the solution up to $t = 1$. The numerical flux \widehat{u}_x we first test is

$$(5.2) \quad \widehat{u}_x = D_x(u) = \frac{[u]}{\Delta x} + \overline{u}_x.$$

DDG methods based on p^k polynomial approximations with $k = 0, 1, 2$ are tested. We list the L^2 and L^∞ errors in Table 1. Note that in this example and the rest L^∞ error is evaluated on many sample points (200 points per cell). We obtain clean 1st and 2nd order accuracy for p^0 and p^1 approximations. However, we obtain only 2nd order convergence for p^2 approximations.

For higher order polynomial approximations, the proposed numerical flux formula suggests that interface values necessarily involve higher order derivatives of the solution. We first test the scheme (2.12) with $h = \theta\Delta x$, called θ -scheme:

$$(5.3) \quad D_x u = \frac{7}{12\theta} \frac{[u]}{\Delta x} + \overline{u}_x + \frac{\theta\Delta x}{6} [u_{xx}].$$

In Table 2 we compute p^2 approximations for problem (5.1) with numerical flux (5.3) and list the L^∞ errors and orders with different θ values in interval $(0, 2)$. We would like to point out that almost all θ -schemes give us 2nd order convergence for p^2 polynomial approximations except the one $\theta = 0.5$, i.e., (2.12), which can fully recover the order of 3. Numerically, we observe that the scheme with any fixed β_1 is not sensitive to the coefficient before $\frac{[u]}{\Delta x}$, i.e., β_0 , as long as the numerical flux is still admissible.

Numerical results for p^3 approximations with these θ -schemes are displayed in Table 3. Different from the p^2 approximations, all schemes give 4th order convergence. This is in sharp contrast to the p^2 approximations, which give the desired order of 3 only in the case of $\beta_1 = 1/12$.

TABLE 1

Computational domain Ω is $[0, 2\pi]$. L^2 and L^∞ errors at $t = 1.0$. p^k polynomial approximations with $k = 0, 1, 2$. Numerical flux (5.2) is used.

k		N=10		N=20		N=40		N=80	
		error	error	order	error	order	error	order	
0	L^2	4.8602E-02	2.3771E-02	1.03	1.1818E-02	1.00	5.9007E-03	1.00	
	L^∞	1.1743E-01	5.8023E-02	1.01	2.8923E-02	1.00	1.4450E-02	1.00	
1	L^2	1.3400E-02	3.3494E-03	2.00	8.3726E-04	2.00	2.0931E-04	2.00	
	L^∞	3.0145E-02	7.5004E-03	2.00	1.8871E-03	2.00	4.7252E-04	2.00	
2	L^2	8.5278E-03	2.1377E-03	1.99	5.3476E-04	1.99	1.3371E-04	2.00	
	L^∞	1.2082E-02	2.9989E-03	2.00	7.5475E-04	1.99	1.8900E-04	2.00	

TABLE 2

L^∞ errors for p^2 approximation with sample θ values in $(0, 2)$ at $t = 1.0$. Numerical flux (5.3) is used.

θ	N=10		N=20		N=40		N=80	
	error	error	order	error	order	error	order	
0.1	1.6671E-03	4.1669E-04	2.00	1.0385E-04	2.00	2.5941E-05	2.00	
0.3	2.3926E-03	6.2056E-04	1.95	1.5549E-04	2.00	3.8894E-05	2.00	
0.49	7.2472E-04	1.0088E-04	2.84	1.6683E-05	2.60	3.4529E-06	2.27	
0.5	7.4892E-04	9.1995E-05	3.02	1.1450E-05	3.00	1.4296E-06	3.00	
0.51	8.1560E-04	1.0970E-04	2.89	1.7893E-05	2.61	3.6315E-06	2.30	
0.8	9.8839E-03	2.4693E-03	2.00	6.2113E-04	1.99	1.5553E-04	1.99	
1.5	5.6436E-02	1.5037E-02	1.90	3.8570E-03	1.96	9.7047E-04	1.99	

TABLE 3

L^∞ errors for p^3 approximation with sample θ values in $(0, 2)$ at $t = 1.0$. Numerical flux (5.3) is used.

θ	N=10		N=20		N=40		N=80	
	error	error	order	error	order	error	order	
0.1	8.1835E-05	5.0752E-06	4.01	3.2055E-07	3.98	2.0087E-08	3.99	
0.3	5.2604E-05	3.2520E-06	4.01	2.0522E-07	3.98	1.2857E-08	4.00	
0.5	1.3097E-04	8.0993E-06	4.01	5.0895E-07	3.99	3.1853E-08	3.99	
0.8	5.0961E-04	3.1639E-05	4.00	1.9942E-06	3.98	1.2491E-07	3.99	
1.5	2.0972E-03	1.4010E-04	3.90	8.8607E-06	3.98	5.5517E-07	3.99	

Next we test the η -scheme with

$$(5.4) \quad D_x u = \frac{1}{\sqrt{4\pi}} \frac{[u]}{\eta \Delta x} + \bar{u}_x + \frac{\eta \Delta x}{\sqrt{\pi}} [u_{xx}].$$

Similar to the θ -schemes, only $\eta = \frac{\sqrt{\pi}}{12}$ gives fully 3rd order convergence for p^2 polynomial approximations. In Table 4 we list the L^∞ errors with different η values, and the numerical results are comparable to the θ -schemes. Note that careful verification shows that a large class of the θ -schemes and η -schemes satisfy the admissible condition (2.8).

In summary, $\beta_1 = \frac{1}{12}$ numerically gives the optimal $(k + 1)$ th order of convergence for both p^2 and p^3 polynomial approximations. In Table 5 we use the numerical flux (5.5) to compute the problem with p^2 , p^3 , and p^4 polynomial approximations:

$$(5.5) \quad D_x u = \frac{[u]}{\Delta x} + \bar{u}_x + \frac{\Delta x}{12} [u_{xx}].$$

TABLE 4

L^∞ errors and orders comparisons for η -schemes at $t = 1.0$. p^2 polynomial approximation. Numerical flux (5.4) is used.

η	N=10		N=20		N=40		N=80	
	error	error	order	error	order	error	order	
0.05	1.4733E-03	3.5730E-04	2.04	8.8847E-05	2.00	2.2182E-05	2.00	
0.1	1.4427E-03	3.4865E-04	2.05	8.6743E-05	2.00	2.1660E-05	2.00	
$\frac{\sqrt{\pi}}{12}$	7.3278E-04	9.1488E-05	3.00	1.1434E-05	3.00	1.4291E-06	3.00	
0.2	3.1376E-03	7.7015E-04	2.02	1.9054E-04	2.01	4.7511E-05	2.00	
0.5	4.7350E-02	1.2413E-02	1.93	3.1738E-03	1.96	7.9794E-04	1.99	

TABLE 5

L^2 and L^∞ errors at $t = 1.0$ with p^k approximations $k = 2, 3, 4$. Numerical flux (5.5) is used.

k		N=10		N=20		N=40		N=80	
		error	error	order	error	order	error	order	
2	L^2	3.9238E-04	4.7037E-05	3.06	5.8181E-06	3.01	7.2535E-07	3.00	
	L^∞	7.5595E-04	9.2213E-05	3.03	1.1456E-05	3.00	1.4298E-06	3.00	
3	L^2	9.2531E-05	5.7809E-06	4.00	3.6128E-07	4.00	2.2579E-08	4.00	
	L^∞	1.5351E-04	9.5014E-06	4.01	5.9750E-07	3.99	3.7403E-08	3.99	
4	L^2	5.5070E-05	3.4999E-06	3.97	2.1965E-07	3.99	1.3743E-08	3.99	
	L^∞	7.7911E-05	4.8932E-06	3.99	3.0974E-07	3.98	1.9422E-08	3.99	

Similar to the p^2 case we lose one order of accuracy for p^4 approximations. These results together indicate that for even-order, $k = 2m$, polynomial approximations, the coefficient β_m seems indispensable.

Example 5.2 (1D linear diffusion equation with higher order polynomial approximations). We study the same problem as the one in Example 5.1 with the numerical flux chosen as

$$(5.6) \quad D_x u = 2 \frac{[u]}{\Delta x} + \overline{u}_x.$$

As discussed in section 2, we find out that with β_0 (the coefficient before $[u]/\Delta x$) big enough the numerical flux formula (2.7) with the first two terms is admissible. In this example, we test the DDG scheme with higher order polynomial approximations $p^k, k = 2, 3, 4, 5, 6, 7$. Errors and orders are listed in Table 6. We obtain k th order accuracy for even k and $(k + 1)$ th order accuracy for odd k .

Example 5.3 (1D linear diffusion equation with nonuniform mesh). Again, we study the same problem as the one in Example 5.1. Here the partition of the domain $[0, 2\pi]$ consists of a repeated pattern of $1.1\Delta x$ and $0.9\Delta x$ for odd and even numbers of indexes $i = 1, \dots, N$, respectively, where $\Delta x = 2\pi/N$ with even number N . The numerical flux we use is

$$D_x(u) = \frac{[u]}{\Delta x} + \overline{u}_x.$$

We obtain similar results as Example 5.2. Errors and orders are listed in Table 7.

Example 5.4 (1D nonlinear diffusion equations).

$$(5.7) \quad U_t - (2UU_x)_x = 0 \quad \text{in } [-12, 12].$$

TABLE 6

High order polynomial approximations ($p^k, k = 2, 3, 4, 5, 6, 7$) with numerical flux (5.6). L^2 and L^∞ errors at $t = 1.0$.

k		N=4	N=8		N=12		N=16	
		error	error	order	error	order	error	order
2	L^2	2.3913E-02	6.5160E-03	1.88	2.9383E-03	1.96	1.6610E-03	1.98
	L^∞	3.0219E-02	8.9725E-03	1.75	4.1066E-03	1.93	2.3335E-03	1.96
3	L^2	3.6671E-03	2.2958E-04	4.00	4.5459E-05	3.99	1.4397E-05	4.00
	L^∞	4.5777E-03	3.6566E-04	3.65	7.5484E-05	3.90	2.4253E-05	3.95
4	L^2	3.5708E-04	2.7557E-05	3.60	5.6506E-06	3.91	1.8111E-06	3.96
	L^∞	4.4087E-04	3.7131E-05	3.70	7.8142E-06	3.85	2.5288E-06	3.93
5	L^2	3.9466E-05	6.4001E-07	5.95	5.6637E-08	5.98	1.0109E-08	5.99
	L^∞	4.6456E-05	9.2965E-07	5.64	8.4966E-08	5.90	1.5332E-08	5.95
6	L^2	1.8891E-06	4.1364E-08	5.51	3.8499E-09	5.86	6.9911E-10	5.93
	L^∞	2.4795E-06	5.5327E-08	5.49	5.3001E-09	5.78	9.7347E-10	5.89
7	L^2	2.1144E-07	8.9249E-10	7.89	3.5478E-11	7.95	3.6571E-12	7.90
	L^∞	2.5297E-07	1.2566E-09	7.65	5.1137E-11	7.90	5.3087E-12	7.87

TABLE 7

Nonuniform mesh test. L^2 and L^∞ errors at $t = 1.0$ with $k = 0, 1, 2, 3, 4, 5$.

k		N=10	N=20		N=40		N=80	
		error	error	order	error	order	error	order
0	L^2	4.8970E-02	2.3903E-02	1.03	1.1879E-02	1.00	5.9304E-03	1.00
	L^∞	1.3021E-01	6.3933E-02	1.02	3.1828E-02	1.00	1.5897E-02	1.00
1	L^2	1.4011E-02	3.4807E-03	2.00	8.6898E-04	2.00	2.1717E-04	2.00
	L^∞	3.8329E-02	9.3268E-03	2.00	2.3522E-03	1.99	5.8881E-04	2.00
2	L^2	8.7915E-03	2.2023E-03	1.99	5.5083E-04	2.00	1.3772E-04	2.00
	L^∞	1.2720E-02	3.0946E-03	2.03	7.7858E-04	1.99	1.9475E-04	2.00
3	L^2	1.9041E-04	1.1836E-05	4.00	7.3911E-07	4.00	4.6186E-08	4.00
	L^∞	3.9591E-04	2.4148E-05	4.03	1.5144E-06	4.00	9.4854E-08	4.00
4	L^2	2.5721E-05	1.6464E-06	3.97	1.0353E-07	4.00	6.4802E-09	4.00
	L^∞	3.7917E-05	2.3196E-06	4.03	1.4645E-07	3.99	9.1649E-09	4.00
5	L^2	3.7561E-07	5.8458E-09	6.00	9.1163E-11	6.00	1.4244E-12	6.00
	L^∞	7.5379E-07	1.1563E-08	6.02	1.8114E-10	6.00	2.8303E-12	6.00

The Barenblatt’s solution with compact support is given as

$$(5.8) \quad U(x, t) = \begin{cases} (t + 1)^{-\frac{1}{3}} \left(6 - \frac{x^2}{12(t+1)^{\frac{2}{3}}} \right), & |x| < 6(t + 1)^{\frac{1}{3}}, \\ 0, & |x| \geq 6(t + 1)^{\frac{1}{3}}. \end{cases}$$

We take the following numerical flux for this nonlinear problem:

$$\widehat{2uu}_x = \widehat{(u^2)}_x = \frac{[u^2]}{\Delta x} + \overline{(u^2)}_x + \frac{\Delta x}{12} [(u^2)_{xx}].$$

Both L^2 and L^∞ errors at $t = 1$ are evaluated in domain $[-6, 6]$ where the solution is smooth. Accuracy data are listed in Table 8. We have $(k + 1)$ th order accuracy with p^k polynomial approximations. Propagation of the compact wave using both P^1 and P^2 elements is plotted in Figure 2. A zoomed-in figure of the left corner at $t = 4$ is plotted in Figure 3. The DDG scheme can sharply capture the contacts with discontinuous derivatives.

TABLE 8

Computational domain Ω is $[-12, 12]$. Exact solution is given as (5.8). L^2 and L^∞ errors are computed in smooth region $[-6, 6]$ with $k = 0, 1, 2$ at $t = 1.0$.

k		N=40	N=80		N=160		N=320	
		error	error	order	error	order	error	order
0	L^2	3.8184E-02	1.8363E-02	1.05	9.0076E-03	1.03	4.4613E-03	1.01
	L^∞	1.5739E-01	7.7079E-02	1.03	3.8038E-02	1.02	1.8887E-02	1.01
1	L^2	2.8239E-03	6.8004E-04	2.05	1.6478E-04	2.04	4.0616E-05	2.02
	L^∞	1.3169E-02	2.8427E-03	2.21	6.6660E-04	2.09	1.6294E-04	2.03
2	L^2	2.2321E-04	1.2519E-05	4.15	1.1960E-06	3.38	1.4516E-07	3.04
	L^∞	2.5443E-03	8.3480E-05	4.92	3.7225E-06	4.48	4.1537E-07	3.16

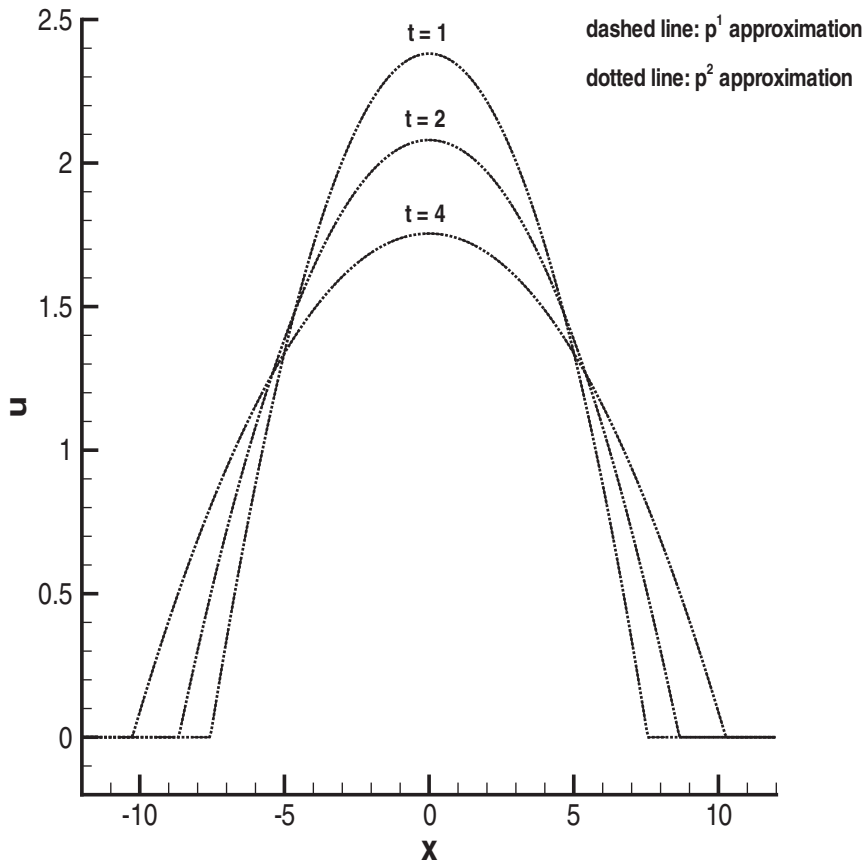


FIG. 2. Nonlinear diffusion equation (5.7). Piecewise linear (p^1) and piecewise quadratic (p^2) approximations with mesh $N = 400$.

Example 5.5 (2D linear diffusion equation).

$$(5.9) \quad U_t - (U_{xx} + U_{yy}) = 0 \quad \text{in } (0, 2\pi) \times (0, 2\pi)$$

with initial condition $U(x, y, 0) = \sin(x + y)$ and periodic boundary conditions. The exact solution is $U(x, y, t) = e^{-2t}\sin(x + y)$. We compute the solution up to $t = 1$ on the uniform rectangular mesh $I_{ij} = I_i \times I_j$. L^2 and L^∞ errors are listed in Table 9. $k + 1$ orders of convergence are obtained for p^k elements with $k \leq 3$.

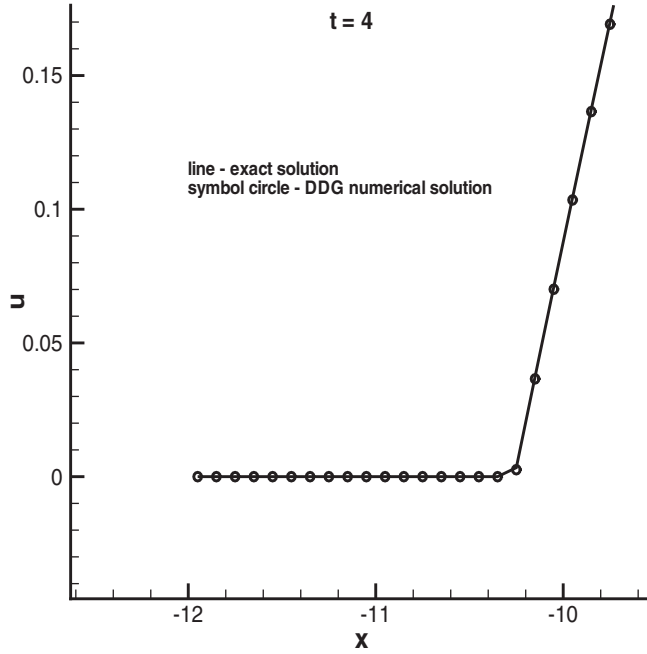


FIG. 3. Zoom in of Figure 2 at $t = 4$ at the left corner where the solution has a discontinuous derivative.

TABLE 9

Computational domain Ω is $[0, 2\pi] \times [0, 2\pi]$ with rectangular mesh $N \times N$. L^2 and L^∞ errors at $t = 1.0$. p^k polynomials with $k = 0, 1, 2, 3, 4$. Numerical flux (5.10) is used.

k		N=10		N=20		N=40		N=80	
		error	error	order	error	order	error	order	
0	L^2	2.5993E-02	1.2456E-02	1.06	6.1599E-03	1.01	3.0714E-03	1.00	
	L^∞	8.4874E-02	4.2512E-02	1.00	2.1258E-02	1.00	1.0629E-02	1.00	
1	L^2	1.0305E-02	2.5482E-03	2.01	6.3522E-04	2.00	1.5869E-04	2.00	
	L^∞	3.5728E-02	8.8609E-03	2.01	2.2234E-03	1.99	5.5637E-04	2.00	
2	L^2	1.1503E-03	1.4042E-04	3.03	1.7437E-05	3.00	2.1759E-06	3.00	
	L^∞	7.4943E-03	9.4734E-04	2.98	1.1872E-04	2.99	1.4849E-05	3.00	
3	L^2	1.0777E-04	6.2265E-06	4.11	3.8168E-07	4.02	2.3740E-08	4.00	
	L^∞	5.7940E-04	3.8896E-05	3.89	2.4742E-06	3.97	1.5531E-07	3.99	
4	L^2	5.3967E-05	3.4795E-06	3.95	2.1932E-07	3.98	1.3738E-08	3.99	
	L^∞	8.5886E-05	5.1584E-06	4.05	3.1396E-07	4.03	1.9488E-08	4.01	

The DDG scheme in 2D with rectangular mesh is a straightforward extension of the 1D scheme. The numerical flux \widehat{u}_x at $x_{i+1/2}$ used in this example is defined as follows:

$$(5.10) \quad \widehat{u}_x|_{x_{i+1/2}} = \frac{[u]}{\Delta x} + \overline{u}_x + \frac{\Delta x}{12}[u_{xx}].$$

Flux \widehat{u}_y at $y_{j+1/2}$ is defined in a similar fashion.

6. Concluding remarks. We have proposed a new DG finite element method for solving diffusion problems. The scheme is formulated using the direct weak for-

mulation for parabolic equations, combined with a careful design of interface values of the solution derivative. Unlike the traditional LDG method, the method in this paper is applied without introducing any auxiliary variables or rewriting the original equation into a 1st order system. The proposed numerical flux formula for solution derivatives is consistent and conservative. A concept of admissibility is further introduced to identify a class of numerical fluxes so that the nonlinear stability for both 1D and multidimensional problems are ensured. For the 1D linear case, k th order accuracy in an energy norm is proven when using k th degree polynomials. A series of numerical examples are presented to demonstrate the high order accuracy of the method and its capacity to sharply capture solutions with discontinuous derivatives. In particular, the optimal $(k + 1)$ th order accuracy is attained for $k = 0, 1, 2, 3$. The method maintains the usual features of DG methods such as high order accuracy and easiness to handle complicated geometry. Moreover, our DDG method has an advantage of easier formulation and implementation and efficient computation of solutions. The compactness of the scheme allows efficient parallelization and hp-adaptivity.

The numerical tests show the strong dependence of the order of convergence of the DDG method on the choice of numerical fluxes. The development of even higher order DDG methods with further analysis of optimal choices for $\beta_i, i \geq 1$, will be studied in a future work. The DDG method for convection-diffusion problems can be defined by applying the procedure described above for the diffusion term combined with numerical fluxes for the convection term developed previously for hyperbolic conservation laws.

7. Appendix. The Bramble–Hilbert lemma. Let Ω be a simply connected Lipschitz domain in R^d and $(m, k) \in \mathbb{N}^2, p, q \in [1, \infty]$. If (p, q, k, m) satisfies

$$(7.1) \quad \frac{1}{q} > \frac{1}{p} - \frac{k + 1 - m}{d},$$

then the Sobolev space $W^{k+1,p}(\Omega)$ is continuously embedded into $W^{m,q}(\Omega)$. With this setting we recall the celebrated Bramble–Hilbert lemma.

LEMMA 7.1 (Bramble–Hilbert). *Let l be a linear operator mapping $W^{k+1,p}(\Omega)$ into $W^{m,q}(\Omega)$ and $P_k(\Omega) \subset Ker(l)$. Then there exists a constant $C(\Omega) > 0$ such that for all $u \in W^{m,q}(\Omega)$*

$$(7.2) \quad |l(v)|_{m,q} \leq C|v|_{k+1,p}.$$

The assumption $P_k(\Omega) \subset Ker(l)$ is used to ensure the Sobolev quotient norm in $W^{k+1,p}/P_k(\Omega)$ to be equivalent to the Sobolev seminorm in $W^{k+1,p}(\Omega)$.

In the 1D case I_j is affine equivalent to $\omega = [0, 1]$ through a linear mapping

$$x = x_{j-1/2} + \xi\Delta x, \quad \xi \in [0, 1].$$

Taking $l = I - \mathbb{P}$ and using the scaling argument, the Bramble–Hilbert lemma enables us to obtain

$$|v - \mathbb{P}v|_{m,q,I_j} \leq C(\Delta x)^{k+1-m+\frac{1}{q}-\frac{1}{p}}|v|_{k+1,p,I_j}, \quad v \in W^{k+1,p}(I_j).$$

Here C depends only on $[0, 1]$ and the projection operator \mathbb{P} but is independent of Δx . Estimates in Lemma 3.1 are immediate if

- (i) we set $p = q = 2$ and assume $m < k + 1$;
- (ii) we set $q = \infty$ and $p = 2$ and assume $m < k + 1/2$.

Solution gradient for the heat equation. Consider the heat equation $u_t = u_{xx}$ with smooth initial data g , having only one discontinuity at $x = 0$. A straightforward calculation from the solution formula

$$u(x, t) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} e^{-(x-y)^2/(4t)} g(y) dy$$

gives

$$(7.3) \quad \begin{aligned} u_x(0, t) &= \sum \frac{2^{m-1}}{(2m-1)!!} t^m [\partial_x^{2m} g] / \sqrt{\pi t} + \sum \frac{2^m}{(2m)!!} t^m \overline{\partial_x^{2m+1} g} \\ &= \frac{1}{\sqrt{4\pi t}} [g] + \overline{\partial_x g} + \sqrt{\frac{t}{\pi}} [\partial_x^2 g] + t \overline{\partial_x^3 g} + \dots, \end{aligned}$$

where the jump or the average of g and its derivatives are involved to evaluate u_x at $x = 0$.

Acknowledgments. The authors thank the anonymous referees who provided valuable comments resulting in improvements in this paper.

REFERENCES

- [1] D. N. ARNOLD, F. BREZZI, B. COCKBURN, AND L. D. MARINI, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.
- [2] D. N. ARNOLD, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.
- [3] G. A. BAKER, *Finite element methods for elliptic equations using nonconforming elements*, Math. Comp., 31 (1977), pp. 45–59.
- [4] F. BASSI AND S. REBAY, *A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations*, J. Comput. Phys., 131 (1997), pp. 267–279.
- [5] C. E. BAUMANN AND J. T. ODEN, *A discontinuous hp finite element method for convection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 175 (1999), pp. 311–341.
- [6] F. BREZZI, B. COCKBURN, L. D. MARINI, AND E. SÜLI, *Stabilization mechanisms in discontinuous Galerkin finite element methods*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 3293–3310.
- [7] P. CASTILLO, B. COCKBURN, I. PERUGIA, AND D. SCHÖTZAU, *An a priori error analysis of the local discontinuous Galerkin method for elliptic problems*, SIAM J. Numer. Anal., 38 (2000), pp. 1676–1706.
- [8] F. CELIKER AND B. COCKBURN, *Superconvergence of the numerical traces of discontinuous Galerkin and hybridized methods for convection-diffusion problems in one space dimension*, Math. Comp., 76 (2007), pp. 67–96.
- [9] Y. CHENG AND C.-W. SHU, *A discontinuous Galerkin finite element method for time dependent partial differential equations with higher order derivatives*, Math. Comp., 77 (2008), pp. 699–730.
- [10] B. COCKBURN AND C. DAWSON, *Approximation of the velocity by coupling discontinuous Galerkin and mixed finite element methods for flow problems*, Comput. Geosci., 6 (2002), pp. 505–522.
- [11] B. COCKBURN, S. HOU, AND C.-W. SHU, *The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. IV. The multidimensional case*, Math. Comp., 54 (1990), pp. 545–581.
- [12] B. COCKBURN, G. KANSCHAT, AND D. SCHÖTZAU, *A locally conservative LDG method for the incompressible Navier-Stokes equations*, Math. Comp., 74 (2005), pp. 1067–1095.
- [13] B. COCKBURN, G. E. KARNIADAKIS, AND C.-W. SHU, *The development of discontinuous Galerkin methods*, in *Discontinuous Galerkin Methods*, Lect. Notes Comput. Sci. Eng. 11, Springer, Berlin, 2000, pp. 3–50.
- [14] B. COCKBURN, S. Y. LIN, AND C.-W. SHU, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. III. One-dimensional systems*, J. Comput. Phys., 84 (1989), pp. 90–113.

- [15] B. COCKBURN AND C.-W. SHU, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. II. General framework*, Math. Comp., 52 (1989), pp. 411–435.
- [16] B. COCKBURN AND C.-W. SHU, *The Runge-Kutta local projection P^1 -discontinuous-Galerkin finite element method for scalar conservation laws*, RAIRO Modél. Math. Anal. Numér., 25 (1991), pp. 337–361.
- [17] B. COCKBURN AND C.-W. SHU, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.
- [18] B. COCKBURN AND C.-W. SHU, *The Runge-Kutta discontinuous Galerkin method for conservation laws. V. Multidimensional systems*, J. Comput. Phys., 141 (1998), pp. 199–224.
- [19] B. COCKBURN AND C.-W. SHU, *Runge-Kutta discontinuous Galerkin methods for convection-dominated problems*, J. Sci. Comput., 16 (2001), pp. 173–261.
- [20] B. COCKBURN, *Discontinuous Galerkin methods for convection-dominated problems*, in High-order Methods for Computational Physics, Lect. Notes Comput. Sci. Eng. 9, Springer, Berlin, 1999, pp. 69–224.
- [21] G. GASSNER, F. LÖRCHER, AND C.-D. MUNZ, *A contribution to the construction of diffusion fluxes for finite volume and discontinuous Galerkin schemes*, J. Comput. Phys., 224 (2007), pp. 1049–1063.
- [22] D. LEVY, C.-W. SHU, AND J. YAN, *Local discontinuous Galerkin methods for nonlinear dispersive equations*, J. Comput. Phys., 196 (2004), pp. 751–772.
- [23] H. LIU AND J. YAN, *A local discontinuous Galerkin method for the Korteweg-de Vries equation with boundary effect*, J. Comput. Phys., 215 (2006), pp. 197–218.
- [24] J. T. ODEN, I. BABUŠKA, AND C. E. BAUMANN, *A discontinuous hp finite element method for diffusion problems*, J. Comput. Phys., 146 (1998), pp. 491–519.
- [25] W. H. REED AND T. R. HILL, *Triangular Mesh Methods for the Neutron Transport Equation*, Technical report LA-UR-73-479, Los Alamos Scientific Laboratory, Los Alamos, NM, 1973.
- [26] C.-W. SHU AND S. OSHER, *Efficient implementation of essentially nonoscillatory shock-capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.
- [27] C.-W. SHU AND S. OSHER, *Efficient implementation of essentially nonoscillatory shock-capturing schemes. II*, J. Comput. Phys., 83 (1989), pp. 32–78.
- [28] C.-W. SHU, *Different formulations of the discontinuous Galerkin method for the viscous terms*, Advances in Scientific Computing, Z.-C. Shi, M. Mu, W. Xue, and J. Zou, eds., Science Press, Beijing, China, 2001, pp. 144–155.
- [29] B. VAN LEER AND S. NOMURA, *Discontinuous Galerkin for diffusion*, in Proceedings of the 17th AIAA Computational Fluid Dynamics Conference, Toronto, Canada, 2005, AIAA-2005-5108.
- [30] M. F. WHEELER, *An elliptic collocation-finite element method with interior penalties*, SIAM J. Numer. Anal., 15 (1978), pp. 152–161.
- [31] Y. XU AND C.-W. SHU, *Local discontinuous Galerkin methods for nonlinear Schrödinger equations*, J. Comput. Phys., 205 (2005), pp. 72–97.
- [32] J. YAN AND C.-W. SHU, *A local discontinuous Galerkin method for KdV type equations*, SIAM J. Numer. Anal., 40 (2002), pp. 769–791.

FINITE ELEMENT APPROXIMATION OF THE THREE-FIELD FORMULATION OF THE STOKES PROBLEM USING ARBITRARY INTERPOLATIONS*

RAMON CODINA†

Abstract. The stress-displacement-pressure formulation of the elasticity problem may suffer from two types of numerical instabilities related to the finite element interpolation of the unknowns. The first is the classical pressure instability that occurs when the solid is incompressible, whereas the second is the lack of stability in the stresses. To overcome these instabilities, there are two options. The first is to use different interpolation for all the unknowns satisfying two inf-sup conditions. Whereas there are several displacement-pressure interpolations that render the pressure stable, less possibilities are known for the stress interpolation. The second option is to use a stabilized finite element formulation instead of the plain Galerkin approach. If this formulation is properly designed, it is possible to use arbitrary interpolation for all the unknowns. The purpose of this paper is precisely to present one of such formulations. In particular, it is based on the decomposition of the unknowns into their finite element component and a subscale, which will be approximated and whose goal is to yield a stable formulation. A singular feature of the method to be presented is that the subscales will be considered orthogonal to the finite element space. We describe the design of the formulation and present the results of its numerical analysis.

Key words. Stokes problem, stress-displacement-pressure, stabilized finite elements

AMS subject classifications. 65N12, 65N30, 76D07

DOI. 10.1137/080712726

1. Introduction. The analysis of the three-field formulation of the linear elastic incompressible problem is probably not a goal by itself, but rather a simple model to study problems in which it is important to interpolate the stresses independently from the displacements and, in the case we will consider, also the pressure. Perhaps the most salient problem that requires the interpolation of the (deviatoric) stresses is the viscoelastic one. In this case, the algebraic constitutive equation (linear or nonlinear) that relates stresses and strains has to be replaced by an evolution equation (see [3] for a review).

The problem we will study in this paper is the simple Stokes problem arising in linear elasticity or creeping flows, taking as unknowns the displacement field (or velocity field, in a fluid problem), the pressure, and the deviatoric part of the stresses. In particular, we shall consider that the material is *incompressible*.

When the finite element approximation of the problem is undertaken, it is well known that incompressibility poses a stringent requirement in the way the pressure is interpolated with respect to the displacement field. The displacement and pressure finite element spaces have to satisfy the classical inf-sup condition [8]. Several interpolations are known that satisfy this condition and yield a stable displacement-pressure numerical solution. However, less is known about another inf-sup condition that needs to be satisfied when the stresses are interpolated independently from the displacement. This inf-sup condition is trivially satisfied for the continuous problem, but only a few interpolations are known that verify it for the discrete case. It is

*Received by the editors January 8, 2008; accepted for publication (in revised form) September 17, 2008; published electronically January 16, 2009.

<http://www.siam.org/journals/sinum/47-1/71272.html>

†Department de Resistència de Materias i Estructures a l'Enginyeria, Universitat Politècnica de Catalunya, Jordi Girona 1-3, Edifici C1, 08034 Barcelona, Spain (ramon.codina@upc.edu).

discussed, for example, in [25]. In the context of viscoelastic flows, a popular stable three-field interpolation was introduced in [23], and the numerical analysis was undertaken in [15]. See also [28, 26] for other contributions proposing different stable finite element interpolations.

The inf-sup conditions for the displacement-pressure and stresses-displacement interpolations are needed if the standard Galerkin method is used for the space discretization. However, there is also the possibility to resort to a *stabilized* finite element method, in which the discrete variational form of the Galerkin formulation is modified in order to enhance its stability. The purpose of this paper is precisely to present one of such formulations. In particular, the one proposed here is based on the decomposition of the unknowns into their finite element component and a subscale, that is, the component of the continuous unknown that cannot be captured by the finite element mesh. Obviously, this subscale needs to be approximated in one way or another. This idea was proposed in the finite element context in [20, 21] and termed *variational multiscale* approximation, although there are similar concepts developed in different situations (both in physical and numerical modeling).

The important property of the formulation to be presented here is that the subscale will be considered *orthogonal* to the appropriate finite element space. This idea was first applied to the Stokes problem in displacement-pressure form in [9], and subsequently applied to general incompressible flows in [10]. Likewise, we will introduce a way to motivate an expression for the subscales on the element boundaries. These will allow us to consider discontinuous interpolations for either the pressure or the stress, or both. We will restrict ourselves to conforming approximations, and thus the displacement interpolation will be considered continuous.

Other stabilization methods based on projecting the pressure or the pressure gradient *to deal with the incompressibility constraint* can be found in the literature. A simple method based on projecting onto discontinuous pressure spaces of lower order can be found in [13]. In [4] a method based on projecting onto pressures defined on patches of elements is proposed, which can be also interpreted (after appropriate approximations) in the variational multiscale framework [7]. See also [24] for an abstract analysis and generalization of these type of methods. Nevertheless, some conditions on the finite element mesh are often required that are difficult to meet in practical unstructured finite element meshes.

Different stabilized formulations for the three-field Stokes problem can be found in the literature. The GLS (Galerkin/least-squares) method is used, for example, in [5, 16, 27]. In [19, 14] the authors propose what they call EVSS (elastic-viscous-split-stress), which is related to the formulation proposed in this paper in what concerns the way to stabilize the stress interpolation. An analysis of both approaches, GLS and EVSS, is presented in [6].

Even though our interest is to consider incompressible materials and therefore to include the pressure as a variable, a similar formulation to the one proposed here could be applied to other versions of the elasticity problem. The difficulty to devise stable total stress-displacement interpolations is well known (see, for example, [2] and also the general approach adopted in [1]). A stabilized formulation for the stress-displacement-rotation formulation can be found in [17] (in 2D) and [18] (in 3D). In these references the stability of the Galerkin formulation is also enhanced by adding some least-square-type terms. The application of the formulation to be presented to different versions of the elasticity problem would be straightforward.

The paper is organized as follows. In the following section we present the problem to be solved and its Galerkin finite element approximation, explaining the sources of

numerical instability. Then we present the stabilized finite element formulation we propose, for which we present a complete numerical analysis in section 4. The paper concludes with some final remarks.

2. Problem statement and Galerkin finite element discretization.

2.1. Boundary value problem. Let Ω be the computational domain of \mathbb{R}^d ($d = 2$ or 3) occupied by the solid (or fluid), assumed to be bounded and polyhedral, and let $\partial\Omega$ be its boundary. If \mathbf{u} is the displacement field, p the pressure (taken as positive in compression), and $\boldsymbol{\sigma}$ the deviatoric component of the stress field, the field equations to be solved in the domain Ω are

$$(2.1) \quad -\nabla \cdot \boldsymbol{\sigma} + \nabla p = \mathbf{f},$$

$$(2.2) \quad \nabla \cdot \mathbf{u} = 0,$$

$$(2.3) \quad \frac{1}{2\mu} \boldsymbol{\sigma} - \nabla^S \mathbf{u} = \mathbf{0},$$

where \mathbf{f} is the vector of body forces, μ the shear modulus, and $\nabla^S \mathbf{u}$ the symmetrical part of $\nabla \mathbf{u}$. For simplicity, we shall consider the simplest boundary condition $\mathbf{u} = \mathbf{0}$ on $\partial\Omega$.

2.2. Variational form. To write the weak form of problem (2.1)–(2.3) we need to introduce some functional spaces. Let $\mathcal{V} = (H_0^1(\Omega))^d$, $\mathcal{Q} = L^2(\Omega)/\mathbb{R}$, and $\mathcal{T} = (L^2(\Omega))_{\text{sym}}^{d \times d}$, the space of symmetric tensors of rank two with square-integrable components. If we call $U = (\mathbf{u}, p, \boldsymbol{\sigma})$, $\mathcal{X} = \mathcal{V} \times \mathcal{Q} \times \mathcal{T}$, the weak form of the problem consists in finding $U \in \mathcal{X}$ such that

$$(2.4) \quad B(U, V) = L(V),$$

for all $V = (\mathbf{v}, q, \boldsymbol{\tau}) \in \mathcal{X}$, where

$$(2.5) \quad B(U, V) = (\nabla^S \mathbf{v}, \boldsymbol{\sigma}) - (p, \nabla \cdot \mathbf{v}) + (q, \nabla \cdot \mathbf{u}) + \frac{1}{2\mu} (\boldsymbol{\sigma}, \boldsymbol{\tau}) - (\nabla^S \mathbf{u}, \boldsymbol{\tau}),$$

$$(2.6) \quad L(V) = \langle \mathbf{f}, \mathbf{v} \rangle,$$

where (\cdot, \cdot) is the L^2 inner product and $\langle \cdot, \cdot \rangle$ is the duality pairing between \mathcal{V} and its dual, $(H^{-1}(\Omega))^d$, where \mathbf{f} is assumed to belong.

2.3. Stability of the Galerkin finite element discretization. Let us consider a finite element partition \mathcal{P}_h of the domain Ω of diameter h . For simplicity, we will consider quasi-uniform refinements, and thus all the element diameters can be bounded above and below by constants multiplying h . The extension of the following analysis to general shape-regular meshes (also called nondegenerate meshes) can be done using the strategy developed in [11].

From the finite element partition we may build up conforming finite element spaces $\mathcal{V}_h \subset \mathcal{V}$, $\mathcal{Q}_h \subset \mathcal{Q}$, and $\mathcal{T}_h \subset \mathcal{T}$ in the usual manner. If $\mathcal{X}_h = \mathcal{V}_h \times \mathcal{Q}_h \times \mathcal{T}_h$ and $U_h = (\mathbf{u}_h, p_h, \boldsymbol{\sigma}_h)$, the Galerkin finite element approximation consists in finding $U_h \in \mathcal{X}_h$ such that

$$(2.7) \quad B(U_h, V_h) = L(V_h),$$

for all $V_h = (\mathbf{v}_h, q_h, \boldsymbol{\tau}_h) \in \mathcal{X}_h$.

In principle, we have posed no restrictions on the choice of the finite element spaces. However, let us analyze the numerical stability of problem (2.7). If we take $V_h = U_h$, it is found that

$$(2.8) \quad B(U_h, U_h) = \frac{1}{2\mu} \|\boldsymbol{\sigma}_h\|^2,$$

where $\|\cdot\|$ is the $L^2(\Omega)$ norm. It is seen from (2.8) that B_h is not coercive in \mathcal{X}_h , the displacement and the pressure being out of control. Moreover, the inf-sup condition

$$\inf_{U_h \in \mathcal{X}_h} \sup_{V_h \in \mathcal{X}_h} \frac{B(U_h, V_h)}{\|U_h\|_{\mathcal{X}} \|V_h\|_{\mathcal{X}}} \geq \beta$$

is *not* satisfied for any positive constant β unless the two conditions

$$(2.9) \quad \inf_{q_h \in \mathcal{Q}_h} \sup_{\mathbf{v}_h \in \mathcal{V}_h} \frac{(q_h, \nabla \cdot \mathbf{v}_h)}{\|q_h\|_{\mathcal{Q}_h} \|\mathbf{v}_h\|_{\mathcal{V}_h}} \geq C_1,$$

$$(2.10) \quad \inf_{\boldsymbol{\tau}_h \in \mathcal{V}_h} \sup_{\boldsymbol{\tau}_h \in \mathcal{T}_h} \frac{(\boldsymbol{\tau}_h, \nabla^S \mathbf{v}_h)}{\|\boldsymbol{\tau}_h\|_{\mathcal{T}_h} \|\mathbf{v}_h\|_{\mathcal{V}_h}} \geq C_2,$$

hold for positive constants C_1 and C_2 (see, for example, [25]). In all the expressions above, $\|\cdot\|_{\mathcal{Y}}$ stands for the appropriate norm in space \mathcal{Y} .

Conditions (2.9) and (2.10) pose stringent requirements on the choice of the finite element spaces. Our intention in this paper is to present a stabilized finite element formulation that avoids the need for such conditions and, in particular, allows equal interpolation for all the unknowns. However, we will consider the most general case, and we will assume that \mathcal{V}_h , \mathcal{Q}_h , and \mathcal{T}_h are constructed from finite element interpolations of degree k_u , k_p , and k_σ , respectively, being the functions in \mathcal{V}_h continuous but the stress and pressure interpolation possibly discontinuous.

Before closing this section, let us introduce some notation. The finite element partition will be denoted by $\mathcal{P}_h = \{K\}$, and summation over all the elements will be indicated as \sum_K . The collection of all *interior edges* (faces, for $d = 3$) will be denoted by $\mathcal{E}_h = \{E\}$ and, as for the elements, summation over all these edges will be indicated as \sum_E . The symbol $\langle f, g \rangle_D$ will be used to denote the integral of the product of functions f and g over D , with $D = K$ (an element), $D = \partial K$ (an element boundary), or $D = E$ (an edge). Likewise, $\|f\|_D^2 := \langle f, f \rangle_D$. Suppose now that elements K_1 and K_2 share an edge E , and let \mathbf{n}_1 and \mathbf{n}_2 be the normals to E exterior to K_1 and K_2 , respectively. For a scalar function f , possibly discontinuous across E , we define its jump as $\llbracket \mathbf{n} f \rrbracket_E := \mathbf{n}_1 f|_{\partial K_1 \cap E} + \mathbf{n}_2 f|_{\partial K_2 \cap E}$, and for a vector or tensor \mathbf{v} , $\llbracket \mathbf{n} \cdot \mathbf{v} \rrbracket_E := \mathbf{n}_1 \cdot \mathbf{v}|_{\partial K_1 \cap E} + \mathbf{n}_2 \cdot \mathbf{v}|_{\partial K_2 \cap E}$.

3. Design of the stabilized finite element approximation using subscales. In this section we describe the finite element formulation proposed. The arguments in this design step are necessarily heuristic. Their validity depends on the numerical performance of the formulation, which will not be checked here (see the final remarks in section 5), and on the numerical analysis to be presented in the following section.

3.1. Decomposition of the unknowns. Let us start by explaining the basic idea of the multiscale formulation proposed in [20] and applying it to our problem. If we split $U = U_h + U'$, where U_h belongs to the finite element space \mathcal{X}_h and U' to any space \mathcal{X}' to complement \mathcal{X}_h in \mathcal{X} , problem (2.4) is exactly equivalent to

$$(3.1) \quad B(U_h + U', V_h) = L(V_h) \quad \forall V_h \in \mathcal{X}_h,$$

$$(3.2) \quad B(U_h + U', V') = L(V') \quad \forall V' \in \mathcal{X}'.$$

In essence, the goal of all subscale methods, including the approximation with bubble functions, is to approximate U' in one way or another and end up with a problem for U_h alone.

Integrating some terms by parts and using the fact that $\mathbf{u}_h = \mathbf{u}' = \mathbf{0}$ on $\partial\Omega$, it is easy to see that (3.1) in our case can be written as

$$(3.3) \quad \begin{aligned} & B(U_h, V_h) + (\nabla^S \mathbf{v}_h, \boldsymbol{\sigma}') - (p', \nabla \cdot \mathbf{v}_h) + \frac{1}{2\mu} (\boldsymbol{\sigma}', \boldsymbol{\tau}_h) \\ & + \sum_E \langle \mathbf{u}'_E, \llbracket \mathbf{n}q_h - \mathbf{n} \cdot \boldsymbol{\tau}_h \rrbracket \rangle_E + \sum_K \langle \mathbf{u}'_K, -\nabla q_h + \nabla \cdot \boldsymbol{\tau}_h \rangle_K = L(V_h), \end{aligned}$$

where we have distinguished between the displacement subscale in the elements interiors, \mathbf{u}'_K , and on the edges, \mathbf{u}'_E . The stress and pressure subscales are required only in the element interiors (recall that they may be discontinuous).

On the other hand, integrating back some terms by parts in (3.2) it is found that

$$(3.4) \quad \begin{aligned} & \sum_K \langle \mathbf{v}', -\mathbf{n}p + \mathbf{n} \cdot \boldsymbol{\sigma} \rangle_{\partial K} + \sum_K \langle \mathbf{v}', \nabla p - \nabla \cdot \boldsymbol{\sigma} \rangle_K \\ & + (q', \nabla \cdot \mathbf{u}) + \frac{1}{2\mu} (\boldsymbol{\sigma}, \boldsymbol{\tau}') - (\nabla^S \mathbf{u}, \boldsymbol{\tau}') = L(V'), \end{aligned}$$

which yield as Euler–Lagrange equations the original differential equations projected onto \mathcal{X}' , together with the continuity of $-\mathbf{n}p + \mathbf{n} \cdot \boldsymbol{\sigma}$ across interelement boundaries in the corresponding trace space.

Let us denote by P_h the projection with respect to

$$(3.5) \quad (f, g)_h := \sum_K \langle f, g \rangle_K,$$

for f and g such that the integral of their product in each $K \in \mathcal{P}_h$ is well defined. Observe that $(f, g)_h$ coincides with the $L^2(\Omega)$ inner product when $f, g \in L^2(\Omega)$.

With this definition, (3.4) and the continuity of the stresses across interelement boundaries imply

$$(3.6) \quad \left. \begin{aligned} -\nabla \cdot \boldsymbol{\sigma}' + \nabla p' = \mathbf{r}_u & := \mathbf{f} + \nabla \cdot \boldsymbol{\sigma}_h - \nabla p_h + \boldsymbol{\xi}_u \\ \nabla \cdot \mathbf{u}'_K = r_p & := -\nabla \cdot \mathbf{u}_h + \xi_p \\ \frac{1}{2\mu} \boldsymbol{\sigma}' - \nabla^S \mathbf{u}'_K = \mathbf{r}_\sigma & := -\frac{1}{2\mu} \boldsymbol{\sigma}_h + \nabla^S \mathbf{u}_h + \boldsymbol{\xi}_\sigma \end{aligned} \right\} \text{ in each } K \in \mathcal{P}_h$$

$$(3.7) \quad \left. \begin{aligned} \mathbf{u}' & = \mathbf{u}'_E \\ \llbracket \mathbf{n}p - \mathbf{n} \cdot \boldsymbol{\sigma} \rrbracket_E & = \mathbf{0} \end{aligned} \right\} \text{ on each } E \in \mathcal{E}_h,$$

where $\boldsymbol{\xi}_u$, ξ_p , and $\boldsymbol{\xi}_\sigma$ are orthogonal to \mathcal{V}' , \mathcal{Q}' , and \mathcal{T}' , respectively, with respect to projection P_h . These vectors are responsible to enforce that the previous equations hold in the space for the subscales, which still needs to be approximated (see [10] for more details). Clearly, if (3.6) is to be understood in a classical sense, \mathbf{f} should be more regular than required up to now and, likewise, the subscales need to be more regular than required. Nevertheless, for the moment we may assume as much regularity as needed. We will see that the final problem (3.18)–(3.19) is well defined in the functional framework introduced earlier.

The way to approximate the solution of problems (3.6)–(3.7) and to choose the space for the subscales is the topic of the following subsection. The objective is to

obtain a closed form expression for $\boldsymbol{\sigma}'$, p' , and \mathbf{u}'_K defined on the element interiors and for \mathbf{u}'_E defined on the interior edges. Without any further simplification, the problem is as complex as the original one. The essential approximation step consists of approximating (3.6) without taking into account \mathbf{u}'_E and then approximating this unknown assuming the subscales on the element interiors are known.

3.2. Approximation of the subscales in the element interiors. There are several possibilities to deal with problem (3.6). As in [10], we will approximate $\boldsymbol{\sigma}'$, p' , and \mathbf{u}' by using an (approximate) Fourier analysis of the problem. We start explaining the basic idea and then we apply it to problem (3.6).

Let us consider a linear differential equation of the form $\mathcal{L}U = F$ posed in each element domain K , where U is in general a vector unknown corresponding to a subscale, \mathcal{L} a linear differential operator, and F a given vector function. Let us denote the Fourier transform by $\widehat{\cdot}$. Scaling the wave number as \mathbf{k}/h , with \mathbf{k} dimensionless and h being the diameter of K , the basic heuristic assumption is to assume that U is highly fluctuating, and thus dominated by high wave numbers. Thus, the boundary term in the Fourier transform of the derivatives can be considered negligible compared with the term involving the integral in K , since the former is $\mathcal{O}(1)$ and the latter $\mathcal{O}(|\mathbf{k}|)$. This essential approximation amounts to evaluating the Fourier transform of the equation as for functions vanishing on ∂K (and extended to \mathbb{R}^d by zero).

Suppose now that the differential equations are written in such a way that the product $F^t U$ is dimensionally well defined; that is to say, all the terms in the sum have the same dimension. Here and in what follows we assume that U , possibly with a subscript, is an element in the domain of \mathcal{L} and F , may be also with a subscript, is an element in the range of \mathcal{L} . It is obvious that the products $F_1^t F_2$ and $U_1^t U_2$ may not be dimensionally well defined. Let M be a scaling matrix, symmetric, positive-definite, and possibly diagonal, which makes the products $F_1^t M F_2$ and $U_1^t M^{-1} U_2$ dimensionally consistent. We will denote $|F|_M^2 = F^t M F$ and $|U|_{M^{-1}}^2 = U^t M^{-1} U$ and refer to these quantities as the squared M -norm of F and the squared M^{-1} -norm of U , respectively. Likewise, we denote by $\|F\|_{L_M^2(K)}$ the $L^2(K)$ norm of $|F|_M$.

Our purpose is to approximate $\mathcal{L}U \approx \Lambda U$ in a certain sense, with Λ a matrix which has to be determined and that will be called *matrix of stabilization parameters*. We propose to do this imposing that *the induced $L_M^2(K)$ norm of Λ is an upper bound for the induced $L_M^2(K)$ norm of \mathcal{L}* ; that is to say, $\|\mathcal{L}\|_{L_M^2(K)} \leq \|\Lambda\|_{L_M^2(K)}$. The symbol \leq has to be understood up to constants and holding independently of the equation coefficients.

According to the approximation explained, we may write the Fourier transform of $\mathcal{L}U$ as $\widehat{\mathcal{L}(\mathbf{k})\widehat{U}(\mathbf{k})}$, where $\widehat{\mathcal{L}(\mathbf{k})}$ is an algebraic operator. The approximate upper bound of $\|\mathcal{L}\|_{L_M^2(K)}$ can be obtained as follows. For any U in the domain of \mathcal{L} we have

$$\begin{aligned} \|\mathcal{L}U\|_{L_M^2(K)}^2 &= \int_K |\mathcal{L}U|_M^2 d\mathbf{x} \\ &\approx \int_{\mathbb{R}^d} |\widehat{\mathcal{L}(\mathbf{k})\widehat{U}(\mathbf{k})}|_M^2 d\mathbf{k} \\ &\leq \int_{\mathbb{R}^d} |\widehat{\mathcal{L}(\mathbf{k})}|_M^2 |\widehat{U}(\mathbf{k})|_M^2 d\mathbf{k} \\ &= |\widehat{\mathcal{L}(\mathbf{k}^0)}|_M^2 \int_{\mathbb{R}^d} |\widehat{U}(\mathbf{k})|_M^2 d\mathbf{k} \\ &\approx |\widehat{\mathcal{L}(\mathbf{k}^0)}|_M^2 \|U\|_{L_M^2(K)}^2. \end{aligned}$$

In the first and in the last steps we have used Plancherel’s formula for the approximate Fourier transform, whereas \mathbf{k}^0 is a wave number whose existence is guaranteed by the mean value theorem. From the previous result it follows that $\|\mathcal{L}\|_{L^2_M(K)} \leq |\widehat{\mathcal{L}}(\mathbf{k}^0)|_M$ for a certain wave number, still denoted \mathbf{k}^0 . Therefore, *our proposal is to choose Λ such that $|\widehat{\mathcal{L}}(\mathbf{k}^0)|_M = |\Lambda|_M$* . Obviously, the value \mathbf{k}^0 is unknown. Its components have to be understood in this context as algorithmic coefficients.

The norm $|\widehat{\mathcal{L}}(\mathbf{k}^0)|_M$ can be computed as the square root of the maximum eigenvalue (in module) of the generalized eigenvalue problem $\widehat{\mathcal{L}}(\mathbf{k}^0)^t M \widehat{\mathcal{L}}(\mathbf{k}^0) X = \lambda M^{-1} X$. This leads to an effective way to determine the expression of matrix Λ .

The general idea exposed allows one to obtain the correct matrix of stabilization parameters for several problems (see [12] for an obtention of this matrix in the context of the hyperbolic wave equation). In particular, we will apply it now to the design of this matrix for the problem considered in this paper. Furthermore, we will show that in this particular case a simple *dimensional argument is enough to obtain Λ if we assume this matrix is diagonal*.

For the sake of simplicity, let us consider the case $d = 2$ (being obvious the extension to $d = 3$) and let us organize the unknowns as $U = (u_1, u_2, p, \sigma_{11}, \sigma_{12}, \sigma_{22})$. The first point is to choose matrix M . If $[\cdot]$ denotes a dimensional group, from (3.6) it is readily checked that

$$[\mathbf{r}_u]^2 \begin{bmatrix} h^2 \\ \mu^2 \end{bmatrix} = [r_p]^2 = [\mathbf{r}_\sigma]^2, \quad [\mathbf{u}']^2 \begin{bmatrix} \mu^2 \\ h^2 \end{bmatrix} = [p']^2 = [\boldsymbol{\sigma}']^2,$$

and therefore we may take

$$(3.8) \quad M = \text{diag}(m, m, 1, 1, 1, 1), \quad m := \frac{h^2}{\mu^2}.$$

Let us consider matrix Λ of the form

$$\Lambda = \text{diag}(\Lambda_u, \Lambda_u, \Lambda_p, \Lambda_\sigma, \Lambda_\sigma, \Lambda_\sigma).$$

If we apply the strategy presented above to determine Λ_u , Λ_p , and Λ_σ , it turns out that these parameters are uniquely determined by dimensionality. To see this, let us start by noting that if \mathcal{L} is now the operator associated to (3.6), it can be checked that the eigenvalue of the problem

$$M \widehat{\mathcal{L}}(\mathbf{k}^0)^t M \widehat{\mathcal{L}}(\mathbf{k}^0) X = \lambda X$$

has dimensions $[\lambda] = [\mu]^{-2}$, and therefore

$$M \Lambda M \Lambda = \text{diag}(\Lambda_u^2 m^2, \Lambda_u^2 m^2, \Lambda_p^2, \Lambda_\sigma^2, \Lambda_\sigma^2, \Lambda_\sigma^2)$$

has to have all the diagonal entries of dimension $[\mu]^{-2}$. Being μ the only parameter of the equation, this immediately implies that

$$\Lambda_u^{-1} = \alpha_u \frac{h^2}{\mu}, \quad \Lambda_p^{-1} = \alpha_p 2\mu, \quad \Lambda_\sigma^{-1} = \alpha_\sigma 2\mu,$$

where α_u , α_p , and α_σ are constants that play the role of the algorithmic parameters of the formulation. This allows us to approximate the solution of (3.6) as

$$(3.9) \quad \mathbf{u}'_K = \alpha_u \frac{h^2}{\mu} \mathbf{r}_u,$$

$$(3.10) \quad p' = \alpha_p 2\mu r_p,$$

$$(3.11) \quad \boldsymbol{\sigma}' = \alpha_\sigma 2\mu \mathbf{r}_\sigma.$$

These are the expressions we were looking for.

It only remains to determine which is the space of the subscales, that is, to choose the functions $\boldsymbol{\xi}_u$, ξ_p , and $\boldsymbol{\xi}_\sigma$. Our particular choice is *to take the space for the subscales P_h orthogonal to the finite element space* (see (3.5) for the definition of P_h). In view of (3.9)–(3.11), this implies that \mathbf{r}_u , r_p , and \mathbf{r}_σ must be orthogonal to \mathcal{V}_h , \mathcal{Q}_h , and \mathcal{T}_h , respectively. Denoting by P_u , P_p , and P_σ , the P_h projections onto these spaces and by P_u^\perp , P_p^\perp , and P_σ^\perp the orthogonal projections, we will have that

$$\begin{aligned} \boldsymbol{\xi}_u &= -P_u(\mathbf{f} + \nabla \cdot \boldsymbol{\sigma}_h - \nabla p_h) & \text{and} & & \mathbf{u}'_K &= \alpha_u \frac{h^2}{\mu} P_u^\perp(\mathbf{f} + \nabla \cdot \boldsymbol{\sigma}_h - \nabla p_h), \\ \xi_p &= -P_p(-\nabla \cdot \mathbf{u}_h) & \text{and} & & p' &= \alpha_p 2\mu P_p^\perp(-\nabla \cdot \mathbf{u}_h), \\ \boldsymbol{\xi}_\sigma &= -P_\sigma\left(-\frac{1}{2\mu}\boldsymbol{\sigma}_h + \nabla^S \mathbf{u}_h\right) & \text{and} & & \boldsymbol{\sigma}' &= \alpha_\sigma 2\mu P_\sigma^\perp\left(-\frac{1}{2\mu}\boldsymbol{\sigma}_h + \nabla^S \mathbf{u}_h\right). \end{aligned}$$

Clearly, we have that $P_\sigma^\perp(-\boldsymbol{\sigma}_h) = \mathbf{0}$. We may also assume for simplicity that the body force belongs to the finite element space, and thus $P_u^\perp(\mathbf{f}) = \mathbf{0}$. Hence, the expressions for the subscales we finally propose are

$$(3.12) \quad \mathbf{u}'_K = \alpha_u \frac{h^2}{\mu} P_u^\perp(\nabla \cdot \boldsymbol{\sigma}_h - \nabla p_h),$$

$$(3.13) \quad p' = -\alpha_p 2\mu P_p^\perp(\nabla \cdot \mathbf{u}_h),$$

$$(3.14) \quad \boldsymbol{\sigma}' = \alpha_\sigma 2\mu P_\sigma^\perp(\nabla^S \mathbf{u}_h).$$

3.3. Approximation of the displacement subscale on the interelement boundaries. The objective now is to propose an expression for \mathbf{u}'_E in (3.7). Let K_1 and K_2 be two elements sharing an edge E (face, for $d = 3$). The idea is to assume that the expressions (3.12)–(3.14) just obtained for \mathbf{u}'_{K_i} , p'_i , and $\boldsymbol{\sigma}'_i$ on element K_i , $i = 1, 2$, hold up to a distance $\delta = \delta_0 h$, $0 < \delta_0 < 1/2$, to the edge E , and that the normal derivative of \mathbf{u}' on E can be approximated as

$$(3.15) \quad \mathbf{n}_i \cdot \nabla \mathbf{u}'|_{\partial K_i \cap E} \approx \frac{1}{\delta} (\mathbf{u}'_E - \mathbf{u}'_{K_i}), \quad i = 1, 2,$$

which will contribute to the stress on $\partial K_i \cap E$ with

$$\mathbf{n}_i \cdot \boldsymbol{\sigma}'_E|_{\partial K_i \cap E} = 2\mu \mathbf{A}(\mathbf{n}_i \cdot \nabla \mathbf{u}'|_{\partial K_i \cap E}),$$

where tangential derivatives \mathbf{u}' on $\partial K_i \cap E$ have been disregarded and \mathbf{A} is a symmetric and positive-definite matrix which comes from the fact that $\boldsymbol{\sigma}'|_{\partial K_i \cap E}$ has to be approximated by the symmetric gradient of \mathbf{u}' on $\partial K_i \cap E$.

Calling also \mathbf{u}'_{K_i} , p'_i , and $\boldsymbol{\sigma}'_i$, the extension of the subgrid displacement, pressure, and stress computed in the interior of element K_i ($i = 1, 2$) and extended to the boundary, the continuity of the total stress expressed in (3.7) implies

$$\begin{aligned} 0 &= \llbracket \mathbf{n}(p_h + p') - \mathbf{n} \cdot (\boldsymbol{\sigma}_h + \boldsymbol{\sigma}' + \boldsymbol{\sigma}'_E) \rrbracket_E \\ &= \llbracket \mathbf{n}(p_h + p') - \mathbf{n} \cdot (\boldsymbol{\sigma}_h + \boldsymbol{\sigma}') \rrbracket_E - 2\mu \mathbf{A} \llbracket \mathbf{n} \cdot \nabla \mathbf{u}' \rrbracket_E, \end{aligned}$$

and using (3.15)

$$(3.16) \quad \mathbf{u}'_E = \{\mathbf{u}'_K\}_E + \frac{\delta}{2\mu} \mathbf{A}^{-1} \llbracket \mathbf{n}(p_h + p') - \mathbf{n} \cdot (\boldsymbol{\sigma}_h + \boldsymbol{\sigma}') \rrbracket_E,$$

where $\{\mathbf{u}'_K\}_E = (\mathbf{u}'_{K_1} + \mathbf{u}'_{K_2})|_E/2$ is the average of the displacement subscales computed in the element interiors and extended to edge E .

Expression (3.16) can be used as subscale on the element boundaries. In fact, all the analysis presented in section 4 carries over when it is used. However, both from numerical experiments and from the numerical analysis presented later on it turns out that it suffices to use a simpler expression, obtained by keeping the dominant finite element terms in (3.16) and replacing \mathbf{A} by the identity (recall that this is a symmetric and positive-definite matrix). The bottom line is expression

$$(3.17) \quad \mathbf{u}'_E = \frac{\delta}{2\mu} \llbracket \mathbf{n}p_h - \mathbf{n} \cdot \boldsymbol{\sigma}_h \rrbracket_E,$$

which will be used in the following.

3.4. Stabilized finite element problem. Once the approximation for subscales in the element interiors (3.12)–(3.14) and for the displacement subscale on the interior edges (3.17) have been derived, the stabilized finite element problem is obtained by inserting these approximations into (3.3). Noting that $(\boldsymbol{\sigma}', \boldsymbol{\tau}_h) = 0$, the result is the following: Find $U_h \in \mathcal{X}_h$ such that

$$(3.18) \quad B_{\text{stab}}(U_h, V_h) = L(V_h),$$

for all $V_h \in \mathcal{X}_h$, where

$$(3.19) \quad \begin{aligned} B_{\text{stab}}(U_h, V_h) &:= B(U_h, V_h) \\ &+ \alpha_\sigma 2\mu (P_\sigma^\perp(\nabla^S \mathbf{v}_h), P_\sigma^\perp(\nabla^S \mathbf{u}_h)) + \alpha_p 2\mu (P_p^\perp(\nabla \cdot \mathbf{v}_h), P_p^\perp(\nabla \cdot \mathbf{u}_h)) \\ &+ \alpha_u \frac{h^2}{\mu} \sum_K \langle P_u^\perp(\nabla q_h - \nabla \cdot \boldsymbol{\tau}_h), P_u^\perp(\nabla p_h - \nabla \cdot \boldsymbol{\sigma}_h) \rangle_K \\ &+ \frac{\delta_0 h}{2\mu} \sum_E \langle \llbracket \mathbf{n}q_h - \mathbf{n} \cdot \boldsymbol{\tau}_h \rrbracket, \llbracket \mathbf{n}p_h - \mathbf{n} \cdot \boldsymbol{\sigma}_h \rrbracket \rangle_E. \end{aligned}$$

The stabilized finite element method we propose and whose stability and convergence properties are established in the following section is (3.18). In expression (3.19) for the stabilized bilinear form some orthogonal projections are used to highlight the symmetry of the resulting formulation. If P^\perp is any of the orthogonal projections appearing in (3.19) and $P = I - P^\perp$, in the implementation of the method for any discrete functions f_h and g_h one may compute $(P^\perp(f_h), P^\perp(g_h)) = (f_h, g_h - P(g_h))$ and treat $P(g_h)$ either implicitly or in an iterative way, that is, evaluated at a previous iteration of an iterative scheme of any type. For example, denoting with a superscript the iteration counter, in the simplest case $(P^\perp(f_h), P^\perp(g_h^i))$ could be approximated by $(f_h, g_h^i - P(g_h^{i-1}))$ (see [11] for more comments on implementation issues of a similar formulation).

Finally, let us comment on the choice of the constants α_σ , α_p , α_u , and δ_0 . The analysis to be presented next can be applied for any set of values. In some numerical tests using linear and quadratic elements, with both continuous and discontinuous stresses and pressures (although with the same interpolation for $\boldsymbol{\sigma}_h$ and p_h) we have

observed that these parameters can be taken in a wide range with little influence in the results. By default, we use $\alpha_\sigma = \alpha_p = 1$, $\alpha_u = 4$, and $\delta_0 = 1/10$ in our numerical tests.

4. Numerical analysis of the formulation. We present here the numerical analysis of the method proposed in the previous section using heuristic arguments. The norm in which the results will be first presented is

$$(4.1) \quad \begin{aligned} \|V_h\|^2 &:= \frac{1}{2\mu} \|\boldsymbol{\tau}_h\|^2 + \alpha_\sigma 2\mu \|\nabla^S \mathbf{v}_h\|^2 + \alpha_p 2\mu \|\nabla \cdot \mathbf{v}_h\|^2 \\ &+ \alpha_u \frac{h^2}{\mu} \sum_K \|\nabla q_h - \nabla \cdot \boldsymbol{\tau}_h\|_K^2 + \delta_0 \frac{h}{\mu} \sum_E \|[\![\mathbf{n} q_h - \mathbf{n} \cdot \boldsymbol{\tau}_h]\!] \|_E^2, \end{aligned}$$

although later on we will transform our results to “natural” norms. In fact, the term multiplied by α_p is unnecessary, since it already appears in the term multiplied by α_σ . However, we will keep it for generality, to see the effect of the subscale associated to the pressure introduced in the previous section. Moreover it would be essential in the case of some nonconforming elements (not considered in this work) for which the discrete Korn’s inequality does not hold in general (see [22]). In all what follows we will assume that all the numerical parameters α_σ , α_p , α_u and δ_0 are positive.

As it has been mentioned in section 2, we will consider for the sake of conciseness quasi-uniform finite element partitions. Therefore, we assume that there is a constant C_{inv} , independent of the mesh size h (the maximum of all the element diameters), such that

$$(4.2) \quad \|\nabla v_h\|_K \leq C_{\text{inv}} h^{-1} \|v_h\|_K,$$

for all finite element functions v_h defined on $K \in \mathcal{P}_h$. This inequality can be used for scalars, vectors, or tensors. Similarly, the trace inequality

$$(4.3) \quad \|v\|_{\partial K}^2 \leq C_{\text{trace}} (h^{-1} \|v\|_K^2 + h \|\nabla v\|_K^2),$$

is assumed to hold for functions $v \in H^1(K)$, $K \in \mathcal{P}_h$. The last term can be dropped if v is a polynomial on the element domain K . Thus, if φ_h is a piecewise discontinuous polynomial (the pressure or the stresses, in our case) and ψ_h a continuous one, it follows that

$$(4.4) \quad \sum_E \|[\![\mathbf{n} \varphi_h]\!] \|_E^2 \leq 2C_{\text{trace}} h^{-1} \sum_K \|\varphi_h\|_K^2,$$

$$(4.5) \quad \sum_E \|\psi_h\|_E^2 \leq \frac{1}{2} C_{\text{trace}} h^{-1} \sum_K \|\psi_h\|_K^2.$$

In all what follows, C , with or without subscript, will denote a positive constant, independent of the discretization and the physical coefficient μ , and possibly different at different occurrences.

We start proving what is in fact the key result, which states that the formulation proposed is stable in the norm (4.1). This stability is presented in the form of an inf-sup condition:

THEOREM 4.1 (stability). *There is a constant $C > 0$ such that*

$$(4.6) \quad \inf_{U_h \in \mathcal{X}_h} \sup_{V_h \in \mathcal{X}_h} \frac{B_{\text{stab}}(U_h, V_h)}{\|U_h\| \|V_h\|} \geq C.$$

Proof. Let us start noting that, for any function $U_h \in \mathcal{X}_h$, we have

$$(4.7) \quad \begin{aligned} B_{\text{stab}}(U_h, U_h) &= \frac{1}{2\mu} \|\boldsymbol{\sigma}_h\|^2 + \alpha_\sigma 2\mu \|P_\sigma^\perp(\nabla^S \mathbf{u}_h)\|^2 + \alpha_p 2\mu \|P_p^\perp(\nabla \cdot \mathbf{u}_h)\|^2 \\ &+ \alpha_u \frac{h^2}{\mu} \sum_K \|P_u^\perp(\nabla p_h - \nabla \cdot \boldsymbol{\sigma}_h)\|_K^2 + \frac{\delta_0 h}{2\mu} \sum_E \|[\![\mathbf{n} p_h - \mathbf{n} \cdot \boldsymbol{\sigma}_h]\!] \|_E^2. \end{aligned}$$

The basic idea is to obtain control on the components on the finite element space for the terms whose orthogonal components appear in this expression. The key point is that this control comes from the Galerkin terms in the bilinear form B_{stab} .

Let us consider $V_{h1} := \alpha_u \frac{h^2}{\mu} (P_u(\nabla p_h - \nabla \cdot \boldsymbol{\sigma}_h), 0, \mathbf{0})$. Recall that P_u is defined based on elementwise integrals, and thus $P_u(\nabla p_h - \nabla \cdot \boldsymbol{\sigma}_h)$ is well defined. We will use the abbreviation $\mathbf{v}_1 \equiv P_u(\nabla p_h - \nabla \cdot \boldsymbol{\sigma}_h)$. A straightforward application of Schwarz's inequality and the inverse estimate (4.2) leads to

$$(4.8) \quad \begin{aligned} B_{\text{stab}}(U_h, V_{h1}) &\geq B(U_h, V_{h1}) - \alpha_\sigma 2\mu \alpha_u \frac{h^2}{\mu} \frac{C_{\text{inv}}}{h} \|\mathbf{v}_1\| \|P_\sigma^\perp(\nabla^S \mathbf{u}_h)\| \\ &- \alpha_p 2\mu \alpha_u \frac{h^2}{\mu} \frac{C_{\text{inv}}}{h} \|\mathbf{v}_1\| \|P_p^\perp(\nabla \cdot \mathbf{u}_h)\|. \end{aligned}$$

On the other hand,

$$\begin{aligned} B(U_h, V_{h1}) &= \alpha_u \frac{h^2}{\mu} \sum_K (\langle \nabla^S \mathbf{v}_1, \boldsymbol{\sigma}_h \rangle_K - \langle \nabla \cdot \mathbf{v}_1, p_h \rangle_K) \\ &= \alpha_u \frac{h^2}{\mu} \sum_K (-\langle \mathbf{v}_1, \nabla \cdot \boldsymbol{\sigma}_h \rangle_K + \langle \mathbf{v}_1, \nabla p_h \rangle_K) - \alpha_u \frac{h^2}{\mu} \sum_E \langle \mathbf{v}_1, [\![\mathbf{n} p_h - \mathbf{n} \cdot \boldsymbol{\sigma}_h]\!] \rangle_E \\ &\geq \alpha_u \frac{h^2}{\mu} \sum_K \|\mathbf{v}_1\|_K^2 - \alpha_u \frac{h^2}{\mu} \sum_E \|\mathbf{v}_1\|_E \|[\![\mathbf{n} p_h - \mathbf{n} \cdot \boldsymbol{\sigma}_h]\!] \|_E \\ &\geq \alpha_u \frac{h^2}{2\mu} \sum_K \|\mathbf{v}_1\|_K^2 - \alpha_u \frac{h C_{\text{trace}}}{4\mu} \sum_E \|[\![\mathbf{n} p_h - \mathbf{n} \cdot \boldsymbol{\sigma}_h]\!] \|_E^2, \end{aligned}$$

where Young's inequality and (4.5) have been used in the last step. Using this in (4.8) and making use again of Young's inequality, it follows that there exist constants C_{1j} , $j = 1, 2, 3, 4$, such that

$$(4.9) \quad \begin{aligned} B_{\text{stab}}(U_h, V_{h1}) &\geq C_{11} \alpha_u \frac{h^2}{\mu} \|P_u(\nabla p_h - \nabla \cdot \boldsymbol{\sigma}_h)\|^2 - C_{12} \alpha_u \frac{h}{\mu} \sum_E \|[\![\mathbf{n} p_h - \mathbf{n} \cdot \boldsymbol{\sigma}_h]\!] \|_E^2 \\ &- C_{13} \alpha_u \alpha_\sigma^2 \mu \|P_\sigma^\perp(\nabla^S \mathbf{u}_h)\|^2 - C_{14} \alpha_u \alpha_p^2 \mu \|P_p^\perp(\nabla \cdot \mathbf{u}_h)\|^2. \end{aligned}$$

Consider now $V_{h2} := \alpha_p 2\mu (\mathbf{0}, q_2, \mathbf{0})$, where $q_2 \equiv P_p(\nabla \cdot \mathbf{u}_h)$. Note that this function may be discontinuous across interelement boundaries. It turns out that

$$\begin{aligned} B_{\text{stab}}(U_h, V_{h2}) &= \alpha_p 2\mu \|q_2\|^2 + \alpha_u \frac{h^2}{2\mu} \sum_K \langle \nabla q_2, P_u^\perp(\nabla p_h - \nabla \cdot \boldsymbol{\sigma}_h) \rangle_K \\ &+ \delta_0 \frac{h}{\mu} \alpha_p 2\mu \sum_E \langle [\![q_2]\!] , [\![\mathbf{n} p_h - \mathbf{n} \cdot \boldsymbol{\sigma}_h]\!] \rangle_E \end{aligned}$$

The same strategy as before, now using (4.4) to deal with the last term in this expression, leads to the existence of certain constants C_{2j} , $j = 1, 2, 3$, such that

$$(4.10) \quad \begin{aligned} B_{\text{stab}}(U_h, V_{h2}) &\geq C_{21}\alpha_p\mu\|P_p(\nabla \cdot \mathbf{u}_h)\|^2 - C_{22}\alpha_p\alpha_u^2\frac{h^2}{\mu}\|P_u^\perp(\nabla p_h - \nabla \cdot \boldsymbol{\sigma}_h)\|^2 \\ &\quad - C_{23}\alpha_p\delta_0^2\frac{h}{\mu}\sum_E\|[\mathbf{n}p_h - \mathbf{n} \cdot \boldsymbol{\sigma}_h]\|_E^2. \end{aligned}$$

Finally, taking $V_{h3} := \alpha_\sigma 2\mu(\mathbf{0}, 0, -P_\sigma(\nabla^S \mathbf{u}_h))$ we obtain that there exist constants C_{3j} , $j = 1, 2, 3, 4$, such that

$$(4.11) \quad \begin{aligned} B_{\text{stab}}(U_h, V_{h3}) &\geq C_{31}\alpha_\sigma\mu\|P_\sigma(\nabla^S \mathbf{u}_h)\|^2 - C_{32}\alpha_\sigma\frac{1}{\mu}\|\boldsymbol{\sigma}_h\|^2 \\ &\quad - C_{33}\alpha_\sigma\alpha_u^2\frac{h^2}{\mu}\|P_u^\perp(\nabla p_h - \nabla \cdot \boldsymbol{\sigma}_h)\|^2 - C_{34}\alpha_\sigma\delta_0^2\frac{h}{\mu}\sum_E\|[\mathbf{n}p_h - \mathbf{n} \cdot \boldsymbol{\sigma}_h]\|_E^2. \end{aligned}$$

Let $V_h = U_h + \beta_1 V_{h1} + \beta_2 V_{h2} + \beta_3 V_{h3}$, with V_{hi} , $i = 1, 2, 3$, introduced above. Adding up inequalities (4.9)–(4.11) multiplied by β_1 , β_2 , and β_3 , respectively, and adding also (4.7), it is trivially verified that the coefficients β_i , $i = 1, 2, 3$, can be chosen large enough so as to obtain

$$(4.12) \quad B_{\text{stab}}(U_h, V_h) \geq C\|U_h\|^2.$$

On the other hand, we have that

$$\begin{aligned} \|V_{h1}\|^2 &\leq 2\alpha_u^2(\alpha_p + \alpha_\sigma)C_{\text{inv}}^2\frac{h^2}{\mu}\|\nabla p_h - \nabla \cdot \boldsymbol{\sigma}_h\|^2 \leq C\|U_h\|^2, \\ \|V_{h2}\|^2 &\leq 2\mu\alpha_p^2(2\alpha_u C_{\text{inv}}^2 + 4\delta_0 C_{\text{trace}})\|\nabla \cdot \mathbf{u}_h\|^2 \leq C\|U_h\|^2, \\ \|V_{h3}\|^2 &\leq 2\alpha_\sigma^2\mu(1 + 2\alpha_u C_{\text{inv}}^2 + 4\delta_0 C_{\text{trace}})\|\nabla^S \mathbf{u}_h\|^2 \leq C\|U_h\|^2, \end{aligned}$$

from where it follows that $\|V_h\| \leq C\|U_h\|$. Using this fact in (4.12) we have shown that for each $U_h \in \mathcal{X}_h$ there exists $V_h \in \mathcal{X}_h$ such that $B_{\text{stab}}(U_h, V_h) \geq C\|U_h\|\|V_h\|$, from where the theorem follows. \square

Once stability is established, a more or less standard procedure leads to convergence. To prove it, we need two preliminary lemmas. The first concerns the consistency of the formulation:

LEMMA 4.2 (consistency). *Let $U \in \mathcal{X}$ be the solution of the continuous problem and $U_h \in \mathcal{X}_h$ the finite element solution of (3.18). If $\mathbf{f} \in \mathcal{V}_h$ and U is regular enough, so that $B_{\text{stab}}(U, V_h)$ is well defined, then*

$$(4.13) \quad B_{\text{stab}}(U - U_h, V_h) = 0 \quad \forall V_h \in \mathcal{X}_h.$$

Proof. This lemma is a trivial consequence of the consistency of the finite element method proposed (considering the force term \mathbf{f} in the finite element space). Note that all the terms added to B in the definition (3.19) of B_{stab} vanish if U_h is replaced by U (recall that $\boldsymbol{\sigma}_h$ could have been added to $\nabla^S \mathbf{u}_h$, since $P_\sigma^\perp(\boldsymbol{\sigma}_h) = \mathbf{0}$). \square

Remark 4.1. If $P_u^\perp(\mathbf{f}) \neq \mathbf{0}$ there are two options. The first is to include this orthogonal projection in the definition of the method, and therefore to modify the right-hand side of (3.18). All the analysis carries over to this case. The second is to take into account the consistency error coming from \mathbf{f} in (4.13). It is easy to see that

in this case this equation can be replaced by $B_{\text{stab}}(U - U_h, V_h) \leq CE(h)\|V_h\|$, where $E(h)$ is introduced below, and the following results can be immediately adapted.

The second preliminary lemma concerns an interpolation error in terms of the norm $\|\cdot\|$ and the bilinear form B_{stab} for the continuous solution $U = (\mathbf{u}, p, \boldsymbol{\sigma}) \in \mathcal{X}$, assumed to have enough regularity. Let \mathcal{W}_h be a finite element space of degree k_v . For any function $v \in H^{k'_v+1}(\Omega)$ and for $i = 0, 1$, we define the interpolation errors $\varepsilon_i(v)$ from the interpolation estimates

$$(4.14) \quad \inf_{v_h \in \mathcal{W}_h} \sum_K \|v - v_h\|_{H^i(K)} \leq Ch^{k''_v+1-i} \sum_K \|v\|_{H^{k''_v+1}(K)} =: \varepsilon_i(v),$$

where $k''_v = \min(k_v, k'_v)$. We will denote by \tilde{v}_h the best approximation of v in \mathcal{W}_h . Clearly, we have that $\varepsilon_0(v) = h\varepsilon_1(v)$. We will use this notation for $v = \mathbf{u}$ (displacement), $v = p$ (pressure) and $v = \boldsymbol{\sigma}$ (stresses), being the respective orders of interpolation k_u, k_p and k_σ .

This notation will allow us to prove that the error function of the method is

$$(4.15) \quad E(h) := \sqrt{\mu}\varepsilon_1(\mathbf{u}) + \frac{1}{\sqrt{\mu}}\varepsilon_0(p) + \frac{1}{\sqrt{\mu}}\varepsilon_0(\boldsymbol{\sigma}).$$

This is indeed the interpolation error:

LEMMA 4.3 (interpolation error). *Let $U \in \mathcal{X}$ be the continuous solution, assumed to be regular enough, and $\tilde{U}_h \in \mathcal{X}_h$ its best finite element approximation. Then, the following inequalities hold:*

$$(4.16) \quad B_{\text{stab}}(U - \tilde{U}_h, V_h) \leq CE(h)\|V_h\|,$$

$$(4.17) \quad \|U - \tilde{U}_h\| \leq CE(h),$$

where $E(h)$ is given in (4.15).

Proof. Let us start considering a general discontinuous finite element interpolation of a function v . Using the trace inequality (4.3) we have that

$$(4.18) \quad \begin{aligned} \sum_E \|[\![\mathbf{n}(v - \tilde{v}_h)]\!] \|_E^2 &\leq 2 \sum_K \|v - \tilde{v}_h\|_{\partial K}^2 \\ &\leq 2C_{\text{trace}} \sum_K (h^{-1}\|v - \tilde{v}_h\|_K^2 + h\|\nabla v - \nabla \tilde{v}_h\|_K^2) \\ &\leq C (h^{-1}\varepsilon_0^2(v) + h\varepsilon_1^2(v)). \end{aligned}$$

The same estimate holds for a continuous interpolation:

$$(4.19) \quad \sum_E \|(v - \tilde{v}_h)\|_E^2 \leq C (h^{-1}\varepsilon_0^2(v) + h\varepsilon_1^2(v)).$$

Let us prove (4.17). By the definition (4.1) of the norm $\|\cdot\|$ and the result just obtained it is immediately checked that

$$\begin{aligned} \|U - \tilde{U}_h\|^2 &\leq C \left[\frac{1}{2\mu}\varepsilon_0^2(\boldsymbol{\sigma}) + \alpha_\sigma 2\mu\varepsilon_1^2(\mathbf{u}) + \alpha_p 2\mu\varepsilon_1^2(\mathbf{u}) \right. \\ &\quad \left. + \alpha_u \frac{h^2}{\mu}\varepsilon_1^2(p) + \alpha_u \frac{h^2}{\mu}\varepsilon_1^2(\boldsymbol{\sigma}) + \delta_0 \frac{h^2}{\mu}\varepsilon_1^2(p) + \delta_0 \frac{h^2}{\mu}\varepsilon_1^2(\boldsymbol{\sigma}) \right], \end{aligned}$$

and (4.17) follows.

Let $\mathbf{e}_u = \mathbf{u} - \tilde{\mathbf{u}}_h$, $e_p = p - \tilde{p}_h$, and $\mathbf{e}_\sigma = \boldsymbol{\sigma} - \tilde{\boldsymbol{\sigma}}_h$. The proof of (4.16) is as follows:

$$\begin{aligned}
B_{\text{stab}}(U - \tilde{U}_h, V_h) &= (\nabla^S \mathbf{v}_h, \mathbf{e}_\sigma) - (e_p, \nabla \cdot \mathbf{v}_h) + \frac{1}{2\mu}(\boldsymbol{\tau}_h, \mathbf{e}_\sigma) \\
&\quad - \sum_K \langle -\nabla q_h + \nabla \cdot \boldsymbol{\tau}_h, \mathbf{e}_u \rangle_K + \sum_E \langle \llbracket \mathbf{n}q_h - \mathbf{n} \cdot \boldsymbol{\tau}_h \rrbracket, \mathbf{e}_u \rangle_E \\
&\quad + \alpha_\sigma (P_\sigma^\perp(\nabla^S \mathbf{v}_h), P_\sigma^\perp(2\mu \nabla^S \mathbf{e}_u - \mathbf{e}_\sigma)) + \alpha_p 2\mu (P_\sigma^\perp(\nabla \cdot \mathbf{v}_h), P_\sigma^\perp(\nabla \cdot \mathbf{e}_u)) \\
&\quad + \delta_0 \frac{h}{\mu} \sum_E \langle \llbracket \mathbf{n}q_h - \mathbf{n} \cdot \boldsymbol{\tau}_h \rrbracket, \llbracket \mathbf{n}e_p - \mathbf{n} \cdot \mathbf{e}_\sigma \rrbracket \rangle_E \\
&\leq C \left[\sqrt{\mu} \|\nabla^S \mathbf{v}_h\| \frac{1}{\sqrt{\mu}} \|\mathbf{e}_\sigma\| + \sqrt{\mu} \|\nabla \cdot \mathbf{v}_h\| \frac{1}{\sqrt{\mu}} \|e_p\| + \frac{1}{2\sqrt{\mu}} \|\boldsymbol{\tau}_h\| \frac{1}{\sqrt{\mu}} \|\mathbf{e}_\sigma\| \right. \\
&\quad + \sum_K \frac{h}{\sqrt{\mu}} \|\nabla q_h - \nabla \cdot \boldsymbol{\tau}_h\|_K \frac{\sqrt{\mu}}{h} \|\mathbf{e}_u\|_K + \sum_E \frac{\sqrt{h}}{\sqrt{\mu}} \|\llbracket \mathbf{n}q_h - \mathbf{n} \cdot \boldsymbol{\tau}_h \rrbracket\|_E \frac{\sqrt{\mu}}{\sqrt{h}} \|\mathbf{e}_u\|_E \\
&\quad + \sqrt{\mu} \|\nabla^S \mathbf{v}_h\| \sqrt{\mu} \|\nabla^S \mathbf{e}_u\| + \sqrt{\mu} \|\nabla^S \mathbf{v}_h\| \frac{1}{\sqrt{\mu}} \|\mathbf{e}_\sigma\| + \sqrt{\mu} \|\nabla \cdot \mathbf{v}_h\| \sqrt{\mu} \|\nabla \cdot \mathbf{e}_u\| \\
&\quad \left. + \sum_E \frac{\sqrt{h}}{\sqrt{\mu}} \|\llbracket \mathbf{n}q_h - \mathbf{n} \cdot \boldsymbol{\tau}_h \rrbracket\|_E \frac{\sqrt{h}}{\sqrt{\mu}} (\|\llbracket \mathbf{n}e_p \rrbracket\|_E + \|\llbracket \mathbf{n} \cdot \mathbf{e}_\sigma \rrbracket\|_E) \right].
\end{aligned}$$

All the terms have been organized to see that, after making use of (4.18) and (4.19), they are all bounded by $CE(h)\|V_h\|$, from where (4.16) follows. \square

We are finally in a position to prove convergence. The proof is standard, but we include it for completeness.

THEOREM 4.4 (convergence). *Let $U = (\mathbf{u}, p, \boldsymbol{\sigma}) \in \mathcal{X}$ be the solution of the continuous problem. Then, there is a constant $C > 0$ such that*

$$\|U - U_h\| \leq CE(h),$$

where $E(h)$ is given in (4.15).

Proof. Consider the finite element function $\tilde{U}_h - U_h \in \mathcal{X}_h$ where, as in Lemma 4.3, $\tilde{U}_h \in \mathcal{X}_h$ is the best finite element approximation to U . Starting from the inf-sup condition (4.6), it follows that there exists $V_h \in \mathcal{X}_h$ such that

$$\begin{aligned}
C\|\tilde{U}_h - U_h\| \|V_h\| &\leq B_{\text{stab}}(\tilde{U}_h - U_h, V_h) \\
&= B_{\text{stab}}(\tilde{U}_h - U, V_h) \quad (\text{from the consistency (4.13)}) \\
&\leq CE(h)\|V_h\| \quad (\text{from (4.16)}),
\end{aligned}$$

from where $\|\tilde{U}_h - U_h\| \leq CE(h)$. The theorem follows now from the triangle inequality $\|U - U_h\| \leq \|U - \tilde{U}_h\| + \|\tilde{U}_h - U_h\|$ and the interpolation error estimate (4.17). \square

Clearly, this convergence result is optimal.

Remark 4.2. From the expression of the error function (4.15) it follows that all the terms have the same order in h if $k_u = k_p + 1 = k_\sigma + 1$. However, Theorem 4.4 holds without any restriction on the interpolation order of the different unknowns.

The next step will be to prove stability and convergence in natural norms, that is to say, in the norm of the space where the continuous problem is posed, and not in the mesh dependent norm (4.1). Even though the results to be presented are the

expected ones, the analysis presented up to this point has highlighted the role played by the stabilization terms of the formulation.

THEOREM 4.5 (stability and convergence in natural norms). *The solution of the discrete problem $U_h = (\mathbf{u}_h, p_h, \boldsymbol{\sigma}_h) \in \mathcal{X}_h$ can be bounded as*

$$(4.20) \quad \sqrt{\mu} \|\mathbf{u}_h\|_{H^1(\Omega)} + \frac{1}{\sqrt{\mu}} \|\boldsymbol{\sigma}_h\| + \frac{1}{\sqrt{\mu}} \|p_h\| \leq \frac{C}{\sqrt{\mu}} \|\mathbf{f}\|_{H^{-1}(\Omega)}.$$

Moreover, if the solution of the continuous problem $U = (\mathbf{u}, p, \boldsymbol{\sigma}) \in \mathcal{X}$ is regular enough, the following error estimate holds:

$$(4.21) \quad \sqrt{\mu} \|\mathbf{u} - \mathbf{u}_h\|_{H^1(\Omega)} + \frac{1}{\sqrt{\mu}} \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\| + \frac{1}{\sqrt{\mu}} \|p - p_h\| \leq CE(h).$$

Proof. Let us first recall that Korn’s inequality implies that $\|\nabla^S \mathbf{v}\|$ is a norm in \mathcal{V} equivalent to $\|\mathbf{v}\|_{H^1(\Omega)}$, and this property is inherited by the conforming approximation considered. On the other hand, it is clear that

$$\langle \mathbf{f}, \mathbf{v}_h \rangle \leq \frac{C}{\sqrt{\mu}} \|\mathbf{f}\|_{H^{-1}(\Omega)} \sqrt{\mu} \|\mathbf{v}_h\|_{H^1(\Omega)} \leq \frac{C}{\sqrt{\mu}} \|\mathbf{f}\|_{H^{-1}(\Omega)} \|V_h\|,$$

where $V_h = (\mathbf{v}_h, q_h, \boldsymbol{\tau}_h) \in \mathcal{X}_h$ is arbitrary. Therefore the inf-sup condition proved in Theorem 4.1 implies that $\|U_h\| \leq \frac{C}{\sqrt{\mu}} \|\mathbf{f}\|_{H^{-1}(\Omega)}$, which, together with the definition of $\|\cdot\|$ in (4.1), yields the bound (4.20) for the first two terms in the left-hand side of this inequality. More precisely, we have that

$$(4.22) \quad \begin{aligned} & \mu \|\mathbf{u}_h\|_{H^1(\Omega)}^2 + \frac{1}{\mu} \|\boldsymbol{\sigma}_h\|^2 \\ & + \frac{h^2}{\mu} \sum_K \|\nabla p_h - \nabla \cdot \boldsymbol{\sigma}_h\|_K^2 + \frac{h}{\mu} \sum_E \|[\![\mathbf{n} p_h - \mathbf{n} \cdot \boldsymbol{\sigma}_h]\!] \|_E^2 \leq \frac{C}{\mu} \|\mathbf{f}\|_{H^{-1}(\Omega)}^2. \end{aligned}$$

On the other hand, using the inverse estimate (4.2) and the trace inequality (4.3) we have

$$\begin{aligned} \frac{h^2}{\mu} \sum_K \|\nabla p_h\|_K^2 & \leq \frac{h^2}{\mu} \sum_K \|\nabla p_h - \nabla \cdot \boldsymbol{\sigma}_h\|_K^2 + \frac{C}{\mu} \|\boldsymbol{\sigma}_h\|^2, \\ \frac{h}{\mu} \sum_E \|[\![\mathbf{n} p_h]\!] \|_E^2 & \leq \frac{h}{\mu} \sum_E \|[\![\mathbf{n} p_h - \mathbf{n} \cdot \boldsymbol{\sigma}_h]\!] \|_E^2 + \frac{C}{\mu} \|\boldsymbol{\sigma}_h\|^2, \end{aligned}$$

so that (4.22) implies

$$(4.23) \quad \mu \|\mathbf{u}_h\|_{H^1(\Omega)}^2 + \frac{1}{\mu} \|\boldsymbol{\sigma}_h\|^2 + \frac{h^2}{\mu} \sum_K \|\nabla p_h\|_K^2 + \frac{h}{\mu} \sum_E \|[\![\mathbf{n} p_h]\!] \|_E^2 \leq \frac{C}{\mu} \|\mathbf{f}\|_{H^{-1}(\Omega)}^2.$$

To prove the L^2 -stability for the pressure we rely on the inf-sup condition between the velocity and pressure spaces that holds for the continuous problem, that is to say, the continuous counterpart of (2.9). If p_h is the solution of the discrete problem, there exists $\mathbf{w} \in \mathcal{V}$ such that

$$C \|p_h\| \|\mathbf{w}\|_{H^1(\Omega)} \leq (p_h, \nabla \cdot \mathbf{w}).$$

Let us choose \mathbf{w} with $\|\mathbf{w}\|_{H^1(\Omega)} = \|p_h\|$ and let $\tilde{\mathbf{w}}_h$ be the best approximation to \mathbf{w} in \mathcal{V}_h , which will satisfy $\|\mathbf{w} - \tilde{\mathbf{w}}_h\| \leq Ch\|p_h\|$. Using (4.3) once again we have that

$$\begin{aligned} C\|p_h\|^2 &\leq (p_h, \nabla \cdot \mathbf{w}) \\ &= - \sum_K \langle \nabla p_h, \mathbf{w} - \tilde{\mathbf{w}}_h \rangle_K + \sum_E \langle \llbracket \mathbf{n} p_h \rrbracket, \mathbf{w} - \tilde{\mathbf{w}}_h \rangle_E \\ &\quad + (\boldsymbol{\sigma}_h, \nabla^S \tilde{\mathbf{w}}_h) - \langle \mathbf{f}, \tilde{\mathbf{w}}_h \rangle \\ &\leq C\|p_h\| \left(h \sum_K \|\nabla p_h\|_K + \sqrt{h} \sum_E \|\llbracket \mathbf{n} p_h \rrbracket\|_E + \|\boldsymbol{\sigma}_h\| + \|\mathbf{f}\|_{H^{-1}(\Omega)} \right). \end{aligned}$$

This, together with (4.23), implies the stability estimate (4.20).

The error estimate can be proved using a similar strategy. First, let us notice that Theorem 4.4 implies the error estimate (4.21) for the displacement and the stresses. We thus have

$$\begin{aligned} (4.24) \quad &\mu \|\mathbf{u} - \mathbf{u}_h\|_{H^1(\Omega)}^2 + \frac{1}{\mu} \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|^2 \\ &\quad + \frac{h^2}{\mu} \sum_K \|\nabla(p - p_h) - \nabla \cdot (\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|_K^2 \\ &\quad + \frac{h}{\mu} \sum_E \|\llbracket \mathbf{n}(p - p_h) - \mathbf{n} \cdot (\boldsymbol{\sigma} - \boldsymbol{\sigma}_h) \rrbracket\|_E^2 \leq CE(h)^2. \end{aligned}$$

On the other hand, using the interpolation estimates (4.14) and (4.18)

$$\begin{aligned} \frac{h^2}{\mu} \sum_K \|\nabla(p - p_h)\|_K^2 &\leq \frac{h^2}{\mu} \sum_K \|\nabla(p - p_h) - \nabla \cdot (\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|_K^2 + \frac{C}{\mu} \varepsilon_0^2(\boldsymbol{\sigma}), \\ \frac{h}{\mu} \sum_E \|\llbracket \mathbf{n}(p - p_h) \rrbracket\|_E^2 &\leq \frac{h}{\mu} \sum_E \|\llbracket \mathbf{n}(p - p_h) - \mathbf{n} \cdot (\boldsymbol{\sigma} - \boldsymbol{\sigma}_h) \rrbracket\|_E^2 + \frac{C}{\mu} \varepsilon_0^2(\boldsymbol{\sigma}), \end{aligned}$$

and, according to (4.24), both terms are bounded by $E(h)^2$. To prove the L^2 -error estimate for the pressure, let now $\mathbf{w} \in \mathcal{V}$, with $\|\mathbf{w}\|_{H^1(\Omega)} = \|p - p_h\|$, be such that $C\|p - p_h\|^2 \leq (p - p_h, \nabla \cdot \mathbf{w})$, and let $\tilde{\mathbf{w}}_h$ be its best approximation in \mathcal{V}_h . We have that

$$\begin{aligned} C\|p - p_h\|^2 &\leq (p - p_h, \nabla \cdot \mathbf{w}) \\ &= - \sum_K \langle \nabla(p - p_h), \mathbf{w} - \tilde{\mathbf{w}}_h \rangle_K + \sum_E \langle \llbracket \mathbf{n}(p - p_h) \rrbracket, \mathbf{w} - \tilde{\mathbf{w}}_h \rangle_E \\ &\quad + (\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \nabla^S \tilde{\mathbf{w}}_h) \\ &\leq C\|p - p_h\| \left(h \sum_K \|\nabla(p - p_h)\|_K + \sqrt{h} \sum_E \|\llbracket \mathbf{n}(p - p_h) \rrbracket\|_E + \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\| \right), \end{aligned}$$

which yields $\|p - p_h\| \leq C\sqrt{\mu}E(h)$. This, together with (4.24), finishes the proof of (4.21). \square

To complete the analysis of the problem, let us obtain an L^2 -error estimate for the displacement, which can be proved using a duality argument.

THEOREM 4.6 (L^2 -error estimate for the velocity). *Suppose that the continuous problem satisfies the elliptic regularity condition*

$$(4.25) \quad \sqrt{\mu} \|\mathbf{u}\|_{H^2(\Omega)} + \frac{1}{\sqrt{\mu}} \|\boldsymbol{\sigma}\|_{H^1(\Omega)} + \frac{1}{\sqrt{\mu}} \|p\|_{H^1(\Omega)} \leq \frac{C}{\sqrt{\mu}} \|\mathbf{f}\|.$$

Then

$$(4.26) \quad \sqrt{\mu} \|\mathbf{u} - \mathbf{u}_h\| \leq Ch \left(\sqrt{\mu} \|\mathbf{u} - \mathbf{u}_h\|_{H^1(\Omega)} + \frac{1}{\sqrt{\mu}} \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\| + \frac{1}{\sqrt{\mu}} \|p - p_h\| \right).$$

Proof. Let $(\boldsymbol{\omega}, \pi, \mathbf{S}) \in \mathcal{X}$ be the solution of the following adjoint problem:

$$(4.27) \quad \nabla \cdot \mathbf{S} - \nabla \pi = \frac{\mu}{\ell^2} (\mathbf{u} - \mathbf{u}_h) \quad \text{in } \Omega,$$

$$(4.28) \quad -\nabla \cdot \boldsymbol{\omega} = 0 \quad \text{in } \Omega,$$

$$(4.29) \quad \frac{1}{2\mu} \mathbf{S} + \nabla^S \boldsymbol{\omega} = \mathbf{0} \quad \text{in } \Omega,$$

with $\boldsymbol{\omega} = \mathbf{0}$ on $\partial\Omega$ and where ℓ is a characteristic length scale of the problem that has been introduced to keep the dimensionality, but that will play no role in the final result. Let also $(\tilde{\boldsymbol{\omega}}_h, \tilde{\pi}_h, \tilde{\mathbf{S}}_h)$ be the best approximation to $(\boldsymbol{\omega}, \pi, \mathbf{S})$ in \mathcal{X}_h . Testing (4.27) with $\mathbf{u} - \mathbf{u}_h$, (4.28) with $p - p_h$, and (4.29) with $\boldsymbol{\sigma} - \boldsymbol{\sigma}_h$, we immediately obtain

$$(4.30) \quad \begin{aligned} \frac{\mu}{\ell^2} \|\mathbf{u} - \mathbf{u}_h\|^2 &= B((\mathbf{u} - \mathbf{u}_h, p - p_h, \boldsymbol{\sigma} - \boldsymbol{\sigma}_h), (\boldsymbol{\omega}, \pi, \mathbf{S})) \\ &= B_{\text{stab}}((\mathbf{u} - \mathbf{u}_h, p - p_h, \boldsymbol{\sigma} - \boldsymbol{\sigma}_h), (\boldsymbol{\omega}, \pi, \mathbf{S})) \\ &\quad - \alpha_\sigma 2\mu \sum_K \left\langle P_\sigma^\perp \left(\frac{1}{2\mu} \mathbf{S} + \nabla^S \boldsymbol{\omega} \right), P_\sigma^\perp(\nabla^S(\mathbf{u} - \mathbf{u}_h)) \right\rangle_K \\ &\quad - \alpha_p 2\mu \sum_K \left\langle P_\sigma^\perp(\nabla \cdot \boldsymbol{\omega}), P_\sigma^\perp(\nabla \cdot (\mathbf{u} - \mathbf{u}_h)) \right\rangle_K \\ &\quad - \alpha_u \frac{h^2}{\mu} \sum_K \left\langle P_u^\perp(\nabla \pi - \nabla \cdot \mathbf{S}), P_u^\perp(\nabla(p - p_h) - \nabla \cdot (\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)) \right\rangle_K \\ &\quad - \delta_0 \frac{h}{2\mu} \sum_E \langle \llbracket \mathbf{n}\pi - \mathbf{n} \cdot \mathbf{S} \rrbracket, \llbracket \mathbf{n}(p - p_h) - \mathbf{n} \cdot (\boldsymbol{\sigma} - \boldsymbol{\sigma}_h) \rrbracket \rangle_E, \end{aligned}$$

where we have made use of the definition (3.19) of B_{stab} . Note that we have included \mathbf{S} in $P_\sigma^\perp(\frac{1}{2\mu} \mathbf{S} + \nabla^S \boldsymbol{\omega})$ because it does not affect the definition of B_{stab} when applied to discrete finite element functions.

The second and third terms in the right-hand side of (4.30) are zero because of (4.29) and (4.28), respectively, and the last one is also zero because of the weak continuity of the stresses associated to problems (4.27)–(4.29). Therefore, only the first and fourth terms need to be bounded.

Using Lemma 4.2, for the first term in (4.30) we have

$$(4.31) \quad \begin{aligned} &B_{\text{stab}}((\mathbf{u} - \mathbf{u}_h, p - p_h, \boldsymbol{\sigma} - \boldsymbol{\sigma}_h), (\boldsymbol{\omega}, \pi, \mathbf{S})) \\ &= B_{\text{stab}}((\mathbf{u} - \mathbf{u}_h, p - p_h, \boldsymbol{\sigma} - \boldsymbol{\sigma}_h), (\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}_h, \pi - \tilde{\pi}_h, \mathbf{S} - \tilde{\mathbf{S}}_h)). \end{aligned}$$

Using the interpolation properties and the shift assumption (4.25) it follows that

$$\begin{aligned} \|\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}_h\|_{H^1(\Omega)} &\leq Ch \|\boldsymbol{\omega}\|_{H^2(\Omega)} \leq Ch \frac{1}{\ell^2} \|\mathbf{u} - \mathbf{u}_h\|, \\ \|\mathbf{S} - \tilde{\mathbf{S}}_h\| &\leq Ch \|\mathbf{S}\|_{H^1(\Omega)} \leq Ch \frac{\mu}{\ell^2} \|\mathbf{u} - \mathbf{u}_h\|, \\ \|\pi - \tilde{\pi}_h\| &\leq Ch \|\pi\|_{H^1(\Omega)} \leq Ch \frac{\mu}{\ell^2} \|\mathbf{u} - \mathbf{u}_h\|. \end{aligned}$$

From these expressions it can be easily checked that (4.31) can be bounded by

$$(4.32) \quad B_{\text{stab}}((\mathbf{u} - \mathbf{u}_h, p - p_h, \boldsymbol{\sigma} - \boldsymbol{\sigma}_h), (\boldsymbol{\omega}, \pi, \mathbf{S})) \leq Ch \frac{\sqrt{\mu}}{\ell^2} \|\mathbf{u} - \mathbf{u}_h\| \left(\sqrt{\mu} \|\mathbf{u} - \mathbf{u}_h\|_{H^1(\Omega)} + \frac{1}{\sqrt{\mu}} \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\| + \frac{1}{\sqrt{\mu}} \|p - p_h\| \right).$$

Let us check this bound for example for the term in $B_{\text{stab}}((\mathbf{u} - \mathbf{u}_h, p - p_h, \boldsymbol{\sigma} - \boldsymbol{\sigma}_h), (\boldsymbol{\omega}, \pi, \mathbf{S}))$ involving boundary integrals, for which we have

$$\begin{aligned} & \delta_0 \frac{h}{\mu} \sum_E \langle \llbracket \mathbf{n}(\tilde{\pi}_h - \pi) - \mathbf{n} \cdot (\tilde{\mathbf{S}}_h - \mathbf{S}) \rrbracket, \llbracket \mathbf{n}(p_h - p) - \mathbf{n} \cdot (\boldsymbol{\sigma}_h - \boldsymbol{\sigma}) \rrbracket \rangle_E \\ & \leq C \frac{h}{\mu} \left[h^{-1/2} (\|\tilde{\pi}_h - \pi\| + \|\tilde{\mathbf{S}}_h - \mathbf{S}\|) + h^{1/2} (\|\tilde{\pi}_h - \pi\|_{H^1(\Omega)} + \|\tilde{\mathbf{S}}_h - \mathbf{S}\|_{H^1(\Omega)}) \right] \\ & \quad \times \left[h^{-1/2} (\|p_h - p\| + \|\boldsymbol{\sigma}_h - \boldsymbol{\sigma}\|) + h^{1/2} (\|p_h - p\|_{H^1(\Omega)} + \|\boldsymbol{\sigma}_h - \boldsymbol{\sigma}\|_{H^1(\Omega)}) \right] \\ & \leq C \frac{h}{\ell^2} \|\mathbf{u} - \mathbf{u}_h\| (\|p - p_h\| + \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|). \end{aligned}$$

The rest of the terms in $B_{\text{stab}}((\mathbf{u} - \mathbf{u}_h, p - p_h, \boldsymbol{\sigma} - \boldsymbol{\sigma}_h), (\boldsymbol{\omega}, \pi, \mathbf{S}))$ can be bounded similarly. We omit the details.

It only remains to bound the fourth term in (4.30). This is again easily done using that $\|\mathbf{S}\|_{H^1(\Omega)} + \|\pi\|_{H^1(\Omega)} \leq C \frac{\mu}{\ell^2} \|\mathbf{u} - \mathbf{u}_h\|$, which yields

$$\begin{aligned} & \alpha_u \frac{h^2}{\mu} \sum_K \langle P_u^\perp(\nabla \pi - \nabla \cdot \mathbf{S}), P_u^\perp(\nabla(p - p_h) - \nabla \cdot (\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)) \rangle_K \\ & \leq C \frac{h^2}{\mu} \frac{\mu}{\ell^2} \|\mathbf{u} - \mathbf{u}_h\| (\|\nabla p - \nabla p_h\| + \|\nabla \cdot \boldsymbol{\sigma} - \nabla \cdot \boldsymbol{\sigma}_h\|). \end{aligned}$$

Using this and (4.32) in (4.30) the theorem follows. \square

5. Concluding remarks. Let us conclude with some remarks concerning the numerical formulation presented in this paper. This formulation is an application of subgrid scale concept to the stress-displacement-pressure formulation of the Stokes problem. Apart from the novelty of this application, a feature of the formulation is to consider the spaces of subgrid scales orthogonal to the finite element spaces. Other ingredients original of this paper are the basis for the design of the parameters of formulation and the introduction of subgrid scales on the element boundaries.

From the point of view of the numerical analysis, the method presented is stable and optimally accurate *using arbitrary interpolations for the displacement, the pressure and the stresses*. Comparing it with the Galerkin method using stable interpolations, exactly the same regularity requirements are needed and the same convergence rates are obtained, also in the same norms. Therefore, the main goal has been achieved.

The accuracy of the method obtained in some numerical experiments is the one expected from the convergence analysis. Theoretical convergence rates are exactly recovered. We have preferred to skip the results of numerical testing in the linear setting analyzed in this paper and to postpone them for a more extensive numerical experimentation in more complex applications.

The practical interest of the problem studied is obvious. As it has been mentioned in the Introduction, this is nothing but a model for more complex situations. Typically, viscoelastic flows are often posed as an example of a problem that requires the

interpolation of the stresses, but this can also be done for nonlinear models such as damage or plasticity in solid mechanics, and non-Newtonian fluids or even turbulence models in fluid mechanics. When designing an extension of the formulations presented here to these more complex situations, the most important idea to bear in mind is which is the stabilization mechanism introduced by the formulations proposed. The analysis dictates that pressure is stabilized by the term proportional to $P_u^\perp(\nabla p_h)$ introduced in the continuity equation, and the displacement gradient is stabilized by the term proportional to $P_\sigma^\perp(\nabla^S \mathbf{u}_h)$ introduced in the momentum equation. This is the essential point. The only condition on the factors that multiply these terms is that they have to yield an adequate scaling and order of convergence.

REFERENCES

- [1] D.N. ARNOLD, G. AWANOU, AND R. WINTHER, *Finite elements for symmetric tensors in three dimensions*, Math. Comp., 77 (2008), pp. 1229–1251.
- [2] D.N. ARNOLD AND R. WINTHER, *Mixed finite elements for elasticity*, Numer. Math., 92 (2002), pp. 401–419.
- [3] F.P.T. BAAIJENS, M.A. HULSEN, AND P.D. ANDERSON, *The use of mixed finite element methods for viscoelastic fluid flow analysis*, Chapter 14 in Encyclopedia of Computational Mechanics, E. Stein, R. de Borst, and T.J.R. Hughes, eds., John Wiley & Sons, New York, 2004, pp. 481–498.
- [4] R. BECKER AND M. BRAACK, *A finite element pressure gradient stabilization for the Stokes equations based on local projections*, Calcolo, 38 (2001), pp. 173–199.
- [5] M. BEHR, L.P. FRANCA, AND T.E. TEZDUYAR, *Stabilized finite element methods for the velocity-pressure-stress formulation of incompressible flows*, Comput. Methods Appl. Mech. Engrg., 104 (1993), pp. 31–48.
- [6] J. BONVIN, M. PICASSO, AND R. STENBERG, *GLS and EVSS methods for a three fields Stokes problems arising from viscoelastic flows*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 3893–3914.
- [7] M. BRAACK AND E. BURMAN, *Local projection stabilization for the Oseen problem and its interpretation as a variational multiscale method*, SIAM J. Numer. Anal., 43 (2006), pp. 2544–2566.
- [8] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
- [9] R. CODINA, *Stabilization of incompressibility and convection through orthogonal sub-scales in finite element methods*, Comput. Methods Appl. Mech. Engrg., 190 (2000), pp. 1579–1599.
- [10] R. CODINA, *Stabilized finite element approximation of transient incompressible flows using orthogonal subscales*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 4295–4321.
- [11] R. CODINA, *Analysis of a stabilized finite element approximation of the Oseen equations using orthogonal subscales*, Appl. Numer. Math., 58 (2008), pp. 264–283.
- [12] R. CODINA, *Finite element approximation of the hyperbolic wave equation in mixed form*, Comput. Methods Appl. Mech. Engrg., 197 (2008), pp. 1305–1322.
- [13] C.R. DOHRMANN AND P.B. BOCHEV, *A stabilized finite element method for the Stokes problem based on polynomial pressure projections*, Int. J. Num. Meth. Fluids, 46 (2004), pp. 183–201.
- [14] M. FORTIN, R. GUÉNETTE, AND R. PIERRE, *Numerical analysis of the modified EVSS method*, Comput. Methods Appl. Mech. Engrg., 143 (1997), pp. 79–95.
- [15] M. FORTIN AND R. PIERRE, *On the convergence of the mixed method of crochet and marchal for viscoelastic flows*, Comput. Methods Appl. Mech. Engrg., 73 (1989), pp. 341–350.
- [16] L. FRANCA AND R. STENBERG, *Error analysis of some Galerkin least-squares methods for the elasticity equations*, SIAM J. Numer. Anal., 28 (1991), pp. 1680–1697.
- [17] G.N. GATICA, *Analysis of a new augmented mixed finite element method for linear elasticity allowing RT_0 - P_1 - P_0 approximations*, ESAIM: Mathematical Modelling and Numerical Analysis, 40 (2006), pp. 1–28.
- [18] G.N. GATICA, A. MÁRQUEZ, AND S. MEDDAHI, *An augmented mixed finite element method of low cost for three-dimensional linear elasticity problems*, preprint 07-21, Departamento de Ingenieria Matematica, Universidad de Concepcion, 2007.
- [19] R. GUÉNETTE AND M. FORTIN, *A new mixed finite element method for computing viscoelastic flows*, J. Non-Newtonian Fluid Mechanics, 60 (1995), pp. 27–52.

- [20] T.J.R. HUGHES, *Multiscale phenomena: Green's function, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized formulations*, Comput. Methods Appl. Mech. Engrg., 127 (1995), pp. 387–401.
- [21] T.J.R. HUGHES, G.R. FEIJÓO, L. MAZZEI, AND J.B. QUINCY, *The variational multiscale method—a paradigm for computational mechanics*, Comput. Methods Appl. Mech. Engrg., 166 (1998), pp. 3–24.
- [22] P. KNOBLOCH AND L. TOBISKA, *On Korn's first inequality for quadrilateral nonconforming finite elements of first order approximation properties*, Int. J. Numer. Anal. Modeling, 2 (2005), pp. 439–458.
- [23] J.M. MARCHAL AND M.J. CROCHET, *A new mixed finite-element for calculating viscoelastic flow*, J. Non-Newtonian Fluid Mechanics, 26 (1987), pp. 77–114.
- [24] G. MATTHIES, P. SKRZYPACZ, AND L. TOBISKA, *A unified convergence analysis for local projection stabilisations applied to the Oseen problem*, Math. Modelling Numer. Anal., 41 (2007), pp. 713–742.
- [25] V. RUAS, *Une méthode mixte contrainte-déplacement-pressure pour la résolution de problèmes de viscoélasticité incompressible en déformations planes*, Comptes Rendus de l'Académie des Sciences. Série 2, 301 (1985), pp. 1171–1174.
- [26] V. RUAS, *Finite element methods for the three-field Stokes system*, RAIRO Modélisation Mathématique et Analyse Numérique, 30 (1996), pp. 489–525.
- [27] V. RUAS, *Galerkin-least-squares finite element methods for the three-field Stokes system in \mathbb{R}^3* , Comput. Methods Appl. Mech. Engrg., 142 (1997), pp. 235–256.
- [28] D. SANDRI, *Analysis of a three-fields approximation of the Stokes problem*, RAIRO Modélisation Mathématique et Analyse Numérique, 23 (1993), pp. 817–841.

ON GENERALIZED GAUSSIAN QUADRATURE RULES FOR SINGULAR AND NEARLY SINGULAR INTEGRALS*

DAAN HUYBRECHS[†] AND RONALD COOLS[†]

Abstract. We construct and analyze generalized Gaussian quadrature rules for integrands with endpoint singularities or near endpoint singularities. The rules have quadrature points inside the interval of integration, and the weights are all strictly positive. Such rules date back to the study of Chebyshev sets, but their use in applications has only recently been appreciated. We provide error estimates, and we show that the convergence rate is unaffected by the singularity of the integrand. We characterize the quadrature rules in terms of two families of functions that share many properties with orthogonal polynomials but that are orthogonal with respect to a discrete scalar product that, in most cases, is not known a priori.

Key words. Gaussian quadrature, Chebyshev sets, orthogonal polynomials

AMS subject classifications. 65D30, 65D32

DOI. 10.1137/080723417

1. Introduction. Gaussian quadrature has many advantages in the numerical integration of

$$\int_a^b w(x)f(x) dx \approx \sum_{j=1}^n w_j f(x_j),$$

with a positive weight function $w(x) > 0 \forall x \in [a, b]$. First, all quadrature points x_j lie inside the interval $[a, b]$ of integration, and the weights w_j are all positive [8, 28]. As a result, applying such quadrature rules is numerically stable. Second, it is well known that among all interpolatory quadrature rules, Gauss-type rules achieve the highest polynomial order. In particular, a Gaussian rule with n points is exact for polynomials up to degree $2n-1$. Convergence is, therefore, quite fast if the integrand is sufficiently smooth. It follows from the Weierstrass approximation theorem and from the positivity of the weights that convergence is guaranteed for all continuous functions f on $[a, b]$. Finally, Gaussian quadrature rules can be computed efficiently owing to their connection to orthogonal polynomials [9, 20], with a computational cost that scales as $O(n^2)$ for traditional algorithms [11] or $O(n)$ for more specialized ones [10]. A disadvantage of Gaussian rules is their inherent lack of adaptivity: different values of n lead to entirely different sets of quadrature points and weights. This is not the case, for example, for Clenshaw–Curtis rules, which otherwise share many of the advantages of Gaussian rules [30]. Several ways have been suggested to remedy the situation, most notably Gauss–Kronrod and Gauss–Kronrod–Patterson extensions [19, 23]. For modest values of n , the issue of adaptivity is less severe. In this paper, we consider the generalization of Gaussian quadrature rules in a different direction, focusing on achieving high accuracy for small n .

*Received by the editors May 7, 2008; accepted for publication (in revised form) September 22, 2008; published electronically January 16, 2009.

<http://www.siam.org/journals/sinum/47-1/72341.html>

[†]Department of Computer Science, Katholieke Universiteit Leuven, 3001 Leuven, Belgium (daan.huybrechs@cs.kuleuven.be, ronald.cools@cs.kuleuven.be). The first author is a Postdoctoral Fellow of the Research Foundation - Flanders (FWO).

The main subject of this paper concerns quadrature rules of Gaussian type for nonsmooth functions f . Though convergence of classical Gauss-type quadrature for such functions is possible, the convergence rate is low and the use of large n in quadrature is not recommended. Instead, research has focused on composite quadrature, singularity-removing transformations [29], graded meshes [26], and, in general, adaptive methods [3]. An efficient alternative was suggested, however, in [22]. Assume the integrand has the general form

$$f(x) = u(x) + v(x)\psi(x),$$

where both u and v are smooth functions and $\psi(x)$ has an integrable singularity of some kind, such as $\psi(x) = \log(x - a)$ or $\psi(x) = (x - a)^\alpha$, with $\alpha > -1$. It was proved in [22] that for many singular choices of ψ , a *generalized Gaussian quadrature* formula exists of the form $\sum_{j=1}^n w_j f(x_j)$ and with the following properties:

1. $x_j \in (a, b)$ and $w_j > 0$, $j = 1, \dots, n$;
2. $\sum_{j=1}^n w_j [x_j^k + x_j^l \psi(x_j)] = \int_a^b w(x) [x^k + x^l \psi(x)] dx$, $k, l = 0, \dots, n - 1$.

The first property indicates that, like classical Gauss-type rules, the quadrature points lie inside the interval $[a, b]$, and the weights are all positive. The second property states that the singularity is integrated exactly if u and v are polynomials up to degree $n - 1$. This rule is said to be Gaussian because $2n$ functions are integrated exactly using only n function evaluations. Note the important property that the rule evaluates only f , and not u or v . It is sufficient that u and v exist—they need not be known explicitly. Thus, the quadrature rule is a numerically stable approach for integrating nonsmooth functions, as long as the lack of smoothness is confined to a known function $\psi(x)$. For this reason, we call f a function with a *confined singularity*.

The existence of generalized Gaussian quadrature rules dates back to Markov in the study of Chebyshev sets [21]. A more recent treatise is given in [16]. It follows from this theory that a quadrature rule with n points exists that integrates $2n$ basis functions ϕ_k exactly,

$$(1.1) \quad \sum_{j=1}^n w_j \phi_k(x_j) = \int_a^b w(x) \phi_k(x) dx, \quad k = 1, \dots, 2n$$

if $\{\phi_k\}_{k=1}^{2n}$ is a Chebyshev set. Functions of the form $x^k + x^l \psi(x)$ are only a special case of this more general setting (albeit possibly a limiting special case if $\psi(x)$ is unbounded [22]).

One of the advantages listed above of classical Gauss-type properties has long been missing: an efficient construction algorithm. Generalized Gaussian quadrature rules have been described for special cases only in the literature, for example, in [12, 25, 7]. Two generally applicable numerical methods for computing these rules were first described in [22, 31]. These authors also introduced the name *generalized Gaussian quadrature*. The proposed methods essentially consist of a continuation approach combined with Newton's method to solve the set of $2n$ nonlinear equations (1.1) for the $2n$ unknowns w_j and x_j . Although not as efficient as orthogonal polynomial-based methods for classical rules, generalized Gaussian quadrature rules can be computed with reasonable efficiency for almost any basis set $\{\phi_k\}$. The results are particularly useful in integral equation methods, which require the evaluation of a large number of integrals with well-understood singular behavior [27, 18, 17, 2].

The purpose of this paper is to analyze generalized Gaussian quadrature rules in the setting of functions with a confined singularity. Though more limited than the

general theory, this setting is very useful in applications. We provide error estimates for generalized Gaussian quadrature rules in section 3. Next, in section 4 we characterize generalized Gaussian quadrature rules in terms of two sequences of functions $R_n(x)$ and $S_n(x)$, which obey certain orthogonality properties and which vanish at the quadrature points. This theory is more comparable to the theory of multivariate cubature formulae than to the theory of univariate Gaussian quadrature [6, 5, 4]. We discuss scaling invariance of the quadrature rules in section 5, and we briefly outline three approaches for the numerical construction of the rules in section 6. We end with some numerical examples in section 7.

2. Preliminaries. We consider in this paper the numerical approximation of the integral

$$(2.1) \quad I[f] := \int_a^b w(x)f(x) dx,$$

where $w(x) > 0$, $x \in (a, b)$, by a quadrature rule $Q[\cdot]$ with n points and weights of the form

$$(2.2) \quad Q[f] := \sum_{j=1}^n w_j f(x_j).$$

This approximation carries an error

$$\epsilon[f] := | I[f] - Q[f] |.$$

2.1. Functions with a confined singularity. We assume that the function f has the form

$$(2.3) \quad f(x) = u(x) + v(x)\psi(x),$$

where u and v lie in $C^k[a, b]$ for some sufficiently large k . We make no assumptions on the smoothness of the function ψ , except that it is possibly unbounded only at one of the endpoints a or b .¹ This most basic case is, arguably, also the most useful case in applications, as it covers integrands with a singularity or near singularity at one of the endpoints. We note, for example, that all rules constructed in [22] fit this pattern. Integrals with an internal singularity, which often appear in boundary element methods, may be treated by breaking the integrand at the singularity.

We introduce some more notation. We denote by P_m the set of polynomials up to degree m , and we define P_{-1} to be the empty set. The sets of functions T_m , $m = 0, 1, \dots$, are defined by

$$(2.4) \quad T_m := \begin{cases} \{1, \psi, x, x\psi, \dots, x^{l-1}\psi, x^l\}, & m = 2l \text{ is even,} \\ \{1, \psi, x, x\psi, \dots, x^{l-1}\psi, x^l, x^l\psi\}, & m = 2l + 1 \text{ is odd.} \end{cases}$$

They form the sequence $\{1\}, \{1, \psi\}, \{1, \psi, x\}, \{1, \psi, x, x\psi\}, \dots$. The corresponding function spaces are defined as

$$V_m := \text{span}\{T_m\}, \quad m = 0, 1, \dots$$

Note that the functions in V_m are not, in general, square integrable because $\psi(x)^2$ may not be integrable on $[a, b]$.

¹This condition appears in the proof of Theorem 3.4. It may conceivably be lifted to allow singularities at both endpoints at the cost of having less nice error estimates.

2.2. Existence of the quadrature rule. We assume in this paper that the function ψ is such that a generalized Gaussian quadrature rule exists for all n . That is, we assume that

$$(2.5) \quad Q[\phi] = I[\phi] \quad \forall \phi \in T_{2n-1}.$$

Expression (2.5) leads to a set of $2n$ nonlinear equations in w_j and x_j —it corresponds exactly to expression (1.1) in our new notation.

Existence and uniqueness of the quadrature rule are guaranteed if T_{2n-1} is a Chebyshev set on $[a, b]$. This is a side result of a more general theory on the geometric properties of the *moment spaces* that are induced by a Chebyshev set (see [21, 16]). More recently, it was proved in [22] that existence and uniqueness is guaranteed if T_{2n-1} is a Chebyshev set on any closed subinterval of (a, b) . The latter generalization allows unbounded singularities at the endpoints. It should be mentioned that these results yield sufficient, but not necessary, conditions.

In both cases, we can define an interpolation operator $\mathcal{P}_{\mathbf{x}}$ for a set of points $\mathbf{x} = \{x_j\}_{j=1}^n$ such that $\mathcal{P}_{\mathbf{x}}[f] \in V_{n-1}$ and

$$(2.6) \quad (\mathcal{P}_{\mathbf{x}}[f])(x_j) = f(x_j), \quad j = 1, \dots, n.$$

Assuming that T_{n-1} is a Chebyshev set on all closed subsets of (a, b) , this operator is the identity on V_{n-1} for all sets \mathbf{x} , with $x_j \in (a, b)$, $j = 1, \dots, n$.

Note that the choice of $\psi(x)$ is not as free as the choice of the weight function $w(x)$. Any weight function that satisfies $w(x) > 0$ on (a, b) will do. On the other hand, it is known that the function $\psi(x)$ should be either monotonically increasing or decreasing, in order to obtain a Chebyshev set. The main choices we have in mind are $\psi(x) = \log(x + \delta)$ and $\psi(x) = (x + \delta)^\alpha$, with $\alpha > -1$ and where δ determines the location of the singularity.

3. Error estimates. The central result in this section is the error estimate, proved in Theorem 3.4:

$$\epsilon[f] \leq \frac{1}{(n-1)!} (b-a)^n \left(W \|u^{(n)}\|_\infty + (2WC_\psi + W_\psi) \|v^{(n)}\|_\infty \right),$$

with the constants defined as in the theorem and depending only on the weight function $w(x)$ and the singularity function $\psi(x)$. The estimate shows that the convergence of the quadrature rule is unaffected by the unboundedness or the lack of smoothness of the singularity function ψ , even though ψ is evaluated implicitly in f and the smooth functions u and v are unknown.

3.1. The Peano kernel. Error estimates for interpolatory quadrature rules are most often given in terms of a derivative of f , with the order of the derivative depending on the polynomial degree of the rule. These estimates can be obtained from error estimates for polynomial interpolation or from the Peano kernel theorem. General error estimates for interpolation by Chebyshev sets are not available. The specific form of the function spaces V_{2n-1} , however, enables the use of the Peano kernel theorem [24]. For a functional $L[f]$ and an integer $k \geq 0$, the Peano kernel is defined by

$$(3.1) \quad K(\theta) = \frac{1}{k!} L_x \left[(x - \theta)_+^k \right],$$

with

$$(3.2) \quad (x - \theta)_+^k = \begin{cases} (x - \theta)^k, & x \geq \theta, \\ 0, & x < \theta. \end{cases}$$

The notation $L_x[\cdot]$ indicates that the functional L operates on a function of x . In the following theorem, $\mathcal{V}[a, b]$ is the space of real-valued functions on $[a, b]$ that is of bounded variation.

THEOREM 3.1 (Peano kernel [24]). *Let k be any nonnegative integer, and let L be a bounded linear functional from $\mathcal{V}[a, b]$ to \mathbb{R} , such that $L[f]$ is zero when f is in P_k , and such that the function $K(\theta)$, $a \leq \theta \leq b$, defined by (3.1), is of bounded variation. Then, if f is in $C^{k+1}[a, b]$, the functional $L[f]$ has the value*

$$(3.3) \quad L[f] = \int_a^b K(\theta) f^{(k+1)}(\theta) d\theta.$$

The proof is based on an expression for the remainder in a Taylor series of f . An estimate follows of the form

$$(3.4) \quad |L[f]| \leq \|K\|_1 \|f^{(k+1)}\|_\infty.$$

In the following section, from Theorem 3.1 we will obtain bounds for the error $L[f] := I[f] - Q[f]$ in terms of a derivative of f .

3.2. Functions with a confined singularity. Let us first apply the Peano kernel theorem to smooth functions $f(x) = u(x)$. The operator

$$(3.5) \quad L_1[u] := I[u] - Q[u]$$

defines the error in the numerical approximation of the integral $I[u]$ by a generalized Gaussian quadrature rule with n points.

LEMMA 3.2. *For $u \in C^n[a, b]$, we have*

$$|I[u] - Q[u]| \leq \frac{1}{(n-1)!} W (b-a)^n \|u^{(n)}\|_\infty,$$

where $W := \int_a^b w(x) dx$.

Proof. The quadrature rule is exact for polynomials up to degree $n-1$. Thus, the Peano kernel (3.1) is given by

$$(3.6) \quad K(\theta) = \frac{1}{(n-1)!} (I[(x-\theta)_+^{n-1}] - Q[(x-\theta)_+^{n-1}]).$$

We have, for $\theta \in [a, b]$,

$$\begin{aligned} I[(x-\theta)_+^{n-1}] &= \int_a^b w(x)(x-\theta)_+^{n-1} dx = \int_\theta^b w(x)(x-\theta)^{n-1} dx \\ &\leq W(b-\theta)^{n-1} \leq W(b-a)^{n-1}. \end{aligned}$$

We also have

$$\begin{aligned} Q[(x-\theta)_+^{n-1}] &= \sum_{j=1}^n w_j (x_j - \theta)_+^{n-1} \\ &\leq \sum_{j=1}^n w_j (b-\theta)^{n-1} = W(b-\theta)^{n-1} \leq W(b-a)^{n-1}. \end{aligned}$$

Note that in the latter derivation, we have used the fact that the weights are all positive and that they sum up to W . Given that both $I[(x-\theta)_+^{n-1}]$ and $Q[(x-\theta)_+^{n-1}]$

in (3.6) are positive, we have

$$|K(\theta)| \leq \frac{1}{(n-1)!} W(b-a)^{n-1}.$$

It follows that

$$\|K\|_1 = \int_a^b |K(\theta)| \, d\theta \leq \frac{1}{(n-1)!} \int_a^b W(b-a)^{n-1} \, d\theta = \frac{1}{(n-1)!} W(b-a)^n.$$

The result now follows from the general error estimate (3.4). \square

Next, we establish an error estimate for functions of the form $f(x) = \psi(x)v(x)$, where $v(x)$ is a smooth function. Define the linear functional

$$L_2[v] := I[\psi v] - Q[\psi v].$$

This functional is exact for polynomials up to degree $n - 1$, and hence, we can again invoke the Peano kernel theorem.

LEMMA 3.3. *If $v \in C^n[a, b]$ and if $\psi(x) > 0, \forall x \in (a, b)$, we have*

$$|L_2[v]| \leq \frac{1}{(n-1)!} W_\psi (b-a)^n \|v^{(n)}\|_\infty,$$

where $W_\psi := \int_a^b w(x)\psi(x) \, dx$.

Proof. The result follows from Lemma 3.2 by defining a weight function of the form $w(x)\psi(x)$. Note that, since $\psi(x)$ is, in this lemma, assumed to be positive, we indeed have

$$\sum_{j=1}^n w_j \psi(x_j) = \int_a^b w(x)\psi(x) \, dx = W_\psi,$$

with all terms in the summation positive, as required in the proof of Lemma 3.2. \square

We can now state the central result of this section.

THEOREM 3.4. *Assume $f(x) = u(x) + v(x)\psi(x)$, with $u, v \in C^n[a, b]$. Then we have*

$$(3.7) \quad \epsilon[f] \leq \frac{1}{(n-1)!} (b-a)^n \left(W \|u^{(n)}\|_\infty + (2WC_\psi + W_\psi) \|v^{(n)}\|_\infty \right),$$

with constants W and W_ψ as defined in Lemma 3.2 and Lemma 3.3 and with

$$(3.8) \quad C_\psi := \min \left(\left| \sup_{x \in [a,b]} \psi(x) \right|, \left| \inf_{x \in [a,b]} \psi(x) \right| \right) \geq 0$$

a positive and bounded constant.

Proof. We can not immediately invoke Lemma 3.3 because the function $\psi(x)$ is not necessarily positive on the open interval (a, b) . We will construct a function $\tilde{\psi}(x) = A\psi(x) + B$ that is positive on (a, b) . Define the values $M^+ = \sup_{x \in [a,b]} \psi(x)$ and $M^- = \inf_{x \in [a,b]} \psi(x)$. Next, define

$$\tilde{\psi}(x) := \begin{cases} \psi(x) - M^-, & \text{if } |M^-| \leq |M^+|, \\ -\psi(x) + M^+, & \text{otherwise.} \end{cases}$$

By our assumption that $\psi(x)$ can be unbounded in at most one endpoint, at least one of M^+ or M^- is finite. We have thus written $\tilde{\psi}(x) = A\psi(x) + B$, with $A = \pm 1$ and $|B| \leq C_\psi$, where C_ψ is finite. By construction, we have $\tilde{\psi}(x) \geq 0$ for $x \in (a, b)$.

We rewrite the function $f(x)$ in terms of $\tilde{\psi}(x)$, using the fact that $1/A = A$:

$$f(x) = u(x) - \frac{B}{A}v(x) + \frac{1}{A}v(x)(A\psi(x) + B) = u(x) - ABv(x) + Av(x)\tilde{\psi}(x).$$

Note that if u and v are polynomials of degree k , then $u(x) - ABv(x)$ and $Av(x)$ are also polynomials of degree k . This means that the generalized Gaussian quadrature rules constructed using either $\psi(x)$ or $\tilde{\psi}(x)$ are the same.

We now apply Lemma 3.2, noting that $C_\psi > |AB| = |B|$,

$$|I[u - ABv] - Q[u - ABv]| \leq \frac{1}{(n-1)!} W(b-a)^n \left(\|u^{(n)}\|_\infty + C_\psi \|v^{(n)}\|_\infty \right).$$

Lemma 3.3 leads to

$$\left| I[\tilde{\psi}Av] - Q[\tilde{\psi}Av] \right| \leq \frac{1}{(n-1)!} W_{\tilde{\psi}}(b-a)^n \|v^{(n)}\|_\infty.$$

We also have

$$W_{\tilde{\psi}} = \int_a^b w(x)(A\psi(x) + B) dx \leq W_\psi + C_\psi W.$$

The combination of the above inequalities proves the result. \square

The importance of Theorem 3.4 is that it shows that the convergence of $Q[f]$ to $I[f]$ depends only on the smoothness of $u(x)$ and $v(x)$, irrespective of the lack of smoothness in $\psi(x)$. An advantage of the current method of proof is that the constants in the error estimate (3.7) are entirely explicit in their dependence on the functions $w(x)$ and $\psi(x)$.

3.3. Functions with multiple singularities. We digress briefly from the case of functions with a single confined singularity to note that the error estimates readily extend to the case of functions with multiple singularities. Consider m functions $\psi_m(x)$ and a function $f(x)$ with multiple singularities of the form

$$(3.9) \quad f(x) = \sum_{m=1}^M u_m(x)\psi_m(x),$$

where $u_m(x)$ are smooth functions, $m = 1, \dots, M$. The function f may, for example, have singularities in both endpoints of the integration interval $[a, b]$. One is led to consider a quadrature rule $Q[f] = \sum_{j=1}^n w_j f(x_j)$ that satisfies

$$(3.10) \quad Q[x^k \psi_m] = I[x^k \psi_m], \quad k = 0, \dots, n_m - 1, \quad m = 1, \dots, M.$$

In the following, we forego the existence question in favor of deriving error estimates. We assume for simplicity that all $\psi_m(x) \geq 0$ and, moreover, that the quadrature rule has positive weights.

LEMMA 3.5. *Assume that all $\psi_m(x) \geq 0 \forall x \in [a, b]$, and define $L_m[u] := Q[u\psi_m] - I[u\psi_m]$. Then for $u \in C^{n_m}[a, b]$, we have*

$$|L_m[u]| \leq \frac{1}{(n_m - 1)!} W_{\psi_m}(b-a)^{n_m} \|u^{(n_m)}\|_\infty,$$

where $W_{\psi_m} = \int_a^b w(x)\psi_m(x) dx$.

The proof of this lemma is exactly like that of Lemma 3.3.

THEOREM 3.6. *Let $Q[f] = \sum_{j=1}^n w_j f(x_j)$ satisfy conditions (3.10) for certain $n_m > 0$, $m = 1, \dots, M$, and let $w_j > 0$, $j = 1, \dots, n$ and $\psi_m(x) \geq 0$, $m = 1, \dots, M$. Then, for functions f of the form (3.9), we have*

$$(3.11) \quad |I[f] - Q[f]| \leq \sum_{m=1}^M \frac{1}{(n_m - 1)!} W_{\psi_m} (b - a)^{n_m} \|u^{(n_m)}\|_{\infty},$$

with W_{ψ_m} defined as in Lemma 3.5.

Proof. We can write

$$L[f] - Q[f] = \sum_{m=1}^M L_m[u_m],$$

where the linear operators L_m are as in Lemma 3.5. The result follows immediately from Lemma 3.5 and from

$$\sum_{m=1}^M L_m[u_m] \leq \sum_{m=1}^M |L_m[u_m]|. \quad \square$$

Note that the assumption $\psi_m(x) \geq 0$ simplifies the error estimate (3.11) compared to the previous estimate (3.7). This comes at a cost of having slightly less general results.

4. A theory of generalized Gaussian quadrature.

4.1. Orthogonal polynomials. It is well known that the points of a classical Gaussian rule are the roots of a polynomial $p_n(x)$ of degree n that is uniquely determined, up to a constant factor, by the orthogonality conditions

$$(4.1) \quad \int_a^b w(x) x^k p_n(x) dx = 0, \quad k = 0, \dots, n - 1.$$

Let us denote the classical Gaussian quadrature rule relative to the weight function $w(x)$ by $Q^G[\cdot]$. The concept of orthogonality derives from an inner product, which is not available in the generalized setting. This, however, is not an essential argument in the characterization of Q^G by p_n . An alternative and more general point of view is that the quadrature rule Q^G is characterized by a set of functions that vanish at the quadrature points. This set, in turn, is characterized by p_n . The meaning of these statements is clarified in the following lemma.

LEMMA 4.1. *Let $I[f]$ be a linear, continuous functional defined on a vector space F of functions on $[a, b]$ and consider a quadrature rule $Q[f] = \sum_{j=1}^n w_j f(x_j)$. For a subspace $F_1 \subset F$, define*

$$F_0 = \{f_0 \in F_1 : f_0(x_j) = 0, j = 1, \dots, n\}.$$

A necessary and sufficient condition for the existence of a quadrature rule that is exact for all $f \in F_1$ is

$$(4.2) \quad f_0 \in F_0 \Rightarrow I[f_0] = 0.$$

Proof. This is only a special case of Theorem 3.1 in [6] (with short proof). \square

An interpolatory quadrature rule with n points x_j is based on interpolating n given function values by a polynomial of degree $n - 1$. It is obvious that such rules

can be exact for polynomials of degree up to $n - 1$, as the function to integrate is recovered exactly by the interpolation. Lemma 4.1 gives conditions for exactness in a larger space F_1 : the functional $I[f]$ has to vanish for all functions in F_1 that vanish at the quadrature points.

Consider, for example, the space $F_1 = P_{2n-1}$ of polynomials up to degree $2n - 1$. Each polynomial that vanishes at all quadrature points can be factorized into a polynomial multiple of p_n . The space F_0 can, therefore, be characterized in terms of p_n by

$$(4.3) \quad F_0 \equiv \text{span} \{p_n(x)x^k\}_{k=0}^{n-1}.$$

Condition (4.2) now corresponds exactly to the orthogonality conditions (4.1).

4.2. Characterizing generalized Gaussian quadrature rules. We return to the setting of a function f with a confined singularity of the form (2.3). Define the space of all functions in V_{2n-1} vanishing at a set $\mathbf{x} = \{x_j\}_{j=1}^n$ of n distinct points in (a, b) as

$$(4.4) \quad F_0(\mathbf{x}) := \{f \in V_{2n-1} | f(x_j) = 0, j = 1, \dots, n\}.$$

The space $F_0(\mathbf{x})$ can not be characterized in terms of a single polynomial that vanishes at the points x_j as in (4.3). It can, however, be characterized in terms of two different functions $R_n(x)$ and $S_n(x)$ that vanish at the set of points. The space $F_0(\mathbf{x})$ then consists of a linear combination of polynomial multiples of $R_n(x)$ and $S_n(x)$. We show this first for the case where $n = 2l$ is even.

LEMMA 4.2. *If $n = 2l$ is even, then each $f_0 \in F_0(\mathbf{x})$ can be written as*

$$(4.5) \quad f_0(x) = p(x)R_n(x) + q(x)S_n(x),$$

with $p, q \in P_{l-1}$ and where

$$(4.6) \quad R_n(x) = x^l - \mathcal{P}_{\mathbf{x}} [x^l], \quad S_n(x) = x^l \psi(x) - \mathcal{P}_{\mathbf{x}} [x^l \psi].$$

Conversely, each function of the form (4.5) with $p, q \in P_{l-1}$ lies in $F_0(\mathbf{x})$.

Proof. Recall that $\mathcal{P}_{\mathbf{x}}$ is an interpolation operator, which is defined by (2.6). It follows from the construction that $R_n(x_j) = S_n(x_j) = 0$. It follows, in turn, that any function of the form $p(x)R_n(x) + q(x)S_n(x) \in F_0(\mathbf{x})$ for $p, q \in P_{l-1}$. It remains to show that all functions $f_0 \in F_0(\mathbf{x})$ can be written this way.

We prove the decomposition, by construction, with a procedure similar to polynomial long division. Any function $f_0 \in F_0(\mathbf{x}) \subset V_{2n-1}$ can be written in the basis T_{2n-1} as

$$f_0(x) = \sum_{k=0}^{n-1} a_k x^k + \sum_{k=0}^{n-1} b_k x^k \psi(x),$$

with suitable coefficients a_k and b_k . We define the function $f_1(x)$ by

$$f_1(x) = f_0(x) - a_{n-1} x^{l-1} R_n(x) - b_{n-1} x^{l-1} \psi(x) S_n(x).$$

Note that we now have $f_1 \in V_{2n-3}$, so we can write

$$f_1(x) = \sum_{k=0}^{n-2} c_k x^k + \sum_{k=0}^{n-2} d_k x^k \psi(x),$$

with suitable coefficients c_k and d_k . We define $f_2(x)$ by

$$f_2(x) = f_1(x) - c_{n-2}x^{l-2}R_n(x) - b_{n-2}x^{l-2}\psi(x)S_n(x)$$

and so on. The procedure can be performed l times until we arrive at

$$f_0(x) = p(x)R_n(x) + q(x)S_n(x) + f_l(x),$$

where $p(x)$ and $q(x)$ are polynomials of degree $l - 1$ and $f_l \in V_{2n-1-2l} = V_{n-1}$. However, since $f_0(x_j) = 0$, we must have $f_l(x_j) = 0$. Therefore, $\mathcal{P}_{\mathbf{x}}[f_l] \equiv 0$, which is only possible if $f_l(x) \equiv 0$. \square

The case where n is odd is analogous, only with small differences in the degree of polynomials involved.

LEMMA 4.3. *If $n = 2l - 1$ is odd, then each $f_0 \in F_0(\mathbf{x})$ can be written as*

$$(4.7) \quad f_0(x) = p(x)R_n(x) + q(x)S_n(x),$$

with $p(x) \in P_{l-2}$ and $q(x) \in P_{l-1}$, and where

$$(4.8) \quad R_n(x) = x^l - \mathcal{P}_{\mathbf{x}}[x^l], \quad S_n(x) = x^{l-1}\psi(x) - \mathcal{P}_{\mathbf{x}}[x^{l-1}\psi].$$

Conversely, each function of the form (4.7) with $p(x) \in P_{l-2}$ and $q(x) \in P_{l-1}$ lies in $F_0(\mathbf{x})$.

From Lemmas 4.2 and 4.3, a set of quadrature points x_j can be characterized as the common roots of the functions $R_n(x)$ and $S_n(x)$. Given a set of points $\mathbf{x} = \{x_j\}_{j=1}^n$, the weights of an interpolatory quadrature rule are found easily. Denote the n Lagrange functions by $\mathcal{L}_j(x) \in P_{n-1}$, $j = 1, \dots, n$, i.e.,

$$\mathcal{L}_j(x_{j'}) = \delta_{j-j'}, \quad j, j' = 1, \dots, n.$$

Then we have

$$(4.9) \quad w_j = I[\mathcal{L}_j].$$

For any set of points \mathbf{x} , expression (4.9) yields a quadrature rule that is exact on V_{n-1} by construction. For the set of generalized Gaussian quadrature points, the rule is exact on V_{2n-1} . Assembling our results, we can prove the following theorem.

THEOREM 4.4. *Let $\mathbf{x} = \{x_j\}_{j=1}^n$ be a set of n distinct points in (a, b) , and define a quadrature rule $Q[f] = \sum_{j=1}^n w_j f(x_j)$ with weights given by (4.9). We have $Q[f] = I[f] \forall f \in V_{2n-1}$ if and only if*

$$(4.10) \quad I[x^k R_n] = 0, \quad k = 0, \dots, l - 1,$$

$$(4.11) \quad I[x^k S_n] = 0, \quad k = 0, \dots, l - 1,$$

if $n = 2l$ is even, and

$$(4.12) \quad I[x^k R_n] = 0, \quad k = 0, \dots, l - 2,$$

$$(4.13) \quad I[x^k S_n] = 0, \quad k = 0, \dots, l - 1,$$

if $n = 2l - 1$ is odd. The functions $R_n(x)$ and $S_n(x)$ are defined by (4.6) for even n and by (4.8) for odd n .

Proof. We consider only the case where $n = 2l$ is even. The case of odd n is proven in an analogous manner.

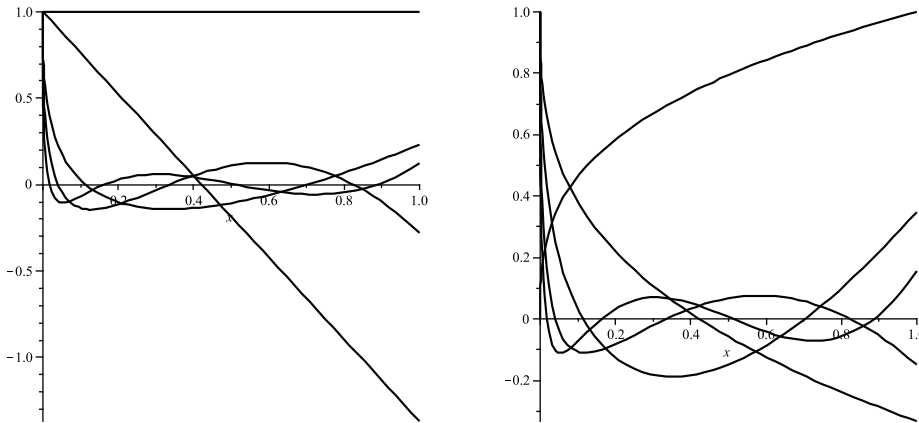


FIG. 4.1. Plots of the functions $R_n(x)$ (left panel) and $S_n(x)$ (right panel) for $n = 0, \dots, 4$, corresponding to the singularity function $\psi(x) = x^{1/3}$ on $[0, 1]$. The functions have been normalized such that $R_n(0) = S_n(0) = 1$, $n = 1, \dots, 4$. Moreover, we have set $R_0(x) = 1$ and $S_0(x) = x^{1/3}$.

Assume for the first direction of the “if and only if” statement that \mathbf{x} is such that $Q[f]$ is exact on V_{2n-1} . It follows from the necessary condition in Lemma 4.1 that we should have

$$I[f_0] = 0 \quad \forall f_0 \in F_0.$$

From Lemma 4.2, we may write $f_0(x) = p(x)R_n(x) + q(x)S_n(x)$. The functions $R_n(x)$ and $S_n(x)$ are well defined. We should have $I[f_0] = 0 \forall p \in P_{l-1}$ and $\forall q \in P_{l-1}$. This corresponds exactly to conditions (4.10)–(4.11).

Conversely, assume that for a given set \mathbf{x} , the conditions (4.10)–(4.11) hold. In that case, we have from Lemma 4.2 that $I[f_0] = 0 \forall f_0 \in F_0$. This, according to Lemma 4.1, is a sufficient condition for the result. \square

Example 4.5. The functions $R_n(x)$ and $S_n(x)$ are illustrated in Figure 4.1 for the case $\psi(x) = x^{1/3}$ on the integration interval $[0, 1]$.

The functions $R_n(x)$ and $S_n(x)$ retain some orthogonality properties: They are orthogonal to polynomials up to a degree of approximately $n/2$. They are not orthogonal to each other: Recall that the product of two functions in V_m for some m may not be integrable, depending on the type of singularity. Yet, the functions $R_n(x)$ and $S_n(x)$ are not independent, as one completely characterizes the other and vice-versa through their common roots at the quadrature points. Denote by $\mathcal{R}(f)$ the set of roots of $f(x)$ on (a, b) for any f with n distinct roots. Then we have the nonlinear relations

$$(4.14) \quad R_n(x) = x^l - \mathcal{P}_{\mathcal{R}(S_n)}[x^l], \quad \text{with } l = \left\lceil \frac{n}{2} \right\rceil,$$

and

$$(4.15) \quad S_n(x) = x^{l'} \psi(x) - \mathcal{P}_{\mathcal{R}(R_n)}[x^{l'} \psi], \quad \text{with } l' = \left\lfloor \frac{n}{2} \right\rfloor.$$

4.3. Discrete orthogonality. In this section we attempt to further clarify the difference between classical Gaussian quadrature rules, connected to orthogonal polynomials, and generalized Gaussian quadrature rules, connected to the functions $R_n(x)$

and $S_n(x)$. We will show that the functions $R_n(x)$ and $S_n(x)$ are orthogonal to all functions in V_{n-1} with respect to a discrete scalar product, defined in terms of the points x_j and weights w_j of the generalized Gaussian quadrature rule as

$$(4.16) \quad u_n(f, g) := \sum_{j=1}^n w_j f(x_j) g(x_j).$$

LEMMA 4.6. *The bilinear form (4.16) is a real scalar product on V_{n-1} .*

Proof. The form is linear and symmetric. It is positive, $u_n(f, f) \geq 0$, because the weights are all (strictly) positive. Finally, it is nondegenerate because $u_n(f, f) = 0$ implies $f(x_j) = 0$, $j = 1, \dots, n$, and since no nonzero function in V_{n-1} vanishes at n distinct points, this, in turn, implies $f(x) \equiv 0$. \square

Consider next the following sequence of orthogonal functions $r_{n,k}(x)$. Denote by $\{\phi_j\}_{j=0}^{n-1}$ a basis for V_{n-1} , for example, $1, \psi, x, x\psi, \dots$. Set

$$(4.17) \quad r_{n,0}(x) := \phi_0(x),$$

and define iteratively

$$(4.18) \quad r_{n,k}(x) = \phi_k(x) - \sum_{j=0}^{k-1} \frac{u_n(\phi_k, r_{n,j})}{u_n(r_{n,j}, r_{n,j})} r_{n,j}(x), \quad k = 1, \dots, n-1.$$

This Gram–Schmidt procedure leads to well-defined functions $r_{n,k}(x)$ that are orthogonal with respect to u_n .

Next, define the functions

$$(4.19) \quad r_{n,n}(x) = x^l - \sum_{j=0}^{n-1} \frac{u_n(x^l, r_{n,j})}{u_n(r_{n,j}, r_{n,j})} r_{n,j}(x),$$

with $l = \lceil \frac{n}{2} \rceil$ and

$$(4.20) \quad s_{n,n}(x) = x^{l'} \psi(x) - \sum_{j=0}^{n-1} \frac{u_n(x^{l'} \psi, r_{n,j})}{u_n(r_{n,j}, r_{n,j})} r_{n,j}(x),$$

with $l' = \lfloor \frac{n}{2} \rfloor$.

THEOREM 4.7. *We have $R_n(x) = r_{n,n}(x)$ and $S_n(x) = s_{n,n}(x)$.*

Proof. We have, by construction, that

$$(4.21) \quad u_n(r_{n,n}, g) = 0 \quad \forall g \in V_{n-1}.$$

Construct the functions $g_j \in V_{n-1}$ such that $g_j(x_i) = \delta_{i,j}$, $i, j = 1, \dots, n$. This is always possible because T_{n-1} is a Chebyshev set. It follows from the definition (4.16) and from the property (4.21) that

$$u_n(r_{n,n}, g_j) = w_j r_{n,n}(x_j) = 0.$$

This implies that $r_{n,n}(x_j) = 0$, $j = 1, \dots, n$. We also have, by construction, that $r_{n,n}(x) \in \text{span}(T_{n-1} \cup \{x^l\})$. Moreover, $r_{n,n}(x)$ is nonzero because the basis function x^l has coefficient 1.

The function $R_n(x) = x^l - \mathcal{P}_x(x^l)$ is nonzero, has coefficient 1 with x^l , and vanishes at the quadrature points. This function is unique because \mathcal{P}_x is invertible on V_{n-1} . It follows that $r_{n,n}(x) = R_n(x)$.

The proof for the statement $S_n(x) = s_{n,n}(x)$ is analogous. \square

Theorem 4.7 implies that both $R_n(x)$ and $S_n(x)$ can be found by a Gram–Schmidt procedure applied to a basis of V_{n-1} using the scalar product u_n . As the scalar product itself is defined in terms of the quadrature rule, however, this only implicitly determines R_n and S_n . In contrast, consider a similar scalar product for classical Gaussian quadrature rules. This scalar product u_n^G can be defined as in (4.16) but using the points and weights of the classical Gaussian quadrature rule. The bilinear form u_n^G coincides with the L_2 inner product for polynomial f and g up to certain degree:

$$u_n^G(f, g) = \int_a^b w(x)f(x)g(x) dx, \quad \forall f \in P_n, \forall g \in P_{n-1}.$$

All computations in the Gram–Schmidt procedure can be performed explicitly, leading to $p_n(x)$. Alternatively, of course, one can employ the three-term recurrence formula of orthogonal polynomials. Both schemes are not available in the setting of generalized Gaussian quadrature.

Example 4.8. An exception to the general case is given by the special case $\psi(x) = \sqrt{x}$. In that case, it is easy to verify that the product of two functions in V_{n-1} lies in V_{2n-1} . The scalar product u_n then coincides with the L_2 inner product because the quadrature rule is exact on V_{2n-1} . The Gram–Schmidt procedure can be performed, and $R_n(x)$ can be determined explicitly for all n .

In this case the generalized Gaussian quadrature rule is closely related to a classical Gaussian quadrature rule with the weight function $w(y) = 2y$. Indeed, consider the substitution $x = y^2$,

$$\int_0^1 f(x) dx = \int_0^1 2yf(y^2) dy.$$

For any $f(x) = u(x) + v(x)\sqrt{x}$ with polynomial u and v , the function $f(y^2)$ is simply a polynomial in y . The generalized Gaussian quadrature rule with points x_j can also be obtained from the classical Gaussian rule with weight function $2y$ and quadrature points y_j by $x_j = y_j^2$.

5. Scaling of the quadrature rule. One additional useful property of generalized Gaussian quadrature rules is that they are invariant to a scaling of the integration interval for a wide variety of functions ψ with a singularity at one of the endpoints. Consider, without loss of generality, a singularity function $\psi(x)$ with a singularity at $x = 0$, and define the integral

$$(5.1) \quad I_b[f] := \int_0^b f(x) dx.$$

We show that for many cases of practical interest, the generalized Gaussian quadrature rule for I_b is invariant to a scaling of b , up to a simple scaling of the weights and quadrature points expressed in (5.3) below.

5.1. Scaling invariant rules. Assume that $f(x) = u(x) + v(x)\psi(x)$ on $[0, b]$. We are interested in the points $x_{j,b}$ and weights $w_{j,b}$ of a generalized Gaussian quadrature rule on $[0, b]$. Rescaling the interval to $[0, 1]$ by letting $x = bt$, we note that $(bt)^\alpha = b^\alpha t^\alpha$ and that $\log(bt) = \log b + \log t$. This motivates the following lemma.

LEMMA 5.1. *If*

$$(5.2) \quad \psi(bt) = A_b \psi(t) + B_b,$$

then

$$(5.3) \quad w_{j,b} = b w_{j,1} \quad \text{and} \quad x_{j,b} = b x_{j,1}.$$

Proof. We rescale the interval to $[0, 1]$. We have $I_b[f] = b I_1[\tilde{f}]$, with

$$\tilde{f}(t) = f(bt) = u(bt) + v(bt)\psi(bt).$$

Using (5.3) we write

$$\tilde{f}(t) = u(bt) + A_b v(bt) + B_b v(bt)\psi(t) =: \tilde{u}(t) + \tilde{v}(t)\psi(t).$$

Note that if u and v are polynomials of degree $n - 1$, then $\tilde{u}(t) = u(bt) + A_b v(bt)$ and $\tilde{v}(t) = B_b v(bt)$ are also polynomials of the same degree. Therefore, if the points and weights $x_{j,1}$ and $w_{j,1}$ define a generalized Gaussian quadrature rule on $[0, 1]$, then the scaled points and weights given by (5.3) define a generalized Gaussian quadrature rule on $[0, b]$ for the integral (5.1). \square

Note that, contrary to alternative approaches where the singularity function $\psi(x)$ has been included into a weight function (see, for example, [15]), in generalized Gaussian quadrature it is not necessary to know the constants A_b and B_b . One evaluates only the function $f(x)$ on $[0, b]$ in the points $b x_{j,1}$.

5.2. Nearly scaling invariant quadrature rules. We say that singularity functions satisfying (5.2) give rise to *scaling invariant* quadrature rules because exactness is retained for $f(x) = u(x) + v(x)\psi(x)$, $x \in [0, b]$, when $u(x)$ and $v(x)$ are polynomials of sufficiently small degree. Less restrictive conditions on ψ than those of Lemma 5.1 may still yield useful results, however, as the following lemma shows.

LEMMA 5.2. *Assume that*

$$(5.4) \quad \psi(bt) = p(t, b)\psi(t) + q(t, b).$$

Then, for $f(x) = u(x) + v(x)\psi(x)$, $x \in [0, b]$, we have

$$\left| I_b[f] - \sum_{j=1}^n w_{j,b} f(x_{j,b}) \right| \leq \frac{1}{(n-1)!} \left(W \|\tilde{u}^{(n)}\|_\infty + (2WC_\psi + W_\psi) \|\tilde{v}^{(n)}\|_\infty \right),$$

with $\tilde{u}(t) = b[u(bt) + q(t, b)]$, $\tilde{v}(t) = bv(bt)p(t, b)$, $t \in [0, 1]$, and with $w_{j,b}$ and $x_{j,b}$ given by (5.3).

Proof. Letting $x = bt$, we obtain

$$\int_0^b f(x) dx = b \int_0^1 [u(bt) + q(t, b) + v(bt)p(t, b)\psi(t)] dt.$$

We then apply Theorem 3.4 using the definitions of \tilde{u} and \tilde{v} . \square

This lemma shows that if $p(t, b)$ and $q(t, b)$ are smooth functions, in the sense that they are sufficiently differentiable and have small derivatives, then the scaled quadrature rule carries small error. The rule is, in general, no longer exact, however, for polynomials u and v . As before, explicit knowledge of the functions $p(t, b)$ and $q(t, b)$ is not required; one simply evaluates $f(x)$.

5.3. Nearly singular integrals. Generalized Gaussian quadrature rules lose some of their appeal in the setting of nearly singular integrals. Consider, for example, the integral

$$\int_0^1 u(x) + v(x)\psi(x + \delta) dx,$$

with $\delta \geq 0$. Let us first note that convergence is not the issue. A generalized Gaussian quadrature rule exists for each value of δ , with the quadrature points $x_j(\delta)$ and weights $w_j(\delta)$ depending on δ . Following Theorem 3.4, the quadrature error is small uniformly in δ if the quantities

$$W_\psi = \int_0^1 w(x)\psi(x + \delta) dx$$

and

$$\min \left(\left| \sup_{x \in [0,1]} \psi(x + \delta) \right|, \left| \inf_{x \in [0,1]} \psi(x + \delta) \right| \right)$$

are bounded in δ or grow only slowly with δ . This can be readily verified for many singularity functions $\psi(x)$ of interest.

Difficulties may arise in applications, however, if integrals appear with a range of values of δ . The points $x_j(\delta_1)$ and $x_j(\delta_2)$ are not related by a simple scaling in this setting. The quadrature rule has to be constructed for each separate value of δ . One can conceivably approximate the functions $x_j(\delta)$ and $w_j(\delta)$ a priori as a function of δ . This approximation is a current subject of further study.

6. Numerical construction methods. A numerical method for the construction of generalized Gaussian quadrature rules was first described in [22]. Starting from a known classical Gaussian quadrature rule, a continuation process is started where the polynomial basis functions are transformed smoothly into the desired Chebyshev set of functions $\{\phi_k\}_{k=1}^{2n}$. At each intermediate stage in the process, generalized Gaussian quadrature rules are computed via Newton's method by solving a set of n nonlinear equations in the unknowns x_{nj} (thereby assuming that this intermediate rule exists, which, in general, need not be the case). The continuation is necessary to provide starting points for the final computation that are sufficiently close to the true solution, in order to ensure the convergence of Newton's method for the quadrature rule one is interested in.

A different approach was proposed in [31] by performing continuation on the weight function. There, the authors solve a nonlinear system of $2n$ equations

$$(6.1) \quad w_1\phi_k(x_1) + w_2\phi_k(x_2) + \cdots + w_n\phi_k(x_n) = I_\delta[\phi_k], \quad k = 1, \dots, 2n,$$

where the weight function depends on the continuation parameter δ . The size of the system is larger, with $2n$ equations rather than n , but the Jacobian assumes a much simpler form, and the method is reported to be more robust.

In this section, we briefly outline three separate approaches for the computation of generalized Gaussian quadrature rules in our framework of functions with an isolated singularity in $\psi(x)$.

6.1. Exploiting orthogonality. The function R_n completely characterizes the generalized Gaussian quadrature rule. Consider the case of even n , $n = 2l$, and recall

from definition (4.6) that $R_n(x) = x^l - \mathcal{P}_{\mathbf{x}}[x^l]$. Since $\mathcal{P}_{\mathbf{x}}[x^l] \in V_{n-1}$, we can write R_n in terms of x^l and a linear combination of the elements of T_{n-1} :

$$R_n(x) = x^l + \sum_{k=0}^{l-1} a_k x^k + \sum_{k=0}^{l-1} b_k x^k \psi(x).$$

This form has n unknowns. However, the function R_n satisfies $l = n/2$ orthogonality conditions (4.10) that result in linear relations in the unknown coefficients a_k and b_k . We may, therefore, write l coefficients in terms of the l other coefficients. The remaining l degrees of freedom are found by imposing the orthogonality conditions (4.11) for S_n .

This approach reduces the size of the nonlinear system of equations from $2n$ to $n/2$ equations. In principle, this is a substantial reduction. However, the Jacobian of this system of equations is rather involved. The method, in particular, requires an implementation of the mapping from R_n to S_n as given by (4.15). As a result, we found that, in practice, it is faster to solve the larger set of equations in all cases we considered. An easier method to exploit the existence of the functions R_n and S_n numerically does not seem apparent.

6.2. A bootstrapping algorithm. From the general theory of Chebyshev sets, one knows that the quadrature points $x_{n,j}$ of various n interlace, i.e.,

$$x_{n,j} \in (x_{n+1,j}, x_{n+1,j+1}).$$

From this, we construct starting points $x_{n+1,j}^*$ as follows. Having computed $x_{n,j}$, we set

$$(6.2) \quad \begin{aligned} x_{n+1,1}^* &= (a + x_{n,1})/2, \\ x_{n+1,j}^* &= (x_{n,j-1} + x_{n,j})/2, \quad j = 2, \dots, n, \\ x_{n+1,n+1}^* &= (x_{n,n} + b)/2. \end{aligned}$$

Newton's method is then used to solve the set of equations (1.1) with \mathbf{x}_{n+1}^* as starting points and starting weights computed from (4.9). The initial one-point rule Q_1 can usually be computed analytically. The weight $w_{1,1}$ is given explicitly by

$$w_{1,1} = \int_a^b w(x) dx,$$

and the corresponding quadrature point $x_{1,1}$ is found from

$$w_{1,1} \psi(x_{1,1}) = \int_a^b w(x) \psi(x) dx,$$

which results in an explicit expression in terms of the inverse of ψ ,

$$x_{1,1} = \psi^{-1} \left(\frac{\int_a^b w(x) \psi(x) dx}{\int_a^b w(x) dx} \right).$$

A small, yet crucial difference with [31] is that we use Newton's method with damping [24] to solve (6.1). This can be described as follows. Consider a general nonlinear system $F(\mathbf{y}) = 0$ with starting value $\mathbf{y}^0 = \mathbf{y}^*$. The typical Newton iteration is

$$\mathbf{y}^{j+1} = \mathbf{y}^j - J_F(\mathbf{y}^j)^{-1} F(\mathbf{y}^j), \quad j = 1, 2, \dots,$$

where $J_F(\mathbf{y})$ is used to denote the Jacobian of F at \mathbf{y} . This is replaced by a damped iteration

$$\mathbf{y}^{j+1} = \mathbf{y}^j - \alpha J_F(\mathbf{y}^j)^{-1} F(\mathbf{y}^j), \quad j = 1, 2, \dots,$$

where $0 < \alpha < 1$ is the damping parameter. The occasional lack of convergence of Newton's method without damping appears to be remedied in our setting by applying a small number of initial iterations with damping.

This approach has the clear advantage that no continuation is necessary. Even though this approach requires the computation of all lower order quadrature rules Q_m for $m = 1, \dots, n$, we found it to be fastest in practice. A disadvantage is that convergence of this approach is not guaranteed, not even when the damping parameter goes to zero. However, we found that the approach converged for all examples that we have implemented so far. Note that this is the method we have used in all numerical examples of this paper.

6.3. A continuation method. If the function $\psi(x)$ is a smooth function away from $x = 0$, an alternative continuation approach becomes viable. One can perform continuation on the parameter δ for functions of the form

$$f(x) = u(x) + v(x)\psi(x + \delta).$$

For large δ , the span of the basis T_{n-1} is close to the span of a polynomial basis. For increasing δ , the generalized Gaussian quadrature rule, therefore, converges to the classical Gaussian quadrature rule. Starting from the classical Gaussian quadrature rule and sufficiently large δ , continuation on δ may be performed until δ has the desired (small) value. Convergence is guaranteed by taking sufficiently small steps.

7. Examples. We end this paper with three numerical examples. We used the bootstrapping method described in section 6.2 to compute all quadrature rules with the following damping approach. If Newton's method without damping failed to converge, we started a new iteration from the starting values using a damping factor 1/2 in the first five iterations only. This process was repeated, halving the damping factor of the first five iterations after each restart, until convergence was achieved. No examples failed to converge with this approach. The majority of computations did not require any damping. Computations were performed in Maple in high-precision arithmetic in order to illustrate the convergence to high accuracy. The evaluation of the quadrature rules was also replicated for all examples in IEEE double precision in Matlab to confirm stability of the computations up to machine precision (this is not shown in the figures).

As our first example, consider the integral

$$I_1 := \int_0^1 H_0^{(1)}(x) dx,$$

where $H_0^{(1)}(x)$ is the Hankel function of the first kind of order zero. The integrand has the form $u(x) + v(x)\log(x)$, but it is not straightforward to obtain expressions for u and v [1]. The convergence rate is shown in Figure 7.1. Machine accuracy in double precision is obtained approximately at $n = 9$ points. The performance of the classical Gauss-Legendre rules on $[0, 1]$ is included in the same figure to illustrate the point that these rules, which completely ignore the singularity, also converge to the right value of the integral, though at a much slower rate.

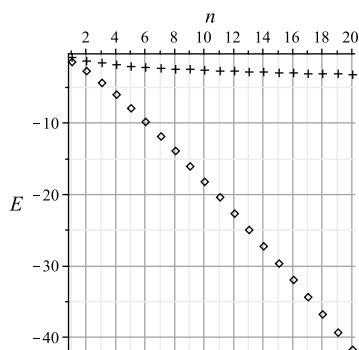


FIG. 7.1. Absolute error E in the approximation of I_1 by a classical (+) and a generalized (\diamond) Gaussian quadrature rule with n points. The error is shown in base-10 logarithmic scale.

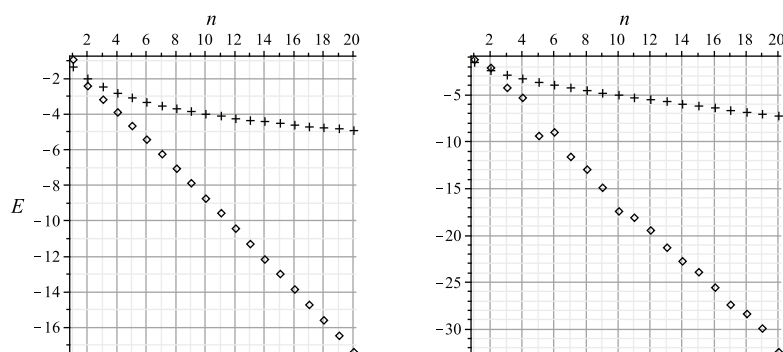


FIG. 7.2. Absolute error (in base-10 logarithmic scale) of classical (+) and generalized (\diamond) Gaussian quadrature for the logarithmically singular integral I_2 with exponential decay (left panel) and for the nearly singular integral I_3 involving a square root (right panel).

In the second example we consider the integral

$$I_2 = \int_0^{\infty} x H_1^{(1)}(x) e^{-(1+i)x} dx,$$

where $H_1^{(1)}$ is the Hankel function of the first kind of order zero. Integrals of this type appeared in computational models for scattering phenomena [14] for the evaluation of oscillatory integrals using a steepest-descent approach [13]. The integrand is continuous at $x = 0$ but has a logarithmic singularity in its derivatives. It decays like e^{-x} for large x . We constructed generalized Gaussian quadrature rules with the singularity function $\psi(x) = \log(x)$ and the weight function $w(x) = e^{-x}$. The results are shown in the left panel of Figure 7.2. The starting values (6.2) were slightly modified in this case because the right endpoint of the integration interval is infinite. As a starting value for the rightmost quadrature point, we used

$$\begin{aligned} x_{2,2}^* &= x_{1,1} + 2, \\ x_{n+1,n+1}^* &= x_{n,n} + (x_{n,n} - x_{n-1,n-1}), \quad n = 2, \dots \end{aligned}$$

We also included the performance of classical Gauss–Laguerre quadrature in this figure. The figure shows that Gauss–Laguerre rules also converge to the right value of the integral, but again, the convergence rate is much slower than that of generalized Gaussian quadrature.

TABLE 7.1

Points and weights of generalized Gaussian quadrature rules for integrands on $[0, \infty)$ with a logarithmic singularity and with the weight function $w(x) = e^{-x}$. The values are rounded to 17 significant digits.

N	x_j	w_j
5	0.26715698737809023 E - 1	0.95610457336708150 E - 1
	0.33502143286158755 E + 0	0.42443007452925864 E + 0
	0.13562361335644682 E + 1	0.39130875750538623 E + 0
	0.35348502336885737 E + 1	0.85922136050594078 E - 1
	0.76316657306249149 E + 1	0.27285745780529126 E - 2
10	0.43429655850744155 E - 2	0.16337243162432338 E - 1
	0.61440851829453254 E - 1	0.10976709295285857 E + 0
	0.28016211921371600 E + 0	0.25898249226845143 E + 0
	0.78997151808506692 E + 0	0.31663066290136217 E + 0
	0.17155826587620794 E + 1	0.21103870117564605 E + 0
	0.31775355330978616 E + 1	0.73891262254266567 E - 1
	0.53108336892605114 E + 1	0.12463634260257873 E - 1
	0.83029731539047433 E + 1	0.87005369176965746 E - 3
	0.12484738636799131 E + 2	0.18796353846581537 E - 4
	0.18673778361448400 E + 2	0.60979108757903096 E - 7
15	0.13999610893490940 E - 2	0.53163416586656096 E - 2
	0.20477656406851898 E - 1	0.38934106518844350 E - 1
	0.97678338044921310 E - 1	0.11259213060566958 E + 0
	0.28882979020397569 E + 0	0.20110058067713363 E + 0
	0.65499962351242915 E + 0	0.24617478546017735 E + 0
	0.12561634158023614 E + 1	0.21010745635601456 E + 0
	0.21493397187772781 E + 1	0.12328344052503329 E + 0
	0.33901241449672296 E + 1	0.48314526729893461 E - 1
	0.50365761536512140 E + 1	0.12149894815521587 E - 1
	0.71552991111532403 E + 1	0.18588202790220021 E - 2
	0.98309561518905044 E + 1	0.16073588144237742 E - 3
	0.13183148125168729 E + 2	0.70482542970423946 E - 5
	0.17402079434042808 E + 2	0.13148642394391024 E - 6
	0.22843724056603261 E + 2	0.75127862808739610 E - 9
	0.30407448759772375 E + 2	0.58258282243283305 E - 12

In the third example we consider the integral

$$I_3 := \int_0^1 \sqrt{0.01 + x + x^2} (\cos(x) + \sin(x)) dx.$$

This example illustrates both the advantages and disadvantages of generalized Gaussian quadrature for nearly singular integrals. The integral behaves as $u(x)\sqrt{x - \epsilon} + v(x)$ for $x \rightarrow \epsilon$, where $\epsilon = -0.0101\dots$ is the root of

$$0.01 + x + x^2 = 0$$

closest to the interval $[0, 1]$. Convergence is illustrated in the right panel of Figure 7.2 using $\psi(x) = \sqrt{x + \epsilon}$. Similar though slightly worse results were obtained by using $\epsilon = 0.01$. The disadvantages for nearly singular integrals are that best results are obtained with a sharp estimate of ϵ and that the quadrature rule depends on ϵ . The advantage is that convergence is very rapid. Machine accuracy in double precision is reached approximately at $n = 9$ points. Gauss–Legendre quadrature for this integral converges exponentially because the singularity is outside the integration interval. However, in this example too, the rate of convergence is significantly slower than that of generalized Gaussian quadrature, as expected.

Finally, Table 7.1 displays some of the quadrature rules that were used to compute the second example of this paper. These rules are useful for the evaluation of singular

and highly oscillatory integrals that may appear in scattering computations [14]. The quadrature rules for integrands with a logarithmic endpoint singularity, as in the first example of this section, are useful in a wide range of applications. They were listed in [22] and are not repeated here.

REFERENCES

- [1] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover Publications, New York, 1965.
- [2] M. CARLEY, *Numerical quadratures for singular and hypersingular integrals in boundary element methods*, SIAM J. Sci. Comput., 29 (2007), pp. 1207–1216.
- [3] R. COOLS AND A. HAEGEMANS, *Algorithm 824: CUBPACK: A package for automatic cubature; framework description*, ACM Trans. Math. Software, 29 (2003), pp. 287–296.
- [4] R. COOLS AND E. NOVAK, *Spherical product algorithms and the integration of smooth functions with one singular point*, SIAM J. Numer. Anal., 39 (2001), pp. 1132–1145.
- [5] R. COOLS AND J. C. SANTOS-LEON, *Cubature formulas of a nonalgebraic degree of precision*, Constr. Approx., 18 (2002), pp. 223–240.
- [6] R. COOLS, *Constructing cubature formulae: The science behind the art*, Acta Numer., 6 (1997), pp. 1–54.
- [7] J. CROW, *Quadrature of integrands with a logarithmic singularity*, Math. Comp., 60 (1993), pp. 297–301.
- [8] P. J. DAVIS AND P. RABINOWITZ, *Methods of numerical integration*, Computer Science and Applied Mathematics, Academic Press, New York, 1984.
- [9] W. GAUTSCHI, *Algorithm 726: ORTHPOL—a package of routines for generating orthogonal polynomials and Gauss-type quadrature rules*, ACM Trans. Math. Software, 20 (1994), pp. 21–62.
- [10] A. GLASER, X. LIU, AND V. ROKHLIN, *A fast algorithm for the calculation of the roots of special functions*, SIAM J. Sci. Comput., 29 (2007), pp. 1420–1438.
- [11] G. H. GOLUB AND J. H. WELSCH, *Calculation of Gauss quadrature rules*, Math. Comp., 23 (1969), pp. 221–230.
- [12] C. G. HARRIS AND W. A. B. EVANS, *Extension of numerical quadrature formulae to cater for end point singular behaviors over finite intervals*, Int. J. Comput. Math. (B), 6 (1977), pp. 219–227.
- [13] D. HUYBRECHS AND S. VANDEWALLE, *On the evaluation of highly oscillatory integrals by analytic continuation*, SIAM J. Numer. Anal., 44 (2006), pp. 1026–1048.
- [14] D. HUYBRECHS AND S. VANDEWALLE, *A sparse discretization for integral equation formulations of high frequency scattering problems*, SIAM J. Sci. Comput., 29 (2007), pp. 2305–2328.
- [15] D. HUYBRECHS AND S. VANDEWALLE, *An efficient implementation of boundary element methods for computationally expensive Green's functions*, Eng. Anal. Bound. Elem., 32 (2008), pp. 621–632.
- [16] S. KARLIN AND W. STUDDEN, *Tchebysheff Systems with Applications in Analysis and Statistics*, Wiley-Interscience, New York, 1966.
- [17] P. KOLM, S. JIANG, AND V. ROKHLIN, *Quadruple and octuple layer potentials in two dimensions I: Analytical apparatus*, Appl. Comput. Harmon. Anal., 14 (2003), pp. 47–74.
- [18] P. KOLM AND V. ROKHLIN, *Numerical quadrature for singular and hypersingular integrals*, Comput. Math. Appl., 41 (2001), pp. 327–352.
- [19] A. S. KRONROD, *Nodes and Weights of Quadrature Formulas*, Consultants Bureau, New York, 1965.
- [20] D. LAURIE, *Computation of Gauss-type quadrature formulas*, J. Comput. Appl. Math., 127 (2001), pp. 201–217.
- [21] A. A. MARKOV, *On the limiting values of integrals in connection with interpolation*, Zap. Imp. Akad. Nauk. Fiz.-Mat. Otd., 6 (1898), pp. 146–230.
- [22] J. MA, V. ROKHLIN, AND S. WANDZURA, *Generalized Gaussian quadrature rules for systems of arbitrary functions*, SIAM J. Numer. Anal., 33 (1996), pp. 971–996.
- [23] T. N. L. PATTERSON, *The optimum addition of points to quadrature formulae*, Math. Comp., 22 (1968), pp. 847–856.
- [24] M. J. D. POWELL, *Approximation Theory and Methods*, Cambridge University Press, Cambridge, 1981.
- [25] H. J. SCHMID, *Interpolatorische Kubaturformeln*, Dissertationes Math. 220, Polish Scientific Publishers, Warszawa, Poland, 1983.

- [26] C. SCHWAB, *Variable order composite quadrature of singular and nearly singular integrals*, Computing, 53 (1994), pp. 173–194.
- [27] R. N. L. SMITH, *Direct Gauss quadrature formulae for logarithmic singularities on isoparametric elements*, Eng. Anal. Bound. Elem., 24 (2000), pp. 161–167.
- [28] A. H. STROUD AND D. SECREST, *Gaussian Quadrature Formulas*, Prentice-Hall, Englewood Cliffs, NJ, 1966.
- [29] H. TAKAHASI AND M. MORI, *Quadrature formulas obtained by variable transformation*, Numer. Math., 21 (1973), pp. 206–219.
- [30] L. N. TREFETHEN, *Is Gauss quadrature better than Clenshaw–Curtis?*, SIAM Rev., 50 (2008), pp. 67–87.
- [31] N. YARVIN AND V. ROKHLIN, *Generalized Gaussian quadratures and singular value decompositions of integral operators*, SIAM J. Sci. Comput., 20 (1998), pp. 699–718.

A POSTERIORI ANALYSIS AND ADAPTIVE ERROR CONTROL FOR MULTISCALE OPERATOR DECOMPOSITION SOLUTION OF ELLIPTIC SYSTEMS I: TRIANGULAR SYSTEMS*

V. CAREY[†], D. ESTEP[‡], AND S. TAVENER[†]

Abstract. In this paper, we perform an a posteriori error analysis of a multiscale operator decomposition finite element method for the solution of a system of coupled elliptic problems. The goal is to compute accurate error estimates that account for the effects arising from multiscale discretization via operator decomposition. Our approach to error estimation is based on a well-known a posteriori analysis involving variational analysis, residuals, and the generalized Green's function. Our method utilizes adjoint problems to deal with several new features arising from the multiscale operator decomposition. In part I of this paper, we focus on the propagation of errors arising from the solution of one component to another and the transfer of information between different representations of solution components. We also devise an adaptive discretization strategy based on the error estimates that specifically controls the effects arising from operator decomposition. In part II of this paper, we address issues related to the iterative solution of a fully coupled nonlinear system.

Key words. a posteriori error analysis, adjoint problem, elliptic system, generalized Green's function, goal-oriented error estimates, multiscale methods, operator decomposition, projection error

AMS subject classifications. 65N15, 65N30, 65N50

DOI. 10.1137/070689917

1. Introduction. Multiscale operator decomposition is a widely used technique for solving multiphysics, multiscale problems [14, 15]. The general approach is to decompose the multiphysics problem into components involving simpler physics over a relatively limited range of scales and then to seek the solution of the entire system through some sort of iterative procedure involving solutions of the individual components. This approach is appealing because there is generally a good understanding of how to solve a broad spectrum of single physics problems accurately and efficiently, and because it provides an alternative to accommodating multiple scales in one discretization. However, multiscale operator decomposition presents an entirely new set of accuracy and stability issues, some of which are obvious and some subtle, and all of which are difficult to correct.

We motivate multiscale operator decomposition for elliptic systems by considering a model of a thermal actuator. A thermal actuator is a microelectronic mechanical switch device (see Figure 1.1). A contact rests on thin braces composed of a conducting material. When a current is passed through the braces, they heat up and consequently expand to close the contact. The system is modeled by a system of three coupled equations, each representing a distinct physical process. They are an

*Received by the editors April 30, 2007; accepted for publication (in revised form) June 6, 2008; published electronically February 4, 2009.

<http://www.siam.org/journals/sinum/47-1/68991.html>

[†]Department of Mathematics, Colorado State University, Fort Collins, CO 80523 (carey@math.colostate.edu, tavener@math.colostate.edu). The work of these authors was supported in part by the Department of Energy (DE-FG02-04ER25620).

[‡]Department of Mathematics and Department of Statistics, Colorado State University, Fort Collins, CO 80523 (estep@math.colostate.edu). This author's work was supported in part by the Department of Energy (DE-FG02-04ER25620, DE-FG02-05ER25699, DE-FC02-07ER54909), the National Aeronautics and Space Administration (NNG04GH63G), the National Science Foundation (DMS-0107832, DMS-0715135, DGE-0221595003, MSPA-CSE-0434354, ECCS-0700559), Idaho National Laboratory (00069249), and the Sandia Corporation (PO299784).

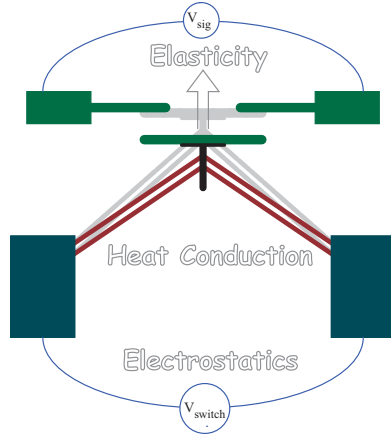


FIG. 1.1. Sketch of a thermal actuator.

electrostatic current equation

$$(1.1) \quad \nabla \cdot (\sigma \nabla V) = 0,$$

governing potential V (where current $J = -\sigma \nabla V$), a steady-state energy equation

$$(1.2) \quad \nabla \cdot (\kappa(T) \nabla T) = \sigma (\nabla V \cdot \nabla V),$$

governing temperature T , and a linear elasticity equation giving the steady-state displacement d ,

$$(1.3) \quad \nabla \cdot (\lambda \operatorname{tr}(E)I + 2\mu E - \beta(T - T_{ref})I) = 0, \quad E = (\nabla d + \nabla d^T)/2.$$

Using multiscale operator decomposition, the complete system (1.1–1.3) is decomposed into three components, each of which is solved with a code specialized to the particular type of physics. Notice that the electric potential V can be calculated independently of T and d . The temperature T can be calculated once the electric potential V is known, while the calculation of displacement d requires prior knowledge of T and therefore of V .

In general, we can write a coupled elliptic system on a domain Ω in the form

$$(1.4) \quad \begin{cases} \mathcal{L}_1(x, u_1, Du_1, \dots, u_n, Du_n) = 0, \\ \vdots \\ \mathcal{L}_n(x, u_1, Du_1, \dots, u_n, Du_n) = 0. \end{cases} \quad x \in \Omega.$$

A natural form of operator decomposition is to split the global multiphysics problem into n “single-physics” components that are solved individually. In general, the solution of each component requires knowledge of the solutions of all the other components; the full problem requires some form of iteration to obtain the solution.

It is possible to impose conditions on the system, the components, and the coupling that allow for an a priori convergence analysis. However, operator decomposition is problematic in practice because it is very difficult to verify such conditions and often impractical to satisfy them. Indeed, numerical solutions obtained via operator decomposition are affected significantly by the specific choice of decomposition. In

this paper, we perform an a posteriori error analysis of a multiscale operator decomposition finite element method for the solution of a system of coupled elliptic problems. The components of the problem are solved in sequence using independent discretizations. The goal is to compute *accurate* computational error estimates that specifically account for the effects arising from operator decomposition. We also devise an adaptive discretization strategy based on the error estimates that controls the effects arising from multiscale operator decomposition.

The a posteriori analysis in this paper is based on a well-known approach involving variational analysis, residuals, and the generalized Green's function solving an adjoint problem [1, 2, 5, 6, 7, 8, 9, 12]. However, we modify this approach to accommodate several new features arising from the operator decomposition. Three important issues addressed here are as follows: (1) Errors in the solution of each component propagate into the solutions of the other components; (2) Transferring information between different discretization representations potentially introduces new error; and (3) The adjoint operators associated with the fully coupled system and an operator decomposition version are not generally equal. In addition, the analysis stays within the "single physics paradigm" by only requiring the solution of adjoint problems associated with the individual components. These issues are characteristic of a broad range of operator decomposition discretizations, e.g., [10, 13], and generally require extensions to the usual a posteriori analysis techniques.

In this paper, we focus attention on analyzing the effects of transferring information between components, which is necessitated by operator decomposition. In order to do so, we consider a "triangular" or one-way coupled system

$$(1.5) \quad \begin{cases} \mathcal{L}_1(x, u_1, Du_1) = 0, \\ \mathcal{L}_2(x, u_1, Du_1, u_2, Du_2) = 0, \\ \mathcal{L}_3(x, u_1, Du_1, u_2, Du_2, u_3, Du_3) = 0, \\ \vdots \\ \mathcal{L}_n(x, u_1, Du_1, u_2, Du_2, u_3, Du_3, \dots, u_n, Du_n) = 0. \end{cases} \quad x \in \Omega.$$

This system can be solved by a finite sequence of component solutions by considering the n problems for $\mathcal{L}_1, \mathcal{L}_2, \dots, \mathcal{L}_n$ sequentially. Such systems are important in practice, e.g., the thermal actuator (1.1)–(1.3) has this form. In part II [3], we consider additional sources of error arising from the iterative procedure required when solving a fully-coupled system via operator decomposition.

We capture the essential features of (1.5) in a two component "one-way" coupled system of the form

$$(1.6) \quad \begin{cases} -\nabla \cdot a_1 \nabla u_1 + \mathbf{b}_1 \cdot \nabla u_1 + c_1 u_1 = f_1(x), & x \in \Omega, \\ -\nabla \cdot a_2 \nabla u_2 + \mathbf{b}_2 \cdot \nabla u_2 + c_2 u_2 = f_2(x, u_1, Du_1), & x \in \Omega, \\ u_1 = u_2 = 0, & x \in \partial\Omega, \end{cases}$$

where a_i, b_i, c_i, f_i are smooth functions on a bounded domain Ω in \mathbb{R}^N with boundary $\partial\Omega$ and the coupling occurs through f_2 . We later generalize to coupling through the coefficients of the elliptic operator for u_2 .

In section 2, we illustrate the main idea by applying the analysis to a linear algebraic system. We perform the transfer error analysis in section 3 and present computational examples when the corresponding discretizations are "related" in the sense that either both computational meshes are identical, or one mesh is generated by

a sequence of mesh refinements on the other mesh. In section 4, we consider the effect of using distinct discretizations for the two components and analyze the additional errors caused by using projections between the components. Additionally, we discuss the use of Monte Carlo integration to estimate these projection errors. We present the full adaptive algorithm in section 5, which we illustrate with several numerical examples.

2. A linear algebra example. We introduce the notation and ideas in the context of a lower triangular linear system of equations. Let U be an approximate solution of the linear system $\mathbf{A}u = b$. We wish to compute a quantity of interest given by a linear functional (ψ, u) . The error $e = u - U$ is not computable, but we can compute the residual $R = b - \mathbf{A}U = \mathbf{A}e$. Using the solution ϕ of the corresponding adjoint equation $\mathbf{A}^\top \phi = \psi$, the error representation for a linear functional of the solution is

$$(\psi, u) - (\psi, U) = (\psi, e) = (\mathbf{A}^\top \phi, e) = (\phi, \mathbf{A}e) = (\phi, R).$$

Now consider the triangular system

$$(2.1) \quad \mathbf{A}u = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = b,$$

with approximate solution

$$U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} \approx \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = u.$$

We estimate the error in a quantity of interest in u_2 only, given by the linear functional

$$(\psi^{(1)}, u) = (\psi_2^{(1)}, u_2), \text{ where } \psi = \begin{pmatrix} 0 \\ \psi_2^{(1)} \end{pmatrix}.$$

We employ the superscript (1), since we later pose additional auxiliary adjoint problems. Clearly, estimates on linear functionals of u_1 are independent of u_2 . The lower triangular structure of \mathbf{A} yields

$$\begin{aligned} \mathbf{A}_{11}u_1 &= b_1, \\ \mathbf{A}_{22}u_2 &= b_2 - \mathbf{A}_{21}u_1, \end{aligned}$$

and the corresponding residuals are

$$\begin{aligned} R_1 &= b_1 - \mathbf{A}_{11}U_1, \\ R_2 &= (b_2 - \mathbf{A}_{21}U_1) - \mathbf{A}_{22}U_2. \end{aligned}$$

The residual R_2 depends upon the solution of the first component, and any attempt to decrease this residual requires a consideration of the accuracy of U_1 . The adjoint problem to (2.1) is

$$\begin{pmatrix} \mathbf{A}_{11}^\top & \mathbf{A}_{21}^\top \\ \mathbf{0} & \mathbf{A}_{22}^\top \end{pmatrix} \begin{pmatrix} \phi_1^{(1)} \\ \phi_2^{(1)} \end{pmatrix} = \begin{pmatrix} 0 \\ \psi_2^{(1)} \end{pmatrix},$$

and the resulting error representation is

$$\begin{aligned}
 (\psi^{(1)}, e) &= (\psi_2^{(1)}, e_2) = (\mathbf{A}_{22}^\top \phi_2^{(1)}, e_2) \\
 &= (\phi_2^{(1)}, \mathbf{A}_{22} u_2) - (\phi_2^{(1)}, \mathbf{A}_{22} U_2) \\
 (2.2) \quad &= (\phi_2^{(1)}, b_2 - \mathbf{A}_{21} u_1) - (\phi_2^{(1)}, \mathbf{A}_{22} U_2) \\
 &= (\phi_2^{(1)}, b_2 - \mathbf{A}_{21} U_1 - \mathbf{A}_{22} U_2) - (\phi_2^{(1)}, \mathbf{A}_{21} e_1) \\
 &= (\phi_2^{(1)}, R_2) - (\phi_2^{(1)}, \mathbf{A}_{21} e_1).
 \end{aligned}$$

The first term of the error representation requires only U_2 and $\phi_2^{(1)}$. Since the adjoint system is upper triangular and

$$\phi_2^{(1)} = \left(\mathbf{A}_{22}^\top \right)^{-1} \psi_2^{(1)}$$

is independent of the first component, the calculation of $(\phi_2^{(1)}, R_2)$ remains within the “single physics paradigm.” The second term $(\phi_2^{(1)}, \mathbf{A}_{21} e_1)$ represents the effect of errors in U_1 on the solution U_2 . At first glance, this term is uncomputable, but we note that it is a *linear functional* of e_1 since

$$(\phi_2^{(1)}, \mathbf{A}_{21} e_1) = (\mathbf{A}_{21}^\top \phi_2^{(1)}, e_1).$$

We therefore form the adjoint problem for the *transfer error*

$$\begin{pmatrix} \mathbf{A}_{11}^\top & \mathbf{A}_{21}^\top \\ \mathbf{0} & \mathbf{A}_{22}^\top \end{pmatrix} \begin{pmatrix} \phi_1^{(2)} \\ \phi_2^{(2)} \end{pmatrix} = \begin{pmatrix} \psi_1^{(2)} \\ 0 \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{21}^\top \phi_2^{(1)} \\ 0 \end{pmatrix}.$$

The upper triangular block structure of \mathbf{A}^\top immediately yields $\phi_2^{(2)} = 0$. As noted earlier, error estimates of u_1 should be independent of u_2 . Thus, $\mathbf{A}_{11}^\top \phi_1^{(2)} = \psi_1^{(2)} = \mathbf{A}_{21}^\top \phi_2^{(1)}$, so that, once again, we can solve for $\phi^{(2)}$ in the “single physics paradigm.” Given $\phi^{(2)}$, we obtain the secondary error representation

$$(2.3) \quad (\psi^{(2)}, e) = (\psi_1^{(2)}, e_1) = (\mathbf{A}_{21}^\top \phi_2^{(1)}, e_1) = (\mathbf{A}_{11}^\top \phi_1^{(2)}, e_1) = (\phi_1^{(2)}, R_1).$$

Combining the first term of (2.2) with (2.3) yields the complete error representation

$$(2.4) \quad (\psi^{(1)}, e) = (\phi_2^{(1)}, R_2) - (\phi_1^{(2)}, R_1),$$

which is a sum of the inner products of “single physics” residuals and adjoint solutions computed using the “single physics” paradigm.

3. Analysis of the discretization error. The corresponding weak form of (1.6) reads as follows: find $u_i \in \tilde{W}_2^1(\Omega)$ satisfying

$$(3.1) \quad \begin{cases} \mathcal{A}_1(u_1, v_1) = (f_1, v_1), \\ \mathcal{A}_2(u_2, v_2) = (f_2(x, u_1), v_2) \end{cases} \quad \forall v_i \in \tilde{W}_2^1(\Omega),$$

where

$$\begin{aligned}
 \mathcal{A}_1(u_1, v_1) &= \mathcal{A}_1(u_1, v_1) \equiv (a_1 \nabla u_1, \nabla v_1) + (\mathbf{b}_1(x) \cdot \nabla u_1, v_1) + (c_1 u_1, v_1), \\
 \mathcal{A}_2(u_2, v_2) &= \mathcal{A}_2(u_2, v_2) \equiv (a_2 \nabla u_2, \nabla v_2) + (\mathbf{b}_2(x) \cdot \nabla u_2, v_2) + (c_2 u_2, v_2)
 \end{aligned}$$

are assumed to be coercive bilinear forms on Ω and $\tilde{W}_p^m(\Omega)$ is the subspace of $W_p^m(\Omega)$ with zero trace on $\partial\Omega$. We suppress the “cross” dependence on the other solutions except in a few remarks below. After introducing (conforming) discretizations $\mathcal{S}_{h,i}(\Omega)$, we solve the discretized system

$$(3.2) \quad \begin{cases} \mathcal{A}_1(U_1, \chi_1) = (f_1, \chi_1), \\ \mathcal{A}_2(U_2, \chi_2) = (f_2(x, U_1, DU_1), \chi_2) \end{cases} \quad \forall \chi_i \in \mathcal{S}_{h,i}(\Omega).$$

In general, however, $\mathcal{S}_{h,1} \subsetneq \mathcal{S}_{h,2}$ (or vice-versa) on Ω , and we may be forced to work with either $\Pi_{1 \rightarrow 2} f_2(U_1)$ or more generally with $f_2(x, \Pi_{1 \rightarrow 2} U_1, \Pi_{1 \rightarrow 2} DU_1)$, where $\Pi_{i \rightarrow j}$ is some projection from $\mathcal{S}_{h,i}$ to $\mathcal{S}_{h,j}$. If the projection is to $\mathcal{S}_{h,i}$ from $\tilde{W}_2^1(\Omega_i)$, then we simply write the projection as Π_i . The resulting discrete system becomes

$$(3.3) \quad \begin{cases} \mathcal{A}_1(U_1, \chi_1) = (f_1, \chi_1), \\ \mathcal{A}_2(U_2, \chi_2) = (f_2(x, \Pi_{1 \rightarrow 2} U_1, \Pi_{1 \rightarrow 2} DU_1), \chi_2) \end{cases} \quad \forall \chi_i \in \mathcal{S}_{h,i}(\Omega).$$

Primary adjoint problem. We seek the error in a quantity of interest representable by a linear functional of the error e_2 , where $u_i - U_i = e_i$ denotes the pointwise errors. Note that a quantity of interest involving only u_1 can be computed without solving for u_2 , hence, there is no loss of generality. The *global* adjoint problem, defined relative to the quantity of interest, is

$$\begin{cases} -\nabla \cdot a_1 \nabla \phi_1^{(1)} - \operatorname{div}(\mathbf{b}_1 \phi_1^{(1)}) + c_1 \phi_1^{(1)} + Lf_2(u_1) \phi_2^{(1)} = 0, \\ -\nabla \cdot a_2 \nabla \phi_2^{(1)} - \operatorname{div}(\mathbf{b}_2 \phi_2^{(1)}) + c_2 \phi_2^{(1)} = \psi_2^{(1)}, \end{cases}$$

where

$$Lf_2(u_1)(u_1 - U_1) = \int_0^1 \frac{\partial f_2}{\partial u_1}(u_1 s + U_1(1 - s)) ds$$

is a linearization of f_2 and $\phi_1^{(1)}$ and $\phi_2^{(1)}$ satisfy homogeneous Dirichlet boundary conditions. The corresponding weak formulation is

$$(3.4) \quad \begin{cases} \mathcal{A}_1^*(\phi_1^{(1)}, v_1) + (Lf_2(u_1) \phi_2^{(1)}, v_1) = 0, \\ \mathcal{A}_2^*(\phi_2^{(1)}, v_2) = (\psi_2^{(1)}, v_2) \end{cases} \quad \forall v_i \in \tilde{W}_2^1(\Omega),$$

where

$$(3.5) \quad \begin{cases} \mathcal{A}_1^*(\phi_1^{(1)}, v_1) = (a_1 \nabla \phi_1^{(1)}, \nabla v_1) - (\operatorname{div}(\mathbf{b}_1 \phi_1^{(1)}), v_1) + (c_1 \phi_1^{(1)}, v_1), \\ \mathcal{A}_2^*(\phi_2^{(1)}, v_2) = (a_2 \nabla \phi_2^{(1)}, \nabla v_2) - (\operatorname{div}(\mathbf{b}_2 \phi_2^{(1)}), v_2) + (c_2 \phi_2^{(1)}, v_2). \end{cases}$$

Using the standard argument, we have the following error representation formula:

$$(3.6) \quad (\psi^{(1)}, e) = (\psi_2^{(1)}, e_2) = \mathcal{A}_2^*(\phi_2^{(1)}, e_2) = (f_2(x, u_1, Du_1), \phi_2^{(1)}) - \mathcal{A}_2(U_2, \phi_2^{(1)}).$$

Observe that $\phi_1^{(1)}$ does not appear in the error representation formula. We define the primary adjoint problem as

$$\mathcal{A}_2^*(\phi_2^{(1)}, v_2) = (\psi_2^{(1)}, v_2) \quad \forall v_2 \in \tilde{W}_2^1(\Omega).$$

Remark 3.1. At first glance, it appears that we need only to solve the second adjoint equation and thus do not need to construct the linearization Lf_2 . However, as seen in the linear algebra example, the analysis takes into account the transfer

of error from the solution of the first component. Estimating this transferred error uses a nonlinear functional of the error to form the right-hand sides in “transfer adjoint problems” (2.3) and (3.11). We approximate this nonlinear functional using the linearization Lf_2 . We evaluate the linearization at the computed solution U , which can be justified by using Taylor’s theorem and assuming that the error $u - U$ is sufficiently small.

Adding and subtracting the projection of $\phi_2^{(1)}$ onto the primal approximation space $(\Pi_2\phi_2^{(1)})$ in (3.6) yields

$$(3.7) \quad (\psi_2^{(1)}, e_2) = (f_2(x, u_1, Du_1), (I - \Pi_2)\phi_2^{(1)}) - \mathcal{A}_2(U_2, (I - \Pi_2)\phi_2^{(1)}) \\ + (f_2(x, u_1, Du_1), \Pi_2\phi_2^{(1)}) - \mathcal{A}_2(U_2, \Pi_2\phi_2^{(1)}).$$

To simplify later constructions, we introduce the notion of the weak residual of a solution component, namely,

$$\mathcal{R}_i(U_i, \chi; \nu) = (f_i(\nu), \chi) - \mathcal{A}_i(U_i, \chi; \nu)$$

and using this notation write (3.6) as

$$(\psi^{(1)}, e) = \mathcal{R}_2(U_2, \phi_2^{(1)}; u_1),$$

indicating that this estimate depends on the solution u_1 .

3.1. Transfer error analysis. Error representation (3.7) is not computable, since u_1 is unknown. We add and subtract $(f_2(x, U_1, DU_1), (I - \Pi_2)\phi_2^{(1)})$ from error representation formula (3.7) and use the definition of approximate weak statement (3.2) to obtain

$$(3.8) \quad (\psi_2^{(1)}, e_2) = (f_2(x, U_1, DU_1), (I - \Pi_2)\phi_2^{(1)}) - \mathcal{A}_2(U_2, (I - \Pi_2)\phi_2^{(1)}) \\ + (f_2(x, u_1, Du_1) - f_2(x, U_1, DU_1), \phi_2^{(1)}) \\ = \mathcal{R}_2(U_2, (I - \Pi_2)\phi_2^{(1)}; U_1) + (f_2(x, u_1, Du_1) - f_2(x, U_1, DU_1), \phi_2^{(1)}).$$

The first term on the right of (3.8) is a traditional dual-weighted residual expression for the error arising from discretization of the second component, while the remaining difference represents the *transfer error* that arises from using an approximation of u_1 in defining the coefficients in the equation for u_2 . The goal now is to estimate this transfer error and its effect on the quantity of interest.

As with the linear algebra example in section 2, we recognize the transfer error expression as a functional of error in u_1 and define

$$(f_2(x, u_1, Du_1) - f_2(x, U_1, DU_1), \phi_2^{(1)})$$

as a new quantity of interest. Then, we construct a secondary adjoint problem to compute the transfer error. In order to obtain a linear functional when f_2 is nonlinear in u_1 , we linearize $f_2(u_1) \approx f_2(U_1) + Df_2(U_1) \times (u_1 - U_1)$, where Df is the Fréchet derivative of f_2 at U_1 . The transfer error term becomes

$$(3.9) \quad (Df_2(U_1) \times e_1, \phi_2^{(1)}),$$

which is a linear functional of the error e_1 that describes the effect of errors in U_1 on the quantity of interest. Note that the Riesz representation theorem guarantees the existence of a $\psi_1^{(2)}$ such that $(\psi_1^{(2)}, e_1)$ equals (3.9), though $\psi_1^{(2)}$ is not needed to evaluate the functional or compute the corresponding adjoint solution.

Transfer error adjoint problem. To estimate the new quantity of interest, we define

$$\begin{cases} (a_1 \nabla \phi_1^{(2)}, \nabla v_1) - (\operatorname{div}(\mathbf{b}_1 \phi_1^{(2)}), v_1) + (c_1 \phi_1^{(2)}, v_1) + (L f_2(u_1) \phi_2^{(2)}, v_1) = \psi_1^{(2)}, \\ (a_2 \nabla \phi_2^{(2)}, \nabla v_2) - (\operatorname{div}(\mathbf{b}_2 \phi_2^{(2)}), v_2) + (c_2 \phi_2^{(2)}, v_2) = 0, \end{cases} \quad \forall v_i \in \tilde{W}_2^1(\Omega).$$

The second equation has the trivial solution, and the secondary adjoint problem reduces to the “transfer error adjoint problem”

$$(3.10) \quad (a_1 \nabla \phi_1^{(2)}, \nabla v_1) - (\operatorname{div}(\mathbf{b}_1 \phi_1^{(2)}), v_1) + (c_1 \phi_1^{(2)}, v_1) = (\psi_1^{(2)}, v_1) \quad \forall v_1 \in \tilde{W}_2^1(\Omega).$$

The transfer error representation formula is given by

$$(3.11) \quad \begin{aligned} (\psi_1^{(2)}, e_1) &= \mathcal{A}_1^*(\phi_1^{(2)}, e_1) = \mathcal{A}_1(e_1, \phi_1^{(2)}) \\ &= (f_1, (I - \Pi_1)\phi_1^{(2)}) - \mathcal{A}_1(U_1, (I - \Pi_1)\phi_1^{(2)}), \end{aligned}$$

where we have used Galerkin orthogonality to introduce the projection of ϕ onto the discretization space (as f_1 does not depend on u). Inserting (3.11) into (3.8) yields

$$(3.12) \quad \begin{aligned} (\psi, e) &= (f_2(x, U_1, DU_1), (I - \Pi_2)\phi_2^{(1)}) - \mathcal{A}_2(U_2, (I - \Pi_2)\phi_2^{(1)}) \\ &\quad + (f_1, (I - \Pi_1)\phi_1^{(2)}) - \mathcal{A}_1(U_1, (I - \Pi_1)\phi_1^{(2)}) \end{aligned}$$

or

$$(\psi, e) = \mathcal{R}_2(U_2, (I - \Pi_2)\phi_2^{(1)}; U_1) + \mathcal{R}_1(U_1, (I - \Pi_1)\phi_1^{(2)}).$$

Remark 3.2. If the model problem includes coupling in the coefficients of the second differential operator, i.e.,

$$(3.13) \quad \begin{cases} -\nabla \cdot a_1(x) \nabla u_1 + \mathbf{b}_1(x) \cdot \nabla u_1 + c_1(x) u_1 = f_1(x), & x \in \Omega, \\ -\nabla \cdot a_2(x, u_1) \nabla u_2 + \mathbf{b}_2(x, u_1) \cdot \nabla u_2 + c_2(x, u_1) u_2 = f_2(x, u_1, Du_1), & x \in \Omega, \\ u_1 = u_2 = 0, & x \in \partial\Omega, \end{cases}$$

then the error representation formula for a quantity of interest that depends on u_2 alone is

$$(\psi, e) = \mathcal{R}_2(U_2, (I - \Pi_2)\phi_2^{(1)}; u_1).$$

Since this is not computable, we replace each term in the weak residual with the same term evaluated at U_1 , yielding

$$\begin{aligned} (\psi, e) &= \mathcal{R}_2(U_2, (I - \Pi_2)\phi_2^{(1)}; U_1) + (f_2(u_1) - f_2(U_1), \phi_2^{(1)}) \\ &\quad - ((a_2(u_1) - a_2(U_1))U_2, \phi^{(1)}) - ((\mathbf{b}_2(u_1) - \mathbf{b}_2(U_1)) \cdot \nabla U_2, \phi^{(1)}) \\ &\quad - ((c_2(u_1) - c_2(U_1))U_2, \phi^{(1)}). \end{aligned}$$

We linearize f_2 , a_2 , b_2 , and c_2 around U_1 to obtain an approximate transfer error term

$$(Df_2(e_1), \phi_2^{(1)}) + (Da_2(e_1)\nabla U_2, \nabla\phi_2^{(1)}) + (D\mathbf{b}_2(e_1) \cdot \nabla U_2, \phi_2^{(1)}) + (Dc_2(e_1)U_2, \phi_2^{(1)}).$$

This is a linear functional on $L_2(\Omega)$, which we use as data to define the “transfer” error adjoint problem and derive a corresponding a posteriori error representation. For details on how to compute a quantity of interest that depends on u_1 and u_2 (so that the choice of linearizations for the coefficients in the equation for u_2 enter directly into the “primary” error contribution), see [9].

Remark 3.3. For a “lower triangular” one-way coupled system of N elliptic equations and a quantity of interest based on the N th component, we solve N total “single physics” adjoint problems and construct the error representation

$$(\psi_N, e_N) = \sum_{i=1}^N \mathcal{R}_{N-i+1}(U_{N-i+1}, \phi^{(i)}; U).$$

We then solve a sequence of adjoint problems, as the corresponding linear functional for the i th adjoint problem ($i > 1$) can be defined recursively (assuming the coupling occurs only through the right-hand side) as

$$\sum_{j=1}^{i-1} \left(\frac{Df_{N+1-j}}{Du_i} \Big|_U (e_i), \phi^{(j)} \right).$$

This extends to coupling in all of the coefficients as above.

3.2. Numerical examples. The following three numerical examples highlight the features of the analysis and the importance of accounting for the transfer error. In the following computations, we approximately solve all adjoint problems using continuous, piecewise quadratic elements in order to be able to evaluate the interpolants arising from Galerkin orthogonality. We denote these approximate adjoints solutions by Φ and use them in place of ϕ in error representation (3.12). For adaptive mesh refinement, we write the estimate as a sum of element contributions and derive a bound by introducing norms. We base the adaptive mesh refinement on the standard optimization approach using the principle of equidistribution [6] applied to the bound. We refine elements whose element contribution to the error bound is greater than half a standard deviation from the mean error contribution or refine a fixed fraction of the elements with the greatest element contributions, whichever criterion yields the greater refinement. We do not do any mesh coarsening, smoothing, or edge flips.

Example 3.1. This example demonstrates the fact that the transfer error can be significant even if the individual components u_1 and u_2 are well resolved. We consider a simple system

$$(3.14) \quad \begin{cases} -\Delta u_1 = \sin(4\pi x) \sin(\pi y), & (x, y) \in \Omega, \\ -\Delta u_2 = \mathbf{b} \cdot \nabla u_1, & (x, y) \in \Omega, \\ u = 0, & (x, y) \in \partial\Omega, \end{cases}$$

where

$$\mathbf{b} = \frac{2}{\pi} \begin{pmatrix} 25 \sin(4\pi x) \\ \sin(\pi x) \end{pmatrix}, \quad \mathbf{f}(u) = \begin{pmatrix} \sin(4\pi x) \sin(\pi y) \\ \mathbf{b} \cdot \nabla u_1 \end{pmatrix}, \quad \Omega = ([0, 1], [0, 1]).$$

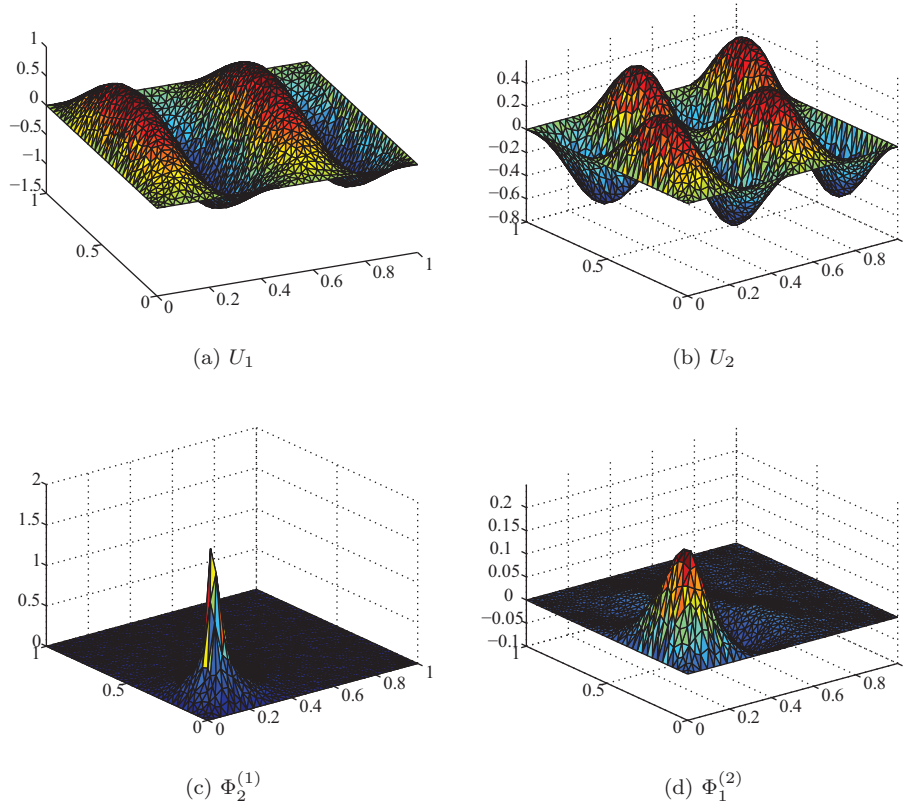


FIG. 3.1. Example 3.1. Primary and (nonzero) adjoint solutions computed on uniformly fine meshes. The adjoint solutions are largest near the region of the quantity of interest $u_2(.25, .25)$.

The quantity of interest is the solution value of u_2 at $(.25, .25)$, which we estimate using a smooth delta function approximation with localized support. The corresponding global adjoint problem is

$$(3.15) \quad \begin{cases} -\Delta\phi_1^{(1)} + Lf(u_1)\phi_2^{(1)} = 0, & (x, y) \in \Omega, \\ -\Delta\phi_2^{(1)} = \delta_{\tilde{x}}^{\text{reg}}, & (x, y) \in \Omega, \\ \phi = 0, & (x, y) \in \partial\Omega, \end{cases}$$

where $\delta_{\tilde{x}}^{\text{reg}}$ is a regularized delta function and $\tilde{x} = (.25, .25)$. Our primary adjoint problem is

$$-\Delta\phi_2^{(1)} = \delta_{\tilde{x}}^{\text{reg}}, \quad (x, y) \in \Omega, \quad \phi = 0, (x, y) \in \partial\Omega.$$

The secondary adjoint problem is

$$(3.16) \quad \begin{cases} \Delta\phi_1^{(2)} = \nabla \cdot (\mathbf{b}\phi_2^{(1)}), & (x, y) \in \Omega, \\ \phi^{(2)} = 0, & (x, y) \in \partial\Omega. \end{cases}$$

The primal system was solved using identical standard continuous piecewise linear finite element discretizations for u_1 and u_2 . We plot the results in Figure 3.1 and show

TABLE 3.1
Error contributions for Example 3.1.

Primary error	Transfer error
0.0042	0.0006

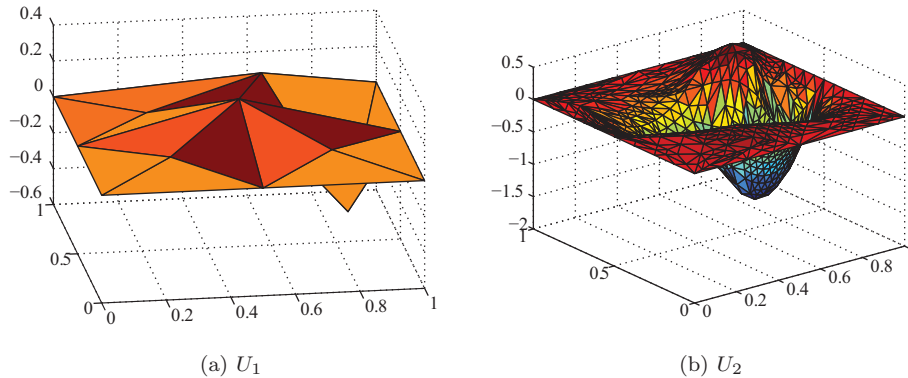


FIG. 3.2. Example 3.2. Adaptivity based on the standard discretization error estimate for the “primary” error, ignoring the “transfer” error. Only the mesh for U_2 is refined.

TABLE 3.2
Error contributions for Example 3.2.

Primary error	Transfer error
0.00005	0.110

the error contributions in Table 3.1. While the adjoint solution $\Phi_1^{(2)}$ in Figure 3.1(d) is concentrated near the location of the quantity of interest, it has nontrivial spatial structure, and the transfer error represents 14% of the total error.

Example 3.2. This example illustrates the importance of computing the transfer error, since, for this problem, simply forcing the “primary” error contribution to be small (by refining the second mesh only) does not provide *any* accuracy in the desired quantity of interest. We reconsider (3.14) but with quantity of interest equal to the average value of u_2 over the whole domain. The exact solution has zero average value on Ω . We solve both components of the primary problem on an identical coarse initial mesh, but adapt and refine only the mesh for u_2 using the traditional weighted residual, the first “primary” error term in (3.12), while neglecting the second “transfer error” term in (3.12). We show the results in Figure 3.2 and Table 3.2.

Ignoring the transfer error and the implied need to refine the first component produces a completely unsuitable adaptive procedure. It is clear from Figure 3.2 that the average value of the second component is far from zero, and the actual computational value is -0.2245 . The estimated transfer error of 0.1 is, in fact, an underestimate since $\Phi_1^{(2)}$ is based on the highly inaccurate solution U_1 , which is computed on a very coarse mesh. The transfer error dominates the computation, and this error *cannot* be reduced without refining the mesh for u_1 .

Example 3.3. The third example shows that an “optimal” adaptive mesh for the quantity of interest that depends only on u_2 may actually involve a richer discretiza-

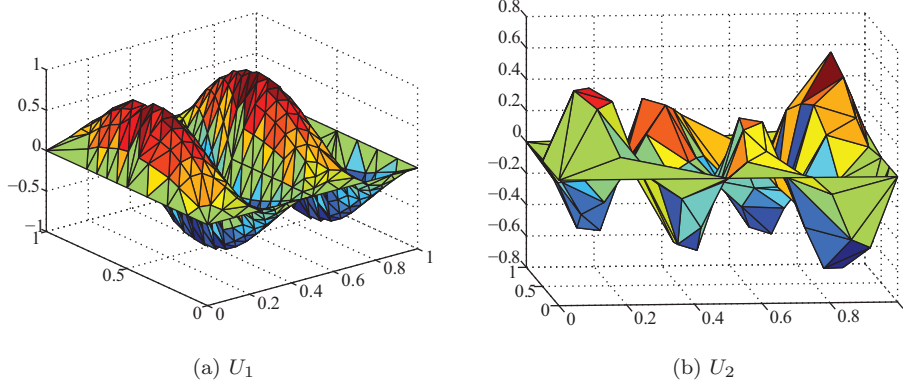


FIG. 3.3. Example 3.3. Adaptivity based on the full estimate that accounts for “primary” and “transfer” errors. The quantity of interest $U_2(.25, .25)$ is more sensitive to errors in U_1 than U_2 .

tion of u_1 than u_2 . We consider system (3.14) with the quantity of interest equal to the average value of u_2 over the whole domain and initial coarse meshes as in the previous example, but we use the transfer error contribution to adapt the mesh for u_1 and the primary contribution to adapt the mesh for u_2 so that the total error is less than 10^{-4} . The resulting meshes are shown in Figure 3.3 and illustrate that despite the fact that the quantity of interest involves only u_2 , the error inherited from u_1 is the most important contribution to consider. In this problem, the strong influence of the transfer error is a result of the dependence of u_2 on the *gradient* of u_1 , which a priori has lower order accuracy. Similar behavior could also arise when u_2 just depends on u_1 .

4. Interpolation error analysis. We use a multiscale discretization for the “fully” adaptive Example 3.3, i.e., the components u_1 and u_2 were computed on different meshes; see Figure 3.3. This raises the issue of understanding the effect of translating one component onto the mesh of the other component when performing the integration necessary to form the discrete equations. Integration involving functions defined on different meshes can cause problems because these quantities may be complicated, as illustrated in Figure 4.1.

In particular, traditional quadrature formulae based on sets of specific points may not preserve the accuracy required for effective computation because a function defined on a different mesh is generally not sufficiently smooth. For example, the

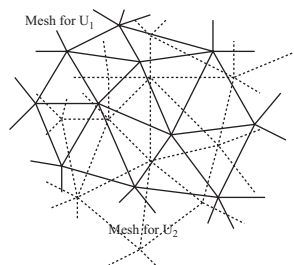


FIG. 4.1. The problem of translation between meshes. Finite element functions on one mesh are generally not smooth on another mesh.

integrand (f_2, χ) is piecewise discontinuous on every element τ_i of mesh 2 in Example 3.3, as $\mathbf{b} \cdot \nabla U_1$ is continuous only within elements of the mesh for U_1 . In general, if the meshes are not congruent, the integrand is C^0 at best. Using a “traditional” higher order quadrature rule will not necessarily lead to the expected increase in accuracy as the integrand (f_2, χ) does not have sufficient regularity. Possible solutions include either the determination of local intersections of simplices and/or hexahedra or the construction of a global union mesh. However, both solutions are computationally expensive, and the global solution often requires several times more memory than the storage of the two individual meshes, especially for three-dimensional problems.

4.1. Projections from mesh 1 to mesh 2. Instead of constructing a union mesh, we use a projection $\Pi_{1 \rightarrow 2}$ from $\mathcal{S}_{1,h}$ to $\mathcal{S}_{2,h}$ and solve the discrete system given by (3.3). This introduces additional sources of error. Starting from error representation formula (3.6), we add and subtract

$$f_2(x, \Pi_{1 \rightarrow 2} U_1, \Pi_{1 \rightarrow 2} DU_1, (I - \Pi_2) \phi_2^{(1)}),$$

yielding

$$\begin{aligned} (\psi^{(1)}, e) &= (\psi_2^{(1)}, e_2) \\ &= (f_2(x, \Pi_{1 \rightarrow 2} U_1, \Pi_{1 \rightarrow 2} DU_1), (I - \Pi_2) \phi_2^{(1)}) - \mathcal{A}_2(U_2, (I - \Pi_2) \phi_2^{(1)}) \\ &\quad + (f_2(x, u_1, Du_1) - f_2(x, \Pi_{1 \rightarrow 2} U_1, \Pi_{1 \rightarrow 2} DU_1), \phi_2^{(1)}). \end{aligned}$$

Adding and subtracting $(f_2(x, U_1, DU_1), \phi_2^{(1)})$ produces

$$\begin{aligned} (\psi^{(1)}, e) &= (f_2(x, \Pi_{1 \rightarrow 2} U_1, \Pi_{1 \rightarrow 2} DU_1), (I - \Pi_2) \phi_2^{(1)}) - \mathcal{A}_2(U_2, (I - \Pi_2) \phi_2^{(1)}) \\ &\quad + (f_2(x, u_1, Du_1) - f_2(x, U_1, DU_1), \phi_2^{(1)}) \\ &\quad + (f_2(x, U_1, DU_1) - f_2(x, \Pi_{1 \rightarrow 2} U_1, \Pi_{1 \rightarrow 2} DU_1), \phi_2^{(1)}). \end{aligned}$$

The first two terms on the right represent the primary discretization error for a functional of the the second component, the third term on the right represents transfer error (3.11), and the fourth term is a new expression that represents the error from the projection $\Pi_{1 \rightarrow 2}$. The projection error can be decomposed as

$$(4.1) \quad (\Pi_{1 \rightarrow 2} f_2(x, U_1, DU_1) - f_2(x, \Pi_{1 \rightarrow 2} U_1, \Pi_{1 \rightarrow 2} DU_1), \phi_2^{(1)}) \\ + (I - \Pi_{1 \rightarrow 2})(f_2(x, U_1, DU_1), \phi_2^{(1)}).$$

The first inner product in (4.1) can be computed (with some effort) on $\Omega_{2,h}$. However, computing the second term raises the same numerical issues that caused the adoption of the projection $\Pi_{1 \rightarrow 2}$ in the first place! We handle this term using the Monte Carlo techniques described in section 4.3.

4.2. Projections from mesh 2 to mesh 1. Complications from the use of projections also arise in computations with the solution of the secondary adjoint problem. The secondary adjoint problem domain is $\Omega_{1,h}$, but $\phi_2^{(1)}$ is computed naturally on $\Omega_{2,h}$.

The new error representation formula for the transfer error becomes

$$(Df_2(U_1) \times e_1, \Pi_{2 \rightarrow 1} \phi_2^{(1)}) + (Df_2(U_1) \times e_1, (I - \Pi_{2 \rightarrow 1}) \phi_2^{(1)}),$$

which is the error contribution arising from the transfer as well as an additional term that is large when the approximation spaces are significantly different. For example, this term is important when the original system is multiscale. The implicit $\psi^{(2)}$ for the transfer error adjoint is now

$$(f_2(u_1) - f_2(U_1), \phi_2^{(1)}) = (Df_2(U_1) \times e_1, \Pi_{2 \rightarrow 1} \phi_2^{(1)}) = (\psi_1^{(2)}, e_1).$$

The additional term $(Df_2(U_1) \times e_1, (I - \Pi_{2 \rightarrow 1}) \phi_2^{(1)})$ is a linear functional, so we may define an additional “tertiary” adjoint problem to estimate this quantity.

Projection (“tertiary”) error adjoint problem. This problem has the same form as transfer error adjoint (3.10), but with data $\psi_1^{(3)}$ satisfying

$$(\psi_1^{(3)}, e_1) = (Df_2(U_1) \times e_1, (I - \Pi_{2 \rightarrow 1}) \phi_2^{(1)}).$$

The resulting error representation formula is

$$(4.2) \quad (\psi_1^{(3)}, e_1) = (f_1, (I - \Pi_1) \phi_1^{(3)}) - \mathcal{A}_1(U_1, (I - \Pi_1) \phi_1^{(3)}) = (\mathcal{R}_1, (I - \Pi_{2 \rightarrow 1}) \phi_1^{(3)}).$$

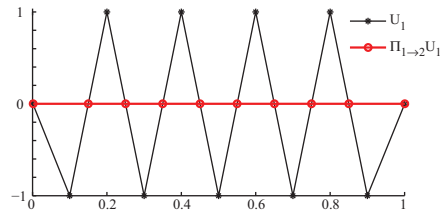
The error representation is therefore

$$(4.3) \quad (\psi_2^{(1)}, e_2) = \mathcal{R}_2(U_2, (I - \Pi_2) \phi_2^{(1)}; U_1) + \mathcal{R}_1(U_1, (I - \Pi_1)(\phi_1^{(2)} + \phi_1^{(3)})) \\ + (\Pi_{1 \rightarrow 2} f_2(U_1) - f_2(\Pi_{1 \rightarrow 2} U_1), \phi_2^{(1)}) + ((I - \Pi_{1 \rightarrow 2}) f_2(U_1), \phi_2^{(1)}).$$

Remark 4.1. Traditional simplex-based numerical integration methods that interrogate U_1 at cubature points can be thought of as projecting the integrand $f(U_1) \chi_2$ into a specific polynomial space \mathcal{P}_τ defined on each simplex τ of the mesh for U_2 and then integrating exactly. We may express this “cubature error” as a projection error and construct a corresponding error representation formula in a similar manner. Cubature error resulting from the fact that integration was not performed on a “union” mesh of two piecewise polynomial spaces may always be viewed as projection error.

Remark 4.2. In this discussion, we assume that the adjoint problems are solved using approximation spaces that are compatible with the corresponding primal approximation space, e.g., using higher order Lagrange elements on the same mesh. In practice, different meshes may be used for the primal and adjoint solves. However, this introduces new projection operators between the corresponding approximation spaces as well as the additional terms due to the loss of Galerkin orthogonality. We confine ourselves to merely alluding to the notational complexities and length of the resulting error representation.

4.3. Monte Carlo Integration. Interpolation-based projections suffer from mesh-aliasing difficulties. An extreme example is given in Figure 4.2. For a more practical example, we construct two quasi-uniform, unstructured meshes 1 and 2,

FIG. 4.2. Interpolation errors for two meshes on $\Omega = [0, 1]$.TABLE 4.1
Errors in various approximations of I_e .

$ I_e - I_1 $	$ I_1 - I_{gauss} $	$ I_1 - I_{\Pi} $	$ I_1 - I_{Samp} $
0.000187	0.000246	0.0060	0.00041

both of size h on $\Omega = [0, 1] \times [0, 1]$ and take the piecewise linear interpolant f_{I_1} of the function $f = \sin(20hx) \sin(20hy)$ on mesh 1. We first compute $I_e = \int_{\Omega} f dx$ and $I_1 = \int_{\Omega} f_{I_1} dx$ exactly and then construct three different approximations I_{gauss} , I_{Π} , and I_{Samp} as follows:

1. I_{gauss} . Using a third order, four-point quadrature rule [16] on the triangles of mesh 2 by interpolating f_{I_1} at the corresponding quadrature points.
2. I_{Π} . Projecting f_{I_1} onto mesh 2 by interpolating f_{I_1} at the nodes of mesh 2 and then using exact integration.
3. I_{Samp} . Performing the integration via a uniform weight quadrature rule using the quadrature points corresponding to the four-point quadrature rule employed by I_{gauss} .

We show the accuracy in Table 4.1. Note that the work for all three methods is roughly the same. The smallest of the projection errors $|I_1 - I_{gauss}|$ is larger than the interpolation error $|I_e - I_1|$. The error in $|I_1 - I_{\Pi}|$ is a factor of 10 larger than $|I_1 - I_{gauss}|$ and $|I_1 - I_{Samp}|$, which, for this problem, amounts to a factor of h^{-1} . Note that the four-point Gauss quadrature rule is only slightly more accurate than the sampling rule I_{Samp} .

Motivated by the example, we employ pseudorandom Monte Carlo integration using p random uniformly distributed sample points on the reference element. The main difficulty (and computational expense) when integrating on $\Omega_{2,h}$ is the evaluation of U_1 at each random sample point, since this involves locating the point in the appropriate element in $\Omega_{1,h}$. Nominally, this process requires $(O(N))$ operations per sample point, where N is the number of degrees of freedom for U_1 , hence $O(MN)$ operations for the integration, where M is the number of degrees of freedom for U_2 . However, this approach may be greatly accelerated by using a geometric implementation of the assembly and point search algorithms.

We illustrate the search algorithm in Figure 4.3. We generate a random integration point p_1^1 in $\tau_1 \in \Omega_{2,h}$ and determine the containing element of $\Omega_{1,h}$. This could potentially involve a full search of $\Omega_{1,h}$, but as this is the initial element, a good starting guess for element location could be provided as an input. Once a matching simplex is found in $\Omega_{1,h}$, the computation is performed, and the next integration point p_2^1 is generated. Moreover, the last matching simplex is stored, so the geometric search using edge/face neighbors and barycentric coordinates to guide neighbor

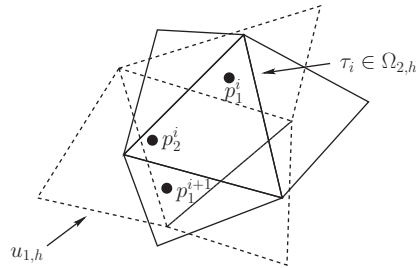


FIG. 4.3. Monte Carlo integration point search.

selection for the next point is very fast. When the integration is finished, we select the next element to be an edge/face neighbor. Now when we generate p_1^2 , we have a good starting point, namely, the last match in p_1^1 which should be “close” to the real element containing p_1^2 . The assembly routine keeps selecting edge neighbors until it has looped over all elements recursively.

This algorithm works even with a primitive data structure as long as recursion is employed. If the number of mesh elements is large, however, this may not be practical due to recursion limits. A nonrecursive algorithm could lead to termination before all element contributions for the mesh were calculated, as the next element returned by the search could have all edge/face neighbors whose element contributions had already been calculated. The algorithm would have to “restart” from an element that has not been computed. On quasi-uniform meshes with no fine scale features in the geometry, the number of “restarts” also grows logarithmically with the number of elements. Of course, with a more sophisticated data structure, either octree based or, for example, a mesh where the elements had been ordered by the use of a space-filling curve, the need for restarting would be eliminated.

When the meshes for $\Omega_{1,h}$ and $\Omega_{2,h}$ are both quasi-uniform on Ω , the number of elements tested in Ω is bounded by some h -independent constant for each integration point. Obviously, this is not the case for general adapted or anisotropic meshes, but in practice, the number of searches grows at most logarithmically with the numbers of degrees of freedom in u_1 . The convergence of this Monte Carlo integration scheme follows from standard results (see [11]) as the integrand can always be defined as the sum of integrals of continuous functions on individual simplices of the union mesh of $\Omega_{1,h}$ and $\Omega_{2,h}$.

4.4. Numerical examples. We demonstrate the significance of the projection errors with two examples.

Example 4.1. The first example illustrates how the projection error can influence a typical computation. We consider a system defined by (3.14), with two randomly generated initial meshes for u_1 and u_2 . The initial mesh for u_1 is finer than for u_2 in order to reduce the transfer error. The quantity of interest in this computation is the average value of u_2 . We show the results in Figure 4.4 and Table 4.2.

We use a local projector $\Pi_{1 \rightarrow 2, \tau}$ given by interpolation at the Gauss points (third-order three-point simplex rule) of simplices τ in $\mathcal{S}_{h,2}$. Use of this projector would integrate (U_1, U_2) exactly if the meshes were identical. The solution using this projector is given by Figure 4.4(b). This is compared against a 16-point Monte Carlo computation illustrated by Figure 4.4(c).

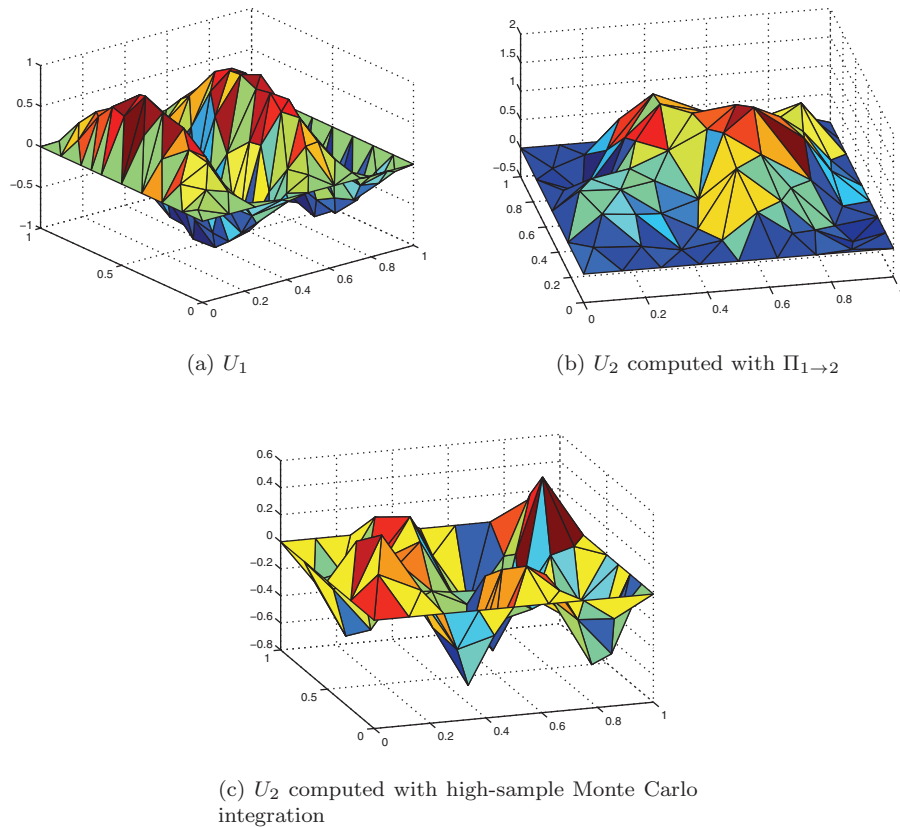


FIG. 4.4. *Example 4.1. The role of projection errors in nonaligned meshes. Note that the magnitude and oscillation of U_2 computed with $\Pi_{1 \rightarrow 2, \tau}$, shown in (b) are incorrect (a fine scale U_2 is given by Figure 3.1).*

TABLE 4.2

Example 4.1. Error contributions for computation shown in Figure 4.4.

Primary error	Transfer error	Projection error
0.003533	0.021589	0.007908

As discussed in section 4.2, projection from $\mathcal{S}_{h,2}^2$ to $\mathcal{S}_{h,1}^2$ can also lead to significant inaccuracies in computing the transfer error, necessitating the computation of tertiary adjoint problem (4.2).

Example 4.2. As discussed in section 4.2, projection from $\mathcal{S}_{h,2}^2$ to $\mathcal{S}_{h,1}^2$ can also lead to significant inaccuracies in computing the transfer error, necessitating the computation of tertiary adjoint problem (4.2). This example shows that computations with significant differences in mesh scale can contribute significantly to the error. We again use the system in Example 3.1 with the quantity of interest point value at $(.15, 15)$, starting with a coarse identical initial mesh for u_1 and u_2 but refining only the mesh for u_2 . There is no projection error as $\mathcal{S}_{h,2} \subseteq \mathcal{S}_{h,1}$. However, when we compute the transfer error, we ignore the fact that a natural choice of decomposition for the computation is integration over the simplices of $S_{h,2}$. Instead, we use the

TABLE 4.3

Example 4.2. Error contributions for the computation shown in Figure 4.5.

Primary error	Transfer error	Projection error	Tertiary error
0.000713	0.0905	0	0.0325

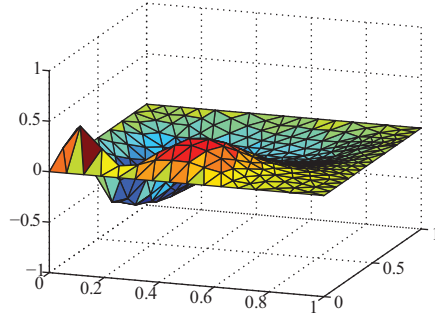


FIG. 4.5. Example 4.2. Tertiary adjoint solution $\Phi_1^{(3)}$ which estimates the projection error in computing the transfer error.

interpolation of $\phi_2^{(1)}$ at the quadrature points at the simplices of $S_{h,1}$. To compute $(I - \Pi)$, we employ the actual nesting of the two meshes to perform an accurate (up to quadrature error on the fine scale mesh) computation of $\phi_1^{(3)}$. We show the results in Table 4.3 and Figure 4.5.

5. An adaptive algorithm for the operator decomposition finite element method. Given tolerance TOL on the error in the quantity of interest, an adaptive algorithm that takes into account all the possible sources of error is given below.

```

while (the total error is less than TOL) do
  Compute  $U_1$  using standard integration.
  Compute  $U_2$  using 16-point M.C. integration for the coupling term.
  Compute  $\Phi_2^{(1)}$  using standard integration.
  Compute  $\Phi_1^{(2)}$  for given adjoint data using 16-point M.C. integration.
  if (the sum of two error contributions is greater than TOL) then
    Refine both meshes based on the primary error contributions for  $U_2$  and the
    transfer error contributions for  $U_1$ .
  else
    Compute the projection error by comparing with a 64-point M.C. integration.
    Compute  $\Phi_1^{(3)}$ .
    if (the total error is greater than TOL) then
      Refine both meshes based on the primary and projection error contributions
      for  $U_2$  and the transfer and tertiary error contributions for  $U_1$ .
    end if
  end if
end while.
    
```

The algorithm drives the primary and transfer error contributions to within a specified error tolerance and then checks for projection error by using 64-point Monte Carlo integration as an approximation to the identity operator I in (4.3) and at-

tempts to correct the projection error by refinement as well. Any projector could be substituted for the M.C. integration used in computing U_2 and $\Phi_1^{(2)}$.

We select the use of 16 sample points for the Monte Carlo integration based on our experience from a series of numerical experiments where different functions were interpolated on a quasi-uniform mesh, and then integrated. This interpolant was then integrated using Monte Carlo with 2^N sample points per simplex on a different quasi-uniform mesh (with the same approximate h); $N = 4$ gave the best tradeoff between speed and accuracy.

5.1. Examples. We describe two applications of the algorithm to one-way coupled systems using different meshes for each solution component. In both examples, we start with identical coarse initial meshes (quasi-uniform with $h \approx .125$) and adapt each mesh until both the primary and transfer error formulas are less than 10^{-4} . We control projection error using Monte Carlo integration.

Example 5.1. In the first example, we approximate the value of u_2 at $(.25, .25)$, where (u_1, u_2) solves

$$\begin{cases} -\Delta u_1 = 64\pi^2 \sin 4\pi(x - .75 + |x - .75|) \sin 4\pi(y - .75 + |y - .75|), & (x, y) \in \Omega, \\ -\Delta u_2 = u_1, & (x, y) \in \Omega, \\ u_1 = u_2 = 0, & (x, y) \in \partial\Omega, \end{cases}$$

with $\Omega = ([0, 1], [0, 1])$. The corresponding adjoint problem is

$$\begin{cases} -\Delta \phi_2^{(1)} = \delta_{\text{reg}}(x_0), & (x, y) \in \Omega, \\ \phi_2^{(1)} = 0, & (x, y) \in \partial\Omega, \end{cases}$$

with $x_0 = (.25, .25)$. The transfer error adjoint problem is

$$\begin{cases} -\Delta \phi_1^{(2)} = \phi_2^{(1)}, & (x, y) \in \Omega, \\ \phi_1^{(2)} = 0, & (x, y) \in \partial\Omega. \end{cases}$$

The accurate computation of the quantity of interest $u_2(0.25, 0.25)$ does not require the fine scale features of u_1 near $(0.75, 0.75)$ to be resolved. However, a quantity of interest equal to the value of u_2 at $(0.9, 0.9)$ near the localized features of u_1 requires better resolution of the details of u_1 . The adapted solutions U_1 for both quantities of interest are given in Figure 5.1(c) and Figure 5.1(d), respectively.

Example 5.2. We now consider an example where convection in component u_1 creates the need for refinement in u_1 remote from the the goal-oriented refinement in u_2 .

$$(5.1) \quad \begin{cases} -\Delta u_1 - \mathbf{b} \cdot \nabla u_1 = 10^3 e^{-100\|x-x_0\|^2}, & x \in \Omega, \\ -\Delta u_2 = 10^3 e^{-100\|x-x_1\|^2} u_1, & x \in \Omega, \\ u_1 = u_2 = 0, & x \in \partial\Omega, \end{cases}$$

where $\mathbf{b} = (100, 40)^\top$, $x_0 = (.75, .75)$, $x_1 = (.1, .5)$, and the quantity of interest is the point value $u_2(x_2)$, $x_2 = (.2, .5)$. The corresponding adjoint problem for the primary error contribution is

$$\begin{cases} -\Delta \phi_2^{(1)} = \delta_{\text{reg}}(x_2), & x \in \Omega, \\ \phi_2^{(1)} = 0, & x \in \partial\Omega, \end{cases}$$

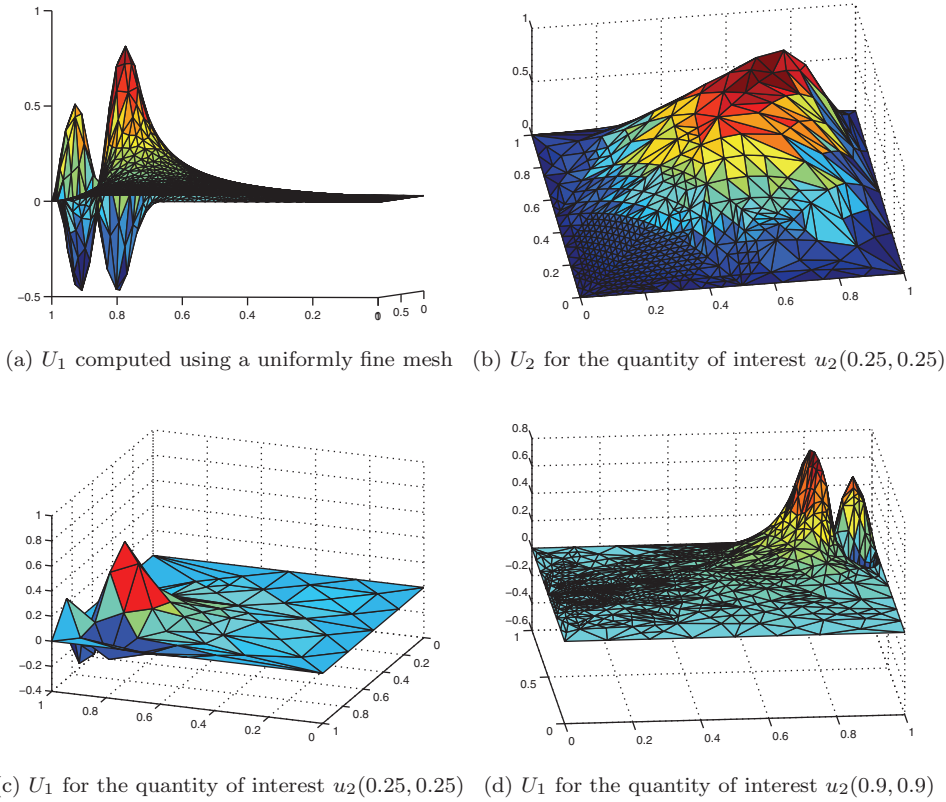


FIG. 5.1. Example 5.1. Example of computational efficiency: U_1 may be computed on a coarse discretization, yet U_2 may be determined with sufficient accuracy.

while the corresponding transfer error adjoint problem is

$$\begin{cases} -\Delta\phi_1^{(2)} + \mathbf{b} \cdot \nabla\phi_1^{(2)} = 10^3 e^{-100\|x-x_1\|^2} \phi_2^{(1)}, & x \in \Omega, \\ \phi_1^{(2)} = 0, & x \in \partial\Omega. \end{cases}$$

The adjoint solution $\Phi_1^{(2)}$ in Figure 5.2(c) shows the influence of the convection term in the equation for u_1 . When the quantity of interest is a value of u_2 in the convective region of influence of the localized source term in the equation for u_1 , the solution for u_1 is resolved “upstream” of the location of the quantity of interest as shown in Figure 5.2(a).

When the quantity of interest is a value of u_2 away from the convective region of influence of the localized source term in the equation for u_1 , the adjoint solution $\phi_1^{(2)}$ has a similar structure to that shown in Figure 5.2(c) but has much smaller magnitude. The resulting mesh for U_1 need not even be detailed enough to eliminate the numerical oscillation (from not satisfying the corresponding Péclet mesh condition). This situation is illustrated by Figure 5.2(d), where the choice of quantity of interest is u_2 at $(0.15, 0.15)$.

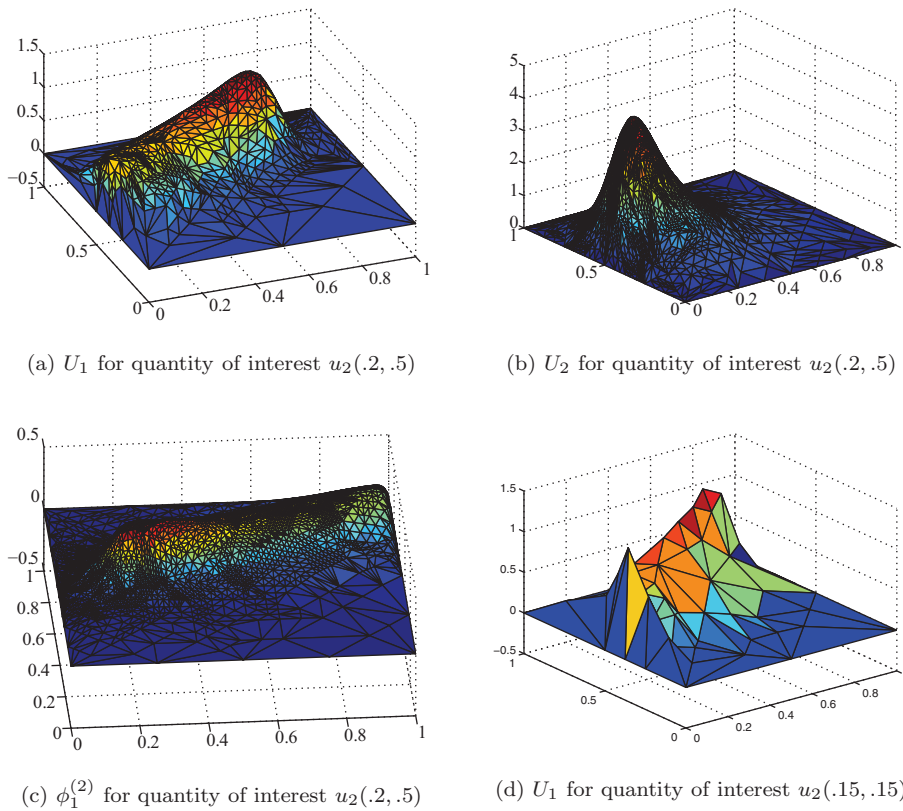


FIG. 5.2. *The role of convection in Example 5.2. Note that altering the location of the quantity of interest alters the density and location of the resulting adapted meshes (same adaptive criteria).*

6. Conclusion. In this paper, we perform an a posteriori error analysis of a multiscale operator decomposition finite element method for the solution of a system of one-way coupled elliptic problems. The analysis specifically accounts for the effects arising from multiscale operator decomposition, including the following issues: (1) Errors in the solution of each component propagate into the solutions of the other components; and (2) Transferring information between different representations potentially introduces new error. We estimate the various sources of errors by defining auxiliary adjoint problems whose data are related to errors in the information passed between components. Through a series of examples, we demonstrate the importance of accounting for the contributions to the error arising from multiscale operator decomposition. We also devise an adaptive discretization strategy based on the error estimates that specifically controls the effects arising from operator decomposition. Finally, we demonstrate the usefulness of Monte Carlo integration methods for dealing with a mismatch between discretizations of different components.

We extend this analysis to a “fully coupled” system in the form of (1.4) in part II of this paper [3]. We address the important issue that the adjoint operator associated with the fully coupled system and an operator decomposition solution are not generally equal. This difference requires additional strategies for error control. We consider the use of noninterpolatory projectors based on averaging to reduce both transfer and projection error in [4].

REFERENCES

- [1] W. BANGERTH AND R. RANNACHER, *Adaptive Finite Element Methods for Differential Equations*, Birkhauser-Verlag, New York, 2003.
- [2] R. BECKER AND R. RANNACHER, *An optimal control approach to a posteriori error estimation in finite element methods*, Acta Numer., 10 (2001), pp. 1–102.
- [3] V. CAREY, D. ESTEP, AND S. TAVENER, *A posteriori analysis and adaptive error control for operator decomposition methods for elliptic systems II: Fully coupled systems*, Internat. J. Numer. Methods Engrg., submitted.
- [4] V. CAREY, D. ESTEP, AND S. TAVENER, *Averaging based projections in operator decomposition methods for elliptic systems*, 2008, manuscript.
- [5] K. ERIKSSON, D. ESTEP, P. HANSBO, AND C. JOHNSON, *Introduction to adaptive methods for differential equations*, Acta Numer., 4 (1995), pp. 105–158.
- [6] K. ERIKSSON, D. ESTEP, P. HANSBO, AND C. JOHNSON, *Computational Differential Equations*, Cambridge University Press, Cambridge, 1996.
- [7] D. ESTEP, *A posteriori error bounds and global error control for approximation of ordinary differential equations*, SIAM J. Numer. Anal., 32 (1995), pp. 1–48.
- [8] D. ESTEP, M. HOLST, AND M. LARSON, *Generalized Green's functions and the effective domain of influence*, SIAM J. Sci. Comput., 26 (2005), pp. 1314–1339.
- [9] D. ESTEP, M. G. LARSON, AND R. D. WILLIAMS, *Estimating the error of numerical solutions of systems of reaction-diffusion equations*, Mem. Amer. Math. Soc., 146 (2000), pp. viii+109.
- [10] D. ESTEP, S. TAVENER, AND T. WILDEY, *A posteriori analysis and improved accuracy for an operator decomposition solution of a conjugate heat transfer problem*, SIAM J. Numer. Anal., 46 (2008), pp. 2068–2089.
- [11] G. S. FISHMAN, *Monte Carlo: Concepts, algorithms, and applications*, Springer Ser. Oper. Res., Springer-Verlag, New York, 1996. Concepts, algorithms, and applications.
- [12] M. GILES AND E. SÜLI, *Adjoint methods for PDEs: A posteriori error analysis and postprocessing by duality*, Acta Numer., 11 (2002), pp. 145–236.
- [13] V. GINTING, D. ESTEP, J. SHADID, AND S. TAVENER, *An a posteriori analysis of operator splitting*, SIAM J. Numer. Anal., 46 (2008), pp. 1116–1146.
- [14] G.I. MARCHUK, *On the theory of the splitting-up method*, in Proceedings of the Second Symposium on Numerical Solution of Partial Differential Equations, SVNSPADE, Academic Press, New York, 1970, pp. 469–500.
- [15] G.I. MARCHUK, *Splitting and alternating direction methods*, in Handbook of Numerical Analysis, Vol. I, P. G. Ciarlet and J. L. Lions, eds., North-Holland, New York, 1990, pp. 197–462.
- [16] G. STRANG AND G. J. FIX, *An Analysis of the Finite Element Method*, Prentice-Hall Series in Automatic Computation, Prentice-Hall, Englewood Cliffs, NJ, 1973.

UNIFORM APPROXIMATION OF PIECEWISE r -SMOOTH AND GLOBALLY CONTINUOUS FUNCTIONS*

LESZEK PLASKOTA[†] AND GRZEGORZ W. WASILKOWSKI[‡]

Abstract. We study the uniform (Chebyshev) approximation of continuous and piecewise r -smooth ($r \geq 2$) functions $f : [0, T] \rightarrow \mathbb{R}$ with a finite number of singular points. The approximation algorithms use only n function values at adaptively or nonadaptively chosen points. We construct a *nonadaptive* algorithm $\mathcal{A}_{r,n}^{\text{non}}$ that, for the functions with at most one singular point, enjoys the best possible convergence rate n^{-r} . This is in sharp contrast to results concerning discontinuous functions. For $r \geq 3$, this optimal rate of convergence holds only in the asymptotic sense, i.e., it occurs only for sufficiently large n that depends on f in a way that is practically impossible to verify. However, it is enough to modify $\mathcal{A}_{r,n}^{\text{non}}$ by using $(r+1)\lfloor(r-1)/2\rfloor$ extra function evaluations to obtain an *adaptive* algorithm $\mathcal{A}_{r,n}^{\text{ada}}$ with error satisfying $\|f - \mathcal{A}_{r,n}^{\text{ada}} f\|_C \leq C_r T^r \|f^{(r)}\|_L^\infty n^{-r}$ for all $n \geq n_0$ and n_0 independent of f . This result cannot be achieved for functions with more than just one singular point. However, the convergence rate n^{-r} can be recovered asymptotically by a *nonadaptive* algorithm $\overline{\mathcal{A}}_{r,n}^{\text{non}}$ that is a slightly modified $\mathcal{A}_{r,n}^{\text{non}}$. Specifically, $\limsup_{n \rightarrow \infty} \|f - \overline{\mathcal{A}}_{r,n}^{\text{non}} f\|_C \cdot n^r \leq C_r T^r \|f^{(r)}\|_L^\infty$ for all r -smooth functions f with finitely many singular points.

Key words. function approximation, adaptive algorithms, singular functions

AMS subject classifications. 65D05, 65Y20

DOI. 10.1137/070708937

1. Introduction. This paper deals with the approximation of scalar functions $f : [0, T] \rightarrow \mathbb{R}$ that are r -smooth except for an (unknown) *singular point* $s_f \in (0, T)$. The approximations (algorithms) are based on finitely many function evaluations, and the functions being approximated are assumed to be continuous with discontinuity starting at some derivative of order less than r . Such problems occur in a number of applications, see, e.g., [8], but the traditional algorithms, developed for nonsingular functions, do not work well. This is why there are a number of results studying singular functions and, in particular, the detection/localization of singular points. One of the approaches is based on the assumption that we have at our disposal Fourier coefficients with respect to some orthogonal basis or wavelet coefficients. See, e.g., papers [3, 4, 5, 6, 7, 9, 12, 13] and papers cited therein. The approach in the present paper is based on the assumption that only a finite (presumably small) number of function evaluations are available. In particular, the considered algorithms cannot use Fourier coefficients.

This problem was studied in [11] under the assumption that the functions or their derivatives may be discontinuous at s_f . Such functions cannot be approximated in the uniform (Chebyshev) norm with error converging to zero as the number n of function evaluations goes to infinity. It was shown that convergent algorithms exist when the errors are measured in a weaker metric such as L^p , with $1 \leq p < \infty$, or

*Received by the editors November 26, 2007; accepted for publication (in revised form) September 8, 2008; published electronically February 4, 2009.

<http://www.siam.org/journals/sinum/47-1/70893.html>

[†]Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland (leszekp@mimuw.edu.pl). This author's research was supported by the Polish Ministry of Science and Higher Education under grant 1 P03A 039 28.

[‡]Department of Computer Science, University of Kentucky, 773 Anderson Hall, Lexington, KY 40506-0046 (greg@cs.uky.edu). This author's research was partially supported by the National Science Foundation under grant DMS-0608727, and part of this work was done when he visited the University of Warsaw in the summer of 2007.

the Skorohod metric. Moreover, to obtain convergence n^{-r} it is necessary to use adaption; the best one can get from nonadaptive methods is $n^{-1/p}$ for L^p and n^{-1} for the Skorohod metric. Recall that nonadaptive methods are those with *a priori* fixed sampling points, whereas adaptive methods choose the consecutive sampling points based on already computed function values.

A similar problem was studied in [1] under the additional assumption that the functions are globally continuous and that their first-order derivatives are discontinuous at s_f . This assumption allowed the authors to prove the very surprising and nontrivial result that the error of their nonadaptive algorithm using the equispaced grid of size h is bounded by $C_r \|f^{(r)}\|_{L^\infty} h^r$ in the Chebyshev norm, however, only for *sufficiently small* h . Actually, it is an intrinsic weakness of all nonadaptive methods that, except for $r = 2$, their errors are bounded as above only for h sufficiently small, $h \leq h_0$, depending on the size of the discontinuity of the first derivative of f . Unfortunately, the estimate given in [1] requires nontrivial discontinuity of f' , i.e., h_0 tends to zero as the size of that discontinuity jump decreases. This means, in particular, that estimates of [1] are of no interest for singular functions with continuous f and f' since then h_0 would have to be zero.

The purpose of the present paper is to extend the results of [1] as well as [11]. As for the former paper, the extensions are in considering functions whose discontinuity may start at higher than the first-order derivative. For instance, f may be globally three times continuously differentiable, and only $f^{(4)}$ may be discontinuous at s_f . Moreover, we also consider adaptive algorithms since, as it will be clear later, they allow us to remove one very crucial and practically impossible to verify assumption relating the number n of sampling points to the size of the corresponding discontinuity jump. This yields positive results on the worst case errors of adaptive methods for a number of function classes. As for the latter paper, the present paper focuses on the class of globally continuous functions which is a subset of the class considered in [11]. Although positive results of this paper for adaptive algorithms could be obtained from those of [11] for the Skorohod metric, they would require some extra assumptions. In particular, the worst case results in [11] hold under the assumption that the considered functions have uniformly bounded $\|f'\|_{L^\infty}$, which is not needed now. More importantly, global continuity leads to positive results in the asymptotic setting for nonadaptive methods. Indeed, nonadaptive methods constructed in this paper have the optimal convergence rate which is proportional to n^{-r} ; recall that nonadaptive methods are superior to adaptive ones for discontinuous functions.

We now discuss the results of this paper in more detail. As already mentioned, we consider globally continuous and r -smooth functions with at most one singular point s_f . For such a class of functions, we construct a nonadaptive algorithm $\mathcal{A}_{r,n}^{\text{non}}$ that asymptotically recovers the error bound of [1]. As it was already noticed in [1], the asymptotic nature of this result is an intrinsic property of nonadaptive methods. This is no longer true if adaptive selection of samples is allowed. Indeed, a small modification of $\mathcal{A}_{r,n}^{\text{non}}$ with very few extra evaluations at adaptively chosen points yields an adaptive algorithm $\mathcal{A}_{r,n}^{\text{ada}}$ such that

$$(1) \quad \|f - \mathcal{A}_{r,n}^{\text{ada}} f\|_C \leq C_r T^r \|f^{(r)}\|_{L^\infty} n^{-r} \quad \text{for all } n \geq n_0,$$

where n_0 depends “weakly” on f ; see Theorem 1. More precisely, n_0 has to be such that the distance of s_f from the end points of the domain is not smaller than $(r - 1)$ times the step size $h = T/n_0$. Actually, this assumption is only for simplicity of presentation. We show how it can be replaced by a number of other conditions independent

of the location of s_f . One of such assumptions is that f is defined on the whole real line \mathbb{R} , an assumption adopted in [1]. This means, in particular, that the minimal *worst case error* of adaptive algorithms over the class of functions with uniformly bounded r th derivative is of order n^{-r} , while this error for nonadaptive algorithms (for $r \geq 3$) equals infinity; see Theorem 2. An exception is the class of functions with uniformly bounded first and r th derivatives; however, adaptive algorithms continue having superior worst case errors also in this class; see Theorem 3.

The algorithms $\mathcal{A}_{r,n}^{\text{non}}$ and $\mathcal{A}_{r,n}^{\text{ada}}$ are constructed by amplifying ideas developed in [1, 10, 11]. (However, we mostly use new proof techniques.) At first, divided differences of order r corresponding to the equispaced grid $ih = iT/m$ ($i = 0, \dots, m$) are used to detect and locate the singularity within an interval of length at most rh . If such an interval is not found, which happens when all of the divided differences are small enough, then the usual piecewise interpolation of degree $r - 1$ is applied. Otherwise, the interpolation is used everywhere except for the interval where the singularity has been detected. In that interval, an extrapolation from the left and from the right with a specially chosen break point is applied. The nonadaptive and adaptive algorithms differ by how the break point is chosen. It is chosen without additional function evaluations in the nonadaptive version and with at most $(r + 1)\lfloor (r - 1)/2 \rfloor$ additional evaluations in the adaptive version.

Analogous results, both positive and negative, hold when the L_p norms ($p < \infty$) are used instead of the uniform norm to measure the errors; see Theorems 4 and 5.

The results discussed so far depend very much on the fact that the functions being approximated have at most one singular point. Indeed, in section 7, we briefly consider classes of functions with a finite number of singularities and show that, instead of (1), one can only have algorithms whose worst case errors are at best proportional to n^{-1} ; this holds already for $r \geq 2$ and functions with three singular points or for $r \geq 3$ and functions with two singular points only; see Theorems 6 and 7. Fortunately, as explained in section 7.2, the optimal rate of n^{-r} can be regained in the asymptotic setting, even for functions with arbitrarily large (but finite) number of singularities. This is achieved even by nonadaptive algorithms. Indeed, a modification of $\mathcal{A}_{r,n}^{\text{non}}$ yields algorithms $\overline{\mathcal{A}}_{r,n}^{\text{non}}$ with the errors satisfying

$$\limsup_{n \rightarrow \infty} \left\| f - \overline{\mathcal{A}}_{r,n}^{\text{non}} f \right\|_C n^{-r} \leq C_r T^r \|f^{(r)}\|_{L^\infty}$$

for all globally continuous and r -smooth functions with a finite number of singular points. Here the constant C_r does not depend on f or on the number of singularities of f ; it only depends on r .

2. Preliminaries. For $r \geq 1$ and $a < b$, denote by $W_r(a, b)$ the space of r -smooth functions defined as

$$W_r(a, b) = \left\{ f \in C^{r-1}([a, b]) \mid f^{(r-1)} \text{ absolutely continuous, } \|f^{(r)}\|_{L^\infty(a,b)} < \infty \right\}.$$

Given $T > 0$, let $F_r = F_r(0, T)$ be the class of functions satisfying the following condition: either $f \in W_r(0, T)$ or there exist $s_f \in (0, T)$ and functions $f_- \in W_r(0, s_f)$ and $f_+ \in W_r(s_f, T)$ such that

$$(2) \quad f(x) = \begin{cases} f_-(x), & 0 \leq x < s_f, \\ f_+(x), & s_f \leq x \leq T. \end{cases}$$

Equivalently, $f \in F_r$ iff

$$(3) \quad f(x) = g(x) + \mathbf{1}_{[s_f, T]}(x) \sum_{j=0}^{r-1} \Delta_f^{(j)} \frac{(x - s_f)^j}{j!}, \quad 0 \leq x \leq T,$$

where $g \in W_r(0, T)$ and $\Delta_f^{(j)}$ are *discontinuity jumps* of the successive derivatives of f at the *singular point* s_f ,

$$\Delta_f^{(j)} = f_+^{(j)}(s_f) - f_-^{(j)}(s_f), \quad 0 \leq j \leq r - 1.$$

Here and elsewhere $\mathbf{1}_A$ denotes the indicator function of A :

$$\mathbf{1}_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

Note that for any $0 \leq a < b \leq T$ we have $\|f^{(r)}\|_{L^\infty(a,b)} = \|g^{(r)}\|_{L^\infty(a,b)}$.

Approximation of functions from $f \in F_r$ (with a possible discontinuity at s_f) was studied in [11]. In the present paper, we concentrate on the subclass

$$G_r = G_r(0, T) := F_r(0, T) \cap C([0, T])$$

of *continuous* functions. Note that $f \in G_r$ iff $f \in F_r$ and

$$\Delta_f^{(0)} = 0.$$

Since $G_1(0, T) = W_1(0, T)$ is not an interesting case, we also assume that the regularity

$$r \geq 2.$$

We are interested in *uniform* approximation of $f \in G_r$ that is based on finitely many evaluations of f . A *nonadaptive* method of approximation (algorithm) is given as

$$\mathcal{A}f = \varphi(f(x_1), f(x_2), \dots, f(x_n))$$

for some $x_j \in [0, T]$ and $\varphi : \mathbb{R}^n \rightarrow C([0, T])$. We will also allow a more general class of *adaptive* methods, where the points x_j and the number n of them can be chosen based on the previously obtained values $f(x_i)$ for $1 \leq i \leq j - 1$. (For more details and a general discussion of adaptive and nonadaptive methods, see, e.g., [14].)

Uniform approximation means that for any $f \in G_r$ the error of approximation is given by

$$\|f - \mathcal{A}f\|_{C([0, T])} = \max_{a \leq x \leq b} |f(x) - (\mathcal{A}f)(x)|.$$

For brevity, we write $\|\cdot\|_C$ and $\|\cdot\|_{L^\infty}$ whenever the norms pertain to the interval $[0, T]$. Although both norms are identical for functions considered in this paper, to help the reader we write $\|g\|_C$ if the function g is continuous and $\|g\|_{L^\infty}$ otherwise.

3. Generic algorithm. Denote by $\mathcal{P}_r(f; a, b)$ the polynomial of degree at most $r-1$ interpolating f at knots $a+(j-1)(b-a)/(r-1)$, $1 \leq j \leq r$. For given $h = (b-a)/m$ with $m \geq r-1$, let $\overline{\mathcal{P}}_r^h(f; a, b)$ be the piecewise polynomial interpolation of f on $[a, b]$ that is based on the mesh of size h . That is,

$$\overline{\mathcal{P}}_r^h(f; a, b) = \sum_{i=1}^k \mathcal{P}_r(f; z_{i-1}, z_i) \mathbf{1}_{[z_{i-1}, z_i]} + \mathcal{P}_r(f; b - (r-1)h, b) \mathbf{1}_{[z_k, b]},$$

where $z_i = a + ih(r-1)$ and k is the largest integer satisfying $kh(r-1) < b-a$. We also let $t_i = ih$ and denote by $f[t_i, t_{i+1}, \dots, t_{i+r}]$ the r th order divided difference of f .

Our generic algorithm $\mathcal{A}_{r,m}^*$ combines ideas developed in [1, 10, 11]. That is, it first uses divided differences of f corresponding to the grid of size h to find out whether the singularity can be ignored. If “yes,” then the piecewise interpolation of degree $r-1$ is applied based on the uniform grid. If “no,” then a subinterval (u, v) of length at most rh containing the essential singularity (if it exists) is identified. The final approximation is the piecewise polynomial interpolation, except for (u, v) . In this subinterval, polynomial extrapolation from the left on (u, ξ) and from the right on $[\xi, v)$ are applied, where ξ is a specially chosen point supposed to approximate the singularity s_f .

The parameters of the algorithm $\mathcal{A}_{r,m}^*$ are smoothness r and an arbitrary number

$$0 \leq D < \infty.$$

The parameter D plays the role of a threshold and is used to decide whether the singularity can be ignored or not. Another important ingredient of the algorithm is the function $\text{SINGULAR}(u, v)$ that returns the approximation ξ of s_f . This function can be defined in different ways; for the time being we treat it as a *black box*.

00 GENERIC ALGORITHM;

01 **begin**

02 input $m \geq 2r-1$; $h := T/m$;

03 **for** $i := 0$ **to** $m-r$ **do** $d_i := f[t_i, t_{i+1}, \dots, t_{i+r}]$;

04 $i^* := \arg \max_i |d_i|$;

05 **if** $|d_{i^*}| \leq D$ **then** $(\mathcal{A}_{r,m}^* f)(x) := \overline{\mathcal{P}}_r^h(f; 0, T)$ **else**

06 **begin**

07 $i := \max(i^*, r-1)$; $j := \min(i^* + r, m-r+1)$;

08 $u := t_i$; $v := t_j$; $u_1 := t_{i-(r-1)}$; $v_1 := t_{j+(r-1)}$;

09 $p_- := \mathcal{P}_r(f; u_1, u)$; $p_+ := \mathcal{P}_r(f; v, v_1)$;

10 $\xi := \text{SINGULAR}(u, v)$; $\kappa := (p_+(\xi) - p_-(\xi))/2$;

$$11 \quad (\mathcal{A}_{r,m}^* f)(x) := \begin{cases} \overline{\mathcal{P}}_r^h(f; 0, u_1) & 0 \leq x \leq u_1 \\ p_-(x) + \kappa(x - u_1)/(\xi - u_1), & u_1 < x < \xi \\ p_+(x) - \kappa(x - v_1)/(\xi - v_1), & \xi \leq x < v_1 \\ \overline{\mathcal{P}}_r^h(f; v_1, T) & v_1 \leq x \leq T \end{cases};$$

12 **end**

13 **end.**

Note that $\mathcal{A}_{r,m}^* f$ is a well defined continuous function. The final approximation uses only function values from the uniform (nonadaptive) grid. However, the whole method of approximation is nonadaptive only if the strategy of choosing ξ is also nonadaptive.

4. Error analysis. In this section, we analyze the error of the Generic Algorithm. The analysis is done under the assumption that the mesh size h is small enough so that all of the divided differences around s_f are well defined. Specifically, we assume that

$$(4) \quad h = \frac{T}{m} \leq \delta_0(f), \quad \text{where} \quad \delta_0(f) := \frac{1}{r-1} \min(s_f, T - s_f).$$

Note that without (4) (or without some other assumptions, see, e.g., Theorems 2 and 3), we would not be able to obtain nontrivial upper bounds, as shown by the following example.

Example 1. Consider an arbitrary algorithm \mathcal{A} and a positive number A . Let x_i , $1 \leq i \leq n$, be the sampling points used by \mathcal{A} that correspond to the following function values: $f(x_i) = A$ if $x_i \neq 0$ and $f(x_i) = 0$ when $x_i = 0$. Let τ be the smallest positive point among the x_i 's. For any $a > 0$, define f_a by $f_a(x) = Ax/a$ for $0 \leq x \leq a$ and $f_a(x) = A$ for $a < x \leq T$. Obviously, for all $a \leq \tau$, $f_a \in F_r$ and $f_a(x_i) = A$ when $x_i \neq 0$. Hence the algorithm \mathcal{A} does not distinguish between f_τ and f_ε for any $0 < \varepsilon < \tau$. Since $\|f_\tau - f_\varepsilon\|_C = A(1 - \varepsilon/\tau)$ and ε can be arbitrarily small, the error of \mathcal{A} is at least $A/2$, independently of how many function evaluations are performed.

Define

$$D_f := \max \left(D, \frac{\|f^{(r)}\|_{L^\infty}}{r!} \right).$$

We first consider the case when all of the divided differences satisfy

$$(5) \quad |f[t_i, t_{i+1}, \dots, t_{i+r}]| \leq D_f, \quad 0 \leq i \leq m - r.$$

It turns out that if condition (5) is satisfied, then the singularity can be ignored. In this case, the error of piecewise polynomial interpolation is bounded similarly as for functions with no singularities. This fact was already noticed in [11] (see also [10]) for L^p norms. Here we use different arguments from those in [11].

Define

$$(6) \quad c_{r,k} := (k - r + 1)(k - r + 2) \dots (k - 1)k.$$

Letting

$$l_s(t) := \prod_{s \neq k=0}^{r-1} \frac{t - k}{s - k}$$

be the s th Lagrange polynomial, we also define

$$(7) \quad \Omega_r := \sum_{s=0}^{r-1} |l_s(r)| = 2^r - 1.$$

PROPOSITION 1. *Let $f \in G_r$. If the condition (5) is satisfied and $h \leq \delta_0(f)$, then*

$$(8) \quad \|f - \mathcal{A}_{r,m}^* f\|_C \leq C_r D_f h^r,$$

where

$$C_r = 2(r! + c_{r,3r-2} \Omega_r).$$

Proof. If the polynomial interpolation is applied on an interval $(a, b) = (t_i, t_{i+r-1})$, then by Lemma 2 of the appendix with $\Delta_f^{(0)} = 0$ and $B = D_f$ we have

$$\|f - \mathcal{P}_r(f; a, b)\|_{C([a,b])} \leq \left(\|f^{(r)}\|_{L^\infty} + c_{r,2r-2}\Omega_r D_f \right) h^r.$$

This already implies (8) in the case $D_f = D \geq \|f^{(r)}\|_{L^\infty}/r!$ since then $\mathcal{A}_{r,m}^* f = \overline{\mathcal{P}}_r(f; 0, T)$.

However, if $D_f = \|f^{(r)}\|_{L^\infty}/r! > D$, then we have to consider, in addition, the case when an interval (u, v) in line 08 of the Generic Algorithm is created and the extrapolation is applied. Then, using again Lemma 2, for any $u_1 \leq x \leq v_1$ we have

$$|f(x) - p_\pm(x)| \leq \left(\|f^{(r)}\|_{L^\infty} + c_{r,3r-2}\Omega_r D_f \right) h^r.$$

Since $\kappa = [(f(\xi) - p_-(\xi)) - (f(\xi) - p_+(\xi))]/2$, we have the same upper bound for $|\kappa|$. Hence

$$\begin{aligned} \|f - \mathcal{A}_{r,m}^* f\|_{C([u_1,\xi])} &\leq \|f - \mathcal{P}_r(f; u_1, u)\|_{C([u_1,\xi])} + |\kappa| \\ &\leq 2(1 + c_{r,3r-2}\Omega_r/r!) \|f^{(r)}\|_{L^\infty} h^r \end{aligned}$$

and similarly for the interval $[\xi, v_1]$. This completes the proof. \square

In summary, if (5) holds, then the approximation error is small regardless of the choice of $\xi \in (u, v)$. The strategy of choosing ξ is crucial only when (5) does not hold. In what follows, we consider one nonadaptive and one adaptive strategy.

4.1. Nonadaptive strategy. Consider first a nonadaptive version $\mathcal{A}_{r,m}^{\text{non}}$ of our algorithm in which no additional function values are used to determine ξ . Specifically, we set

$$(9) \quad \xi := \arg \min_{u \leq x \leq v} |p_+(x) - p_-(x)|,$$

where p_+ and p_- are interpolation polynomials defined in line 09 of the Generic Algorithm. Note that $\mathcal{A}_{r,m}^{\text{non}}$ is then nonadaptive and uses $n = m + 1$ function evaluations.

For given $f \in G_r$, let

$$\mathcal{T}_f(t) := \sum_{j=1}^{r-1} \Delta_f^{(j)} \frac{t^j}{j!}$$

and

$$(10) \quad \omega_0(f) := \sup\{\omega > 0 \mid \mathcal{T}_f(t) \text{ is monotone on } [-r\omega, 0] \text{ and } [0, r\omega]\}.$$

We also define

$$\beta_r := \frac{1}{r!} \prod_{j=0}^{r-1} (2r - 1 - j) = \binom{2r - 1}{r - 1}.$$

Note that by the error formula of Lagrange interpolation/extrapolation we have

$$\|g - \mathcal{P}_r(g; t_i, t_{i+r-1})\|_{C([t_i, t_{i+2r-1}])} \leq \beta_r \|g^{(r)}\|_{L^\infty} h^r \quad \text{for all } g \in W_r.$$

PROPOSITION 2. Let $f \in G_r$. If the singularity s_f is essential, i.e., (5) does not hold, and

$$h \leq \min(\delta_0(f), \omega_0(f)),$$

then

$$\|f - \mathcal{A}_{r,m}^{\text{non}} f\|_C \leq 3.5 \beta_r \|f^{(r)}\|_{L^\infty} h^r.$$

Proof. Suppose for a moment that the singular point s_f is not in the selected interval (u, v) . Then all of the divided differences $f[t_i, \dots, t_{i+r}]$ with $s_f \in [t_i, t_{i+r}]$ have their absolute values not greater than $|d_{i^*}|$. On the other hand, $|d_{i^*}| \leq \|f^{(r)}\|_{L^\infty}/r!$ since f is r -smooth in (u_1, v_1) . Hence Proposition 1 holds with D_f replaced by $\|f^{(r)}\|_{L^\infty}/r!$.

Consider now the case of $s_f \in (u, v)$. It suffices to consider the error in (u_1, v_1) . Assume without loss of generality that $u < s_f \leq \xi \leq v$ (the other case is symmetric), and denote for brevity

$$e_h := \beta_r \|f^{(r)}\|_{L^\infty} h^r.$$

Observe first that

$$\begin{aligned} |\kappa| &\leq \frac{1}{2} |p_+(s_f) - p_-(s_f)| \\ &\leq \frac{1}{2} (|f(s_f) - p_-(s_f)| + |f(s_f) - p_+(s_f)|) \\ &\leq \frac{1}{2r!} \max_{r-1 \leq x \leq 2r-1} \left(\prod_{j=0}^{r-1} |x-j| + \prod_{j=2r-1}^{3r-2} |x-j| \right) \|f^{(r)}\|_{L^\infty} h^r \\ &\leq \frac{1}{2} e_h, \end{aligned}$$

with the last inequality due to the fact that the second derivative of the function $f(x) = \prod_{j=0}^{r-1} (x-j) + \prod_{j=2r-1}^{3r-2} (j-x)$ is nonnegative for $x \in [r-1, 2r-1]$, and hence f attains the above maximum at $x = r-1$ and/or $x = 2r-1$. This implies

$$|f(x) - (\mathcal{A}_{r,m}^{\text{non}} f)(x)| \leq |f(x) - p_-(x)| + |\kappa| \leq \frac{3}{2} e_h \quad \text{for } u_1 \leq x \leq s_f.$$

The same bound obviously holds for $\xi \leq x \leq v_1$.

Consider the remaining case $s_f < x < \xi$. For such x ,

$$f(x) = g(x) + \mathcal{T}_f(x - s_f),$$

where g is as in (3). This and monotonicity of \mathcal{T}_f yield

$$\begin{aligned} |f(x) - (\mathcal{A}_{r,m}^{\text{non}} f)(x)| &\leq |f(x) - p_-(x)| + |\kappa| \\ &\leq |\mathcal{T}_f(x - s_f)| + |g(x) - p_-(x)| + |\kappa| \\ &\leq |\mathcal{T}_f(\xi - s_f)| + \frac{3}{2} e_h. \end{aligned}$$

To estimate $|\mathcal{T}_f(\xi - s_f)|$ observe that

$$p_+(x) - p_-(x) = \mathcal{T}_f(x - s_f) + (g(x) - p_-(x)) - (f(x) - p_+(x)).$$

Using this with $x = \xi$ we finally obtain

$$|\mathcal{T}_f(\xi - s_f)| \leq 2|\kappa| + |g(\xi) - p_-(\xi)| + |f(\xi) - p_+(\xi)| \leq 2e_h$$

as claimed. \square

In the case $r = 2$, for any f the function \mathcal{T}_f is linear. Hence $\omega_0(f) = \infty$, and the assumption $h \leq \omega_0(f)$ can be dropped. It turns out that using a slightly different estimation we can also reduce the constant $3.5\beta_2 = 10.5$ at $\|f^{(2)}\|_{L^\infty} h^2$ in the error formula of Proposition 2.

PROPOSITION 3. *Let $f \in G_2$. If the singularity s_f is essential, i.e., (5) does not hold, and $h \leq \delta_0(f)$, then*

$$\|f - \mathcal{A}_{2,m}^{\text{non}} f\|_C \leq 7.5 \|f^{(2)}\|_{L^\infty} h^2.$$

Proof. Since $r = 2$, the polynomial $p_+ - p_-$ is of degree at most 1. Hence

$$|p_+(x) - p_-(x)| \leq |p_+(s_f) - p_-(s_f)| \leq e_h,$$

where now

$$e_h = \beta_2 \|f''\|_C h^2 = 3 \|f''\|_C h^2.$$

Hence

$$\begin{aligned} |f(x) - (\mathcal{A}_{2,m}^{\text{non}} f)(x)| &\leq |f(x) - p_-(x)| + |\kappa| \\ &\leq |f(x) - p_+(x)| + |p_+(x) - p_-(x)| + |\kappa| \\ &\leq 2.5 e_h. \end{aligned}$$

This completes the proof. \square

The assumption $h \leq \omega_0(f)$ cannot be omitted when $r \geq 3$. This is shown by the following example, which is a small modification of the corresponding example from [1]. We present it here for completeness.

Example 2. Consider a nonadaptive algorithm that uses arbitrary n points x_1, \dots, x_n . Then one can find an interval $(a, b) \subset (0, T)$ such that $x_j \notin (a, b)$ for all j , and $|b - a| \geq T/(n + 1)$. Let $g(x) = (x - a)(x - b)$ and $f_a = cg\mathbf{1}_{(-\infty, a)}$, $f_b = cg\mathbf{1}_{(-\infty, b)}$, where $c \in \mathbb{R}$ is arbitrary. For $r \geq 3$ we have $\|f_a^{(r)}\|_C = \|f_b^{(r)}\|_C = 0$. Since f_a, f_b share the same information and

$$\|f_b - f_a\|_C = \frac{|c|(b - a)^2}{4} \geq \frac{|c|T^2}{4(n + 1)^2},$$

the error for f_a or for f_b is at least $|c|T^2(n + 1)^{-2}/8$. Note also that $\omega_0(f_a) = \omega_0(f_b) = (b - a)/(2r)$.

4.2. Adaptive strategy. A closer inspection of the proof of Proposition 3 shows that assumption (4) is not needed for the error to be bounded by $2.5\beta_r \|f^{(r)}\|_{L^\infty} h^r$ for arbitrary $r \geq 2$, provided that in the interval $[\min(s_f, \xi), \max(s_f, \xi)]$ the polynomial $|p_+(x) - p_-(x)|$ takes its maximum at $x = s_f$. As shown in Example 2, any nonadaptive strategy of choosing ξ gives no guarantee that this condition is satisfied. However, it turns out that it is possible to force that condition (up to a constant arbitrarily close to 1) by choosing ξ based on a few additional adaptive function evaluations.

Specifically, let $\gamma > 0$. First we construct a set S of points as follows.

```

00 ADAPTIVE POINTS;
01 begin
02    $S := \{t_i \mid i^* \leq i \leq i^* + r\}$ ;
03    $p := p_+ - p_-$ ;  $P := (u, v)$ ;
04   while  $\langle$  there exists a local max of  $|p|$  in  $P$   $\rangle$  do
05     begin
06        $x := \langle$  largest local max of  $|p|$  in  $P$   $\rangle$ ;
07        $(x_-, x_+) := \langle$  an interval in  $P$  such that  $x \in (x_-, x_+)$  and
08          $|p(x)| \leq (1 + \gamma)|p(t)$  for all  $t \in [x_-, x_+]$   $\rangle$ ;
09        $S := S \cup \langle$   $(r + 1)$  arbitrary different points in  $[x_-, x_+]$   $\rangle$ ;
10        $P := P \setminus [x_-, x_+]$ 
11     end
12 end;

```

Suppose that $S = \{\tau_i\}_{j=0}^k$ with $t_{i^*} = \tau_0 < \tau_1 < \dots < \tau_k = t_{i^*+r}$. Then, similarly as in the Generic Algorithm, we select

$$i^{**} := \arg \max_{0 \leq i \leq k-r} |f[\tau_i, \tau_{i+1}, \dots, \tau_{i+r}]|$$

and define $u^* := \max(u, \tau_{i^{**}})$, $v^* := \min(\tau_{i^{**}+r}, v)$. Finally,

$$\xi := \arg \min_{u^* \leq x \leq v^*} |p_+(x) - p_-(x)|.$$

This adaptive version of the Generic Algorithm will be denoted by $\mathcal{A}_{r,m}^{\text{ada}}$. Since a polynomial of degree $r - 1$ has at most $\lfloor (r - 1)/2 \rfloor$ local maxima, $\mathcal{A}_{r,m}^{\text{ada}}$ uses no more than

$$(r + 1) \left\lfloor \frac{r - 1}{2} \right\rfloor$$

adaptive function evaluations, in addition to $m + 1$ nonadaptive points from the initial uniform grid. We also want to stress that for $r = 2$ no adaptive points are constructed, and therefore $\mathcal{A}_{2,m}^{\text{ada}} = \mathcal{A}_{2,m}^{\text{non}}$.

PROPOSITION 4. *Let $f \in G_r$. If the singularity s_f is essential, i.e., (5) does not hold, and $h \leq \delta_0(f)$, then*

$$\|f - \mathcal{A}_{r,m}^{\text{ada}} f\|_C \leq (2.5 + \gamma) \beta_r \|f^{(r)}\|_{L^\infty} h^r.$$

Proof. Observe first that in view of Lemma 1 of the appendix we have

$$|f[\tau_{i^{**}}, \dots, \tau_{i^{**}+r}]| \geq |f[t_{i^*}, \dots, t_{i^*+r}]| > D_f,$$

which means that $s_f \in (u^*, v^*)$. Note also that $[u^*, v^*]$ contains exactly $r + 1$ points from S .

As noticed earlier, it suffices to show that for $s_f < x < \xi$ we have

$$|p(x)| \leq (1 + \gamma) |p(s_f)|, \quad p := p_+ - p_-.$$

Indeed, if there are arguments in (s_f, ξ) for which $|p|$ is larger than $|p(s_f)|$, then there are some local maxima of $|p|$ in (s_f, ξ) . Let x^* be the largest such maximum. Then we have two cases: either x^* was one of the local maxima considered in line 06 of Adaptive Points or not.

TABLE 1
Results for $z_1 = \pi$, $z_2 = \pi + 1.0$.

m	$\mathcal{P}_{2,m}^{\text{non}}$	$\mathcal{A}_{2,m}^{\text{non}}$	$\mathcal{A}_{2,m}^{\text{ada}}$	$\mathcal{P}_{4,m}^{\text{non}}$	$\mathcal{A}_{4,m}^{\text{non}}$	$\mathcal{A}_{4,m}^{\text{ada}}$
10^2	1.5898	2.1156	2.1156	1.8338	infinity	infinity
10^3	2.8695	3.8217	3.8217	2.9336	infinity	infinity
10^4	3.6170	5.4717	5.4717	3.7438	infinity	infinity
10^5	5.1670	7.6526	7.6526	4.8494	infinity	infinity
10^6	5.7100	9.8947	9.8947	5.8127	infinity	infinity
10^7	6.6451	11.5004	11.5004	6.7849	infinity	infinity
10^8	7.6043	13.4427	13.4427	7.8385	infinity	infinity

TABLE 2
Results for $z_1 = \pi$, $z_2 = \pi + 10^{-3}$.

m	$\mathcal{P}_{2,m}^{\text{non}}$	$\mathcal{A}_{2,m}^{\text{non}}$	$\mathcal{A}_{2,m}^{\text{ada}}$	$\mathcal{P}_{4,m}^{\text{non}}$	$\mathcal{A}_{4,m}^{\text{non}}$	$\mathcal{A}_{4,m}^{\text{ada}}$
10^2	2.6020	-0.9553	-0.9553	3.1310	6.6020	infinity
10^3	4.6020	-1.0426	-1.0426	4.9480	6.6020	infinity
10^4	6.4688	5.9212	5.9212	6.7860	6.6020	infinity
10^5	8.1638	7.7640	7.7640	7.8249	infinity	infinity
10^6	8.7069	9.9052	9.9052	8.8139	infinity	infinity
10^7	9.6449	11.5020	11.5020	9.7847	infinity	infinity
10^8	10.6042	13.4429	13.4429	10.8385	infinity	infinity

In the first case, the interval $[x_-, x_+]$ chosen in line 07 of Adaptive Points had to contain ξ or s_f since, otherwise, there would be more than $r + 1$ points in $[u^*, v^*]$. Since $|p(\xi)| \leq |p(s_f)|$, then $|p(x^*)| \leq (1 + \gamma)|p(s_f)|$.

In the other case, there was another local maximum x^{**} outside of (s_f, ξ) for which the corresponding interval $[x_-, x_+]$ contained x^* . Then either ξ or s_f was in $[x_-, x_+]$. This implies that $|p(x^*)| \leq |p(x^{**})| \leq (1 + \gamma)|p(s_f)|$, as claimed. \square

We end this section with two simple numerical tests showing the described algorithms in action. The tests are motivated by Example 2. We want to approximate $f : [0, 10] \rightarrow \mathbb{R}$ defined as

$$f(x) = \begin{cases} 0, & 0 \leq x \leq z_2, \\ (x - z_1)(x - z_2), & z_2 < x \leq 10, \end{cases}$$

where $z_1 = \pi$, and $z_2 = \pi + 1.0$ in the first test and $z_2 = \pi + 10^{-3}$ in the second test. The results are presented, correspondingly, in Tables 1 and 2. For $r = 2, 4$ and for each algorithm $\mathcal{P}_{r,m}^{\text{non}}$ (which is the piecewise polynomial interpolation of degree $r - 1$ based on equispaced grid), $\mathcal{A}_{r,m}^{\text{non}}$, and $\mathcal{A}_{r,m}^{\text{ada}}$, the errors are given in the logarithmic scale, i.e., $-\log_{10} \|f - \mathcal{A}f\|_{C([0,10])}$. Exact approximations are marked with “infinity.”

Observe that the results essentially depend on the distance between z_1 and z_2 . Indeed, since for our test function $z_2 - z_1 = \Delta_f^{(1)}$, the jump of f' is relatively big in the first test and relatively small in the second test. Therefore the asymptotic superiority of $\mathcal{A}_{r,m}^{\text{non}}$ over $\mathcal{P}_{r,m}^{\text{non}}$ for $r = 2, 4$ appears very quickly in Table 1 and rather late in Table 2.

The distance $z_2 - z_1$ is also crucial for comparison of $\mathcal{A}_{4,m}^{\text{non}}$ and $\mathcal{A}_{4,m}^{\text{ada}}$. (Recall that both algorithms coincide for $r = 2$.) Indeed, since $p(x) := p_+(x) - p_-(x) = (x - z_1)(x - z_2)$, the nonadaptive algorithm wrongly chooses z_1 as the approximation to $s_f = z_2$ as long as the resolution is not smaller than $z_2 - z_1$. On the other hand, the adaptive algorithm recognizes the right point at the cost of 5 extra adaptive function evaluations.

5. Worst case results. Combining Propositions 1, 2, 3, and 4 we obtain the following theorem.

THEOREM 1. (a) *There exist C'_r such that for $r \geq 3$*

$$\|f - \mathcal{A}_{r,m}^{\text{non}} f\|_C \leq C'_r D_f m^{-r} \quad \forall f \in G_r \quad \forall m \geq \frac{T}{\min(\delta_0(f), \omega_0(f))}$$

and for $r = 2$

$$\|f - \mathcal{A}_{2,m}^{\text{non}} f\|_C \leq C'_2 D_f m^{-2} \quad \forall f \in G_2 \quad \forall m \geq \frac{T}{\delta_0(f)}.$$

(b) *There exist C''_r such that for $r \geq 3$*

$$\|f - \mathcal{A}_{r,m}^{\text{ada}} f\|_C \leq C''_r D_f m^{-r} \quad \forall f \in G_r \quad \forall m \geq \frac{T}{\delta_0(f)}.$$

This should be compared with the quality of (nonadaptive) piecewise polynomial interpolation $\overline{\mathcal{P}}_r^h(f; 0, T)$ of functions $f \in W_r(0, T)$ (no singularities), where the error is upper bounded by $\hat{C}_r \|f^{(r)}\|_{L^\infty} m^{-r}$ for all $m \geq r-1$. Unfortunately, Examples 1 and 2 prove that such *unconditional* bounds are not available in G_r , i.e., the convergence rate m^{-r} is only asymptotic for any $f \in G_r$. This has serious consequences for the *worst case setting*.

Recall that the worst case error of an algorithm \mathcal{A} with respect to a class \mathcal{F} is defined as

$$e_\infty(\mathcal{F}; \mathcal{A}) := \sup_{f \in \mathcal{F}} \|f - \mathcal{A}f\|_C.$$

Let $e_\infty^{\text{non}}(\mathcal{F}; n)$ and $e_\infty^{\text{ada}}(\mathcal{F}; n)$ be the minimal (or the infima of) worst case errors that can be achieved by, respectively, nonadaptive and adaptive algorithms using no more than n function evaluations. Obviously, $e_\infty^{\text{ada}}(\mathcal{F}; n) \leq e_\infty^{\text{non}}(\mathcal{F}; n)$.

Let $0 < L_r < \infty$. From what we said it follows that for the class

$$\mathcal{G}_r^0 := \left\{ f \in W_r(0, T) \mid \|f^{(r)}\|_{L^\infty} \leq L_r \right\}$$

we have $e_\infty^{\text{non}}(\mathcal{G}_r^0; n) \asymp n^{-r}$. However, in the presence of singularities, i.e., for

$$\mathcal{G}_r := \left\{ f \in G_r(0, T) \mid \|f^{(r)}\|_{L^\infty} \leq L_r \right\},$$

we have

$$(11) \quad e_\infty^{\text{non}}(\mathcal{G}_r; n) = e_\infty^{\text{ada}}(\mathcal{G}_r; n) = \infty.$$

This negative result can be improved by narrowing down the function class. Consider

$$\begin{aligned} \mathcal{G}_r^a &:= \{ f \in \mathcal{G}_r \mid \delta_0(f) \geq \delta \} \quad (\delta > 0), \\ \mathcal{G}_r^b &:= \left\{ f \in \mathcal{G}_r \mid f^{(j)}(0) = f^{(j)}(T), 0 \leq j \leq r-1 \right\}, \end{aligned}$$

and

$$\begin{aligned} \mathcal{G}_r^c &:= \left\{ f : \mathbb{R} \rightarrow \mathbb{R} \mid f|_{[0, T]} \in \mathcal{G}_r, \|f^{(r)}\|_{L^\infty(\mathbb{R})} \leq L_r, \right. \\ &\quad \left. f|_{(-\infty, s_f)} \in W_r(-\infty, s_f), f|_{(s_f, \infty)} \in W_r(s_f, \infty) \right\}. \end{aligned}$$

THEOREM 2. Let \mathcal{G}_r^* be one of the function classes \mathcal{G}_r^a , \mathcal{G}_r^b , or \mathcal{G}_r^c . Then

$$\begin{aligned} e_\infty^{\text{non}}(\mathcal{G}_2^*; n) &\asymp n^{-2}, \\ e_\infty^{\text{non}}(\mathcal{G}_r^*; n) &= \infty, \quad r \geq 3, \end{aligned}$$

and

$$e_\infty^{\text{ada}}(\mathcal{G}_r^*; n) \asymp n^{-r}, \quad r \geq 2.$$

Hence if smoothness $r \geq 3$, then adaptive algorithms are much better than nonadaptive algorithms for the class \mathcal{G}_r^* .

Proof. Additional assumptions on f remove the condition $m \geq T/\delta_0(f)$ since now it is possible to evaluate divided differences in the vicinity of s_f , independently of its location. In particular, functions $f \in \mathcal{G}_r^b$ can be treated as defined on \mathbb{R} and such that all $f^{(j)}$, $0 \leq j \leq r-1$, are T -periodic. (Note that Example 1 does not work anymore.) The negative results for nonadaptive algorithms remain valid because Example 2 can be constructed also for \mathcal{G}_r^* . \square

The problem with condition $m \geq T/\delta_0(f)$ can also be removed by assuming uniform boundedness of the first derivative. That is, consider the class

$$\mathcal{G}_r^d := \left\{ f \in \mathcal{G}_r \mid \|f'\|_{L^\infty} \leq L_1, \|f^{(r)}\|_{L^\infty} \leq L_r \right\} \subset \mathcal{G}_r.$$

THEOREM 3. Let $0 < L_1 < \infty$ and $0 \leq L_r < \infty$ (with positive L_r for $r = 2$). We have

$$\begin{aligned} e_\infty^{\text{non}}(\mathcal{G}_2^d; n) &\asymp n^{-2}, \\ e_\infty^{\text{non}}(\mathcal{G}_r^d; n) &\asymp n^{-2}, \quad r \geq 3, \end{aligned}$$

and

$$e_\infty^{\text{ada}}(\mathcal{G}_r^d; n) \asymp n^{-r}, \quad r \geq 2.$$

Hence if smoothness $r \geq 3$, then adaptive algorithms are much better than nonadaptive algorithms for the class \mathcal{G}_r^d .

Proof. To obtain the lower bounds for nonadaptive algorithms we construct, as in Example 2, an interval (a, b) and functions $f_a = cg\mathbf{1}_{(-\infty, a)}$, $f_b = cg\mathbf{1}_{(-\infty, b)}$ that share the same information and $\|f_a - f_b\|_C \geq |c|T^2(n+1)^{-2}/4$. Taking $c = L_1/(2T)$ for $r \geq 3$ and $c = \min(L_1/(2T), L_2/2)$ for $r = 2$, we have that $f_a, f_b \in \mathcal{G}_r^d$. Hence for one of these functions the error of approximation is at least

$$\frac{\min(L_1T, L_2T^2)}{16(n+1)^2}$$

as claimed.

To meet the upper bounds, we can use correspondingly $\mathcal{A}_{2,m}^{\text{non}}$ and $\mathcal{A}_{r,m}^{\text{ada}}$ with slightly modified initial (nonadaptive) sampling for x close to the boundary of $[0, T]$. That sampling is described in detail in [11, section 5.1] and relies on using higher and higher resolution as the end points are approached, finally reaching resolution τ of order m^{-r} . Then piecewise linear interpolation on intervals $[0, (r-1)\tau]$ and $[T - (r-1)\tau, T]$ gives error of order m^{-r} . \square

6. (Remarks on) L^p approximation. In this (and only this) section we assume that, instead of the uniform norm, the L^p norm with $1 \leq p < \infty$ is used to measure the error of approximation. That is, for any f the error of an algorithm \mathcal{A} is given as

$$\|f - \mathcal{A}f\|_{L^p} = \left(\int_0^T |f(x) - (\mathcal{A}f)(x)|^p dx \right)^{1/p}.$$

We denote by $e_p^{\text{non}}(\mathcal{F}; n)$ and $e_p^{\text{ada}}(\mathcal{F}; n)$ the corresponding n th minimal worst case L^p errors of nonadaptive and adaptive algorithms with respect to a class \mathcal{F} .

6.1. Continuous functions. It is easy to see that in the case of L^p approximation the negative result (11) remains valid for nonadaptive algorithms and $r \geq 2$, as well as for adaptive algorithms and $r \geq 3$.

Indeed, let \mathcal{A}^{non} be an arbitrary nonadaptive algorithm. Let τ be the smallest positive sampling point used by \mathcal{A}^{non} . Let $f_a(x) = cx/a$ for $0 \leq x \leq a$ and $f_a(x) = c$ otherwise. Then for any c we have that $f_{\tau/2}$ and f_τ are in \mathcal{G}_r , both functions share the same approximation, and the L^p distance between them goes to infinity as $c \rightarrow \infty$.

For adaptive \mathcal{A}^{ada} and $r \geq 3$, we choose τ as the smallest positive sampling point used by \mathcal{A}^{ada} for $f \equiv 0$. The two functions with the properties as above are given by $f_\pm(x) = \pm cx(x - \tau)$ for $0 \leq x \leq \tau$ and $f_\pm(x) = 0$ otherwise.

Surprisingly, for smoothness $r = 2$ it is possible to construct an adaptive algorithm $\mathcal{B}_{2,m}^{\text{ada}}$, $m \geq 2$, with the worst case L^p error in the class \mathcal{G}_2 proportional to h^2 , where $h = T/m$. This algorithm is defined as follows. For $x \in (h, T - h)$ we have $(\mathcal{B}_{2,m}^{\text{ada}}f)(x) = (\mathcal{A}_{2,m}^{\text{non}}f)(x)$. Actually, since the approximations are now in L^p space, we take $\kappa = 0$ in the Generic Algorithm.

For $x \in [T - h, T]$ we proceed as follows. Let w_- be the polynomial of the first degree interpolating f at $T - 2h$ and $T - h$. Fix $A > 0$, and define

$$(12) \quad \tau := \min \left(h, \left(\frac{Ah^2}{|f(T) - w_-(T)|} \right)^p \right).$$

Then on $[T - \tau, T]$ we apply the polynomial w_+ of the first degree interpolating f at $T - \tau$ and T , and on $[T - h, T - \tau]$ we apply our (nonadaptive) extrapolation procedure with w_- and w_+ as polynomials extrapolating f from the left and from the right, respectively.

For $x \in [0, h]$ we proceed symmetrically to the case $x \in [T - h, T]$.

Observe that $\mathcal{B}_{2,m}^{\text{ada}}$ is adaptive; however, it uses only *two* adaptively chosen samples.

To estimate the error of $\mathcal{B}_{r,m}^{\text{ada}}$ we need to know that the extrapolation procedure works properly on a given interval even when the singularity s_f is not in that interval, and the jump $\Delta_f^{(1)}$ is arbitrary. (Due to assumption (4), this case was not present in the analysis of $\mathcal{A}_{2,m}^{\text{non}}$.)

Indeed, suppose that the extrapolation procedure is applied on an interval $[t_{-1}, t_1]$ of length at most $2h$. Let p_- and p_+ be correspondingly the polynomials interpolating f at t_{-2}, t_{-1} and t_1, t_2 , with $0 < t_{-1} - t_{-2} \leq h$ and $0 < t_2 - t_1 \leq h$. Assume without loss of generality that $s_f \in (t_1, t_2)$. Then it suffices to consider the error for $\xi \leq x \leq t_1$, where ξ is the approximation of s_f defined as in (9) with $[u, v] = [t_{-1}, t_1]$. For such x

we have

$$\begin{aligned} |f(x) - (\mathcal{B}_{2,m}^{\text{ada}} f)(x)| &= |f(x) - p_+(x)| \\ &\leq |f(x) - p_-(x)| + |p_+(x) - p_-(x)| \\ &\leq |f(x) - p_-(x)| + |f(t_1) - p_-(t_1)| \\ &\leq 2e_h, \end{aligned}$$

where $e_h := \beta_2 \|f^{(2)}\|_{L^\infty} h^2$.

This and the results for $\mathcal{A}_{2,m}^{\text{non}}$ yield that on the interval $(h, T - h)$ the error is uniformly bounded by Ce_h with some constant C . Hence it remains to consider the intervals $[0, h]$ and $[T - h, T]$. Since the situation is symmetric, we concentrate on the second interval.

The same arguments as before yield the estimate Ce_h for $x \in [T - h, T - \tau)$. Consider the last interesting case when both x and s_f are in $[T - \tau, T]$. Then, using decomposition $f = g + \Delta_f^{(1)}(\cdot - s_f)\mathbf{1}_{[s_f, T]}$ with $g \in W_2(0, T)$, we have

$$\begin{aligned} |f(x) - w_-(x)| &\leq |g(x) - w_-(x)| + \left| \Delta_f^{(1)} \right| (x - s_f)\mathbf{1}_{[s_f, T]}(x) \\ &\leq e_h + \left| \Delta_f^{(1)} \right| (T - s_f) \\ &\leq 2e_h + |f(T) - w_-(T)|. \end{aligned}$$

We also have

$$\begin{aligned} |w_+(x) - w_-(x)| &\leq \max(|f(t - \tau) - w_-(T - \tau)|, |f(T) - w_-(T)|) \\ &\leq \max(e_h, |f(T) - w_-(T)|). \end{aligned}$$

Hence

$$\begin{aligned} |f(x) - (\mathcal{B}_{2,m}^{\text{ada}} f)(x)| &= |f(x) - w_+(x)| \\ &\leq |f(x) - w_-(x)| + |w_-(x) - w_+(x)| \\ &\leq 4 \max(e_h, |f(T) - w_-(T)|). \end{aligned}$$

This and definition (12) of τ yield

$$\|f - \mathcal{B}_{2,m}^{\text{ada}} f\|_{L^p(T-\tau, \tau)} \leq 4h^2 \max(\beta_2 \|f^{(2)}\|_{L^\infty}, A).$$

Thus we have shown the following result.

THEOREM 4. *Let $1 \leq p < \infty$. Then*

$$\begin{aligned} e_p^{\text{non}}(\mathcal{G}_2; n) &= \infty, \\ e_p^{\text{ada}}(\mathcal{G}_2; n) &\asymp n^{-2}, \end{aligned}$$

and for $r \geq 3$

$$e_p^{\text{non}}(\mathcal{G}_r; n) = e_p^{\text{ada}}(\mathcal{G}_r; n) = \infty.$$

6.2. Discontinuous functions. In this subsection, we relax the requirement that the approximated functions are continuous. That is, we return to the class $F_r = F_r(0, T)$, defined by (2), of functions f for which the jump $\Delta_f^{(0)}$ is not necessarily

zero. Note that f is right-continuous at s_f , i.e., $f(s_f) = f_+(s_f)$. We also naturally extend the definition of \mathcal{T}_f to

$$\mathcal{T}_f(t) := \sum_{j=0}^{r-1} \Delta_f^{(j)} \frac{t^j}{j!}$$

so that $f(x) = g(x) + \mathcal{T}_f(x - s_f)\mathbf{1}_{[s_f, T]}(x)$.

Approximation of such functions was studied in [11], where it was noticed that it is impossible to have algorithms with error converging to zero in the uniform norm. Therefore the authors proposed an algorithm with error converging at speed n^{-r} in L^p for any $f \in F_r$. Moreover, the algorithm from [11] has the worst case L^p error proportional to n^{-r} with respect to the class

$$\mathcal{F}_r^d := \left\{ f \in F_r \mid \|f^{(r)}\|_{L^\infty} \leq L_r, \|f'\|_{L^\infty} \leq L_1, \left| \Delta_f^{(0)} \right| \leq D_0 \right\}.$$

Using proof techniques of the present paper, similar results can be obtained with the condition $\|f'\|_{L^\infty} \leq L_1$ replaced by some other assumptions, cf. Theorems 2 and 3. For that purpose, we first generalize Propositions 2, 3, and 4 to the case of functions with discontinuity at s_f .

Let $\mathcal{A}_{r,m}^{\text{non}}$ be the nonadaptive version of our algorithm with the only difference that $\kappa = 0$ in the Generic Algorithm. Then we have the following results parallel to Propositions 2 and 3.

PROPOSITION 5. *Let $f \in F_r$. If the singularity s_f is essential and $h \leq \min(\delta_0(f), \omega_0(f))$, then*

$$\|f - \mathcal{A}_{r,m}^{\text{non}} f\|_{C([u,v])} \leq \left| \Delta_f^{(0)} \right| + 3\beta_r \|f^{(r)}\|_{L^\infty} h^r.$$

Proof. Assume without loss of generality that $u < s_f \leq \xi \leq v$. Then, in the critical interval $s_f < x < \xi$, we have

$$|f(x) - (\mathcal{A}_{r,m}^{\text{non}} f)(x)| = |f(x) - p_-(x)| \leq |\mathcal{T}_f(x - s_f)| + e_h.$$

We now have two cases depending on the maximum of $|\mathcal{T}_f|$ in $[0, \xi - s_f]$. If the maximum is attained at 0, then the error above is bounded by $|\mathcal{T}_f(0)| + e_h = |\Delta_f^{(0)}| + e_h$. Otherwise, by monotonicity condition (10), the maximum is at $\xi - s_f$. Then $|f(x) - p_-(x)| \leq |\mathcal{T}_f(\xi - s_f)| + e_h$. Note that

$$\mathcal{T}_f(\xi - s_f) = (p_+(\xi) - p_-(\xi)) + (f(\xi) - p_+(\xi)) - (g(\xi) - p_-(\xi))$$

which gives

$$|\mathcal{T}_f(\xi - s_f)| \leq \left| \Delta_f^{(0)} \right| + 2e_h$$

and completes the proof. \square

PROPOSITION 6. *Let $f \in F_2$. If the singularity s_f is essential and $h \leq \delta_0(f)$, then*

$$\|f - \mathcal{A}_{2,m}^{\text{non}} f\|_{C([u,v])} \leq \left| \Delta_f^{(0)} \right| + 6 \|f^{(2)}\|_{L^\infty} h^2.$$

Proof. Indeed, for $s_f < x < \xi$ we have

$$\begin{aligned} |f(x) - p_-(x)| &\leq |p_+(x) - p_-(x)| + |f(x) - p_+(x)| \\ &\leq |p_+(s_f) - p_-(s_f)| + e_h \\ &\leq \left| \Delta_f^{(0)} \right| + 2e_h, \end{aligned}$$

where $e_h = \beta_2 \|f^{(2)}\|_{L^\infty} h^2 = 3 \|f^{(2)}\|_{L^\infty} h^2$. \square

We now switch to the adaptive version $\mathcal{A}_{r,m}^{\text{ada}}$ of our algorithm (again with $\kappa = 0$). Proceeding exactly as in the proof of Proposition 4 we obtain the following result.

PROPOSITION 7. *Let $f \in F_r$. If the singularity s_f is essential and $h \leq \delta_0(f)$, then*

$$\|f - \mathcal{A}_{r,m} f\|_{C([u^*, v^*])} \leq (1 + \gamma) \left| \Delta_f^{(0)} \right| + (2 + \gamma) \beta_r \|f^{(r)}\|_{L^\infty} h^r.$$

It now follows that the L_p errors of $\mathcal{A}_{r,m}^{\text{non}}$ and $\mathcal{A}_{r,m}^{\text{ada}}$, under the assumptions of corresponding propositions, are upper bounded by $C_r(\|f^{(r)}\|_{L^\infty} m^{-r} + |\Delta_f^{(0)}| m^{-1/p})$. This bound can be improved by using the adaptive (bisection-like) procedure, described in detail in [11, section 5], that locates an interval of length at most m^{-rp} containing the essential singularity (if it exists) and by applying the extrapolation on that interval. Such modified algorithms use $n = o(m)$ function evaluations and have the L^p error bounds proportional to

$$\left(\|f^{(r)}\|_{L^\infty} + \left| \Delta_f^{(0)} \right| \right) n^{-r}.$$

Our analysis yields the following result for discontinuous functions that is parallel to Theorem 2, where continuous functions are considered. Let

$$\mathcal{F}_r := \left\{ f \in F_r(0, T) \mid \|f^{(r)}\|_{L^\infty} \leq L_r, \left| \Delta_f^{(0)} \right| \leq D_0 \right\}.$$

Let \mathcal{F}_r^* , $*$ $\in \{a, b, c\}$, be defined as \mathcal{G}_r^* in section 5, with \mathcal{G}_r replaced by \mathcal{F}_r .

THEOREM 5. *We have*

$$\begin{aligned} e_p^{\text{non}}(\mathcal{F}_2^*; n) &\asymp n^{-2}, \\ e_p^{\text{non}}(\mathcal{F}_r^*; n) &= \infty, \quad r \geq 3, \end{aligned}$$

and

$$e_p^{\text{ada}}(\mathcal{F}_r^*; n) \asymp n^{-r}, \quad r \geq 2.$$

7. Multiple singularities. The results of the previous sections rely very much on the fact that the functions being approximated have at most one singular point. In this section, we consider a more general case by allowing multiple singularities.

Let $F_r^\infty = F_r^\infty(0, T)$ be the set of functions $f : [0, T] \rightarrow \mathbb{R}$ that are piecewise r -smooth. That is, $f \in F_r^\infty$ iff there are a function $g \in W_r(0, T)$, an integer $k = k_f \geq 0$, points $0 = s_0 < s_1 < \dots < s_k < s_{k+1} = T$, and numbers $\Delta_i^{(j)}$, $i = 1, \dots, k$, $j = 1, \dots, r - 1$, such that

$$f(x) = g(x) + \sum_{i=1}^k \mathbf{1}_{[s_i, T]}(x) \sum_{j=0}^{r-1} \Delta_i^{(j)} \frac{(x - s_i)^j}{j!}.$$

Note that s_i , $1 \leq i \leq k$, are the singularities of f and $\Delta_i^{(j)}$, $0 \leq j \leq r - 1$, are the corresponding discontinuity jumps. We are interested in approximating continuous functions

$$f \in G_r^\infty = G_r^\infty(0, T) := F_r^\infty(0, T) \cap C([0, T]).$$

Obviously, $f \in G_r^\infty$ iff $f \in F_r^\infty$ and $\Delta_i^{(0)} = 0$ for all $1 \leq i \leq k$.

We distinguish in G_r^∞ the classes G_r^ℓ of functions with no more than ℓ singular points:

$$G_r^\ell := \{f \in G_r^\infty \mid k_f \leq \ell\}.$$

In particular, $G_r^1 = G_r$ is the class considered in the previous sections. Obviously,

$$G_r^0 \subset G_r^1 \subset \dots \subset G_r^\ell \subset \dots \subset G_r^\infty \quad \text{and} \quad G_r^\infty = \bigcup_{\ell=0}^\infty G_r^\ell.$$

7.1. Worst case setting. Let

$$\mathcal{G}_r^\ell := \left\{ f \in G_r^\ell \mid \|f^{(r)}\|_{L^\infty} \leq L_r \right\}.$$

Let $\mathcal{G}_r^{\ell,*}$, $* \in \{a, b, c\}$ be the classes defined as \mathcal{G}_r^* in section 5 with \mathcal{G}_r replaced by \mathcal{G}_r^ℓ and with

$$\delta_0(f) := \frac{1}{r-1} \min(s_1, T - s_k).$$

THEOREM 6. *Even for $L_r = 0$, we have*

$$e_\infty^{\text{ada}}(\widehat{\mathcal{G}}_r^2; n) = \infty, \quad \text{where} \quad \widehat{\mathcal{G}}_r^2 = \mathcal{G}_r^{2,a} \cap \mathcal{G}_r^{2,b} \cap \mathcal{G}_r^{2,c}.$$

Proof. Let \mathcal{A}_n be an adaptive algorithm. Choose an arbitrary $A > 0$, and define functions $\psi_0(x) = Ax$ and $\psi_T(x) = A(x - T)$. We construct points x_i and intervals (a_i, b_i) as follows. Initially, $(a_0, b_0) = (\delta, T - \delta)$. If the first sampling point $x_1 \notin (a_0, b_0)$, then $(a_1, b_1) = (a_0, b_0)$. Otherwise, $(a_1, b_1) = (a_0, x_1)$ or $(a_1, b_1) = (x_1, b_0)$, whichever is longer. The second sampling point x_2 is selected using the value $\psi_0(x_1)$ if $x_1 \leq a_1$ and $\psi_T(x_1)$ if $x_1 \geq b_1$. The subinterval (a_2, b_2) equals (a_1, b_1) if $x_2 \notin (a_1, b_1)$; otherwise, it is the longer interval between (a_1, x_2) and (x_2, b_1) . Repeating this process inductively n times we get an interval $(a, b) = (a_n, b_n) \subseteq (\delta, T - \delta)$ whose interior does not include any of the points x_i , and $b - a \geq (T - 2\delta)2^{-n}$. Moreover, x_i are the sampling points for any function f satisfying $f(x) = \psi_0(x)$ for $x \leq a$, and $f(x) = \psi_T(x)$ for $x \geq b$.

Consider now

$$f_1(x) = \begin{cases} \psi_0(x), & x \leq a, \\ \psi_0(a)\frac{x-c}{a-c} + \psi_T(c)\frac{x-a}{c-a}, & a < x < c, \\ \psi_T(x), & c \leq x, \end{cases}$$

and

$$f_2(x) = \begin{cases} \psi_0(x), & x \leq c, \\ \psi_0(c)\frac{x-b}{c-b} + \psi_T(b)\frac{x-c}{b-c}, & c < x < b, \\ \psi_T(x), & b \leq x, \end{cases}$$

where $c = (a + b)/2$. Clearly, each f_k , $k = 1, 2$, is continuous, its derivative is discontinuous at exactly two points, and

$$\delta_0(f_k) \geq \delta.$$

Moreover, $f_k^{(j)}$ are T -periodic for all $j \geq 0$. Hence, indeed, $f_k \in \widehat{\mathcal{G}}_r^2$.

Since $f_1(x_i) = f_2(x_i)$ for all i , the algorithm \mathcal{A}_n cannot distinguish between the two functions. This and the fact that $\|f_1 - f_2\|_C = f_1(c) - f_2(c) = AT$ imply that for at least one of the functions the error is at least $AT/2$. Hence since A can be arbitrarily large, the worst case error in $\widehat{\mathcal{G}}_r^2$ equals infinity. \square

The situation changes for the class

$$\mathcal{G}_r^{\ell,d} := \{ f \in \mathcal{G}_r^\ell \mid \|f'\|_{L^\infty} \leq L_1 \}.$$

THEOREM 7. *We have*

$$\begin{aligned} e_\infty^{\text{non}}(\mathcal{G}_2^{2,d}; n) &\asymp n^{-1}, \\ e_\infty^{\text{ada}}(\mathcal{G}_2^{2,d}; n) &\asymp n^{-2}. \end{aligned}$$

In all of the other cases, i.e., for $r = 2$ and $\ell \geq 3$, or $r \geq 3$ and $\ell \geq 2$,

$$e_\infty^{\text{ada}}(\mathcal{G}_r^{\ell,d}; n) \asymp e_\infty^{\text{non}}(\mathcal{G}_r^{\ell,d}; n) \asymp n^{-1}.$$

Proof. We first show all of the lower bounds.

For $\ell \geq 3$, the lower bound of any adaptive algorithm \mathcal{A} is obtained as follows. Let $x_i, 1 \leq i \leq n$, be the sampling points used by \mathcal{A} for $f \equiv 0$. Then we can find an interval (a, b) of length $T/(n + 1)$ that does not contain any x_i . Let f^* be the ‘‘hat’’ function defined as $f^*(x) = 0$ for $x \notin (a, b)$, $f^*(x) = L_1(x - a)$ for $a \leq x \leq c$, and $f^*(x) = -L_1(x - b)$ for $c < x \leq b$, where $c = (a + b)/2$. Then f^* and $-f^*$ share the same information and are in $\mathcal{G}_r^{\ell,d}$. Hence for one of them the error is at least $\|f^*\|_{L^\infty} = L_1T/(2(n + 1))$.

For $\ell = 2$ and $r \geq 3$, we construct the interval (a, b) as before and define $f^*(x) = L_1(x - a)(x - b)/(b - a)$ for $x \in (a, b)$ and $f^*(x) = 0$ for $x \notin (a, b)$. The error for f^* or $-f^*$ is at least $\|f^*\|_{L^\infty} = L_1T/(4(n + 1))$.

For $\ell = 2$ and $r = 2$, we again select (a, b) as before. In the case of adaption we take $f^* = 0$ for $x \notin (a, b)$ and $f^*(x) = A(x - a)(x - b)$, with $A = \min(L_2/2, (b - a)^{-1})$. Then the error for f^* or $-f^*$ is at least $A(b - a)^2/4 \asymp n^{-2}$. In the case of nonadaptive algorithms we take the two functions as follows: $f_1^*(x) = L_1(x - a)$ for $x \leq c$ and $f_1^*(x) = -L_1(x - b)$ for $x > c$; $f_2^*(x) = f_1^*(x)$ for $x \notin (a, b)$ and $f_2^*(x) = 0$ for $x \in (a, b)$. For one of the functions the error is at least $\|f_1^* - f_2^*\|_\infty/2 = L_1T/(4(n + 1))$.

The upper bounds for the minimal error of nonadaptive algorithms is achieved by piecewise linear interpolation based on equispaced points $x_i = (i - 1)/(n - 1)$, $1 \leq i \leq n$, where the error for any $f \in \mathcal{G}_r^{\ell,d}$ is at most $L_1T/(2(n - 1))$.

It remains to construct an adaptive algorithm with the worst case error of order n^{-2} in the class $\mathcal{G}_2^{2,d}$. For an initial resolution $h = T/m$, the algorithm divides the interval $[0, T]$ into two sets V and $W = [0, T] \setminus V$, where V is the sum of some subintervals $[t_i, t_{i+1}]$. Then the piecewise polynomial interpolation of degree 1 with resolution h^2 is applied on V and with resolution h on \overline{W} . Specifically, V is defined as follows. Let $D \geq 0$ be arbitrary. Let

$$|d_{i_1}| \geq |d_{i_2}| \geq |d_{i_3}|$$

be the three largest divided differences among $|d_i| := |f[t_i, t_{i+1}, t_{i+2}]|$, $0 \leq i \leq m - 2$.

```

00 V-CONSTRUCTION;
01 begin
02   V := [0, h] ∪ [T - h, T];
03   if |di1| > D then V := V ∪ [ti1, ti1+2];
04   if |di2| > D then
05     begin
06       V := V ∪ [ti2, ti2+2];
07       if (|i2 - i1| = 1) and (|di3| > D) then V := V ∪ [ti3, ti3+2]
08     end
09 end.

```

From the construction it follows that all of the divided differences $|d_i|$ “covering” points $x \notin V$ do not exceed $D_f := \max(D, \|f^{(2)}\|_{L^\infty}/2)$. Then Lemma 3 of the appendix gives that the error in W is of order h^2 . Since in V the resolution is h^2 , the error in this set is of order h^2 as well. The proof completes with the observation that the described algorithm uses more than $n = (m + 1) + 7(m - 1) = 8m + 6$ samples. \square

Thus, unlike for just one singularity, for multiple singularities the worst case convergence n^{-r} can be obtained only when $r = \ell = 2$.

Remark 1. It is easy to check that the algorithm for $r = \ell = 2$ described in the proof of Theorem 7 has the worst case error of order n^{-2} in the more general than $\mathcal{G}_2^{2,d}$ class:

$$\mathcal{G}_r^{2,e} := \left\{ f \in \mathcal{G}_2^2 \mid \max \left(\left| \Delta_1^{(1)} \right|, \left| \Delta_2^{(1)} \right| \right) \leq D_1 \right\}.$$

However, for $r \geq 4$ the minimal error $e_\infty^{\text{ada}}(\mathcal{G}_r^{2,e}; n) = \infty$. To see this, it suffices to repeat the proof of Theorem 6 with modified f_1 and f_2 . For $a < x < c$ and $c < x < b$, these functions are, respectively, defined as polynomials of degree 3 interpolating data $\psi_0(a), \psi'_0(a), \psi_T(c), \psi'_T(c)$ and $\psi_0(c), \psi'_0(c), \psi_T(b), \psi'_T(b)$.

7.2. Asymptotic setting. The problem with multiple singularities in the worst case setting relies on the fact that the singular points can be arbitrarily close to one another, and therefore it is impossible to separate them using a prescribed number n of function evaluations. This problem disappears in the *asymptotic setting* where the optimal rate n^{-r} of convergence can be regained. Recall that in the asymptotic setting we investigate how fast the error of approximation converges to zero for any $f \in G_r^\infty$ as the number of samples increases to infinity.

The rate n^{-r} is already obtained by the adaptive algorithm $\overline{\mathcal{A}}_{r,m}^{\text{ada}}$ from [11, section 6.2], with obvious modification related to the fact that $\overline{\mathcal{A}}_{r,m}^{\text{ada}}$ was originally designed for F_r^∞ , i.e., for functions with possible discontinuities. We refer to [11] for a precise description and analysis of this algorithm. Here we mention only that $\overline{\mathcal{A}}_{r,m}^{\text{ada}}$ relies, roughly speaking, on the application of an adaptive detection mechanism on $\ell = \ell(m)$ disjoint subintervals corresponding to ℓ largest divided differences, where $\ell(m)$ “slowly” increases to ∞ .

It follows from [11, Theorems 4 and 5] that $\overline{\mathcal{A}}_{r,m}^{\text{ada}}$ uses $n = O(m)$ samples and

$$(13) \quad \lim_{m \rightarrow \infty} \left\| f - \overline{\mathcal{A}}_{r,m}^{\text{ada}} f \right\|_C \cdot m^r = \alpha_r T^r \|f^{(r)}\|_{L^\infty} \quad \text{for all } f \in G_r^\infty,$$

where

$$\alpha_r := \frac{1}{r!} \max_{0 \leq t \leq 1} \prod_{i=1}^r \left| t - \frac{i-1}{r-1} \right|.$$

We want to stress that for continuous functions the convergence n^{-r} can be achieved by nonadaptive algorithms as well. Indeed, it is possible to modify $\overline{\mathcal{A}}_{r,m}^{\text{ada}}$, replacing the adaptive detection mechanism by the nonadaptive version (9) of the extrapolation procedure, to get a nonadaptive algorithm $\overline{\mathcal{A}}_{r,m}^{\text{non}}$ with the following properties.

THEOREM 8. *The nonadaptive algorithm $\overline{\mathcal{A}}_{r,m}^{\text{non}}$ uses $n = m + 1$ samples and*

$$(14) \quad \limsup_{m \rightarrow \infty} \left\| f - \overline{\mathcal{A}}_{r,m}^{\text{non}} f \right\|_C \cdot m^r \leq 3.5 \beta_r T^r \|f^{(r)}\|_{L^\infty} \quad \forall f \in C_r^\infty.$$

We end this section by comparing the asymptotic error bounds (13) and (14). Since the adaptive algorithm $\overline{\mathcal{A}}_{r,m}^{\text{ada}}$ asymptotically uses $m(r - 1)$ samples, it is only fair to compare it to the nonadaptive algorithm $\overline{\mathcal{A}}_{r,m(r-1)}^{\text{non}}$. Then

$$\frac{\left\| f - \overline{\mathcal{A}}_{r,m(r-1)}^{\text{non}} f \right\|_C}{\left\| f - \overline{\mathcal{A}}_{r,m}^{\text{ada}} f \right\|_C} \leq R(r) (1 + o(1)) \quad \text{as } m \rightarrow \infty,$$

where

$$R(r) = \frac{3.5 \beta_r}{\alpha_r (r - 1)^r}.$$

The exact value of α_r can easily be computed for small values of r , i.e., we have $\alpha_2 = 1/8$, $\alpha_3 = \sqrt{3}/216$, and $\alpha_4 = 1/1944$. This yields

$$R(2) = 84, \quad R(3) = 315\sqrt{3} = 545.596 \dots, \quad \text{and} \quad R(4) = 2940.$$

For larger values of r we proceed as follows. Substituting $t = 1/(2(r - 1))$ in the product $\prod_{i=1}^r |t - \frac{i-1}{r-1}|$ defining the constant α_r , we get

$$\alpha_r \geq \frac{\prod_{i=1}^{r-1} (2i - 1)}{r! (2(r - 1))^r} = \frac{(2r - 2)!}{r! 2^{2r-1} (r - 1)! (r - 1)^r}.$$

Since $\beta_r = (2r - 1)!/(r!(r - 1)!)$, we then have

$$R(r) \leq 1.75 (2r - 1) 4^r.$$

On the other hand,

$$R(r) \geq 4r \binom{2r - 1}{r - 1},$$

which follows from a well-known fact that

$$r! \alpha_r \leq \frac{(r - 1)!}{4 (r - 1)^r}.$$

This means that the ratio between the errors of the nonadaptive and adaptive algorithms could be proportional to 4^r . However, this is not so bad since those errors are inversely proportional to the number of evaluations raised to power r . Hence for the same errors in both algorithms it suffices to let the nonadaptive algorithm use four times as many evaluation points.

Appendix.

LEMMA 1. Let $f : [a, b] \rightarrow \mathbb{R}$ and $a \leq \tau_0 < \tau_1 < \dots < \tau_k \leq b$ with $k \geq r$. Then for any subsequence $0 \leq j_0 < j_1 < \dots < j_r \leq k$ we have

$$|f[\tau_{j_0}, \tau_{j_1}, \dots, \tau_{j_r}]| \leq \max_{0 \leq i \leq k-r} |f[\tau_i, \tau_{i+1}, \dots, \tau_{i+r}]|.$$

Proof. (The lemma follows, in particular, from [2, Proposition 23]. For completeness, we provide a simple and independent proof.)

Suppose the lemma is not true. We can assume without loss of generality that

$$f[\tau_{j_0}, \tau_{j_1}, \dots, \tau_{j_r}] > 0.$$

Let p_f be the polynomial of degree r interpolating f at t_{j_i} for $0 \leq i \leq r$. Then

$$(f - p_f)[\tau_i, \tau_{i+1}, \dots, \tau_{i+r}] < 0 \quad \text{for } 0 \leq i \leq k - r.$$

This implies that the $(r - 1)$ st order divided differences $(f - p_f)[\tau_i, \tau_{i+1}, \dots, \tau_{i+r-1}]$, $0 \leq i \leq k + 1 - r$, change sign at most once (where hitting zero is considered a sign change), the $(r - 2)$ nd order divided differences change sign at most twice, and so on. Finally, $(f - p_f)(\tau_i)$, $0 \leq i \leq k$, change sign at most r times. This is a contradiction since $f(\tau_{j_i}) = p_f(\tau_{j_i})$ for $0 \leq i \leq r$. \square

LEMMA 2. Let $h = T/m$ and $t_j = jh$ for all j . Let $f \in F_r^1(0, T)$ with singularity $s_f \in [t_{r-1}, t_{m-r+1}]$ if $\Delta_f^{(0)} = 0$ and $s_f \in (t_{r-1}, t_{m-r+1})$ if $\Delta_f^{(0)} \neq 0$. If the divided differences satisfy

$$|f[t_j, t_{j+1}, \dots, t_{j+r}]| \leq B, \quad 0 \leq j \leq m - r,$$

then for any $0 \leq i \leq m - r + 1$ and $0 \leq x \leq T$ the error of polynomial interpolation/extrapolation

$$|f(x) - \mathcal{P}_r(f; t_i, t_{i+r-1})(x)| \leq \left(\|f^{(r)}\|_{L^\infty} + C_x B \right) h^r,$$

where $C_x = c_{r, \ell(x)} \Omega_r$,

$$\ell(x) = 2(r - 1) + \lceil \text{dist}(x, [t_i, t_{i+r-1}]) / h \rceil,$$

and $c_{r,k}$ and Ω_r are defined by (6) and (7), respectively. (Here $\text{dist}(x, [u_1, u_2])$ is the distance of x from the interval $[u_1, u_2]$.)

In particular, if $|f[t_j, t_{j+1}, \dots, t_{j+r}]| \leq B$ holds for $i - r + 1 \leq j \leq i + r - 2$, then the error of interpolation

$$\|f - \mathcal{P}_r(f; t_i, t_{i+r-1})\|_{C([t_i, t_{i+r-1}])} \leq \left(\|f^{(r)}\|_{L^\infty} + c_{r, 2(r-1)} \Omega_r B \right) h^r.$$

Proof. By the assumption about the location of s_f , there exists $0 \leq j \leq m - r + 1$ such that $x \in (t_{j-1}, t_{j+r})$, $s_f \notin (\min(t_j, x), \max(t_{j+r-1}, x))$, and f is left-continuous at $\max(t_{j+r-1}, x)$. Indeed, we could take j such that $x \in (t_{j-1}, t_j]$ when $s_f \leq x$ and $x \in [t_{j+r-1}, t_{j+r})$ when $s_f > x$.

Denoting by q_f the polynomial of degree at most $r - 1$ interpolating f at t_j, \dots, t_{j+r-1} we have

$$(15) \quad |f(x) - q_f(x)| \leq \|f^{(r)}\|_{L^\infty} h^r.$$

Let, for brevity, $p_f = \mathcal{P}_r(f; t_i, t_{i+r-1})$. We now estimate the error $|f(t_k) - p_f(t_k)|$ for $k = j, j+1, \dots, j+r-1$. To that end, assume without loss of generality that $i \leq j$. Since $(f - p_f)(t_l) = 0$ for $i \leq l \leq i+r-1$, by definition of the divided differences we have

$$(f - p_f)[t_i, t_{i+1}, \dots, t_{i+r-1}, t_k] = \frac{(f - p_f)(t_k)}{(k - r + 1)(k - r + 2) \cdots (k - 1)kh^r}.$$

On the other hand, by Lemma 1 we also have

$$|(f - p_f)[t_i, t_{i+1}, \dots, t_{i+r-1}, t_k]| \leq \max_{i \leq l \leq k-r} |(f - p_f)[t_l, \dots, t_{l+r}]| \leq B.$$

Hence

$$|(f - p_f)(t_k)| \leq c_{r, k-i} B h^r.$$

Since $p_f - q_f$ is a polynomial of degree at most $r - 1$, we now have

$$\begin{aligned} |p_f(x) - q_f(x)| &\leq \sum_{k=j}^{j+r-1} \left| (p_f - q_f)(t_k) l_{k-j} \left(\frac{x - t_j}{h} \right) \right| \\ &\leq B h^r \sum_{k=j}^{j+r-1} c_{r, k-i} \left| l_{k-j} \left(\frac{x - t_j}{h} \right) \right| \\ (16) \qquad \qquad &\leq C B h^r, \end{aligned}$$

where $C = c_{r, j-i+r-1} \Omega_r = C_x$.

Combining (15) and (16) we finally obtain

$$\begin{aligned} |f(x) - p_f(x)| &\leq |(f - q_f)(x)| + |p_f(x) - q_f(x)| \\ &\leq \left(\|f^{(r)}\|_{L^\infty} + C_x B \right) h^r \end{aligned}$$

as claimed. \square

LEMMA 3. Let $f \in G_2^2(0, T)$. Let $1 \leq i \leq m - 2$. If

$$|f[t_j, t_{j+1}, t_{j+2}]| \leq B \quad \text{for } j = i - 1, i,$$

then

$$\|f - \mathcal{P}_2(f; t_i, t_{i+1})\|_{C([t_i, t_{i+1}])} \leq \left(\frac{9}{8} \|f^{(2)}\|_{L^\infty} + 2B \right) h^2.$$

Proof. Let $s_1 < s_2$ be the singular points of f . Denote $p_f := \mathcal{P}_2(f; t_i, t_{i+1})$. If s_1 and s_2 are not in (t_i, t_{i+1}) , then the error of interpolation is at most $\|f^{(2)}\|_{L^\infty} h^2 / 8$. Suppose $s_1 \in (t_i, t_{i+1})$. (The case $s_2 \in (t_i, t_{i+1})$ is symmetric.) Then for $t_i \leq x \leq s_1$ the error is obtained as in the proof of Lemma 2. That is, denoting by q_f the polynomial of degree 1 interpolating f at t_{i-1} and t_i we have $|f(x) - q_f(x)| \leq \|f^{(2)}\|_{L^\infty} h^2$. We also have $|f(t_{i-1}) - p_f(t_{i-1})| \leq 2Bh^2$. Hence

$$\begin{aligned} |f(x) - p_f(x)| &\leq |f(x) - q_f(x)| + |q_f(x) - p_f(x)| \\ &\leq \left(\|f^{(2)}\|_{L^\infty} + 2B \right) h^2. \end{aligned}$$

The same bound is obtained for $s_2 < x \leq t_{i+1}$ (if $s_2 < t_{i+1}$). Consider the remaining case $s_1 < x \leq s^* := \min(t_{i+1}, s_2)$. Let v_f be the polynomial of degree 1 interpolating f at s_1 and s^* . Then $|f(x) - v_f(x)| \leq \|f^{(2)}\|_{L^\infty} h^2/8$ and

$$\begin{aligned} |v_f(x) - p_f(x)| &\leq \max(|v_f(s_1) - p_f(s_1)|, |v_f(s^*) - p_f(s^*)|) \\ &\leq \left(\|f^{(2)}\|_{L^\infty} + 2B \right) h^2. \end{aligned}$$

Hence

$$\begin{aligned} |f(x) - p_f(x)| &\leq |f(x) - v_f(x)| + |v_f(x) - p_f(x)| \\ &\leq \left(\frac{9}{8} \|f^{(2)}\|_{L^\infty} + 2B \right) h^2 \end{aligned}$$

as claimed. \square

REFERENCES

- [1] F. ARANDIGA, A. COHEN, R. DONAT, AND N. DYN, *Interpolation and approximation of piecewise smooth functions*, SIAM J. Numer. Anal., 43 (2005), pp. 41–57.
- [2] C. DE BOOR, *Divided differences*, Surv. Approx. Theory, 1 (2005), pp. 46–69.
- [3] E.J. CANDÈS AND D.L. DONOHO, *New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities*, Comm. Pure Appl. Math., 57 (2004), pp. 219–266.
- [4] D.L. DONOHO, M. VETTERLI, R.A. DEVORE, AND I. DAUBECHIES, *Data compression and harmonic analysis*, IEEE Trans. Inform. Theory, 44 (1998), pp. 2435–2476.
- [5] K.S. ECKHOFF, *Accurate reconstructions of functions of finite regularity from truncated Fourier series expansions*, Math. Comp., 64 (1995), pp. 671–690.
- [6] K.S. ECKHOFF, *On a high order numerical method for functions with singularities*, Math. Comp., 67 (1998), pp. 1063–1087.
- [7] S. ENGELBERG AND E. TADMOR, *Recovery of edges from spectral data with noise—a new perspective*, SIAM J. Numer. Anal., 46 (2008), pp. 2620–2635.
- [8] A. HARTEN, *ENO schemes with subcell resolution*, J. Comput. Phys., 83 (1994), pp. 148–184.
- [9] H.N. MHASKAR AND J. PRESTIN, *Polynomial frames: A fast tour*, in Approximation Theory XI, Gatlinburg 2004, C.K. Chui, M. Neamtu, and L.L. Schumaker, eds., Nashboro Press, Brentwood, TN, 2005, pp. 287–318.
- [10] L. PLASKOTA AND G.W. WASILKOWSKI, *Adaption allows efficient integration of functions with unknown singularities*, Numer. Math., 102 (2005), pp. 123–144.
- [11] L. PLASKOTA, G.W. WASILKOWSKI, AND Y. ZHAO, *The power of adaption for approximating functions with singularities*, Math. Comp., 77 (2008), pp. 2309–2338.
- [12] E.B. SAFF AND V. TOTIK, *Polynomial approximation of piecewise analytic functions*, J. London Math. Soc., 39 (1989), pp. 487–498.
- [13] E. TADMOR, *Filters, mollifiers and the computation of the Gibbs phenomenon*, Acta Numer., 16 (2007), pp. 305–378.
- [14] J.F. TRAUB, G.W. WASILKOWSKI, AND H. WOŹNIAKOWSKI, *Information-Based Complexity*, Academic Press, New York, 1988.

EXPONENTIAL ROSENBRUCK-TYPE METHODS*

MARLIS HOCHBRUCK[†], ALEXANDER OSTERMANN[‡], AND JULIA SCHWEITZER[†]

Abstract. We introduce a new class of exponential integrators for the numerical integration of large-scale systems of stiff differential equations. These so-called Rosenbrock-type methods linearize the flow in each time step and make use of the matrix exponential and related functions of the Jacobian. In contrast to standard integrators, the methods are fully explicit and do not require the numerical solution of linear systems. We analyze the convergence properties of these integrators in a semigroup framework of semilinear evolution equations in Banach spaces. In particular, we derive an abstract stability and convergence result for variable step sizes. This analysis further provides the required order conditions and thus allows us to construct pairs of embedded methods. We present a third-order method with two stages, and a fourth-order method with three stages, respectively. The application of the required matrix functions to vectors are computed by Krylov subspace approximations. We briefly discuss these implementation issues, and we give numerical examples that demonstrate the efficiency of the new integrators.

Key words. exponential Rosenbrock-type methods, exponential integrators, stiff order conditions, stability bounds, convergence bounds, embedded methods of high order, variable step size implementation

AMS subject classifications. 65M12, 65L06

DOI. 10.1137/080717717

1. Introduction. In this paper, we are concerned with a new class of numerical methods for the time integration of large systems of stiff differential equations

$$(1.1) \quad u'(t) = F(t, u(t)), \quad u(t_0) = u_0.$$

Such equations typically arise from spatial discretizations of nonlinear time dependent partial differential equations. The numerical work when solving (1.1) by standard integrators like implicit Runge–Kutta methods or backward differentiation formulas is often dominated by the numerical linear algebra, which is required for the solution of the arising nonlinear systems of equations. For a collection of ODE solvers, test problems, and related references, we refer to [21]. In particular, we point out the codes VODEPK [1, 2] and ROWMAP [28], where the linear algebra is based on Krylov subspace methods. Runge–Kutta discretizations of nonlinear evolution equations have been studied in [19, 20, 22].

Exponential integrators, on the other hand, require the matrix exponential and related functions of a certain matrix. Most exponential integrators analyzed so far in literature [5, 6, 9, 14, 16, 17, 18, 23, 26] make use of a (rough) a priori linearization

$$(1.2) \quad u'(t) = Au(t) + f(t, u(t))$$

of the nonlinear problem (1.1). The matrix A then *explicitly* enters the formulation of the exponential integrator as the argument where the matrix functions are evaluated.

*Received by the editors March 5, 2008; accepted for publication (in revised form) September 17, 2008; published electronically February 4, 2009. This work was supported by the Deutsche Forschungsgemeinschaft through the Transregio-SFB 18.

<http://www.siam.org/journals/sinum/47-1/71771.html>

[†]Mathematisches Institut, Heinrich-Heine Universität Düsseldorf, Universitätsstr. 1, D-40225 Düsseldorf, Germany (marlis@am.uni-duesseldorf.de, schweitzer@am.uni-duesseldorf.de).

[‡]Institut für Mathematik, Universität Innsbruck, Technikerstr. 13, A-6020 Innsbruck, Austria (alexander.ostermann@uibk.ac.at).

Such an approach is justified in situations where the remainder f is small, or at least bounded in terms of A . The latter is the case for semilinear parabolic problems, if f is relatively bounded with respect to A . In particular, if A has a simple structure, it is possible to compute the product of a matrix function with a vector in a fast and reliable way. For instance, if A is the semidiscretization of the Laplacian on a regular rectangular mesh, these functions can be computed by fast Fourier transform techniques. Such an approach has been used in [16].

On the other hand, a fixed linearization like (1.2) can also lead to problems. As the remainder f is integrated explicitly by standard exponential methods, a badly chosen linearization can cause a severe step size restriction. This, for instance, is the case if the numerical solution stays near an equilibrium point (e.g., a saddle point) of the problem for a long time. If the linearization (1.2) is performed far from this equilibrium point, the integrator is forced to take small steps due to stability requirements. This will cause computational inefficiency.

In order to avoid these problems, we propose a new class of exponential integrators that linearize (1.1) in each integration step. The linearization can be computed either analytically or numerically. We first presented this approach in [15]. Here we give a rigorous stability and convergence proof, we discuss a possible variable step size implementation, and we give numerical comparisons. Related ideas have been used in [12] and [27]. Since the Jacobian of the problem changes from step to step, FFT techniques can no longer be used to compute the products of matrix functions with vectors. We will use Krylov subspace approximations instead [7, 11].

The outline of our paper is as follows. In section 2, we introduce the method class and discuss a reformulation of the method which allows an efficient implementation with Krylov subspace methods. An implementation using Leja points was proposed in [3]. Since the reformulation speeds up the Krylov implementation considerably, we will not consider Leja point methods in this paper. In section 3, we introduce the analytic framework and derive preliminary error bounds. We work in a framework of C_0 semigroups that covers many abstract semilinear evolution equations in Banach spaces. In contrast to exponential Runge–Kutta methods [14], the new class of Rosenbrock-type methods produces smaller defects when inserting the exact solution into the numerical scheme. This is due to the linearization. It facilitates the derivation of the order conditions and gives much simpler conditions than in [14]. In particular, it is possible to construct a fourth-order integrator with an embedded third-order method, using three stages only. Since the Jacobian varies from step to step, the stability estimate of the discrete evolution operator is crucial. The necessary stability bounds for variable step size discretizations are derived in section 3.3.

In section 4, we give a convergence bound for methods up to order four. Particular methods of order three and four are given in section 5, and a generalization to nonautonomous problems is discussed in section 6. In section 7, we briefly describe an implementation based on Krylov subspace approximations, and we present two numerical examples: a two-dimensional advection-diffusion-reaction problem and a Schrödinger equation with time dependent potential. The possible extension for analytic semigroups is sketched in the appendix.

2. Exponential Rosenbrock-type methods. In this paper, we consider the time discretization of (possibly abstract) differential equations in autonomous form

$$(2.1) \quad u'(t) = F(u(t)), \quad u(t_0) = u_0.$$

The precise assumptions on the problem class will be stated in section 3 below. The numerical schemes considered are based on a continuous linearization of (2.1) along the numerical solution. For a given point u_n in the state space, this linearization is

$$(2.2a) \quad u'(t) = J_n u(t) + g_n(u(t)),$$

$$(2.2b) \quad J_n = DF(u_n) = \frac{\partial F}{\partial u}(u_n), \quad g_n(u(t)) = F(u(t)) - J_n u(t),$$

with J_n denoting the Jacobian of F and g_n the nonlinear remainder, evaluated at u_n , respectively. The numerical schemes will make *explicit* use of these quantities.

2.1. Method class. Let u_n denote the numerical approximation to the solution of (2.1) at time t_n . Its value at t_0 is given by the initial condition. Applying an explicit exponential Runge–Kutta scheme [14] to (2.2a), we obtain the following class of explicit one-step methods:

$$(2.3a) \quad U_{ni} = e^{c_i h_n J_n} u_n + h_n \sum_{j=1}^{i-1} a_{ij} (h_n J_n) g_n(U_{nj}), \quad 1 \leq i \leq s,$$

$$(2.3b) \quad u_{n+1} = e^{h_n J_n} u_n + h_n \sum_{i=1}^s b_i (h_n J_n) g_n(U_{ni}).$$

Here, $h_n > 0$ denotes a positive time step, and u_{n+1} is the numerical approximation to the exact solution at time $t_{n+1} = t_n + h_n$.

The method is built on s internal stages U_{ni} that approximate the solution at $t_n + c_i h_n$. The real numbers c_i are called nodes of the method. The method is fully explicit and does not require the solution of linear or nonlinear systems of equations. As usual in exponential integrators, the weights $b_i(z)$ are linear combinations of the entire functions

$$(2.4) \quad \varphi_k(z) = \int_0^1 e^{(1-\sigma)z} \frac{\sigma^{k-1}}{(k-1)!} d\sigma, \quad k \geq 1.$$

These functions satisfy the recurrence relations

$$(2.5) \quad \varphi_k(z) = \frac{\varphi_{k-1}(z) - \varphi_{k-1}(0)}{z}, \quad \varphi_0(z) = e^z.$$

The coefficients $a_{ij}(z)$ will be chosen as linear combinations of the related functions $\varphi_k(c_i z)$. Henceforth, the methods (2.3) will be called *exponential Rosenbrock methods*.

Without further mentioning, we will assume throughout the paper that the methods fulfill the following simplifying assumptions:

$$(2.6) \quad \sum_{i=1}^s b_i(z) = \varphi_1(z), \quad \sum_{j=1}^{i-1} a_{ij}(z) = c_i \varphi_1(c_i z), \quad 1 \leq i \leq s.$$

Note that (2.6) implies $c_1 = 0$ and consequently $U_{n1} = u_n$.

Methods that satisfy the simplifying assumptions (2.6) possess several interesting features. They preserve equilibria of (2.1), they have small defects which in turn leads to simple order conditions for stiff problems (section 3.1), they allow a reformulation for efficient implementation (see below), and they can easily be extended to nonautonomous problems (section 6).

2.2. Reformulation of the method. For the implementation of an exponential Rosenbrock method, it is crucial to approximate the application of matrix functions to vectors efficiently. We, therefore, suggest to express the vectors $g_n(U_{nj})$ as

$$g_n(U_{nj}) = g_n(u_n) + D_{nj}, \quad 2 \leq j \leq s.$$

A similar approach was used in [27]. Due to the simplifying assumptions (2.6), the method (2.3) takes the equivalent form

$$(2.7a) \quad U_{ni} = u_n + c_i h_n \varphi_1(c_i h_n J_n) F(u_n) + h_n \sum_{j=2}^{i-1} a_{ij}(h_n J_n) D_{nj},$$

$$(2.7b) \quad u_{n+1} = u_n + h_n \varphi_1(h_n J_n) F(u_n) + h_n \sum_{i=2}^s b_i(h_n J_n) D_{ni}.$$

The main motivation for this reformulation is that the vectors D_{ni} are expected to be small in norm. When computing the application of matrix functions to these vectors with some Krylov subspace method, this should be possible in a low-dimensional subspace. Consequently, only one computationally expensive Krylov approximation will be required in each time step, namely, that involving $F(u_n)$. A similar idea has also been used to make the code `exp4` efficient [12].

3. Analytic framework and preliminary error analysis. For the error analysis of (2.3), we work in a semigroup framework. Background information on semigroups can be found in the textbooks [8, 24]. Let

$$(3.1) \quad J = J(u) = DF(u) = \frac{\partial F}{\partial u}(u)$$

be the Fréchet derivative of F in a neighborhood of the exact solution of (2.1). Throughout the paper we consider the following assumptions.

Assumption C.1. The linear operator J is the generator of a strongly continuous semigroup e^{tJ} on a Banach space X . More precisely, we assume that there exist constants C and ω such that

$$(3.2) \quad \|e^{tJ}\|_{X \leftarrow X} \leq C e^{\omega t}, \quad t \geq 0$$

holds uniformly in a neighborhood of the exact solution of (2.1).

Recall that the analytic functions $b_i(z)$ and $a_{ij}(z)$ are linear combinations of $\varphi_k(z)$ and $\varphi_k(c_i z)$, respectively. These functions are related to the exponential function through (2.4). Assumption C.1 thus guarantees that the coefficients $b_i(hJ)$ and $a_{ij}(hJ)$ of the method are bounded operators. This property is crucial in our proofs.

In the subsequent analysis, we restrict our attention to semilinear problems

$$(3.3) \quad u'(t) = F(u(t)), \quad F(u) = Au + f(u), \quad u(t_0) = u_0.$$

This implies that (2.2b) takes the form

$$(3.4) \quad J_n = A + \frac{\partial f}{\partial u}(u_n), \quad g_n(u(t)) = f(u(t)) - \frac{\partial f}{\partial u}(u_n)u(t).$$

Our main hypothesis on the nonlinearity f is the following.

Assumption C.2. We suppose that (3.3) possesses a sufficiently smooth solution $u : [0, T] \rightarrow X$, with derivatives in X and that $f : X \rightarrow X$ is sufficiently often

Fréchet differentiable in a strip along the exact solution. All occurring derivatives are supposed to be uniformly bounded.

By Assumption C.2, the Jacobian (3.1) satisfies the Lipschitz condition

$$(3.5) \quad \|J(u) - J(v)\|_{X \leftarrow X} \leq C \|u - v\|$$

in a neighborhood of the exact solution.

Remark. If the semigroup generated by J is not only strongly continuous but analytic, more general nonlinearities can be analyzed. To keep our presentation simple, we restrict ourselves to strongly continuous semigroups for the moment and sketch the possible extension to analytic semigroups later in Appendix A.

Examples will be considered in section 7.

3.1. Defects. For brevity, we denote $G_n(t) = g_n(u(t))$. Inserting the exact solution into the numerical scheme gives

$$(3.6a) \quad u(t_n + c_i h_n) = e^{c_i h_n J_n} u(t_n) + h_n \sum_{j=1}^{i-1} a_{ij}(h_n J_n) G_n(t_n + c_j h_n) + \Delta_{ni},$$

$$(3.6b) \quad u(t_{n+1}) = e^{h_n J_n} u(t_n) + h_n \sum_{i=1}^s b_i(h_n J_n) G_n(t_n + c_i h_n) + \delta_{n+1},$$

with defects Δ_{ni} and δ_{n+1} . The computation and estimation of the defects is carried out in the same way as in our previous paper [14, section 4.1]. In particular, expressing the left-hand side of (3.6a) by the variation-of-constants formula

$$u(t_n + c_i h_n) = e^{c_i h_n J_n} u(t_n) + \int_0^{c_i h_n} e^{(c_i h_n - \tau) J_n} G_n(t_n + \tau) d\tau$$

and then expanding G_n into a Taylor series at t_n yields

$$(3.7) \quad \Delta_{ni} = h_n \psi_{1,i}(h_n J_n) G_n(t_n) + h_n^2 \psi_{2,i}(h_n J_n) G'_n(t_n) + \Delta_{ni}^{[2]},$$

with

$$(3.8) \quad \psi_{j,i}(z) = \varphi_j(c_i z) c_i^j - \sum_{k=1}^{i-1} a_{ik}(z) \frac{c_k^{j-1}}{(j-1)!}$$

and remainders $\Delta_{ni}^{[2]}$ satisfying

$$(3.9) \quad \|\Delta_{ni}^{[2]}\| \leq C h_n^3.$$

Small defects in the internal stages facilitate our convergence proofs considerably. This gives a further reason for requiring (2.6), which implies $\psi_{1,i}(z) \equiv 0$. Unfortunately, explicit methods *cannot* have $\psi_{2,i}(z) \equiv 0$ for all i . Nevertheless, the second term on the right-hand side of (3.7) turns out to be small. This is seen from the identity

$$G'_n(t_n) = \frac{\partial g_n}{\partial u}(u(t_n)) u'(t_n) = \left(\frac{\partial f}{\partial u}(u(t_n)) - \frac{\partial f}{\partial u}(u_n) \right) u'(t_n),$$

which itself is a consequence of linearizing at each step; cf. (3.4). By Assumption C.2, this relation implies

$$(3.10) \quad \|G'_n(t_n)\| \leq C \|e_n\|,$$

TABLE 3.1

Stiff order conditions for exponential Rosenbrock methods applied to autonomous problems.

No.	Condition in defect	Order condition	Order
1	$\psi_1(z) \equiv 0$	$\sum_{i=1}^s b_i(z) = \varphi_1(z)$	1
2	$\psi_{1,i}(z) \equiv 0$	$\sum_{j=1}^{i-1} a_{ij}(z) = c_i \varphi_1(c_i z), \quad 2 \leq i \leq s$	2
3	$\psi_3(z) \equiv 0$	$\sum_{i=2}^s b_i(z) c_i^2 = 2\varphi_3(z)$	3
4	$\psi_4(z) \equiv 0$	$\sum_{i=2}^s b_i(z) c_i^3 = 6\varphi_4(z)$	4

with $e_n = u_n - u(t_n)$, and the defects of the internal stages thus obey the bound

$$(3.11) \quad \|\Delta_{ni}\| \leq Ch_n^2 \|e_n\| + Ch_n^3.$$

Similarly, we get for the defects δ_{n+1} at time t_{n+1}

$$(3.12) \quad \delta_{n+1} = \sum_{j=1}^q h_n^j \psi_j(h_n J_n) G_n^{(j-1)}(t_n) + \delta_{n+1}^{[q]},$$

with

$$(3.13) \quad \psi_j(z) = \varphi_j(z) - \sum_{k=1}^s b_k(z) \frac{c_k^{j-1}}{(j-1)!}$$

and remainders $\delta_{n+1}^{[q]}$ satisfying

$$(3.14) \quad \|\delta_{n+1}^{[q]}\| \leq Ch_n^{q+1}.$$

Again, small defects are desirable. Due to (2.6), we have $\psi_1(z) \equiv 0$. To obtain higher order bounds for δ_{n+1} , first observe that the h^2 -term in (3.12) is small due to (3.10). Additional terms vanish if $\psi_j = 0, j \geq 3$.

All conditions encountered so far are collected in Table 3.1. They will later turn out to be the order conditions for methods up to order 4.

LEMMA 3.1. *If the order conditions of Table 3.1 are satisfied up to order $p \leq 4$, we obtain*

$$(3.15) \quad \|\delta_{n+1}\| \leq Ch_n^2 \|e_n\| + Ch_n^{p+1}.$$

Proof. This at once follows from (3.12). \square

3.2. Preliminary error bounds. Let

$$e_n = u_n - u(t_n) \quad \text{and} \quad E_{ni} = U_{ni} - u(t_n + c_i h_n)$$

denote the differences between the numerical solution and the exact solution. Subtracting (3.6) from the numerical method (2.3) gives the error recursion

$$(3.16a) \quad E_{ni} = e^{c_i h_n J_n} e_n + h_n \sum_{j=1}^{i-1} a_{ij}(h_n J_n) (g_n(U_{nj}) - G_n(t_n + c_j h_n)) - \Delta_{ni},$$

$$(3.16b) \quad e_{n+1} = e^{h_n J_n} e_n + h_n \sum_{i=1}^s b_i(h_n J_n) (g_n(U_{ni}) - G_n(t_n + c_i h_n)) - \delta_{n+1}.$$

We will derive bounds for these errors.

LEMMA 3.2. *Under Assumption C.2, we have*

$$(3.17a) \quad \|g_n(U_{ni}) - G_n(t_n + c_i h_n)\| \leq C (h_n + \|e_n\| + \|E_{ni}\|) \|E_{ni}\|,$$

$$(3.17b) \quad \|g_n(u_n) - G_n(t_n)\| \leq C \|e_n\|^2,$$

$$(3.17c) \quad \left\| \frac{\partial g_n}{\partial u}(u(t_n)) \right\|_{X \leftarrow X} \leq C \|e_n\|,$$

as long as the errors E_{ni} and e_n remain in a sufficiently small neighborhood of 0.

Proof. The last bound (3.17c) is a direct consequence of the linearization and the Lipschitz condition (3.5). Using Taylor series expansion, we get

$$\begin{aligned} g_n(U_{ni}) - G_n(t_n + c_i h_n) &= \frac{\partial g_n}{\partial u}(u(t_n + c_i h_n)) E_{ni} \\ &\quad + \int_0^1 (1 - \tau) \frac{\partial^2 g_n}{\partial u^2}(u(t_n + c_i h_n) + \tau E_{ni})(E_{ni}, E_{ni}) \, d\tau. \end{aligned}$$

Setting $i = 1$ at once proves (3.17b). To derive (3.17a), we expand the first term on the right-hand side once more at t_n and use the identity

$$\frac{\partial g_n}{\partial u}(u(t_n)) = - \int_0^1 \frac{\partial^2 g_n}{\partial u^2}(u(t_n) + \tau e_n) e_n \, d\tau.$$

This finally proves (3.17a). \square

Using this result, we can establish an error bound for the internal stages.

LEMMA 3.3. *Under Assumptions C.1 and C.2, we have*

$$\|E_{ni}\| \leq C \|e_n\| + C h_n^3,$$

as long as the global errors e_n remain in a bounded neighborhood of 0.

Proof. The assertion at once follows from (3.16a), Lemma 3.2, and (3.11). \square

3.3. Stability bounds. In order to establish convergence bounds, we have to solve recursion (3.16b). For this purpose, stability bounds for the discrete evolution operators are crucial. In a first step, we will show stability along the exact solution.

We commence with two auxiliary results.

LEMMA 3.4. *Let the initial value problem (3.3) satisfy Assumptions C.1 and C.2, and let $\widehat{J}_n = DF(u(t_n))$. Then, for any $\tilde{\omega} > \omega$, there exists a constant C_L independent of h_{n-1} such that*

$$(3.18) \quad \left\| e^{t\widehat{J}_n} - e^{t\widehat{J}_{n-1}} \right\|_{X \leftarrow X} \leq C_L h_{n-1} e^{\tilde{\omega}t}, \quad t \geq 0.$$

Proof. Applying the variation-of-constants formula to the initial value problem

$$v'(t) = \widehat{J}_n v(t) = \widehat{J}_{n-1} v(t) + \left(\widehat{J}_n - \widehat{J}_{n-1} \right) v(t)$$

shows the representation

$$(3.19) \quad e^{t\widehat{J}_n} - e^{t\widehat{J}_{n-1}} = \int_0^1 t e^{(1-\sigma)t\widehat{J}_{n-1}} \left(\widehat{J}_n - \widehat{J}_{n-1} \right) e^{\sigma t\widehat{J}_n} \, d\sigma.$$

The required estimate now follows from (3.5) and the smoothness of $u(t)$. \square

LEMMA 3.5. *Under the assumptions of Lemma 3.4, the relation*

$$(3.20) \quad \|x\|_n = \sup_{t \geq 0} e^{-\tilde{\omega}t} \left\| e^{t\hat{J}_n} x \right\|, \quad x \in X$$

defines for any $n = 0, 1, 2, \dots$ a norm on X . This norm is equivalent to $\|\cdot\|$ and satisfies the bound

$$(3.21) \quad \|x\|_n \leq (1 + C_L h_{n-1}) \|x\|_{n-1}, \quad n \geq 1.$$

Proof. Obviously, we have $\|x\| \leq \|x\|_n$. On the other hand, the bound (3.2) yields $\|x\|_n \leq C \|x\|$. Thus, the two norms are equivalent.

For arbitrary $x \in X$, we have

$$\begin{aligned} \|x\|_n &= \sup_{t \geq 0} e^{-\tilde{\omega}t} \left\| \left(e^{t\hat{J}_n} - e^{t\hat{J}_{n-1}} + e^{t\hat{J}_{n-1}} \right) x \right\| \\ &\leq \|x\|_{n-1} + \sup_{t \geq 0} e^{-\tilde{\omega}t} \left\| e^{t\hat{J}_n} - e^{t\hat{J}_{n-1}} \right\|_{X \leftarrow X} \|x\| \\ &\leq (1 + C_L h_{n-1}) \|x\|_{n-1} \end{aligned}$$

by Lemma 3.4 and the equivalence of the norms. \square

The following lemma proves the stability of the discrete evolution operators along the exact solution.

LEMMA 3.6. *Under the assumptions of Lemma 3.4, there exists a constant C such that*

$$(3.22) \quad \left\| e^{h_n \hat{J}_n} \dots e^{h_0 \hat{J}_0} \right\|_{X \leftarrow X} \leq C e^{\Omega(h_0 + \dots + h_n)},$$

with $\Omega = C_L + \tilde{\omega}$.

Proof. By (3.20) and Lemma 3.5, we have

$$\begin{aligned} \left\| e^{h_n \hat{J}_n} \dots e^{h_0 \hat{J}_0} x \right\|_n &= \sup_{t \geq 0} \left\| e^{-\tilde{\omega}t} e^{t\hat{J}_n} e^{-\tilde{\omega}h_n} e^{\tilde{\omega}h_n} e^{h_n \hat{J}_n} \dots e^{h_0 \hat{J}_0} x \right\| \\ &\leq \sup_{t \geq 0} \left\| e^{-\tilde{\omega}t} e^{t\hat{J}_n} e^{\tilde{\omega}h_n} e^{h_{n-1} \hat{J}_{n-1}} \dots e^{h_0 \hat{J}_0} x \right\| \\ &= e^{\tilde{\omega}h_n} \left\| e^{h_{n-1} \hat{J}_{n-1}} \dots e^{h_0 \hat{J}_0} x \right\|_n \\ &\leq e^{\tilde{\omega}h_n} (1 + C_L h_{n-1}) \left\| e^{h_{n-1} \hat{J}_{n-1}} \dots e^{h_0 \hat{J}_0} x \right\|_{n-1}. \end{aligned}$$

Thus, the estimate $1 + C_L h_{n-1} \leq e^{C_L h_{n-1}}$ together with an induction argument proves the lemma. \square

We now turn our attention to the operators $J_n = DF(u_n)$ that result from the linearization process (2.2). These operators constitute an essential component of the numerical scheme (2.3). The triangle inequality shows that

$$(3.23) \quad \|u_n - u_{n-1}\| \leq C h_{n-1} + \|e_n\| + \|e_{n-1}\|.$$

We now repeat the above estimations with J_n in the role of \hat{J}_n and, in particular, use (3.23) in the proof of Lemma 3.4. This gives the following stability result for the discrete evolution operators on X .

THEOREM 3.7. *Let the initial value problem (3.3) satisfy Assumptions C.1 and C.2. Then, for any $\tilde{\omega} > \omega$, there exist constants C and C_E such that*

$$(3.24) \quad \left\| e^{h_n J_n} \dots e^{h_0 J_0} \right\|_{X \leftarrow X} \leq C e^{\Omega(h_0 + \dots + h_n) + C_E \sum_{j=1}^n \|e_j\|},$$

with $\Omega = C_L + \tilde{\omega}$. The bound holds as long as the numerical solution u_n stays in a sufficiently small neighborhood of the exact solution of (3.3).

The stability bound (3.24) requires some attention. Strictly speaking, stability is only guaranteed if the term $\sum_{j=1}^n \|e_j\|$ is uniformly bounded in n for $t_0 \leq t_n \leq T$. This condition can be considered as a (weak) restriction on the employed step size sequence; see the discussion in section 4 below.

4. Error bounds. We are now ready to present the main result of our paper. We will show that the conditions of Table 3.1 are sufficient to obtain convergence up to order 4 under a mild restriction on the employed step size sequence.

THEOREM 4.1. *Let the initial value problem (3.3) satisfy Assumptions C.1 and C.2. Consider for its numerical solution an explicit exponential Rosenbrock method (2.3) that fulfills the order conditions of Table 3.1 up to order p for some $2 \leq p \leq 4$. Further, let the step size sequence h_j satisfy the condition*

$$(4.1) \quad \sum_{k=1}^{n-1} \sum_{j=0}^{k-1} h_j^{p+1} \leq C_H,$$

with a constant C_H that is uniform in $t_0 \leq t_n \leq T$. Then, for C_H sufficiently small, the numerical method converges with order p . In particular, the numerical solution satisfies the error bound

$$(4.2) \quad \|u_n - u(t_n)\| \leq C \sum_{j=0}^{n-1} h_j^{p+1}$$

uniformly on $t_0 \leq t_n \leq T$. The constant C is independent of the chosen step size sequence satisfying (4.1)

Proof. From (3.16b), we obtain the error recursion

$$(4.3) \quad e_{n+1} = e^{h_n J_n} e_n + h_n \varrho_n - \delta_{n+1}, \quad e_0 = 0,$$

with

$$\varrho_n = \sum_{i=1}^s b_i(h_n J_n) (g_n(U_{ni}) - G_n(t_n + c_i h_n)).$$

Solving this recursion and using $e_0 = 0$ yields

$$(4.4) \quad e_n = \sum_{j=0}^{n-1} h_j e^{h_{n-1} J_{n-1}} \dots e^{h_{j+1} J_{j+1}} (\varrho_j - h_j^{-1} \delta_{j+1}).$$

Employing Lemmas 3.1, 3.2, and 3.3, we obtain the bound

$$(4.5) \quad \|\varrho_j\| + h_j^{-1} \|\delta_{j+1}\| \leq C \left(h_j \|e_j\| + \|e_j\|^2 + h_j^p \right).$$

Inserting this into (4.4) and using the stability estimate (3.24) yields

$$(4.6) \quad \|e_n\| \leq C \sum_{j=0}^{n-1} h_j \left(\|e_j\|^2 + h_j \|e_j\| + h_j^p \right).$$

The constant in this estimate is uniform as long as

$$(4.7) \quad \sum_{j=1}^{n-1} \|e_j\| \leq C_A$$

uniformly holds on $t_0 \leq t_n \leq T$. The application of a discrete Gronwall lemma to (4.6) then shows the desired bound (4.2).

It still remains to verify that condition (4.7) holds with a uniform bound C_A . This follows now recursively from (4.2) and our assumption on the step size sequence (4.1) with C_H sufficiently small. \square

In the remainder of this section, we discuss the encountered restriction (4.1) on the step size sequence. For *constant* step sizes, this condition evidently holds with

$$C_H = \frac{1}{2} h^{p-1} (t_n - t_0)^2.$$

Since $p \geq 2$, the size of C_H tends to zero for $h \rightarrow 0$.

A similar bound holds for *quasi-uniform* step size sequences where the ratio between the maximal and minimal step length is uniformly bounded. For sequences with increasing step sizes, condition (4.1) is fulfilled as well.

In practice, a problem with (4.1) might occur if the step size suddenly drops by several orders of magnitude. In that case, however, it is possible to modify the above stability analysis and to relax the condition on the step sizes. We briefly explain the idea, but we do not work out all details. If the error at time t_j , say, is large compared to the actual step length, one should rather compare the numerical solution with a smooth trajectory that passes close to u_j . Although u_j might be a nonsmooth initial value, such trajectories exist. Then the previous stability proof can be applied once more, at the possible price of increasing the constant C in (3.23) and thus the constants C_L and Ω . As long as this is done only a fixed number of times, stability in (3.24) is still guaranteed.

5. Methods of order up to four. The well-known exponential Rosenbrock–Euler method is given by

$$(5.1) \quad \begin{aligned} u_{n+1} &= e^{h_n J_n} u_n + h_n \varphi_1(h_n J_n) g_n(u_n) \\ &= u_n + h_n \varphi_1(h_n J_n) F(u_n). \end{aligned}$$

It is computationally attractive, since it requires only one matrix function per step. The method obviously satisfies condition 1 of Table 3.1, while condition 2 is void. Therefore, it is second-order convergent for problems satisfying our analytic framework. A possible error estimator for (5.1) is described in [3].

From the order conditions of Table 3.1, it is straightforward to construct pairs of embedded methods of order 3 and 4. For our variable step size implementation, we consider (2.3b) together with an embedded approximation

$$(5.2) \quad \hat{u}_{n+1} = e^{h_n J_n} u_n + h \sum_{i=1}^s \hat{b}_i(h J_n) g_n(U_{ni}),$$

which relies on the same stages U_{ni} . The methods given below were first introduced in [15]. They will be used in the numerical experiments in section 7.

The scheme `exprb32` consists of a third-order exponential Rosenbrock method with a second-order error estimator (the exponential Rosenbrock–Euler method). Its coefficients are

$$\begin{array}{c|cc} c_1 & & \\ c_2 & a_{21} & \\ \hline & b_1 & b_2 \\ & \widehat{b}_1 & \end{array} = \begin{array}{c|cc} 0 & & \\ 1 & \varphi_1 & \\ \hline & \varphi_1 - 2\varphi_3 & 2\varphi_3 \\ & \varphi_1 & \end{array}$$

The scheme `exprb43` is a fourth-order method with a third-order error estimator. Its coefficients are

$$\begin{array}{c|ccc} c_1 & & & \\ c_2 & a_{21} & & \\ c_3 & a_{31} & a_{32} & \\ \hline & b_1 & b_2 & b_3 \\ & \widehat{b}_1 & \widehat{b}_2 & \widehat{b}_3 \end{array} = \begin{array}{c|ccc} 0 & & & \\ \frac{1}{2} & \frac{1}{2}\varphi_1(\frac{1}{2}\cdot) & & \\ 1 & 0 & & \varphi_1 \\ \hline & \varphi_1 - 14\varphi_3 + 36\varphi_4 & 16\varphi_3 - 48\varphi_4 & -2\varphi_3 + 12\varphi_4 \\ & \varphi_1 - 14\varphi_3 & 16\varphi_3 & -2\varphi_3 \end{array}$$

Note that the internal stages of the above methods are just exponential Rosenbrock–Euler steps. This leads to simple methods that can cheaply be implemented.

Evidently, the order conditions of Table 3.1 imply that the weights of any third-order method have to depend on φ_3 , whereas that of any fourth-order method depend on φ_3 and φ_4 (in addition to φ_1).

6. Nonautonomous problems. The proposed method can easily be extended to nonautonomous problems

$$(6.1) \quad u' = F(t, u), \quad u(t_0) = u_0$$

by rewriting the problem in autonomous form:

$$(6.2a) \quad U' = \mathcal{F}(U), \quad U = \begin{bmatrix} t \\ u \end{bmatrix}, \quad \mathcal{F}(U) = \begin{bmatrix} 1 \\ F(t, u) \end{bmatrix},$$

with Jacobian

$$(6.2b) \quad \mathcal{J}_n = \begin{bmatrix} 0 & 0 \\ v_n & J_n \end{bmatrix}, \quad v_n = \frac{\partial}{\partial t}F(t_n, u_n), \quad J_n = \frac{\partial}{\partial u}F(t_n, u_n).$$

This transformation is standard for Rosenbrock methods as well (see [10]), but it changes a linear nonautonomous problem into a nonlinear one.

In order to apply our method to the autonomous system (6.2), we have to compute the matrix functions of \mathcal{J}_n . Using Cauchy’s integral formula and exploiting the special structure of \mathcal{J} , we get

$$\varphi(h\mathcal{J}) = \begin{bmatrix} \varphi(0) & 0 \\ h\widehat{\varphi}(hJ)v & \varphi(hJ) \end{bmatrix}, \quad \widehat{\varphi}(z) = \frac{\varphi(z) - \varphi(0)}{z}.$$

For the particular functions in our method, we obtain from (2.5) the relation

$$(6.3) \quad \widehat{\varphi}_i(hJ) = \varphi_{i+1}(hJ).$$

In our formulation, we will work again with the smaller quantities

$$(6.4) \quad D_{nj} = g_n(t_n + c_j h_n, U_{nj}) - g_n(t_n, u_n),$$

where

$$g_n(t, u) = F(t, u) - J_n u - v_n t.$$

Applying method (2.7) to the autonomous formulation (6.2), we get

$$(6.5a) \quad \begin{aligned} U_{ni} &= u_n + h_n c_i \varphi_1(c_i h_n J_n) F(t_n, u_n) \\ &+ h_n^2 c_i^2 \varphi_2(c_i h_n J_n) v_n + h_n \sum_{j=2}^{i-1} a_{ij}(h_n J_n) D_{nj}, \end{aligned}$$

$$(6.5b) \quad u_{n+1} = u_n + h_n \varphi_1(h_n J_n) F(t_n, u_n) + h_n^2 \varphi_2(h_n J_n) v_n + h_n \sum_{i=2}^s b_i(h_n J_n) D_{ni}.$$

This is the format of an exponential Rosenbrock method for nonautonomous problems (6.1).

7. Numerical experiments. We have implemented the exponential Rosenbrock methods `exprb32` and `exprb43` in MATLAB with adaptive time stepping. We employ a standard step size selection strategy based on the local error [10, pp. 28–31]. The error is estimated with the help of the corresponding embedded method from section 5. Our implementation involves two different options for dealing with the matrix φ -functions: For small examples, we employ diagonalization or Padé approximation for the explicit computation of the matrix functions. For large problems, Krylov subspace methods are used for approximating the product of the matrix functions with the corresponding vectors. For autonomous problems, we use the reformulation (2.7), which requires one Krylov subspace with the vector $F(u_n)$ and $s - 1$ Krylov subspaces with the vectors D_{ni} , $i = 2, \dots, s$. Due to $\|D_{ni}\| = \mathcal{O}(h_n^2)$, these approximations can be computed in very low dimensional subspaces. For nonautonomous problems, the format (6.5) requires one additional Krylov subspace with the vector v_n . Since the term involving v_n is multiplied with h_n^2 (compared to h_n for the other vectors), this subspace will be low-dimensional as well.

Example 7.1. As a first example we consider a two-dimensional advection-diffusion-reaction equation for $u = u(x, y, t)$:

$$(7.1) \quad \partial_t u = \varepsilon(\partial_{xx} u + \partial_{yy} u) - \alpha(u_x + u_y) + \gamma u \left(u - \frac{1}{2}\right) (1 - u), \quad (x, y) \in (0, 1)^2,$$

with homogeneous Neumann boundary conditions and the initial value

$$u(x, y, 0) = 256((1 - x)x(1 - y)y)^2 + 0.3,$$

where $\varepsilon = 1/100$, $\alpha = -10$, and $\gamma = 100$. The spatial discretization was done with finite differences using 101 grid points in each direction.

This example is taken from [3], where FORTRAN implementations of `exprb43`, combined with the real Leja point method [4], and of the Runge–Kutta–Chebyshev method RKC from [25] were compared. Here, we compare MATLAB implementations of RKC, `exprb43`, `exp4` from [12], and Krogstad’s method [17]. The latter three make use of Krylov subspace approximations. To improve the efficiency of the Krogstad

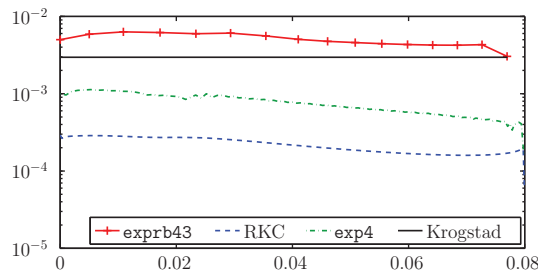


FIG. 7.1. Step sizes for the advection-diffusion-reaction equation (7.1) for $t \in [0, 0.08]$

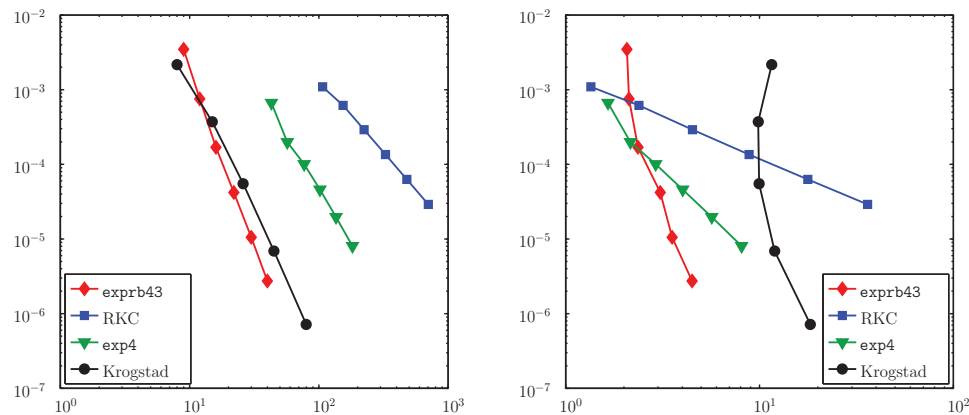


FIG. 7.2. Number of time steps versus accuracy (left) and CPU time versus accuracy (right) for the advection-diffusion-reaction example (7.1) for $t = 0.08$

method, we reused information from previously computed Krylov subspaces, an approach proposed in [13]. Since an adaptive step size control based on embedding is not possible for Krogstad's method, we ran this method with constant step size. For this particular example, the step size control of the other schemes also lead to almost constant steps sizes; see Figure 7.1. All simulations achieved a final accuracy of about 0.004 at $t = 0.08$. It can be seen that, due to the large advection part, the exponential methods can take much larger steps than RKC with `exprb43` taking the largest ones. In total, `exprb43` takes only 18 steps, Krogstad's method takes 27 steps, `exp4` takes 119 steps, while RKC uses 383 steps.

In Figure 7.2, we compare the performance of the Krylov implementations of `exp4`, `exprb43`, and Krogstad's method with a MATLAB implementation of RKC. Our implementation of RKC is based on the well-established FORTRAN code by Sommeijer available from the `netlib` repository. Our implementations of `exp4` and `exprb43` allow a maximum dimension of the Krylov subspaces of 36, which is the default value suggested in [12]. The codes were run with tolerances $ATOL = RTOL = 10^{-4}, 10^{-4.5}, \dots, 10^{-6.5}$ (except for Krogstad's method, which was used with constant step size). In the left diagram, we plot the achieved accuracy as a function of the required number of steps. It turns out that, for a given accuracy, the exponential Rosenbrock method `exprb43` uses significantly larger time steps than `exp4` and RKC. The number of time steps required for Krogstad's method is about the same as for `exprb43`.

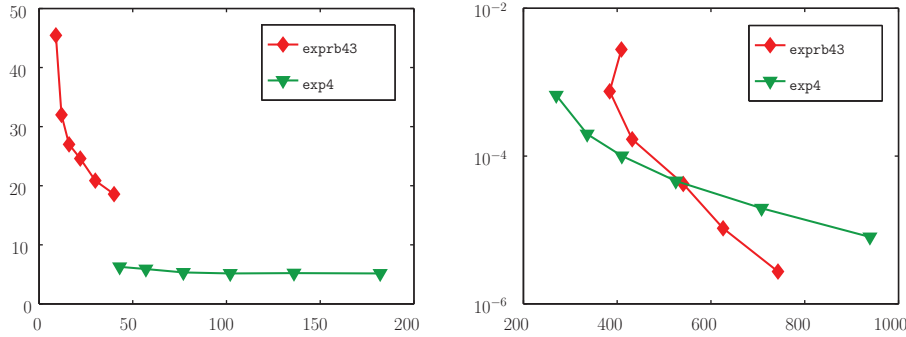


FIG. 7.3. Number of time steps versus average number of Krylov steps (left) and number of Krylov steps versus accuracy (right) for the advection-diffusion-reaction example (7.1) for $t = 0.08$

However, the efficiency of a code should also take the cost per time step into account. Therefore, we next consider the CPU time required to achieve a certain accuracy. We are fully aware of the fact that comparing CPU times strongly depends on the available computer architecture, the implementation, and the programming language. Nevertheless, we think that MATLAB comparisons might be of interest.

In Figure 7.2 we show the achieved accuracy as a function of the required CPU time. It can be seen that for moderate tolerances, **exp4** is faster than **exprb43** while for more stringent tolerances, **exprb43** requires less CPU time. This can be explained by considering the number of Krylov steps used by these methods. In the left diagram in Figure 7.3, we plotted the average number of Krylov steps over the total number of time steps. Since **exprb43** uses significantly larger time steps, we know from the convergence analysis of Krylov subspace methods [7, 11] that this requires more Krylov steps. The right diagram of Figure 7.3 shows the achieved accuracy versus the total number of Krylov steps. Since the Krylov approximations dominate the computational cost, this explains the right diagram of Figure 7.2. Note that it is impossible to give a reformulation of Krogstad’s method in such a way that only one expensive Krylov subspace is required in each step. The gain achieved by reusing previously computed Krylov subspaces [13] does not compensate for this disadvantage. Moreover, Krogstad’s method has four stages and uses even more matrix functions than **exprb43**.

Example 7.2. As a second example, we consider the one-dimensional Schrödinger equation [12] for $\psi = \psi(x, t)$:

$$(7.2a) \quad i \frac{\partial \psi}{\partial t} = H(x, t)\psi,$$

with the time-dependent Hamiltonian

$$(7.2b) \quad H(x, t) = -\frac{1}{2} \frac{\partial^2}{\partial x^2} + \kappa \frac{x^2}{2} + \mu(\sin t)^2 x .$$

We used the parameter values $\kappa = 10$ and $\mu = 100$. The initial value was chosen as $\psi(x, 0) = e^{-\sqrt{\kappa}x^2/2}$, which corresponds to the ground state of the unforced harmonic oscillator. Semidiscretization in space was done by a pseudospectral method with 512 Fourier modes on the interval $[-10, 10]$ with periodic boundary conditions.

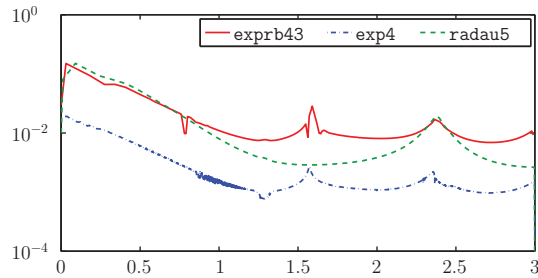


FIG. 7.4. Step sizes taken by `exp4`, `radau5`, and `exprb43` for the laser example (7.2) for $t \in [0, 3]$.

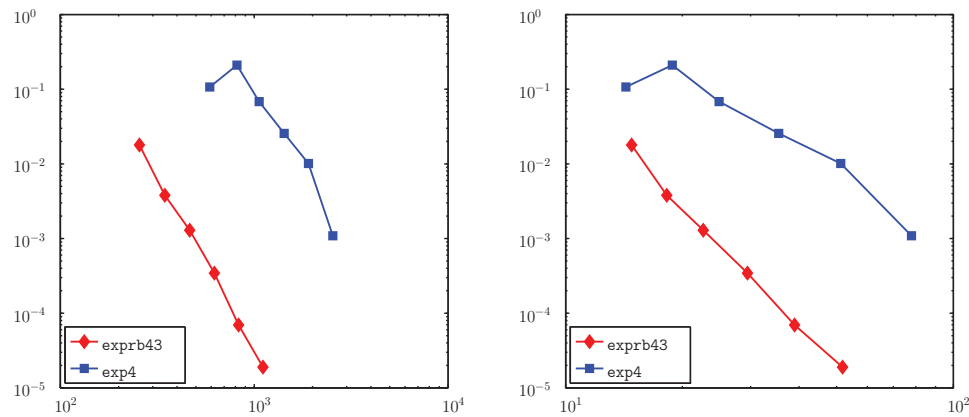


FIG. 7.5. Number of time steps versus accuracy (left) and CPU time versus accuracy (right) for the laser example (7.2) for $t = 3$.

It was shown in [12] that the MATLAB implementation of `exp4` outperforms MATLAB's standard nonstiff `ode45` method and matrix-free implementations of the stiff solvers `radau5` and `ode15s`. We refer to [12] for details. Here, we use exactly the same spatial discretization but run the simulation until $t = 3$.

In Figure 7.4, we display the step sizes chosen by the adaptive step size control for `exp4`, `radau5`, and `exprb43`. The tolerances were set in such a way that all methods achieved a final accuracy of about 0.05. As illustrated in Figure 7.4, `exprb43` advances with larger step sizes than the other two methods. In total, `exprb43` uses 256 steps, `exp4` uses 1906 steps, and `radau5` uses 537 steps. In our implementation of `radau5`, the linear systems arising within the Newton iteration are solved directly, while `exp4` and `exprb43` are used with Krylov subspace approximations. The direct solution of the linear systems arising in the `radau5` code result in a total CPU time which is more than 10 times longer than `exprb43`. Since it has been shown in [12] that a much more efficient W-version of `radau5` was still slower than `exp4`, we did not include `radau5` into our runtime comparisons.

In Figure 7.5, we compare the performance of the Krylov implementations of `exp4` and `exprb43`. Both codes were run with tolerances $ATOL = RTOL = 10^{-4}, 10^{-4.5}, \dots, 10^{-6.5}$. The diagrams show that the exponential Rosenbrock method `exprb43` uses significantly larger step sizes than `exp4`. Moreover, it is also much faster in terms of total CPU time.

8. Concluding remarks. In this paper, we have analyzed the convergence properties of exponential Rosenbrock-type methods in an abstract framework of C_0 semigroups. A local error analysis revealed the stiff order conditions, which in turn enabled us to construct methods of orders three and four with embedded error estimates of orders two and three, respectively. To control the error propagation, we derived stability bounds for variable step sizes. This enabled us to give a variable step size convergence proof. We implemented the methods in MATLAB, using Krylov subspace methods to approximate the applications of matrix functions to vectors. The numerical results clearly demonstrate the efficiency of the new integrators.

Appendix A. Analytic semigroups.

So far, we restricted our attention to strongly continuous semigroups. This framework, however, limits the class of possible nonlinearities due to Assumption C.2. If the semigroup is even analytic, we can allow more general nonlinearities. In this appendix, we sketch how to extend our analysis to this case. For the theoretical background of analytic semigroups, we refer to [8, 24].

Assumption A.1. The linear operator A in (3.3) is the generator of an analytic semigroup.

Without loss of generality, we can assume that A is invertible (otherwise, we shift it by an appropriate multiple of the identity). Therefore, fractional powers of A are well defined. We choose $0 \leq \alpha < 1$ and define $V = \mathcal{D}(A^\alpha) \subset X$. The linear space V is a Banach space with norm $\|v\|_V = \|A^\alpha v\|$.

Our basic assumptions on f are the following.

Assumption A.2. We suppose that (3.3) possesses a sufficiently smooth solution $u : [0, T] \rightarrow V$ with derivatives in V and that $f : V \rightarrow X$ is sufficiently often Fréchet differentiable in a strip along the exact solution. All occurring derivatives are supposed to be uniformly bounded.

A consequence of Assumption A.1 is that there exist constants C and ω such that

$$(A.1) \quad \|e^{tJ}\|_{V \leftarrow V} + \|t^\alpha e^{tJ}\|_{V \leftarrow X} \leq C e^{\omega t}, \quad t \geq 0$$

holds in a neighborhood of the exact solution.

With these assumptions at hand, we derive once more the bounds of section 3. Instead of (3.11), we now get

$$(A.2) \quad \|\Delta_{ni}\|_X + h_n^\alpha \|\Delta_{ni}\|_V \leq Ch_n^2 \|e_n\|_V + Ch_n^3,$$

and (3.15) is replaced by

$$(A.3) \quad \|\delta_{n+1}\|_X + h_n^\alpha \|\delta_{n+1}\|_V \leq Ch_n^2 \|e_n\|_V + Ch_n^{p+1}.$$

The same arguments as in the proofs of Lemma 3.2 and 3.3 show the following refined estimates.

LEMMA A.1. *Under Assumptions A.1 and A.2, we have*

$$(A.4a) \quad \|g_n(U_{ni}) - G_n(t_n + c_i h_n)\|_X \leq C (h_n + \|e_n\|_V + \|E_{ni}\|_V) \|E_{ni}\|_V,$$

$$(A.4b) \quad \|g_n(u_n) - G_n(t_n)\|_X \leq C \|e_n\|_V^2,$$

$$(A.4c) \quad \left\| \frac{\partial g_n}{\partial u}(u(t_n)) \right\|_{X \leftarrow V} \leq C \|e_n\|_V,$$

and

$$(A.4d) \quad \|E_{ni}\|_V \leq C \|e_n\|_V + Ch_n^{3-\alpha},$$

as long as the errors E_{ni} and e_n remain in a sufficiently small neighborhood of 0. \square

Further, Assumption A.2 implies

$$(A.5) \quad \left\| \widehat{J}_n - \widehat{J}_{n-1} \right\|_{X \leftarrow V} \leq Ch_{n-1}, \quad n \geq 1,$$

with a constant C that is independent of h_{n-1} . The same arguments as in the proof of Lemma 3.4 with (3.2) replaced by (A.1) now show that

$$(A.6) \quad \left\| e^{t\widehat{J}_n} - e^{t\widehat{J}_{n-1}} \right\|_{V \leftarrow V} \leq C_L h_{n-1} e^{\tilde{\omega}t}.$$

This implies the desired stability estimate in V . For the convergence proof, we need an additional stability result that reflects the parabolic smoothing.

LEMMA A.2. *Let the initial value problem (3.3) satisfy Assumptions A.1 and A.2, and let $\widehat{J}_n = DF(u(t_n))$. Then, for any $\tilde{\omega} > \omega$, there exists a constant C independent of h_{n-1} such that*

$$(A.7) \quad \left\| e^{h_n \widehat{J}_n} \dots e^{h_0 \widehat{J}_0} \right\|_{V \leftarrow X} \leq C \frac{e^{\Omega(h_0 + \dots + h_n)}}{(h_0 + \dots + h_n)^\alpha},$$

with $\Omega = C_L + \tilde{\omega}$ and C_L from (A.6).

Proof. Using the same arguments as in [22, section 5] shows this bound. \square

We are now in the position to state the convergence proof for exponential Rosenbrock methods in the framework of analytic semigroups. For notational simplicity, we formulate the result for constant step sizes only.

THEOREM A.3. *Let the initial value problem (3.3) satisfy Assumptions A.1 and A.2 and consider for its numerical solution an explicit exponential Rosenbrock method (2.3) with constant step size h . Assume that the order conditions of Table 3.1 hold up to order p with $p = 2$ or $p = 3$. Then, for h sufficiently small, the numerical method converges with order p . In particular, the numerical solution u_n satisfies the uniform error bound*

$$\|u_n - u(t_n)\|_V \leq Ch^p.$$

The constant C depends on T , but it is independent of n and h for $0 \leq nh \leq T - t_0$.

Proof. We proceed as in the proof of Theorem 4.1. Due to (A.3) and (A.4), we can bound

$$(A.8) \quad \|\varrho_n\|_X + h^{-1} \|\delta_{n+1}\|_X \leq C \left(h \|e_n\|_V + \|e_n\|_V^2 + h^p \right).$$

By the stability estimate, we now have

$$\|e_n\|_V \leq C \sum_{j=0}^{n-1} \frac{h}{(t_n - t_{j+1})^\alpha} \left(h \|e_j\|_V + \|e_j\|_V^2 + h^p \right).$$

The desired error bound thus follows from the application of a discrete Gronwall lemma with weakly singular kernel. \square

Remark. For $p \geq 4$, the analysis is much more delicate. Due to (A.4d), the bound (A.8) now contains a term of the order $h^{4-\alpha}$. Under additional assumptions on f , this order reduction can be avoided. For exponential Runge–Kutta methods, this has been detailed in [14]. We do not elaborate this point here.

REFERENCES

- [1] P.N. BROWN, G.D. BYRNE, AND A.C. HINDMARSH, *VODE: A variable-coefficient ODE solver*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 1038–1051.
- [2] G.D. BYRNE, *Pragmatic experiments with Krylov methods in the stiff ODE setting*, in Computational Ordinary Differential Equations, J.R. Cash and I. Gladwell, eds., Clarendon Press, Oxford, 1992, pp. 323–356.
- [3] M. CALIARI AND A. OSTERMANN, *Implementation of exponential Rosenbrock-type integrators*, Appl. Numer. Math., to appear.
- [4] M. CALIARI, M. VIANELLO, AND L. BERGAMASCHI, *Interpolating discrete advection-diffusion propagators at Leja sequences*, J. Comput. Appl. Math., 172 (2004), pp. 79–99.
- [5] M.P. CALVO AND C. PALENCIA, *A class of explicit multistep exponential integrators for semilinear problems*, Numer. Math., 102 (2006), pp. 367–381.
- [6] S.M. COX AND P.C. MATTHEWS, *Exponential time differencing for stiff systems*, J. Comput. Phys., 176 (2002), pp. 430–455.
- [7] V.L. DRUSKIN AND L.A. KNIZHNERMAN, *Krylov subspace approximations of eigenpairs and matrix functions in exact and computer arithmetic*, Numer. Linear Algebra Appl., 2 (1995), pp. 205–217.
- [8] K.-J. ENGEL AND R. NAGEL, *One-Parameter Semigroups for Linear Evolution Equations*, Springer, New York, 2000.
- [9] A. FRIEDLI, *Verallgemeinerte Runge–Kutta Verfahren zur Lösung steifer Differentialgleichungssysteme*, in Numerical Treatment of Differential Equations, Lect. Notes in Math. 631, R. Bulirsch, R.D. Grigorieff, and J. Schröder, eds., Springer, Berlin, 1978.
- [10] E. HAIRER AND G. WANNER, *Solving Ordinary Differential Equations II, Stiff and Differential-Algebraic Problems*, 2nd rev. ed., Springer, New York, 1996.
- [11] M. HOCHBRUCK AND CH. LUBICH, *On Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 34 (1997), pp. 1911–1925.
- [12] M. HOCHBRUCK, CH. LUBICH, AND H. SELHOFER, *Exponential integrators for large systems of differential equations*, SIAM J. Sci. Comput., 19 (1998), pp. 1552–1574.
- [13] M. HOCHBRUCK AND J. NIEHOFF, *Approximation of matrix operators applied to multiple vectors*, Math. Comput. Simulation, 79 (2008), pp. 1270–1283.
- [14] M. HOCHBRUCK AND A. OSTERMANN, *Explicit exponential Runge–Kutta methods for semilinear parabolic problems*, SIAM J. Numer. Anal., 43 (2005), pp. 1069–1090.
- [15] M. HOCHBRUCK AND A. OSTERMANN, *Exponential integrators of Rosenbrock-type*, Oberwolfach Rep. 3 (2006), pp. 1107–1110.
- [16] A.-K. KASSAM AND L.N. TREFETHEN, *Fourth-order time stepping for stiff PDEs*, SIAM J. Sci. Comput., 26 (2005), pp. 1214–1233.
- [17] S. KROGSTAD, *Generalized integrating factor methods for stiff PDEs*, J. Comput. Phys., 203 (2005), pp. 72–88.
- [18] J.D. LAWSON, *Generalized Runge–Kutta processes for stable systems with large Lipschitz constants*, SIAM J. Numer. Anal., 4 (1967), pp. 372–380.
- [19] CH. LUBICH AND A. OSTERMANN, *Linearly implicit time discretization of non-linear parabolic equations*, IMA J. Numer. Anal., 15 (1995), pp. 555–583.
- [20] CH. LUBICH AND A. OSTERMANN, *Runge–Kutta approximation of quasi-linear parabolic equations*, Math. Comp., 64 (1995), pp. 601–627.
- [21] F. MAZZIA (Coordinator), *Test Set for IVP Solvers*, <http://pitogora.dm.uniba.it/~testset>.
- [22] A. OSTERMANN AND M. THALHAMMER, *Convergence of Runge–Kutta methods for nonlinear parabolic equations*, Appl. Numer. Math., 42 (2002), pp. 367–380.
- [23] A. OSTERMANN, M. THALHAMMER, AND W. WRIGHT, *A class of explicit exponential general linear methods*, BIT, 46 (2006), pp. 409–431.
- [24] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, New York, 1983.
- [25] B. SOMMEIJER, L. SHAMPINE, AND J. VERWER, *RKC: An explicit solver for parabolic PDEs*, J. Comput. Appl. Math., 88 (1998), pp. 315–326.
- [26] K. STREHMEL AND R. WEINER, *B-convergence results for linearly implicit one step methods*, BIT, 27 (1987), pp. 264–281.
- [27] M. TOKMAN, *Efficient integration of large stiff systems of ODEs with exponential propagation iterative (EPI) methods*, J. Comput. Phys., 213 (2006), pp. 748–776.
- [28] R. WEINER, B.A. SCHMITT, AND H. PODHAISKY, *ROWMAP—a ROW-code with Krylov techniques for large stiff ODEs*, Appl. Numer. Math., 25 (1997), pp. 303–319.

HIGHER-ORDER FINITE ELEMENT METHODS AND POINTWISE ERROR ESTIMATES FOR ELLIPTIC PROBLEMS ON SURFACES*

ALAN DEMLOW†

Abstract. We define higher-order analogues to the piecewise linear surface finite element method studied in [G. Dziuk, “Finite elements for the Beltrami operator on arbitrary surfaces,” in *Partial Differential Equations and Calculus of Variations*, Springer-Verlag, Berlin, 1988, pp. 142–155] and prove error estimates in both pointwise and L_2 -based norms. Using the Laplace–Beltrami problem on an implicitly defined surface Γ as a model PDE, we define Lagrange finite element methods of arbitrary degree on polynomial approximations to Γ which likewise are of arbitrary degree. Then we prove a priori error estimates in the L_2 , H^1 , and corresponding pointwise norms that demonstrate the interaction between the “PDE error” that arises from employing a finite-dimensional finite element space and the “geometric error” that results from approximating Γ . We also consider parametric finite element approximations that are defined on Γ and thus induce no geometric error. Computational examples confirm the sharpness of our error estimates.

Key words. Laplace–Beltrami operator, surface finite element methods, a priori error estimates, boundary value problems on surfaces, pointwise and maximum norm error estimates

AMS subject classifications. 58J32, 65N15, 65N30

DOI. 10.1137/070708135

1. Introduction. The numerical solution of partial differential equations (PDEs) defined on surfaces arises naturally in many applications (cf. [CDR03], [CDD+04], [BMN05], [He06], and [DE07a], among many others). We consider the following model problem in order to focus on basic issues arising in the definition and analysis of such numerical methods. Let Γ be a smooth n -dimensional surface ($n = 2, 3$) without boundary embedded in \mathbb{R}^{n+1} . Let f be given data satisfying $\int_{\Gamma} f \, d\sigma = 0$ where $d\sigma$ is the surface measure, and let u solve

$$-\Delta_{\Gamma} u = f \text{ on } \Gamma.$$

Here Δ_{Γ} is the Laplace–Beltrami operator on Γ , and we require $\int_{\Gamma} u \, d\sigma = 0$ in order to guarantee uniqueness.

Several methods for defining suitable triangulations of Γ and corresponding finite element spaces have been proposed. For example, one may use the manifold structure of Γ (cf. [Ho01]) or a global parametric representation (cf. [AP05]) to triangulate Γ . In this work we focus on the method originally considered in [Dz88] in which Γ is represented as a level set of a smooth signed distance function d . In [Dz88], Γ is approximated by a polyhedral surface Γ_h having triangular faces, and the equations for defining a piecewise linear finite element approximation to u are conveniently defined and solved on Γ_h . This method has several advantages when compared with approaches relying on global or local parametrizations of Γ . These include its flexibility in handling various surfaces and its direct extension to problems in which the surface under consideration evolves in an unknown fashion and a parametrization is

*Received by the editors November 14, 2007; accepted for publication (in revised form) October 7, 2008; published electronically February 6, 2009. This material is based upon work partially supported under National Science Foundation grants DMS-0303378 and DMS-0713770.

<http://www.siam.org/journals/sinum/47-2/70813.html>

†Department of Mathematics, University of Kentucky, 715 Patterson Office Tower, Lexington, KY 40506-0027 (demlow@ms.uky.edu).

thus not available. The paradigm example of such an evolution problem is motion of a surface by mean curvature flow; cf. [Dz91], [DDE05].

In the present work we focus on two goals. The first is to define higher-order analogues to the surface finite element method defined in [Dz88]. Higher-order approximations are desirable in many situations because of their increased computational efficiency versus piecewise linear finite element methods. In order to obtain such approximations, it is generally necessary to approximate Γ to higher order in addition to employing higher-order finite element spaces. We thus construct parametric finite element spaces of arbitrary degree that are defined on arbitrary-degree polynomial approximations to Γ . In addition, we describe fully parametric finite element spaces defined directly on Γ via local transformations from the faces of Γ_h so that no error arises from approximating Γ . It should be noted that in both of these cases, we require explicit knowledge of the distance function d (either through an analytical formula or by a numerical approximation) in order to construct our algorithm.

Our second main goal is to carry out a thorough error analysis for finite element methods for the Laplace–Beltrami operator on surfaces. The original work of Dziuk in [Dz88] contains proofs of optimal-order convergence of the piecewise linear surface finite element method in the L_2 and energy norms. Here we prove optimal-order estimates for pointwise errors in function values and gradients and for local energy errors in addition to the L_2 and energy errors. These estimates are valid for arbitrary degrees of finite element spaces and polynomial approximations to Γ . As in [Dz88], we split the overall error into a “geometric error” arising from the approximation of Γ and a standard finite element “almost-best-approximation” error which arises from approximating an infinite-dimensional function space by a finite-dimensional finite element space. Roughly speaking, when employing finite element spaces of degree r on polynomial surface approximations of degree k , we have

$$\begin{aligned}\|\nabla_{\Gamma}(u - u_h)\|_{L_2(\Gamma)} &\leq Ch^r \|u\|_{H^{r+1}(\Gamma)} + Ch^{k+1} \|u\|_{H^1(\Gamma)}, \\ \|u - u_h\|_{L_2(\Gamma)} &\leq Ch^{r+1} \|u\|_{H^{r+1}(\Gamma)} + Ch^{k+1} \|u\|_{H^1(\Gamma)},\end{aligned}$$

where u_h is the finite element solution, ∇_{Γ} is the tangential gradient on Γ , and C depends on geometric properties of Γ . We also prove similar estimates in L_{∞} and W_{∞}^1 . As we verify via numerical experiments, one must thus choose $k + 1 \geq r$ to achieve optimal-order convergence in W_p^1 norms and $k \geq r$ to achieve optimal-order convergence in L_p norms.

We finally note that approximating Γ via higher-degree polynomials has the added benefit that the curvatures of the approximating surface Γ_h have a natural pointwise definition and converge to those of Γ . The availability of a simple curvature approximation is beneficial in applications where the weak form of the PDE under consideration, and thus also the finite element method, explicitly employs curvature information (as, for example, in the image processing application in [CDR03]). Curvature information also was used in the a posteriori error estimates given in [DD07]. However, pointwise curvatures are not naturally defined on the piecewise linear discrete surfaces employed in [Dz88], and ad-hoc reconstruction methods must be used to define suitable curvatures if they are explicitly required in calculations (cf. [CDR03]).

An outline of the paper is as follows. Section 2 contains definitions and preliminaries. In section 3 we prove abstract error estimates in various norms. In section 4, we demonstrate how these abstract estimates may be applied to various finite element methods on surfaces and give computational results illustrating the basic error behavior of the methods. In section 5 we give a brief discussion of conditions under which

our error analysis may be extended to more general classes of PDEs on surfaces and manifolds.

2. Preliminaries. In this section we record a number of preliminaries concerning geometry, transformations of functions between the continuous and discrete surfaces Γ and Γ_h , analytical results, and finite element approximation theory.

2.1. Geometric and analytical preliminaries on Γ . We assume throughout that Γ is a compact, oriented, C_∞ , two- or three-dimensional surface without boundary which is embedded in \mathbb{R}^3 or \mathbb{R}^4 , respectively. Our results may be extended to higher-dimensional surfaces of codimension one if appropriate results from finite element approximation theory can be proved; we restrict ourselves to lower-dimensional manifolds so that we may employ the Lagrange interpolant in our analysis.

Let d be the oriented distance function for Γ . For concreteness, let $d < 0$ on the interior of Γ and $d > 0$ on the exterior of Γ . $\vec{\nu} = \nabla d$ is then the outward-pointing unit normal, and $\mathbf{H} = \nabla^2 d$ is the Weingarten map. Here we express these quantities in the coordinates of the embedding space \mathbb{R}^{n+1} ($n = 2, 3$). For $x \in \Gamma$, the n eigenvalues $\kappa_1, \dots, \kappa_n$ of \mathbf{H} corresponding to eigenvectors perpendicular to $\vec{\nu}$ are the principal curvatures at x . Let $U \subset \mathbb{R}^{n+1}$ be a strip of width δ about Γ , where $\delta > 0$ is sufficiently small to ensure that the decomposition

$$a(x) = x - d(x)\vec{\nu}(x)$$

onto Γ is unique. We also require that $\delta < \min_{i=1, \dots, n} \frac{1}{\|\kappa_i\|_{L_\infty(\Gamma)}}$; cf. [GT98, section 14.6] and [DD07].

Let $\mathbf{P} = \mathbf{I} - \vec{\nu} \otimes \vec{\nu}$ be the projection onto the tangent plane at x , where \otimes is the outer product defined by $(\vec{a} \otimes \vec{b})\vec{c} = \vec{a}\vec{b} \cdot \vec{c}$. Then $\nabla_\Gamma = \mathbf{P}\nabla$ is the tangential gradient, $\text{div}_\Gamma = \nabla_\Gamma \cdot$ is the tangential divergence, and $\Delta_\Gamma = \text{div}_\Gamma \nabla_\Gamma$ is the Laplace–Beltrami operator. We shall use standard notation ($H^1(\Gamma)$, $W_p^j(\Gamma)$, etc.) for Sobolev spaces and norms of functions possessing j tangential derivatives lying in L_p .

Next we state some analytical results. Let

$$(2.1) \quad L(u, v) = \int_\Gamma \nabla_\Gamma u \nabla_\Gamma v \, d\sigma,$$

and let (\cdot, \cdot) be the L_2 inner product over Γ .

LEMMA 2.1. *Let $f \in L_2(\Gamma)$ satisfy $\int_\Gamma f \, d\sigma = 0$. Then the problem $L(u, v) = (f, v) \forall v \in H^1(\Gamma)$ has a unique weak solution u satisfying $\int_\Gamma u \, d\sigma = 0$, and*

$$(2.2) \quad \|u\|_{H_2^1(\Gamma)} \leq C\|f\|_{L_2(\Gamma)}.$$

Proof. See [Aub82, Chapter 4] for a proof of existence and uniqueness. Inequality (2.2) may be proved by local transformations to subsets of \mathbb{R}^n and a covering argument. \square

The proofs of our pointwise error estimates also rely on properties of the Green’s function. We denote by $\alpha(x, y)$ the surface distance between $x, y \in \Gamma$.

LEMMA 2.2. *There exists a function $G(x, y)$, unique up to a constant, such that for all functions $\phi \in C^2(\Gamma)$,*

$$\phi(x) = \frac{1}{|\Gamma|} \int_\Gamma \phi \, d\sigma + \int_\Gamma G(x, y)(-\Delta_\Gamma \phi(y)) \, d\sigma.$$

In addition, for $x, y \in \Gamma$ with $x \neq y$,

$$(2.3) \quad G(x, y) \leq \begin{cases} C(1 + \log \alpha(x, y)), & n = 2, \\ C\alpha(x, y)^{2-n}, & n > 2. \end{cases}$$

Also, let $|\gamma + \beta| > 0$, where γ and β are multi-indices. Then

$$(2.4) \quad |D_{\Gamma,y}^\gamma D_{\Gamma,x}^\beta G(x,y)| \leq C\alpha(x,y)^{2-n-|\gamma+\beta|}.$$

Proof. Existence of the Green's function G , (2.3), and (2.4) for $1 \leq |\alpha| \leq 2$ and $|\beta| = 0$ are contained in Theorem 4.13 of [Aub82]. Inequality (2.4) may be easily extended to arbitrary α, β with $|\alpha + \beta| > 0$ by using the representation (17) on p. 109 of [Aub82]. \square

Finally, let $\gamma_\Gamma > 0$ be the largest positive number such that all balls $B_{\gamma_\Gamma}(x_0) = \{x \in \Gamma : \alpha(x, x_0) < \gamma_\Gamma\}$ of radius γ_Γ map smoothly to domains in \mathbb{R}^n . Such a number γ_Γ exists since Γ is a smooth, compact surface.

2.2. The discrete surface Γ_h . Let $\Gamma_h \subset U$ be a polyhedron having triangular faces ($n = 2$) or a polytope having tetrahedral cells ($n = 3$) whose vertices lie on Γ and whose faces (cells) are shape-regular and quasi-uniform of diameter h . We shall denote by $\tilde{\mathcal{T}}_h$ the set of triangular faces of Γ_h and by \mathcal{T}_h the image under a of $\tilde{\mathcal{T}}_h$ (i.e., \mathcal{T}_h consists of curved simplices lying on Γ). Let $\vec{\nu}_h$ be the outward unit normal on Γ_h .

We will analyze finite element methods defined on Γ_h , on Γ , and on higher-order polynomial approximations of Γ , but Γ_h will play a central role in defining and analyzing all of them. From a programming standpoint in particular, Γ_h is fundamental to our methods in that the faces $\tilde{\mathcal{T}}_h$ of Γ_h always constitute the "base" triangulation of Γ , with parametric finite element spaces then being defined over $\tilde{\mathcal{T}}_h$.

2.3. Higher-order polynomial approximations to Γ . Next we describe a family Γ_h^k ($k \geq 1$) of polynomial approximations to Γ . The higher-order finite element spaces we use here are largely described in [He05] and also are similar to the surface element spaces described in [Ne76]. First let $\Gamma_h = \Gamma_h^1$ be a polyhedral approximation to Γ as in the preceding subsection. For $k \geq 2$ and for a given element $\tilde{T} \in \tilde{\mathcal{T}}_h$, let $\phi_1^k, \dots, \phi_{n_k}^k$ be the Lagrange basis functions of degree k on \tilde{T} corresponding to the nodal points x^1, \dots, x^{n_k} . For $x \in \tilde{T}$, we then define the discrete projection

$$a_k(x) = \sum_{j=1}^{n_k} a(x^j) \phi_j^k(x).$$

Employing the above definition on each element $\tilde{T} \in \tilde{\mathcal{T}}_h$ yields a continuous piecewise polynomial map on Γ_h . We then define the corresponding discrete surface

$$\Gamma_h^k = \{a_k(x) : x \in \Gamma_h\}.$$

Thus each component of a_k is the Lagrange interpolant of the corresponding component of the projection a restricted to Γ_h . Let $\hat{\mathcal{T}}_h^k$ be the image under a_k of $\tilde{\mathcal{T}}_h$, i.e., for $\hat{T} \in \hat{\mathcal{T}}_h^k$, $\hat{T} = a_k(\tilde{T})$ for some $\tilde{T} \in \tilde{\mathcal{T}}_h$. Let also \mathcal{T}_h^k be the image under a of $\hat{\mathcal{T}}_h^k$.

Next we discuss the computation of geometric quantities on Γ_h^k . Note first that Γ_h^k is defined *parametrically*, not *implicitly* as is Γ . Thus practical computation of geometric quantities such as normals and curvatures on Γ_h^k may involve somewhat different formulas than does computation of the corresponding quantities on Γ .

Let $\vec{\nu}_h^k$ be the (piecewise smooth) unit normal on Γ_h^k . In order to compute $\vec{\nu}_h^k$ in a practical situation, we let K be a unit simplicial reference element lying in \mathbb{R}^n . Let $\hat{T} \in \hat{\mathcal{T}}_h^k$ with $\hat{T} = a_k(\tilde{T})$ where $\tilde{T} \in \tilde{\mathcal{T}}_h$, and let $\mathbf{M} : K \rightarrow \hat{T}$ be an affine coordinate transformation with $\mathbf{M}(K) = \hat{T}$. A typical finite element code allows easy access to the quantities $\hat{a}_{k,x_1}, \dots, \hat{a}_{k,x_n}$, where x_1, \dots, x_n are the standard Euclidean

coordinates on K and $\hat{a}_k = a_k \circ \mathbf{M}$. $\vec{\nu}_h^k$ is then the outward-pointing unit vector that is perpendicular to $\hat{a}_{k,x_1}, \dots, \hat{a}_{k,x_n}$. If $n = 2$, we thus have for $x \in K$

$$(2.5) \quad \vec{\nu}_h^k(\hat{a}_k(x)) = \pm \frac{\hat{a}_{k,x_1}(x) \times \hat{a}_{k,x_2}(x)}{|\hat{a}_{k,x_1}(x) \times \hat{a}_{k,x_2}(x)|}.$$

One advantage of employing higher-order approximations to Γ is that in contrast to piecewise linear approximations, such surfaces have naturally defined pointwise curvatures. This information is explicitly needed in the weak (and thus finite element) formulations of various equations. Fix a point $\hat{a}_k(x) \in \Gamma_h^k$, where $x \in K$ with K and \hat{a}_k as above. The second fundamental form with respect to the basis $\{\hat{a}_{k,x_1}, \dots, \hat{a}_{k,x_n}\}$ of the tangent space $T_{\hat{a}_k(x)}$ is given by $II = [\hat{a}_{k,x_i x_j} \cdot \vec{\nu}_h^k]$, and the metric tensor is given by $G = [\hat{a}_{k,x_i} \cdot \hat{a}_{k,x_j}]$. The Weingarten map with respect to the basis $\{\hat{a}_{k,x_1}, \dots, \hat{a}_{k,x_n}\}$ is then $\mathbf{H}_{tan} = IIG^{-1}$. It is often desirable to express the Weingarten map with respect to the coordinates of the embedding space \mathbb{R}^{n+1} instead of with respect to the basis of the tangent space induced by \hat{a}_k . We thus compute

$$\mathbf{H}_h^k = \left[\hat{a}_{k,x_1} \dots \hat{a}_{k,x_n} \right] \mathbf{H}_{tan} \mathbf{P}_n \left[\hat{a}_{k,x_1} \dots \hat{a}_{k,x_n} \vec{\nu}_h^k \right]^{-1},$$

where \mathbf{P}_n is defined by $(x_1, \dots, x_n, x_{n+1}) \rightarrow (x_1, \dots, x_n)$. The principal curvatures and corresponding eigenbasis of the tangent space may be computed from \mathbf{H}_h^k . An alternative when $n = 2$ is to apply the formula $\mathbf{H}_h^k = \nabla_{\Gamma_h^k} \vec{\nu}_h^k$ to (2.5).

We now state results concerning the approximation of Γ by Γ_h^k .

PROPOSITION 2.3. *For h small enough, $\tilde{T} \in \tilde{\mathcal{T}}_h$, $\hat{T} \in \hat{\mathcal{T}}_h^k$, and $1 \leq i \leq k$,*

$$(2.6) \quad \|d\|_{L_\infty(\Gamma_h^k)} \leq \|a - a_k\|_{L_\infty(\Gamma_h)} \leq Ch^{k+1},$$

$$(2.7) \quad \|a - a_k\|_{W_\infty^i(\tilde{T})} \leq Ch^{k+1-i},$$

$$(2.8) \quad \|\vec{\nu} - \vec{\nu}_h^k\|_{L_\infty(\Gamma_h^k)} \leq Ch^k,$$

$$(2.9) \quad \|\mathbf{H} \circ a - \mathbf{H}_h^k\|_{L_\infty(\hat{T})} \leq Ch^{k-1}.$$

The constants C above depend upon the distance function d and its derivatives.

Proof. Inequalities (2.6) and (2.7) follow directly from the definition of a_k as the Lagrange interpolant of a and the definition of d (cf. [BS02] for standard results concerning finite element interpolation theory). To prove (2.8), consider a point $\hat{x} \in \Gamma_h^k$, where $\hat{x} = a_k(\tilde{x})$ for $\tilde{x} \in \tilde{T} \subset \Gamma_h$. Employing (2.6) and the smoothness of Γ , we have

$$\begin{aligned} |\vec{\nu}(\hat{x}) - \vec{\nu}_h^k(\hat{x})| &\leq |\vec{\nu}(a_k(\tilde{x})) - \vec{\nu}(a(\tilde{x}))| + |\vec{\nu}(a(\tilde{x})) - \vec{\nu}_h^k(a_k(\tilde{x}))| \\ &\leq C(\Gamma)h^{k+1} + |\vec{\nu}(a(\tilde{x})) - \vec{\nu}_h^k(a_k(\tilde{x}))|. \end{aligned}$$

Assuming without loss of generality that T lies in the x_1, \dots, x_n -hyperplane, we next note that $\vec{\nu}(a(\tilde{x}))$ is the outward-facing unit vector orthogonal to a_{x_1}, \dots, a_{x_n} and $\vec{\nu}_h^k(a_k(\tilde{x}))$ is the outward-facing unit vector orthogonal to $a_{k,x_1}, \dots, a_{k,x_n}$. From (2.7) we have $|a_{x_i} - a_{k,x_i}| \leq Ch^k$, and it is also not difficult to compute that $|a_{x_i}|$ is bounded from above and below independent of h for $1 \leq i \leq n$. Using these facts, one may then compute in an elementary fashion that $|\vec{\nu}(a(\tilde{x})) - \vec{\nu}_h^k(a_k(\tilde{x}))| \leq Ch^k$, for example, by using the Gram-Schmidt orthonormalization algorithm.

Inequality (2.9) may be proved in a similar fashion after noting that $\|a_{x_i x_j} - a_{k,x_i x_j}\|_{L_\infty(\hat{T})} \leq Ch^{k-1}$ for any element $\hat{T} \subset \Gamma_h^k$. \square

Remark 2.4. Because \mathbf{H}_h^k involves the second derivatives of a C^0 interpolant, it is only defined elementwise. However, for $k \geq 2$ a pointwise definition of \mathbf{H}_h^k on an element interface may be defined by taking the limit of \mathbf{H}_h^k as the interface is approached from any adjacent element. Stitching these elementwise approximations together yields a global, piecewise continuous curvature approximation with $O(h^{k-1})$ error. In particular, while \mathbf{H}_h^k viewed globally is a distribution with singular jump terms on element interfaces, it is not necessary to take these jump terms into account in order to obtain a convergent pointwise curvature approximation for higher-order discrete surfaces.

2.4. The correspondence between Γ_h , Γ_h^k , and Γ . Our analysis requires a number of relationships between functions defined on Γ and Γ_h^k , as in [Dz88] and [DD07]. In addition, proving approximation results for the parametric finite element spaces S_{hk}^r will require establishing similar relationships between functions defined on Γ_h^k and Γ_h .

We first establish relationships between functions defined on the continuous surface Γ and the discrete surfaces Γ_h^k . Let $v \in H^1(\Gamma)$ and define the extension $v^\ell(x) = v(a(x))$ for $x \in U$. For $v_h \in H^1(\Gamma_h^k)$ we define the lift $\tilde{v}_h \in H^1(\Gamma)$ by $\tilde{v}_h(a(\tilde{x})) = v_h(x)$, $\tilde{x} \in \Gamma_h$. For $v_h \in H^1(\Gamma_h^k)$, we then define the extension $v_h^\ell(x) = \tilde{v}_h(a(x))$ for any $x \in U$. Also, for $\hat{x} \in \Gamma_h^k$ let $\mu_{hk}(\hat{x})$ satisfy $\mu_{hk}(\hat{x}) d\sigma_{hk}(\hat{x}) = d\sigma(a(\hat{x}))$, where $d\sigma$ and $d\sigma_{hk}$ are surface measures on Γ and Γ_h^k , respectively.

PROPOSITION 2.5. *Let $x \in \Gamma_h^k$ and $n = 2, 3$. Then*

$$(2.10) \quad \mu_{hk}(\hat{x}) = \vec{\nu}(\hat{x}) \cdot \vec{\nu}_h^k(\hat{x}) \prod_{i=1}^n (1 - d(\hat{x})\kappa_i(\hat{x})).$$

Remark 2.6. For $x \in U$, $\kappa_i(x) = \frac{\kappa_i(a(x))}{1 + d(x)\kappa_i(a(x))}$; cf. [GT98], [DD07].

Proof. Equation (2.10) is proved in [DD07] for $n = 2$ using properties of the cross product, so we sketch a proof for $n = 3$. Let $\hat{T} \subset \mathbb{R}^n$ be a reference simplex. Let also $f = a_k \circ L: \hat{T} \rightarrow \tilde{T} \subset \Gamma_h^k$, where $\tilde{T} = a_k(\bar{T})$ for $\bar{T} \in \tilde{T}_h$ and $L: \hat{T} \rightarrow \bar{T}$ is one of the obvious natural linear transformations. Let f have Jacobian $\mathbf{F} \in \mathbb{R}^{(n+1) \times n}$ with singular values $\sigma_1, \dots, \sigma_n$ and singular value decomposition $\mathbf{F} = \mathbf{U}\Sigma\mathbf{V}^T$. Here \mathbf{U} has orthonormal columns $u_1, \dots, u_n, \vec{\nu}_h^k$, $\Sigma \in \mathbb{R}^{(n+1) \times n}$, and $\mathbf{V} \in \mathbb{R}^{n \times n}$ is orthogonal.

Let dx be a Lebesgue measure on \hat{T} . First we compute $d\sigma_{hk} = |\prod_{i=1}^n \sigma_i| dx$ and $d\sigma = |\det[(\mathbf{P} - d\mathbf{H})\mathbf{F} \vec{\nu}]| dx = |\prod_{i=1}^n (1 - d\kappa_i)| |\det[\mathbf{P}\mathbf{F} \vec{\nu}]| dx$. But $|\det[\mathbf{P}\mathbf{F} \vec{\nu}]| = \sqrt{\det \mathbf{F}^T \mathbf{P} \mathbf{P} \mathbf{F}}$. For $n = 2, 3$, a short computation involving the singular value decomposition yields $\sqrt{\det \mathbf{F}^T \mathbf{P} \mathbf{P} \mathbf{F}} = \vec{\nu} \cdot \vec{\nu}_h^k |\prod_{i=1}^n \sigma_i|$, which completes the proof. \square

Next we state identities regarding tangential gradients on Γ , Γ_h , and Γ_h^k (cf. [Dz88], [DD07]). For $v_h \in H^1(\Gamma_h^k)$, $v \in H^1(\Gamma)$, and $\hat{x} \in \Gamma_h^k$,

$$(2.11) \quad \nabla_{\Gamma_h^k} v^\ell(\hat{x}) = [\mathbf{P}_{h,k}(\hat{x})][(\mathbf{I} - d\mathbf{H})(\hat{x})][\mathbf{P}(\hat{x})] \nabla_{\Gamma} v(a(\hat{x})),$$

$$(2.12) \quad \nabla_{\Gamma} v_h^\ell(a(\hat{x})) = [(\mathbf{I} - d\mathbf{H})(\hat{x})]^{-1} \left[\mathbf{I} - \frac{\vec{\nu}_h^k(\hat{x}) \otimes \vec{\nu}(\hat{x})}{\vec{\nu}_h^k(\hat{x}) \cdot \vec{\nu}(\hat{x})} \right] \nabla_{\Gamma_h^k} v_h(\hat{x}).$$

Here $\mathbf{P}_{h,k} = \mathbf{I} - \vec{\nu}_h^k \otimes \vec{\nu}_h^k$ is the projection onto the tangent space of $\Gamma_{h,k}$. Letting

$$(2.13) \quad \mathbf{A}_{\Gamma}(a(\hat{x})) = \frac{1}{\mu_{hk}(\hat{x})} \mathbf{P}(\hat{x}) [\mathbf{I} - d(\hat{x})\mathbf{H}(\hat{x})] \mathbf{P}_{h,k}(\hat{x}) [\mathbf{I} - d(\hat{x})\mathbf{H}(\hat{x})] \mathbf{P}(\hat{x})$$

for $\hat{x} \in \Gamma_h^k$, (2.11) also yields the integral equality

$$(2.14) \quad \int_{\Gamma_h^k} \nabla_{\Gamma_h^k} u_h \nabla_{\Gamma_h^k} v_h d\sigma_{hk} = \int_{\Gamma} \mathbf{A}_{\Gamma} \nabla_{\Gamma} u_h^\ell \nabla_{\Gamma} v_h^\ell d\sigma.$$

We also shall need to compare Sobolev norms of functions defined on Γ and Γ_h^k . Let $v \in W_p^j(\Gamma)$ with $j \geq 0$ and $1 \leq p < \infty$. Then there exist constants C_j depending on j and Γ such that for h small enough,

$$(2.15) \quad \frac{1}{C_0} \|v\|_{L_p(\Gamma)} \leq \|v^\ell\|_{L_p(\Gamma_h^k)} \leq C_0 \|v\|_{L_p(\Gamma)},$$

$$(2.16) \quad \frac{1}{C_1} \|\nabla_\Gamma v\|_{L_p(\Gamma)} \leq \|\nabla_{\Gamma_h^k} v^\ell\|_{L_p(\Gamma_h^k)} \leq C_1 \|\nabla_\Gamma v\|_{L_p(\Gamma)},$$

$$(2.17) \quad \|D_{\Gamma_h^k}^j v^\ell\|_{L_p(\Gamma_h^k)} \leq C_j \sum_{1 \leq m \leq j} \|D_\Gamma^m v\|_{L_p(\Gamma)}.$$

The first two inequalities follow from (2.11) and (2.12) along with the equivalence of $d\sigma$ and $d\sigma_{hk}$ for h small enough. Inequality (2.17) follows from repeated application of (2.11), Proposition 2.3, and the equivalence of $d\sigma$ and $d\sigma_{hk}$.

Next we establish analogues of (2.15), (2.16), and (2.17) for functions defined on Γ_h^k and Γ_h . In particular, let \tilde{T} be a triangular face of Γ_h , and let $\hat{T} = a_k(\tilde{T}) \subset \Gamma_h^k$. Let also v be defined and piecewise smooth on Γ_h^k , and for $\tilde{x} \in \tilde{T}$ let $\tilde{v}(\tilde{x}) = v(a_k(\tilde{x}))$. Then there exist positive constants $C_{i,j}$ such that for h small enough,

$$(2.18) \quad \frac{1}{C_{0,k}} \|v\|_{L_p(\hat{T})} \leq \|\tilde{v}\|_{L_p(\tilde{T})} \leq C_{0,k} \|v\|_{L_p(\hat{T})},$$

$$(2.19) \quad \frac{1}{C_{1,k}} \|\nabla_{\Gamma_h^k} v\|_{L_p(\hat{T})} \leq \|\nabla_{\Gamma_h} \tilde{v}\|_{L_p(\tilde{T})} \leq C_{1,k} \|\nabla_{\Gamma_h^k} v\|_{L_p(\hat{T})},$$

$$(2.20) \quad \|D_{\Gamma_h^k}^j \tilde{v}\|_{L_p(\tilde{T})} \leq C_j \sum_{1 \leq m \leq j} \|D_{\Gamma_h^k}^m v\|_{L_p(\hat{T})}.$$

We briefly discuss the proof of the above inequalities. Because the transformation $\tilde{x} \rightarrow a_k(\tilde{x})$ is the Lagrange interpolant of $\tilde{x} \rightarrow a(\tilde{x})$, $\|a_k\|_{W_\infty^m(T)} \leq C \|a\|_{W_\infty^m(T)} \leq C$ for $m \geq 0$ and h small enough. Let $\tilde{\mu}_{hk}$ be defined by $\tilde{\mu}_{hk}(\tilde{x}) d\sigma_{h1} = d\sigma_{hk}(a_k(\tilde{x}))$, $\tilde{x} \in \Gamma_h$. Then $|\mu_{h1} - \tilde{\mu}_{hk}| \leq Ch^k$, so that $\tilde{\mu}_{hk} \approx 1$ for h small enough. These two facts taken together immediately give (2.18), (2.20), and the second inequality in (2.19).

In order to establish the first inequality in (2.19), assume for simplicity that $n = 2$ and T lies in the xy -plane. The general case follows by employing an appropriate coordinate transformation and making the obvious adjustments if $n = 3$. We have

$$(2.21) \quad \begin{aligned} \nabla_{\Gamma_h} \tilde{v}(\tilde{x}) &= \nabla_{\Gamma_h} v(a_k(\tilde{x})) \\ &= \begin{bmatrix} a_{k,x} & a_{k,y} & 0 \end{bmatrix}^T \nabla_{\Gamma_h^k} v(a_k(\tilde{x})) \\ &= \left(\begin{bmatrix} a_{k,x} & a_{k,y} & 0 \end{bmatrix}^T + \vec{v}_h^k \otimes \vec{v}_h^k \right) \nabla_{\Gamma_h^k} v(a_k(\tilde{x})). \end{aligned}$$

Let $\mathbf{A} = \begin{bmatrix} a_{k,x}(\tilde{x}) & a_{k,y}(\tilde{x}) & 0 \end{bmatrix}^T + \vec{v}_h^k(\tilde{x}) \otimes \vec{v}_h^k(\tilde{x})$ and $\mathbf{B} = (\mathbf{I} - d\mathbf{H})(\tilde{x}) = \nabla a + \vec{v} \otimes \vec{v}$ for $\tilde{x} \in \Gamma_h$, and let $\|\cdot\|_2$ be the matrix 2-norm. We first use the fact that $\nabla a = \mathbf{P} - d\mathbf{H}$ to calculate that $|a_z| = |\nabla a \cdot \vec{v}_h^1| = |\nabla a \cdot (\vec{v}_h^1 - \vec{v})| \leq Ch$. In addition, $|a_{k,x} - a_x| + |a_{k,y} - a_y| \leq Ch^k$. Next we note that since \mathbf{B} is defined on Γ_h and approaches the identity as $dist(\Gamma_h, \Gamma) \rightarrow 0$, $\|\mathbf{B}\|_2 + \|\mathbf{B}^{-1}\|_2 \leq C$ for h small enough.

Thus employing (2.8), we have (again for h small enough) that

$$\begin{aligned}
 \|\mathbf{A}^{-1}\|_2 &\leq \|\mathbf{A}^{-1} - \mathbf{B}^{-1}\|_2 + \|\mathbf{B}^{-1}\|_2 \\
 (2.22) \qquad &\leq \|\mathbf{A}^{-1}\|_2 \|\mathbf{B} - \mathbf{A}\|_2 \|\mathbf{B}^{-1}\|_2 + C \\
 &\leq Ch \|\mathbf{A}^{-1}\|_2 + C \leq C.
 \end{aligned}$$

Multiplying (2.21) through by \mathbf{A}^{-1} , inserting (2.22) into (2.21), and employing the equivalence of $d\sigma_h$ and $d\sigma_{hk}$ yields the first inequality in (2.19).

2.5. Finite element spaces and approximation theory. We begin by defining a family of Lagrange finite element spaces on Γ_h . Let $\tilde{S}_h^r = \{\tilde{\chi} \in C^0(\Gamma_h) : \tilde{\chi}|_{\tilde{T}} \in \mathbb{P}_r \forall \tilde{T} \in \tilde{\mathcal{T}}_h\}$, where $r \geq 1$ and \mathbb{P}_r is the set of polynomials in n variables of degree r or less. We next define the family \hat{S}_{hk}^r on Γ_h^k by

$$\hat{S}_{hk}^r = \{\hat{\chi} \in C^0(\Gamma_h^k) : \hat{\chi} = \tilde{\chi} \circ a_k^{-1} \text{ for some } \tilde{\chi} \in \tilde{S}_h^r\}.$$

\hat{S}_{hk}^r is an *isoparametric* finite element space if $k = r$, *subparametric* if $k < r$, and *superparametric* if $k > r$. We finally define the corresponding lifted spaces on Γ ,

$$S_h^r = \{\chi \in C^0(\Gamma) : \chi = \tilde{\chi}^\ell \text{ for some } \tilde{\chi} \in \tilde{S}_h^r\}$$

and

$$S_{hk}^r = \{\chi \in C^0(\Gamma) : \chi = \hat{\chi}^\ell \text{ for some } \hat{\chi} \in \hat{S}_{hk}^r\}.$$

Note that because $a \circ a_k \neq a$, $S_{hk}^r \neq S_h^r$.

Next we state results concerning finite element approximation theory. We only consider Lagrange-type interpolants as we only need to approximate functions which are sufficiently smooth (H_2^2) to guarantee the availability of point values for $n \leq 3$. For $v \in H_2^2(\Gamma)$, we define the interpolant $I_h^1 = I_h : C^0(\Gamma) \rightarrow S_h^r$ by

$$I_h v = (\tilde{I}_h v^\ell)^\ell,$$

where $\tilde{I}_h : C^0(\Gamma_h) \rightarrow \tilde{S}_h^r$ is the standard Lagrange interpolant. We also define the interpolant $\hat{I}_h^k : C^0(\Gamma_h^k) \rightarrow \hat{S}_{hk}^r$ by $\hat{I}_h^k v(x) = \tilde{I}_h v(a_k^{-1}(x))$, and

$$I_h^k v = (\hat{I}_h^k v^\ell)^\ell.$$

Note that $I_h \neq I_h^k$ since $a \circ a_k(x) \neq a(x)$ for $x \in \Gamma_h$. This is the case even though the nodal points lying on Γ (and thus nodal values) of the two interpolants are the same.

At several points in our presentation we will consider subdomains $D \subset \Gamma$. Let $D_h = \text{int}(\cup_{T \in \mathcal{T}_h, T \cap \bar{D} \neq \emptyset} \bar{T})$ and $D_{hk} = \text{int}(\cup_{T \in \mathcal{T}_h^k, T \cap \bar{D} \neq \emptyset} \bar{T})$. Also, for a given parameter $\gamma \geq h$, we let $D_\gamma = \{x \in \Gamma : \text{dist}_\Gamma(x, D) < \gamma\}$.

We shall need the following approximation and superapproximation results.

PROPOSITION 2.7. *Assume that $v \in W_p^{r+1}(\Gamma)$ for some $2 \leq p \leq \infty$, let h be small enough, and let $D \subset \Gamma$. Assume that either $I = I_h$, $\tilde{D}_h = D_h$, and $S^r = S_h^r$ or $I = I_h^k$, $\tilde{D}_h = D_{hk}$, and $S^r = S_{hk}^r$. Then for $i = 0, 1$ and $2 \leq m \leq r + 1$,*

$$(2.23) \qquad |v - Iv|_{W_p^i(D)} \leq Ch^{m-i} \|v\|_{W_p^m(\tilde{D}_h)}.$$

Let also $\omega \in W_\infty^r(\Gamma)$. Then for $\chi \in S^r$,

$$(2.24) \quad \begin{aligned} & \|\nabla_\Gamma(\omega\chi - I(\omega\chi))\|_{L_p(D)} \\ & \leq C \left(h^r \|\chi\|_{L_p(\tilde{D}_h)} \|\omega\|_{W_\infty^{r+1}(\Gamma)} + \|\nabla_\Gamma\chi\|_{L_p(\tilde{D}_h)} \sum_{i=1}^r h^i \|\omega\|_{W_\infty^i(\tilde{D}_h)} \right). \end{aligned}$$

Finally, for any $\chi \in S^r$ and any mesh domain \tilde{D}_h ,

$$(2.25) \quad \|\nabla_\Gamma\chi\|_{L_2(\tilde{D}_h)} \leq Ch^{-1} \|\chi\|_{L_2(\tilde{D}_h)}.$$

All constants above depend on sufficiently high derivatives of the distance function d .

Proof. The proof follows by combining (2.15) through (2.20) with standard estimates for the Lagrange interpolant on Γ_h (cf. [BS02]). For example, if $I = I_h^k$, we may prove (2.24) by letting \tilde{T} be a face of Γ_h and $(a \circ a_k)(\tilde{T}) = T \subset \Gamma$. Let $\tilde{\chi}(x) = \chi((a \circ a_k)(x))$ and $\tilde{\omega}(x) = \omega((a \circ a_k)(x))$ for $x \in \tilde{T}$. Inequalities (2.15) and (2.19), standard approximation and inverse results on \tilde{T} , and (2.17) and (2.20) then yield

$$\begin{aligned} \|\nabla_\Gamma(\omega\chi - I_h(\omega\chi))\|_{L_p(T)} & \leq C_1 C_{1,k} \|\nabla_{\Gamma_h}(\tilde{\omega}\tilde{\chi} - \tilde{I}_h(\tilde{\omega}\tilde{\chi}))\|_{L_p(\tilde{T})} \\ & \leq Ch^r |\tilde{\omega}\tilde{\chi}|_{W_p^{r+1}(\tilde{T})} \leq Ch^r \sum_{i=1}^{r+1} |\tilde{\omega}|_{W_\infty^i(\tilde{T})} |\tilde{\chi}|_{W_p^{r+1-i}(\tilde{T})} \\ & \leq C \left(h^r \|\tilde{\chi}\|_{L_p(\tilde{T})} |\tilde{\omega}|_{W_\infty^{r+1}(\tilde{T})} + \|\nabla_{\Gamma_h}\tilde{\chi}\|_{L_p(\tilde{T})} \sum_{i=1}^r h^i |\tilde{\omega}|_{W_\infty^i(\tilde{T})} \right) \\ & \leq CC_{r+1} C_{r+1,k} \left[h^r \|\chi\|_{L_p(T)} \|\omega\|_{W_\infty^{r+1}(T)} + C_1 C_{1,k} \|\nabla_\Gamma\chi\|_{L_p(T)} \sum_{i=1}^r h^i \|\omega\|_{W_\infty^i(T)} \right]. \end{aligned}$$

Summing over $T \cap D \neq \emptyset$ completes the proof of (2.20). The rest of Proposition 2.7 is proved in a similar fashion, with obvious slight simplifications when $I = I_h$. \square

The proofs of our pointwise estimates also employ a discrete δ -function.

PROPOSITION 2.8. *Let $S^r = S_h^r$ or $S^r = S_{hk}^r$, let $x \in T \subset \Gamma$ with T a surface triangle in either \mathcal{T}_h or \mathcal{T}_h^k , and let \vec{n} be a unit vector lying in the tangent plane to Γ at x . Then there exist $\delta_x \in C_0^\infty(T)$ and $\tilde{\delta}_x \in [C_0^\infty(T)]^{n+1}$ such that*

$$(2.26) \quad \|\delta_x\|_{W_p^j(T)} + \|\tilde{\delta}_x\|_{W_p^j(T)} \leq Ch^{-j-n+\frac{n}{p}}$$

for $j = 0, 1$ and $1 \leq p \leq \infty$, and for any $\chi \in S^r$,

$$(2.27) \quad |\chi(x)| \leq C \left| \int_T \delta_x \chi \, d\sigma \right|,$$

$$(2.28) \quad |\nabla_\Gamma\chi(x) \cdot \vec{n}| \leq C \left| \int_T \chi \nabla_\Gamma \cdot \tilde{\delta}_x \, d\sigma \right|.$$

Proof. We prove (2.28) when $S^r = S_h^r$; the other cases are similar. Assume $x = a(\tilde{x})$ for $\tilde{x} \in \tilde{T} \in \tilde{\mathcal{T}}_h$, and $T = a(\tilde{T})$. Then employing (2.12), we have

$$\begin{aligned} |\nabla_\Gamma\chi(x) \cdot \vec{n}| & = \left| [(\mathbf{I} - d\mathbf{H})(\tilde{x})]^{-1} \left[\mathbf{I} - \frac{\vec{\nu}_h(\tilde{x}) \otimes \vec{\nu}(\tilde{x})}{\vec{\nu}_h(\tilde{x}) \cdot \vec{\nu}(\tilde{x})} \right] \nabla_{\Gamma_h}\chi^\ell(\tilde{x}) \cdot \vec{n} \right| \\ & \leq C |\nabla_{\Gamma_h}\chi^\ell(\tilde{x}) \cdot \vec{n}|. \end{aligned}$$

Following [SW95], there exists a smooth function $\delta_{\tilde{x}}$ with support in \tilde{T} and not dependent on χ such that $\|\delta_{\tilde{x}}\|_{W_p^k(T)} \leq Ch^{-k-n+\frac{n}{p}}$ and $\nabla_{\Gamma_h} \chi^\ell(\tilde{x}) \cdot \vec{n} = \int_{\tilde{T}} \nabla_{\Gamma_h} \chi^\ell \cdot \vec{n} \delta_{\tilde{x}} d\sigma_h$. Employing (2.11) and integrating by parts yields

$$\int_{\tilde{T}} \nabla_{\Gamma_h} \chi^\ell \cdot \vec{n} \delta_{\tilde{x}} d\sigma_h = - \int_T \chi \nabla_\Gamma \cdot \left([\mathbf{I} - d\mathbf{H}][\mathbf{P}_h] \vec{n} \frac{1}{\mu_h} \delta_{\tilde{x}}^\ell \right) d\sigma.$$

Setting $\tilde{\delta}_x = \frac{1}{\mu_h} \delta_{\tilde{x}}^\ell [\mathbf{I} - d\mathbf{H}][\mathbf{P}_h] \vec{n}$, we thus have (2.28). The proof of (2.26) is easily accomplished using (2.15) and (2.16). \square

2.6. Finite element methods. In this section we define two main types of finite element methods. The first type is defined on polynomial approximations of Γ using the spaces \hat{S}_{hk}^r . Dziuk's original method in [Dz88] is a special case of this method. The second class of methods involves finite element solutions defined on Γ using the spaces S_h^r and S_{hk}^r .

We first define $\tilde{u}_{hk} \in \hat{S}_{hk}^r$. Let $f_h \in L_2(\Gamma_h^k)$ be an approximation to f^ℓ satisfying $\int_{\Gamma_h^k} f_h d\sigma_{hk} = 0$. Then $\tilde{u}_{hk} \in \hat{S}_{hk}^r$ uniquely satisfies $\int_{\Gamma_h^k} \tilde{u}_{hk} d\sigma_{hk} = 0$ and

$$(2.29) \quad \int_{\Gamma_h^k} \nabla_{\Gamma_h^k} \tilde{u}_{hk} \nabla_{\Gamma_h^k} v_h d\sigma_{hk} = \int_{\Gamma_h^k} f_h v_h d\sigma_{hk} \quad \forall v_h \in \hat{S}_{hk}^r.$$

Dziuk's original method results if we take $k = r = 1$ and $f_h = f^\ell - \frac{1}{|\Gamma_h|} \int_{\Gamma_h} f^\ell d\sigma_{h1}$. Using (2.14) while recalling the definition (2.13) of A_Γ and the definition (2.1) of L , we have the perturbed Galerkin orthogonality relationship

$$L(u - \tilde{u}_{hk}^\ell, \chi^\ell) = \int_\Gamma (\mathbf{A}_\Gamma - \mathbf{P}) \nabla_\Gamma \tilde{u}_{hk}^\ell \nabla_\Gamma \chi^\ell d\sigma + \int_\Gamma \left(f - \frac{f_h^\ell}{\mu_{hk}^\ell} \right) \chi^\ell d\sigma, \quad \chi \in \hat{S}_{hk}^r.$$

We next define two methods directly on Γ . The first of these methods employs the spaces S_h^r that are defined by lifting polynomial spaces directly from Γ_h . In particular, let $u_{h,\Gamma} \in S_h^r$ satisfy $\int_\Gamma u_{h,\Gamma} d\sigma_h = 0$ and

$$(2.30) \quad \int_\Gamma \nabla_\Gamma u_{h,\Gamma} \nabla_\Gamma v_h d\sigma = \int_\Gamma f v_h d\sigma \quad \forall v_h \in S_h^r.$$

$u_{h,\Gamma}$ satisfies the Galerkin orthogonality relationship

$$L(u - u_{h,\Gamma}, \chi) = 0, \quad \chi \in S_h^r.$$

So long as one has ready access to the projection a , it is not difficult to program the method (2.30). Indeed, from (2.12) we see that (2.30) may be viewed as a finite element method over Γ_h for an elliptic problem with nonconstant elliptic coefficient matrix. Equation (2.30) may thus be regarded as an alternative to our generalized version (2.29) of Dziuk's method which does not involve any geometric error. We emphasize, however, that there are cases where one has access only to a polynomial approximation of Γ , and employing (2.30) is not possible in these cases.

In addition, we let $u_{hk} \in S_{hk}^r$ satisfy $\int_\Gamma u_{hk} = 0$,

$$(2.31) \quad \int_\Gamma \nabla_\Gamma u_{hk} \nabla_\Gamma v_h d\sigma = \int_\Gamma f v_h d\sigma \quad \forall v_h \in S_{hk}^r.$$

u_{hk} satisfies the Galerkin orthogonality relationship

$$L(u - u_{hk}, \chi) = 0, \quad \chi \in S_{hk}^r.$$

We employ (2.31) only as a theoretical tool in duality arguments used to prove error bounds in non-energy norms and do not foresee any practical use for it.

3. Abstract error analysis. In this section we prove error estimates for surface finite element methods. Our analysis is carried out under the assumption that the approximation properties proved for the spaces S_h^r and S_{hk}^r in section 2.5 hold. We prove our results under general assumptions, as we wish our analysis to apply in other situations. In particular, these assumptions will hold if the approximating surfaces Γ_h and Γ_h^k have nodes that lie within $O(h^{k+1})$ of Γ instead of on Γ . It is reasonable to expect that this would be the case when using isoparametric spaces to compute evolving surfaces as in [Dz91], for example.

3.1. Assumptions on the finite element space and solution. We denote by S^r a generic finite element space of degree r . Depending on the error estimate to be proven, we shall require some or all of the following approximation properties:

- A1: *Basic approximation.* We assume that there exists a linear interpolation operator $I : H_2^2(\Gamma) \rightarrow S^r$ satisfying (2.23).
- A2: *Superapproximation.* Inequality (2.24) holds for any $\chi \in S^r$.
- A3: *Inverse inequality.* Inequality (2.25) holds for any $\chi \in S^r$.
- A4: *Discrete δ function.* There exist discrete δ -functions satisfying the properties (2.26), (2.27), and (2.28).

Finally we assume that the finite element approximation $u_h \in S^r$ to u satisfies the perturbed Galerkin orthogonality relationship

$$(3.1) \quad \int_{\Gamma} \nabla_{\Gamma}(u - u_h) \nabla_{\Gamma} \chi \, d\sigma = F(\chi) \quad \forall \chi \in S^r,$$

where F is assumed to be a continuous linear functional on $H^1(\Gamma)/\mathbb{R}$. Here we shall think of F as encoding a geometric error resulting from the discrete approximation of the surface Γ . Thus $F \equiv 0$ for the methods (2.30) and (2.31) defined directly on Γ , while for the method (2.29) defined on polynomial approximations to Γ we have $F(\chi) = \int_{\Gamma} (\mathbf{A}_{\Gamma} - \mathbf{I}) \nabla_{\Gamma} \tilde{u}_{hk}^{\ell} \nabla_{\Gamma} \chi \, d\sigma + \int_{\Gamma} (f - f_h / \mu_{hk}^{\ell}) \chi \, d\sigma$. (The latter version of F is continuous on $H^1(\Gamma)/\mathbb{R}$ because $\int_{\Gamma} (f - f_h / \mu_{hk}^{\ell}) \, d\sigma = 0$.) Such a linear functional F may also be employed to analyze other error sources such as the inexact evaluation of integrals due to numerical quadrature or nonlinearities (cf. the classical work [NS74] and the discussion in [De07]).

3.2. H^1 and L_2 estimates. Here we give local and global H^1 and L_2 estimates. Before doing so, we define the norms

$$\| \| F \| \|_{H^{-j}} = \sup_{u \in H^j(\Gamma)/\mathbb{R}, \|u\|_{H^j(\Gamma)/\mathbb{R}}=1} F(u)$$

and

$$\| \| F \| \|_{H^{-1}(D)} = \sup_{u \in H_0^1(D), \|\nabla_{\Gamma} u\|_{L_2(D)}=1} F(u), \quad D \subsetneq \Gamma$$

on linear functionals $F : H^1(\Gamma)/\mathbb{R} \rightarrow \mathbb{R}$.

THEOREM 3.1. *Assume that $u \in H^1(\Gamma)$ and $u_h \in S^r$ satisfy $L(u - u_h, v_h) = F(v_h) \forall v_h \in S^r$, where F is a continuous linear functional on $H^1(\Gamma)/\mathbb{R}$. Then*

$$(3.2) \quad \|\nabla_{\Gamma} u_h\|_{L_2(\Gamma)} \leq \|\nabla_{\Gamma} u\|_{L_2(\Gamma)} + C \| \| F \| \|_{H^{-1}},$$

$$(3.3) \quad \|\nabla_{\Gamma}(u - u_h)\|_{L_2(\Gamma)} \leq \min_{\chi \in S^r} \|\nabla_{\Gamma}(u - \chi)\|_{L_2(\Gamma)} + C \| \| F \| \|_{H^{-1}}.$$

Let $D \subset \Gamma$ be a subdomain, and let $Kh \leq \gamma \leq \gamma_\Gamma$ with K sufficiently large and γ_Γ defined as in section 2.1. Then if A.1, A.2, and A.3 hold,

$$(3.4) \quad \begin{aligned} \|\nabla_\Gamma(u - u_h)\|_{L_2(D)} &\leq C \min_{\chi \in S^r} \left(\|\nabla_\Gamma(u - \chi)\|_{L_2(D_\gamma)} + \frac{1}{\gamma} \|u - \chi\|_{L_2(D_\gamma)} \right) \\ &\quad + \frac{1}{\gamma} \|u - u_h\|_{L_2(D_\gamma)} + \|F\|_{H^{-1}(D_\gamma)}. \end{aligned}$$

Finally, let $\overline{u - u_h} = \frac{1}{|\Gamma|} \int_\Gamma (u - u_h) \, d\sigma$. Then if A.1 is satisfied,

$$(3.5) \quad \|u - u_h - \overline{u - u_h}\|_{L_2(\Gamma)} \leq C(h \min_{\chi \in S^r} \|\nabla(u - \chi)\|_{H^1(\Gamma)} + h \|F\|_{H^{-1}} + \|F\|_{H^{-2}}).$$

Proof. In order to prove (3.2), we calculate that

$$\begin{aligned} \|\nabla_\Gamma u_h\|_{L_2(\Gamma)}^2 &= \int_\Gamma \nabla_\Gamma u \nabla_\Gamma u_h \, d\sigma - F(u_h) \\ &\leq \|\nabla_\Gamma u\|_{L_2(\Gamma)} \|\nabla_\Gamma u_h\|_{L_2(\Gamma)} + \|F\|_{H^{-1}} \|u_h\|_{H^1(\Gamma)/\mathbb{R}} \\ &\leq (\|\nabla_\Gamma u\|_{L_2(\Gamma)} + C \|F\|_{H^{-1}}) \|\nabla_\Gamma u_h\|_{L_2(\Gamma)}, \end{aligned}$$

where C arises from a Poincaré inequality. Dividing through by $\|\nabla_\Gamma u_h\|_{L_2(\Gamma)}$ completes the proof of (3.2). Inequality (3.3) may be proved by writing $u - u_h = (u - \chi) - (u_h - \chi)$.

We next prove (3.4). Let $\{D_i\}_{i=1}^N$ be a cover of D consisting of balls of radius $\frac{\gamma}{4}$, and let $D_{i,\gamma/2} = \{x \in \Gamma : \text{dist}_\Gamma(x, D_i) < \frac{\gamma}{4}\}$. We may choose the cover so that the balls $D_{i,\gamma/2}$ have finite overlap. Finally let $\omega_i \in C_0^\infty(D_{i,\gamma/2})$ with $\omega_i|_{D_i} \equiv 1$ and $\|\omega_i\|_{W_\infty^j(\Gamma)} \leq C\gamma^{-j}$, $0 \leq j \leq r+1$. Such a cutoff function ω exists for $\gamma \leq \gamma_\Gamma$. Fixing $\chi \in S^r$, we set $\psi_i = \omega_i^2(\chi - u_h)$ and compute

$$(3.6) \quad \begin{aligned} \|\nabla_\Gamma(u - u_h)\|_{L_2(D)}^2 &\leq \sum_{i=1}^N L(\omega_i(u - u_h), \omega_i(u - u_h)) \\ &= \sum_{i=1}^N L(u - u_h, \omega_i^2(u - u_h)) + \int_{D_{i,\gamma/2}} |\nabla_\Gamma \omega_i|^2 (u - u_h)^2 \, d\sigma \\ &\leq \sum_{i=1}^N [L(u - u_h, \omega_i^2(u - \chi)) + L(u - u_h, \psi_i - I\psi_i) + F(I\psi_i)] \\ &\quad + \frac{C}{\gamma^2} \|u - u_h\|_{L_2(D_\gamma)}^2. \end{aligned}$$

Next we bound the terms in the last sum in (3.6). For any $1 \geq \epsilon > 0$,

$$(3.7) \quad \begin{aligned} L(u - u_h, \omega_i^2(u - \chi)) &= \int_\Gamma \nabla_\Gamma(\omega_i(u - u_h)) [\omega_i \nabla_\Gamma(u - \chi) + 2(u - \chi) \nabla_\Gamma \omega_i] \, d\sigma \\ &\quad - \int_\Gamma \omega_i(u - u_h) \nabla_\Gamma \omega_i \nabla_\Gamma(u - \chi) \, d\sigma - 2 \int_\Gamma |\nabla_\Gamma \omega_i|^2 (u - u_h)(u - \chi) \, d\sigma \\ &\leq \epsilon \|\nabla_\Gamma(\omega_i(u - u_h))\|_{L_2(\Gamma)}^2 + \frac{C}{\epsilon} \|\nabla_\Gamma(u - \chi)\|_{L_2(D_{i,\gamma/2})}^2 \\ &\quad + \frac{C}{\gamma^2 \epsilon} (\|u - u_h\|_{L_2(D_{i,\gamma/2})}^2 + \|u - \chi\|_{L_2(D_{i,\gamma/2})}^2). \end{aligned}$$

Applying (2.24) and (2.25) while recalling that $h \leq \gamma$ and $\|\omega_i\|_{W_\infty^j(\Gamma)} \leq C\gamma^{-j}$ yields

$$\begin{aligned}
(3.8) \quad & \|\nabla_\Gamma(\psi_i - I_h\psi_i)\|_{L_2(\Gamma)} \\
& \leq C\frac{h}{\gamma} \left(\frac{1}{\gamma} \|\chi - u_h\|_{L_2((D_{i,\gamma/4})_h)} + \|\nabla_\Gamma(\chi - u_h)\|_{L_2((D_{i,\gamma/4})_h)} \right) \\
& \leq \frac{C}{\gamma} (\|u - \chi\|_{L_2(D_{i,\gamma/2})} + \|u - u_h\|_{L_2(D_{i,\gamma/2})}).
\end{aligned}$$

Applying the first line of the previous inequality, we find

$$\begin{aligned}
(3.9) \quad & L(u - u_h, \psi_i - I_h\psi_i) \leq C\frac{h}{\gamma} \|\nabla_\Gamma(u - u_h)\|_{L_2(D_{i,\gamma/2})}^2 + C\|\nabla_\Gamma(u - \chi)\|_{L_2(D_{i,\gamma/2})}^2 \\
& \quad + \frac{C}{\gamma^2} (\|u - u_h\|_{L_2(D_{i,\gamma/2})}^2 + \|u - \chi\|_{L_2(D_{i,\gamma/2})}^2).
\end{aligned}$$

Applying the second line of (3.8) and noting that $\|\nabla_\Gamma\psi_i\|_{L_2(D_{i,\gamma/2})} \leq \|\nabla_\Gamma(u - \chi)\|_{L_2(D_{i,\gamma/2})} + \|\nabla_\Gamma(\omega_i(u - u_h))\|_{L_2(D_{i,\gamma/2})} + \frac{1}{\gamma}\|u - u_h\|_{L_2(D_{i,\gamma/2})}$, we finally compute

$$\begin{aligned}
(3.10) \quad & \sum_{i=1}^N F(I\psi_i) = F \left(\sum_{i=1}^N I\psi_i \right) \leq \|F\|_{H^{-1}(D_{\gamma/2})} \sum_{i=1}^N \|\nabla_\Gamma I\psi_i\|_{L_2(D_{i,\gamma/2})} \\
& \leq \|F\|_{H^{-1}(D_{\gamma/2})} \left[\sum_{i=1}^N \|\nabla_\Gamma(I\psi_i - \psi_i)\|_{L_2(D_{i,\gamma/2})} + \|\nabla_\Gamma\psi_i\|_{L_2(D_{i,\gamma/2})} \right] \\
& \leq \frac{C}{\epsilon} \|F\|_{H^{-1}(D_{\gamma/2})}^2 + \frac{C}{\gamma^2} (\|u - \chi\|_{L_2(D_{\gamma/2})}^2 + \|u - u_h\|_{L_2(D_{\gamma/2})}^2) \\
& \quad + C\|\nabla_\Gamma(u - \chi)\|_{L_2(D_{\gamma/2})}^2 + \epsilon \sum_{i=1}^N \|\nabla_\Gamma(\omega_i(u - u_h))\|_{L_2(\Gamma)}^2.
\end{aligned}$$

Combining (3.7), (3.9), and (3.10) into (3.6) yields

$$\begin{aligned}
(3.11) \quad & \sum_{i=1}^N \|\nabla_\Gamma(\omega_i(u - u_h))\|_{L_2(D_{i,\gamma/2})}^2 \leq C(\epsilon) \left[\frac{1}{\gamma^2} (\|u - \chi\|_{L_2(D_{\gamma/2})}^2 \right. \\
& \quad \left. + \|u - u_h\|_{L_2(D_{\gamma/2})}^2) + \|\nabla_\Gamma(u - \chi)\|_{L_2(D_{\gamma/2})}^2 + \|F\|_{H^{-1}(D_{\gamma/2})}^2 \right] \\
& \quad + \frac{Ch}{\gamma} \|\nabla_\Gamma(u - u_h)\|_{L_2(D_{\gamma/2})}^2 + 2\epsilon \sum_{i=1}^N \|\nabla_\Gamma(\omega_i(u - u_h))\|_{L_2(D_{i,\gamma/2})}^2.
\end{aligned}$$

The last term in (3.11) may be kicked back by taking $\epsilon = \frac{1}{4}$, yielding

$$\begin{aligned}
(3.12) \quad & \|\nabla_\Gamma(u - u_h)\|_{L_2(D)}^2 \leq C \left[\frac{1}{\gamma^2} (\|u - \chi\|_{L_2(D_{\gamma/2})}^2 + \|u - u_h\|_{L_2(D_{\gamma/2})}^2) \right. \\
& \quad \left. + \|\nabla_\Gamma(u - \chi)\|_{L_2(D_{\gamma/2})}^2 + \|F\|_{H^{-1}(D_{\gamma/2})}^2 + \frac{h}{\gamma} \|\nabla_\Gamma(u - u_h)\|_{L_2(D_{\gamma/2})}^2 \right].
\end{aligned}$$

The term $\frac{h}{\gamma} \|\nabla_\Gamma(u - u_h)\|_{L_2(D_{\gamma/2})}^2$ above may be eliminated by iterating (3.12) with $D_{\gamma/2}$ and D_γ replacing D and $D_{\gamma/2}$, respectively. This results in a term $\frac{h^2}{\gamma^2} \|\nabla_\Gamma(u -$

$\chi) + \nabla_\Gamma(\chi - u_h)\|_{L_2(D_\gamma)}^2$ which may be eliminated by using the triangle inequality and an inverse inequality.

In order to prove (3.5), we first let $z \in H^1(\Gamma)$ solve $L(v, z) = (v, e - \bar{e})_\Gamma$, $\int_\Gamma z \, d\sigma = 0$, where $e = u - u_h$ and $\bar{e} = \overline{u - u_h}$. Then using (2.23), (2.2), and (3.3) yields

$$\begin{aligned} \|e - \bar{e}\|_{L_2(\Gamma)}^2 &= (e - \bar{e}, -\Delta_\Gamma z) = L(e, z - I_h z) + F(I_h z - z) + F(z) \\ &\leq C \|\nabla_\Gamma e\|_{L_2(\Gamma)} \|\nabla_\Gamma(z - I_h z)\|_{L_2(\Gamma)} + \|F\|_{H^{-1}} \|z - I_h z\|_{H^1(\Gamma)} \\ &\quad + \|F\|_{H^{-2}} \|z\|_{H_2^2(\Gamma)} \\ &\leq C(h \min_{\chi \in S^r} \|\nabla_\Gamma(u - \chi)\|_{L_2(\Gamma)} + h \|F\|_{H^{-1}} + \|F\|_{H^{-2}}) \|z\|_{H_2^2(\Gamma)} \\ &\leq C(h \min_{\chi \in S^r} \|\nabla_\Gamma(u - \chi)\|_{L_2(\Gamma)} + h \|F\|_{H^{-1}} + \|F\|_{H^{-2}}) \|e - \bar{e}\|_{L_2(\Gamma)}. \end{aligned}$$

Dividing through by $\|e - \bar{e}\|_{L_2(\Gamma)}$ completes the proof. \square

3.3. Pointwise estimates: Statement of results. In this subsection we state pointwise stability and error estimates. Following [Sch98], let $\sigma_x(y) = \frac{h}{\alpha(x,y)+h}$, where we recall that $\alpha(x, y)$ is the surface distance on Γ . We then define the weighted norm

$$\|u\|_{W_p^j, x, s} = \sum_{0 \leq |\alpha| \leq j} \|\sigma_x^s D^\alpha u\|_{L_p(\Gamma)}.$$

Letting q be the conjugate exponent to p , we define the weighted norm

$$(3.13) \quad \|F\|_{W_p^{-j}, x, s} = \sup_{\|v\|_{W_q^j, x, -s} = 1} F(v).$$

We shall drop the subscripts x and s in (3.13) when $s = 0$.

THEOREM 3.2. *Let $0 \leq s \leq r - 1$ and $0 \leq t \leq r$, and assume that A1, A2, A3, and A4 all hold. Then for any $x \in \Gamma$,*

$$(3.14) \quad \begin{aligned} |(u - u_h - \overline{u - u_h})(x)| \\ \leq C \ell_{h,s} \inf_{\chi \in S^r} (h \|\nabla_\Gamma(u - \chi)\|_{L_\infty, x, s} + \|u - \chi\|_{L_\infty, x, s}) \\ + C(h \ell_{h,s} \|F\|_{W_\infty^{-1}, x, s} + \ell_h \|F\|_{W_\infty^{-2}}), \end{aligned}$$

and

$$(3.15) \quad |\nabla_\Gamma u_h(x)| \leq C(\ell_{h,t} \|\nabla_\Gamma u\|_{L_\infty, x, t} + \ell_h \|F\|_{W_\infty^{-1}}),$$

$$(3.16) \quad |\nabla_\Gamma(u - u_h)(x)| \leq C(\ell_{h,t} \inf_{\chi \in S^r} \|\nabla_\Gamma(u - \chi)\|_{L_\infty, x, t} + \ell_h \|F\|_{W_\infty^{-1}}).$$

Here $\ell_h = \ln \frac{1}{h}$, $\ell_{h,t} = \ell_h$ if $t = r$ and $\ell_{h,t} = 1$ otherwise, and $\ell_{h,s} = \ell_h$ if $s = r - 1$ and $\ell_{h,s} = 1$ otherwise.

Taking $s = t = 0$ and taking a maximum of (3.14) and (3.16) over Γ yields quasi-optimal L_∞ and W_∞^1 error estimates, modulo analysis of perturbation terms involving F . When $s > 0$ (3.14) shows that the pointwise gradient error at x is localized to x in that the weight σ_x^s deemphasizes the approximation error $\nabla(u - \chi)(y)$ by a factor of h^s when $\alpha(x, y) \approx 1$. No localization occurs in errors for function values in the piecewise linear case as $s = r - 1 = 0$ in this case (cf. [De04] for a counterexample). Note that (3.14) and (3.16) are very similar to the results in [Sch98] for domains in \mathbb{R}^n . Details peculiar to the fact that we are working on surfaces are hidden in the functional F .

3.4. Proof of Theorem 3.2. We shall prove (3.15) in full detail. The proof of (3.16) follows from (3.15) by writing $\nabla_\Gamma(u - u_h) = \nabla_\Gamma(u - \chi) - \nabla_\Gamma(u_h - \chi)$. The proof of (3.14) is similar but slightly simpler, and we only sketch its proof.

We proceed via a duality argument. Fix a point $x \in \Gamma$, and let \vec{n} be a unit vector lying in the tangent plane to Γ at x . Let $\tilde{\delta}_x$ satisfy the properties (2.26) and (2.28), and let g^x be a discrete Green's function satisfying $L(v, g^x) = (v, \nabla_\Gamma \cdot \tilde{\delta}_x)$ for all $v \in H^1(\Gamma)$ and $\int_\Gamma g^x \, d\sigma = 0$. (Note that $\int_\Gamma \nabla_\Gamma \cdot \tilde{\delta}_x = 0$.) Let also $g_h^x \in S^r$ be its finite element approximation satisfying $L(v_h, g^x - g_h^x) = 0 \, \forall v_h \in S^r$ and $\int_\Gamma g_h^x \, d\sigma = 0$. Then

$$\begin{aligned} |\nabla_\Gamma u_h(x) \cdot \vec{n}| &\leq C \int_\Gamma u_h \nabla_\Gamma \cdot \tilde{\delta}_x \, d\sigma \\ &= L(u_h, g_h^x) = L(u, g_h^x) - F(g_h^x) \\ &= L(u, g_h^x - g^x) + L(u, g^x) - F(g_h^x) \\ &\leq \|\nabla_\Gamma u\|_{L_\infty, x, t} \|\nabla_\Gamma(g^x - g_h^x)\|_{L_1(\Gamma), x, -t} + \int_T u \nabla_\Gamma \cdot \tilde{\delta}_x \, d\sigma \\ &\quad + \|F\|_{W_\infty^{-1}} \|g_h^x\|_{W_1^1(\Gamma)} \\ &\leq C \|\nabla_\Gamma u\|_{L_\infty, x, t} (1 + \|\nabla_\Gamma(g^x - g_h^x)\|_{L_1(\Gamma), x, -t}) \\ &\quad + C \|F\|_{W_\infty^{-1}} \|\nabla_\Gamma g_h^x\|_{L_1(\Gamma)}, \end{aligned}$$

where we have used a Poincaré inequality in the last step.

Similarly, fix $x \in \Gamma$, and let \hat{g}^x satisfy $\int_\Gamma \hat{g}^x \, d\sigma = 0$ and $L(v, \hat{g}^x) = (v, \delta_x - \overline{\delta_x})$ for δ_x satisfying (2.26) and (2.27). Also let $\hat{g}_h^x \in S^r$ satisfy $L(\hat{g}^x - \hat{g}_h^x, \chi) = 0 \, \forall \chi \in S^r$ and $\int_\Gamma \hat{g}_h^x \, d\sigma = 0$. Let also $x \in T$. Then for $\chi \in S^r$,

$$\begin{aligned} |(u - u_h)(x) - \overline{u - u_h}| &\leq |(u - \chi)(x)| + C \left| \int_\Gamma (\chi - u_h - \overline{u - u_h}) \delta_x \, d\sigma \right| \\ &\leq C (\|u - \chi\|_{L_\infty(T)} + |L(u - u_h, \hat{g}^x)|) \\ &\leq (\|\nabla_\Gamma(u - \chi)\|_{L_\infty, x, s} + \|F\|_{W_\infty^{-1}, x, s}) \|\hat{g}^x - \hat{g}_h^x\|_{W_1^1, x, -s} \\ &\quad + C \|u - \chi\|_{L_\infty(T)} + \|F\|_{W_\infty^{-2}} \|\hat{g}^x\|_{W_1^2(\Gamma)}. \end{aligned}$$

The heart of our proof consists of the following lemma.

LEMMA 3.3. *Under the assumptions of section 2 and Theorem 3.2,*

$$(3.17) \quad \|\nabla_\Gamma(g^x - g_h^x)\|_{L_1, x, -t} \leq C \ell_{h, t},$$

$$(3.18) \quad \|\hat{g}^x - \hat{g}_h^x\|_{W_1^1, x, -s} \leq C h \ell_{h, s},$$

$$(3.19) \quad \|\nabla_\Gamma g^x\|_{L_1(\Gamma)} + \|\hat{g}^x\|_{W_1^2(\Gamma)} \leq C \ell_h.$$

The proof of (3.16) will be complete once we prove Lemma 3.3.

3.5. Proof of Lemma 3.3. The proof of Lemma 3.3 is similar to that given for domains in \mathbb{R}^n in [Sch98] (though the fact that we consider here an indefinite bilinear form complicates matters slightly). Thus we omit some details from our proof.

Note first that $g^x - g_h^x$ satisfies the error estimates of Theorem 3.1 with $F \equiv 0$. We then decompose Γ into annular subdomains about the point x . For a parameter $M > 0$ which we shall later take to be large enough, we fix $\Gamma_0 = B_{Mh}(x)$ and define

$\gamma_j = 2^j Mh$. Let J be the largest integer such that $\gamma_J \leq \frac{2\pi}{2}$, where γ_Γ is defined in section 2.1. For $0 < j < J$, we define the annuli $\Gamma_j = \{y \in \Gamma : \gamma_{j-1} < \alpha(x, y) < \gamma_j\}$ and then finally define $\Gamma_J = \Gamma \setminus \cup_{0 \leq j < J} \overline{\Gamma_j}$. Thus $\Gamma = \cup_{0 \leq j \leq J} \Gamma_j$. Also, we let $\Gamma'_j = \text{int}(\overline{\Gamma_{j-1}} \cup \overline{\Gamma_j} \cup \overline{\Gamma_{j+1}})$, $\Gamma''_j = \Gamma'_{j-1} \cup \Gamma'_j \cup \Gamma'_{j+1}$, and $\Gamma'''_j = \Gamma''_{j-1} \cup \Gamma''_j \cup \Gamma''_{j+1}$.

We then use (3.4), Hölder's inequality, and (2.23) to find that

$$\begin{aligned}
& \|\nabla_\Gamma(g^x - g_h^x)\|_{L_{1,x,-t}} \\
& \leq C(M)h^{n/2}\|\nabla_\Gamma(g^x - g_h^x)\|_{L_2(\Gamma_0)} + C \sum_{j=1}^J \left(\frac{\gamma_j}{h}\right)^t \gamma_j^{n/2} \|\nabla_\Gamma(g^x - g_h^x)\|_{L_2(\Gamma_j)} \\
(3.20) \quad & \leq C(M)h^{n/2}[\|\nabla_\Gamma(g^x - g_h^x)\|_{L_2(\Gamma_0)} + h^{-1}\|g^x - g_h^x\|_{L_2(\Gamma_0)} \\
& \quad + \min_{\chi \in \mathcal{S}^r} (\|\nabla_\Gamma(g^x - \chi)\|_{L_2(\Gamma_0)} + h^{-1}\|g^x - \chi\|_{L_2(\Gamma_0)})] \\
& \quad + \sum_{j=1}^J \left[\left(\frac{\gamma_j}{h}\right)^t \gamma_j^n h^r \|g^x\|_{W_\infty^{r+1}(\Gamma_{j,h})} + \left(\frac{\gamma_j}{h}\right)^t \gamma_j^{n/2-1} \|g^x - g_h^x\|_{L_2(\Gamma_j)} \right].
\end{aligned}$$

Let $\omega_j \in C_0^\infty(\Gamma'_j)$ be a cutoff function satisfying $0 \leq \omega_j \leq 1$ and $\omega_j \equiv 1$ on Γ_j . Let $C_j = \frac{1}{|\Gamma|} \int_{\Gamma'_j} \omega_j^2 (g^x - g_h^x) d\sigma$, and let $w \in H^2(\Gamma)$ with $\int_\Gamma w d\sigma = 0$ solve

$$L(w, v) = (\omega_j^2 (g^x - g_h^x) - C_j, v) \quad \forall v \in H^1(\Gamma).$$

Using (2.23) and recalling that $\int_\Gamma (g^x - g_h^x) d\sigma = 0$, we compute

$$\begin{aligned}
(3.21) \quad & \|g^x - g_h^x\|_{L_2(\Gamma_j)}^2 \leq \|\omega_j (g^x - g_h^x)\|_{L_2(\Gamma)}^2 \\
& = (\omega_j^2 (g^x - g_h^x) - C_j, g^x - g_h^x) \\
& = L(w, g^x - g_h^x) \\
& = L(w - I_h w, g^x - g_h^x) \\
& \leq C(h\|w\|_{H^2(\Gamma''_j)} \|\nabla_\Gamma(g^x - g_h^x)\|_{L_2(\Gamma''_j)} \\
& \quad + h^r \|w\|_{W_\infty^{r+1}(\Gamma \setminus \Gamma'_j)} \|\nabla_\Gamma(g^x - g_h^x)\|_{L_1(\Gamma)}).
\end{aligned}$$

Noting that $w(y) = \int_\Gamma G^y(z) \omega_j^2 (g^x - g_h^x) d\sigma(z)$ since $\int_\Gamma G^y(z) C_j d\sigma(z) = 0$, we use (2.4) to calculate that for any multi-index β with $|\beta| \leq r+1$ and any $y \in \Gamma \setminus \Gamma'_j$,

$$\begin{aligned}
(3.22) \quad & D^\beta w(y) = \int_\Gamma D_y^\beta G^y(z) [\omega_j^2 (g^x - g_h^x)] d\sigma(z) \\
& \leq \sqrt{|\Gamma_j|} \|\omega_j^2 (g^x - g_h^x)\|_{L_2(\Gamma'_j)} \|D_y^\beta G^y\|_{L_\infty(\Gamma'_j)} \\
& \leq C \gamma_j^{n/2} \|\omega_j (g^x - g_h^x)\|_{L_2(\Gamma'_j)} \gamma_j^{1-n-r}.
\end{aligned}$$

Inserting (3.22) into (3.21) and using the regularity estimate (2.2) yields

$$\begin{aligned}
& \|\omega_j (g^x - g_h^x)\|_{L_2(\Gamma)}^2 \leq C[h\|\nabla_\Gamma(g^x - g_h^x)\|_{L_2(\Gamma''_j)} \\
& \quad + \gamma_j^{-n/2+1} \left(\frac{h}{\gamma_j}\right)^r \|\nabla_\Gamma(g^x - g_h^x)\|_{L_1(\Gamma)}] \|\omega_j (g^x - g_h^x)\|_{L_2(\Gamma)},
\end{aligned}$$

so that

$$(3.23) \quad \begin{aligned} \|g^x - g_h^x\|_{L_2(\Gamma_j)} &\leq Ch \|\nabla_\Gamma(g^x - g_h^x)\|_{L_2(\Gamma_j'')} \\ &\quad + \gamma_j^{-n/2+1} \left(\frac{h}{\gamma_j}\right)^r \|\nabla_\Gamma(g^x - g_h^x)\|_{L_1(\Gamma)}. \end{aligned}$$

Recalling (2.26), we next compute that for $y \in \Gamma_{j,h}$ and β with $|\beta| = r + 1$,

$$(3.24) \quad \begin{aligned} D^\beta g^x(y) &= - \int_\Gamma \nabla_{\Gamma,z} D_y^\beta G^y(z) \tilde{\delta}_x(z) \, d\sigma(z) \\ &\leq \|\nabla_\Gamma D_y^\beta G^y\|_{L_\infty(\text{supp}(\tilde{\delta}_x))} \|\tilde{\delta}_x\|_{L_1(\Gamma)} \\ &\leq C \gamma_j^{-n-r}. \end{aligned}$$

Finally, employing (3.3), (3.5), (2.23), (2.2), and (2.26) yields

$$(3.25) \quad \begin{aligned} C(M)h^{n/2} [\|\nabla_\Gamma(g^x - g_h^x)\|_{L_2(\Gamma_0)} + h^{-1}\|g^x - g_h^x\|_{L_2(\Gamma_0)} \\ + \min_{\chi \in S_h^r} (\|\nabla_\Gamma(g^x - \chi)\|_{L_2(\Gamma_0)} + h^{-1}\|g^x - \chi\|_{L_2(\Gamma_0)})] \\ \leq Ch^{n/2+1} \|\nabla_\Gamma \cdot \tilde{\delta}_x\|_{L_2(\Gamma)} \leq C. \end{aligned}$$

Inserting (3.23), (3.24), and (3.25) into (3.20), rearranging terms, and finally employing (3.25) yields

$$\begin{aligned} \|\nabla_\Gamma(g^x - g_h^x)\|_{L_{1,x,-t}} &\leq C + C \sum_{j=1}^J \left(\frac{\gamma_j}{h}\right)^t \gamma_j^{n/2} \|\nabla_\Gamma(g^x - g_h^x)\|_{L_2(\Gamma_j)} \\ &\leq C + C \sum_{j=1}^J \left(\frac{\gamma_j}{h}\right)^t \gamma_j^n h^r \gamma_j^{-r-n} + C \sum_{j=1}^J \left(\frac{\gamma_j}{h}\right)^t \gamma_j^{n/2} \frac{h}{\gamma_j} \|\nabla_\Gamma(g^x - g_h^x)\|_{L_2(\Gamma_j)} \\ &\quad + C \|\nabla_\Gamma(g^x - g_h^x)\|_{L_1} \sum_{j=1}^J \left(\frac{\gamma_j}{h}\right)^t \left(\frac{h}{\gamma_j}\right)^r \\ &\leq C + C(1 + \|\nabla_\Gamma(g^x - g_h^x)\|_{L_1(\Gamma)}) \sum_{j=1}^J \left(\frac{h}{\gamma_j}\right)^{r-t} \\ &\quad + \frac{C}{M} \sum_{j=1}^J \left(\frac{\gamma_j}{h}\right)^t \gamma_j^{n/2} \|\nabla_\Gamma(g^x - g_h^x)\|_{L_2(\Gamma_j)}. \end{aligned}$$

The last term above may be kicked back (to the last term in the first line) for M large enough. In addition, we note that $\sum_{j=1}^J \left(\frac{h}{\gamma_j}\right)^{r-t} \leq C \ell_{h,t} \frac{1}{M^{r-t}}$. Thus

$$(3.26) \quad \|\nabla_\Gamma(g^x - g_h^x)\|_{L_{1,x,-t}} \leq C + \frac{C}{M^{r-t}} \ell_{h,t} \|\nabla_\Gamma(g^x - g_h^x)\|_{L_1(\Gamma)}.$$

Applying (3.26) with $t = 0$ and taking M large enough to kick back the last term yields

$$(3.27) \quad \|\nabla_\Gamma(g^x - g_h^x)\|_{L_1} \leq C.$$

Inserting (3.27) into (3.26) completes the proof of (3.17).

In order to prove the inequality $\|\nabla_{\Gamma} g^x\|_{L_1(\Gamma)} \leq \ell_h$ from (3.19), we first note the easily proven regularity estimate

$$\|\nabla_{\Gamma} g^x\|_{L_2(\Gamma)} \leq C\|\tilde{\delta}_x\|_{L_2(\Gamma)} \leq Ch^{-n/2}.$$

Computing as in (3.24) yields $D^{\alpha} g^x(y) \leq C\alpha(x, y)^{-2}$ for $|\alpha| = 1$ and $\alpha(x, y) \geq 3h$. We thus find that

$$\begin{aligned} \|\nabla_{\Gamma} g^x\|_{L_1(\Gamma)} &\leq Ch^{n/2}\|\nabla_{\Gamma} g^x\|_{L_2(\Gamma)} + \|\nabla_{\Gamma} g^x\|_{L_1(\Gamma \setminus B_{3h}(x))} \\ &\leq C + \int_{3h}^C y^{-1} dy \leq C\ell_h. \end{aligned}$$

The proofs of (3.18) and the inequality $\|\hat{g}^x\|_{W_1^2(\Gamma)} \leq C\ell_h$ are very similar to the corresponding proofs for the appropriate norms of $g^x - g_h^x$ and \hat{g}^x and also to the proofs given in [Sch98], so we only make a couple of notes. First, (3.18) requires us to bound a weighted W_1^1 norm of $\hat{g}^x - \hat{g}_h^x$, not just an L_1 norm of the gradient as in (3.17). However, if we carry out the computation in (3.20) with $\hat{g}^x - \hat{g}_h^x$ and s in place of $g^x - g_h^x$ and t , respectively, then the last line of (3.20) can easily be shown to bound $\|\hat{g}^x - \hat{g}_h^x\|_{L_{1,x,-s}}$. Second, the right-hand side $\delta_x - \overline{\delta_x}$ is not locally supported, which requires a modification when performing computations similar to (3.22) and (3.24). In particular, we note that $\hat{g}^x(y) = \int_{\Gamma} G^y(z)(\delta_x - \overline{\delta_x}) d\sigma(z) = \int_{\Gamma} G^y(z)\delta_x d\sigma(z)$ and then proceed essentially as in (3.24). \square

4. Error analysis of specific methods and numerical results. In this section we apply the abstract error analysis in section 3 to the methods (2.29) and (2.30) in section 2.6. In the case of the method (2.29) defined on polynomial approximations to Γ , the resulting error bounds consist of a ‘‘PDE’’- or ‘‘almost-best-approximation’’-type term that arises in essentially every finite element approximation, plus a geometric error term arising from the approximation of Γ by Γ_h^k . We also briefly describe numerical experiments that confirm the structure of our H^1 and L_2 estimates.

4.1. Error estimates for FEM on polynomial approximations to Γ . We first state a fundamental geometric error bound which is an extension of a bound found in [Dz88] to higher-order approximations of Γ .

PROPOSITION 4.1.

$$(4.1) \quad \|\mathbf{A}_{\Gamma} - \mathbf{P}\|_{L_{\infty}(\Gamma)} \leq Ch^{k+1}.$$

Proof. Recalling that $\|d\|_{L_{\infty}(\Gamma_{h,k})} \leq Ch^{k+1}$ and noting from (2.10) that $|1 - \frac{1}{\mu_{hk}}| \leq Ch^{k+1} + C|1 - \vec{\nu} \cdot \vec{\nu}_h^k| \leq Ch^{k+1} + C|\vec{\nu} - \vec{\nu}_h^k|^2 \leq Ch^{k+1}$, we have $|\mathbf{A}_{\Gamma} - \mathbf{P}| \leq |\mathbf{P}\mathbf{P}_{h,k}\mathbf{P} - \mathbf{P}| + Ch^{k+1}$. But $|\mathbf{P}\mathbf{P}_{h,k}\mathbf{P} - \mathbf{P}| = |(\vec{\nu}_h^k - \vec{\nu} \cdot \vec{\nu}_h^k \vec{\nu}) \otimes (\vec{\nu}_h^k - \vec{\nu} \cdot \vec{\nu}_h^k \vec{\nu})| \leq Ch^{2k}$, which completes the proof. \square

Next we give H^1 and L_2 estimates.

COROLLARY 4.2. *Let \tilde{u}_{hk} satisfy (2.29) with $f_h = \mu_{hk} f^{\ell}$. Then if $u \in H^{r+1}(\Gamma)$,*

$$(4.2) \quad \|\nabla_{\Gamma}(u - \tilde{u}_{hk}^{\ell})\|_{L_2(\Gamma)} \leq C(h^r \|u\|_{H^{r+1}(\Gamma)} + h^{k+1} \|\nabla_{\Gamma} u\|_{L_2(\Gamma)}),$$

$$(4.3) \quad \|u - \tilde{u}_{hk}^{\ell} - \overline{u - \tilde{u}_{hk}^{\ell}}\|_{L_2(\Gamma)} \leq C(h^{r+1} \|u\|_{H^{r+1}(\Gamma)} + h^{k+1} \|\nabla_{\Gamma} u\|_{L_2(\Gamma)}),$$

where C depends on d and its derivatives.

Remark 4.3. The geometric error in the L_2 estimate (3.5) has the form $h\|F\|_{-1} + \|F\|_{-2}$. However, we cannot take advantage of the fact that the norm $\|\cdot\|_{-2}$ is weaker than the norm $\|\cdot\|_{-1}$ in order to achieve a higher order of convergence h^{k+2} for the geometric error in our L_2 estimates. Computational experiments in section 4 confirm that the geometric error is indeed of order h^{k+1} for both the L_2 and energy errors.

Remark 4.4. It is possible to show that $|\overline{u - \tilde{u}_{hk}^\ell}| = |\overline{\tilde{u}_{hk}^\ell}| \leq Ch^{k+1}\|\nabla_\Gamma u\|_{L_2(\Gamma)}$ for h small enough, so that in fact (4.3) holds with $\|u - \tilde{u}_{hk}^\ell\|_{L_2(\Gamma)}$ on the left-hand side. We state (4.3) as we do both to maintain consistency with [Dz88] and because we wish to emphasize that (4.3) is sharp with respect to the order of the geometric error.

Proof. Note first that if $f_h = \mu_{hk}f^\ell$, \tilde{u}_{hk} satisfies (3.1) with $F(\chi) = \int_\Gamma(\mathbf{A}_\Gamma - \mathbf{P})\nabla_\Gamma\tilde{u}_{hk}^\ell\nabla_\Gamma\chi\,d\sigma$. Combining (3.2) and (4.1) yields

$$\begin{aligned} \|F\|_{H^{-1}} &\leq Ch^{k+1}\|\nabla_\Gamma\tilde{u}_{hk}^\ell\|_{L_2(\Gamma)} \\ &\leq Ch^{k+1}(\|\nabla_\Gamma u\|_{L_2(\Gamma)} + C\|F\|_{H^{-1}}). \end{aligned}$$

Taking h small enough to kick back the last term above yields

$$(4.4) \quad \|F\|_{H^{-1}} \leq Ch^{k+1}\|\nabla_\Gamma u\|_{L_2(\Gamma)},$$

which when combined with (3.3) and (2.23) completes the proof of (4.2).

Noting that $\|F\|_{H^{-2}} \leq \|F\|_{H^{-1}}$ and then inserting (4.4) into (3.5) while recalling (2.23) completes the proof of (4.3). \square

We now give pointwise error estimates.

COROLLARY 4.5. *Let \tilde{u}_{hk} satisfy (2.29) with $f_h = \mu_{hk}f^\ell$. Let also $0 \leq s \leq r - 1$ and $0 \leq t \leq r$. Then for any $x \in \Gamma$,*

$$(4.5) \quad \begin{aligned} &|(u - \tilde{u}_{hk}^\ell)(x) - \overline{u - \tilde{u}_{hk}^\ell}| \\ &\leq C\ell_{h,s} \inf_{\chi \in S_{hk}^r} (h\|\nabla_\Gamma(u - \chi)\|_{L_\infty,x,s} + \|u - \chi\|_{L_\infty,x,s}) + Ch^{k+1}\ell_h\|\nabla_\Gamma u\|_{L_\infty(\Gamma)}, \end{aligned}$$

$$(4.6) \quad |\nabla_\Gamma(u - \tilde{u}_{hk}^\ell)(x)| \leq C(\ell_{h,t} \inf_{\chi \in S_{hk}^r} \|\nabla_\Gamma(u - \chi)\|_{L_\infty,x,t} + h^{k+1}\ell_h\|\nabla_\Gamma u\|_{L_\infty(\Gamma)}).$$

Here C depends on d and its derivatives, and ℓ_h , $\ell_{h,t}$, and $\ell_{h,s}$ are defined as in Theorem 3.2.

Proof. We recall that $F(\chi) = \int_\Gamma(\mathbf{A}_\Gamma - \mathbf{P})\nabla_\Gamma\tilde{u}_{hk}^\ell\nabla_\Gamma\chi\,d\sigma$ and then use (3.15) with $t = 0$ and (4.1) to find that for h small enough,

$$\begin{aligned} \|\nabla_\Gamma\tilde{u}_{hk}^\ell\|_{L_\infty(\Gamma)} &\leq C(\|\nabla_\Gamma u\|_{L_\infty(\Gamma)} + \ell_h\|\mathbf{A}_\Gamma - \mathbf{P}\|_{L_\infty(\Gamma)}\|\nabla_\Gamma\tilde{u}_{hk}^\ell\|_{L_\infty(\Gamma)}) \\ &\leq C(\|\nabla_\Gamma u\|_{L_\infty(\Gamma)} + h^{k+1}\ell_h\|\nabla_\Gamma\tilde{u}_{hk}^\ell\|_{L_\infty(\Gamma)}) \\ &\leq C\|\nabla_\Gamma u\|_{L_\infty(\Gamma)}. \end{aligned}$$

Here we have kicked back the last term on the right-hand side by taking h sufficiently small. Thus $\|F\|_{W_\infty^{-1},x,s} + \|F\|_{W_\infty^{-2}} \leq Ch^{k+1}\|\nabla_\Gamma u\|_{L_\infty(\Gamma)}$, which when inserted into (3.14) and (3.16) yields (4.5) and (4.6), respectively. \square

Taking the maximum of (4.5) and (4.6) with $t = s = 0$ leads to standard quasi-optimal pointwise error estimates. In addition, one can easily use (2.23) and elementary manipulations to prove asymptotic error expansion inequalities similar to those given in [Sch98] for domains in \mathbb{R}^n .

COROLLARY 4.6. *Under the conditions of Corollary 4.5,*

$$\|u - \tilde{u}_{hk}^\ell - \overline{u - \tilde{u}_{hk}^\ell}\|_{L_\infty(\Gamma)} \leq C(\tilde{\ell}_h h^{r+1} \|u\|_{W_\infty^{r+1}(\Gamma)} + Ch^{k+1} \ell_h \|\nabla_\Gamma u\|_{L_\infty(\Gamma)}),$$

$$\|\nabla_\Gamma(u - \tilde{u}_{hk}^\ell)\|_{L_\infty(\Gamma)} \leq C(h^r \|u\|_{W_\infty^{r+1}(\Gamma)} + Ch^{k+1} \ell_h \|\nabla_\Gamma u\|_{L_\infty(\Gamma)}),$$

where $\tilde{\ell}_h = \ell_h$ if $r = 1$ and $\tilde{\ell}_h = 1$ otherwise. In addition for $0 \leq s \leq r - 1$, $0 \leq t \leq r$, and $x \in \Gamma$,

$$\begin{aligned} |(u - \tilde{u}_{hk}^\ell)(x) - \overline{u - \tilde{u}_{hk}^\ell}| &\leq C\ell_{h,s} h^{r+1} \left[\sum_{1 \leq |\beta| \leq r+1} |D_\Gamma^\beta u(x)| \right. \\ &\quad \left. + \sum_{r+2 \leq |\beta| \leq r+s} h^{|\beta|-r-1} |D_\Gamma^\beta u(x)| + h^s \|u\|_{W_\infty^{r+1+s}(\Gamma)} \right], \\ |\nabla_\Gamma(u - \tilde{u}_{hk}^\ell)(x)| &\leq C\ell_{h,r} h^r \left[\sum_{1 \leq |\beta| \leq r+1} |D_\Gamma^\beta u(x)| \right. \\ &\quad \left. + \sum_{r+2 \leq |\beta| \leq r+t} h^{|\beta|-r-1} |D_\Gamma^\beta u(x)| + h^t \|u\|_{W_\infty^{r+1+t}(\Gamma)} \right]. \end{aligned}$$

4.2. Error estimates for finite element methods defined on Γ . In order to obtain error estimates for the method (2.30), we simply apply Theorems 3.1 and 3.2 with $F \equiv 0$ while recalling (2.23).

COROLLARY 4.7. *Let $u_{h,\Gamma}$ defined by (2.30), and assume $u \in H^{r+1}(\Gamma)$. Then*

$$\begin{aligned} \|\nabla_\Gamma(u - u_{h,\Gamma})\|_{L_2(\Gamma)} &\leq Ch^r \|u\|_{H^{r+1}(\Gamma)}, \\ \|u - u_{h,\Gamma}\|_{L_2(\Gamma)} &\leq Ch^{r+1} \|u\|_{H^{r+1}(\Gamma)}. \end{aligned}$$

For $x \in \Gamma$, $0 \leq s \leq r - 1$, and $0 \leq t \leq r$,

$$\begin{aligned} |(u - u_{h,\Gamma})(x)| &\leq C\ell_{h,s} \inf_{\chi \in S_h^r} (h \|\nabla_\Gamma(u - \chi)\|_{L_\infty,x,s} + \|u - \chi\|_{L_\infty,x,s}), \\ |\nabla_\Gamma(u - u_{h,\Gamma})(x)| &\leq C\ell_{h,t} \inf_{\chi \in S_h^r} \|\nabla_\Gamma(u - \chi)\|_{L_\infty,x,t}. \end{aligned}$$

Here $\ell_{h,s}$ and $\ell_{h,t}$ are as defined in Theorem 3.2.

4.3. Numerical experiments. In our numerical experiments we let $\Gamma = \{x \in \mathbb{R}^3 : x_1^2 + x_2^2 + \frac{x_3^2}{9} = 1\}$; that is, Γ is an ellipsoid having principal axes of lengths 1, 1, and 3. Also, we let $u = x_1$. (Note that $\Delta_\Gamma u \neq 0$ on Γ , even though $u(x) = x_1$ is a harmonic function on \mathbb{R}^3 .) Computations were performed on a sequence of uniformly refined meshes in all cases, with high-order quadrature being employed. We refer to [DD07] for more implementation details, in particular the numerical approximation of a when, as in the current case, d is not explicitly available. All methods were implemented using the finite element toolbox ALBERTA [SS05].

In Figure 1 we display plots of $\|\nabla_\Gamma(u - u_h)\|_{L_2(\Gamma)}$ versus the number of degrees of freedom (DOF), where $u_h = \tilde{u}_{h1}^\ell$, $u_h = \tilde{u}_{h2}^\ell$, and $u_h = u_{h,\Gamma}$ are the finite element approximations defined on a polyhedral approximation to Γ (via (2.29) with $k = 1$), a quadratic approximation to Γ (via (2.29) with $k = 2$), and Γ (via (2.30)), respectively.

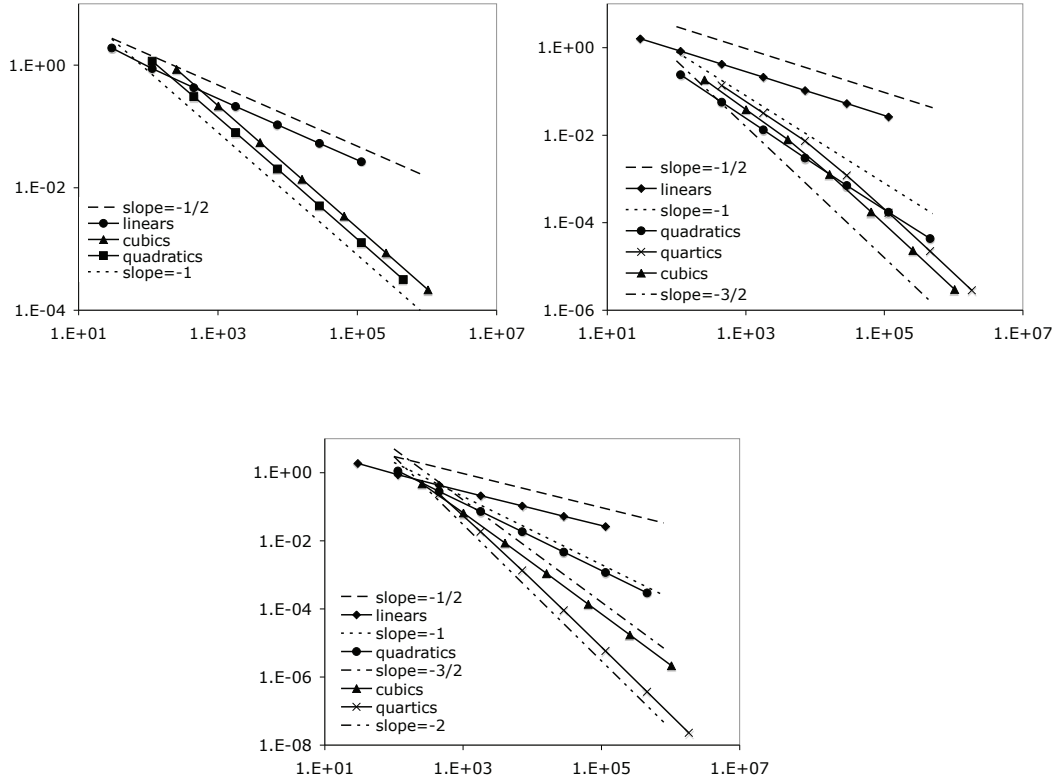


FIG. 1. Plots of $\|\nabla_{\Gamma}(u - u_h)\|_{L_2(\Gamma)}$ vs. the number of degrees of freedom: Finite element method defined on Γ_h (upper left), Γ_h^2 (upper right), and Γ (bottom).

Optimal-order decrease for $\|\nabla_{\Gamma}(u - u_h)\|_{L_2(\Gamma)}$ is $DOF^{-r/2}$, so we display logarithmic lines of various slopes for comparison with computed error trends.

The effect of the geometric error is clearly seen. When $k = 1$ (upper left of Figure 1), we obtain optimal-order convergence when $r = 1$ and $r = 2$ so that $h^{k+1} \leq h^r$. Suboptimal convergence is obtained when $r \geq 3$, as expected. When $k = 2$ (upper right) we obtain optimal convergence for $r \leq 3$, but not for $r = 4$. Thus (4.2) is sharp with respect to the geometric error $h^{k+1}\|\nabla_{\Gamma}u\|_{L_2(\Gamma)}$. Finally, in the bottom plot of Figure 1 we observe optimal-order convergence for all polynomial degrees $r \leq 4$ when defining the finite element method directly on Γ via (2.30). We note, however, that our experiments use high-order quadrature, and the quadrature error is likely to be more pronounced when using (2.30) in practical situations, as this formulation essentially involves an elliptic problem with a nonconstant coefficient matrix.

Similar plots of the L_2 error on linear and quadratic surface approximations are displayed in Figure 2. These plots confirm the sharpness of the error estimate (4.3).

5. Extensions. In this section we briefly discuss extensions of our methods and analysis to more general situations.

5.1. More general surface approximations. Our definitions in section 2 require that the nodes of the discrete surfaces Γ_h and Γ_h^k lie on Γ . This is a reasonable assumption for stationary problems, but not for geometric evolution problems

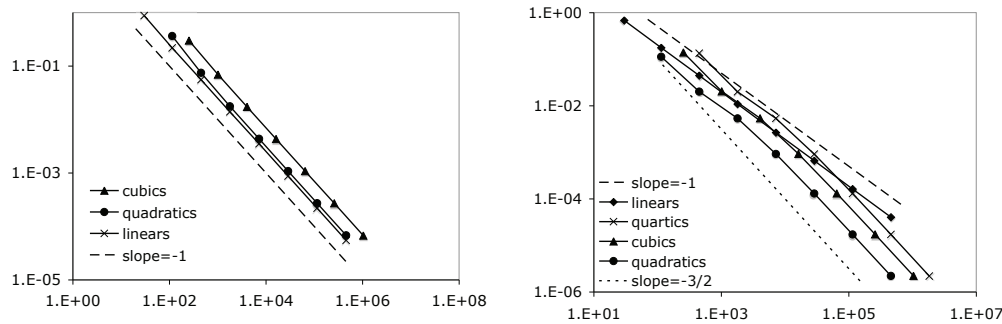


FIG. 2. Plots of $\|u - u_h - \overline{u - u_h}\|_{L_2(\Gamma)}$ vs. the number of degrees of freedom: Finite element method defined on Γ_h (left) and Γ_h^2 (right).

such as mean curvature flow where the goal is to approximate an unknown surface Γ (cf. [Dz91]). Instead of assuming that the nodes of the discrete surfaces lie on Γ , it is reasonable to assume that they lie within $O(h^{k+1})$ of Γ ; cf. the comments at the beginning of section 3.

5.2. Surfaces with boundary. Our development may be carried out for surfaces Γ with boundary $\partial\Gamma$ modulo “variational crimes” that arise when $S^r \not\subset H^1(\Gamma)$, just as for domains in \mathbb{R}^n . Note that variational crimes do not arise if $\partial\Gamma$ is “curvi-polygonal” in the sense that $a(\partial\Gamma_h) = \partial\Gamma$ (cf. [DD07]). In a few situations, $\partial\Gamma$ may be both smooth and “curvi-polygonal” in this sense (e.g., if Γ is a half-sphere).

5.3. General second-order elliptic PDE. Many applications involve general second-order linear elliptic problems of the form $-\operatorname{div}_\Gamma(\mathcal{D}\nabla_\Gamma u) + \tilde{\mathbf{b}} \cdot \nabla_\Gamma u + cu = f$. If we make the natural assumption that $\mathcal{D}\vec{\tau} \cdot \vec{\nu} = \tilde{\mathbf{b}} \cdot \vec{\nu} = 0$ for $\vec{\tau} \cdot \vec{\nu} = 0$ (cf. [DE07b]), then the H^1 and L_2 error estimates of sections 3 and 4 hold for this problem if the associated bilinear form is coercive and the coefficients sufficiently smooth. In particular, one can show that the geometric error is still of order h^{k+1} in the more general case. Our pointwise estimates hold if a Green’s function satisfying the identities and inequalities in Lemma 2.2 exists (note that [Aub82] considers only the Laplace–Beltrami operator).

5.4. C^2 surfaces. In many situations of interest, Γ is not infinitely differentiable. The essential assumption that the orthogonal projection a exists generally requires that Γ be C^2 , and situations where Γ is less regular cannot be considered without substantial modification to our methodology. If Γ is merely C^2 , the abstract energy and L_2 error estimates of Theorem 3.1 hold verbatim, but the order of the geometric error in Corollary 4.2 is naturally restricted by the smoothness of Γ . We also expect the abstract pointwise estimates of Theorem 3.2 to hold if Γ is only C^2 so long as $s = 0$ and $t \leq 1$. Proving such a statement using our techniques requires the establishment of pointwise estimates for the Green’s function as in Lemma 2.2. This can likely be accomplished using an elementary mapping argument, though we have not checked the details.

5.5. Manifolds. The abstract error analysis of section 3 relies on two classes of assumptions: those concerning the finite element triangulation and space, and those concerning the underlying PDEs. The PDE assumptions employed in section 3 hold with slight modification if one considers smooth Riemannian manifolds without boundary instead of smooth surfaces without boundary. Thus if one can construct

finite element spaces on manifolds satisfying the assumptions A1 through A4, the results of section 3 should hold as well.

REFERENCES

- [AP05] T. APEL AND C. PESTER, *Clement-type interpolation on spherical domains—interpolation error estimates and application to a posteriori error estimation*, IMA J. Numer. Anal., 25 (2005), pp. 310–336.
- [Aub82] T. AUBIN, *Nonlinear Analysis on Manifolds. Monge-Ampère Equations*, Grundlehren Math. Wiss. 252, Springer-Verlag, New York, 1982.
- [BMN05] E. BÄNSCH, P. MORIN, AND R. H. NOCHETTO, *A finite element method for surface diffusion: The parametric case*, J. Comput. Phys., 203 (2005), pp. 321–343.
- [BS02] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, 2nd ed., Texts Appl. Math. 15, Springer-Verlag, New York, 2002.
- [CDD+04] U. CLARENZ, U. DIEWALD, G. DZIUK, M. RUMPF, AND R. RUSU, *A finite element method for surface restoration with smooth boundary conditions*, Comput. Aided Geom. Design, 21 (2004), pp. 427–445.
- [CDR03] U. CLARENZ, U. DIEWALD, AND M. RUMPF, *A multiscale fairing method for textured surfaces*, in Visualization and Mathematics III, Math. Vis., Springer-Verlag, Berlin, 2003, pp. 245–260.
- [DDE05] K. DECKELNICK, G. DZIUK, AND C. M. ELLIOTT, *Computation of geometric partial differential equations and mean curvature flow*, Acta Numer., 14 (2005), pp. 139–232.
- [De04] A. DEMLOW, *Piecewise linear finite element methods are not localized*, Math. Comp., 73 (2004), pp. 1195–1201.
- [De07] A. DEMLOW, *Sharply localized pointwise and W_∞^{-1} estimates for finite element methods for quasilinear problems*, Math. Comp., 76 (2007), pp. 1725–1741.
- [DD07] A. DEMLOW AND G. DZIUK, *An adaptive finite element method for the Laplace–Beltrami operator on implicitly defined surfaces*, SIAM J. Numer. Anal., 45 (2007), pp. 421–442.
- [Dz88] G. DZIUK, *Finite elements for the Beltrami operator on arbitrary surfaces*, in Partial Differential Equations and Calculus of Variations, Lecture Notes in Math. 1357, Springer-Verlag, Berlin, 1988, pp. 142–155.
- [Dz91] G. DZIUK, *An algorithm for evolutionary surfaces*, Numer. Math., 58 (1991), pp. 603–611.
- [DE07a] G. DZIUK AND C. M. ELLIOTT, *Finite elements on evolving surfaces*, IMA J. Numer. Anal., 27 (2007), pp. 262–292.
- [DE07b] G. DZIUK AND C. M. ELLIOTT, *Surface finite elements for parabolic equations*, J. Comput. Math., 25 (2007), pp. 385–407.
- [GT98] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, 1998.
- [He05] C.-J. HEINE, *Isoparametric Finite Element Approximation of Curvature on Hypersurfaces*, preprint, 2005.
- [He06] C.-J. HEINE, *Computations of form and stability of rotating drops with finite elements*, IMA J. Numer. Anal., 26 (2006), pp. 723–751.
- [Ho01] M. HOLST, *Adaptive numerical treatment of elliptic systems on manifolds. A posteriori error estimation and adaptive computational methods*, Adv. Comput. Math., 15 (2001), pp. 139–191.
- [Ne76] J.-C. NÉDÉLEC, *Curved finite element methods for the solution of singular integral equations on surfaces in R^3* , Comput. Methods Appl. Mech. Engrg., 8 (1976), pp. 61–80.
- [NS74] J. A. NITSCHKE AND A. H. SCHATZ, *Interior estimates for Ritz–Galerkin methods*, Math. Comp., 28 (1974), pp. 937–958.
- [Sch98] A. H. SCHATZ, *Pointwise error estimates and asymptotic error expansion inequalities for the finite element method on irregular grids. I. Global estimates*, Math. Comp., 67 (1998), pp. 877–899.
- [SW95] A. H. SCHATZ AND L. B. WAHLBIN, *Interior maximum-norm estimates for finite element methods, Part II*, Math. Comp., 64 (1995), pp. 907–928.
- [SS05] A. SCHMIDT AND K. G. SIEBERT, *Design of Adaptive Finite Element Software*, Lect. Notes Comput. Sci. Eng. 42, Springer-Verlag, Berlin, 2005.

CONVERGENCE ANALYSIS OF PROJECTION METHODS FOR THE NUMERICAL SOLUTION OF LARGE LYAPUNOV EQUATIONS*

V. SIMONCINI[†] AND V. DRUSKIN[‡]

Abstract. The numerical solution of large-scale continuous-time Lyapunov matrix equations is of great importance in many application areas. Assuming that the coefficient matrix is positive definite, but not necessarily symmetric, in this paper we analyze the convergence of projection-type methods for approximating the solution matrix. Under suitable hypotheses on the coefficient matrix, we provide new asymptotic estimates for the error matrix when a Galerkin method is used in a Krylov subspace. Numerical experiments confirm the good behavior of our upper bounds when linear convergence of the solver is observed.

Key words. Lyapunov equation, Krylov subspace, matrix exponential, Faber polynomials

AMS subject classifications. 65F10, 93B40

DOI. 10.1137/070699378

1. The problem. We are interested in the approximate solution of the following Lyapunov matrix equation:

$$(1.1) \quad AX + XA^\top = BB^\top,$$

with A a real matrix of large dimension and B a real tall matrix. Here A^\top indicates the transpose of A . We assume that the $n \times n$ matrix A is either symmetric and positive definite or nonsymmetric with positive definite symmetric part, that is, $(A + A^\top)/2$ is positive definite. In the following we mostly deal with the case of B having a single column, that is, $B = b$, and we assume that b has unit Euclidean norm, that is, $\|b\| = 1$. Nonetheless, our results can be extended to the multiple vector case.

This problem arises in a large variety of applications, such as signal processing and system and control theory. The symmetric solution X carries important information on the stability and energy of an associated dynamical linear system and on the feasibility of order reduction techniques [2], [6], [8]. The analytic solution of (1.1) can be written as

$$(1.2) \quad X = \int_0^\infty e^{-tA} BB^\top e^{-tA^\top} dt = \int_0^\infty xx^\top dt,$$

where we have set $x = e^{-tA}B$. Let α_{\min} be the smallest eigenvalue of the symmetric part of A , $\alpha_{\min} = \lambda_{\min}((A + A^\top)/2) > 0$. Then it can be shown that $\|x\| \leq \exp(-t\alpha_{\min})\|B\|$; see, e.g., [8, Lemma 3.2.1].

Projection-type methods seek an approximate solution X_m in a subspace of \mathbb{R}^n by requiring, e.g., that the residual $BB^\top - (AX_m + X_mA^\top)$ be orthogonal to this subspace. A particularly effective choice as approximation space is given by (here for $B = b$) the Krylov subspace $K_m(A, b) = \text{span}\{b, Ab, \dots, A^{m-1}b\}$ of dimension $m \leq n$

*Received by the editors August 8, 2007; accepted for publication (in revised form) October 14, 2008; published electronically February 6, 2009.

<http://www.siam.org/journals/sinum/47-2/69937.html>

[†]Dipartimento di Matematica, Università di Bologna, Piazza di Porta S. Donato 5, I-40127 Bologna, Italy (valeria@dm.unibo.it).

[‡]Schlumberger-Doll Research, 1 Hampshire St., Cambridge, MA 02139 (druskin1@boston.oilfield.slb.com).

[22], [23], [31]; we also refer to a richer bibliographic account collected in [2], [9], while we point to [33] for recent algorithmic progress within the Krylov subspace context. Abundant experimental evidence over the years has shown that the use of the space $K_m(A, b)$ allows one to often obtain a satisfactorily accurate approximation X_m , in a space of much lower dimension than n . A particularly attractive feature is that X_m may be written as a low rank matrix, $X_m = U_m U_m^\top$ with U_m of low column rank, so that only the matrix U_m needs to be stored.

To the best of our knowledge, no asymptotic convergence analysis of this Galerkin method is available in the literature. The aim of this paper is to fill this gap. We also refer to [30] for a priori estimates on the residual norm when solving the Sylvester equation with projection-type methods; there, the role of α_{\min} is also emphasized, although the bound derived in [30, Proposition 4.1] for the residual norm is of greater value as a nonstagnation condition of the procedure, rather than as an estimate of the actual convergence behavior.

To derive our error estimates, we shall use the integral representation (1.2) for both X and X_m and explicitly bound the norm of the error matrix $X - X_m$; we refer to [31] for early considerations in this direction. Our approach is highly inspired by, and fully relies on, the papers [13], [24], where general estimates for the error in approximating matrix operators by polynomial methods are derived. We provide explicit estimates when A is symmetric, and when A is nonsymmetric with its field of values (or spectrum) contained in certain not necessarily convex sets of \mathbb{C}^+ .

We also show that the convergence of the Galerkin method is closely related to that of Galerkin methods for solving the linear system $(A + \alpha_{\min} I)d = b$.

Our estimates are asymptotic, and thus linear; that is, they do not capture the possibly superlinear convergence behavior of the method that is sometimes observed [29]. In the linear system setting, the superlinear behavior is due to the fact that Krylov-based methods tend to adapt to the (discrete) spectrum of A , accelerating convergence as spectral information is gained while enlarging the space. Recent results for A symmetric have been derived, which completely describe the behavior of Krylov subspace solvers in the presence of superlinear convergence [4], [5]; see also [34] for a discussion and more references.

Throughout the paper we assume exact arithmetic.

2. Numerical solution and preliminary considerations. Given the Krylov subspace $K_m(A, b)$ and a matrix V_m whose orthonormal columns span $K_m(A, b)$, with $b = V e_1$, we seek an approximation in the form $X_m = V_m Y_m V_m^\top$. Here and in the following, e_i denotes the i th column of the identity matrix of given dimension. Imposing that the residual $R_m = b b^\top - (A X_m + X_m A^\top)$ be orthogonal to the given space, the so-called Galerkin condition, yields the equation

$$V_m^\top R_m V_m = 0 \quad \Leftrightarrow \quad T_m Y + Y T_m^\top = e_1 e_1^\top,$$

where $T_m = V_m^\top A V_m$; see, e.g., [2], [31]. The $m \times m$ matrix Y_m can thus be computed by solving the resulting small-size Lyapunov equation.

The matrix X_m can be equivalently written in integral form. Indeed, let $x_m = x_m(t) = V_m e^{-t T_m} e_1$ be the so-called Krylov approximation to $x = x(t)$ in $K_m(A, b)$. Then X_m can be written as

$$\begin{aligned} X_m &= V_m \left(\int_0^\infty e^{-t T_m} e_1 e_1^\top e^{-t T_m^\top} dt \right) V_m^\top \\ &= \int_0^\infty V_m e^{-t T_m} e_1 e_1^\top e^{-t T_m^\top} V_m^\top dt = \int_0^\infty x_m x_m^\top dt. \end{aligned}$$

We are interested in finding a priori bounds for the 2-norm of the error matrix, that is, for $\|X - X_m\|$, where the 2-norm is the matrix norm induced by the vector Euclidean norm. We start by observing that $\|X - X_m\| = \|\int_0^\infty (xx^\top - x_m x_m^\top) dt\|$, and that

$$\|xx^\top - x_m x_m^\top\| = \|x(x - x_m)^\top + (x - x_m)x_m^\top\| \leq (\|x\| + \|x_m\|)\|x - x_m\|.$$

It holds that $\lambda_{\min}((T_m + T_m^\top)/2) \geq \alpha_{\min}$. Using $\|x_m\| \leq \exp(-t\lambda_{\min}((T_m + T_m^\top)/2)) \leq \exp(-t\alpha_{\min})$, we have

$$\begin{aligned} \|X - X_m\| &\leq \int_0^\infty \|xx^\top - x_m x_m^\top\| dt \leq \int_0^\infty (\|x\| + \|x_m\|)\|x - x_m\| dt \\ (2.1) \quad &\leq 2 \int_0^\infty e^{-t\alpha_{\min}} \|x - x_m\| dt. \end{aligned}$$

We notice that

$$\begin{aligned} e^{-t\alpha_{\min}} \|x - x_m\| &= \|\exp(-t(A + \alpha_{\min}I))b - V_m \exp(-t(T_m + \alpha_{\min}I))e_1\| \\ &=: \|\hat{x} - \hat{x}_m\|, \end{aligned}$$

which is the error in the approximation of the exponential of the *shifted* matrix $A + \alpha_{\min}I$ with the Krylov subspace solution. Therefore,

$$(2.2) \quad \|X - X_m\| \leq 2 \int_0^\infty \|\hat{x} - \hat{x}_m\| dt.$$

In the following we will bound $\|X - X_m\|$ by judiciously integrating an upper bound of the integrand function. In fact, estimates for the error norm $\|\hat{x} - \hat{x}_m\|$ are available in the literature, which show superlinear convergence of the Krylov approximation x_m to the exponential vector x ; see, e.g., [12], [39], [36], [21]. However, these bounds are not appropriate when used in the generalized integral above.

The matrix $V_m = [v_1, \dots, v_m]$ can be generated one vector at the time, by means of the following Arnoldi recursion:

$$(2.3) \quad AV_m = V_m T_m + v_{m+1} t_{m+1,m} e_m^\top, \quad v_1 = b/\|b\|,$$

where $V_{m+1} = [V_m, v_{m+1}]$ has orthonormal columns and spans $K_{m+1}(A, b)$. In general, T_m is upper Hessenberg, and it is symmetric, and thus tridiagonal, when A is itself symmetric.

We conclude this section with a technical lemma, whose proof is included for completeness; see, e.g., [24] for a similar result in finite precision arithmetic.

LEMMA 2.1. *Let P_k be a polynomial of degree at most k . Let $f(z) = \sum_{k=0}^\infty f_k P_k(z)$ be a convergent series expansion of the analytic function f and assume that the expansions of $f(A)$ and of $f(T_m)$ are also well defined. Then*

$$\|f(A)b - V_m f(T_m)e_1\| \leq \sum_{k=m}^\infty |f_k| (\|P_k(A)\| + \|P_k(T_m)\|).$$

Proof. We have

$$\begin{aligned} f(A)b - V_m f(T_m)e_1 &= \sum_{k=0}^{m-1} f_k (P_k(A)b - V_m P_k(T_m)e_1) \\ &\quad + \sum_{k=m}^\infty f_k (P_k(A)b - V_m P_k(T_m)e_1). \end{aligned}$$

Using the Arnoldi relation and the fact that T_m is upper Hessenberg, $A^k V_m e_1 = V_m T_m^k e_1$ for $k = 1, \dots, m - 1$, and thus $P_k(A)b = P_k(A)V_m e_1 = V_m P_k(T_m)e_1$, $k = 1, \dots, m - 1$, so that

$$f(A)b - V_m f(T_m)e_1 = \sum_{k=m}^{\infty} f_k(P_k(A)b - V_m P_k(T_m)e_1).$$

Taking norms, the result follows. \square

3. The symmetric case. In the symmetric case, we show that the asymptotic convergence rate of the Krylov subspace solver is the same as that of the conjugate gradient method applied to the shifted system $(A + \alpha_{\min}I)x = b$, where $\alpha_{\min} = \lambda_{\min}$, the smallest eigenvalue of the positive definite matrix A [18]; see also section 5.

PROPOSITION 3.1. *Let A be symmetric and positive definite, and let λ_{\min} be the smallest eigenvalue of A . Let $\hat{\lambda}_{\min}, \hat{\lambda}_{\max}$ be the extreme eigenvalues of $A + \lambda_{\min}I$ and $\hat{\kappa} = \hat{\lambda}_{\max}/\hat{\lambda}_{\min}$. Then*

$$(3.1) \quad \|X - X_m\| \leq 4 \frac{\sqrt{\hat{\kappa}} + 1}{\hat{\lambda}_{\min} \sqrt{\hat{\kappa}}} \left(\frac{\sqrt{\hat{\kappa}} - 1}{\sqrt{\hat{\kappa}} + 1} \right)^m.$$

Proof. Using (2.1) we are left to estimate $\int_0^\infty e^{-t\alpha_{\min}} \|x - x_m\| dt$. Let λ_{\max} be the largest eigenvalue of A . Formula (4.2) in [12] shows that both x and x_m may be written as Chebyshev series,¹ e.g., for x we have

$$x = 2 \exp\left(-t \frac{\lambda_{\max} + \lambda_{\min}}{2}\right) \sum_{k=0}^{\infty} I_k\left(t \frac{\lambda_{\max} - \lambda_{\min}}{2}\right) \mathcal{T}_k(A')b,$$

where I_k is the Bessel function of an imaginary argument, or modified Bessel function, \mathcal{T}_k is the Chebyshev polynomial of degree k , and $A' = (\lambda_{\max} + \lambda_{\min})/(\lambda_{\max} - \lambda_{\min})I - 2/(\lambda_{\max} - \lambda_{\min})A$ so that $\|\mathcal{T}_k(A')\| \leq 1$ holds; see also [1, formula (9.6.34)]. Since polynomials of degree up to $k - 1$ are exactly represented in the Krylov subspace of dimension k (see [12] and also Lemma 2.1), it thus follows that

$$\|x - x_m\| \leq 4 \exp\left(-t \frac{\lambda_{\max} + \lambda_{\min}}{2}\right) \sum_{k=m}^{\infty} I_k\left(t \frac{\lambda_{\max} - \lambda_{\min}}{2}\right).$$

Therefore, setting $p = (3\lambda_{\min} + \lambda_{\max})/(\lambda_{\max} - \lambda_{\min}) = (\hat{\kappa} + 1)/(\hat{\kappa} - 1)$ and $\rho = p + \sqrt{p^2 - 1}$, we have

$$(3.2) \quad \begin{aligned} \|X - X_m\| &\leq 2 \int_0^\infty \|\hat{x} - \hat{x}_m\| dt \\ &\leq 8 \sum_{k=m}^{\infty} \int_0^\infty \exp\left(-t \left(\frac{3}{2}\lambda_{\min} + \frac{1}{2}\lambda_{\max}\right)\right) I_k\left(t \frac{\lambda_{\max} - \lambda_{\min}}{2}\right) dt \\ &= \frac{8}{\sqrt{\left(\frac{3}{2}\lambda_{\min} + \frac{\lambda_{\max}}{2}\right)^2 - \frac{(\lambda_{\max} - \lambda_{\min})^2}{4}}} \sum_{k=m}^{\infty} \frac{1}{\left(p + \sqrt{p^2 - 1}\right)^k} \\ &= \frac{8(\hat{\kappa} + 1)}{\sqrt{\hat{\kappa}}(3\lambda_{\min} + \lambda_{\max})} \sum_{k=m}^{\infty} \frac{1}{\left(p + \sqrt{p^2 - 1}\right)^k} \\ &= \frac{4(\hat{\kappa} + 1)}{\sqrt{\hat{\kappa}}(3\lambda_{\min} + \lambda_{\max})} \frac{2\rho}{\rho - 1} \frac{1}{\rho^m}. \end{aligned}$$

¹The prime in the series indicates that the first term is divided by two.

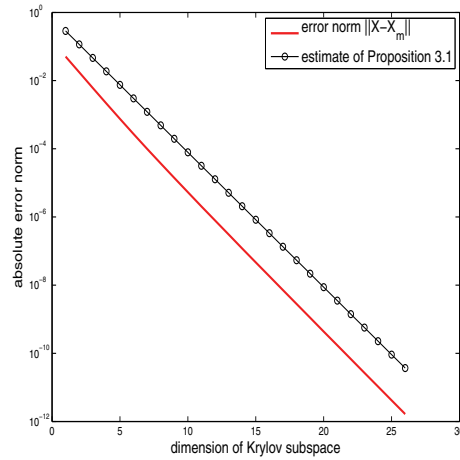


FIG. 3.1. Example of section 3. 400×400 diagonal matrix with uniformly distributed eigenvalues in $[1, 10]$. True error norm and its estimate of Proposition 3.1 for the Krylov subspace approximation of the Lyapunov solution.

To get (3.2) we used the following integral formula for Bessel functions in [19, Formula (6.611.4)]:

$$\int_0^\infty e^{-\alpha t} I_\nu(\beta t) dt = \frac{\beta^\nu}{\sqrt{\alpha^2 - \beta^2}(\alpha + \sqrt{\alpha^2 - \beta^2})^\nu} \quad \text{for } \Re \nu > -1 \text{ and } \Re \alpha > |\Re \beta|.$$

Standard algebraic manipulations give

$$\rho = \frac{\sqrt{\hat{\kappa}} + 1}{\sqrt{\hat{\kappa}} - 1}, \quad \frac{2\rho}{\rho - 1} = \sqrt{\hat{\kappa}} + 1. \quad \square$$

In Figure 3.1 we report the behavior of the bound of Proposition 3.1 for a 400×400 diagonal matrix A having uniformly distributed eigenvalues between 1 and 10. Here $\alpha_{\min} = \lambda_{\min} = 1$. The vector b is the normalized vector of all ones.

We explicitly observe that the linearity of the convergence rate is exactly reproduced by the upper bound of Proposition 3.1.

4. The nonsymmetric case. For A nonsymmetric, the result of the previous section can be generalized whenever the field of values of A is contained in a “well-behaved” set of the complex plane. We recall that the field of values of a real matrix A in the Euclidean inner product is defined as $F(A) = \{x^* Ax, x \in \mathbb{C}^n, \|x\| = 1\}$, where x^* is the conjugate transpose of x . The location of the field of values plays a crucial role in the behavior and analysis of polynomial-type methods for the solution of linear systems; see, e.g., [15], [27].

The following results make use of the theory of Faber polynomials and of recently obtained results that have been used in the context of linear systems. To this end, we need some definitions on conformal mappings. Let $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$, and let $D(0, 1) = \{|\tau| \leq 1\}$ be the closed unit disk centered at zero. Given a bounded set Ω such that its complement is simply connected, define the conformal mapping ϕ that maps the complement of Ω onto the exterior of the unit disk $D(0, 1)$, and such that $\phi(\infty) = \infty$ and $\phi'(\infty) > 0$; see, e.g., [35, section 1.2]. Let ψ denote the inverse of ϕ .

The principal (polynomial) part of the Laurent series of ϕ^k is the Faber polynomial Φ_k , of exact degree k . Under these hypotheses, it was recently shown by Beckermann

that for any z in a convex and compact set of \mathbb{C} , it holds that $|\Phi_k(z)| \leq 2$. Assume that $f(\lambda) = \exp(-\lambda t)$ is regular in $\Omega = \psi(D(0, r_2))$, and let

$$f(\lambda) \equiv \exp(-\lambda t) = \sum_{k=0}^{\infty} f_k \Phi_k(\lambda)$$

be the expansion of $\exp(-\lambda t)$ in Faber series in Ω with $1 < r_2 < \infty$. For $1 < r < r_2$, the expansion coefficients are given as

$$(4.1) \quad f_k = \frac{1}{2\pi i} \int_{|\tau|=r} \frac{\exp(-t\psi(\tau))}{\tau^{k+1}} d\tau, \quad |f_k| \leq \frac{1}{r^k} \sup_{|\tau|=r} |\exp(-t\psi(\tau))|;$$

see, e.g., [35, sect. 2.1.3], [37]. Note that $f_k = \overline{f_k(t)}$.

4.1. Field of values contained in an ellipse. The case in which the field of values is contained in an ellipse is a particularly natural generalization of the symmetric case.

PROPOSITION 4.1. *Assume the field of values of the real matrix A is contained in the ellipse $E \subset \mathbb{C}^+$ of center $(c, 0)$, foci $(c \pm d, 0)$, and semiaxes a_1 and a_2 , so that $d = \sqrt{a_1^2 - a_2^2}$. Then*

$$\|X - X_m\| \leq \frac{8}{\sqrt{(\alpha_{\min} + c)^2 - d^2}} \frac{r_2}{r_2 - 1} \left(\frac{1}{r_2}\right)^m,$$

where $r_2 = \frac{c + \alpha_{\min}}{2r} + \frac{1}{2r} \sqrt{(c + \alpha_{\min})^2 - d^2}$, and $r = \frac{a_1 + a_2}{2}$.

Proof. For $\lambda \in E$, and setting $\hat{r} = 2r/d$, we have $\Phi_k(\lambda) = 2(\hat{r})^{-k} T_k(\frac{\lambda - c}{d})$ (see, e.g., [38], [17]); therefore we can explicitly write the Faber series on E via Chebyshev ones as

$$e^{-\lambda t} = 2 \exp(-tc) \sum_{k=0}^{\infty} I_k(td) T_k\left(\frac{\lambda - c}{d}\right) = \exp(-tc) \sum_{k=0}^{\infty} I_k(td) \hat{r}^k \Phi_k(\lambda).$$

Using Lemma 2.1, the bounds $\|\Phi_k(T_m)\| \leq 2$, $\|\Phi_k(A)\| \leq 2$ obtained in [3], and the same integral formula for Bessel functions as in the proof of Proposition 3.1, we obtain

$$\begin{aligned} & \|X - X_m\| \\ & \leq 2 \int_0^{\infty} \|\hat{x} - \hat{x}_m\| dt \\ & \leq 8 \sum_{k=m}^{\infty} \int_0^{\infty} e^{(-\alpha_{\min} - c)t} I_k(td) \hat{r}^k dt \\ & = \frac{8}{\sqrt{(\alpha_{\min} + c)^2 - d^2}} \sum_{k=m}^{\infty} \left(\frac{1}{r_2}\right)^k = \frac{8}{\sqrt{(\alpha_{\min} + c)^2 - d^2}} \frac{r_2}{r_2 - 1} \left(\frac{1}{r_2}\right)^m. \quad \square \end{aligned}$$

We show the quality of the estimate with a few numerical examples.

Example 4.2. We consider a 400×400 (normal) diagonal matrix A whose eigenvalues are $\lambda = c + a_1 \cos \theta + ia_2 \sin \theta$, θ uniformly distributed in $[0, 2\pi]$ and $c = 20$, semiaxes $a_1 = 10$ and $a_2 = 2$, so that the eigenvalues are on an elliptic curve with center c and focal distance $d = \sqrt{a_1^2 - a_2^2} = \sqrt{96}$. Here $\alpha_{\min} \approx 10.001$, yielding

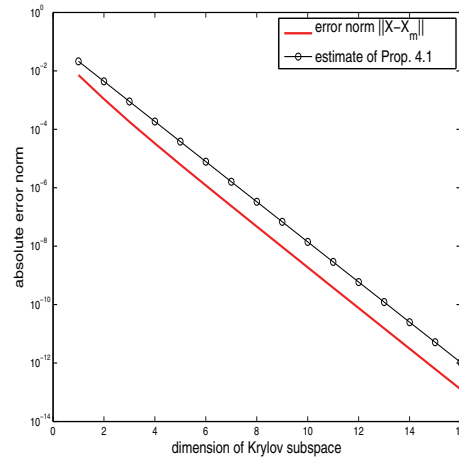


FIG. 4.1. Example 4.2. True error and its estimate of Proposition 4.1 for the Krylov subspace solver of the Lyapunov equation.

$1/r_2 \approx 0.2056$ for Proposition 4.1. The vector b is the vector of all ones, normalized to have unit norm. In Figure 4.1 we report the error associated with the Krylov subspace approximation of the Lyapunov solution, and the estimate of Proposition 4.1. The agreement is impressive, as should be expected since the spectrum lies exactly on the elliptic curve and the matrix is normal, so that the field of values coincides with the associated convex hull.

Example 4.3. We next consider the 400×400 matrix A stemming from the centered finite difference discretization of the operator $\mathcal{L}(u) = -\Delta u + 40(x+y)u_x + 200u$ in the unit square, with Dirichlet boundary conditions. The spectrum of A , together with its field of values (computed with the MATLAB function `fv.m` in [20]) and a surrounding ellipse, is shown in the left plot of Figure 4.2. Here $\alpha_{\min} = 0.4533$. The ellipse has parameters $c = 4.4535$, $a_1 = c - \alpha_{\min}$, $a_2 = 3.7$, a_1, a_2 being the semiaxes' length, and focal distance $d = \sqrt{a_1^2 - a_2^2} \approx 1.52$, yielding $1/r_2 \approx 0.8044$. The right plot of Figure 4.2 shows the convergence history of the Krylov solver, together with the asymptotic factor $(1/r_2)^m$ in Proposition 4.1. The initial asymptotic convergence rate is reasonably well captured by the estimate.

Example 4.4. We consider the 400×400 bidiagonal matrix A with uniformly distributed diagonal elements in the interval $[10, 110]$ and unit upper diagonal. In this case $\alpha_{\min} = 9.4692$. The vector b is the normalized vector of all ones. Our numerical computation reported in the left plot of Figure 4.3 showed that the field of values of A (computed once again with `fv.m` [20]) is contained in an ellipse with center $c = 60$, semiaxes $a_1 = 50.8$, $a_2 = 4.2$, and focal distance $d = \sqrt{a_1^2 - a_2^2} \approx 50.62$, yielding $1/r_2 = 0.4699$. The right plot of Figure 4.3 shows the convergence history of the Krylov solver, together with the asymptotic factor in Proposition 4.1. Once again, the asymptotic rate is a good estimate of the actual convergence rate. Even more accurate bounds for this example might be obtained by using more appropriate conformal mappings than the ellipse. It may be possible to include the field of values into a rectangle, for which the mapping ψ could be numerically estimated [14], [16]; see also Example 4.9.

4.2. Field of values contained in a more general region. For a more general region, we employ the general expansion in Faber series. We will proceed as

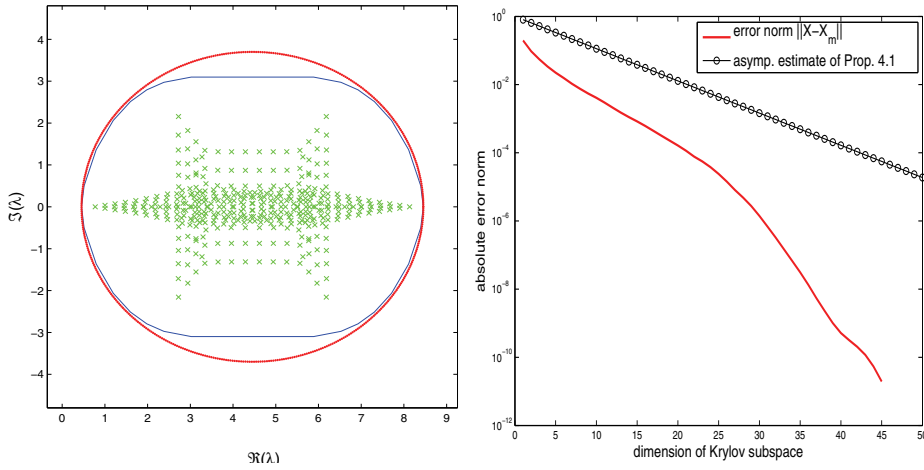


FIG. 4.2. Example 4.3. Left plot: Spectrum of A , field of values (thin solid curve), and smallest computed elliptic curve including the field of value (thick solid curve). Right plot: True error and its asymptotic factor in the estimate of Proposition 4.1 for the Krylov subspace solver of the Lyapunov equation.

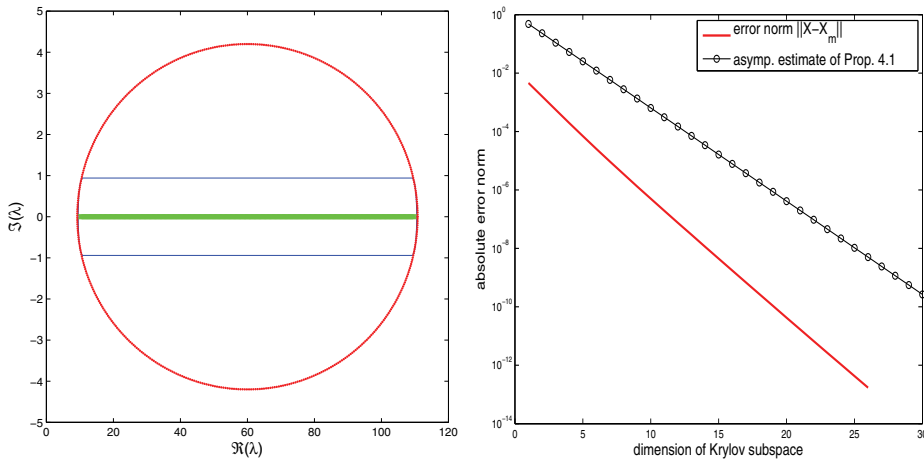


FIG. 4.3. Example 4.4. Left plot: Real spectrum, field of values (thin solid curve), and smallest computed elliptic curve including the field of value (thick solid curve). Right plot: True error and its estimate of Proposition 4.1 for the Krylov subspace solver of the Lyapunov equation.

follows. Using Lemma 2.1, we write

$$\|\hat{x} - \hat{x}_m\| \leq \sum_{k=m}^{\infty} |f_k| (\|\Phi_k(A + \alpha_{\min}I)\| + \|\Phi_k(T_m + \alpha_{\min}I)\|).$$

If we consider a convex set containing the field of values of $A + \alpha_{\min}I$, the result in [3] allows us to write $\|\Phi_k(A + \alpha_{\min}I)\| \leq 2$ and $\|\Phi_k(T_m + \alpha_{\min}I)\| \leq 2$, so that

$$\|\hat{x} - \hat{x}_m\| \leq 4 \sum_{k=m}^{\infty} |f_k|,$$

and we can conclude by using (4.1), once appropriate estimates for the sup function and for r_2 are identified. More precisely, if $\mathcal{M} = \mathcal{M}(t) > 0$ is such that $|f_k| \leq \mathcal{M}r_2^{-k}$ for all k , and $\int_0^\infty \mathcal{M}dt$ converges, then

$$\|X - X_m\| \leq 8 \left(\int_0^\infty \mathcal{M}dt \right) \frac{r_2}{r_2 - 1} \left(\frac{1}{r_2} \right)^m.$$

In the next few corollaries we derive a result of the same type, with a choice of r_2 such that the generalized integral converges.

In case we wish to work only with a set containing the spectrum, but not necessarily the field of values of $A + \alpha_{\min}I$, we can relax the convexity assumption and differently bound the norm of the Faber polynomials in A , at the price of keeping the condition number of the eigenvector matrix in the convergence estimate. This case will be analyzed at the end of this section, and one example will be given around Corollary 4.10.

We start by considering once again the case when the field of values is contained in an ellipse, for which the result is qualitatively the same as that in Proposition 4.1. The reason for reproducing the result in the case of the ellipse is precisely to appreciate the limited loss of accuracy given by the bound, when the more general approach is used, and to explicitly show the calculations in the case of an easy-to-handle mapping.

COROLLARY 4.5. *Assume the field of values of the real matrix A is contained in an ellipse $E \subset \mathbb{C}^+$ of center $(c, 0)$ and semiaxes a_1 and a_2 , $a_1 > a_2$. Let $\alpha_{\min} = \lambda_{\min}((A + A^T)/2)$. Then for ϵ satisfying $0 < \epsilon \leq 2\alpha_{\min}$,*

$$\|X - X_m\| \leq \frac{8}{\epsilon} \frac{r_2}{r_2 - 1} \left(\frac{1}{r_2} \right)^m,$$

where

$$r_2 = \frac{c + \alpha_{\min} - \epsilon}{2r} + \frac{1}{2r} \sqrt{(c + \alpha_{\min} - \epsilon)^2 - d^2}, \quad r = \frac{a_1 + a_2}{2}, \quad d = \sqrt{a_1^2 - a_2^2}.$$

Proof. Let $\alpha = \alpha_{\min}$ and let \hat{E} be the selected ellipse containing the field of values of $A + \alpha I$. We consider the mapping whose boundary image of the unit disk is $\partial \hat{E}$, $\psi(\tau) = c + \alpha + r\tau + \frac{(d/2)^2}{r\tau}$, with $\tau = e^{i\theta} \in D(0, 1)$, so that $\psi(|\tau| = 1) = \partial \hat{E}$. For $\epsilon > 0$, we define $r_2 := |\psi^{-1}(\epsilon)|$, so that

$$\exp(-t\epsilon) = \max_{|\tau|=r_2} |\exp(-t\psi(\tau))|,$$

and for $1 < \hat{r} < r_2$,

$$(4.2) \quad \frac{1}{2\pi} \int_0^{2\pi} |f(\psi(\hat{r}e^{i\theta}))| d\theta \leq \exp(-t\epsilon) =: \mathcal{M}(t).$$

Since \hat{E} is convex, it follows that $\|\Phi_k(A + \alpha I)\| \leq 2$ for $k = 0, 1, \dots$; see [3]. The same holds for $\|\Phi_k(T_m + \alpha I)\|$, since the field of values of $T_m + \alpha I$ is included in that of $A + \alpha I$. Therefore, Lemma 2.1 ensures that

$$\|\hat{x} - \hat{x}_m\| \leq \sum_{k=m}^\infty |f_k| (\|\Phi_k(A + \alpha I)\| + \|\Phi_k(T_m + \alpha I)\|) \leq 8 \exp(-t\epsilon) \frac{r_2}{r_2 - 1} \left(\frac{1}{r_2} \right)^m.$$

Finally, using $\int_0^\infty \exp(-t\epsilon)dt = \epsilon^{-1}$,

$$\|X - X_m\| \leq \int_0^\infty \|\hat{x} - \hat{x}_m\|dt \leq \frac{8}{\epsilon} \frac{r_2}{r_2 - 1} \left(\frac{1}{r_2}\right)^m,$$

which completes the proof. \square

The ideal result for $\|x - x_m\|$ would set r_2 to be equal to $r_2 = \psi^{-1}(0)$ and not to $r_2 = \psi^{-1}(\epsilon)$ in the proof. However, this would make \mathcal{M} in (4.2) equal to one, and the generalized integral would not converge. The result above can be compared to the sharper one in Proposition 4.1. In practice, however, the asymptotic result is not affected by the use of ϵ , since it is sufficient to take ϵ small compared to α_{\min} , and the same asymptotic rate as in Proposition 4.1 is recovered; only the multiplicative factor increases. Therefore, setting $r_{2,0} = \psi^{-1}(0)$, the result above shows that

$$(4.3) \quad \|X - X_m\| = O\left(\left(\frac{1}{r_{2,0}}\right)^m\right).$$

The following mapping is a modified version of the external mapping used, for instance, in [21]:

$$(4.4) \quad \psi(\tau) = \gamma_1 - \gamma_2 \left(1 - \frac{1}{\tau}\right)^{2-\theta} \tau, \quad \tau = \sigma e^{i\omega}, \quad |\tau| \geq 1,$$

for $0 < \theta < 1$ and $\gamma_1, \gamma_2 \in \mathbb{R}^+$. The function ψ maps the exterior of the disc $D(0,1)$ onto a wedge-shaped convex set Ω in \mathbb{C}^+ . The following result holds.

COROLLARY 4.6. *Let $\hat{\Omega} \subset \mathbb{C}^+$ be the wedge-shaped set which is the image through $\hat{\psi}$ of the disk $D(0,1)$, where $\hat{\psi}$ is as in (4.4). Assume the field of values of the matrix $A + \alpha_{\min}I$, with $\alpha_{\min} = \lambda_{\min}((A + A^T)/2)$, is contained in $\hat{\Omega}$. For $0 < \epsilon < 2\alpha_{\min}$, let $r_2 = |\hat{\psi}^{-1}(\epsilon)|$. Then*

$$\|X - X_m\| \leq \frac{8}{\epsilon} \frac{r_2}{r_2 - 1} \left(\frac{1}{r_2}\right)^m.$$

Proof. The proof follows the same steps as that of Corollary 4.5. \square

Example 4.7. We consider the 400×400 (normal) diagonal matrix A whose eigenvalues are on the curve $\psi(\tau) = 2 - 2(1 - 1/\tau)^{2-\omega}\tau$ for $\tau \in D(0,1)$ with $\omega = 0.3$ (see the left plot of Figure 4.4). Here $\alpha_{\min} = 1.9627$. The image of the mapping $\hat{\psi}(\tau) = \alpha_{\min} + \psi(\tau)$, $\tau \in D(0,1)$, thus contains the spectrum of $A + \alpha_{\min}I$. Numerical computation yields $r_{2,0} = |\hat{\psi}^{-1}(0)| \approx 3.5063$. The vector b is the normalized vector of all ones. The right plot of Figure 4.4 shows the convergence history of the Krylov solver, together with the asymptotic factor $(1/r_{2,0})^m$ in the estimate of Corollary 4.6. The linear asymptotic convergence is fully captured by the estimate.

In our next examples we numerically determine a contour bounding the field of values of the coefficient matrix. Indeed, more general mappings than in the examples above can be obtained and numerically approximated within the class of Schwarz–Christoffel conformal mappings [10]. In all cases, the vector b was taken to be the normalized vector of all ones.

Example 4.8. We consider the 200×200 Toeplitz matrix

$$A = \text{Toeplitz}(-1, -1, \underline{2}, 0.1).$$

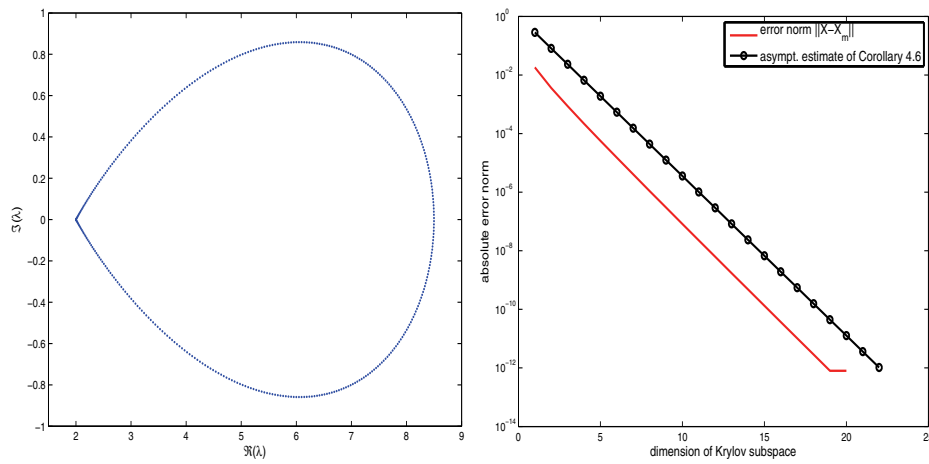


FIG. 4.4. *Example 4.7. Left plot: Spectrum of A . Right plot: True error and its asymptotic factor associated with the asymptotic estimate $(1/r_{2,0})^m$ related to Corollary 4.6 for the Krylov subspace solver of the Lyapunov equation.*

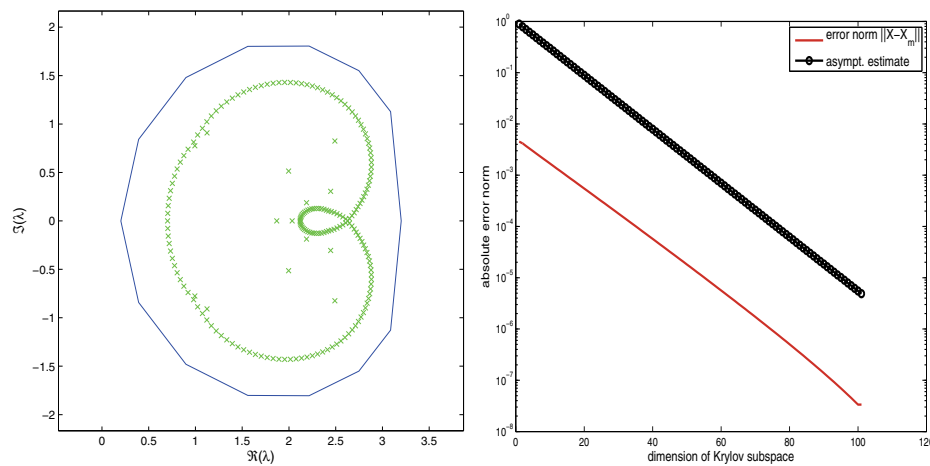


FIG. 4.5. *Example 4.8. Left: Spectrum (“x”) and approximated field of values (solid line). Right: True convergence rate and asymptotic estimate $(1/r_{2,0})^m$.*

In this case, the asymptotic convergence rate was numerically determined. To this end, we used the Schwarz–Christoffel mapping Toolbox [11] in MATLAB to numerically compute a conformal mapping whose image was an approximation to the boundary of the field of values of A (cf. left plot of Figure 4.5). A polygon with few vertices approximating $\partial F(A + \alpha_{\min}I)$ was obtained with `fv.m`, and this was then injected into the Schwarz–Christoffel inverse mapping function to construct the sought-after mapping and the value of $r_{2,0}$ according to (4.3). The asymptotic rate was determined to be $1/r_{2,0} \approx 0.8859$. The right plot in Figure 4.5 shows the extremely good agreement between the true error and the asymptotic rate for this numerically determined mapping.

Example 4.9. We consider once again the matrix in Example 4.4 and use the Schwarz–Christoffel mapping Toolbox to generate a sharper estimate of the polygon

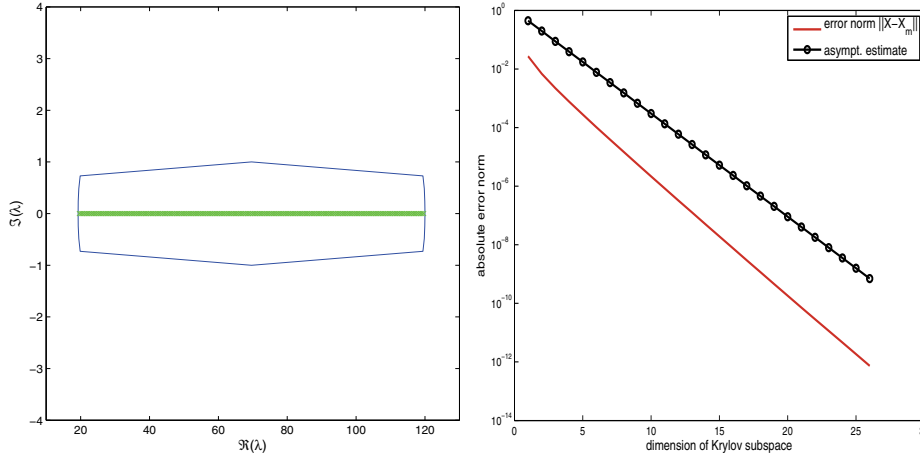


FIG. 4.6. Example 4.9. Left: Spectrum (“x”) and approximated field of values (solid line). Right: True convergence rate and asymptotic estimate $(1/r_{2,0})^m$.

including the field of values. This provides a refined numerical mapping and a more accurate convergence rate. The polygon approximating the field of values of $A + \alpha_{\min}I$ is shown in the left plot of Figure 4.6, while the history of the error norm and the estimate for the numerically computed value $1/r_{2,0} \approx 0.4445$ (cf. (4.3)) are reported in the right plot of Figure 4.6. The estimated convergence rate is clearly higher, that is, $1/r_{2,0}$ is smaller, than the value computed with the ellipse, which was $1/r_2 \approx 0.4699$.

The following mapping was analyzed in [26] and is associated with a nonconvex domain; the specialized case of an annular sector is discussed, for instance, in [7]. Given a set Ω , assume that $\partial\Omega$ is an analytic Jordan curve. If Ω is of bounded (or finite) boundary rotation, then

$$\max_{z \in \Omega} |\Phi_k(z)| \leq \frac{V(\Omega)}{\pi},$$

where $V(\Omega)$ is the boundary rotation of Ω , defined as the total variation of the angle between the positive real axis and the tangent of $\partial\Omega$. In particular, this bound is scale-invariant, so that it also holds that $V(s\Omega) = V(\Omega)$ [26]. These important properties ensure that for a diagonalizable matrix A , $\|\Phi_k(A + \alpha_{\min}I)\|$ is bounded independently of k , on a nonconvex set with bounded boundary rotation. Indeed, letting $A = Q\Lambda Q^{-1}$ be the spectral decomposition of A , then $\|\Phi_k(A + \alpha_{\min}I)\| \leq \kappa(Q)\|\Phi_k(\Lambda + \alpha_{\min}I)\|$, where $\kappa(Q) = \|Q\| \|Q^{-1}\|$, and the estimate above can be applied.

COROLLARY 4.10. Assume that A is diagonalizable, and let $A = Q\Lambda Q^{-1}$ be its spectral decomposition. Assume the spectrum of $A + \alpha_{\min}I$ is contained in the set $s\Omega \in \mathbb{C}^+$, with $s > 0$, whose boundary is the “bratwurst” image for $|\tau| = 1$ of

$$\psi(\tau) = \frac{(\rho\tau - \lambda N)(\rho\tau - \lambda M)}{(N - M)\rho\tau + \lambda(NM - 1)} \in \mathbb{C}^+,$$

where $\tau \in D(0, r)$, $r \geq 1$, while N, M, ρ , and λ are given and such that $\psi(D(0, 1)) \subset \mathbb{C}^+$. Then, for $0 < \epsilon < \min_{|\tau|=1} \Re(\psi(\tau))$,

$$\|X - X_m\| \leq \frac{8V(\Omega)\kappa(Q)}{\pi\epsilon} \frac{r_2}{r_2 - 1} \left(\frac{1}{r_2}\right)^m,$$

where $r_2 \geq 1$ is the smallest radius such that $\epsilon = \Re(\psi(r_2 \exp(i\theta)))$ for some θ .

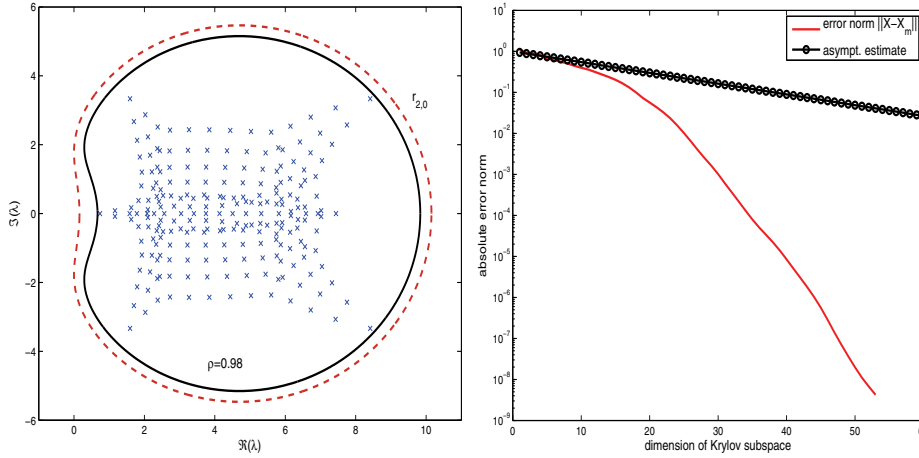


FIG. 4.7. Example 4.11. Left plot: Spectrum and “bratwurst” curves associated with disks of different radius. Right plot: True error and the asymptotic factor of its estimate in Corollary 4.10 for the Krylov subspace solver of the Lyapunov equation.

Proof. Proceeding as in Corollary 4.5 we have

$$\begin{aligned} \|\hat{x} - \hat{x}_m\| &\leq \sum_{k=m}^{\infty} |f_k| (\|\Phi_k(A + \alpha I)\| + \|\Phi_k(T_m + \alpha I)\|) \\ &\leq 4\mathcal{M}(t) \frac{V(\Omega)\kappa(Q)}{\pi} \frac{r_2}{r_2 - 1} \left(\frac{1}{r_2}\right)^m. \end{aligned}$$

Here $\mathcal{M}(t) = \exp(-t\epsilon)$. Finally,

$$\|X - X_m\| \leq 2 \int_0^{\infty} \|\hat{x} - \hat{x}_m\| dt \leq \frac{8V(\Omega)\kappa(Q)}{\pi} \int_0^{\infty} \mathcal{M}(t) dt \frac{r_2}{r_2 - 1} \left(\frac{1}{r_2}\right)^m,$$

from which the result follows. \square

Example 4.11. This example is taken from [25]; see also [26] for more details. In this case, A is the 225×225 matrix PDE225 of the Matrix Market repository [28] and it is such that $\alpha_{\min} \approx 0.08249$. The spectrum of $A + \alpha_{\min}I$ is included in the set 2Ω whose boundary is the bratwurst image of ψ as in Corollary 4.10, with $\lambda = -1$, $N = 1.0508$, $\rho = 0.98$, $M = 0.6626$ (exact to the first decimal digits; the other parameters defined in [25] were set at the different values $\theta = 5/4\pi$, $e = 1.40$). The left plot of Figure 4.7 shows the spectrum of $A + \alpha_{\min}I$ as “x”; the solid curve corresponds to the boundary of $\psi(D(0, 1))$, enclosing the whole spectrum. Let $r_{2,0} \geq 1$ be the smallest radius such that $\psi(r_{2,0}e^{i\theta}) = 0$ for some θ . Then the dashed curve is the boundary of $\psi(D(0, r_{2,0}))$. The right plot of Figure 4.7 shows the convergence curve of the Krylov subspace solver, together with the asymptotic quantity $(\frac{1}{r_{2,0}})^m$, $m = 1, 2, \dots$, associated with Corollary 4.10. We observe that the initial convergence phase is well captured by the estimate. As expected, the estimate cannot reproduce the superlinear convergence of the solver at later stages.

5. Connections to linear system solvers and further considerations. The relation

$$z^{-1} = \int_0^{\infty} e^{-tz} dt$$

can be used to show a close connection between our estimates and the solution of the linear system $(A + \alpha_{\min}I)d = b$ in the Krylov subspace. Let $V_m(T_m + \alpha_{\min}I)^{-1}e_1$ be the Galerkin approximation to the linear system solution d in the Krylov subspace $K_m(A, b) = K_m(A + \alpha_{\min}I, b)$. Then the system error can be written as

$$\begin{aligned} & (A + \alpha_{\min}I)^{-1}b - V_m(T_m + \alpha_{\min}I)^{-1}e_1 \\ &= \int_0^\infty (\exp(-t(A + \alpha_{\min}I))b - V_m \exp(-t(T_m + \alpha I))e_1) dt. \end{aligned}$$

Comparing the last integral with the error bound in (2.2) shows that the error norm $\|(A + \alpha_{\min}I)^{-1}b - V_m(T_m + \alpha_{\min}I)^{-1}e_1\|$ may be bounded by exactly the same tools we have used for the Lyapunov error and that the two initial integral bounds differ only by a factor of two. Indeed, the estimates of Proposition 3.1 (symmetric case) and of Proposition 4.1 (spectrum contained in an ellipse) employ the same asymptotic factors that characterize the convergence rate of methods such as the conjugate gradients in the symmetric case, and FOM or GMRES in the nonsymmetric case, when applied to the system $(A + \alpha_{\min}I)d = b$; see, e.g., [32]. Therefore, we have shown that the convergence of a Galerkin procedure in the Krylov subspace for solving (1.1) has the same convergence factor as a corresponding Krylov subspace method for the shifted (single vector) linear system.

As a natural consequence of the discussion above, the previous results can be generalized to the case when b is replaced by a matrix B , with more than one column. A Galerkin approximation may be obtained by first generating the “block” Krylov subspace $\mathcal{K}_m(A, B) = \text{span}\{B, AB, \dots, A^{m-1}B\}$ and then proceeding as described in section 2; see, e.g., [2]. Let $B = [b_1, \dots, b_s]$. Setting $Z = \exp(-tA)B$ and letting $Z_m \in \mathcal{K}_m(A, B)$ be the associated Krylov approximation to the exponential, we can bound $\|ZZ^\top - Z_mZ_m^\top\|$, for instance, as

$$\|ZZ^\top - Z_mZ_m^\top\| \leq \sum_{k=1}^s \|z_m^{(k)}(z_m^{(k)})^\top - z_m^{(k)}(z_m^{(k)})^\top\|,$$

where $Z = [z^{(1)}, \dots, z^{(s)}]$ and $Z_m = [z_m^{(1)}, \dots, z_m^{(s)}]$. The results of the previous sections can be thus applied to each term in the sum. Refined bounds may possibly be obtained by using the theory of matrix polynomials, but this is beyond the scope of this work; see, e.g., [32].

We also observe that our convergence results can be generalized to the case of accelerated methods, such as that described in [33], by using the theoretical matrix function framework described in [13].

Acknowledgments. We are deeply indebted to Leonid Knizhnerman for several insightful comments which helped improve a previous version of this paper. We also thank the referee, whose criticism helped us improve this paper.

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of Mathematical Functions*, Dover, New York, 1965.
 [2] A. C. ANTOUNAS, *Approximation of Large-Scale Dynamical Systems*, Adv. Des. Control 6, SIAM, Philadelphia, 2008.
 [3] B. BECKERMANN, *Image numérique, GMRES et polynômes de Faber*, C. R. Acad. Sci. Paris Ser. I, 340 (2005), pp. 855–860.

- [4] B. BECKERMANN AND A. B. J. KUIJLAARS, *Superlinear convergence of conjugate gradients*, SIAM J. Numer. Anal., 39 (2001), pp. 300–329.
- [5] B. BECKERMANN AND A. B. J. KUIJLAARS, *Superlinear CG convergence for special right-hand sides*, Electron. Trans. Numer. Anal., 14 (2002), pp. 1–19.
- [6] P. BENNER, *Control theory*, in Handbook of Linear Algebra, Chapman & Hall/CRC, Boca Raton, FL, 2006, Chapter 57.
- [7] J. P. COLEMAN AND N. J. MYERS, *The Faber polynomials for annular sectors*, Math. Comp., 64 (1995), pp. 181–203.
- [8] M. J. CORLESS AND A. E. FRAZHO, *Linear Systems and Control—An Operator Perspective*, Pure Appl. Math., Marcel Dekker, New York, Basel, 2003.
- [9] B. N. DATTA, *Krylov subspace methods for large-scale matrix problems in control*, Future Generation Computer Systems, 19 (2003), pp. 1253–1263.
- [10] T. A. DRISCOLL AND L. N. TREFETHEN, *Schwarz-Christoffel Mapping*, Cambridge Monogr. Appl. Comput. Math. 8, Cambridge University Press, Cambridge, UK, 2002.
- [11] T. DRISCOLL, *Algorithm 756: A MATLAB Toolbox for Schwarz-Christoffel mapping*, ACM Trans. Math. Software, 22 (1996), pp. 168–186.
- [12] V. DRUSKIN AND L. KNIZHNERMAN, *Two polynomial methods of calculating functions of symmetric matrices*, U.S.S.R. Comput. Math. Math. Phys., 29 (1989), pp. 112–121.
- [13] V. DRUSKIN AND L. KNIZHNERMAN, *Extended Krylov subspaces: Approximation of the matrix square root and related functions*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 755–771.
- [14] M. EIERMANN, *On semiiterative methods generated by Faber polynomials*, Numer. Math., 56 (1989), pp. 139–156.
- [15] M. EIERMANN, *Field of values and iterative methods*, Linear Algebra Appl., 180 (1993), pp. 167–197.
- [16] S. W. ELLACOTT, *Computation of Faber series with application to numerical polynomial approximation in the complex plane*, Math. Comp., 40 (1983), pp. 575–587.
- [17] K. O. GEDDES, *Near-minimax polynomial approximation in an elliptical region*, SIAM J. Numer. Anal., 15 (1978), pp. 1225–1233.
- [18] G. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.
- [19] I. S. GRADSHTEYN AND I. M. RYZHIK, *Table of Integrals, Series, and Products* (corrected and enlarged edition), Academic Press, San Diego, CA, 1980.
- [20] N. J. HIGHAM, *The Matrix Computation Toolbox*, <http://www.ma.man.ac.uk/~higham/mctoolbox>.
- [21] M. HOCHBRUCK AND C. LUBICH, *On Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 34 (1997), pp. 1911–1925.
- [22] I. M. JAIMOUKHA AND E. M. KASENALLY, *Krylov subspace methods for solving large Lyapunov equations*, SIAM J. Numer. Anal., 31 (1994), pp. 227–251.
- [23] K. JBILOU AND A. J. RIQUET, *Projection methods for large Lyapunov matrix equations*, Linear Algebra Appl., 415 (2006), pp. 344–358.
- [24] L. KNIZHNERMAN, *Calculus of functions of unsymmetric matrices using Arnoldi’s method*, Comput. Math. Math. Phys., 31 (1991), pp. 1–9.
- [25] T. KOCH AND J. LIESEN, *The conformal “bratwurst” maps and associated Faber polynomials*, Numer. Math., 86 (2000), pp. 173–191.
- [26] J. LIESEN, *Construction and Analysis of Polynomial Iterative Methods for Non-Hermitian Systems of Linear Equations*, Ph.D. thesis, Fakultät für Mathematik, Universität Bielefeld, 1998.
- [27] T. A. MANTEUFFEL, *The Tchebychev iteration for nonsymmetric linear systems*, Numer. Math., 28 (1977), pp. 307–327.
- [28] MATRIX MARKET, *A Visual Repository of Test Data for Use in Comparative Studies of Algorithms for Numerical Linear Algebra*, Mathematical and Computational Sciences Division, National Institute of Standards and Technology; available online at <http://math.nist.gov/MatrixMarket>.
- [29] O. NEVANLINNA, *Convergence of Iterations for Linear Equations*, Birkhäuser, Basel, 1993.
- [30] M. ROBBÉ AND M. SADKANE, *A convergence analysis of GMRES and FOM for Sylvester equations*, Numer. Algorithms, 30 (2002), pp. 71–89.
- [31] Y. SAAD, *Numerical solution of large Lyapunov equations*, in Signal Processing, Scattering, Operator Theory, and Numerical Methods, Proceedings of the International Symposium MTNS-89, Vol. III, M. A. Kaashoek, J. H. van Schuppen, and A. C. Ran, eds., Birkhäuser, Boston, 1990, pp. 503–511.
- [32] Y. SAAD, *Iterative Methods for Sparse Linear Systems*, PWS, Boston, 1996.

- [33] V. SIMONCINI, *A new iterative method for solving large-scale Lyapunov matrix equations*, SIAM J. Sci. Comput., 29 (2007), pp. 1268–1288.
- [34] V. SIMONCINI AND D. B. SZYLD, *On the occurrence of superlinear convergence of exact and inexact Krylov subspace methods*, SIAM Rev., 47 (2005), pp. 247–272.
- [35] V. I. SMIRNOV AND N. A. LEBEDEV, *Functions of a Complex Variable, Constructive Theory*, MIT Press, Cambridge, MA, 1968.
- [36] D. E. STEWART AND T. S. LEYK, *Error estimates for Krylov subspace approximations of matrix exponentials*, J. Comput. Appl. Math., 72 (1996), pp. 359–369.
- [37] P. K. SUETIN, *Fundamental properties of Faber polynomials*, Russian Math. Surv., 19 (1964), pp. 121–149.
- [38] P. K. SUETIN, *Series of Faber Polynomials (Analytical Methods and Special Functions)*, Gordon and Breach Science Publishers, Amsterdam, 1998 (translated by E. V. Pankratiev).
- [39] H. TAL-EZER, *Spectral methods in time for parabolic problems*, SIAM J. Numer. Anal., 26 (1989), pp. 1–11.

CAN THE NONLOCAL CHARACTERIZATION OF SOBOLEV SPACES BY BOURGAIN ET AL. BE USEFUL FOR SOLVING VARIATIONAL PROBLEMS?*

GILLES AUBERT[†] AND PIERRE KORNPÖBST[‡]

Abstract. We question whether the recent characterization of Sobolev spaces by Bourgain, Brezis, and Mironescu (2001) could be useful to solve variational problems on $W^{1,p}(\Omega)$. To answer this, we introduce a sequence of functionals so that the seminorm is approximated by an integral operator involving a differential quotient and a radial mollifier. Then, for the approximated formulation, we prove existence, uniqueness, and convergence of the solution to the unique solution of the initial formulation. We show that these results can also be extended in the BV -case. Interestingly, this approximation leads to a unified implementation, for Sobolev spaces (including with high p -values) and for the BV space. Finally, we show how this theoretical study can indeed lead to a numerically tractable implementation, and we give some image diffusion results as an illustration.

Key words. calculus of variation, functional analysis, Sobolev spaces, BV , variational approach, integral approximations, nonlocal formulations

AMS subject classifications. 35J, 45E, 49J, 65N, 68W

DOI. 10.1137/070696751

1. Introduction. The goal of this work is to propose a new unifying method for solving variational problems defined on the Sobolev spaces $W^{1,p}(\Omega)$ or on the space of functions of bounded variations $BV(\Omega)$ of the form

$$(1.1) \quad \inf_{u \in W^{1,p}(\Omega)} F(u),$$

with

$$F(u) = \int_{\Omega} |\nabla u(x)|^p dx + \int_{\Omega} h(x, u(x)) dx.$$

To solve this problem numerically, particularly in the case when $p = 1$, several methods have been proposed; see, e.g., [8, 13, 14, 7, 18, 19]. These methods mainly rely on regularization or duality results.

In this article we propose an alternative method based on a recent new characterization of the Sobolev spaces by Bourgain, Brezis, and Mironescu [5], and further extended by Ponce [16] in the BV -case. In [5] the authors showed that the Sobolev seminorm of a function f can be approximated by a sequence of integral operators involving a differential quotient of f and a suitable sequence of radial mollifiers:

$$\lim_{n \rightarrow \infty} \int_{\Omega} \int_{\Omega} \frac{|u(x) - u(y)|^p}{|x - y|^p} \rho_n(|x - y|) dx dy = K_{N,p} \int_{\Omega} |\nabla u|^p dx.$$

*Received by the editors July 10, 2007; accepted for publication (in revised form) October 15, 2008; published electronically February 6, 2009.

<http://www.siam.org/journals/sinum/47-2/69675.html>

[†]Laboratoire J. A. Dieudonné, UMR 6621 CNRS, Université de Nice-Sophia Antipolis, 06108 Nice Cedex 2, France (gaubert@math.unice.fr).

[‡]INRIA Sophia Antipolis, Projet Odyssee, 2004 Route des Lucioles, 06902 Sophia Antipolis, France (pierre.kornprobst@inria.fr).

In this paper, our main contribution is to show how this characterization can be used to approximate the variational formulation (1.1) by defining the sequence of functionals

$$F_n(u) = \int_{\Omega} \int_{\Omega} \frac{|u(x) - u(y)|^p}{|x - y|^p} \rho_n(|x - y|) dx dy + \int_{\Omega} h(x, u(x)) dx.$$

To do this, we prove that the sequence of minimizers of F_n converges to the solution of the original variational formulation. We prove this result for any $p \geq 1$, so that the BV -case is also covered (thanks to results by Ponce [16]). Note that approximation is not constrained by the fidelity attach term (see [7]). Numerically, we propose a unified subgradient approach for all $p \geq 1$, and we show how to discretize the nonlocal singular term with a finite element-type method.

Interestingly, the nonlocal term in F_n has some similarities to recent contributions by Gilboa and Osher [12] and Gilboa et al. [11], who propose to minimize nonlocal functionals of the type

$$\int_{\Omega} \int_{\Omega} \phi(|u(x) - u(y)|) w(|x, y|) dx dy,$$

where ϕ is a convex positive function and w is a weighting function. The authors propose a general formalism for nonlocal smoothing terms but define them heuristically for their applications in image processing (see also the link to neighborhood filters [6]). In our contribution, the nonlocal term that we propose comes from the approximation of a seminorm, so that we will show some regularity results on the solution. Notice that one related major difference is the weighting function, which is in our case singular.

This paper is organized as follows. In section 2, we recall the main results from [5] that we will use herein and define the sequence of the approximating functional F_n . In section 3, we present the most significant results of the paper, considering the case $p > 1$: we prove existence and uniqueness of a minimizer u_n of F_n , characterize its regularity, derive the optimality condition, and finally show that u_n converges to the unique solution of the initial formulation. In section 4, we describe how those results can be extended to the case $p = 1$, which corresponds to the BV -case. Finally, we show in section 5 how this theoretical study can indeed lead to a numerically tractable implementation, and we give some image diffusion results as an illustration.

2. The Bourgain–Brezis–Mironescu result. Let us first recall the result of Bourgain, Brezis, and Mironescu [5].

PROPOSITION 2.1. *Assume $1 \leq p < \infty$ and $u \in W^{1,p}(\Omega)$, and let $\rho \in L^1(\mathbb{R})$, $\rho \geq 0$. Then*

$$(2.1) \quad \int_{\Omega} \int_{\Omega} \frac{|u(x) - u(y)|^p}{|x - y|^p} \rho(|x - y|) dx dy \leq C \|u\|_{W^{1,p}}^p \|\rho\|_{L^1(\mathbb{R})},$$

where $\|u\|_{W^{1,p}}^p$ denotes the (semi)norm defined by $\|u\|_{W^{1,p}}^p = \int_{\Omega} |\nabla u|^p dx$ and C depends only on p and Ω .

Now let us suppose that (ρ_n) is a sequence of radial mollifiers, i.e.,

$$(2.2) \quad \rho_n \geq 0, \quad \int_{\mathbb{R}^N} \rho_n(|x|) dx = 1,$$

and for every $\delta > 0$, we assume that

$$(2.3) \quad \lim_{n \rightarrow \infty} \int_{\delta}^{\infty} \rho_n(r) r^{N-1} dr = 0.$$

With conditions (2.2) and (2.3), which we will assume throughout this article, we have the following proposition.

PROPOSITION 2.2. *If $1 < p < \infty$ and $u \in W^{1,p}(\Omega)$, then*

$$(2.4) \quad \lim_{n \rightarrow \infty} \int_{\Omega} \int_{\Omega} \frac{|u(x) - u(y)|^p}{|x - y|^p} \rho_n(|x - y|) dx dy = K_{N,p} \|u\|_{W^{1,p}}^p,$$

where $K_{N,p}$ depends only on p and N .

In this paper, we propose to apply Propositions 2.1 and 2.2 for solving general variational problems of the form

$$(2.5) \quad \inf_{u \in W^{1,p}(\Omega)} F(u),$$

with

$$(2.6) \quad F(u) = \int_{\Omega} |\nabla u(x)|^p dx + \int_{\Omega} h(x, u(x)) dx, u \in W^{1,p}(\Omega).$$

To do this, following [5], we introduce the nonlocal formulation

$$(2.7) \quad \inf_{u \in L^p(\Omega)} F_n(u),$$

with

$$(2.8) \quad F_n(u) = \int_{\Omega} \int_{\Omega} \frac{|u(x) - u(y)|^p}{|x - y|^p} \rho_n(|x - y|) dx dy + \int_{\Omega} h(x, u(x)) dx.$$

Our goal is to establish in which sense formulation (2.7)–(2.8) approximates the initial formulation (2.5)–(2.6).

3. Approximation of variational problems on $W^{1,p}(\Omega)$, $p > 1$. Thanks to Proposition 2.1, functional $F_n(u)$ is well-defined on $W^{1,p}(\Omega)$. However, one cannot prove directly that F_n admits a unique minimizer on $W^{1,p}(\Omega)$, since minimizing sequences cannot be bounded in that space. Thus we need to consider the minimization over the larger space $L^p(\Omega)$, and problem (2.7) is in fact an unbounded problem in $L^p(\Omega)$.

In this section, we prove the following results:

- For n fixed, we show in section 3.1 that problem (2.7) admits a unique solution $u_n \in L^p(\Omega)$.
- Then we show in section 3.2 that u_n is more regular and belongs to the Sobolev space $W^{s,p}(\Omega)$ with $1/2 < s < 1$. Moreover, we show that all minimizing sequences are bounded on $W^{s,p}(\Omega)$. The main consequence is that minimizing sequences $(u_n^l)_l$ indeed converge strongly to u_n . This additional regularity will also enable us to consider problems with Dirichlet boundary conditions, since one can give a meaning to the trace operator on that space.
- The previous regularity result will be fundamental in section 3.3 when we consider that n tends to infinity. Applying some results by Ponce [16], we will show that u_n converges to the unique solution u of the original formulation (2.5).
- In section 3.4 we establish the expression of the Euler–Lagrange equation.

Remark. Note that throughout this section and in the proofs, we will denote by C a universal constant that may be different from one line to the other. If the constant depends on n , for example, it will be denoted by $C(n)$.

3.1. Existence and uniqueness of a solution u_n in $L^p(\Omega)$. Now, let us show that functional (2.8) admits a unique minimizer. It is clear by using again Proposition 2.1 and the fact that $\|\rho_n\|_{L^1(\mathbb{R})} = 1$ that we have for all v in $W^{1,p}(\Omega)$

$$\inf_{u \in L^p(\Omega)} F_n(u) \leq \inf_{u \in W^{1,p}(\Omega)} F_n(u) \leq F_n(v) \leq C\|v\|_{W^{1,p}}^p + \int_{\Omega} h(x, v(x)) \, dx,$$

from which we deduce that $\inf_{u \in L^p(\Omega)} F_n(u)$ is bounded by a finite constant (independent of n).

PROPOSITION 3.1. *Assume that $h \geq 0$, the function $x \mapsto h(x, u(x))$ is in $L^1(\Omega)$ for all u in $L^p(\Omega)$, h is convex with respect to its second argument, and, for each n , the function $t \mapsto \rho_n(t)$ is nonincreasing. Then functional (2.8) admits a unique minimizer in $L^p(\Omega)$.*

Before proving this proposition, let us recall a technical lemma from Bourgain, Brezis, and Mironescu (Lemma 2 in [5]) that we will use in the proof of Proposition 3.1.

LEMMA 3.2. *Let $g, k : (0, \delta) \rightarrow \mathbb{R}_+$. Assume $g(t) \leq g(t/2)$ for $t \in (0, \delta)$, and that k is nonincreasing. Then for all $M > 0$, there exists a constant $C(M) > 0$ such that*

$$(3.1) \quad \int_0^\delta t^{M-1} g(t) k(t) dt \geq C(M) \delta^{-M} \int_0^\delta t^{M-1} g(t) dt \int_0^\delta t^{M-1} k(t) dt.$$

Proof of Proposition 3.1. Let us consider a minimizing sequence u_n^l of $F_n(u)$ with $n > 0$ fixed. Since $h \geq 0$ and $\inf_{u \in L^p(\Omega)} F_n(u)$ is bounded, then there exists a constant C such that

$$(3.2) \quad \int_{\Omega} \int_{\Omega} \frac{|u_n^l(x) - u_n^l(y)|^p}{|x - y|^p} \rho_n(|x - y|) dx dy \leq C.$$

We are going to apply techniques borrowed from Bourgain, Brezis, and Mironescu [5, Theorem 4]. Without loss of generality, we may assume that $\Omega = \mathbb{R}^N$ and that the support of u_n^l is included in a ball B of diameter 1. This can be achieved by extending each function u_n^l by reflection across the boundary in a neighborhood of $\partial\Omega$. We may also assume the normalization condition $\int_{\Omega} u_n^l(x) dx = 0$ for all n and l . Let us define for each $n, l, t > 0$

$$(3.3) \quad E_n^l(t) = \int_{S^{N-1}} \int_{\mathbb{R}^N} |u_n^l(x + tw) - u_n^l(x)|^p dx dw,$$

where S^{N-1} denotes the unit sphere of \mathbb{R}^N . Straightforward changes of variables show that

$$\int_{\Omega} \int_{\Omega} \frac{|u_n^l(x) - u_n^l(y)|^p}{|x - y|^p} \rho_n(|x - y|) dx dy = \int_0^1 t^{N-1} \frac{E_n^l(t)}{t^p} \rho_n(t) dt,$$

and thus (3.2) can be equivalently expressed as

$$(3.4) \quad \int_0^1 t^{N-1} \frac{E_n^l(t)}{t^p} \rho_n(t) dt \leq C.$$

Now since we have supposed that u_n^l is of zero mean, we can write

$$u_n^l(x) = u_n^l(x) - \frac{1}{|B|} \int_B u_n^l(y) dy.$$

Thus

$$\int |u_n^l(x)|^p dx = \int \left| u_n^l(x) - \frac{1}{|B|} \int_B u_n^l(y) dy \right|^p dx = \frac{1}{|B|^p} \int \left| \int_B u_n^l(x) - u_n^l(y) dy \right|^p dx,$$

and, thanks to the Hölder inequality, there exists a constant C such that

$$(3.5) \quad \int |u_n^l(x)|^p dx \leq C \int_{|h| \leq 1} \left(\int |u_n^l(x+h) - u_n^l(x)|^p dx \right) dh = C \int_0^1 t^{N-1} E_n^l(t) dt.$$

Now, an interesting property of E_n^l is that

$$(3.6) \quad E_n^l(2t) \leq 2^p E_n^l(t).$$

Inequality (3.6) follows from the triangle inequality $|a + b|^p \leq 2^{p-1}(|a|^p + |b|^p)$:

$$\begin{aligned} E_n^l(2t) &= \int_{S^{N-1}} \int_{\mathbb{R}^N} |u_n^l(x + 2tw) - u_n^l(x)|^p dx dw \\ &= \int_{S^{N-1}} \int_{\mathbb{R}^N} |u_n^l(x + 2tw) - u_n^l(x + tw) + u_n^l(x + tw) - u_n^l(x)|^p dx dw \\ &\leq 2^{p-1} \left(\int_{S^{N-1}} \int_{\mathbb{R}^N} |u_n^l(x + 2tw) - u_n^l(x + tw)|^p dx dw \right. \\ (3.7) \quad &\quad \left. + \int_{S^{N-1}} \int_{\mathbb{R}^N} |u_n^l(x + tw) - u_n^l(x)|^p dx dw \right) \\ &\leq 2^p E_n^l(t), \end{aligned}$$

since both integrals in (3.7) are equal (up to a change of variable).

To conclude we apply Lemma 3.2 with $M = N$, $\delta = 1$, $k(t) = \rho_n(t)$, and $g(t) = \frac{E_n^l(t)}{t^p}$ (this choice is valid thanks to the hypotheses on ρ_n and property (3.6)). We obtain

$$\begin{aligned} \int_0^1 t^{N-1} \rho_n(t) \frac{E_n^l(t)}{t^p} dt &\geq C \int_0^1 t^{N-1} \rho_n(t) dt \int_0^1 t^{N-1} \frac{E_n^l(t)}{t^p} dt \\ (3.8) \quad &\geq C \int_0^1 t^{N-1} \rho_n(t) dt \int_0^1 t^{N-1} E_n^l(t) dt, \end{aligned}$$

where we have used in the last inequality the fact that $0 < t < 1$. Let us denote $d(n) = \int_0^1 t^{N-1} \rho_n(t) dt > 0$; we obtain, thanks to (3.4), (3.5), and (3.8), that there exists a constant $C(n) > 0$ (but which is independent of l) such that

$$(3.9) \quad \|u_n^l\|_{L^p(\Omega)} \leq C(n).$$

From (3.9), we deduce that, up to a subsequence, u_n^l tends weakly in $L^p(\Omega)$ to some $u_n \in L^p(\Omega)$ as $l \rightarrow +\infty$. Then we deduce that the sequence $w_n^l(x, y) = u_n^l(x) - u_n^l(y)$ tends weakly in $L^p(\Omega \times \Omega)$ to $w_n(x, y) = u_n(x) - u_n(y)$. Since the functional

$$w \rightarrow \int_{\Omega} \int_{\Omega} |w(x, y)|^p \frac{\rho_n(|x - y|)}{|x - y|^p} dx dy$$

is nonnegative, convex, and lower semicontinuous from $L^p(\Omega \times \Omega) \rightarrow \bar{R}$, we easily get

$$F_n(u_n) \leq \liminf_{l \rightarrow \infty} F_n(u_n^l) = \inf_{u \in L^p(\Omega)} F_n(u),$$

where the symbol \liminf denotes the lower limit. Therefore u_n is a minimizer of F_n . Moreover it is unique since the function $t \mapsto |t|^p$ is strictly convex for $p > 1$. \square

3.2. Regularity result for u_n . We have obtained the existence of a minimizer in $L^p(\Omega)$. Let us show that the solution is in fact more regular than just L^p .

As for $W^{1,p}(\Omega)$, the space $W^{s,p}(\Omega)$ can be characterized by a differential quotient. For $0 < s < 1$ and $1 \leq p < \infty$, we define

$$W^{s,p}(\Omega) = \left\{ u \in L^p(\Omega); \frac{|u(x) - u(y)|}{|x - y|^{s+N/p}} \in L^p(\Omega \times \Omega) \right\},$$

endowed with the norm

$$|u|_{W^{s,p}(\Omega)}^p = \int_{\Omega} |u|^p dx + \int_{\Omega} \int_{\Omega} \frac{|u(x) - u(y)|^p}{|x - y|^{sp+N}} dx dy.$$

Let us consider n fixed and let us denote by $C(n)$ a universal positive constant depending on n (i.e., $C(n)$ may be different from one line to the next). Let $(u_n^l)_l$ be a minimizing sequence of (2.7) so that

$$(3.10) \quad \int_{\Omega} \int_{\Omega} \frac{|u_n^l(x) - u_n^l(y)|^p}{|x - y|^p} \rho_n(|x - y|) dx dy \leq C(n).$$

Then we would like to prove that (3.10) implies

$$(3.11) \quad \int_{\Omega} \int_{\Omega} \frac{|u_n^l(x) - u_n^l(y)|^p}{|x - y|^{sp+N}} dx dy \leq C(n)$$

for some $1/2 < s < 1$ and some other constant $C(n)$, thus showing that u_n^l belongs to $W^{s,p}(\Omega)$.

PROPOSITION 3.3. *Let q be a real number such that $\frac{p}{2} < q < p$ and $(p-1) \leq q$, and let us assume that ρ_n verifies (2.2)–(2.3) and also that conditions of Proposition 3.1 are fulfilled. Moreover let us suppose that the functions $t \rightarrow \rho_n(t)$ and $t \rightarrow t^{q+2-p}\rho_n(t)$ are nonincreasing for $t \geq 0$. Then $u_n^l \in W^{q/p,p}(\Omega)$ for all l .*

Proof. Without loss of generality, let us prove Proposition 3.3 for the case $N = 2$. Equivalently, thanks to (3.3) of E_n^l , we can rewrite (3.10) and (3.11) so that one needs to prove that

$$(3.12) \quad \int_0^1 t \frac{E_n^l(t)}{t^p} \rho_n(t) dt \leq C(n)$$

implies

$$\int_0^1 t \frac{E_n^l(t)}{t^{sp+2}} dt \leq C(n).$$

Let us apply Lemma 3.2 with $M = \delta = 1$, $g(t) = \frac{E_n^l(t)}{t^{q+1}}$, $k(t) = t^{q+2-p}\rho_n(t)$. Assuming the hypothesis on $g(t)$ is true, Lemma 3.2 gives

$$(3.13) \quad \int_0^1 \frac{E_n^l(t)\rho_n(t)}{t^{p-1}} dt \geq C(M) \int_0^1 \frac{E_n^l(t)}{t^{q+1}} dt \int_0^1 t^{q+2-p}\rho_n(t) dt.$$

Therefore

$$\int_0^1 \frac{E_n^l(t)}{t^{q+1}} dt \leq \frac{1}{C(M) \int_0^1 t^{q+2-p}\rho_n(t) dt} \int_0^1 \frac{E_n^l(t)\rho_n(t)}{t^{p-1}} dt,$$

and according to (3.12), we get

$$\int_0^1 \frac{E_n^l(t)}{t^{q+1}} dt \leq \frac{C(n)/C(M)}{\int_0^1 t^{q+2-p} \rho_n(t) dt},$$

where the right-hand term is bounded independently of l . Thus $u_n^l \in W^{s,p}(\Omega)$ with $s = \frac{q}{p}$, and since we have supposed $\frac{q}{2} < q < p$ we have $\frac{1}{2} < s < 1$.

So it remains to show that function $g(t)$ verifies the hypothesis of Lemma 3.2. We have to check $g(t) \leq g(t/2)$. Since $g(t) = \frac{E_n^l(t)}{t^{q+1}}$ then $g(t/2) = \frac{E_n^l(t/2)}{t^{q+1}} 2^{q+1} \geq 2^{q+1-p} \frac{E_n^l(t)}{t^{q+1}} = 2^{q+1-p} g(t)$ (thanks to (3.3)). Thus we get $g(t/2) \geq g(t)$ if $q+1-p \geq 0$, i.e., if $q \geq (p-1)$. \square

Depending on p , one needs to find a function $\rho_n(t)$ so that $\rho_n(t)$ and $t^{q+2-p} \rho_n(t)$ are decreasing, and verify (2.2) and (2.3). Let us show that such a ρ_n function exists. We define

$$(3.14) \quad \rho_n(t) = Cn^2 \rho(nt) \quad \text{with} \quad C = \frac{1}{\int_{\mathbb{R}^2} \rho(|x|) dx}$$

and, depending on the values of p , we propose the following functions:

$$(3.15) \quad \rho(t) = \begin{cases} \exp(-t)/t^{q+1} & \text{if } p = 1, \text{ with } 0.5 < q < 1, \\ \exp(-t)/t^q & \text{if } p = 2, \text{ with } 1 < q < 2, \\ \exp(-t)/t & \text{if } p > 2, \text{ with } q = p - 1. \end{cases}$$

As a consequence, we have the following proposition.

PROPOSITION 3.4. *Let $(u_n^l)_l$ be a minimizing sequence of (2.7). Let us suppose that h verifies the conditions of Proposition 3.1 and the coercivity condition $h(x, u) \geq a|u|^p + b$, with $a > 0$. Then the sequence $(u_n^l)_l$ is bounded in $W^{q/p,p}(\Omega)$ uniformly with respect to l . Therefore, up to a subsequence, u_n^l tends weakly to u_n in $W^{q/p,p}(\Omega)$ (and strongly in $L^p(\Omega)$).*

Another direct consequence of Proposition 3.3 is the following.

LEMMA 3.5. *We have $\inf_{u \in L^p(\Omega)} F_n(u) = \inf_{u \in W^{s,p}(\Omega)} F_n(u)$, and the solution of the problem posed on $L^p(\Omega)$ is also the solution of the problem posed in $W^{s,p}(\Omega)$.*

Proof. Since $W^{s,p}(\Omega) \subset L^p(\Omega)$, then

$$\inf_{u \in L^p(\Omega)} F_n(u) \leq \inf_{u \in W^{s,p}(\Omega)} F_n(u).$$

By definition, since u_n is the minimizer of F_n in $L^p(\Omega)$, we have

$$F_n(u_n) = \inf_{u \in L^p(\Omega)} F_n(u) \leq \inf_{u \in W^{s,p}(\Omega)} F_n(u),$$

but as $u_n \in W^{s,p}(\Omega)$, we have finally

$$\inf_{u \in W^{s,p}(\Omega)} F_n(u) \leq F_n(u_n) = \inf_{u \in L^p(\Omega)} F_n(u) \leq \inf_{u \in W^{s,p}(\Omega)} F_n(u),$$

which concludes the proof. \square

Remark. Yet another consequence of Proposition 3.3 is that one can also consider problems with Dirichlet boundary conditions if necessary: If one needs to solve problem (2.5) with a Dirichlet boundary condition $u = \varphi$ on $\partial\Omega$, then one can impose the minimizing sequence of (2.7) to verify $u_n^l = \varphi$ on $\partial\Omega$ (which has a meaning thanks to this regularity result), so that, by continuity of the trace operator, we have $u_n = \varphi$ on $\partial\Omega$. Thus u_n is the unique minimizer in $W^{q/p,p}(\Omega)$ of problem (2.7), also verifying the Dirichlet boundary condition.

3.3. Study of the $\lim_{n \rightarrow \infty} u_n$. In section 3 we proved the existence of a unique solution u_n for problem (2.7), with n fixed, which is in fact in $W^{s,p}(\Omega)$. Now, we are going to examine the asymptotic behavior of (2.7) as $n \rightarrow \infty$. Throughout this section we will suppose the hypotheses stated in Propositions 3.3 and 3.4 hold. By definition of a minimizer, we have, for all $v \in W^{q/p,p}(\Omega)$,

$$(3.16) \quad F_n(u_n) \leq F_n(v) = \int_{\Omega} \int_{\Omega} \frac{|v(x) - v(y)|^p}{|x - y|^p} \rho_n(|x - y|) dx dy + \int_{\Omega} h(x, v(x)) dx.$$

Thus by using (2.1) and the fact that $|\rho_n|_{L^1} = 1$ we deduce from (3.16) that $F_n(u_n)$ is bounded uniformly with respect to n . In particular, we get for some constant $C > 0$

$$\int_{\Omega} \int_{\Omega} \frac{|u_n(x) - u_n(y)|^p}{|x - y|^p} \rho_n(|x - y|) dx dy \leq C.$$

By using the same technique as in Proposition 3.3, we still have that (u_n) is bounded in $W^{q/p,p}(\Omega)$. Therefore there exists u such that (up to a subsequence) $u_n \rightarrow u$ in $L^p(\Omega)$ -strong. Moreover, by applying Theorem 4 from [5], we obtain that $u \in W^{1,p}(\Omega)$. We claim that u is the unique solution of problem (2.5), i.e., for all $v \in W^{1,p}(\Omega)$,

$$(3.17) \quad \int_{\Omega} |\nabla u(x)|^p dx + \int_{\Omega} h(x, u(x)) dx \leq \int_{\Omega} |\nabla v(x)|^p dx + \int_{\Omega} h(x, v(x)) dx.$$

To prove (3.17) we refer the reader to the paper by Ponce [16]. In this paper the author studies in the same spirit as [5] new characterizations of Sobolev spaces and also of the space $BV(\Omega)$ of functions of bounded variations (see also section 4). The author considers more general differential quotients than the ones in [5], namely, functionals of the form

$$E_n(u) = \int_{\Omega} \int_{\Omega} w \left(\frac{|u(x) - u(y)|}{|x - y|} \right) \rho_n(|x - y|) dx dy.$$

By studying the asymptotic behavior, Ponce [16] obtained new characterizations of $W^{1,p}(\Omega)$ but also of $BV(\Omega)$. In particular, for $w(t) = |t|^p$ the author proved that $E_n(u)$ Γ -converge (up to a multiplicative constant) to $E(u) = \int_{\Omega} |\nabla u|^p dx$.

We have the following proposition.

PROPOSITION 3.6.

(i) *The sequence of functionals*

$$F_n(u) = E_n(u) + \int_{\Omega} h(x, u(x)) dx$$

Γ -converges (up to a multiplicative constant) to

$$F(u) = E(u) + \int_{\Omega} h(x, u(x)) dx.$$

(ii) *The sequence u_n of minimizers of $F_n(u)$, which is precompact in $L^p(\Omega)$, converges to the unique minimizer of $F(u)$.*

Proof. Item (i) is the Γ -convergence result shown by Ponce [16]. Item (ii) is a direct consequence of general Γ -convergence properties, since we proved that the sequence (u_n) is bounded in $W^{s,p}(\Omega)$, and thus converges strongly in $L^p(\Omega)$ to u (up to a subsequence). \square

3.4. Euler–Lagrange equation. Since u_n is a global minimizer of $F_n(u)$ it necessarily verifies $F'_n(u_n) = 0$, i.e., an Euler–Lagrange equation. The Euler–Lagrange equation is given in the following proposition.

PROPOSITION 3.7. *If function h is differentiable, verifies conditions of Propositions 3.1 and 3.4, and verifies for all u and a.e. x an inequality of the form $|\frac{\partial h(x,u)}{\partial u}| \leq l(x) + b|u|^{p-1}$ for some function $l(x) \in L^1(\Omega)$, $l(x) > 0$ and some $b > 0$, then the unique minimizer u_n of $F_n(u)$ verifies for a.e. x*

$$(3.18) \quad 2p \int_{\Omega} \frac{|u_n(x) - u_n(y)|^{p-2}}{|x - y|^p} (u_n(x) - u_n(y)) \rho_n(|x - y|) dy + \frac{\partial h(x, u_n(x))}{\partial u} = 0.$$

Proof. Let us focus on the smoothing term and denote

$$E_n(u_n) = \int_{\Omega} \int_{\Omega} \frac{|u_n(x) - u_n(y)|^p}{|x - y|^p} \rho_n(|x - y|) dx dy,$$

and let us consider for all v in $W^{1,p}(\Omega)$ the differential quotient

$$D_v(t) = \frac{E_n(u_n + tv) - E_n(u_n)}{t}.$$

We have

$$D_v(t) = \int_{\Omega} \int_{\Omega} \frac{|u_n(x) - u_n(y) + t(v(x) - v(y))|^p - |u_n(x) - u_n(y)|^p}{|x - y|^p} \rho_n(|x - y|) dx dy.$$

Thanks to Taylor’s formula, there exists $c(t, x, y)$ with $|c(t, x, y) - (u_n(x) - u_n(y))| < t|v(x) - v(y)|$ such that

$$D_v(t) = p \int_{\Omega} \int_{\Omega} \frac{(v(x) - v(y))c(t, x, y)|c(t, x, y)|^{p-2}}{|x - y|^p} \rho_n(|x - y|) dx dy.$$

Moreover, we have, as $t \rightarrow 0$,

$$\begin{aligned} & \frac{(v(x) - v(y))c(t, x, y)|c(t, x, y)|^{p-2}}{|x - y|^p} \rho_n(|x - y|) \\ & \rightarrow \frac{(v(x) - v(y))(u_n(x) - u_n(y))|u_n(x) - u_n(y)|^{p-2}}{|x - y|^p} \rho_n(|x - y|). \end{aligned}$$

On the other hand

$$|c(t, x, y)|^{p-1} \leq 2^p(|u_n(x) - u_n(y)|^{p-1} + |v(x) - v(y)|^{p-1}).$$

Thus

$$(3.19) \quad \begin{aligned} & \left| \frac{(v(x) - v(y))c(t, x, y)|c(t, x, y)|^{p-2}}{|x - y|^p} \rho_n(|x - y|) \right| \\ & \leq 2^p \left(\frac{|v(x) - v(y)||u_n(x) - u_n(y)|^{p-1}}{|x - y|^p} \rho_n(|x - y|) + \frac{|v(x) - v(y)|^p}{|x - y|^p} \rho_n(|x - y|) \right). \end{aligned}$$

Let us discuss the integrability of the right-hand side terms denoted, respectively, by A and B . The second term B is bounded by an integrable function because $v \in W^{1,p}(\Omega)$ and thanks to Proposition 2.1. The first term A gives

$$A = \frac{|v(x) - v(y)|}{|x - y|} \rho_n^{\frac{1}{p}}(x - y) \left| \frac{u_n(x) - u_n(y)}{|x - y|} \right|^{p-1} \rho_n^{\frac{p-1}{p}}(x - y),$$

where

$$\frac{|v(x) - v(y)|}{|x - y|} \rho_n^{\frac{1}{p}}(x - y)$$

is in $L^p(\Omega)$ since $v \in W^{1,p}(\Omega)$ and thanks to Proposition 2.1, and

$$\left| \frac{u_n(x) - u_n(y)}{|x - y|} \right|^{p-1} \rho_n^{\frac{p-1}{p}}(x - y)$$

is in $L^{\frac{p}{p-1}}(\Omega)$ since u_n is a minimizing sequence. So A is also bounded by an integrable function.

Therefore we can apply Lebesgue’s dominated convergence theorem (n is fixed) and get

$$\langle E'_n(u_n), v \rangle = p \int_{\Omega} \frac{|u_n(x) - u_n(y)|^{p-2}}{|x - y|^p} (v(x) - v(y))(u_n(x) - u_n(y)) \rho_n(|x - y|) dy.$$

The computation of the derivative of $\int_{\Omega} h(x, u(x)) dx$ is classical. Thus the desired result (3.18) by remarking that the function $(x, y) \mapsto \frac{|u_n(x) - u_n(y)|^{p-2} (u_n(x) - u_n(y))}{|x - y|^p}$ is antisymmetric with respect to (x, y) . \square

4. Extension of previous results to the $BV(\Omega)$ -case ($p = 1$). A similar result to that of Proposition 2.2 holds if $p = 1$; see [16]. In this case we need to search for a solution for problem (2.5) in $BV(\Omega)$, the space of functions of bounded variations [1, 10]. In fact most results are still valid in this case with some adaptations. We do not reproduce here details of their proofs, which rely upon the work by Ponce [16], who has, as said before, generalized to $BV(\Omega)$ the results of [5] stated in the $W^{1,p}(\Omega)$ case.

Let us recall the main steps and show how the results can be extended.

- The first point is that the proof of Proposition 3.1 does not apply in the case $p = 1$ since we cannot extract from a sequence bounded in $L^1(\Omega)$ a weakly converging subsequence. Thus we have to show that a minimizing sequence u_n^l of $F_n(u)$ is bounded in the Sobolev space $W^{q,1}(\Omega)$, with $0.5 < q < 1$. To do that, we use the same proof as in Proposition 3.3. Then, thanks to the two-dimensional Rellich–Kondrachov theorem $W^{q,1}(\Omega) \subset L^r(\Omega)$ with compact injection for $1 \leq r < \frac{2}{2-q}$ (note that if $0.5 < q < 1$, then $4/3 < \frac{2}{2-q} < 2$). Therefore, up to a subsequence, $u_n^l(x)$ tends, a.e., to some function $u_n(x)$. Then by using Fatou’s lemma we get $F_n(u_n) \leq \liminf_{l \rightarrow \infty} F_n(u_n^l)$; i.e., u_n is a minimizer of F_n .
- The result when n tends to infinity is again obtained thanks to the Γ -convergence result by Ponce and the compactness of the sequence u_n in $L^r(\Omega)$. As a result, u_n converges strongly in $L^1(\Omega)$ to $u \in BV(\Omega)$.
- Finally, the Euler–Lagrange equation (3.18) is no longer true in the case $p = 1$ since the function $t \rightarrow |t|$ is not differentiable. However, it is subdifferentiable. Therefore (3.18) changes into an inclusion

$$(4.1) \quad 0 \in \partial E_n(u_n) + \frac{\partial h}{\partial u}(x, u_n),$$

where $E_n(u) = \int_{\Omega} \int_{\Omega} \frac{|u(x) - u(y)|}{|x - y|} \rho_n(|x - y|) dx dy$. In (4.1), we can choose any element of the subdifferential, and, for example,

$$(4.2) \quad 2 \int_{\Omega} \frac{1}{|x - y|} \text{sign}(u_n(x) - u_n(y)) \rho_n(|x - y|) dy,$$

where

$$(4.3) \quad \text{sign}(s) = \begin{cases} -1 & \text{if } s < 0, \\ 0 & \text{if } s = 0, \\ 1 & \text{if } s > 0. \end{cases}$$

5. Implementation details and results.

5.1. A unified discrete implementation. In this section, we give the implementation details to solve the general variational problem (2.7) in a unified way (for n fixed) for both Sobolev and BV spaces.

The goal is to solve the differential inclusion

$$0 \in \partial F_n(u_n),$$

with a standard subgradient descent approach [17, 4]:

$$(5.1) \quad \begin{cases} u^{k+1}(x) = u^k(x) - \alpha^k g^k(x), \\ u^0(x) = u_0(x) \quad \forall x \in \Omega, \end{cases}$$

where α^k is the k th step size and g^k is any subgradient in $\partial F_n(u_n)$.

Taking into account the expression of the gradient or subgradient, we have here

$$(5.2) \quad u^{k+1}(x) = u^k(x) + \alpha^k \left(- \frac{\partial h}{\partial u}(x, u^k(x)) - 2pI_{u^k}(x) \right),$$

with

$$(5.3) \quad I_{u^k}(x) = \int_{\Omega} \frac{|u^k(x) - u^k(y)|^{p-1}}{|x - y|^p} \text{sign}(u^k(x) - u^k(y)) \rho_n(|x - y|) dy \quad \forall p.$$

Note that (5.3) is a unified expression which corresponds to the gradient when $p > 1$ (see the Euler–Lagrange equation in section 5.1), or a given element of the subdifferential in the BV -case (see section 4). We remind the reader that the definition of ρ_n also depends on p (see (3.15)).

Now the problem is to discretize in space the integral $I_{u^k}(x)$, which has a singular kernel, not defined when $x = y$. Let us introduce the function J_{u^k} such that

$$(5.4) \quad I_{u^k}(x) = \int_{\Omega} \frac{J_{u^k}(x, y)}{|x - y|} dy,$$

with

$$J_{u^k}(x, y) = \frac{|u^k(x) - u^k(y)|^{p-1}}{|x - y|^{p-1}} \text{sign}(u^k(x) - u^k(y)) \rho_n(|x - y|).$$

Because of the singularity, simple schemes using finite differences and integral approximations, for example, will fail. Here we propose to do the following:

- Discretize the space using a triangulation. We denote by \mathcal{T} the family of triangles covering Ω (see Figure 1).
- Interpolate linearly the function $J_{u^k}(x, y)$ on each triangle (x fixed).
- Find explicit expressions for the integral $J_{u^k}(x, y)/|x - y|$ on each triangle. Note that this kind of estimation also appears, for instance, in electromagnetism problems such as MEG-EEG (see, e.g., [9]), where one needs to estimate such singular integrals on meshed domains (three-dimensional domains here).

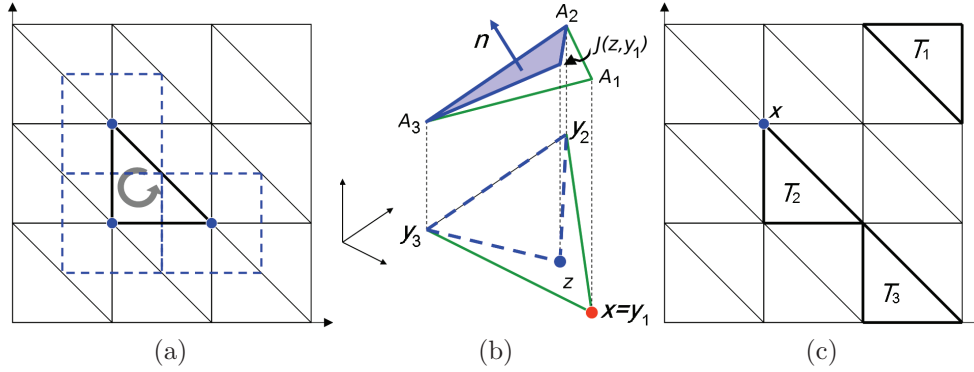


FIG. 1. (a) Mesh definition. Pixels are represented by the dashed squares. The circles correspond to the centers of the pixels defining the nodes of the mesh. Four nodes define two triangles. (b) In the special case when x is a node ($x = y_1$ in the figure), one needs an interpolation to define $J_{u^k}(x, y)$. In that situation, another point z close to the node is introduced and a linear interpolation is estimated. (c) Different cases depending on the situation of x with respect to T_i . Triangle T_1 has no edge aligned with x ; for triangle T_2 , x is one node; for T_3 , x is aligned with one edge.

Let us now detail each step. First, integral (5.4) becomes

$$(5.5) \quad I_{u^k}(x) = \sum_{T_i \in \mathcal{T}} \int_{T_i} \frac{J_{u^k}(x, y)}{|x - y|} dy.$$

Then let us approximate $J_{u^k}(x, y)$ on each triangle by a linear interpolation. We assume that x is given and fixed. Given one triangle $T \in \mathcal{T}$, let us denote the three nodes of T by $\{y_i = (y_i^1, y_i^2)^T\}_{i=1..3}$, where the subscript indicates the component. Then we define $\{A_i\}_{i=1..3}$ to be the three-dimensional points

$$A_i = (y_i^1, y_i^2, J_{u^k}(x, y_i))^T.$$

Note that as soon as $x \neq y_i$, $J_{u^k}(x, y_i)$ is well-defined. Otherwise, if x is in fact a node of T , for example, y_1 (see Figure 1(b)), then we use a linear interpolation algorithm: We introduce one point $z \in T$ close to y_1 , estimate the value of $J_{u^k}(z, y_1)$ at this point, and deduce the value of $J_{u^k}(x, y_1)$ by interpolation.

So, given $\{A_i\}_{i=1..3}$, we can in fact choose any node y_j and write

$$(5.6) \quad J_{u^k}(x, y) = J_{u^k}(x, y_j) - \frac{1}{n^3} \begin{pmatrix} n^1 \\ n^2 \end{pmatrix} (y - y_j),$$

where n is the normal to the triangle $A_1A_2A_3$ (see Figure 1(b)). With (5.6) we obtain

$$(5.7) \quad \begin{aligned} \int_T \frac{J_{u^k}(x, y)}{|x - y|} dy &= J_{u^k}(x, y_j) \int_T \frac{1}{|x - y|} dy - \frac{1}{n^3} \begin{pmatrix} n^1 \\ n^2 \end{pmatrix} \int_T \frac{(y - y_j)}{|x - y|} dy \\ &= J_{u^k}(x, y_j) \int_T \frac{1}{|x - y|} dy \\ &\quad - \frac{1}{n^3} \begin{pmatrix} n^1 \\ n^2 \end{pmatrix} \left[\int_T \frac{(y - x)}{|x - y|} dy + (x - y_j) \int_T \frac{1}{|x - y|} dy \right]. \end{aligned}$$

So, in order to estimate the integral over triangle T , one need only estimate

$$(5.8) \quad \int_T \frac{1}{|x - y|} dy \quad \text{and} \quad \int_T \frac{(y - x)}{|x - y|} dy.$$

If we introduce the distance function

$$\text{Dist}(x, y) = |x - y| = \sqrt{(x^1 - y^1)^2 + (x^2 - y^2)^2},$$

so that

$$\begin{aligned} \nabla_y \text{Dist}(x, y) &= \frac{y - x}{|x - y|}, \\ \Delta_y \text{Dist}(x, y) &= \frac{1}{\text{Dist}(x, y)}, \end{aligned}$$

then we have the following relations:

$$(5.9) \quad \int_T \frac{1}{|x - y|} dy = \int_T \Delta_y \text{Dist}(x, y) dy = \sum_{i=1,2} \int_{\partial T} \frac{\partial \text{Dist}}{\partial y^i}(x, y) N^i ds,$$

$$(5.10) \quad \int_T \frac{(y - x)}{|x - y|} dy = \int_T \nabla_y \text{Dist}(x, y) dy = \int_{\partial T} \text{Dist}(x, y) N ds,$$

where N is the normal to the edges of the triangle T . So we need to estimate the two kinds of integrals defined on the boundaries of the triangles. This can be done explicitly, as follows.

LEMMA 5.1. *Let us consider a segment $S = (\alpha, \beta)$ of extremities $\alpha = (\alpha^1, \alpha^2)$, $\beta = (\beta^1, \beta^2)$, N the normal to this segment, and x a fixed given point. Let us define*

$$\begin{aligned} a &= |\alpha\beta|, & \delta &= a^2b^2 - c^2, & l_1 &= c/\sqrt{\delta}, \\ b &= |x\alpha|, & d &= x\vec{\alpha} \cdot N, & l_2 &= (a^2 + c)/\sqrt{\delta}, \\ c &= x\vec{\alpha} \cdot \vec{\alpha}\beta. \end{aligned}$$

Then we have

$$(5.11) \quad \sum_{i=1,2} \int_S \frac{\partial \text{Dist}}{\partial y^i}(x, y) N^i ds = \begin{cases} 0 & \text{if } x \text{ is aligned with } S, \\ d(\text{asinh}(l_2) - \text{asinh}(l_1)) & \text{otherwise,} \end{cases}$$

and

$$(5.12) \quad \int_S \text{Dist}(x, y) N ds = \begin{cases} a^2/2 & \text{if } x = \alpha \text{ or } x = \beta, \\ a^2/2 + c & \text{if } c = ab \text{ (} x \text{ aligned with } \vec{\alpha}\beta \text{) and } c > 0, \\ -a^2/2 - c & \text{if } c = -ab \text{ (} x \text{ aligned with } \vec{\alpha}\beta \text{) and } c < 0, \\ \delta/a^2 \left(l_2\sqrt{1+l_2^2} + \text{asinh}(l_2) - l_1\sqrt{1+l_1^2} - \text{asinh}(l_1) \right) & \text{otherwise.} \end{cases}$$

Proof. Let us show how to obtain (5.11) when x, α , and β are not aligned. To do this, let us parametrize the segment $S = [\alpha, \beta]$ so that

$$S = \left\{ y(t) = t \begin{pmatrix} \beta^1 \\ \beta^2 \end{pmatrix} + (1-t) \begin{pmatrix} \alpha^1 \\ \alpha^2 \end{pmatrix}; t \in (0, 1) \right\}.$$

The unitary normal vector of the segment S is given by

$$N = \left(\begin{array}{c} -(\beta^2 - \alpha^2) \\ \beta^1 - \alpha^1 \end{array} \right) \frac{1}{\sqrt{(\beta^1 - \alpha^1)^2 + (\beta^2 - \alpha^2)^2}}.$$

So we have

$$I = \sum_{i=1,2} \int_S \frac{\partial \text{Dist}}{\partial y^i}(x, y) N^i ds = \sum_{i=1,2} \int_0^1 \frac{y^i(t) - x^i}{|x - y(t)|} N^i |\alpha\beta| ds.$$

After some algebraic computations, we get

$$I = \alpha\beta \cdot x\alpha^\perp \int_0^1 \frac{dt}{\sqrt{t^2|\alpha\beta|^2 + |x\alpha|^2 + 2t x\alpha \cdot \alpha\beta}},$$

with $x\alpha^\perp = \left(\begin{array}{c} -(\alpha^2 - x^2) \\ \alpha^1 - x^1 \end{array} \right)$. Using the notation defined in Lemma 5.1, and since $\delta > 0$ (x , α , and β are not aligned), we have

$$I = \alpha\beta \cdot x\alpha^\perp \frac{a}{\sqrt{\delta}} \int_0^1 \frac{dt}{\sqrt{\frac{a^4}{\delta} \left(t + \frac{c}{a^2}\right)^2 + 1}}.$$

We can explicitly compute the integral with the change of variable

$$z = \frac{a^2}{\sqrt{\delta}} \left(t + \frac{c}{a^2}\right),$$

so that we obtain

$$I = \frac{\alpha\beta \cdot x\alpha^\perp}{|\alpha\beta|} (\text{asinh}(l_2) - \text{asinh}(l_1)),$$

which concludes the proof. Other cases follow from similar arguments. \square

With Lemma 5.1, one can estimate (5.9) and (5.10) and thus (5.7). By summing over all the squares and for a given x , we obtain the estimation of the integral $I_{u^k}(x)$ (5.5), and then we can iterate (5.2).

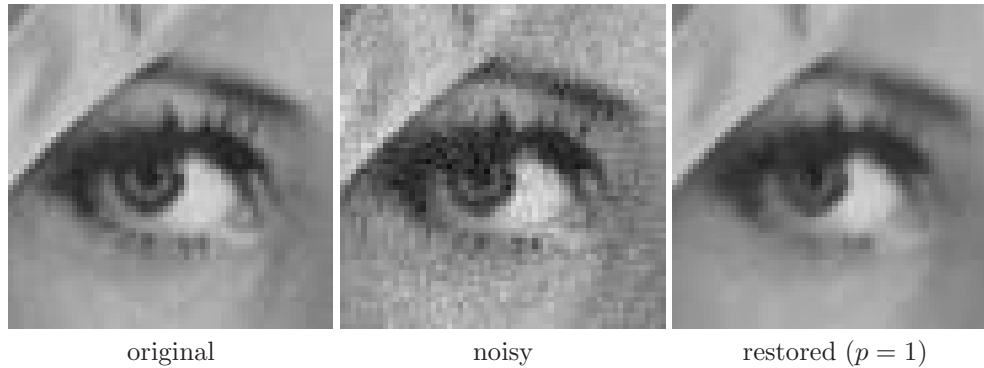
5.2. Experiments on image restoration. Let $u : \Omega \subset R^2 \rightarrow R$ be an original image describing a real scene, and let u_0 be the observed image of the same scene (i.e., a degradation of u). We assume that

$$(5.13) \quad u_0 = Ru + \eta,$$

where η stands for a white additive Gaussian noise and where R is a linear operator representing the blur (usually a convolution). Given u_0 , the problem is then to reconstruct u knowing (5.13). Supposing that η is a white Gaussian noise, and according to the maximum likelihood principle, we can find an approximation of u by solving the least-squares problem

$$\inf_u \int_\Omega |u_0 - Ru|^2 dx,$$

where Ω is the domain of the image. However, this is well known to yield to an ill-posed problem [15, 3].

FIG. 2. *Example of image restoration.*

A classical way to overcome ill-posed minimization problems is to add a regularization term to the energy so that the problem is to minimize

$$(5.14) \quad F(u) = \int_{\Omega} |u_0 - Ru|^2 \, dx + \lambda \int_{\Omega} |\nabla u|^p \, dx.$$

The first term in $F(u)$ measures the fidelity to the data. The second is a smoothing term. In other words, we search for a u that best fits the data so that its gradient is low (so that noise will be removed). The parameter λ is a positive weighting constant. For $p = 1$ we have in fact a BV -norm which leads to discontinuous solutions (see [2] for a review).

Remark that (5.14) is of the form (2.5), with $h(x, u(x)) = |u_0(x) - Ru(x)|^2$. Without loss of generality, we will assume that the operator R is the identity operator. So, in this section, we show some numerical results considering the minimization of the nonlocal functional

$$(5.15) \quad F_n(u) = \int_{\Omega} |u_0 - u|^2 \, dx + \lambda \int_{\Omega} \int_{\Omega} \frac{|u(x) - u(y)|^p}{|x - y|^p} \rho_n(|x - y|) \, dx \, dy$$

for a given n .

The first result, shown in Figure 2, illustrates an image restoration result on a real noisy image for $p = 1$. The result is as expected, which is very close to classical TV results. We recall that this approximation of the BV regularization problem is indeed independent of the fidelity attach term.

The second result, shown in Figure 3, is another image restoration result on a simple synthetic step image, which illustrates the effect of the parameter p on the edges. For example, we recover the classical observation for $p = 1$ or $p = 2$. More importantly, we show that our approximation can be successfully used to handle variational problems posed on $W^{1,p}(\Omega)$ with high values of p which, to our knowledge, generally leads to numerically unstable schemes.

6. Conclusion. Our main contribution was to show that the characterization result due to Bourgain, Brezis, and Mironescu [5] for the Sobolev seminorm can indeed be successfully applied to solve variational problems. It was not a priori straightforward that this characterization of $W^{1,p}$ could be useful in the theoretical and numerical analysis of problems of calculus of variations.

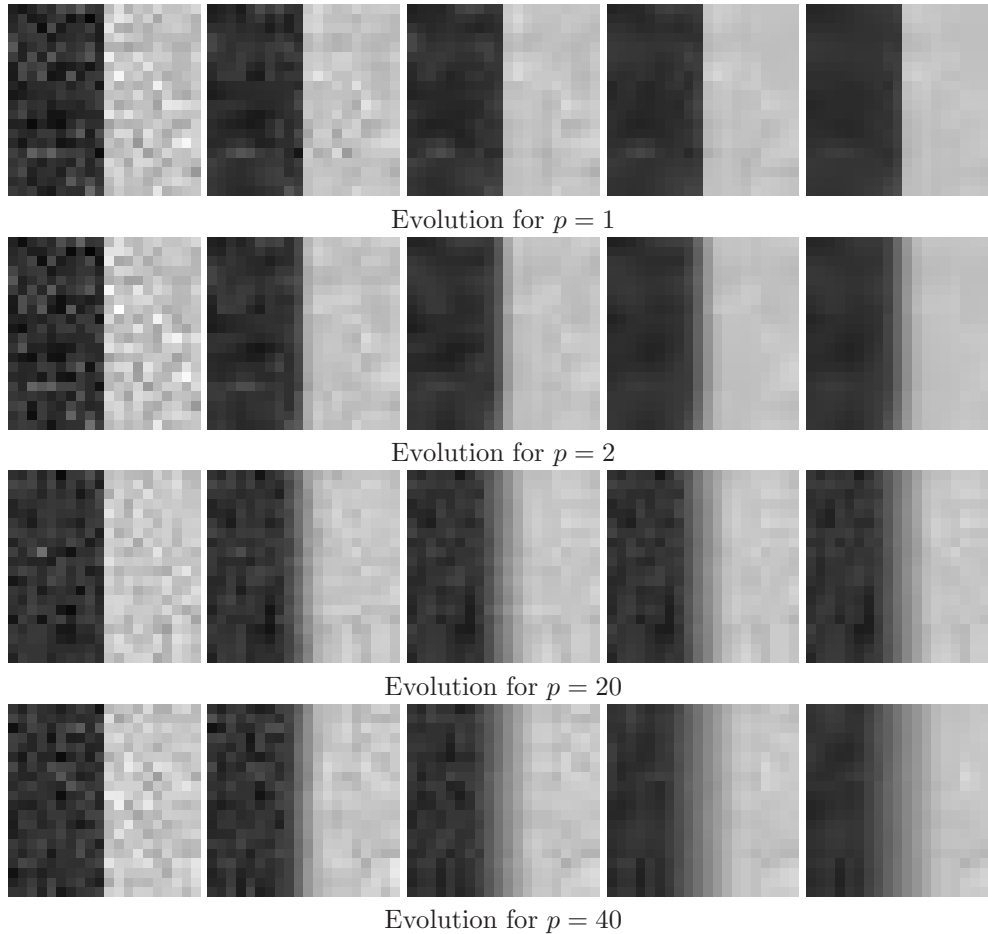


FIG. 3. Example of evolutions with various values of p applied to a synthetic noisy image.

A step further, we proved that our results can be extended also in the BV -case, thanks to Ponce's results [16]. Note that the BV -case is not a simple extension from the $W^{1,p}$ -case, and it requires some adaptations.

Interestingly, we show that this approach allows us to treat problems posed in $W^{1,p}$ with high values of p , which is a challenging problem as far as we know.

Finally, our contribution does not target a particular field of application, and image restoration was proposed here as an illustration: We wanted also to show that this alternative formulation, which leads to nonlocal terms with singular kernels, can be implemented.

REFERENCES

- [1] L. AMBROSIO, N. FUSCO, AND D. PALLARA, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford Math. Monogr., Oxford University Press, New York, 2000.
- [2] G. AUBERT AND P. KORNPBST, *Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations*, 2nd ed., Appl. Math. Sci. 147, Springer-Verlag, New York, 2006.
- [3] M. BERTERO AND P. BOCCACCI, *Introduction to Inverse Problems in Imaging*, Institute of Physics Publishing, Bristol, 1998.

- [4] D. P. BERTSEKAS, *Nonlinear Programming*, 2nd ed., Athena Scientific, Nashua, NH, 1999.
- [5] J. BOURGAIN, H. BREZIS, AND P. MIRONESCU, *Another look at Sobolev spaces*, in *Optimal Control and Partial Differential Equations*, J. L. Menaldi, E. Rofman, and A. Sulem, eds., IOS Press, 2001, pp. 439–455.
- [6] A. BUADES, B. COLL, AND J. M. MOREL, *Neighborhood filters and PDE's*, *Numer. Math.*, 105 (2006), pp. 1–34.
- [7] A. CHAMBOLLE, *An algorithm for total variation minimization and applications*, *J. Math. Imaging Vision*, 20 (2004), pp. 89–97.
- [8] T. F. CHAN, G. H. GOLUB, AND P. MULET, *A nonlinear primal-dual method for total variation-based image restoration*, *SIAM J. Sci. Comput.*, 20 (1999), pp. 1964–1977.
- [9] E. DARVE, *Méthodes multipôles rapides: Résolution des équations de Maxwell par formulations intégrales*, Ph.D. thesis, Université de Paris 6, 1999.
- [10] L. C. EVANS AND R. F. GARIEPY, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.
- [11] G. GILBOA, J. DARBON, S. OSHER, AND T. F. CHAN, *Nonlocal Convex Functionals for Image Regularization*, Technical Report 06-57, UCLA CAM Report, UCLA, Los Angeles, CA, 2006.
- [12] G. GILBOA AND S. OSHER, *Nonlocal linear image regularization and supervised segmentation*, *Multiscale Model. Simul.*, 6 (2007), pp. 595–630.
- [13] M. HINTERMÜLLER AND K. KUNISCH, *Total bounded variation regularization as a bilaterally constrained optimization problem*, *SIAM J. Appl. Math.*, 64 (2004), pp. 1311–1333.
- [14] M. HINTERMÜLLER AND G. STADLER, *An infeasible primal-dual algorithm for total bounded variation-based inf-convolution-type image restoration*, *SIAM J. Sci. Comput.*, 28 (2006), pp. 1–23.
- [15] A. KIRSCH, *An Introduction to the Mathematical Theory of Inverse Problems*, *Appl. Math. Sci.* 120, Springer-Verlag, New York, 1996.
- [16] A. PONCE, *A new approach to Sobolev spaces and connections to γ -convergence*, *Calc. Var. Partial Differential Equations*, 19 (2004), pp. 229–255.
- [17] N. Z. SHOR, *Minimization Methods for Nondifferentiable Functions*, Springer Ser. Comput. Math. 3, Springer-Verlag, Berlin, 1985.
- [18] C. R. VOGEL AND M. E. OMAN, *Fast, robust total variation-based reconstruction of noisy, blurred images*, *IEEE Trans. Image Process.*, 7 (1998), pp. 813–824.
- [19] P. WEISS, L. BLANC-FÉRAUD, AND G. AUBERT, *Efficient schemes for total variation minimization under constraints in image processing*, *SIAM J. Sci. Comput.*, to appear.

A GOAL-ORIENTED ADAPTIVE FINITE ELEMENT METHOD WITH CONVERGENCE RATES*

MARIO S. MOMMER[†] AND ROB STEVENSON[‡]

Abstract. An adaptive finite element method is analyzed for approximating functionals of the solution of symmetric elliptic second order boundary value problems. We show that the method converges and derive a favorable upper bound for its convergence rate and computational complexity. We illustrate our theoretical findings with numerical results.

Key words. adaptive finite element method, convergence rates, computational complexity, quantity of interest, a posteriori error estimators

AMS subject classifications. 65N30, 65N50, 65N15, 65Y20, 41A25

DOI. 10.1137/060675666

1. Introduction. Adaptive finite element methods (AFEMs) have become a standard tool for the numerical solution of partial differential equations. Although used successfully for more than 25 years, in more than one space dimension, even for the most simple case of symmetric elliptic equations of second order $a(u, v) = f(v)$ ($\forall v$), their convergence was not demonstrated before the works of Dörfler [Dör96] and Morin, Nochetto, and Siebert [MNS00]. Convergence alone, however, does not show that the use of an AFEM for a solution that has singularities improves upon, or even competes with, that of a nonadaptive FEM. Recently, after the derivation of such a result by Binev, Dahmen, and DeVore [BDD04] for an AFEM extended with a so-called coarsening routine, in [Ste07] it was shown that standard AFEMs converge with the best possible rate in linear complexity.

The aforementioned works all deal with AFEMs in which the error is measured in the energy norm $\|\cdot\|_E := a(\cdot, \cdot)^{\frac{1}{2}}$. In many applications, however, one is not so much interested in the solution u as a whole, but rather in a (linear) *functional* $g(u)$ of the solution, often being referred to as a *quantity of interest*. With u_τ denoting the finite element approximation of u with respect to a partition τ , from $|g(u) - g(u_\tau)| \leq \|g\|_{E'} \|u - u_\tau\|_E$, obviously it follows that convergence of u_τ towards u with respect to $\|\cdot\|_E$ implies that of $g(u_\tau)$ towards $g(u)$ with at least the same rate. It is, however, generally observed that with adaptive methods especially designed for the approximation of this quantity of interest, known as *goal-oriented adaptive methods*, convergence of $g(u_\tau)$ towards $g(u)$ takes place at a higher rate. Examples of such methods can be found in the monographs [AO00, BR03, BS01], and in references cited therein. So far these goal-oriented adaptive methods are usually not proven to converge. An exception is the method from [DKV06], however, in which adaptivity is purely driven by energy norm minimalization of the error in the *dual problem* $a(v, z) = g(v)$ ($\forall v$). Another exception is the goal-oriented method from [MvSST06], which is

*Received by the editors November 22, 2006; accepted for publication (in revised form) October 16, 2008; published electronically February 6, 2009. This work was supported by the Netherlands Organization for Scientific Research and by the European Community's Human Potential Programme under contract HPRN-CT-2002-00286.

<http://www.siam.org/journals/sinum/47-2/67566.html>

[†]Interdisciplinary Center for Scientific Computing (IWR), Universität Heidelberg, Im Neuenheimer Feld 368, 69120 Heidelberg, Germany (mario.mommer@iwr.uni-heidelberg.de).

[‡]Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands (R.P.Stevenson@uva.nl).

proven to converge with a rate equal to what we will demonstrate (for piecewise linears), where in [MvSST06] the strong assumption $u, z \in C^3(\overline{\Omega})$ was made.

The starting point of our method is the well-known upper bound

$$(1.1) \quad |g(u) - g(u_\tau)| = |a(u - u_\tau, z - z_\tau)| \leq \|u - u_\tau\|_E \|z - z_\tau\|_E,$$

where z_τ is the finite element approximation with respect to τ of z . Having available an AFEM that is convergent with respect to the energy norm, in view of (1.1) an obvious approach would be to use it for finding partitions τ_p and τ_d such that the corresponding finite element approximations u_{τ_p} and z_{τ_d} have, say, both energy norm errors less than $\sqrt{\varepsilon}$. Indeed, then the product of the errors in primal and dual finite element approximations with respect to the smallest common refinement of τ_p and τ_d —and thus the error in the approximation of the quantity of interest—is less than ε . This approach, however, would not benefit from the situation in which, quantitatively or qualitatively, either primal or dual solution is easier to approximate by finite element functions.

The alternative method we propose here works, in essence, as follows. On the k th iteration, we start from a partition τ_k and compute on it the solutions of the primal and dual problems. To advance the iteration, this partition is refined in such a way that the product $\|u - u_\tau\|_E \|z - z_\tau\|_E$ is reduced by a constant factor. To achieve this, we consider the effort needed to reduce each of $\|u - u_\tau\|_E$ and $\|z - z_\tau\|_E$ by the same constant factor, which we do by separately computing suitable refinement sets. The smallest of these sets is then applied to τ_k to obtain τ_{k+1} .

We can show that this method is convergent. In particular, we prove that if, for whatever $s, t > 0$, the solutions of the primal and dual problems can be approximated in energy norm to any accuracy $\delta > 0$ from partitions of cardinality $\mathcal{O}(\delta^{-1/s})$ or $\mathcal{O}(\delta^{-1/t})$, respectively, then given $\varepsilon > 0$, our method constructs a partition of cardinality $\mathcal{O}(\varepsilon^{-1/(s+t)})$ such that

$$|g(u) - g(u_\tau)| \leq \|u - u_\tau\|_E \|z - z_\tau\|_E \leq \varepsilon.$$

In view of the assumptions, this order of cardinality realizing $\|u - u_\tau\|_E \|z - z_\tau\|_E \leq \varepsilon$ is optimal. Moreover, by solving the arising linear systems only inexactly, we show that the overall cost of the algorithm is of order $\mathcal{O}(\varepsilon^{-1/(s+t)})$.

The convergence rate $s + t$ of our goal-oriented method is thus the sum of the rates s and t of the best approximations in energy norm for primal and dual problems. With the approach of approximating both primal and dual problem within tolerance $\sqrt{\varepsilon}$, the rate would be $2 \min(s, t)$. Another alternative approach, namely, to solve each of the problems to an accuracy of $\varepsilon^{s/(s+t)}$ and $\varepsilon^{t/(s+t)}$, respectively, would also result in the rate $s + t$. This approach, however, is not feasible, since the values s and t are generally unknown. Our method converges at the rate $s + t$ without previous knowledge about the regularity of the solutions.

Concerning the value of s (and similarly t), when applying finite elements of order p , for s up to p/n , a rate s is guaranteed when the solution has “ ns orders of smoothness” in $L_\tau(\Omega)$ for some $\tau > (\frac{1}{2} + s)^{-1}$ (instead of in $L_2(\Omega)$ required for nonadaptive approximation) (cf. [BDDP02]).

Our method is based on minimizing an *upper bound* for the error in the functional, which under certain circumstances can be crude. Actually, in all available goal-oriented adaptive methods the decision of which elements have to be refined is based on some upper bound for the error. Unlike the error in energy norm, there exists no computable two-sided bound for the error in a functional of the solution.

This leaves open the possibility that some bounds are “usually” sharper than others. An argument against the upper bound (1.1) brought up in [BR03] is that it is based on the application of a global Cauchy–Schwarz inequality, whereas the dual weighted residual method advocated there would better respect the local information. The contribution of the current paper is that we *prove* a rate that is generally observed with goal-oriented methods. When applying finite element spaces of equal order at primal and dual sides, we neither expect (see Remark 5.1 for details) nor observe in our experiments that on average our bound gets increasingly more pessimistic when the iteration proceeds.

This paper is organized as follows: In section 2, we describe the model boundary value problem that we will consider. The finite element spaces and the refinement rules based on bisections of n -simplices are discussed in section 3. In section 4, we give results on residual-based a posteriori energy error estimators. In section 5, we present our goal-oriented AFEM under the simplifying assumption that the right-hand sides of both primal and dual problems are piecewise polynomial with respect to the initial finite element partition. We derive the aforementioned bound on the cardinality of the output partition. In section 6, the method is extended to general right-hand sides. By replacing the exact solutions of the arising linear systems by inexact ones, it is further shown that the required number of arithmetic operations and storage locations satisfies the same favorable bound as the cardinality of the output partition. Finally, in section 7, we present numerical results obtained with the method. To apply our approach also to unbounded functionals, here we recall the use of extraction functionals, an approach introduced in [BS01].

In this paper, by $C \lesssim D$ we will mean that C can be bounded by a multiple of D , independently of parameters upon which C and D may depend. Similarly, $C \gtrsim D$ is defined as $D \lesssim C$, and $C \approx D$ as $C \lesssim D$ and $C \gtrsim D$.

2. The model problem. Let $\Omega \subset \mathbb{R}^n$ be a polygonal domain. We consider the following model boundary value problem in variational form: Given $f \in H^{-1}(\Omega)$, find $u \in H_0^1(\Omega)$ such that

$$(2.1) \quad a(u, v) := \int_{\Omega} \mathbf{A} \nabla u \cdot \nabla v = f(v) \quad (v \in H_0^1(\Omega)),$$

where $\mathbf{A} \in L_{\infty}(\Omega)$ is a symmetric $n \times n$ matrix with $\text{ess inf}_{x \in \Omega} \lambda_{\min}(\mathbf{A}(x)) > 0$. We assume that \mathbf{A} is piecewise constant with respect to an initial finite element partition τ_0 of Ω specified below. To keep the exposition simple, we do not attempt to derive results that hold uniformly in the size of jumps of $\rho(\mathbf{A})$ over element interfaces, although, under some conditions, this is likely possible; cf. [Ste05]. For $f \in L_2(\Omega)$, we interpret $f(v)$ as $\int_{\Omega} f v$.

Given some $g \in H^{-1}(\Omega)$, we will be interested in $g(u)$. With $z \in H_0^1(\Omega)$ we will denote the solution of the dual problem

$$(2.2) \quad a(v, z) = g(v) \quad (v \in H_0^1(\Omega)).$$

We set the energy norm on $H_0^1(\Omega)$ and dual norm on $H^{-1}(\Omega)$ by

$$\|v\|_E = a(v, v)^{\frac{1}{2}} \quad \text{and} \quad \|h\|_{E'} = \sup_{0 \neq v \in H_0^1(\Omega)} \frac{|h(v)|}{\|v\|_E},$$

respectively.

3. Finite element spaces. Given an essentially disjoint subdivision τ of $\bar{\Omega}$ into (closed) n -simplices, called a partition, we will search approximations for u and z from the finite element space

$$\mathbb{V}_\tau := H_0^1(\Omega) \cap \prod_{T \in \tau} P_p(T),$$

where $0 < p \in \mathbb{N}$ is some fixed constant. For approximating the functionals f and g , we will make use of spaces

$$\mathbb{V}_\tau^* := \prod_{T \in \tau} P_{p-1}(T).$$

Although it is not a finite element space in the usual sense, we also use

$$(3.1) \quad \mathbb{W}_\tau^* := \prod_{T \in \tau} \{\mathbf{h} \in H(\operatorname{div}; T) : \llbracket \mathbf{h} \cdot \mathbf{n} \rrbracket_{\partial T} \in L_2(\partial T)\},$$

with \mathbf{n} being a unit vector normal to ∂T , and $\llbracket \cdot \rrbracket_{\partial T}$ denoting the jump of its argument over ∂T in the direction of \mathbf{n} , defined to be zero on $\partial\Omega$. Obviously, $[\mathbb{V}_\tau^*]^n \subset \mathbb{W}_\tau^*$.

Below, we specify the type of (nested) partitions we will consider, and we recall some results from [Ste08], generalizing upon known results for *newest vertex bisection* in two dimensions.

For $0 \leq k \leq n - 1$, a (closed) simplex spanned by $k + 1$ vertices of an n -simplex T is called a hyperface of T . For $k = n - 1$, it will be called a *true hyperface*. A partition τ is called *conforming* when the intersection of any two different $T, T' \in \tau$ is either empty or a hyperface of both simplices. Different simplices T, T' that share a true hyperface will be called *neighbors*. (Actually, when $\Omega \neq \operatorname{int}(\bar{\Omega})$, the above definition of a conforming partition can be unnecessarily restrictive. We refer to [Ste08] for a discussion of this matter.)

Simplices will be refined by means of bisection. In order to guarantee uniform shape regularity of all descendants, a proper cyclic choice of the refinement edges should be made. To that end, given $\{x_0, \dots, x_n\} \subset \mathbb{R}^n$, not on a joint $(n - 1)$ -dimensional hyperplane, we distinguish between $n(n + 1)!$ *tagged* simplices given by all possible *ordered* sequences $(x_0, x_1, \dots, x_n)_\gamma$ and *types* $\gamma \in \{0, \dots, n - 1\}$. Given a tagged simplex $T = (x_0, x_1, \dots, x_n)_\gamma$, its children are the tagged simplices

$$(x_0, \frac{x_0+x_n}{2}, x_1, \dots, x_\gamma, x_{\gamma+1}, \dots, x_{n-1})_{(\gamma+1) \bmod n}$$

and

$$(x_n, \frac{x_0+x_n}{2}, x_1, \dots, x_\gamma, x_{n-1}, \dots, x_{\gamma+1})_{(\gamma+1) \bmod n},$$

where the sequences $(x_{\gamma+1}, \dots, x_{n-1})$ and (x_1, \dots, x_γ) should be read as being void for $\gamma = n - 1$ and $\gamma = 0$, respectively. So these children are defined by bisecting the edge $\overline{x_0x_n}$ of T —i.e., by connecting its midpoint with the other vertices x_1, \dots, x_{n-1} —by an appropriate ordering of their vertices and by having type $(\gamma + 1) \bmod n$. See Figure 3.1 for an illustration. This bisection process was introduced in [Tra97] and, using different notation, in [Mau95]. The edge $\overline{x_0x_n}$ is called the *refinement edge* of T . In the $n = 2$ case, the vertex opposite this edge is known as the *newest vertex*.

Corresponding to a tagged simplex $T = (x_0, \dots, x_n)_\gamma$, we set

$$T_R = (x_n, x_1, \dots, x_\gamma, x_{n-1}, \dots, x_{\gamma+1}, x_0)_\gamma,$$

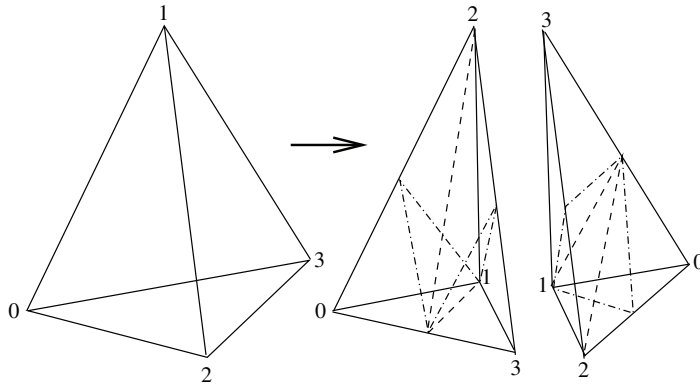


FIG. 3.1. Bisection of a tagged tetrahedron of type 0 with the next two-level cuts indicated.

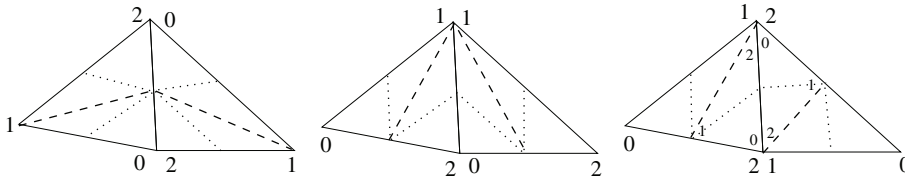


FIG. 3.2. Matching neighbors for $n = 2$, and their level 1 and 2 descendants. The neighbors in the rightmost picture are not reflected neighbors, but the pair of their neighboring children are.

which is the tagged simplex that has the same set of children as T , and in this sense is equal to T . So actually we distinguish between $\frac{1}{2}n(n + 1)!$ tagged simplices.

Given a fixed conforming initial partition τ_0 of tagged simplices of some fixed type γ ,

we will exclusively consider partitions that can be created from τ_0 by recurrent bisections of tagged simplices, in short, descendants of τ_0 .

Simplices that can be created in this way are uniformly shape regular, dependent only on τ_0 and n . For the case that Ω might have slits, we assume that

$$\partial\Omega \text{ is the union of true hyperfaces of } T \in \tau_0.$$

We will assume that the simplices from τ_0 are tagged in such a way that any two neighbors $T = (x_0, \dots, x_n)_\gamma$, $T' = (x'_0, \dots, x'_n)_\gamma$ from P_0 match in the sense that if $\overline{x_0x_n}$ or $\overline{x'_0x'_n}$ is on $T \cap T'$, then either T and T' are reflected neighbors, meaning that the ordered sequence of vertices of either T or T_R coincides with that of T' on all but one position, or the pair of neighboring children of T and T' are reflected neighbors. See Figure 3.2 for an illustration. It is known (see [BDD04] and the references therein) that for any conforming partition into triangles there exists a local numbering of the vertices so that the matching condition is satisfied. We do not now whether the corresponding statement holds in more space dimensions. Yet we showed that any conforming partition of n -simplices can be refined, inflating the number of simplices by not more than an absolute constant factor, into a conforming partition τ_0 that allows a local numbering of the vertices so that the matching condition is satisfied.

For applying a posteriori error estimators, we will require that the partitions τ underlying the approximation spaces be conforming. So in the following

$\tau, \tau', \hat{\tau}$, etc., will always denote conforming partitions.

Bisecting one or more simplices in a conforming partition τ generally results in a nonconforming partition ϱ . Conformity has to be restored by (recursively) bisecting any simplex $T \in \varrho$ that contains a vertex v of a $T' \in \varrho$ that does not coincide with any vertex of T (such a v is called a hanging vertex). This process, called completion, results in the smallest conforming refinement of ϱ .

Our adaptive method will be of the following form:

```

for  $j := 1$  to  $M$ 
do create some, possibly nonconforming refinement  $\varrho_j$  of  $\tau_{j-1}$ 
   complete  $\varrho_j$  to its smallest conforming refinement  $\tau_j$ 
endfor
    
```

As we will see, we will be able to bound $\sum_{j=1}^M \#\varrho_j - \#\tau_{j-1}$. Because of the additional bisections made in the completion steps, however, generally $\#\tau_M - \#\tau_0$ will be larger. The following crucial result, which relies on the matching condition in the initial partition, shows that these additional bisections inflate the total number of simplices by at most an absolute constant factor.

THEOREM 3.1 (generalizes upon [BDD04, Theorem 2.4] for $n = 2$).

$$\#\tau_M - \#\tau_0 \lesssim \sum_{j=1}^M \#\varrho_j - \#\tau_{j-1},$$

dependent only on τ_0 and n , and in particular thus independently of M .

Remark 3.2. Note that this result in particular implies that any descendant ϱ of τ_0 has a conforming refinement τ with $\#\tau \lesssim \#\varrho$, dependent only on τ_0 and n .

We end this section by introducing two more notations. For partitions τ', τ , we write $\tau' \supseteq \tau$ ($\tau' \supset \tau$) to denote that τ' is a (proper) refinement of τ . The smallest common refinement of τ and τ' will be denoted as $\tau \cup \tau'$.

4. A posteriori estimators for the energy error. Given a partition τ , and with u_τ denoting the solution in \mathbb{V}_τ of

$$(4.1) \quad a(u_\tau, v_\tau) = f(v_\tau) \quad (v_\tau \in \mathbb{V}_\tau),$$

in this section we discuss properties of the common residual-based a posteriori error estimator for $\|u - u_\tau\|_E$. Since $a(\cdot, \cdot)$ is symmetric, an analogous result will apply to $\|z - z_\tau\|_E$, with z_τ denoting the solution in \mathbb{V}_τ of

$$(4.2) \quad a(v_\tau, z_\tau) = g(v_\tau) \quad (v_\tau \in \mathbb{V}_\tau).$$

By formally viewing $H_0^1(\Omega)$ as \mathbb{V}_τ corresponding to the infinitely uniformly refined partition $\tau = \infty$, at some places we interpreted results derived for u_τ to hold for the solution u of (2.1) by substituting $\tau = \infty$.

For developing an AFEM that reduces the error in each iteration, it will be necessary to approximate the right-hand side by discrete functions. Loosely speaking, in [MNS00] the error in this approximation is called *data oscillation*. Being on a partition τ , it will be allowed to use functions from $\mathbb{V}_\tau^* + \text{div}[\mathbb{V}_\tau^*]^n$, where $\text{div} := (-\nabla)' : L_2(\Omega)^n \rightarrow H^{-1}(\Omega)$. Depending on the right-hand side at hand, it might be

more convenient to approximate it by functions from \mathbb{V}_τ^* or from $\text{div}[\mathbb{V}_\tau^*]^n$, or by a combination of these. In view of this, we will write

$$(4.3) \quad f = f^1 + \text{div} \mathbf{f}^2,$$

where $f^1 \in H^{-1}(\Omega)$ and $\mathbf{f}^2 \in L_2(\Omega)^n$ are going to be approximated by functions from \mathbb{V}_τ^* or from $\text{div}[\mathbb{V}_\tau^*]^n$, respectively. Similarly, we write $g = g^1 + \text{div} \mathbf{g}^2$.

Remark 4.1. Obviously, any $f \in H^{-1}(\Omega)$ can be written in the above form with vanishing \mathbf{f}^2 . On the other hand, by taking $\mathbf{f}^2 = -\nabla w$ with $w \in H_0^1(\Omega)$ being the solution of $\int_\Omega \nabla w \cdot \nabla v = f(v)$ ($v \in H_0^1(\Omega)$), we see that we can equally well consider a vanishing f^1 .

For $\bar{u}_\tau \in \mathbb{V}_\tau$, $\bar{f}^1 \in L_2(\Omega)$, and $\bar{\mathbf{f}}^2 \in \mathbb{W}_\tau^*$ (see (3.1)), where we have in mind approximations to u_τ , f^1 , and \mathbf{f}^2 , respectively, and $T \in \tau$, we set the local error indicator

$$\begin{aligned} \eta_T(\bar{f}^1, \bar{\mathbf{f}}^2, \bar{u}_\tau) &:= \text{diam}(T)^2 \|\bar{f}^1 + \nabla \cdot [\mathbf{A} \nabla \bar{u}_\tau + \bar{\mathbf{f}}^2]\|_{L_2(T)}^2 \\ &\quad + \text{diam}(T) \|\llbracket [\mathbf{A} \nabla \bar{u}_\tau + \bar{\mathbf{f}}^2] \cdot \mathbf{n} \rrbracket_{\partial T}\|_{L_2(\partial T)}^2. \end{aligned}$$

Note that the first term is the weighted local residual of the equation in strong form. We set the energy error estimator

$$\mathcal{E}(\tau, \bar{f}^1, \bar{\mathbf{f}}^2, \bar{u}_\tau) := \left[\sum_{T \in \tau} \eta_T(\bar{f}^1, \bar{\mathbf{f}}^2, \bar{u}_\tau) \right]^{\frac{1}{2}}.$$

The following Proposition 4.2 is a generalization of [Ste07, Theorem 4.1] valid for $\mathbf{A} = \text{Id}$, $\mathbf{f}^2 = 0$, and polynomial degree $p = 1$. This result in turn was a generalization of [BMN02, Lemma 5.1, eq. (5.4)] (see also [Ver96]) in the sense that instead of $\|u - u_\tau\|_E$, the difference $\|u_{\tau'} - u_\tau\|_E$ for any $\tau' \supset \tau$ is estimated. Proposition 4.2 tells us that this difference can be bounded from above by the square root of the sum of the local error indicators corresponding to those simplices from τ that either are not in τ' since they were refined or have nonempty intersection with such simplices. By taking $\tau' = \infty$, this result yields the known bound for $\|u - u_\tau\|_E$.

PROPOSITION 4.2. *Let $\tau' \supset \tau$ be partitions, and let $f^1 \in L_2(\Omega)$, $\mathbf{f}^2 \in \mathbb{W}_\tau^*$, and*

$$G = G(\tau, \tau') := \{T \in \tau : T \cap \tilde{T} \neq \emptyset \text{ for some } \tilde{T} \in \tau, \tilde{T} \notin \tau'\}.$$

Then we have

$$\|u_{\tau'} - u_\tau\|_E \leq C_1 \left[\sum_{T \in G} \eta_T(f^1, \mathbf{f}^2, u_\tau) \right]^{\frac{1}{2}}$$

for some absolute constant $C_1 > 0$. Note that $\#G \lesssim \#\tau' - \#\tau$.

In particular, by taking $\tau' = \infty$, we have

$$(4.4) \quad \|u - u_\tau\|_E \leq C_1 \mathcal{E}(\tau, f^1, \mathbf{f}^2, u_\tau).$$

Proof. We have $\|u_{\tau'} - u_\tau\|_E = \sup_{0 \neq v_{\tau'} \in \mathbb{V}_{\tau'}} \frac{|a(u_{\tau'} - u_\tau, v_{\tau'})|}{\|v_{\tau'}\|_E}$. For any $v_{\tau'} \in \mathbb{V}_{\tau'}$, $v_\tau \in \mathbb{V}_\tau$, we have

$$\begin{aligned} a(u_{\tau'} - u_\tau, v_{\tau'}) &= a(u_{\tau'} - u_\tau, v_{\tau'} - v_\tau) \\ &= \sum_T \int_T f^1(v_{\tau'} - v_\tau) - \mathbf{f}^2 \cdot \nabla(v_{\tau'} - v_\tau) - \mathbf{A} \nabla u_{\tau'} \cdot \nabla(v_{\tau'} - v_\tau) \\ &= \sum_T \left\{ (f^1 + \nabla \cdot [\mathbf{A} \nabla u_\tau + \mathbf{f}^2])(v_{\tau'} - v_\tau) - \int_{\partial T} [\mathbf{A} \nabla u_\tau + \mathbf{f}^2] \cdot \mathbf{n}(v_{\tau'} - v_\tau) \right\}, \end{aligned}$$

where the last line follows by integration by parts. By taking v_τ to be a suitable local quasi-interpolant of $v_{\tau'}$ as in [Ste07] (for $p > 1$, one may consult [KS08]) or, alternatively, a Clément-type interpolator, and applying a Cauchy–Schwarz inequality, one completes the proof. \square

Remark 4.3. For the lowest order elements, i.e., $p = 1$, a statement similar to Proposition 4.2 is valid with error indicators consisting of the jump terms over the interfaces only. As a consequence, along the lines that we will follow for elements of general degree p , for $p = 1$ a cheaper goal-oriented AFEM can be developed that has similar properties. Details can be found in Appendix A of the extended preprint version [MS08] of this work.

Next we study whether the error estimator also provides a lower bound for $\|u - u_\tau\|_E$ and, when τ' is a sufficient refinement of τ , for $\|u_{\tau'} - u_\tau\|_E$. In order to derive such estimates, for the moment we further restrict the type of right-hand sides. The proof of the following proposition will be derived along the lines of the proof of [BMN02, Lemma 5.3], where the Stokes problem is considered (see also [MNS00, Lemma 4.2] for the case $p = 1$ and $\mathbf{f}^2 = 0$). For convenience of the reader we include it here.

PROPOSITION 4.4. *Let $\tau \subset \tau'$ be partitions, and let $f^1 \in \mathbb{V}_\tau^*$, $\mathbf{f}^2 \in [\mathbb{V}_\tau^*]^n$, and $\bar{u}_\tau \in \mathbb{V}_\tau$.*

(a) *If $T \in \tau$ contains a vertex of τ' in its interior, then*

$$\text{diam}(T)^2 \|f^1 + \nabla \cdot [\mathbf{A}\nabla \bar{u}_\tau + \mathbf{f}^2]\|_{L_2(T)}^2 \lesssim |u_{\tau'} - \bar{u}_\tau|_{H^1(T)}^2.$$

(b) *If a joint true hyperface e of $T_1, T_2 \in \tau$ contains a vertex of τ' in its interior, then*

$$\begin{aligned} \text{diam}(e) \| [[\mathbf{A}\nabla \bar{u}_\tau + \mathbf{f}^2] \cdot \mathbf{n}]_e \|_{L_2(e)}^2 &\lesssim |u_{\tau'} - \bar{u}_\tau|_{H^1(T_1 \cup T_2)}^2 \\ &+ \sum_{i=1}^2 \text{diam}(T_i)^2 \|f^1 + \nabla \cdot [\mathbf{A}\nabla \bar{u}_\tau + \mathbf{f}^2]\|_{L_2(T_i)}^2. \end{aligned}$$

Proof. Let $\phi_T \in H_0^1(\Omega) \cap \prod_{T' \in \tau'} P_1(T')$ be the canonical nodal basis function associated to a vertex of τ' inside T . Writing $R_T = (f^1 + \nabla \cdot [\mathbf{A}\nabla \bar{u}_\tau + \mathbf{f}^2])|_T \in P_{d-1}(T)$, and $v_{\tau'} = R_T \phi_T \in \mathbb{V}_{\tau'}$, using the fact that $\text{supp } v_{\tau'} \subset T$, by integration by parts we get

$$\begin{aligned} \int_T R_T^2 &\lesssim \int_T R_T^2 \phi_T = \int_T R_T v_{\tau'} = (f_1 + \text{div } \mathbf{f}^2)(v_{\tau'}) - \int_T \mathbf{A}\nabla \bar{u}_\tau \cdot \nabla v_{\tau'} \\ &= \int_T \mathbf{A}\nabla (u_{\tau'} - \bar{u}_\tau) \cdot \nabla v_{\tau'}, \end{aligned}$$

and so by $|v_{\tau'}|_{H^1(T)} \lesssim \text{diam}(T)^{-1} \|v_{\tau'}\|_{L_2(T)} \lesssim \text{diam}(T)^{-1} \|R_T\|_{L_2(T)}$, we infer (a).

Let $\phi_e \in H_0^1(\Omega) \cap \prod_{T' \in \tau'} P_1(T')$ be the canonical nodal basis function associated to a vertex interior to e . Writing $J_e = [[\mathbf{A}\nabla \bar{u}_\tau + \mathbf{f}^2] \cdot \mathbf{n}]_e \in P_{d-1}(e)$, let $\bar{J}_e \in P_{d-1}(T_1 \cup T_2)$ denote its extension constant in the direction normal to e , and let $v_{\tau'} = \bar{J}_e \phi_e \in \mathbb{V}_{\tau'}$. Using the fact that $\text{supp } v_{\tau'} \subset T_1 \cup T_2$, by integration by parts we get

$$\int_e J_e^2 \lesssim \int_e J_e^2 \phi_e = \int_e J_e v_{\tau'} = \int_{T_1 \cup T_2} (\mathbf{A}\nabla \bar{u}_\tau + \mathbf{f}^2) \cdot \nabla v_{\tau'} + \int_{T_1 \cup T_2} \nabla \cdot (\mathbf{A}\nabla \bar{u}_\tau + \mathbf{f}^2) v_{\tau'}.$$

From

$$\int_{T_1 \cup T_2} \mathbf{f}^2 \cdot \nabla v_{\tau'} = -\operatorname{div} \mathbf{f}^2(v_{\tau'}) = -a(u_{\tau'}, v_{\tau'}) + \int_{T_1 \cup T_2} f^1 v_{\tau'},$$

we infer

$$\begin{aligned} \int_e J_e^2 &\lesssim a(\bar{u}_\tau - u_{\tau'}, v_{\tau'}) + \int_{T_1 \cup T_2} (f^1 + \nabla \cdot (\mathbf{A} \nabla \bar{u}_\tau + \mathbf{f}^2)) v_{\tau'} \\ &\lesssim \left[|\bar{u}_\tau - u_{\tau'}|_{H^1(T_1 \cup T_2)} \operatorname{diam}(e)^{-1} + \sum_{i=1}^2 \|R_{T_i}\|_{L_2(T_i)} \right] \|v_{\tau'}\|_{L_2(T_1 \cup T_2)}. \end{aligned}$$

Using the fact that $\|v_{\tau'}\|_{L_2(T_1 \cup T_2)} \approx \|\bar{J}_e\|_{L_2(T_1 \cup T_2)} \approx \operatorname{diam}(e)^{\frac{1}{2}} \|J_e\|_{L_2(e)}$, we infer item (b) of the proposition. \square

In view of this last result, we will call a (possibly nonconforming) $\varrho \supset \tau$ a *full refinement with respect to $T \in \tau$* when

T , and its neighbors in τ , as well as all true hyperfaces of T , all contain a vertex of ϱ in their interiors.

As a direct consequence of Proposition 4.4 we have the following.

COROLLARY 4.5. *Let τ be a partition, let $f^1 \in \mathbb{V}_\tau^*$, $\mathbf{f}^2 \in [\mathbb{V}_\tau^*]^n$, and $\bar{u}_\tau \in \mathbb{V}_\tau$, and let $\tau' \supset \tau$ be a full refinement of τ with respect to all T from some $F \subset \tau$. Then*

$$(4.5) \quad c_2 \left[\sum_{T \in F} \eta_T(f^1, \mathbf{f}^2, \bar{u}_\tau) \right]^{\frac{1}{2}} \leq \|u_{\tau'} - \bar{u}_\tau\|_E$$

for some absolute constant $c_2 > 0$. In particular, we have

$$(4.6) \quad c_2 \mathcal{E}(\tau, f^1, \mathbf{f}^2, \bar{u}_\tau) \leq \|u - \bar{u}_\tau\|_E.$$

Next, we investigate the stability of the energy error estimator.

PROPOSITION 4.6. *Let τ be a partition, and let $f^1 \in L_2(\Omega)$, $\mathbf{f}^2 \in \mathbb{W}_\tau^*$, and $v_\tau, w_\tau \in \mathbb{V}_\tau$. Then*

$$c_2 |\mathcal{E}(\tau, f^1, \mathbf{f}^2, v_\tau) - \mathcal{E}(\tau, f^1, \mathbf{f}^2, w_\tau)| \leq \|v_\tau - w_\tau\|_E.$$

Proof. For $\tilde{f}^1 \in L_2(\Omega)$, $\tilde{\mathbf{f}}^2 \in \mathbb{W}_\tau^*$, and $v_\tau, w_\tau \in \mathbb{V}_\tau$, by two applications of the triangle inequality in the form $\| \|\cdot\| - \|\cdot\| \|^2 \leq \| \cdot \cdot \|^2$, first for vectors and then for functions, we have

$$|\mathcal{E}(\tau, f^1, \mathbf{f}^2, v_\tau) - \mathcal{E}(\tau, \tilde{f}^1, \tilde{\mathbf{f}}^2, w_\tau)| \leq \mathcal{E}(\tau, f^1 - \tilde{f}^1, \mathbf{f}^2 - \tilde{\mathbf{f}}^2, v_\tau - w_\tau).$$

By substituting $\tilde{f}^1 = f^1$ and $\tilde{\mathbf{f}}^2 = \mathbf{f}^2$, and by applying (4.6) the proof is complete. \square

5. An idealized goal-oriented AFEM. From (2.2) and $u - u_\tau \perp_{a(\cdot, \cdot)} \mathbb{V}_\tau \ni z_\tau$, we have

$$(5.1) \quad |g(u) - g(u_\tau)| = |a(u - u_\tau, z)| = |a(u - u_\tau, z - z_\tau)| \leq \|u - u_\tau\|_E \|z - z_\tau\|_E.$$

We will develop an adaptive method for minimizing the right-hand side of this expression.

Remark 5.1. A question that naturally arises is whether there is something to be gained from using finite elements of different orders for the dual and the primal problems. Note that the derivation of (5.1) remains valid if the dual solution is computed in a lower order space, or for that matter in any space that is a subspace of \mathbb{V}_τ . But this will result in a larger $\|z - z_\tau\|_E$, worsening our error estimate without changing the actual error $|g(u) - g(u_\tau)|$.

And how about using a higher order space for the dual problem? In this case, (5.1) no longer holds. As $g(u) = f(z)$, we can approximate it by $f(z_\tau)$ with

$$(5.2) \quad |f(z) - f(z_\tau)| = |a(u, z - z_\tau)| = |a(u - u_\tau, z - z_\tau)| \leq \|u - u_\tau\|_E \|z - z_\tau\|_E.$$

Thus, as before, we obtain a worse error estimate than if we had used the same higher order space for the primal problem as well.

We conclude that with our approach there is no gain from using different orders and, accordingly, will consider here only spaces of equal order.

Up to and including Lemma 5.3, we start with discussing a method for reducing $\|u - u_\tau\|_E$ or similarly $\|z - z_\tau\|_E$ separately. For some *fixed*

$$\theta \in \left(0, \frac{c_2}{C_1}\right),$$

we will make use of the following routine to mark simplices for refinement:

MARK $[\tau, \bar{f}^1, \bar{f}^2, \bar{u}_\tau] \rightarrow F$

% $\bar{f}^1 \in L_2(\Omega)$, $\bar{f}^2 \in \mathbb{W}_\tau^*$, $\bar{u}_\tau \in \mathbb{V}_\tau$.

Select, in $\mathcal{O}(\#\tau)$ operations, a set $F \subset \tau$ with, up to some absolute factor, minimal cardinality such that

$$(5.3) \quad \sum_{T \in F} \eta_T(\bar{f}^1, \bar{f}^2, \bar{u}_\tau) \geq \theta^2 \mathcal{E}(\tau, \bar{f}^1, \bar{f}^2, \bar{u}_\tau)^2.$$

Remark 5.2. Selecting F that satisfies (5.3) with truly minimal cardinality would require the sorting of all $\eta_T = \eta_T(\bar{f}^1, \bar{f}^2, \bar{u}_\tau)$, which takes $\mathcal{O}(\#\tau \log(\#\tau))$ operations. The log-factor can be avoided by performing an approximate sorting based on binning that we recall here: With $N := \#\tau$, we may discard all $\eta_T \leq (1 - \theta^2)\mathcal{E}(\tau, \bar{f}^1, \bar{f}^2, \bar{u}_\tau)^2/N$. With $M := \max_{T \in \tau} \eta_T$, and q the smallest integer with $2^{-q-1}M \leq (1 - \theta^2)\mathcal{E}(P^c, \bar{f}^1, \bar{f}^2, w_{P^c})^2/N$, we store the others in $q + 1$ bins depending on whether η_T is in $[M, \frac{1}{2}M)$, $[\frac{1}{2}M, \frac{1}{4}M)$, \dots , or $[2^{-q}M, 2^{-q-1}M)$. Then we build F by extracting η_T from the bins, starting with the first bin, moving to the second bin when the first is empty, and so on until (5.3) is satisfied. Let the resulting \tilde{F} now contain η_T from the ℓ th bin, but not from further bins. Then a minimal set \tilde{F} that satisfies (5.3) contains all η_T from the bins up to the $(\ell - 1)$ th one. Since any two η_T in the ℓ th bin differ at most by a factor of 2, we infer that the cardinality of the contribution from the ℓ th bin to F is at most twice as large as that to \tilde{F} , so that $\#F \leq 2\#\tilde{F}$. Assuming that each evaluation of η_T takes $\mathcal{O}(1)$ operations, the number of operations and storage locations required by this procedure is $\mathcal{O}(q + \#\tau)$, with $q < \log_2(MN/[(1 - \theta^2)\mathcal{E}(\tau, \bar{f}^1, \bar{f}^2, \bar{u}_\tau)^2]) \leq \log_2(N/(1 - \theta^2)) \lesssim \log_2(\#\tau) < \#\tau$. The assumption on the cost of evaluating η_T is satisfied when $f^1 \in \mathbb{V}_\tau^*$ and $f^2 \in [\mathbb{V}_\tau^*]^n$, as will be the case in our applications.

Having a set of marked elements F , the next step is to apply the following:

REFINE $[\tau, F] \rightarrow \tau'$

% Determines the smallest $\tau' \supseteq \tau$ which is a full refinement
 % with respect to all $T \in F$.

The cost of the call is $\mathcal{O}(\#\tau')$ operations.

Using the results on the a posteriori error estimator derived in the previous section, we have the following result.

LEMMA 5.3. *Let $f^1 \in \mathbb{V}_\tau^*$, $\mathbf{f}^2 \in [\mathbb{V}_\tau^*]^n$. Then for $F = \mathbf{MARK}[\tau, f^1, \mathbf{f}^2, u_\tau]$ and $\tau' \supseteq \mathbf{REFINE}[\tau, F]$, we have*

$$(5.4) \quad \|u - u_{\tau'}\|_E \leq \left[1 - \frac{c_2^2 \theta^2}{C_1^2}\right]^{\frac{1}{2}} \|u - u_\tau\|_E.$$

Furthermore

$$\#F \lesssim \#\hat{\tau} - \#\tau_0$$

for any partition $\hat{\tau}$ for which

$$\|u - u_{\hat{\tau}}\|_E \leq \left[1 - \frac{C_1^2 \theta^2}{c_2^2}\right]^{\frac{1}{2}} \|u - u_\tau\|_E.$$

Proof. Since this is a key result, for convenience of the reader we recall the arguments from [Ste07].

From

$$\|u - u_\tau\|_E^2 = \|u - u_{\tau'}\|_E^2 + \|u_{\tau'} - u_\tau\|_E^2$$

and, by (4.5), (5.3), and (4.4),

$$\|u_{\tau'} - u_\tau\|_E \geq c_2 \theta \mathcal{E}(\tau, f^1, \mathbf{f}^2, u_\tau) \geq \frac{c_2 \theta}{C_1} \|u - u_\tau\|_E,$$

we conclude (5.4).

With $\hat{\tau}$ being a partition as in the statement of the theorem, let $\check{\tau} = \tau \cup \hat{\tau}$. Then, as τ and $\hat{\tau}$, the partition $\check{\tau}$ is a conforming descendant of τ_0 , $\|u - u_{\check{\tau}}\|_E \leq \|u - u_{\hat{\tau}}\|_E$, and

$$\#\check{\tau} - \#\tau \leq \#\hat{\tau} - \#\tau_0.$$

To see the last statement, note that each simplex in $\check{\tau}$ that is not in τ is in $\hat{\tau}$. Therefore, since $\tau \supset \tau_0$, the number of bisections needed to create $\check{\tau}$ from τ , whose number is equal to $\#\check{\tau} - \#\tau$, is not larger than the number of bisections needed to create $\hat{\tau}$ from τ_0 , whose number is equal to $\#\hat{\tau} - \#\tau_0$.

With $G = G(\tau, \check{\tau})$ from Proposition 4.2, we have

$$\begin{aligned} C_1^2 \sum_{T \in G} \eta_T(f^1, \mathbf{f}^2, u_\tau) &\geq \|u_{\check{\tau}} - u_\tau\|_E^2 = \|u - u_\tau\|_E^2 - \|u - u_{\check{\tau}}\|_E^2 \\ &\geq \frac{C_1^2 \theta^2}{c_2^2} \|u - u_\tau\|_E^2 \geq C_1^2 \theta^2 \mathcal{E}(\tau, f^1, \mathbf{f}^2, u_\tau)^2 \end{aligned}$$

by (4.6). By construction of F , we conclude that

$$\#F \lesssim \#G \lesssim \#\check{\tau} - \#\tau \leq \#\hat{\tau} - \#\tau_0,$$

which completes the proof. \square

The idea of the goal-oriented AFEM will be to mark sets of simplices for refinement corresponding to both primal and dual problems, and then to perform the actual refinement corresponding to that set of marked simplices that has the smallest cardinality. In order to assess the quality of the method, we first introduce the approximation classes \mathcal{A}^s .

For $s > 0$, we define

$$\mathcal{A}^s = \left\{ u \in H_0^1(\Omega) : |u|_{\mathcal{A}^s} := \sup_{\varepsilon > 0} \varepsilon \inf_{\{\tau: \|u - u_\tau\|_E \leq \varepsilon\}} [\#\tau - \#\tau_0]^s < \infty \right\}$$

and equip it with norm $\|u\|_{\mathcal{A}^s} := \|u\|_E + |u|_{\mathcal{A}^s}$. So \mathcal{A}^s is the class of functions that can be approximated within any given tolerance $\varepsilon > 0$ in $\|\cdot\|_E$ by a continuous piecewise polynomial of degree p on a partition τ with $\#\tau - \#\tau_0 \leq \varepsilon^{-1/s} |u|_{\mathcal{A}^s}^{1/s}$.

Remark 5.4. Although in the definition of \mathcal{A}^s we consider only conforming descendants τ of τ_0 , in view of Remark 3.2, we note that these approximation classes would remain the same if we would replace τ by any descendant ϱ of τ_0 , conforming or not.

While the \mathcal{A}^s contain \mathbb{V}_τ for any s , and thus are never empty, only the range $s \leq p/n$ is of interest, as even C^∞ functions are only guaranteed to belong to \mathcal{A}^s for this range. Classical estimates show that for $s \leq p/n$, $H^{1+p}(\Omega) \cap H_0^1(\Omega) \subset \mathcal{A}^s$, where it is sufficient to consider uniform refinements. The class \mathcal{A}^s is much larger than $H^{1+p}(\Omega) \cap H_0^1(\Omega)$, which is the reason to consider adaptive methods in the first place. A (near) characterization of \mathcal{A}^s for $s \leq p/n$ in terms of Besov spaces can be found in [BDDP02] (although there the case $n = 2$ and $p = 1$ is considered, results easily generalize).

We now consider the following adaptive algorithm:

```

GOAFEM[ $f^1, \mathbf{f}^2, g^1, \mathbf{g}^2, \varepsilon$ ]  $\rightarrow$  [ $\tau_n, u_{\tau_n}, z_{\tau_n}$ ]
% For this preliminary version of the goal-oriented AFEM,
% it is assumed that  $f^1, g^1 \in \mathbb{V}_{\tau_0}^*$  and  $\mathbf{f}^2, \mathbf{g}^2 \in [\mathbb{V}_{\tau_0}^*]^n$ .
 $k := 0$ 
while  $C_1 \mathcal{E}(\tau_k, f^1, \mathbf{f}^2, u_{\tau_k}) \cdot C_1 \mathcal{E}(\tau_k, g^1, \mathbf{g}^2, z_{\tau_k}) > \varepsilon$  do
     $F_p := \mathbf{MARK}[\tau_k, f^1, \mathbf{f}^2, u_{\tau_k}]$ 
     $F_d := \mathbf{MARK}[\tau_k, g^1, \mathbf{g}^2, z_{\tau_k}]$ 
    With  $F$  being the smallest of  $F_p$  and  $F_d$ ,  $\tau_{k+1} := \mathbf{REFINE}[\tau_k, F]$ 
     $k := k + 1$ 
end do
 $n := k$ 
    
```

THEOREM 5.5. *Let $f^1, g^1 \in \mathbb{V}_{\tau_0}^*$ and $\mathbf{f}^2, \mathbf{g}^2 \in [\mathbb{V}_{\tau_0}^*]^n$. Then $[\tau_n, u_{\tau_n}, z_{\tau_n}] = \mathbf{GOAFEM}[f^1, \mathbf{f}^2, g^1, \mathbf{g}^2, \varepsilon]$ terminates, and $\|u - u_{\tau_n}\|_E \|z - z_{\tau_n}\|_E \leq \varepsilon$. If $u \in \mathcal{A}^s$ and $z \in \mathcal{A}^t$, then*

$$\#\tau_n - \#\tau_0 \lesssim \varepsilon^{-1/(s+t)} (|u|_{\mathcal{A}^s} |z|_{\mathcal{A}^t})^{1/(s+t)},$$

dependent only on τ_0 , and on s or t when they tend to 0 or ∞ .

Remark 5.6. Assuming only that $u \in \mathcal{A}^s$ and $z \in \mathcal{A}^t$, given a partition τ , the generally smallest upper bound for the product of the errors in energy norm in primal and dual solutions that can be expected is $[\#\tau - \#\tau_0]^{-s} |u|_{\mathcal{A}^s} [\#\tau - \#\tau_0]^{-t} |z|_{\mathcal{A}^t}$. Setting this expression equal to ε , one finds $\#\tau - \#\tau_0 = \varepsilon^{-1/(s+t)} (|u|_{\mathcal{A}^s} |z|_{\mathcal{A}^t})^{1/(s+t)}$.

We conclude that the partition produced by **GOAFEM** is at most a constant factor larger than the generally smallest partition τ for which $\|u - u_\tau\|_E \|z - z_\tau\|_E$ is less than the prescribed tolerance.

Proof. Let $E_k := \|u - u_{\tau_k}\|_E \|z - z_{\tau_k}\|_E$. Then $E_{k+1} \leq [1 - \frac{c_2^2 \theta^2}{C_1^2}]^{\frac{1}{2}} E_k$ by (5.4), and $c_2 \mathcal{E}(\tau_k, f^1, \mathbf{f}^2, u_{\tau_k}) c_2 \mathcal{E}(\tau_k, g^1, \mathbf{g}^2, z_{\tau_k}) \leq E_k$ by (4.6). So **GOAFEM** $[f^1, \mathbf{f}^2, g^1, \mathbf{g}^2, \varepsilon]$ terminates, with $E_n \leq C_1 \mathcal{E}(\tau_n, f^1, \mathbf{f}^2, u_{\tau_n}) C_1 \mathcal{E}(\tau_n, g^1, \mathbf{g}^2, z_{\tau_n}) \leq \varepsilon$ by (4.4).

With F_k being the set of marked cells inside the k th call of **REFINE**, Lemma 5.3 and the assumptions $u \in \mathcal{A}^s, z \in \mathcal{A}^t$ show that

$$\begin{aligned} \#F_k &\leq \min \left\{ \left[1 - \frac{C_1^2 \theta^2}{c_2^2} \right]^{-\frac{1}{2s}} \|u - u_{\tau_{k-1}}\|_E^{-1/s} |u|_{\mathcal{A}^s}^{1/s}, \left[1 - \frac{C_1^2 \theta^2}{c_2^2} \right]^{-\frac{1}{2t}} \|z - z_{\tau_{k-1}}\|_E^{-1/t} |z|_{\mathcal{A}^t}^{1/t} \right\} \\ &\lesssim \min \{ \|u - u_{\tau_{k-1}}\|_E^{-1/s} |u|_{\mathcal{A}^s}^{1/s}, \|z - z_{\tau_{k-1}}\|_E^{-1/t} |z|_{\mathcal{A}^t}^{1/t} \} \\ &\leq \max_{\delta \eta \geq E_{k-1}} \min \{ \delta^{-1/s} |u|_{\mathcal{A}^s}^{1/s}, \eta^{-1/t} |z|_{\mathcal{A}^t}^{1/t} \} = E_{k-1}^{-1/(s+t)} (|u|_{\mathcal{A}^s} |z|_{\mathcal{A}^t})^{1/(s+t)}. \end{aligned}$$

The partition τ_k is the smallest conforming refinement of the generally nonconforming ϱ_k , defined as the smallest refinement of τ_{k-1} which is a full refinement with respect to all $T \in F_k$. From Theorem 3.1, $\#\varrho_k - \#\tau_{k-1} \lesssim \#F_k$, the majorized linear convergence of $k \mapsto E_{k-1}$, and $E_{n-1} > \frac{c_2^2}{C_1^2} \varepsilon$, we conclude that

$$\begin{aligned} \#\tau_n - \#\tau_0 &\lesssim \sum_{k=1}^n \#F_k \lesssim E_{n-1}^{-1/(s+t)} (|u|_{\mathcal{A}^s} |z|_{\mathcal{A}^t})^{1/(s+t)} \\ &\lesssim \varepsilon^{-1/(s+t)} (|u|_{\mathcal{A}^s} |z|_{\mathcal{A}^t})^{1/(s+t)}. \quad \square \end{aligned}$$

6. A practical goal-oriented AFEM. So far, we assumed that $f = f^1 + \text{div} \mathbf{f}^2, g = g^1 + \text{div} \mathbf{g}^2$, with $f^1, g^1 \in \mathbb{V}_\tau^*, \mathbf{f}^2, \mathbf{g}^2 \in [\mathbb{V}_\tau^*]^n$ for any partition τ that we encountered; i.e., we assumed that $f^1, g^1 \in \mathbb{V}_{\tau_0}^*, \mathbf{f}^2, \mathbf{g}^2 \in [\mathbb{V}_{\tau_0}^*]^n$. From now on, given a partition τ , we will approximate $f, g \in H^{-1}(\Omega)$ by $f_\tau^1 + \text{div} \mathbf{f}_\tau^2, g_\tau^1 + \text{div} \mathbf{g}_\tau^2$, respectively, where $f_\tau^1, g_\tau^1 \in \mathbb{V}_{\tau'}^*, \mathbf{f}_\tau^2, \mathbf{g}_\tau^2 \in [\mathbb{V}_{\tau'}^*]^n$ and either $\tau' = \tau$ or, when it is needed to have a smaller approximation error, $\tau' \supset \tau$. We will set

$$f_{\tau'} := f_\tau^1 + \text{div} \mathbf{f}_\tau^2, \quad g_{\tau'} := g_\tau^1 + \text{div} \mathbf{g}_\tau^2.$$

To be able to distinguish between primal or dual solutions corresponding to different right-hand sides, we introduce operators $L : H_0^1(\Omega) \rightarrow H^{-1}(\Omega)$ by $(Lv)(w) = a(v, w)$ ($v, w \in H_0^1(\Omega)$), and $L_\tau : \mathbb{V}_\tau \rightarrow \mathbb{V}_\tau'$ by $(L_\tau v_\tau)(w_\tau) = a(v_\tau, w_\tau)$ ($v_\tau, w_\tau \in \mathbb{V}_\tau$). The solutions u, z, u_τ, z_τ of (2.1), (2.2), (4.1), (4.2) can now be written as $L^{-1}f, (L')^{-1}g, L_\tau^{-1}f, (L'_\tau)^{-1}g$, respectively. Since in our case $L' = L$ and $L'_\tau = L_\tau$, for notational convenience we will drop the prime. Note that $\|L \cdot\|_{E'} = \|\cdot\|_E, \|L_\tau^{-1}\|_{E' \rightarrow E} \leq 1$, and $\|(L^{-1} - L_\tau^{-1})\|_{E' \rightarrow E} \leq 1$.

Furthermore, in view of controlling the cost of our adaptive solver, from now on we will solve the arising Galerkin systems only approximately.

The following lemma generalizes upon Lemma 5.3, relaxing both the condition that the right-hand side is in $\mathbb{V}_\tau^* + \text{div}[\mathbb{V}_\tau^*]^n$ and the assumption that we have the exact Galerkin solution available, assuming that the deviations from that ideal situation are sufficiently small in a relative sense.

LEMMA 6.1 (see [Ste07, Lemmas 6.1 and 6.2]). *There exist positive constants $\omega = \omega(\theta, C_1, c_2)$ and $\lambda = \lambda(\omega, C_1, c_2)$ such that for any $f \in H^{-1}(\Omega)$, partition τ ,*

$f_\tau^1 \in \mathbb{V}_\tau^*$, $\mathbf{f}_\tau^2 \in [\mathbb{V}_\tau^*]^n$, $\bar{u}_\tau \in \mathbb{V}_\tau$ with

$$(6.1) \quad \|f - f_\tau\|_{E'} + \|L_\tau^{-1} f_\tau - \bar{u}_\tau\|_E \leq \omega \mathcal{E}(\tau, f_\tau^1, \mathbf{f}_\tau^2, \bar{u}_\tau),$$

$F := \mathbf{MARK}[\tau, f_\tau^1, \mathbf{f}_\tau^2, \bar{u}_\tau]$ satisfies

$$\#F \lesssim \#\hat{\tau} - \#\tau_0$$

for any partition $\hat{\tau}$ for which

$$\|u - u_{\hat{\tau}}\|_E \leq \lambda \|u - \bar{u}_\tau\|_E.$$

Furthermore, given a

$$\mu \in \left(\left[1 - \frac{c_2^2 \theta^2}{C_1^2} \right]^{\frac{1}{2}}, 1 \right),$$

there exists an $\omega = \omega(\mu, \theta, C_1, c_2) > 0$, such that if (6.1) is valid for this ω , and for $\tau' \supseteq \mathbf{REFINE}[\tau, F]$, $f_{\tau'} \in H^{-1}(\Omega)$ and $\bar{u}_{\tau'} \in \mathbb{V}_{\tau'}$,

$$\|f - f_{\tau'}\|_{E'} + \|L_{\tau'}^{-1} f_{\tau'} - \bar{u}_{\tau'}\|_E \leq \omega \mathcal{E}(\tau, f_\tau^1, \mathbf{f}_\tau^2, \bar{u}_\tau),$$

then

$$\|u - \bar{u}_{\tau'}\|_E \leq \mu \|u - \bar{u}_\tau\|_E.$$

For solving the Galerkin systems approximately, we assume that we have an iterative solver of optimal type available:

GALSOLVE $[\tau, f_\tau, u_\tau^{(0)}, \delta] \rightarrow \bar{u}_\tau$

% $f_\tau \in (\mathbb{V}_\tau)'$ and $u_\tau^{(0)} \in \mathbb{V}_\tau$, the latter being an initial approximation for an
% iterative solver. The output $\bar{u}_\tau \in \mathbb{V}_\tau$ satisfies

$$\|L_\tau^{-1} f_\tau - \bar{u}_\tau\|_E \leq \delta.$$

% The call requires $\lesssim \max\{1, \log(\delta^{-1} \|L_\tau^{-1} f_\tau - u_\tau^{(0)}\|_E)\} \#\tau$
% arithmetic operations.

Multigrid methods with local smoothing, or their additive variants (Bramble–Pasciak–Xu) as preconditioners in conjugate gradients, are known to be of this type.

A routine called **RHS_f**, and analogously **RHS_g**, will be needed to find a sufficiently accurate approximation to the right-hand side f of the form $f_\tau^1 + \text{div} \mathbf{f}_\tau^2$ with $f_\tau^1 \in \mathbb{V}_\tau^*$, $\mathbf{f}_\tau^2 \in [\mathbb{V}_\tau^*]^n$. Since this might not be possible with respect to the current partition, a call of **RHS_f** may result in further refinement.

RHS_f $[\tau, \delta] \rightarrow [\tau', f_{\tau'}^1, \mathbf{f}_{\tau'}^2]$

% $\delta > 0$. The output consists of $f_{\tau'}^1 \in \mathbb{V}_{\tau'}^*$, and $\mathbf{f}_{\tau'}^2 \in [\mathbb{V}_{\tau'}^*]^n$, where $\tau' = \tau$ or,
% if necessary, $\tau' \supset \tau$, such that $\|f - f_{\tau'}\|_{E'} \leq \delta$.

Assuming that $u \in \mathcal{A}^s$ for some $s > 0$, the cost of approximating the right-hand side f using **RHS_f** will generally not dominate the other costs of our adaptive

method only if there is some constant c_f such that for any $\delta > 0$ and any partition τ , for $[\tau', \cdot, \cdot] := \mathbf{RHS}_f[\tau, \delta]$, it holds that

$$\#\tau' - \#\tau \leq c_f^{1/s} \delta^{-1/s},$$

and the number of arithmetic operations required by the call is $\lesssim \#\tau'$. We will refer to such an \mathbf{RHS}_f as s -optimal with constant c_f . Obviously, given s , such a routine can exist only when $f \in \bar{\mathcal{A}}^s$, defined by

$$\bar{\mathcal{A}}^s = \left\{ f \in H^{-1}(\Omega) : \sup_{\varepsilon > 0} \inf_{\{\tau: \inf_{f_\tau^1 \in \mathbb{V}_\tau^*, f_\tau^2 \in [\mathbb{V}_\tau^*]^n} \|f - f_\tau\|_{E'} \leq \varepsilon\}} [\#\tau - \#\tau_0]^s < \infty \right\}.$$

On the one hand, $u \in \mathcal{A}^s$ implies that $f \in \bar{\mathcal{A}}^s$. Indeed, for any partition τ , let $f_\tau^2 := -\mathbf{A}\nabla u_\tau$. Then $f_\tau^2 \in [\mathbb{V}_\tau^*]^n$ and $\|f - \operatorname{div} f_\tau^2\|_{E'} = \|u - u_\tau\|_E$. On the other hand, knowing that $f \in \bar{\mathcal{A}}^s$ is a different thing than knowing how to construct suitable approximations. If $s \in [\frac{1}{n}, \frac{p+1}{n}]$ and $f \in H^{sn-1}(\Omega)$, then the best approximations f_τ^1 to f from \mathbb{V}_τ^* with respect to $L_2(\Omega)$ using uniform refinements τ of τ_0 are known to converge with the required rate. For general $f \in \bar{\mathcal{A}}^s$, however, a realization of a suitable routine \mathbf{RHS}_f has to depend on the functional f at hand.

Remark 6.2. When u and f are smooth, then $u \in \mathcal{A}^{p/n}$ and $f \in \bar{\mathcal{A}}^{(p+1)/n}$. Indeed, u is approximated by piecewise polynomials of degree p , and f by those of degree $p - 1$ (apart from possible approximations from $\operatorname{div}[\mathbb{V}_\tau^*]^n$), whereas the errors are measured in $H_0^1(\Omega)$ or $H^{-1}(\Omega)$, respectively. Also for less smooth u and f , one can expect that usually $u \in \mathcal{A}^s$ and $f \in \bar{\mathcal{A}}^{s'}$ for some $s' > s$.

In our adaptive method, given some partition τ , for both computing the error estimator and setting up the Galerkin system, we will *replace* f by an approximation from $\mathbb{V}_{\tau'}^* + \operatorname{div}[\mathbb{V}_{\tau'}^*]^n$ where $\tau' \supseteq \tau$ (and similarly for g). This has the advantages that we can consider $f \notin L_2(\Omega) + \operatorname{div}\mathbb{W}_\tau^*$, for which thus the error estimator is not defined, and that we don't have to worry about quadrature errors in various places in the algorithm.

Assuming $f \in L_2(\Omega) + \operatorname{div}\mathbb{W}_\tau^n$ for any τ , another option, followed in [MNS00], is not to replace f by an approximation, but to check whether, on the current partition, the error in the best approximation for f from $\mathbb{V}_\tau^* (+\operatorname{div}[\mathbb{V}_\tau^*]^n)$, called *data oscillation*, is sufficiently small relative to the error in the current approximation to u , and, if not, to refine τ to achieve this. Convergence of this approach was shown, and it can be expected that by applying suitable quadrature and inexact Galerkin solves, optimal computational complexity can be shown as well. The observations at the beginning of this remark indicate that “usually,” at least asymptotically, there will be no refinements needed to reduce the data oscillation. This explains why common adaptive methods that ignore data oscillation usually converge with optimal rates.

In addition to being s -optimal, we will have to assume that \mathbf{RHS}_f is *linearly convergent*, by which we mean that for any $d \in (0, 1)$, there exists a $D > 0$ such that for any $\delta > 0$, partitions τ and $\tau' \supseteq \hat{\tau}$ where $[\hat{\tau}, \cdot, \cdot] := \mathbf{RHS}_f[\tau, \delta]$, the output $[\tau'', \cdot, \cdot] := \mathbf{RHS}_f[\tau', d\delta]$ satisfies $\#\tau'' \leq D\#\tau'$.

Remark 6.3. Usually, a realization of $[\hat{\tau}, \cdot, \cdot] := \mathbf{RHS}_f[\tau, \delta]$ will be based on the selection of $\hat{\tau}$ such that an *upper bound* for the error is less than the prescribed tolerance. Since this upper bound will be an algebraically decreasing function of $\#\hat{\tau} - \#\tau_0$, linear convergence is obtained.

We now have the ingredients in hand to define our practical adaptive goal-oriented finite element routine **GOAFEM**. Compared to the idealized version from the previous section, we will have to deal with the fact that when solving the Galerkin systems

only inexactly, and applying inexact right-hand sides, C_1 times the a posteriori error estimator $\mathcal{E}(\cdot)$ is not necessarily an upper bound for the energy norm of the error. We have to add correction terms to obtain an upper bound. Furthermore, after applying **REFINE** on either the primal or dual side, we have to specify a tolerance for the error in the new approximation of the right-hand side and in that of the new approximate Galerkin solution. In order to know that a subsequent **REFINE** results in an error reduction, in view of Lemma 6.1 we would like to choose this tolerance smaller than ω times the new error estimator, which, however, is not known yet. Although we can expect that usually the new estimator is only some moderate factor less than the existing one, it cannot be excluded that the new estimator is arbitrarily small, e.g., when we happen to have reached a partition on which the solution can be exactly represented. In this case, an error reduction is immediate, and so we don't have to rely on **REFINE** to achieve it.

```

GOAFEM[ $f, g, \delta_p, \delta_d, \varepsilon$ ]  $\rightarrow$  [ $\tau, \bar{u}_\tau, \bar{z}_\tau$ ]
% Let  $\omega \in (0, c_2)$  be a constant not larger than the constants  $\omega(\theta, C_1, c_2)$  and
%  $\omega(\mu, \theta, C_1, c_2)$  for some  $\mu \in ([1 - \frac{c_2^2 \theta^2}{C_1^2}]^{\frac{1}{2}}, 1)$  mentioned in Lemma 6.1.
% Let  $0 < \beta < [(\frac{2+3C_1c_2^{-1}}{2+C_1c_2^{-1}} + C_1c_2^{-1})(2 + C_1(c_2^{-1} + 2\omega^{-1}))]^{-1}$  be a constant.
 $\tau := \tau_0$ , [ $\tau_p, f_{\tau_p}^1, \mathbf{f}_{\tau_p}^2$ ] := RHS $_f[\tau, \delta_p]$ , [ $\tau_d, g_{\tau_d}^1, \mathbf{g}_{\tau_d}^2$ ] := RHS $_g[\tau, \delta_d]$ 
 $\bar{u}_{\tau_p} := \bar{z}_{\tau_d} := 0$ 
do
   $\bar{u}_{\tau_p} :=$  GALSOLVE[ $\tau_p, f_{\tau_p}, \bar{u}_{\tau_p}, \delta_p$ ]
   $\bar{z}_{\tau_d} :=$  GALSOLVE[ $\tau_d, g_{\tau_d}, \bar{z}_{\tau_d}, \delta_d$ ]
   $\sigma_p := (2 + C_1c_2^{-1})\delta_p + C_1\mathcal{E}(\tau_p, f_{\tau_p}^1, \mathbf{f}_{\tau_p}^2, \bar{u}_{\tau_p})$ 
   $\sigma_d := (2 + C_1c_2^{-1})\delta_d + C_1\mathcal{E}(\tau_d, g_{\tau_d}^1, \mathbf{g}_{\tau_d}^2, \bar{z}_{\tau_d})$ 
  if  $\sigma_p\sigma_d \leq \varepsilon$  then  $\tau := \tau_p \cup \tau_d$ ,  $\bar{u}_\tau := \bar{u}_{\tau_p}$ ,  $\bar{z}_\tau := \bar{z}_{\tau_d}$  stop endif
  if  $2\delta_p \leq \omega\mathcal{E}(\tau_p, f_{\tau_p}^1, \mathbf{f}_{\tau_p}^2, \bar{u}_{\tau_p})$  then  $F_p :=$  MARK[ $\tau, f_{\tau_p}^1, \mathbf{f}_{\tau_p}^2, \bar{u}_{\tau_p}$ ]
  else  $F_p := \emptyset$  endif
  if  $2\delta_d \leq \omega\mathcal{E}(\tau_d, g_{\tau_d}^1, \mathbf{g}_{\tau_d}^2, \bar{z}_{\tau_d})$  then  $F_d :=$  MARK[ $\tau, g_{\tau_d}^1, \mathbf{g}_{\tau_d}^2, \bar{z}_{\tau_d}$ ]
  else  $F_d := \emptyset$  endif
  if  $\#\tau_p - \#\tau + \#F_p \leq \#\tau_d - \#\tau + \#F_d$ 
  then  $\tau :=$  REFINE[ $\tau_p, F_p$ ],  $\delta_p := \min(\delta_p, \beta\sigma_p)$ 
    [ $\tau_p, f_{\tau_p}^1, \mathbf{f}_{\tau_p}^2$ ] := RHS $_f[\tau, \delta_p]$ ,  $\tau_d := \tau \cup \tau_d$ 
  else  $\tau :=$  REFINE[ $\tau_d, F_d$ ],  $\delta_d := \min(\delta_d, \beta\sigma_d)$ 
     $\tau_p := \tau \cup \tau_p$ , [ $\tau_d, g_{\tau_d}^1, \mathbf{g}_{\tau_d}^2$ ] := RHS $_g[\tau, \delta_d]$ 
  endif
enddo

```

THEOREM 6.4. [$\tau, \bar{u}_\tau, \bar{z}_\tau$] = **GOAFEM**[$f, g, \underline{\delta}_p, \underline{\delta}_d, \varepsilon$] terminates, and

$$\|u - \bar{u}_\tau\|_E \|z - \bar{z}_\tau\|_E \leq \varepsilon.$$

If $u \in \mathcal{A}^s$, $z \in \mathcal{A}^t$, **RHS** $_f$ (**RHS** $_g$) is s -optimal (t -optimal) with constant c_f (c_g), $\underline{\delta}_p > c_f$, and $\underline{\delta}_d > c_g$, then

$$\#\tau \lesssim \#\tau_0 + \varepsilon^{-1/(s+t)} [(|u|_{\mathcal{A}^s}^{1/s} + c_f^{1/s})^s (|z|_{\mathcal{A}^t}^{1/t} + c_g^{1/t})^t]^{1/(s+t)}.$$

If, additionally, $\|f\|_{E'} \lesssim \underline{\delta}_p$, $\|g\|_{E'} \lesssim \underline{\delta}_d$, and $\underline{\delta}_p \underline{\delta}_d \lesssim \|u - u_{\tau_0}\|_E \|z - z_{\tau_0}\|_E + \varepsilon$, then the number of arithmetic operations and storage locations required by the call

are bounded by some absolute multiple of the same expression. The constant factors involved in these bounds may depend only on τ_0 , and on s or t when they tend to 0 or ∞ , and concerning the cost, on the constants involved in the additional assumptions.

Remark 6.5. The condition $\underline{\delta}_p > c_f$ implies that for a call $[\tau', \cdot, \cdot] = \mathbf{RHS}[\tau, \underline{\delta}_p]$, we have $\tau' = \tau$.

Proof. We start with collecting a few useful estimates. At evaluation of σ_p , by (4.4) and Proposition 4.6, we have

$$\begin{aligned}
 \|u - \bar{u}_{\tau_p}\|_E &\leq \|u - L^{-1}f_{\tau_p}\|_E + \|(L^{-1} - L_{\tau_p}^{-1})f_{\tau_p}\|_E + \|L_{\tau_p}^{-1}f_{\tau_p} - \bar{u}_{\tau_p}\|_E \\
 &\leq \delta_p + C_1\mathcal{E}(\tau_p, f_{\tau_p}^1, \mathbf{f}_{\tau_p}^2, L_{\tau_p}^{-1}f_{\tau_p}) + \|L_{\tau_p}^{-1}f_{\tau_p} - \bar{u}_{\tau_p}\|_E \\
 &\leq \delta_p + C_1\mathcal{E}(\tau_p, f_{\tau_p}^1, \mathbf{f}_{\tau_p}^2, \bar{u}_{\tau_p}) + (C_1c_2^{-1} + 1)\|L_{\tau_p}^{-1}f_{\tau_p} - \bar{u}_{\tau_p}\|_E \\
 (6.2) \quad &\leq (2 + C_1c_2^{-1})\delta_p + C_1\mathcal{E}(\tau_p, f_{\tau_p}^1, \mathbf{f}_{\tau_p}^2, \bar{u}_{\tau_p}) =: \sigma_p
 \end{aligned}$$

and, by Corollary 4.5,

$$\begin{aligned}
 \mathcal{E}(\tau_p, f_{\tau_p}^1, \mathbf{f}_{\tau_p}^2, \bar{u}_{\tau_p}) &\leq c_2^{-1}\|L^{-1}f_{\tau_p} - \bar{u}_{\tau_p}\|_E \\
 &\leq c_2^{-1}[\|u - u_{\tau_p}\|_E + \|(L^{-1} - L_{\tau_p}^{-1})(f_{\tau_p} - f)\|_E + \|L_{\tau_p}^{-1}f_{\tau_p} - \bar{u}_{\tau_p}\|_E] \\
 (6.3) \quad &\leq c_2^{-1}\|u - u_{\tau_p}\|_E + c_2^{-1}2\delta_p.
 \end{aligned}$$

So if $2\delta_p \leq \omega\mathcal{E}(\tau_p, f_{\tau_p}^1, \mathbf{f}_{\tau_p}^2, \bar{u}_{\tau_p})$, then $\mathcal{E}(\tau_p, f_{\tau_p}^1, \mathbf{f}_{\tau_p}^2, \bar{u}_{\tau_p}) \leq [c_2 - \omega]^{-1}\|u - u_{\tau_p}\|_E$, and so

$$(6.4) \quad \sigma_p \leq D\|u - u_{\tau_p}\|_E, \quad \text{where } D := \frac{(1+\frac{1}{2}C_1c_2^{-1})\omega+C_1}{c_2-\omega}.$$

Now we are ready to show majorized linear convergence of $\sigma_p\sigma_d$. Consider any two instances $\sigma_p^{(A)}$ and $\sigma_p^{(B)}$ of σ_p , where $\sigma_p^{(A)}$ has been computed preceding $\sigma_p^{(B)}$. With $\delta_p^{(A)}$, $\delta_p^{(B)}$ and $\tau_p^{(A)}$, $\tau_p^{(B)}$ being the corresponding tolerances and partitions, from (6.3), $\delta_p^{(B)} \leq \delta_p^{(A)}$ and $\tau_p^{(B)} \supseteq \tau_p^{(A)}$, and so $\|u - u_{\tau_p^{(B)}}\|_E \leq \|u - \bar{u}_{\tau_p^{(A)}}\|_E \leq \sigma_p^{(A)}$ by (6.2), and we have

$$\begin{aligned}
 \sigma_p^{(B)} &= (2 + C_1c_2^{-1})\delta_p^{(B)} + C_1\mathcal{E}(\tau_p^{(B)}, f_{\tau_p^{(B)}}^1, \mathbf{f}_{\tau_p^{(B)}}^2, \bar{u}_{\tau_p^{(B)}}) \\
 &\leq (2 + 3C_1c_2^{-1})\delta_p^{(A)} + C_1c_2^{-1}\sigma_p^{(A)} \\
 (6.5) \quad &\leq K\sigma_p^{(A)}, \quad \text{where } K := \frac{2+3C_1c_2^{-1}}{2+C_1c_2^{-1}} + C_1c_2^{-1}.
 \end{aligned}$$

Let us denote by $\tau_p^{(i)}$, $\delta_p^{(i)}$, $f_{\tau_p^{(i)}}^1$, $\mathbf{f}_{\tau_p^{(i)}}^2$, $\bar{u}_{\tau_p^{(i)}}$, $\sigma_p^{(i)}$ the instances of τ_p , δ_p , $f_{\tau_p}^1$, $\mathbf{f}_{\tau_p}^2$, \bar{u}_{τ_p} , σ_p at the moment of the i th call of $\mathbf{REFINE}[\tau_p, F_p]$. If $2\delta_p^{(i)} > \omega\mathcal{E}(\tau_p^{(i)}, f_{\tau_p^{(i)}}^1, \mathbf{f}_{\tau_p^{(i)}}^2, \bar{u}_{\tau_p^{(i)}})$, then for any $k < i$,

$$\sigma_p^{(i)} < (2 + C_1(c_2^{-1} + 2\omega^{-1}))\delta_p^{(i)} \leq (2 + C_1(c_2^{-1} + 2\omega^{-1}))\beta\sigma_p^{(k)}.$$

If, for some $k \in \mathbb{N}_0$, $2\delta_p^{(j)} \leq \omega\mathcal{E}(\tau_p^{(j)}, f_{\tau_p^{(j)}}^1, \mathbf{f}_{\tau_p^{(j)}}^2, \bar{u}_{\tau_p^{(j)}})$ for $j = i, \dots, i - k$, then by (6.4), Lemma 6.1, where we use that $\delta_p^{(j)} \leq \delta_p^{(j-1)}$, and (6.2),

$$\sigma_p^{(i)} \leq D\|u - \bar{u}_{\tau_p^{(i)}}\|_E \leq D\mu^k\|u - \bar{u}_{\tau_p^{(i-k)}}\|_E \leq D\mu^k\sigma_p^{(i-k)}.$$

Since $(2 + C_1(c_2^{-1} + 2\omega^{-1}))\beta < 1/K$ by definition of β , from (6.5) we conclude that for any $\alpha \in (0, 1)$ there exists an M such that $\sigma_p^{(i+M)} \leq \alpha\sigma_p^{(i)}$. Since all results derived so far are equally valid on the dual side, by taking $\alpha < 1/K$ we infer that by $2M$ iterations of the loop inside **GOAFEM**, the product $\sigma_p\sigma_d$ is reduced by a factor $\alpha K < 1$. Indeed, either σ_p or σ_p is reduced by a factor α , whereas the other cannot increase by a factor larger than K .

Next, we bound the cardinality of the output partition. If **GOAFEM** terminates as a result of the first evaluation of the test $\sigma_p\sigma_d \leq \varepsilon$, then by the assumptions that $\underline{\delta}_p > c_f$ and $\underline{\delta}_d > c_g$, the output partition $\tau_p \cup \tau_d = \tau_0$. In the following, we consider the case that initially $\sigma_p\sigma_d > \varepsilon$.

At evaluation of the test $\#\tau_p - \#\tau + \#F_p \leq \#\tau_d - \#\tau + \#F_d$, we have

$$(6.6) \quad \#\tau_p - \#\tau \leq (\beta K^{-1}\sigma_p)^{-1/s} c_f^{1/s}.$$

Indeed, the current $\#\tau_p - \#\tau$ is not larger than this difference at the moment of the most recent call of $\mathbf{RHS}_f[\tau, \delta_p]$. By the assumption of \mathbf{RHS}_f being s -optimal, the latter difference was zero when at that time $\delta_p > c_f$. Otherwise, since $\underline{\delta}_p > c_f$ by assumption, this δ_p was equal to β times the minimum of all values attained by σ_p up to that moment. Using (6.5) and the fact that \mathbf{RHS}_f is s -optimal with constant c_f , we end up with (6.6).

If, at evaluation of the test $\#\tau_p - \#\tau + \#F_p \leq \#\tau_d - \#\tau + \#F_d$, $F_p \neq \emptyset$, i.e., if in the preceding lines $2\delta_p \leq \omega\mathcal{E}(\tau_p, f_{\tau_p}^1, \mathbf{f}_{\tau_p}^2, \bar{u}_{\tau_p})$ and $F_p := \mathbf{MARK}[\tau, f_{\tau_p}^1, \mathbf{f}_{\tau_p}^2, \bar{u}_{\tau_p}]$, an application of Lemma 6.1 and the assumption that $u \in \mathcal{A}^s$ show that then

$$(6.7) \quad \#F_p \lesssim \|u - \bar{u}_{\tau_p}\|_E^{-1/s} |u|_{\mathcal{A}^s}^{1/s} \lesssim \sigma_p^{-1/s} |u|_{\mathcal{A}^s}^{1/s}$$

by (6.4).

Clearly, results analogous to (6.6) and (6.7) are valid on the dual side. Now with $\sigma_{p,j}, \sigma_{d,j}$ being the instances of σ_p, σ_d at the j th evaluation of the test $\#\tau_p - \#\tau + \#F_p \leq \#\tau_d - \#\tau + \#F_d$, with n being the last one, an application of Theorem 3.1 shows that for τ being the output of the call of **REFINE** following this last test, being thus the last call of **REFINE**, we have

$$(6.8) \quad \begin{aligned} \#\tau - \#\tau_0 &\lesssim \sum_{j=1}^n \min\{\sigma_{p,j}^{-1/s} (|u|_{\mathcal{A}^s}^{1/s} + c_f^{1/s}), \sigma_{d,j}^{-1/t} (|z|_{\mathcal{A}^t}^{1/t} + c_g^{1/t})\} \\ &\leq \sum_{j=1}^n (\sigma_{p,j}\sigma_{d,j})^{-1/(s+t)} [(|u|_{\mathcal{A}^s}^{1/s} + c_f^{1/s})^s (|z|_{\mathcal{A}^t}^{1/t} + c_g^{1/t})^t]^{1/(s+t)} \\ &\lesssim \varepsilon^{-1/(s+t)} [(|u|_{\mathcal{A}^s}^{1/s} + c_f^{1/s})^s (|z|_{\mathcal{A}^t}^{1/t} + c_g^{1/t})^t]^{1/(s+t)} \end{aligned}$$

by the majorized linear convergence of $(\sigma_{p,j}\sigma_{d,j})_j$ and $\sigma_{p,n}\sigma_{d,n} > \varepsilon$.

Suppose that this last call of **REFINE** took place on the primal side. Then the output partition of **GOAFEM** is $\tau_p \cup \tau_d$, where $[\tau_p, \cdot, \cdot] := \mathbf{RHS}_f[\tau, \delta_p]$ and $\tau_d := \tau \cup \tau_d$. As we have seen, if $\delta_p \leq c_f$, i.e., if possibly $\tau_p \supsetneq \tau$, then δ_p is larger than βK^{-1} times the current σ_p , which, by its definition, is larger than $2 + C_1 c_2^{-1}$ times the previous value of δ_p , denoted as $\delta_p^{(\text{prev})}$. A call of $\mathbf{RHS}_f[\cdot, \delta_p^{(\text{prev})}]$ has been made inside **GOAFEM**, and so $\tau \supseteq \tau'$ with $[\tau', \cdot, \cdot] := \mathbf{RHS}_f[\cdot, \delta_p^{(\text{prev})}]$. The assumption of \mathbf{RHS}_f being linearly convergent shows that $\#\tau_p \lesssim \#\tau$.

The current $\#\tau_d - \#\tau$ is not larger than this difference at the moment of the last call of **RHS** $_g$, and so analogously we find that $\#\tau_d \lesssim \#\tau$. We conclude that

$$(6.9) \quad \#\tau_p \cup \tau_d \lesssim \#\tau \lesssim \#\tau_0 + \varepsilon^{-1/(s+t)} [(|u|_{\mathcal{A}^s}^{1/s} + c_f^{1/s})^s (|z|_{\mathcal{A}^t}^{1/t} + c_g^{1/t})^t]^{1/(s+t)}.$$

Finally, we have to bound the cost of the algorithm. At the moment of the first call of **GALSOLVE** $[\tau_p, f_{\tau_p}, \bar{u}_{\tau_p}, \delta_p]$, we have

$$\|L_{\tau_p}^{-1} f_{\tau_p} - \bar{u}_{\tau_p}\|_E \leq \|f_{\tau_p} - f\|_{E'} + \|f\|_{E'} \leq \delta_p + \|f\|_{E'} \lesssim \delta_p$$

by assumption. We now consider any further calls. From (6.3), $\|u - u_{\tau_0}\|_E \leq \|f\|_{E'} \lesssim \underline{\delta}_p$ by assumption, and (6.5), we have that the currents δ_p and σ_p at the moment of such a call satisfy $\sigma_p \lesssim \delta_p$. As a consequence, we have

$$\begin{aligned} \|L_{\tau_p}^{-1} f_{\tau_p} - \bar{u}_{\tau_p}\|_E &\leq \|(L^{-1} - L_{\tau_p}^{-1})f_{\tau_p}\|_E + \|L^{-1} f_{\tau_p} - \bar{u}_{\tau_p}\|_E \leq 2\|L^{-1} f_{\tau_p} - \bar{u}_{\tau_p}\|_E \\ &\leq 2[\|f - f_{\tau_p}\|_{E'} + \|u - \bar{u}_{\tau_p}\|_E] \leq 2\delta_p + 2\sigma_p \lesssim \delta_p. \end{aligned}$$

By the assumption of **GALSOLVE** being an optimal iterative solver, we conclude that the cost of these calls is $\mathcal{O}(\#\tau_p)$.

The number of arithmetic operations needed for the calls **MARK** $[\tau, f_{\tau_p}^1, f_{\tau_p}^2, \bar{u}_{\tau_p}]$, $\tau := \mathbf{REFINE}[\tau_p, F_p]$, and $[\tau_p, \cdot, \cdot] := \mathbf{RHS}_f[\tau, \delta_p]$ are $\mathcal{O}(\#\tau)$, $\mathcal{O}(\#\tau)$, and $\mathcal{O}(\#\tau_p)$, respectively. Moreover, we know that $\#\tau_p \lesssim \#\tau$, and that $\#\tau - \#\tau_0$ as a function of the iteration count is majorized by a linearly increasing sequence with upper bound (6.8). From the assumption that $\underline{\delta}_p \underline{\delta}_d \lesssim \|u - u_{\tau_0}\|_E \|z - z_{\tau_0}\|_E + \varepsilon$, the first $\sigma_p \sigma_d \lesssim \|u - u_{\tau_0}\|_E \|z - z_{\tau_0}\|_E + \varepsilon$, meaning that after some absolute constant number of iterations, either the current τ is unequal to τ_0 or the algorithm has terminated. Together, above observations show that the total cost is bounded by some absolute multiple of the right-hand side of (6.9). \square

Remark 6.6. The functions $\bar{u}_\tau, \bar{z}_\tau$ produced by **GOAFEM** are not the exact Galerkin approximations, and so $\|u - \bar{u}_\tau\|_E \|z - \bar{z}_\tau\|_E$ is not necessarily an upper bound for $|g(u) - g(\bar{u}_\tau)|$. Writing

$$g(u) - g(\bar{u}_\tau) = a(u - \bar{u}_\tau, z) = a(u - \bar{u}_\tau, z - z_\tau) = a(u - \bar{u}_\tau, z - \bar{z}_\tau) - a(u - \bar{u}_\tau, z_\tau - \bar{z}_\tau),$$

and using the fact that $\|u - \bar{u}_\tau\|_E \leq \sigma_p$, $\|z - \bar{z}_\tau\|_E \leq \sigma_d$, $\|z_\tau - \bar{z}_\tau\| \leq \delta_d \leq (2 + C_1 c_2^{-1})^{-1} \sigma_d$, and $\sigma_p \sigma_d \leq \varepsilon$, we end up with $|g(u) - g(\bar{u}_\tau)| \leq [1 + (2 + C_1 c_2^{-1})^{-1}] \varepsilon$.

7. Numerical experiments. In this section we will consider the performance of the **GOAFEM** routine in practice. As many real-world problems require the evaluation of functionals that are unbounded on $H_0^1(\Omega)$, we will also consider such a problem. As **GOAFEM** can handle only bounded functionals, we need to do some additional work. Following [BS01], we will apply a so-called *extraction functional*, a technique that we recall below. An alternative approach would be to apply a regularized functional as suggested in [OR76, BR96]. This approach can be applied more generally since no Green's function is needed. On the other hand, it introduces an additional error that can only be controlled in terms of higher order derivatives of the solution beyond those that are needed for the functional to be well defined.

7.1. Extraction functionals. Let \tilde{g} be some functional defined on the solution u of (2.1), but that is unbounded on $H_0^1(\Omega)$. With f being the right-hand side of (2.1), we write $\tilde{g}(u) = g(u) + M(f)$, where $g \in H^{-1}(\Omega)$ and M is a functional on

f . Since u and f are related via an invertible operator, this is always possible, even for any $g \in H^{-1}(\Omega)$. Yet, we would like to do this under the additional constraint that $M(f)$ can be computed within any given tolerance at low cost. Basically, this additional condition requires that a Green’s function for the differential operator is available.

We consider $\mathbf{A} = \text{Id}$, i.e., the Poisson problem, on a two-dimensional domain Ω , and, for some $\bar{x} \in \Omega$, $\tilde{g} = \tilde{g}_{\bar{x}}$ given by

$$\tilde{g}_{\bar{x}}(u) = \frac{\partial u}{\partial x_1}(\bar{x}),$$

assuming that u is sufficiently smooth. With (r, θ) denoting polar coordinates centered at \bar{x} , we have $\Delta \frac{\log r}{2\pi} = \delta_{\bar{x}}$, and so $-\Delta \frac{\cos \theta}{2\pi r} = \tilde{g}_{\bar{x}}$ in the sense that for any smooth test function $\phi \in \mathcal{D}(\mathbb{R}^2)$, $-\int_{\mathbb{R}^2} \frac{\cos \theta}{2\pi r} \Delta \phi = \tilde{g}_{\bar{x}}(\phi)$. Generally, this formula cannot be applied with ϕ replaced by the solution u of (2.1). Indeed, in the general case this function has a nonvanishing normal derivative at the boundary of Ω , and therefore its zero extension is not sufficiently smooth. Therefore, with $w_0^{\bar{x}} := \frac{\cos \theta}{2\pi r}$, $w_1^{\bar{x}}$ being a sufficiently smooth function equal to $w_0^{\bar{x}}$ outside some open $\Sigma \Subset \Omega$ that contains \bar{x} , and $w^{\bar{x}} := w_0^{\bar{x}} - w_1^{\bar{x}}$ for any $\phi \in \mathcal{D}(\mathbb{R}^2)$, we write

$$\begin{aligned} \tilde{g}_{\bar{x}}(\phi) &= - \int_{\mathbb{R}^2} w_1^{\bar{x}} \Delta \phi - \int_{\mathbb{R}^2} w^{\bar{x}} \Delta \phi \\ &= \int_{\mathbb{R}^2} \Delta(-w_1^{\bar{x}}) \phi + \int_{\Omega} w^{\bar{x}} (-\Delta \phi) \\ &=: g_{\bar{x}}(\phi) + M_{\bar{x}}(-\Delta \phi). \end{aligned}$$

Clearly, $g_{\bar{x}}$ extends to a bounded functional on $L_1(\mathbb{R}^2)$, with $g_{\bar{x}}(v) = \int_{\Omega} \Delta(-w_1^{\bar{x}})v$ when $\text{supp } v \subset \Omega$. In particular, $g_{\bar{x}}$ is bounded on $H_0^1(\Omega)$, which enables us to use **GOAFEM** to evaluate it. Moreover, since $\text{supp } w^{\bar{x}} \Subset \Omega$, under some mild conditions the above reformulation can be shown to be applicable to u . The details are as follows.

PROPOSITION 7.1. *If*

- (a) $f \in L_2(\Omega)$,
- (b) u is continuously differentiable at \bar{x} , and
- (c) in a neighborhood of \bar{x} , f is in L^p for some $p > 2$,

then

$$\tilde{g}_{\bar{x}}(u) = g_{\bar{x}}(u) + M_{\bar{x}}(f).$$

Proof. Let $B(\bar{x}; \varepsilon)$ be the ball centered at \bar{x} with radius ε , and small enough such that $B(\bar{x}; \varepsilon) \Subset \Omega$. Since $u, w^{\bar{x}} \in H^1(\Omega \setminus B(\bar{x}; \varepsilon))$, $\Delta u \in L_2(\Omega \setminus B(\bar{x}; \varepsilon))$ by (a), $\Delta w^{\bar{x}} \in L_2(\Omega \setminus B(\bar{x}; \varepsilon))$, and $\text{supp } w^{\bar{x}} \Subset \Omega$, integration by parts shows that

$$(7.1) \quad \int_{\partial B(\bar{x}; \varepsilon)} w^{\bar{x}} \frac{\partial u}{\partial \mathbf{n}} - u \frac{\partial w^{\bar{x}}}{\partial \mathbf{n}} = \int_{\Omega \setminus B(\bar{x}; \varepsilon)} u \Delta w^{\bar{x}} - w^{\bar{x}} \Delta u,$$

where \mathbf{n} is the outward pointing normal of $\partial B(\bar{x}; \varepsilon)$.

We have $\lim_{\varepsilon \downarrow 0} \int_{\Omega \setminus B(\bar{x}; \varepsilon)} u \Delta w^{\bar{x}} = -\lim_{\varepsilon \downarrow 0} \int_{\Omega \setminus B(\bar{x}; \varepsilon)} u \Delta w_1^{\bar{x}} = g_{\bar{x}}(u)$.

Since $|\int_{B(\bar{x}; \varepsilon)} w_0^{\bar{x}} f| \leq \|f\|_{L_p(B(\bar{x}; \varepsilon))} \|w_0^{\bar{x}}\|_{L_q(B(\bar{x}; \varepsilon))} (\frac{1}{p} + \frac{1}{q} = 1)$, and furthermore $\|w_0^{\bar{x}}\|_{L_q(B(\bar{x}; \varepsilon))} = [\int_0^\varepsilon \int_0^{2\pi} |\frac{\cos \theta}{2\pi r}|^q r]^{1/q} \rightarrow 0$ when $\varepsilon \downarrow 0$ and $q < 2$, from (c) we conclude that $-\lim_{\varepsilon \downarrow 0} \int_{\Omega \setminus B(\bar{x}; \varepsilon)} w^{\bar{x}} \Delta u = \int_{\Omega} w^{\bar{x}} f = M_{\bar{x}}(f)$.

The contributions of $w_1^{\bar{x}}$ to the left-hand side of (7.1) vanish when $\varepsilon \downarrow 0$.

From $\int_{\partial B(\bar{x};\varepsilon)} w_0^{\bar{x}} \frac{\partial u}{\partial \mathbf{n}} = \int_0^{2\pi} (\cos\theta \frac{\partial u}{\partial x_1} + \sin\theta \frac{\partial u}{\partial x_2}) \frac{\cos\theta}{2\pi\varepsilon} \varepsilon d\theta$ and (b), we infer that $\lim_{\varepsilon \downarrow 0} \int_{\partial B(\bar{x};\varepsilon)} w_0^{\bar{x}} \frac{\partial u}{\partial \mathbf{n}} = \frac{1}{2} \frac{\partial u}{\partial x_1}(\bar{x})$.

From

$$\begin{aligned} \int_{\partial B(\bar{x};\varepsilon)} u \frac{\partial w_0^{\bar{x}}}{\partial \mathbf{n}} &= \frac{-1}{2\pi\varepsilon} \int_0^{2\pi} \cos\theta u d\theta = \frac{1}{2\pi\varepsilon} \int_0^{2\pi} \sin\theta \frac{\partial u}{\partial \theta} d\theta \\ &= \frac{1}{2\pi} \int_0^{2\pi} \sin\theta \left(-\sin\theta \frac{\partial u}{\partial x_1} + \cos\theta \frac{\partial u}{\partial x_2}\right) d\theta \end{aligned}$$

and (b), we infer that $-\lim_{\varepsilon \downarrow 0} \int_{\partial B(\bar{x};\varepsilon)} w_0^{\bar{x}} \frac{\partial w_0^{\bar{x}}}{\partial \mathbf{n}} = \frac{1}{2} \frac{\partial u}{\partial x_1}(\bar{x})$. Together, the above observations give the proof. \square

7.2. Implementation. The implementation of the **GOAFEM** routine is essentially as described above, with the sole difference that we did not approximate the right-hand sides for setting up the Galerkin systems and computing the a posteriori error estimators, but instead used quadrature directly. This was possible, and in view of Remark 6.2 reasonable, because in our experiments either the right-hand sides are very smooth or they are already in $\mathbb{V}_{\tau_0}^* + \text{div}[\mathbb{V}_{\tau_0}^*]^n$.

For all experiments, we used $p = 2$, i.e., quadratic Lagrange elements.

The **GALSOLVE** routine we use solves the linear systems with the conjugate gradient method using the well-known Bramble–Pasciak–Xu preconditioner.

All routines were implemented in Common Lisp and run using the SBCL compiler and run-time environment. This allowed for a short development time and well-instrumented code. With regards to efficiency, the only effort made in that direction consisted in making sure that the asymptotics were correct. While an efficient implementation would be possible with moderate effort (see [Neu03]), for our purposes convenience and correctness were the most important considerations.

For the experiment in which we use the extraction functional for the partial derivative at a point introduced above, we also have to solve a quadrature problem. For this we used the adaptive cubature routine **Cuhre** [BEG91] as implemented in the **Cuba** cubature package [Hah05].

7.3. Experiments. To test **GOAFEM**, we chose two distinct situations. For the first example, we want to compute a partial derivative at a point of a function given as the solution of a Poisson problem, thus illustrating the applicability of our method to this situation.

In our second example, we consider a problem in which the singularities of the solutions to the primal and dual problems are spatially separated.

Example 7.2. Let $\Omega = (0, 1)^2$. We consider problem (2.1), choosing the right-hand side $f = 1$ (i.e., $f(v) = \int_{\Omega} v dx$). We will test the performance of **GOAFEM** on the task of computing

$$\frac{\partial u}{\partial x_1}(\bar{x}),$$

with $\bar{x} = (\frac{\pi}{7}, \frac{49}{100})$. The initial partition is as indicated in Figure 7.1, with $(\frac{1}{2}, \frac{1}{2})$ being the newest vertex of all 4 triangles.

Following the discussion from subsection 7.1, we take $w_1^{\bar{x}} = \psi w_0^{\bar{x}}$, and thus $w^{\bar{x}} = (1 - \psi)w_0^{\bar{x}}$, with ψ being a sufficiently smooth function, 1 outside some neighborhood

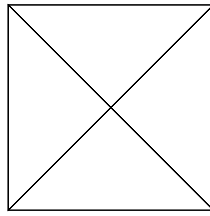
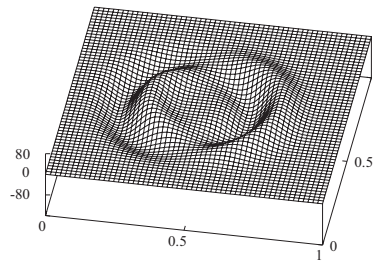
FIG. 7.1. Initial partition τ_0 corresponding to Example 7.2.

FIG. 7.2. Right-hand side of the dual problem corresponding to Example 7.2.

of \bar{x} inside Ω , and 0 on some smaller neighborhood of \bar{x} . Proposition 7.1 shows that $\frac{\partial u}{\partial x_1}(\bar{x}) = \int_{\Omega} u \Delta(-\psi w_0^{\bar{x}}) + \int_{\Omega} (1 - \psi) w_0^{\bar{x}} f$. Writing (θ, r) for the polar coordinates around \bar{x} , we chose

$$(7.2) \quad \psi(\theta, r) := \int_0^r \psi^*(s) ds / \int_0^\infty \psi^*(s) ds,$$

with ψ^* a spline function of order 6, with support $[0.1, 0.45]$.

We evaluated $\int_{\Omega} (1 - \psi) w_0^{\bar{x}} f$ using the adaptive quadrature routine **Cuhre**. To obtain precision of 10^{-12} it needed 216515 integrand evaluations. On current off-the-shelf hardware, it takes only a few seconds.

To approximate $\int_{\Omega} u \Delta(-\psi w_0^{\bar{x}})$ we used **GOAFEM**. Since the right-hand sides 1 and $\Delta(-\psi w_0^{\bar{x}})$ of primal and dual problems are smooth, their solutions are in $\mathcal{A}^{p/n} = \mathcal{A}^1$, so that the error in the functional is $\mathcal{O}([\#\tau - \#\tau_0]^{-2})$. We compared the results with those obtained with the corresponding non-goal-oriented adaptive finite element routine **AFEM** for minimizing the error in energy norm, which is obtained by applying refinements always because of the markings at primal side.

The solutions of the primal and dual problems are in $H^{3-\varepsilon}(\Omega)$ for any $\varepsilon > 0$, but, because the right-hand sides do not vanish at the corners, they are not in $H^3(\Omega)$. Recalling that we use quadratic elements, as a consequence (fully) optimal convergence rates with respect to $\|\cdot\|_E$ are not obtained using uniform refinements. On the other hand, since the (weak) singularities in the primal and dual solutions are solely caused by the shape of the domain, the same local refinements near the corners are appropriate for both primal and dual problem. Therefore, in view of (1.1), we may expect that also with **AFEM** the error in the functional is $\mathcal{O}([\#\tau - \#\tau_0]^{-2})$. On the other hand, since quantitatively the right-hand side, and so the solution of the dual problem, are not that smooth (see Figure 7.2), we may hope that the application of **GOAFEM** yields quantitatively better results.

In Figure 7.3, we show errors in $\int_{\Omega} u \Delta(-\psi w_0^{\bar{x}})$ as a function of $\#\tau - \#\tau_0$. The re-

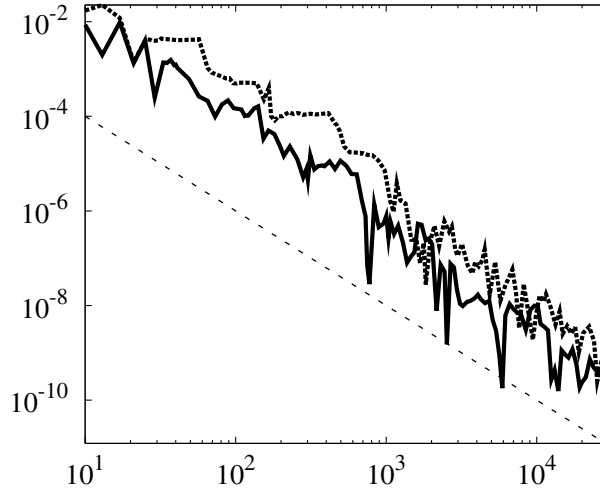


FIG. 7.3. Error in the functional vs. $\#\tau - \#\tau_0$ using **GOAFEM** (solid) and **AFEM** (dashed) corresponding to Example 7.2, and a curve $C[\#\tau - \#\tau_0]^{-2}$.

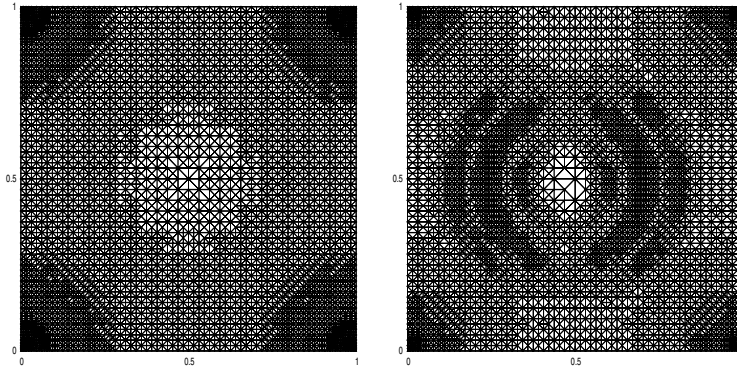


FIG. 7.4. Partitions produced by **AFEM** and **GOAFEM** with nearly equal number of triangles for Example 7.2.

sults confirm that for both **GOAFEM** and **AFEM**, these errors are $\mathcal{O}([\#\tau - \#\tau_0]^{-2})$, where on average for **GOAFEM** the errors are smaller. In Figure 7.4, we show partitions produced by **GOAFEM** and **AFEM**. With **AFEM** local refinements are made only towards the corners, whereas with **GOAFEM** additional local refinements are made in areas where quantitatively the dual solution is nonsmooth due to oscillations in its right-hand side.

Example 7.3. As in Example 7.2, we consider Poisson’s problem on the unit square. We now take as initial partition the one that is obtained from the partition from Figure 7.1 by 2 uniform refinements. We define the right-hand sides f and g of primal and dual problems by

$$(7.3) \quad f(v) = - \int_{T_f} \frac{\partial v}{\partial x_1}, \quad g(v) = - \int_{T_g} \frac{\partial v}{\partial x_1},$$

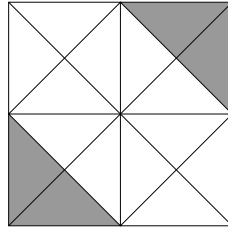


FIG. 7.5. Initial partition τ_0 corresponding to Example 7.3, and T_f (left bottom), T_g (right top).

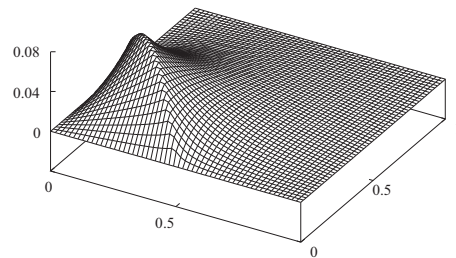


FIG. 7.6. Primal solution corresponding to Example 7.3.

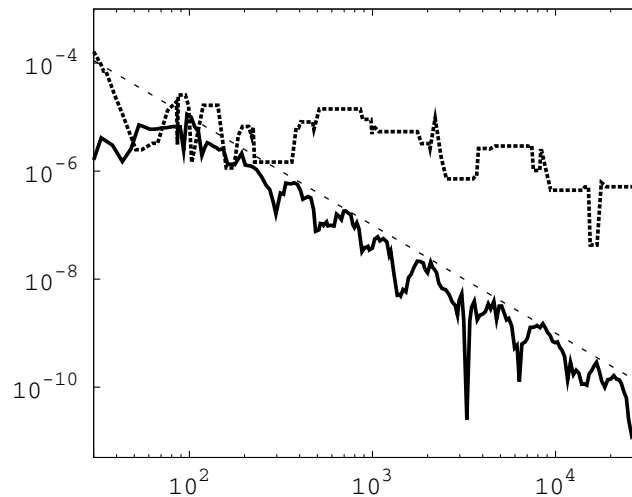


FIG. 7.7. Error in the functional vs. $\#\tau - \#\tau_0$ using **GOAFEM** (solid) and **AFEM** (dashed) corresponding to Example 7.3, and a curve $C[\#\tau - \#\tau_0]^{-2}$.

where T_f and T_g are the simplices $\{(0, 0), (\frac{1}{2}, 0), (0, \frac{1}{2})\}$ and $\{(1, 1), (\frac{1}{2}, 1), (1, \frac{1}{2})\}$, respectively; see Figure 7.5. That is, with χ_f being the characteristic function of T_f , $f = \text{div}[\chi_f \ 0]^T$. So in view of (4.3), here we write f as $f^1 + \text{div} \mathbf{f}^2$ with vanishing f^1 , and benefit from the fact that $\mathbf{f}^2 \in [V_{\tau_0}^*]^2$. Similarly for g .

The primal solution has a singularity along the line connecting the points $(\frac{1}{2}, 0)$ and $(0, \frac{1}{2})$ (see Figure 7.6), and similarly the dual solution has one along the line connecting $(1, \frac{1}{2})$ and $(\frac{1}{2}, 1)$. Since the non-goal-oriented adaptive finite element routine **AFEM** does not see the latter singularity, it behaves much worse than **GOAFEM**, as seen in Figure 7.7. For **GOAFEM** we observe an error $\mathcal{O}([\#\tau - \#\tau_0]^{-2})$, which, since

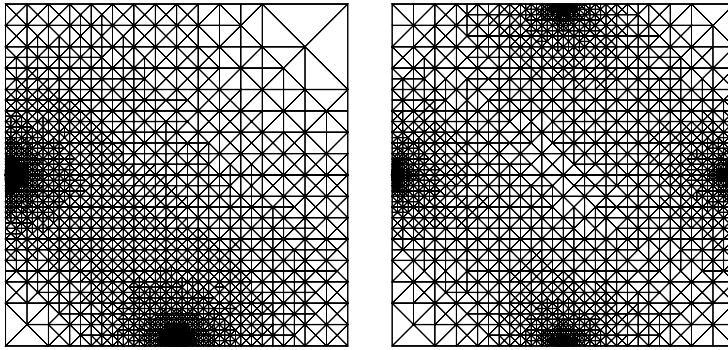


FIG. 7.8. Partitions produced by **AFEM** and **GOAFEM** with nearly equal number of triangles for Example 7.3.

$p/n = 1$, is equal to the best possible rate predicted by Theorem 6.4. In Figure 7.8, we show partitions produced by **AFEM** and **GOAFEM**, respectively.

REFERENCES

- [AO00] M. AINSWORTH AND J. T. ODEN, *A Posteriori Error Estimation in Finite Element Analysis*, Pure Appl. Math., Wiley-Interscience, New York, 2000.
- [BR96] R. BECKER AND R. RANNACHER, *A feed-back approach to error control in finite element methods: Basic analysis and examples*, East-West J. Numer. Math., 4 (1996), pp. 237–264.
- [BDD04] P. BINEV, W. DAHMEN, AND R. DEVORE, *Adaptive finite element methods with convergence rates*, Numer. Math., 97 (2004), pp. 219–268.
- [BDDP02] P. BINEV, W. DAHMEN, R. DEVORE, AND P. PETRUCHEV, *Approximation classes for adaptive methods*, Serdica Math. J., 28 (2002), pp. 391–416.
- [BEG91] J. BERNSTEN, T. O. ESPELID, AND A. GENZ, *An adaptive algorithm for the approximate calculation of multiple integrals*, ACM Trans. Math. Software, 17 (1991), pp. 437–451.
- [BMN02] E. BÄNSCH, P. MORIN, AND R. H. NOCHETTO, *An adaptive Uzawa FEM for the Stokes problem: Convergence without the inf-sup condition*, SIAM J. Numer. Anal., 40 (2002), pp. 1207–1229.
- [BR03] W. BANGERTH AND R. RANNACHER, *Adaptive Finite Element Methods for Differential Equations*, Lectures Math. ETH Zürich, Birkhäuser Verlag, Basel, 2003.
- [BS01] I. BABUŠKA AND T. STROUBOULIS, *The Finite Element Method and Its Reliability*, Numer. Math. Sci. Comput., The Clarendon Press, Oxford University Press, New York, 2001.
- [DKV06] W. DAHMEN, A. KUNOTH, AND J. VORLOEPER, *Convergence of adaptive wavelet methods for goal-oriented error estimation*, in Numerical Mathematics and Advanced Applications, Springer-Verlag, Berlin, 2006, pp. 39–61.
- [Dör96] W. DÖRFLER, *A convergent adaptive algorithm for Poisson’s equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.
- [Hah05] T. HAHN, *Cuba—a library for multidimensional numerical integration*, Comput. Phys. Comm., 168 (2005), pp. 78–95.
- [KS08] Y. KONDRATYUK AND R. P. STEVENSON, *An optimal adaptive finite element method for the Stokes problem*, SIAM J. Numer. Anal., 46 (2008), pp. 747–775.
- [Mau95] J. M. MAUBACH, *Local bisection refinement for n -simplicial grids generated by reflection*, SIAM J. Sci. Comput., 16 (1994), pp. 210–227.
- [MNS00] P. MORIN, R. NOCHETTO, AND K. SIEBERT, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.
- [MS08] M. MOMMER AND R. P. STEVENSON, *A Goal-Oriented Adaptive Finite Element Method with Convergence Rates—Extended Version*, preprint, Korteweg–de Vries Institute for Mathematics, University of Amsterdam, 2008; extended preprint version of current work on <http://staff.science.uva.nl/~rstevens/publ.html>.

- [MvSST06] K.-S. MOON, E. VON SCHWERIN, A. SZEPESSY, AND R. TEMPONE, *Convergence rates for an adaptive dual weighted residual finite element algorithm*, BIT, 46 (2006), pp. 367–407.
- [Neu03] N. NEUSS, *On using Common Lisp in scientific computing*, in Challenges in Scientific Computing–CISC 2002, Springer-Verlag, Berlin, 2003, pp. 237–245.
- [OR76] J. T. ODEN AND J. N. REDDY, *An Introduction to the Mathematical Theory of Finite Elements*, Pure Appl. Math., Wiley-Interscience, New York, 1976.
- [Ste05] R. P. STEVENSON, *An optimal adaptive finite element method*, SIAM J. Numer. Anal., 42 (2005), pp. 2188–2217.
- [Ste07] R. P. STEVENSON, *Optimality of a standard adaptive finite element method*, Found. Comput. Math., 7 (2007), pp. 245–269.
- [Ste08] R. P. STEVENSON, *The completion of locally refined simplicial partitions created by bisection*, Math. Comp., 77 (2008), pp. 227–241.
- [Tra97] C. T. TRAXLER, *An algorithm for adaptive mesh refinement in n dimensions*, Computing, 59 (1997), pp. 115–137.
- [Ver96] R. VERFÜRTH, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley-Teubner, Chichester, UK, 1996.

PRACTICAL VARIANCE REDUCTION VIA REGRESSION FOR SIMULATING DIFFUSIONS*

G. N. MILSTEIN[†] AND M. V. TRETYAKOV[‡]

Abstract. The well-known variance reduction methods—the method of importance sampling and the method of control variates—can be exploited if an approximation of the required solution is known. Here we employ conditional probabilistic representations of solutions together with the regression method to obtain sufficiently inexpensive (although rather rough) estimates of the solution and its derivatives by using the single auxiliary set of approximate trajectories starting from the initial position. These estimates can effectively be used for significant reduction of variance and further accurate evaluation of the required solution. The developed approach is supported by numerical experiments.

Key words. probabilistic representations of solutions of partial differential equations, numerical integration of stochastic differential equations, Monte Carlo technique, variance reduction methods, regression

AMS subject classifications. Primary, 65C05; Secondary, 65C30, 60H10

DOI. 10.1137/060674661

1. Introduction. The stochastic approach to solving problems of mathematical physics is based on probabilistic representations of their solutions by making use of the weak-sense numerical integration of stochastic differential equations (SDEs) and the Monte Carlo (MC) technique. In this approach we have two main errors: the error of SDE numerical integration and the MC error. The first error essentially depends on the choice of a method of numerical integration, and the second one depends on the choice of the probabilistic representation (it is understood that the first error for a chosen method can be reduced by decreasing the step of discretization, and the MC error for a selected probabilistic representation can be reduced by increasing the number of independent trajectories). While the error of numerical integration is well studied in the systematic theory of numerical integration of SDEs, which allows us to propose suitable effective methods for a lot of typical problems (see, e.g., [16]), in connection with the MC error there is a lack of constructive variance reduction methods.

The well-known variance reduction methods (see [12, 16, 21] and the references therein) of importance sampling and of control variates can be exploited only in the case when an approximation of the required solution $u(t, x)$ is known. However, in general even rough approximations of the desired solution $u(t, x)$ and its derivatives $\partial u / \partial x^i(t, x)$, $i = 1, \dots, d$, are unknown beforehand. At first sight, it seems that approximating them roughly is not difficult since they can be found by the MC technique using a comparatively small number of independent trajectories. But this presupposes evaluating them at many points (t_k, x_k) . Computing $u(t_k, x_k)$ and $\partial u / \partial x^i(t_k, x_k)$ by

*Received by the editors November 10, 2006; accepted for publication (in revised form) October 17, 2008; published electronically February 6, 2009. This work was partially supported by the Royal Society International Joint Project-2004/R2-FS grant and UK EPSRC research grant EP/D049792/1. <http://www.siam.org/journals/sinum/47-2/67466.html>

[†]Department of Mathematics, Ural State University, Lenin Str. 51, 620083 Ekaterinburg, Russia (Grigori.Milstein@usu.ru).

[‡]Department of Mathematics, University of Leicester, Leicester LE1 7RH, UK (M.Tretyakov@le.ac.uk). This author's research was partially supported by a Leverhulme Research Fellowship. Part of this work was done while the author was on study leave granted by the University of Leicester.

the MC technique requires different auxiliary sets of approximate trajectories because of the different starting points (t_k, x_k) . This is too expensive, i.e., as a rule, such a procedure is more expensive than simple increase of the number of trajectories starting from the initial position (t_0, x_0) , at which we aim to find the value of the solution u .

So, a suitable method of constructing $u(t_k, x_k)$ and $\partial u / \partial x^i(t_k, x_k)$ should be comparatively inexpensive. Therefore we cannot require a considerable accuracy of the estimates for $u(t_k, x_k)$ and $\partial u / \partial x^i(t_k, x_k)$ because there is a trade-off between accuracy and computational expenses. Our proposition is to exploit conditional probabilistic representations. Their employment together with the regression method allows us to evaluate $u(t_k, x)$ and $\partial u / \partial x^i(t_k, x)$ using the single auxiliary set of approximate trajectories starting from the initial position (t_0, x_0) only. This plays a crucial role in obtaining sufficiently inexpensive (but at the same time useful for variance reduction) estimates $\hat{u}(t_k, x)$ and $\widehat{\partial u} / \partial x^i(t_k, x)$. The construction of \hat{u} and $\widehat{\partial u} / \partial x^i$ is accompanied by a number of errors of a different nature. Although it is impossible to evaluate these errors satisfactorily, the suitability of $\hat{u}(t_k, x)$ and $\widehat{\partial u} / \partial x^i(t_k, x)$ for variance reduction can be directly verified during computations since the MC error can always be estimated. We emphasize that the obtained (even rather rough) estimates can effectively be used for accurately evaluating the function u not only at the position (t_0, x_0) but at many other positions as well.

This paper is most closely connected with [6, 12, 13, 14] (see also the [16]) and with the works [21, 20] by N. Newton. The method of importance sampling from [6, 12] is exploited in [25] for some specific physical applications. Various other aspects of variance reduction related to simulating diffusions are considered, e.g., in [2, 4, 9, 10, 24] (see also the references therein). An extended list of works devoted to variance reduction of MC simulations can be found in [7].

In section 2 we recall some known facts concerning the MC technique for linear parabolic equations and the general scheme of regression method for estimating conditional expectations. Section 3 is devoted to conditional probabilistic representations of solutions of parabolic equations and their derivatives. These representations together with regression approach play a decisive role in the economical estimating of u and $\partial u / \partial x^i$ at all points (t, x) , given the only set of trajectories starting from the initial point (t_0, x_0) . In section 3.2 we obtain the estimate $\hat{u}(s, x)$ and propose to estimate the derivatives $\partial u / \partial x^i(s, x)$ by $\partial \hat{u} / \partial x^i(s, x)$. This estimation of derivatives is inexpensive from the computational point of view, but they are rather rough. Section 3.3 is devoted to the more accurate way of estimating derivatives using a linear regression method directly to find $\widehat{\partial u} / \partial x^i(t_k, x)$. In section 3.4, we obtain $\widehat{\partial u} / \partial x^i(t_k, x)$ in the case of nonsmooth initial data exploiting probabilistic representations for $\partial u / \partial x^i(s, x)$ which rest on the Malliavin integration by parts. To this aim, we derive a conditional version of the Malliavin integration-by-parts formula adapted to our context. It should be noted that if the dimension d is large, the procedures of sections 3.3 and 3.4 are computationally very demanding since they require integration of the d^2 -dimensional system of first-order variation equations whose solution is present in the probabilistic representations for $\partial u / \partial x^i(s, x)$. Therefore, in practice, the inexpensive procedure of section 3.2 is preferable if d is large. In section 4 we give a simple, analytically tractable example to illustrate the benefits of the proposed variance reduction procedure, and we also test it on a one-dimensional array of stochastic oscillators and on the Black–Scholes pricing model for a binary asset-or-nothing call option. Section 5 gives a summary of the proposed approach to variance reduction.

2. Preliminaries. In this section we recall some known facts concerning probabilistic representations of the solutions of parabolic partial differential equations and the regression method of estimating conditional expectations in the form suitable for our purposes.

2.1. Probabilistic representations. Let us consider the Cauchy problem for the linear parabolic equation

$$(2.1) \quad \frac{\partial u}{\partial t} + \frac{1}{2} \sum_{i,j=1}^d a^{ij}(t,x) \frac{\partial^2 u}{\partial x^i \partial x^j} + \sum_{i=1}^d b^i(t,x) \frac{\partial u}{\partial x^i} + c(t,x)u + g(t,x) = 0, \quad t_0 \leq t < T, \quad x \in \mathbf{R}^d,$$

with the initial condition

$$(2.2) \quad u(T,x) = f(x), \quad x \in \mathbf{R}^d.$$

The matrix $a(t,x) = \{a^{ij}(t,x)\}$ in (2.1) is symmetric and at least positive semidefinite. Let $\sigma(t,x)$ be a matrix obtained from the equation

$$a(t,x) = \sigma(t,x)\sigma^\top(t,x).$$

Let $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$, $t_0 \leq t \leq T$, be a filtered probability space. The solution to the problem (2.1)–(2.2) has the following probabilistic representation (the well-known Feynman–Kac formula):

$$(2.3) \quad u(s,x) = E[f(X_{s,x}(T))Y_{s,x,1}(T) + Z_{s,x,1,0}(T)],$$

where $X_{s,x}(t)$, $Y_{s,x,y}(t)$, $Z_{s,x,y,z}(t)$, $t \geq s$, is the solution of the Cauchy problem for the system of SDEs

$$(2.4) \quad \begin{aligned} dX &= b(t,X)dt + \sigma(t,X)dw(t), \quad X(s) = x, \\ dY &= c(t,X)Ydt, \quad Y(s) = y, \\ dZ &= g(t,X)Ydt, \quad Z(s) = z. \end{aligned}$$

Here $w(t) = (w^1(t), \dots, w^d(t))^\top$ is a d -dimensional $\{\mathcal{F}_t\}_{t \geq t_0}$ -adapted standard Wiener process, and Y and Z are scalars. If $y = 1$, $z = 0$, we shall use the notation $Y_{s,x}(t) := Y_{s,x,1}(t)$, $Z_{s,x}(t) := Z_{s,x,1,0}(t)$ (analogous notation will be used later for some other variables). So,

$$(2.5) \quad u(s,x) = E[f(X_{s,x}(T))Y_{s,x}(T) + Z_{s,x}(T)].$$

There are various sets of sufficient conditions ensuring connection between the solutions of the Cauchy problem (2.1)–(2.2) and their probabilistic representations (2.5)–(2.4). For definiteness, we shall keep the following assumptions.

We assume that the coefficients b , σ , c , and g have bounded derivatives up to some order, and additionally c and g are bounded on $[t_0, T] \times \mathbf{R}^d$. Further, we assume that the matrix $a(t,x)$ is positive definite and, moreover, the uniform ellipticity condition holds: there exists $\sigma_0 > 0$ such that

$$\| a^{-1}(t,x) \| = \| (\sigma(t,x)\sigma^\top(t,x))^{-1} \| \leq \sigma_0^{-1}, \quad t_0 \leq t \leq T, \quad x \in \mathbf{R}^d.$$

As for function $f(x)$, it is assumed to grow at infinity not faster than a polynomial function. It can be both smooth and nonsmooth.

We note that the results of this paper can be used under other sets of conditions. For instance, one can consider situations with nonglobally Lipschitz coefficients [18] or with matrix $a(t, x)$ which is positive semidefinite. For example, in section 4.2 we consider a numerical example with nonglobally Lipschitz coefficients and positive semidefinite matrix $a(t, x)$, and the example from section 4.3 has a discontinuous $f(x)$.

The value $u(s, x)$ from (2.5) can be evaluated using the weak-sense numerical integration of the system (2.4) together with the MC technique. More specifically, we have

$$(2.6) \quad \begin{aligned} u(s, x) &\approx E[f(\bar{X}_{s,x}(T))\bar{Y}_{s,x}(T) + \bar{Z}_{s,x}(T)] \\ &\approx \frac{1}{M} \sum_{m=1}^M [f({}_m\bar{X}_{s,x}(T)){}_m\bar{Y}_{s,x}(T) + {}_m\bar{Z}_{s,x}(T)], \end{aligned}$$

where the first approximate equality involves an error due to replacing X, Y, Z by $\bar{X}, \bar{Y}, \bar{Z}$ (the error is related to the approximate integration of (2.4)) and the error in the second approximate equality comes from the MC technique; ${}_m\bar{X}_{s,x}(T), {}_m\bar{Y}_{s,x}(T), {}_m\bar{Z}_{s,x}(T), m = 1, \dots, M$, are independent realizations of $\bar{X}_{s,x}(T), \bar{Y}_{s,x}(T), \bar{Z}_{s,x}(T)$. While the weak-sense integration of SDEs is developed sufficiently well and a lot of different effective weak-sense numerical methods have been constructed (see, e.g., [16]), the methods of reducing the second error in (2.6) are more intricate.

The error of the MC method is evaluated by

$$\bar{\rho} = c \frac{(\text{var}[f(\bar{X}_{s,x}(T))\bar{Y}_{s,x}(T) + \bar{Z}_{s,x}(T)])^{1/2}}{M^{1/2}},$$

where, e.g., the values $c = 1, 2, 3$ correspond to the fiducial probabilities 0.68, 0.95, 0.997, respectively. Introduce

$$(2.7) \quad \Gamma = \Gamma_{s,x} := f(X_{s,x}(T))Y_{s,x}(T) + Z_{s,x}(T),$$

$$(2.8) \quad \bar{\Gamma} = \bar{\Gamma}_{s,x} := f(\bar{X}_{s,x}(T))\bar{Y}_{s,x}(T) + \bar{Z}_{s,x}(T).$$

Since $\text{var}\Gamma_{s,x}$ is close to $\text{var}\bar{\Gamma}_{s,x}$, we can assume that the error of the MC method is estimated by

$$(2.9) \quad \rho = c \frac{(\text{var}\Gamma_{s,x})^{1/2}}{M^{1/2}}.$$

2.2. Variance reduction. If $\text{var}\Gamma_{s,x}$ is large, then to achieve a satisfactory accuracy we have to simulate a very large number of independent trajectories. Clearly, variance reduction is of crucial importance for effectiveness of any MC procedure. To reduce the MC error, one usually exploits some other probabilistic representations of solutions to considered problems. To obtain various probabilistic representations of the solution to the problem (2.1)–(2.2), we introduce the system (see [13, 14, 16])

$$(2.10) \quad \begin{aligned} dX &= b(t, X)dt - \sigma(t, X)\mu(t, X)dt + \sigma(t, X)dw(t), \quad X(s) = x, \\ dY &= c(t, X)Ydt + \mu^\top(t, X)Ydw(t), \quad Y(s) = 1, \\ dZ &= g(t, X)Ydt + F^\top(t, X)Ydw(t), \quad Z(s) = 0, \end{aligned}$$

where μ and F are column-vector functions of dimension d satisfying some regularity conditions (e.g., they have bounded derivatives with respect to x^i up to some order). We should note that X, Y, Z in (2.10) differ from X, Y, Z in (2.4); however, this does not lead to any ambiguity. The formula (2.5), i.e.,

$$(2.11) \quad u(s, x) = E\Gamma_{s,x},$$

remains valid under the new X, Y, Z . While the mean $E\Gamma$ does not depend on the choice of μ and F , the variance $var\Gamma = E\Gamma^2 - (E\Gamma)^2$ does. Thus, μ and F can be used to decrease the variance $var\Gamma$ and, consequently, the MC error can be reduced. The following theorem is proved in [14] (see also [13, 16]).

THEOREM 2.1. *Let μ and F be such that for any $x \in \mathbf{R}^d$ there exists a solution to the system (2.10) on the interval $[s, T]$. Then the variance $var\Gamma$ is equal to*

$$(2.12) \quad var\Gamma = E \int_s^T Y_{s,x}^2(t) \sum_{j=1}^d \left(\sum_{i=1}^d \sigma^{ij} \frac{\partial u}{\partial x^i} + u\mu^j + F^j \right)^2 dt,$$

provided that the expectation in (2.12) exists. In (2.12) all the functions $\sigma^{ij}, \mu^j, F^j, u, \partial u/\partial x^i$ have $(t, X_{s,x}(t))$ as their argument.

In particular, if μ and F are such that

$$(2.13) \quad \sum_{i=1}^d \sigma^{ij} \frac{\partial u}{\partial x^i} + u\mu^j + F^j = 0, \quad j = 1, \dots, d,$$

then $var\Gamma = 0$, i.e., Γ is deterministic.

We recall that if we put here $F = 0$, then we obtain the method of importance sampling (first considered in [6, 12, 24]), and if we put $\mu = 0$, then we obtain the method of control variates (first considered in [21]). Theorem 2.1 establishes the combining method of variance reduction proved in [13]; see also [16].

Obviously, μ and F satisfying (2.13) cannot be constructed without knowing $u(t, x), s \leq t \leq T, x \in \mathbf{R}^d$. Nevertheless, the theorem claims a general possibility of variance reduction by a proper choice of the functions μ^j and $F^j, j = 1, \dots, d$. Theorem 2.1 can be used, for example, if we know a function $\hat{u}(t, x)$ connected with an approximating problem and which is close to $u(t, x)$. In this case we take any $\hat{\mu}^j, \hat{F}^j, j = 1, \dots, d$, satisfying

$$(2.14) \quad \sum_{i=1}^d \sigma^{ij} \frac{\partial \hat{u}}{\partial x^i} + \hat{u}\hat{\mu}^j + \hat{F}^j = 0,$$

and then the variance $var \Gamma$, though not zero, is small.

Let us emphasize that (2.13) serves only as a guidance for getting suitable μ and F (recall that the mean $E\Gamma$ does not depend on the choice of μ and F). In particular, the derivative estimate $\widehat{\partial u/\partial x^i}$ can differ from $\partial \hat{u}/\partial x^i$. In such cases, instead of (2.14) we use

$$(2.15) \quad \sum_{i=1}^d \sigma^{ij} \frac{\partial \widehat{u}}{\partial x^i} + \hat{u}\hat{\mu}^j + \hat{F}^j = 0.$$

It might seem that the problem of at least rough approximation of the functions $u(t, x)$ and $\partial u/\partial x^i(t, x)$ is not difficult since they can be found approximately due to

the Feynman–Kac formula, numerical integration of SDEs, and the MC technique. But then numerical integration of the system (2.10) presupposes evaluating $u(t_k, \bar{X}_k)$ and $\partial u/\partial x^i(t_k, \bar{X}_k)$ at many points (t_k, \bar{X}_k) . Their evaluation by the MC method requires different sets of auxiliary approximate trajectories because of the different starting points (t_k, \bar{X}_k) . This is too expensive; i.e., as a rule, such a procedure is more expensive than simple increase of M in (2.6).

Our aim is to propose a systematic method of approximating the functions u and $\partial u/\partial x^i$, $i = 1, \dots, d$, relatively cheaply, and hence obtain systematic methods of variance reduction. To this end, we exploit the regression method of evaluating $u(t_k, x)$ and $\partial u/\partial x^i(t_k, x)$, which allows us to use only one set of approximate trajectories starting from the initial position (t_0, x_0) .

2.3. Pathwise approach for derivatives $\partial u/\partial x^i(s, x)$. The probabilistic representation for the derivatives

$$\partial^i(s, x) := \frac{\partial u(s, x)}{\partial x^i}, \quad i = 1, \dots, d,$$

can be obtained by the straightforward differentiation of (2.11) (see, e.g., [7, 13]):

$$(2.16) \quad \partial^i(s, x) = E \left(\sum_{j=1}^d \frac{\partial f(X_{s,x}(T))}{\partial x^j} \delta_{s,x}^i X^j(T) Y_{s,x}(T) + f(X_{s,x}(T)) \delta_{s,x}^i Y(T) + \delta_{s,x}^i Z(T) \right),$$

where

$$\begin{aligned} \delta^i X^j(t) &:= \delta_{s,x}^i X^j(t) := \frac{\partial X_{s,x}^j(t)}{\partial x^i}, \quad \delta^i Y(t) := \delta_{s,x}^i Y(t) := \frac{\partial Y_{s,x}(t)}{\partial x^i}, \\ \delta^i Z(t) &:= \delta_{s,x}^i Z(t) := \frac{\partial Z_{s,x}(t)}{\partial x^i}, \quad s \leq t \leq T, \quad i, j = 1, \dots, d, \end{aligned}$$

satisfy the system of variational equations associated with (2.10):

$$(2.17) \quad d\delta^i X = \sum_{j=1}^d \frac{\partial(b(t, X) - \sigma(t, X)\mu(t, X))}{\partial x^j} \delta^i X^j dt + \sum_{j=1}^d \frac{\partial\sigma(t, X)}{\partial x^j} \delta^i X^j dw(t),$$

$$\delta^i X^j(s) = 0 \text{ if } j \neq i, \text{ and } \delta^i X^i(s) = 1,$$

$$(2.18) \quad d\delta^i Y = \sum_{j=1}^d Y \frac{\partial c(t, X)}{\partial x^j} \delta^i X^j dt + c(t, X) \delta^i Y dt$$

$$+ \sum_{j=1}^d Y \frac{\partial \mu^\top(t, X)}{\partial x^j} \delta^i X^j dw(t) + \mu^\top(t, X) \delta^i Y dw(t), \quad \delta^i Y(s) = 0,$$

$$(2.19) \quad d\delta^i Z = \sum_{j=1}^d Y \frac{\partial g(t, X)}{\partial x^j} \delta^i X^j dt + g(t, X) \delta^i Y dt$$

$$+ \sum_{j=1}^d Y \frac{\partial F^\top(t, X)}{\partial x^j} \delta^i X^j dw(t) + F^\top(t, X) \delta^i Y dw(t), \quad \delta^i Z(s) = 0.$$

Introduce a partition of the time interval $[t_0, T]$, for simplicity the equidistant one: $t_0 < t_1 < \dots < t_N = T$ with step size $h = (T - t_0)/N$. Let us apply a weak scheme (see, e.g., [16]) to the systems of SDEs (2.10), (2.17)–(2.19) to obtain independent approximate trajectories $(t_k, {}_m\bar{X}(t_k))$, $m = 1, \dots, M$, all starting from the point (t_0, x) , and ${}_m\bar{Y}(t_k)$, ${}_m\bar{Z}(t_k)$, ${}_m\bar{\delta}^i X(t_k)$, ${}_m\bar{\delta}^i Y(t_k)$, ${}_m\bar{\delta}^i Z(t_k)$ with ${}_m\bar{Y}(t_0) = 1$, ${}_m\bar{Z}(t_0) = 0$, ${}_m\bar{\delta}^i X^j(t_0) = 0$ if $j \neq i$, and ${}_m\bar{\delta}^i X^i(t_0) = 1$, ${}_m\bar{\delta}^i Y(t_0) = 0$, ${}_m\bar{\delta}^i Z(t_0) = 0$. Then we obtain the following MC estimates of the derivatives $\partial u/\partial x^i(t_0, x)$ from (2.16) with $(s, x) = (t_0, x)$:

$$(2.20) \quad \hat{\partial}^i(t_0, x) = \frac{1}{M} \sum_{m=1}^M \left[\sum_{j=1}^d \frac{\partial f({}_m\bar{X}(T))}{\partial x^j} {}_m\bar{\delta}^i X^j(T) {}_m\bar{Y}(T) + f({}_m\bar{X}(T)) {}_m\bar{\delta}^i Y(T) + {}_m\bar{\delta}^i Z(T) \right].$$

Clearly, the estimates $\hat{\partial}^i(t_k, x)$ for derivatives $\partial u/\partial x^i(t_k, x)$ can be obtained analogously.

Theorem 2.1 asserts that the variance in evaluating u by (2.11) can reach zero value for some μ and F . In [13] it is proved that for the same μ and F the variance in evaluating ∂^i by (2.16) is equal to zero as well (we pay attention that not only μ and F but also their derivatives are present in (2.18) and (2.19)).

2.4. Regression method of estimating conditional expectation. Let us recall the general scheme of the linear regression method (see, e.g., [8]). Consider a sample $({}_mX, {}_mV)$, $m = 1, \dots, M_r$, from a generic member (X, V) of the sample, where X is a d -dimensional and V is a one-dimensional random variable. We pay attention that we denote by M_r the size of the sample used in the regression, while M is the number of realizations used for computing the required quantity $u(t_0, x_0)$ (see (2.6)). Let the values of X belong to a domain $\mathbf{D} \subset \mathbf{R}^d$. It is of interest to estimate the regression function

$$(2.21) \quad c(x) = E(V|X = x).$$

Let $\{\varphi_l(x)\}_{l=1}^L$ be a set of basis functions each mapping \mathbf{D} to \mathbf{R} . As an estimate $\hat{c}(x)$ of $c(x)$, we choose the function of the form $\sum_{l=1}^L \alpha_l \varphi_l(x)$ that minimizes the empirical risk:

$$(2.22) \quad \hat{\alpha} = \arg \min_{\alpha \in \mathbf{R}^L} \frac{1}{M_r} \sum_{m=1}^{M_r} \left({}_mV - \sum_{l=1}^L \alpha_l \varphi_l({}_mX) \right)^2.$$

So

$$(2.23) \quad \hat{c}(x) = \sum_{l=1}^L \hat{\alpha}_l \varphi_l(x),$$

where $\hat{\alpha}_l$ satisfy the system of linear algebraic equations

$$(2.24) \quad \begin{aligned} a_{11}\alpha_1 + a_{12}\alpha_2 + \dots + a_{1L}\alpha_L &= b_1 \\ \dots & \\ a_{L1}\alpha_1 + a_{L2}\alpha_2 + \dots + a_{LL}\alpha_L &= b_L \end{aligned}$$

with

$$(2.25) \quad a_{ln} = \frac{1}{M_r} \sum_{m=1}^{M_r} \varphi_l(mX)\varphi_n(mX), \quad b_l = \frac{1}{M_r} \sum_{m=1}^{M_r} \varphi_l(mX) \, {}_mV, \quad l, n = 1, \dots, L.$$

Thus, the usual base material in the field of regression is a sample $({}_mX, {}_mV)$, $m = 1, \dots, M_r$, from a generic member (X, V) of the sample.

Remark 2.2. Although in this paper we use linear regression, in principle other regression methods (see, e.g., [3, 8]) can be exploited as well.

3. Conditional probabilistic representations and methods of evaluating $u(s, x)$ and $\partial u/\partial x^i(s, x)$ by regression. The routine (unconditional) probabilistic representations are ideal for the MC evaluation of $u(t_0, x_0)$ by using a set of trajectories starting from the point (t_0, x_0) . To find $u(s, x)$ by this approach, we need to construct another set of trajectories which starts from (s, x) . However, we can use the previous set starting from (t_0, x_0) to compute $u(s, x)$, $s > t_0$, if we make use of conditional probabilistic representations. In this section we introduce the conditional probabilistic representations for solutions of parabolic equations and for derivatives of the solutions.

3.1. Conditional probabilistic representations for $u(s, x)$ and $\partial u/\partial x^i(s, x)$. Along with the unconditional probabilistic representation (2.11), (2.7), (2.10) for $u(s, x)$, we have the following conditional one:

$$(3.1) \quad \begin{aligned} u(s, x) &= E(f(X_{s,x}(T))Y_{s,x}(T) + Z_{s,x}(T)) \\ &= E(f(X_{s,X}(T))Y_{s,X}(T) + Z_{s,X}(T) \text{ with } X := X_{t_0,x_0}(s)|X_{t_0,x_0}(s) = x). \end{aligned}$$

This formula can be considered as the conditional version of the Feynman–Kac formula.

Analogously to (3.1), we get for $\partial^i(s, x) = \partial u/\partial x^i(s, x)$ (see (2.16))

$$(3.2) \quad \begin{aligned} \partial^i(s, x) &= E \left(\sum_{j=1}^d \frac{\partial f(X_{s,x}(T))}{\partial x^j} \delta_{s,x}^i X^j(T) Y_{s,x}(T) + f(X_{s,x}(T)) \delta_{s,x}^i Y(T) + \delta_{s,x}^i Z(T) \right) \\ &= E \left(\sum_{j=1}^d \frac{\partial f(X_{s,X}(T))}{\partial x^j} \delta_{s,X}^i X^j(T) Y_{s,X}(T) \right. \\ &\quad \left. + f(X_{s,X}(T)) \delta_{s,X}^i Y(T) + \delta_{s,X}^i Z(T) | X := X_{t_0,x_0}(s) = x \right). \end{aligned}$$

So, we have two different probabilistic representations both for $u(s, x)$ and $\partial^i(s, x)$: the first one is in the form of unconditional expectation (see section 2), and the second one (i.e., (3.1) and (3.2)) is in the form of conditional expectation. The first form can be realized naturally by the MC approach and the second one by a regression method. As we discussed before, it is too expensive to run sets of trajectories starting from various initial points (s, x) , and we do have the set of trajectories $(t, {}_mX_{t_0,x_0}(t))$. Taking this into account, the second way (which relies on the conditional probabilistic representations and regression) is more preferable although it is less accurate.

A proof of (3.1) and (3.2) relies on the following assertion: if ζ is $\tilde{\mathcal{F}}$ -measurable, $f(x, \omega)$ is independent of $\tilde{\mathcal{F}}$, and $E f(x, \omega) = \phi(x)$, then $E(f(\zeta, \omega)|\tilde{\mathcal{F}}) = \phi(\zeta)$ (see,

e.g., [11]). From this assertion, for any measurable g it holds (with $\zeta = X_{t_0,x_0}(s)$, $\tilde{\mathcal{F}} = \sigma\{X_{t_0,x_0}(s)\}$, $f(x, \omega) = g(X_{s,x}(T))$) that

$$E(g(X_{s,X}(T)) | X_{t_0,x_0}(s) = x) = Eg(X_{s,x}(T)) \text{ with } X := X_{t_0,x_0}(s),$$

hence (3.1) and (3.2).

3.2. Evaluating $u(s, x)$. In evaluating $u(s, x)$ by regression, the pairs (X, V) and $({}_mX, {}_mV)$ have the form

$$(3.3) \quad \begin{aligned} (X, V) &\sim (X_{t_0,x_0}(s), f(X_{s,X}(T))Y_{s,X}(T) + Z_{s,X}(T)), \\ ({}_mX, {}_mV) &\sim ({}_mX_{t_0,x_0}(s), f({}_mX_{s,mX}(T)) {}_mY_{s,mX}(T) + {}_mZ_{s,mX}(T)). \end{aligned}$$

To realize a regression algorithm, we construct the set of trajectories $(t, {}_mX_{t_0,x_0}(t))$. Of course, we construct them approximately at the time moments $s = t_k$ and store the obtained values. So, in reality we have $(t_k, {}_m\bar{X}_{t_0,x_0}(t_k))$. The time s in (3.3) is equal to that of t_k . We note that

$$(3.4) \quad X_{s,X}(t) = X_{s,X_{t_0,x_0}(s)}(t) = X_{t_0,x_0}(t), \quad t \geq s;$$

i.e., $X_{s,X}(t)$ is a continuation of the base solution starting at the moment t_0 and $X_{s,X}(T)$ in (3.3) is equal to $X_{t_0,x_0}(T)$. It is not so for Y :

$$Y_{s,X}(T) \neq Y_{t_0,x_0}(T).$$

Let us recall that $Y_{s,X}(t)$ is the solution of the equation (see (2.10))

$$(3.5) \quad dY_{s,X} = c(t, X_{s,X}(t))Y_{s,X} dt + \mu^\top(t, X_{s,X}(t))Y_{s,X} dw(t), \quad Y(s) = 1.$$

Clearly,

$$(3.6) \quad Y_{s,X}(t) = \frac{Y_{t_0,x_0}(t)}{Y_{t_0,x_0}(s)}, \quad s \leq t \leq T,$$

hence storing $Y_{t_0,x_0}(t)$, we can get $Y_{s,X}(T)$ in (3.3).

Analogously, $Z_{s,X}(T) \neq Z_{t_0,x_0}(T)$. It is not difficult to find that

$$(3.7) \quad Z_{s,X}(t) = \frac{1}{Y_{t_0,x_0}(s)}(Z_{t_0,x_0}(t) - Z_{t_0,x_0}(s)), \quad Z_{s,X}(T) = \frac{1}{Y_{t_0,x_0}(s)}(Z_{t_0,x_0}(T) - Z_{t_0,x_0}(s)).$$

Therefore

$$u(s, x) = E \left(f(X_{t_0,x_0}(T)) \frac{Y_{t_0,x_0}(T)}{Y_{t_0,x_0}(s)} + \frac{1}{Y_{t_0,x_0}(s)}(Z_{t_0,x_0}(T) - Z_{t_0,x_0}(s)) \mid X_{t_0,x_0}(s) = x \right).$$

Thus, storing ${}_mX_{t_0,x_0}(t)$, ${}_mY_{t_0,x_0}(t)$, ${}_mZ_{t_0,x_0}(t)$, $t_0 \leq t \leq T$ (in fact, storing ${}_m\bar{X}$, ${}_m\bar{Y}$, ${}_m\bar{Z}$ at t_k), we get the pairs $({}_mX, {}_mV)$ from

$$(X, V) \sim \left(X_{t_0,x_0}(s), f(X_{t_0,x_0}(T)) \frac{Y_{t_0,x_0}(T)}{Y_{t_0,x_0}(s)} + \frac{1}{Y_{t_0,x_0}(s)}(Z_{t_0,x_0}(T) - Z_{t_0,x_0}(s)) \right).$$

Having this sample, one can obtain $\hat{u}(s, x)$ by the linear regression method (see section 2.4):

$$(3.8) \quad \hat{u}(s, x) = \sum_{l=1}^L \hat{\alpha}_l \varphi_l(x).$$

From (3.8) it is straightforward to obtain a very simple estimate $\hat{\partial}^i(s, x)$ for $\partial^i(s, x) = \partial u / \partial x^i(s, x)$:

$$(3.9) \quad \hat{\partial}^i(s, x) = \frac{\partial \hat{u}(s, x)}{\partial x^i} = \sum_{l=1}^L \hat{\alpha}_l \frac{\partial \varphi_l(x)}{\partial x^i}.$$

Then from (2.14) we find some $\hat{\mu}(s, x)$, $\hat{F}(s, x)$ for any $t_0 < s < T$ (in reality for any t_k) and construct the variate $\hat{\Gamma}(t_0, x_0)$ (see (2.5) and (2.7)) for $u(t_0, x_0)$ due to the system (2.10) with $\mu = \hat{\mu}$ and $F = \hat{F}$. We repeat that the variate $\hat{\Gamma}(t_0, x_0)$ is unbiased for any $\hat{\mu}$ and \hat{F} . We note that it is sufficient to have rather rough (in comparison with the required accuracy in evaluating $u(t_0, x_0)$) approximations $\hat{\mu}(s, x)$ and $\hat{F}(s, x)$ of some optimal μ and F from (2.13). Therefore, it is natural to use a coarser discretization and fewer MC runs in the regression part of evaluating $\hat{u}(s, x)$ due to (3.8), i.e., to take M_r in (2.22) smaller than M and to construct samples ${}_m X$ in (2.25) with a comparatively rough discretization. Then in computing $u(t_0, x_0)$ with a finer discretization, the necessary values of $\hat{\mu}$ and \hat{F} at the intermediate points can be obtained after, e.g., linear interpolation of \hat{u} with respect to time. The success of any regression-based approach clearly depends on the choice of basis functions. This is known to be a rather complicated problem, both in practice and theory. In fact, it is necessary to use a special basis tailored to each particular problem. Fortunately, the variance can easily be evaluated during simulation. Therefore, it is not very expensive from the computational point of view to check the quality of a given basis if we take coarse discretizations both in the regression part and in the main part of evaluating $u(t_0, x_0)$ and if we take not too large numbers M_r and M of MC runs. This can help in choosing a proper basis.

Remark 3.1. Clearly, $\hat{\alpha}_l$ depend on s (on t_k). Let us note that the number L and the set $\{\varphi_l(x)\}_{l=1}^L$ may depend on t_k as well.

Remark 3.2. It is obvious that in practice we use (2.10) with different μ and F in the implementation of the regression and in computing the required quantity $u(t_0, x_0)$. Indeed, in the regression part of the procedure we can take arbitrary μ and F (e.g., both zero), while in computing $u(t_0, x_0)$ we choose μ and F according to (2.14) with \hat{u} obtained via the regression or according to (2.15) with \hat{u} and $\widehat{\partial u} / \partial x^i$ obtained via the regression.

Remark 3.3. At $s = t_0$ the system (2.24) degenerates into the single equation (we suppose that not all of $\varphi_l(x_0)$ are equal to zero)

$$(3.10) \quad \varphi_1(x_0)\alpha_1 + \dots + \varphi_L(x_0)\alpha_L = \frac{1}{M_r} \sum_{m=1}^{M_r} [f({}_m \bar{X}_{t_0, x_0}(T)) \quad {}_m \bar{Y}_{t_0, x_0}(T) + {}_m \bar{Z}_{t_0, x_0}(T)].$$

Therefore, the coefficients $\alpha_1(t_0), \dots, \alpha_L(t_0)$ cannot be found from (3.10) uniquely. At the same time, the linear combination $\alpha_1(t_0)\varphi_1(x_0) + \dots + \alpha_L(t_0)\varphi_L(x_0)$, i.e., the estimate

$$\hat{u}(t_0, x_0) = \frac{1}{M_r} \sum_{m=1}^{M_r} [f({}_m \bar{X}_{t_0, x_0}(T)) \quad {}_m \bar{Y}_{t_0, x_0}(T) + {}_m \bar{Z}_{t_0, x_0}(T)],$$

is defined uniquely. Clearly, when t_k is close to t_0 (for instance, at t_1), the system (2.24), though not degenerate, is ill-conditioned. Nevertheless, for such t_k and for x

close to x_0 , the estimate

$$\hat{u}(t_k, x) = \alpha_1(t_k)\varphi_1(x) + \cdots + \alpha_L(t_k)\varphi_L(x)$$

can be found sufficiently accurate. However, since it is not possible to satisfactorily determine the coefficients $\alpha_1(t_k), \dots, \alpha_L(t_k)$, we cannot get the derivatives $\partial\hat{u}(t_k, x)/\partial x^i$ by direct differentiation as $\alpha_1(t_k)\partial\varphi_1(x)/\partial x^i + \cdots + \alpha_L(t_k)\partial\varphi_L(x)/\partial x^i$. In addition, let us emphasize that such difficulties are not essential for the whole procedure of variance reduction because the variance is equal to the integral (2.12), and unsatisfactory knowledge of u and $\partial u/\partial x^i$ on short parts of the interval $[t_0, T]$ does not significantly affect the value of the integral.

3.3. Evaluating $\partial u/\partial x^i(s, x)$. The problem of evaluating $\partial u/\partial x^i(s, x)$ is of independent importance due to its connection with numerical computation of Greeks in finance. Many articles are devoted to pathwise methods of estimating Greeks (see [7] and the references therein; see also [13]). In [17] the finite-difference-based method is developed, and [5, 4] suggest using Malliavin calculus for computing Greeks. Several pathwise and finite-difference-based methods for calculating sensitivities of Bermudan options using regression methods and MC simulations are considered in [1] (see also the references therein). In this section we propose a conditional version of the pathwise method, and in section 3.4 we present a conditional version of the approach based on the Malliavin integration by parts for evaluating $\partial u/\partial x^i(s, x)$.

As mentioned previously, differentiating the equality (3.8) gives an estimate for $\partial^i(s, x) = \partial u/\partial x^i(s, x)$ (see (3.9)); however, in general, it is rather rough. A more accurate way is to use the linear regression method directly.

In evaluating $\partial^i(s, x)$ by regression, the pair (X, V^i) has the form (see (3.2))

(3.11)

$$\begin{aligned} X &= X_{t_0, x_0}(s), \\ V^i &= \sum_{j=1}^d \frac{\partial f(X_{s, X}(T))}{\partial x^j} \delta_{s, X}^i X^j(T) Y_{s, X}(T) + f(X_{s, X}(T)) \delta_{s, X}^i Y(T) + \delta_{s, X}^i Z(T). \end{aligned}$$

We already have expressions for $X_{s, X}(T), Y_{s, X}(T), Z_{s, X}(T)$ via $X_{t_0, x_0}(t), Y_{t_0, x_0}(t), Z_{t_0, x_0}(t)$, with t being equal to s and T (see the formulas (3.4), (3.6), (3.7)). Our nearest aim is to express $\delta_{s, X}^i X^j(T), \delta_{s, X}^i Y(T), \delta_{s, X}^i Z(T)$ via $X_{t_0, x_0}(t), Y_{t_0, x_0}(t), Z_{t_0, x_0}(t), \delta_{t_0, x_0}^i X^j(t), \delta_{t_0, x_0}^i Y(t), \delta_{t_0, x_0}^i Z(t)$.

We begin with $\delta_{s, X}^i X^j(t)$. The column-vector $\delta_{s, X}^i X(t)$ is the solution of the linear homogeneous stochastic system (2.17) whose coefficients depend on $X_{s, X}(t) = X_{t_0, x_0}(t)$. Let the matrix

$$\Phi_{s, X}(t) := \{\delta_{s, X}^i X^j(t)\}$$

be the fundamental matrix of solutions of (2.17) normalized at time s , i.e., $\Phi_{s, X}(s) = I$, where I is the identity matrix. Its element on the j th row and i th column is equal to $\delta_{s, X}^i X^j(t)$. Clearly,

$$(3.12) \quad \Phi_{s, X}(t) = \Phi_{t_0, x_0}(t) \Phi_{t_0, x_0}^{-1}(s).$$

Now let us turn to the column-vector $\delta_{s, X} Y(t)$, consisting of components $\delta_{s, X}^i Y(t)$. We have (see (2.18))

$$(3.13) \quad \begin{aligned} d\delta_{s, X} Y &= Y_{s, X}(t) \Phi_{s, X}^\top(t) \nabla c(t, X_{s, X}(t)) dt + c(t, X_{s, X}(t)) \delta_{s, X} Y dt \\ &+ Y_{s, X}(t) \Phi_{s, X}^\top(t) \nabla[\mu^\top(t, X_{s, X}(t)) dw(t)] + \delta_{s, X} Y \mu^\top(t, X_{s, X}(t)) dw(t), \quad \delta_{s, X} Y(s) = 0. \end{aligned}$$

Due to the equality $X_{s,X}(t) = X_{t_0,x_0}(t)$ and (3.6) and (3.12), we get from (3.13)

$$(3.14) \quad \begin{aligned} d\delta_{s,X}Y &= \frac{Y_{t_0,x_0}(t)}{Y_{t_0,x_0}(s)}[\Phi_{t_0,x_0}^{-1}(s)]^\top \Phi_{t_0,x_0}^\top(t) \nabla c(t, X_{t_0,x_0}(t))dt + c(t, X_{t_0,x_0}(t))\delta_{s,X}Y dt \\ &\quad + \frac{Y_{t_0,x_0}(t)}{Y_{t_0,x_0}(s)}[\Phi_{t_0,x_0}^{-1}(s)]^\top \Phi_{t_0,x_0}^\top(t) \nabla[\mu^\top(t, X_{t_0,x_0}(t))]dw(t) \\ &\quad + \delta_{s,X}Y\mu^\top(t, X_{t_0,x_0}(t))dw(t), \quad \delta_{s,X}Y(s) = 0. \end{aligned}$$

Taking into account the equality

$$\begin{aligned} d\delta_{t_0,x_0}Y(t) &= Y_{t_0,x_0}(t)\Phi_{t_0,x_0}^\top(t) \nabla c(t, X_{t_0,x_0}(t))dt + c(t, X_{t_0,x_0}(t))\delta_{t_0,x_0}Y(t)dt \\ &\quad + Y_{t_0,x_0}(t)\Phi_{t_0,x_0}^\top(t) \nabla[\mu^\top(t, X_{t_0,x_0}(t))]dw(t) + \delta_{t_0,x_0}Y(t)\mu^\top(t, X_{t_0,x_0}(t))dw(t), \end{aligned}$$

it is not difficult to verify that

$$(3.15) \quad \delta_{s,X}Y(t) = \frac{1}{Y_{t_0,x_0}(s)}[\Phi_{t_0,x_0}^{-1}(s)]^\top \left(\delta_{t_0,x_0}Y(t) - \frac{Y_{t_0,x_0}(t)}{Y_{t_0,x_0}(s)}\delta_{t_0,x_0}Y(s) \right).$$

In the similar way we obtain

$$(3.16) \quad \begin{aligned} \delta_{s,X}Z(t) &= \frac{1}{Y_{t_0,x_0}(s)}[\Phi_{t_0,x_0}^{-1}(s)]^\top (\delta_{t_0,x_0}Z(t) - \delta_{t_0,x_0}Z(s)) \\ &\quad - \frac{1}{Y_{t_0,x_0}^2(s)}[\Phi_{t_0,x_0}^{-1}(s)]^\top \delta_{t_0,x_0}Y(s) (Z_{t_0,x_0}(t) - Z_{t_0,x_0}(s)). \end{aligned}$$

Hence the column-vector $\partial(s, x)$ with the components $\partial^i(s, x)$ is equal to

$$(3.17) \quad \begin{aligned} \partial(s, x) &= E \left(\frac{Y_{t_0,x_0}(T)}{Y_{t_0,x_0}(s)}[\Phi_{t_0,x_0}^{-1}(s)]^\top \Phi_{t_0,x_0}^\top(T) \nabla f(X_{t_0,x_0}(T)) \right. \\ &\quad \left. + f(X_{t_0,x_0}(T))\delta_{s,X}Y(T) + \delta_{s,X}Z(T) \mid X_{t_0,x_0}(s) = x \right), \end{aligned}$$

where $\delta_{s,X}Y(T)$ and $\delta_{s,X}Z(T)$ are from (3.15) and (3.16).

Thus, storing ${}_mX_{t_0,x_0}(t)$, ${}_mY_{t_0,x_0}(t)$, ${}_mZ_{t_0,x_0}(t)$, ${}_m\Phi_{t_0,x_0}(t)$, ${}_m\delta_{t_0,x_0}Y(t)$, ${}_m\delta_{t_0,x_0}Z(t)$, $t_0 \leq t \leq T$, we get the corresponding samples

$$(3.18) \quad \begin{aligned} ({}_mX, {}_mV^i) &= \left({}_mX_{t_0,x_0}(s), \left(\frac{{}_mY_{t_0,x_0}(T)}{{}_mY_{t_0,x_0}(s)}[{}_m\Phi_{t_0,x_0}^{-1}(s)]^\top {}_m\Phi_{t_0,x_0}^\top(T) \nabla f({}_mX_{t_0,x_0}(T)) \right. \right. \\ &\quad \left. \left. + f({}_mX_{t_0,x_0}(T)) {}_m\delta_{s,{}_mX}Y(T) + {}_m\delta_{s,{}_mX}Z(T) \right)^i \right), \end{aligned}$$

where ${}_m\Phi_{t_0,x_0}(s)$ is a realization of the fundamental matrix $\Phi_{t_0,x_0}(s)$ which corresponds to the same elementary event $\omega \in \Omega$ as the realization ${}_mX_{t_0,x_0}(t)$. We use $({}_mX, {}_mV^i)$ for evaluating $\partial^i(s, x)$, $i = 1, \dots, d$, by the linear regression method:

$$(3.19) \quad \hat{\partial}^i(s, x) = \sum_{l=1}^L \hat{\beta}_l^i \psi_l(x).$$

Remark 3.4. This paper is most closely connected with [6, 12, 13, 14] (see also [16]) and with the works [21, 20] by N. Newton. In [21, 20], both the method of control variates and the method of importance sampling for calculating solutions $u(t, x)$ of parabolic partial differential equations by the MC method are considered. In both cases, a perfect variate (i.e., one which is unbiased and has zero variance) is constructed based on the Funke–Shevlyakov–Haussmann formula (see the corresponding reference and details in [21]; such a formula is usually called as the Clark–Ocone–Haussmann formula). Then some approximation methods of simulating the variates are proposed in [21, 20] to yield unbiased estimators for the desired solution $u(t, x)$ with reduced variances. If the dimension d is large, the most labor-consuming calculations are connected with integration of the d^2 -dimensional system of first-order variation equations. This is required to construct the estimators. In this paper, we use variates in the form (2.11), (2.10) with μ and F satisfying (2.13). Due to Theorem 2.1, these variates are perfect if u and $\partial u/\partial x^i$ are exact. We evaluate u and $\partial u/\partial x^i$ based on conditional probabilistic representations and construct unbiased estimators for $u(t, x)$ using (2.15) or (2.14). We note that (2.14) allows us to avoid estimating $\partial u/\partial x^i$ (see (3.8)–(3.9)) and hence to avoid integration of the equations of first-order variation. In addition, the obtained estimator by (2.14) remains unbiased. In spite of the fact that our approach and that of N. Newton clearly differ, they undoubtedly have profound connections. For example, the Clark–Ocone–Haussmann formula, being the basis for Newton’s approach, can fairly easily be derived using the conditional probabilistic representations (3.1), (3.2).

3.4. Evaluating $\partial u/\partial x^i(s, x)$ using the Malliavin integration by parts.

If $f(x)$ is an irregular function, one can use the procedure recommended in section 3.2, where we do not need direct calculations of derivatives $\partial u/\partial x^i$. Another way consists in approximating f by a smooth function with the consequent use of the procedure from section 3.3. Because we do not pursue a high accuracy in estimating u and $\partial u/\partial x^i$, such approximation of f can be quite satisfactory. For direct calculation of derivatives $\partial u/\partial x^i$ without smoothing f , we can use the conditional version of the integration-by-parts (Bismut–Elworthy–Li) formula. This formula is successfully applied for evaluating deltas in the case of an irregular f (see, e.g., [5, 4, 22]).

For calculating $\partial u/\partial x^i$ in the case of u given by

$$u(s, x) = E\Gamma_{s,x} = E[f(X_{s,x}(T))Y_{s,x}(T) + Z_{s,x}(T)],$$

where $X_{s,x}(T)$, $Y_{s,x}(T)$, $Z_{s,x}(T)$ satisfy system (2.10), the following variant of the integration-by-parts formula can be derived:

$$\begin{aligned} (3.20) \quad \partial^i(s, x) &= \frac{1}{T-s} E\Gamma_{s,x} \int_s^T \left[\sigma^{-1} \frac{\partial X_{s,x}(s')}{\partial x^i} \right]^\top dw(s') \\ &- \frac{1}{T-s} E\Gamma_{s,x} \int_s^T \mu^\top \sigma^{-1} \frac{\partial X_{s,x}(s')}{\partial x^i} ds' + \frac{1}{T-s} E \int_s^T Z_{s,x}(s') \mu^\top \sigma^{-1} \frac{\partial X_{s,x}(s')}{\partial x^i} ds' \\ &+ \frac{1}{T-s} E\Gamma_{s,x} \int_s^T \frac{1}{Y_{s,x}(s')} \frac{\partial Y_{s,x}(s')}{\partial x^i} ds' - \frac{1}{T-s} E \int_s^T \frac{Z_{s,x}(s')}{Y_{s,x}(s')} \frac{\partial Y_{s,x}(s')}{\partial x^i} ds' \\ &- \frac{1}{T-s} E \int_s^T Y_{s,x}(s') F^\top \sigma^{-1} \frac{\partial X_{s,x}(s')}{\partial x^i} ds' + \frac{1}{T-s} E \int_s^T \frac{\partial Z_{s,x}(s')}{\partial x^i} ds' := D^i(s, x), \end{aligned}$$

where μ^\top , σ^{-1} , and F^\top have $(s', X_{s,x}(s'))$ as their arguments. In particular, if $c = 0$, $g = 0$, $\mu = 0$, $F = 0$, we get the well-known integration-by-parts formula (see,

e.g., [22]):

$$(3.21) \quad \partial^i(s, x) = \frac{1}{T-s} Ef(X_{s,x}(T)) \int_s^T \left[\sigma^{-1}(s', X_{s,x}(s')) \frac{\partial X_{s,x}(s')}{\partial x^i} \right]^\top dw(s').$$

As in section 3.1, together with the unconditional probabilistic representation (3.20) for $\partial^i(s, x)$, we have the following conditional one:

$$(3.22) \quad \partial^i(s, x) = E(D^i(s, X)|X := X_{t_0, x_0}(s) = x).$$

Again, the formula (3.20) is natural for the MC approach and (3.22) for a regression method. An implementation of the regression method is based upon the corresponding approximation $({}_mX, {}_mV^i)$ of the pair $(X, V^i) = (X_{t_0, x_0}(s), D^i(s, X_{t_0, x_0}(s)))$ following the ideas of section 3.3.

3.5. Two-run procedure. The straightforward implementation of evaluating $u(s, x)$ and $\partial u/\partial x^i(s, x)$ by regression as described in sections 3.2 and 3.3 requires storing

$${}_m\Lambda(t_k) := ({}_mX_{t_0, x_0}(t_k), {}_mY_{t_0, x_0}(t_k), {}_mZ_{t_0, x_0}(t_k), {}_m\Phi_{t_0, x_0}(t_k), {}_m\delta_{t_0, x_0}Y(t_k), {}_m\delta_{t_0, x_0}Z(t_k))$$

(or, more precisely, their approximations ${}_m\bar{\Lambda}(t_k)$) at all $t_k, k = 1, \dots, N$, in the main computer memory (RAM) until the end of the simulation. This puts a requirement on the RAM size that is too demanding and limits the practicality of the proposed approach since in almost any practical problem a relatively large number of time steps is needed. However, this difficulty can be overcome and we can avoid storing ${}_m\bar{\Lambda}(t_k)$ at all t_k by implementing the two-run procedure described below.

First, we recall that, as a rule, pseudorandom number generators used for MC simulations have the property that the sequence of random numbers obtained by them is easily reproducible (see, e.g., [16] and the references therein). Let us fix a sequence of pseudorandom numbers. The two-run procedure can schematically be presented as follows.

First run:

- simulate M_r number of independent trajectories ${}_m\bar{\Lambda}(t_k), k = 1, \dots, N$, with an arbitrary choice of μ and F (e.g., $\mu = 0$ and $F = 0$);
- compute and store the values ${}_m\bar{\Gamma}$ to form the component V needed for the regression in the second run and compute and store the values

$${}_m\bar{Y}(T) {}_m\bar{\Phi}_{t_0, x_0}^\top(T) \nabla f({}_m\bar{X}(T)) + f({}_m\bar{X}(T)) {}_m\bar{\delta Y}(T) + {}_m\bar{\delta Z}(T)$$

and ${}_m\bar{Y}(T)$ to form the components V^i in the second run.

Second run:

- reinitialize the random number generator so that it produces the same sequence as for the first run;
- for $k = 1, \dots, N$
 - simulate the same ${}_m\bar{\Lambda}(t_k), m = 1, \dots, M_r$, as in the first run (i.e., they correspond to the same sequence of pseudorandom numbers as in the first run), keeping only the current ${}_m\bar{\Lambda}(t_k)$ in RAM;
 - use the values stored in RAM during the first run and ${}_m\bar{\Lambda}(t_k)$ from this run to find $\bar{u}(t_k, x)$ and $\bar{\partial u}/\partial x^i(t_k, x)$ by regression (${}_m\bar{\Lambda}(t_k)$ and ${}_m\bar{\Lambda}(T)$ form the pairs $({}_mX, {}_mV)$ and $({}_mX, {}_mV^i)$ needed for the regression);

- use the found $\bar{u}(t_k, x)$ and $\overline{\partial u / \partial x^i}(t_k, x)$ to obtain $\bar{\mu}(t_k, x)$ and $\bar{F}(t_k, x)$ required for variance reduction (see section 2.2);
- simulate (2.10) with $\mu = \bar{\mu}$ and $F = \bar{F}$ on this step and thus obtain M independent triples

$$\begin{aligned} &({}_m\tilde{X}_{t_0, x_0}(t_k), {}_m\tilde{Y}_{t_0, x_0}(t_k), {}_m\tilde{Z}_{t_0, x_0}(t_k)) = ({}_m\tilde{X}_{t_{k-1}, m\tilde{X}(t_{k-1})}(t_k), \\ &{}_m\tilde{Y}_{t_{k-1}, m\tilde{X}(t_{k-1}), m\tilde{Y}(t_{k-1})}(t_k), {}_m\tilde{Z}_{t_{k-1}, m\tilde{X}(t_{k-1}), m\tilde{Y}(t_{k-1}), m\tilde{Z}(t_{k-1})}(t_k)), \end{aligned}$$

which we keep in RAM until the next step;

- use the obtained $({}_m\tilde{X}_{t_0, x_0}(T), {}_m\tilde{Y}_{t_0, x_0}(T), {}_m\tilde{Z}_{t_0, x_0}(T))$ to get the required $u(t_0, x_0)$ (see (2.6)).

We emphasize that in the two-run procedure at each time moment $s = t_k$ we need to keep in memory only the precomputed values stored at the end of the first run and the values ${}_m\bar{\Lambda}(t_k)$ and $({}_m\tilde{X}_{t_0, x_0}(t_k), {}_m\tilde{Y}_{t_0, x_0}(t_k), {}_m\tilde{Z}_{t_0, x_0}(t_k))$ (only at the current time step k), which is well within RAM limits of a PC.

We note that the two-run realization of the procedure from section 3.2 based on using regression for estimating u only is less computationally demanding (both on processor time and RAM and especially for problems of large dimension d) than the procedures of sections 3.3 and 3.4 which estimate the derivatives of u via regression.

The two-run procedure was used in the numerical experiments of sections 4.2 and 4.3.

4. Examples. The first example is partly illustrative and partly theoretical. The second and third examples are numerical.

4.1. Heat equation. Consider the Cauchy problem

$$\begin{aligned} (4.1) \quad & \frac{\partial u}{\partial t} + \frac{\sigma^2}{2} \frac{\partial^2 u}{\partial x^2} = 0, \quad t_0 \leq t < T, \quad x \in \mathbf{R}, \\ & u(T, x) = x^2. \end{aligned}$$

Its solution is

$$(4.2) \quad u(t, x) = \sigma^2(T - t) + x^2.$$

The probabilistic representation (2.10), (2.11) with $\mu = 0$ takes the form

$$(4.3) \quad u(s, x) = E [X_{s,x}^2(T) + Z_{s,x}(T)] = E\Gamma_{s,x},$$

$$(4.4) \quad dX = \sigma dw(t), \quad X(s) = x,$$

$$(4.5) \quad dZ = F(t, X)dw(t), \quad Z(s) = 0.$$

Due to Theorem 2.1, we have $var\Gamma_{s,x} = var [X_{s,x}^2(T) + Z_{s,x}(T)] = 0$ for the optimal choice of the function $F(t, x) = -\sigma\partial u/\partial x = -2\sigma x$. We note that in this example $\partial u/\partial x$ and the optimal F do not depend on time t .

For the purpose of this illustrative example, we evaluate $u(0, 0) = E\Gamma_{0,0}$. Let us simulate (4.4) exactly (i.e., we have no error of numerical integration):

$$(4.6) \quad X_0 = x, \quad X_{k+1} = X_k + \sigma\Delta_k w, \quad k = 0, \dots, N - 1, \quad \Delta_k w := w(t_{k+1}) - w(t_k).$$

For $F \equiv 0$, we have $u(0, 0) = E\Gamma_{0,0} \approx \hat{u}(0, 0) = \frac{1}{M} \sum_{m=1}^M {}_mX_N^2$, where ${}_mX_N$ are independent realizations of X_N obtained by (4.6). Further, $var\Gamma_{0,0} = 2\sigma^4 T^2$, and

hence the MC error is equal to (see (2.9))

$$(4.7) \quad \rho = c \frac{\sqrt{2}\sigma^2 T}{\sqrt{M}}.$$

For instance, to achieve the accuracy $\rho = 0.0001$ for $c = 3$ (recall that there is no error of numerical integration here) in the case of $\sigma = 1$ and $T = 10$, one needs to perform $M = 18 \times 10^{10}$ MC runs.

To reduce the MC error, we estimate $\partial u/\partial x$ by regression to get $\hat{F}(t_k, x)$ close to the optimal $F = -2\sigma x$. As the basis functions for the regression, we take the first two Hermite polynomials:

$$(4.8) \quad \psi_1(x) = 1, \quad \psi_2(x) = 2x.$$

We note that in this example the required derivative $\partial u/\partial x$ can be expanded in the basis (4.8); i.e., here we do not have any error due to the cut-off of a set of basis functions. In the construction of the estimate for $\partial u/\partial x$, we put $F = 0$ in (4.5).

The variational equation associated with (4.4) has the form (see (2.17)) $d\delta X = 0$, $\delta X(s) = 1$, and hence $\delta X(t) = 1$, $t \geq s$. Thus, the sample from (3.18) takes the form $({}_mX, {}_mV) = ({}_mX_{t_0, x_0}(s), 2 {}_mX_{t_0, x_0}(T))$ and the estimator $\hat{\partial}(t_k, x)$ for $\partial u/\partial x(t_k, x)$ is constructed as

$$(4.9) \quad \hat{\partial}(t_k, x) = \hat{\alpha}_1(t_k) + 2\hat{\alpha}_2(t_k)x, \quad k = 1, \dots, N,$$

where $\hat{\alpha}_1(t_k)$ and $\hat{\alpha}_2(t_k)$ satisfy the system of linear algebraic equations (see (2.24)–(2.25))

$$(4.10) \quad \begin{aligned} a_{11}\alpha_1 + a_{12}\alpha_2 &= b_1, \\ a_{21}\alpha_1 + a_{22}\alpha_2 &= b_2, \end{aligned}$$

$$(4.11) \quad \begin{aligned} a_{11} &= 1, \quad a_{12} = a_{21} := a_{12}(t_k) = \frac{1}{M_r} \sum_{m=1}^{M_r} 2 \times {}_mX(t_k), \\ a_{22} &:= a_{22}(t_k) = \frac{1}{M_r} \sum_{m=1}^{M_r} 4 \times ({}_mX(t_k))^2, \end{aligned}$$

$$b_1 := b_1(t_k) = \frac{1}{M_r} \sum_{m=1}^{M_r} 2 \times {}_mX(T), \quad b_2 := b_2(t_k) = \frac{1}{M_r} \sum_{m=1}^{M_r} 4 \times {}_mX(t_k) \times {}_mX(T).$$

Here ${}_mX(t_k)$, $m = 1, \dots, M_r$, $k = 1, \dots, N$, are independent realizations of $X(t_k)$ obtained by (4.6). Hence

$$(4.12) \quad \hat{\alpha}_1(t_k) = \frac{b_1 a_{22} - b_2 a_{12}}{a_{22} - (a_{12})^2}, \quad \hat{\alpha}_2(t_k) = \frac{b_2 - b_1 a_{12}}{a_{22} - (a_{12})^2}.$$

We define

$$(4.13) \quad \begin{aligned} \hat{F}(0, x) &= -\frac{\sigma}{M_r} \sum_{m=1}^{M_r} 2 \times {}_mX(T), \\ \hat{F}(t, x) &= -\sigma (\hat{\alpha}_1(t_k) + 2\hat{\alpha}_2(t_k)x) \quad \text{for } t \in (t_{k-1}, t_k], \quad k = 1, \dots, N. \end{aligned}$$

We simulate (4.5) with $F = \hat{F}(t, x)$ exactly (i.e., again we have no error of numerical integration):

$$(4.14) \quad Z_0 = 0,$$

$$Z_{k+1} = Z_k - \sigma \hat{\alpha}_1(t_{k+1}) \Delta_k w - 2\sigma^2 \hat{\alpha}_2(t_{k+1}) w(t_k) \Delta_k w - \sigma^2 \hat{\alpha}_2(t_{k+1}) \left[(\Delta_k w)^2 - h \right].$$

The increments $\Delta_k w$ are the same both in (4.6) and in (4.14) and are independent of the ones used to estimate $\hat{\alpha}_1$ and $\hat{\alpha}_2$.

We simulate

$$(4.15) \quad u(0, 0) = E\Gamma_{0,0} = E(X_N^2 + Z_N) \approx \hat{u}(0, 0) = \frac{1}{M_r} \sum_{m=1}^{M_r} ({}_m X_N^2 + {}_m Z_N),$$

where ${}_m X_N$ and ${}_m Z_N$ are independent realizations of X_N and Z_N obtained according to (4.6) and (4.14). We note that the approximation (4.15) does not have the numerical integration error or the error due to the cut-off of the basis; it has the MC error only.

Using Theorem 2.1, one can evaluate $var\Gamma_{0,0}$ in the case of $F = \hat{F}$ defined in (4.13) and obtain $var\Gamma_{0,0} \approx 4\sigma^4 T^2 / M_r$. Then the MC error ρ in this case is equal to (compare with (4.7))

$$(4.16) \quad \rho \approx c \frac{2\sigma^2 T}{\sqrt{M M_r}}.$$

This example illustrates that in the absence of the error due to the cut-off of a set of basis functions used in regression and of the numerical integration error, the MC error is reduced $\sim 1/\sqrt{M_r}$ times by the proposed variance reduction technique. This is, of course, a significant improvement. Indeed, let us return to the example discussed after (4.7). The estimate (4.16) implies that to achieve the accuracy $\rho = 0.0001$ for $c = 3$ in the case of $\sigma = 1$ and $T = 10$, one can take, e.g., $M = M_r = 6 \times 10^5$; i.e., one can run about 10^5 times fewer trajectories than when the variance reduction was not used (see the discussion after (4.7)). The gain of computational efficiency is significant in spite of the fact that there is an overhead cost of solving the linear system (4.10) in the “regression’s runs.”

Remark 4.1. In the above analysis we assumed that “regression’s runs” and the MC runs for computing the desired value $u(0, 0)$ are independent. In practice, this assumption can be dropped, and we can use the same paths $X(t)$ for both the “regression’s runs” and the MC runs. Then, as a rule, we choose $M_r \leq M$.

Remark 4.2. We are expecting (see also experiments in section 4.2) that in the general case the MC error after application of this variance reduction technique has the form

$$(4.17) \quad \rho = O\left(\frac{1}{\sqrt{M M_r}} + \frac{h^{p/2}}{\sqrt{M}} + \frac{err_B}{\sqrt{M}}\right),$$

where the first term has the same nature as in this illustrative example (see (4.16)); the second term is due to the error of numerical integration (it is assumed that a method of weak order p is used); and the third one arises as a result of the use of a finite set of functions as the basis in the regression, while the solution $u(t, x)$ is usually expandable in a basis consisting of an infinite number of functions (i.e., this

error is due to the cut-off of the basis). We note that finding an appropriate basis for regression in applying this variance reduction approach to a particular problem can be a difficult task and requires some knowledge of the solution $u(t, x)$ of the considered problem. Roughly speaking, in the proposed implementation of the variance reduction methods (the method of importance sampling, the method of control variates, or the combining method) we substitute the task of finding an approximate solution to the problem of interest with the task of finding an appropriate basis for the regression.

For complicated systems of SDEs, it is preferable to use regression to approximate the solution $u(t, x)$ and then differentiate this approximation to approximate the derivatives $\partial u / \partial x^i$. In the case of this illustrative example we take the first three Hermite polynomials,

$$(4.18) \quad \psi_1(x) = 1, \quad \psi_2(x) = 2x, \quad \psi_3(x) = 4x^2 - 2,$$

as the basis functions for the regression. In this example the required function $u(t, x)$ can be expanded in the basis (4.18). We construct the estimator $\hat{u}(t_k, x)$ for $u(t_k, x)$:

$$(4.19) \quad \hat{u}(t_k, x) = \hat{\alpha}_1(t_k) + 2\hat{\alpha}_2(t_k)x + \hat{\alpha}_3(t_k) \cdot (4x^2 - 2), \quad k = 1, \dots, N,$$

where $\hat{\alpha}_1(t_k), \hat{\alpha}_2(t_k), \hat{\alpha}_3(t_k)$ satisfy the system of linear algebraic equations (2.24) with the corresponding coefficients. Further, we approximate the derivative $\partial u / \partial x(t_k, x)$,

$$(4.20) \quad \frac{\partial u}{\partial x}(t_k, x) \approx 2\hat{\alpha}_2(t_k) + 8\hat{\alpha}_3(t_k)x,$$

with $\hat{\alpha}_2(t_k)$ and $\hat{\alpha}_3(t_k)$ from (4.19), and we define

$$(4.21) \quad \hat{F}(t, x) := -\sigma(2\hat{\alpha}_2(t_k) + 8\hat{\alpha}_3(t_k)x) \text{ for } t \in [t_{k-1}, t_k], \quad k = 1, \dots, N,$$

which we use for variance reduction by putting $F = \hat{F}$ in (4.5). In the experiments we simulate (4.5) with $F = \hat{F}(t, x)$ exactly (see (4.14)). The new estimator for $u(0, 0)$ has the form (4.15) again but with the new Z_N corresponding to the choice of $\hat{F}(t, x)$ from (4.21).

TABLE 1

Heat equation. Simulation of $u(0, 0)$ for $\sigma = 1$ and $T = 10$ by (4.15) with the corresponding choice of the function F and for various M . The time step $h = 0.1$ and $M_r = M$. The exact value is $u(0, 0) = 10$. The value after “ \pm ” equals two standard deviations of the corresponding estimator and gives the confidence interval for the corresponding value with probability 0.95 (i.e., $c = 2$).

M	$F = 0$	$F = \hat{F}$ from (4.13)	$F = \hat{F}$ from (4.21)
10^3	9.67 ± 0.85	9.993 ± 0.045	9.999 ± 0.101
10^4	9.92 ± 0.28	9.9970 ± 0.0058	9.999 ± 0.012
10^5	9.970 ± 0.089	10.0000 ± 0.0003	10.0014 ± 0.0014

Table 1 gives some results of simulating $u(0, 0)$ by (4.15) with $F = 0$, $F = \hat{F}$ from (4.13), and $F = \hat{F}$ from (4.21). We see that for $F = 0$ the MC error is consistent with (4.7); i.e., it decreases $\sim 1/\sqrt{M}$. When the variance reduction is used, the results in Table 1 approve the MC error estimate (4.16). It is quite obvious that \hat{F} from (4.13) is a more accurate estimator for the exact $F = -2\sigma x$ than \hat{F} from (4.21), and then the MC error in the first case should usually be less than in the second case, which is observed in the experiments as well.

We also did similar experiments in the case of the terminal condition $u(T, x) = x^4$ in (4.1). To estimate $\partial u / \partial x$ by regression, we took the basis consisting of the first four Hermite polynomials. The results were analogous to those given above for the case x^2 .

4.2. Ergodic limit for one-dimensional array of stochastic oscillators.

Consider the one-dimensional array of oscillators [23, 19]:

(4.22)

$$dP^i = -V'(Q^i) dt - \lambda \cdot (2Q^i - Q^{i+1} - Q^{i-1}) dt - \nu P^i dt + \sigma dw_i(t), \quad P^i(0) = p^i,$$

$$dQ^i = P^i dt, \quad Q^i(0) = q^i, \quad i = 1, \dots, n,$$

where periodic boundary conditions are assumed, i.e., $Q^0 := Q^n$ and $Q^{n+1} := Q^1$; $w_i(t)$, $i = 1, \dots, n$, are independent standard Wiener processes; $\nu > 0$ is a dissipation parameter; $\lambda \geq 0$ is a coupling constant; σ is the noise intensity; and $V(z)$, $z \in \mathbf{R}$, is a potential.

The SDEs (4.22) are ergodic with the Gibbs invariant measure μ . We are interested in computing the average of the potential energy with respect to the invariant measure associated with (4.22):

$$E_\mu U(Q) = E_\mu \sum_{i=1}^n \left(V(Q^i) + \frac{\lambda}{2} \cdot (Q^i - Q^{i+1})^2 \right).$$

To this end (see further details in [19]), we simulate the system (4.22) on a long time interval and approximate the ergodic limit $E_\mu U(Q)$ by $EU(Q(T))$ for a large T . To illustrate variance reduction via regression, we simulate

(4.23)
$$u(0, p, q) = EU(Q_{p,q}(T)) = E[U(Q_{p,q}(T)) + Z_{p,q}(T)],$$

where $Z(t)$, $0 \leq t \leq T$, satisfies

(4.24)
$$dZ = F^\top(t, P, Q)dw(t), \quad Z(0) = 0.$$

We choose the n -dimensional vector function $F(t, p, q)$ to be equal to (see (2.14))

(4.25)
$$F^i(t, p, q) = -\sigma \frac{\partial \hat{u}}{\partial p^i}, \quad i = 1, \dots, n,$$

where $\hat{u} = \hat{u}(t, p, q)$ is an approximation of the function

$$u(t, p, q) := EU(Q_{t,p,q}(T)).$$

We simulate (4.22) using the second-order weak quasi-symplectic integrator from [15, 16]:

(4.26)
$$P_0 = p, \quad Q_0 = q,$$

$$\mathcal{P}_{1,k}^i = e^{-\nu h/2} P_k^i, \quad \mathcal{Q}_{1,k}^i = Q_k^i + \frac{h}{2} P_{1,k}^i,$$

$$\mathcal{P}_{2,k}^i = \mathcal{P}_{1,k}^i + h \left\{ -V'(\mathcal{Q}_{1,k}^i) - \lambda \cdot (2\mathcal{Q}_{1,k}^i - \mathcal{Q}_{1,k}^{i+1} - \mathcal{Q}_{1,k}^{i-1}) \right\} + h^{1/2} \sigma \xi_{ik},$$

$$P_{k+1}^i = e^{-\nu h/2} \mathcal{P}_{2,k}^i, \quad Q_{k+1}^i = \mathcal{Q}_{1,k}^i + \frac{h}{2} \mathcal{P}_{2,k}^i, \quad i = 1, \dots, n, \quad k = 0, \dots, N-1,$$

where ξ_{ik} are independent and identically distributed random variables with the law

(4.27)
$$P(\xi = 0) = 2/3, \quad P(\xi = \pm\sqrt{3}) = 1/6.$$

And we approximate (4.24) by the standard second-order weak method (see [16, p. 103]):

(4.28)

$$\begin{aligned}
 Z_0 &= 0, \\
 Z_{k+1} &= Z_k + h^{1/2} \sum_{i=1}^n F^i(t_k, P_k, Q_k) \xi_{ik} + \sigma h \sum_{r=1}^n \sum_{i=1}^n \frac{\partial}{\partial p^i} F^r(t_k, P_k, Q_k) \xi_{irk} \\
 &\quad + \frac{1}{2} h^{3/2} \sum_{i=1}^n \mathcal{L} F^i(t_k, P_k, Q_k) \xi_{ik}, \\
 \xi_{irk} &= \frac{1}{2} \xi_{ik} \xi_{rk} - \frac{1}{2} \gamma_{ir} \zeta_{ik} \zeta_{rk}, \quad \gamma_{ir} = \begin{cases} -1, & i < r, \\ 1, & i \geq r, \end{cases} \\
 \mathcal{L} &:= \frac{\partial}{\partial t} + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \frac{\partial^2}{\partial p^i \partial p^j} + \sum_{i=1}^n (-V'(q^i) - \lambda \cdot (2q^i - q^{i+1} - q^{i-1}) - \nu p^i) \frac{\partial}{\partial p^i} \\
 &\quad + \sum_{i=1}^n p^i \frac{\partial}{\partial q^i},
 \end{aligned}$$

where ξ_{ik} and ζ_{jk} are mutually independent random variables, ξ_{ik} are distributed by the law (4.27), and the ζ_{ik} are distributed by the law $P(\zeta = \pm 1) = 1/2$.

We consider two potentials: the harmonic potential

$$(4.29) \quad V(z) = \frac{1}{2} z^2, \quad z \in \mathbf{R},$$

and the hard anharmonic potential

$$(4.30) \quad V(z) = \frac{1}{2} z^2 + \frac{1}{2} z^4, \quad z \in \mathbf{R}.$$

We define the approximation $\hat{u}(t, p, q)$ used in (4.25) at $t = t_k, k = 0, \dots, N - 1$, as follows. First, it is reasonable to put $\partial \hat{u} / \partial p^i(t, p, q) = 0$ for $0 \leq t \leq T_0$ with some relatively small T_0 since for large T the function $u(t, p, q), 0 \leq t \leq T_0$, is almost constant due to the ergodicity (the expectation in (4.23) is almost independent of the initial condition).

Further, let T_0, T, h, N , and a nonnegative integer \varkappa be such that $T_0 = N_0 h, T = N h, N - N_0 = \varkappa N'$, where N_0 and N' are integers. Introduce $\theta_{k'} = t_{N_0+k'\varkappa}, k' = 1, \dots, N'$.

In the case of harmonic potential the required function $u(t, p, q)$ can be expanded in the basis consisting of the finite number of functions

$$(4.31) \quad \varphi_l \in \{1, p^i, q^i, p^i p^j, q^i q^j, p^i q^j, \quad i, j = 1, \dots, n\}.$$

In our experiments we deal with three oscillators ($n = 3$); the basis (4.31) in this case has 28 functions.

We use the set of functions (4.31) as a set of basis functions for regression in both cases of harmonic and hard anharmonic potentials. Namely, using regression as described in section 3.2, we construct the estimator $\hat{u}(\theta_{k'}, p, q)$ for $u(\theta_{k'}, p, q)$ as

$$(4.32) \quad \hat{u}(\theta_{k'}, p, q) = \sum_{l=1}^L \hat{\alpha}_l(\theta_{k'}) \varphi_l(p, q),$$

where φ_l are defined in (4.31) and $\hat{\alpha}_l(\theta_{k'})$ satisfy the system of linear algebraic equations (2.24). The matrix formed from $\hat{\alpha}_l(\theta_{k'})$ is positive definite, and we solve the system of linear algebraic equations by Cholesky decomposition. To find the estimator \hat{u} , we use M_r independent trajectories.

Then for $T_0 < t_k < T$ we put $\hat{u}(t_k, p, q) = \hat{u}(\theta_{k'}, p, q)$ with $\theta_{k'} \leq t_k < \theta_{k'+1}$. The recalculation of the estimator \hat{u} once per a few number of steps \varkappa reduces the cost of the procedure.

We note that for the basis (4.31) the corresponding function F from (4.25) is such that some terms in the scheme (4.28) are canceled; in particular, it is not required to simulate the ζ_{ik} in this case.

We compute $u(0, p, q)$ in the usual way,

$$(4.33) \quad u(0, p, q) = E [U(Q_{p,q}(T)) + Z_{p,q}(T)] \approx E [U(Q_N) + Z_N] \\ \approx \frac{1}{M} \sum_{m=1}^M [U(mQ_N) + {}_mZ_N],$$

by simulating M independent realizations of Q_N , Z_N from (4.26), (4.28). In these experiments the two-run procedure described in section 3.5 was used.

Suppose we would like to compute $u(0, p, q)$ for the particular set of parameters $n = 3$, $\lambda = 1$, $\nu = 1$, $\sigma = 1$, $T = 10$ and the potentials (4.29) and (4.30) with accuracy of order 10^{-3} . Since we are using the scheme of order two, we can take $h = 0.02$.

Let us first consider the case of harmonic potential (4.29). Without variance reduction (i.e., for $F = 0$), we obtain 0.7500 ± 0.0010 with the fiducial probability 95% by simulating $M = 1.4 \times 10^6$ trajectories, taking ~ 541 sec on a PC. When we use the variance reduction technique as described above, it is sufficient to take $T_0 = 2$, $\varkappa = 2$, $M_r = 2 \times 10^4$, $M = 3 \times 10^4$ to get 0.7496 ± 0.0010 in ~ 64 sec. In this example the procedure with variance reduction requires an eighth of the computational time. All the expenses are taken into account, including the time required for the first run of the two-run procedure, which is less than 10% of the total time. We recall that in this case the required function $u(t, p, q)$ can be expanded in the finite basis (4.31), unlike the case of hard anharmonic potential when such a basis is infinite.

Now consider the case of hard anharmonic potential (4.30). Without variance reduction (i.e., for $F = 0$), we obtain 0.6491 ± 0.0011 with the fiducial probability 95% by simulating $M = 10^6$ trajectories, taking ~ 403 sec on a PC. With variance reduction, we reach the same level of accuracy 0.6491 ± 0.0011 in ~ 98 sec by choosing, e.g., $T_0 = 2$, $\varkappa = 2$, $M_r = 2.5 \times 10^4$, $M = 5.5 \times 10^4$. Thus, the procedure with variance reduction requires a quarter of the computational time.

Some other results of our numerical experiments are presented in Tables 2 and 3. They show dependence of the MC error on M and M_r . The numerical integration error is relatively small here and does not essentially affect the results. The case $M_r = 0$ means that the simulation was done without variance reduction. We observe that in both tables for a fixed M_r the MC error decreases $\sim 1/\sqrt{M}$. Further, we see from Table 2 that the MC error is $\sim 1/\sqrt{M_r}$ for fixed M (for $M_r > 0$, of course), and, consequently, it is $\sim 1/\sqrt{MM_r}$ when the variance reduction is used (we recall that the time step is relatively small here). As noted before, the basis used in the variance reduction is such that the function $u(t, x)$ can be expanded in it in the case of harmonic potential; i.e., err_B in (4.17) is equal to 0. These observations are consistent with the MC error estimate (4.17). For the anharmonic potential, err_B is not equal to zero, and we see in Table 3 that the increase of M_r has less impact on the MC error in this case.

TABLE 2

Harmonic potential. Two standard deviations of the estimator (4.33) in the case of potential (4.29) for different M and M_r . $M_r = 0$ means that variance reduction was not used. The other parameters are $n = 3, \lambda = 1, \nu = 1, \sigma = 1, T = 10$ and $h = 0.01, T_0 = 2, \varkappa = 1$.

	$M_r = 0$	$M_r = 10^3$	$M_r = 10^4$	$M_r = 10^5$
$M = 10^3$	4.0×10^{-2}	2.6×10^{-2}	--	--
$M = 10^4$	1.2×10^{-2}	7.8×10^{-3}	2.3×10^{-3}	--
$M = 10^5$	3.9×10^{-3}	2.3×10^{-3}	7.9×10^{-4}	2.5×10^{-4}
$M = 10^6$	1.2×10^{-3}	8.2×10^{-4}	2.4×10^{-4}	7×10^{-5}

TABLE 3

Hard anharmonic potential. Two standard deviations of the estimator (4.33) in the case of potential (4.30) for different M and M_r . The other parameters are the same as in Table 2.

	$M_r = 0$	$M_r = 10^3$	$M_r = 10^4$	$M_r = 10^5$
$M = 10^3$	3.3×10^{-2}	2.3×10^{-2}	--	--
$M = 10^4$	1.1×10^{-2}	7.4×10^{-3}	3.0×10^{-3}	--
$M = 10^5$	3.5×10^{-3}	2.4×10^{-3}	9.5×10^{-4}	6.7×10^{-4}
$M = 10^6$	1.1×10^{-3}	7.4×10^{-4}	2.9×10^{-4}	2.2×10^{-4}

4.3. Pricing a binary asset-or-nothing call option. Consider the Black-Scholes equation for pricing a binary asset-or-nothing call option:

$$(4.34) \quad \frac{\partial u}{\partial t} + \frac{\nu^2}{2} x^2 \frac{\partial^2 u}{\partial x^2} + r x \frac{\partial u}{\partial x} - r u = 0, \quad 0 \leq t < T, \quad x \in \mathbf{R},$$

$$u(T, x) = f(x) = \begin{cases} 0 & \text{if } x < K, \\ x & \text{if } x \geq K. \end{cases}$$

The solution of this problem for $x > 0$ and $K > 0$ is

$$(4.35) \quad u(t, x) = x \Phi(y_*),$$

where

$$y_* = \frac{1}{\nu \sqrt{T-t}} \left[\ln \frac{x}{K} + \left(r + \frac{\nu^2}{2} \right) (T-t) \right] \quad \text{and} \quad \Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y e^{-z^2/2} dz.$$

The probabilistic representation (with $\mu = 0$) of the solution to (4.34) takes the form

$$(4.36) \quad u(s, x) = E \left[f(X_{s,x}(T)) e^{-r(T-s)} + Z_{s,x}(T) \right],$$

$$(4.37) \quad dX = rX dt + \nu X dw(t), \quad X(s) = x,$$

$$(4.38) \quad dZ = F(t, X) e^{-r(t-s)} dw(t), \quad Z(s) = 0.$$

The purpose of this example is to illustrate that the approach to evaluating $u(s, x)$ introduced in section 3.2 works, in principle, in the case of discontinuous initial conditions $f(x)$. We use, as a set of basis functions for regression, the set consisting of three functions:

$$(4.39) \quad \varphi_1(x) = \frac{K}{\pi} (\arctan(\alpha(x - K)) + \arctan(\alpha K)),$$

$$\varphi_2(x) = \frac{x}{2} + \frac{x(x - 2K)}{4(\sqrt{(x - K)^2/4 + \beta} + \sqrt{K^2/4 + \beta})}, \quad \varphi_3(x) = \frac{x}{\gamma + x^2},$$

where $\alpha > 0$, $\beta > 0$, and $\gamma > 0$ are parameters, which can change from one time layer to another. We note that the functions are chosen so that $\varphi_l(0) = 0$, $l = 1, 2, 3$, and the payoff $f(x)$ is well approximated by $\varphi_1(x) + \varphi_2(x)$ with large α and small β .

In the experiments, we take the volatility $\nu = 0.2$, the interest rate $r = 0.02$, and the maturity time $T = 3$ and approximate the option price $u(0, 1)$, whose exact value due to (4.35) is $u(0, 1) \approx 0.63548$. We define the time-dependent $\alpha = \alpha(t)$ and $\beta = \beta(t)$ via linear interpolation:

$$\alpha(t) = \frac{10t}{T} + \frac{0.01(T - t)}{T}, \quad \beta(t) = \frac{0.0001t}{T} + \frac{0.005(T - t)}{T},$$

and we choose $\gamma = 8$. We simulate (4.37)–(4.38) using the weak Euler scheme with time step $h = T/N = 0.001$. In the first run (see section 3.5 for the description of the algorithm), we put $F = 0$ and store the values $f(m\bar{X}(T))e^{-rT}$, which are needed for the regression in the second run. In the second run, using regression with the set of basis functions (4.39), we construct the estimator $\hat{u}(\theta_{k'}, x)$ for $u(\theta_{k'}, x)$, where $\theta_{k'} = \varkappa k' h$, $k' = 1, \dots, N'$; \varkappa and N' are nonnegative integers such that $\varkappa N' h = T$. We use here $\varkappa = 5$; i.e., we recalculate the estimator \hat{u} only once per five time layers to reduce the computational cost. Further, $\hat{u}(t_k, x)$ is set equal to zero for $0 \leq t_k < 0.01$. In the second run we put $F(t, x) = -\nu \partial \hat{u} / \partial x$. In both runs we simulate $M = 4 \cdot 10^4$ independent trajectories. As a result, we get $u(0, 1) \approx \bar{u}(0, 1) = 0.6358 \pm 0.0018$ with the fiducial probability 95%. To achieve a similar result without variance reduction, namely, $\bar{u}(0, 1) = 0.6342 \pm 0.0019$, one has to simulate $M = 5 \cdot 10^5$ independent trajectories, which requires at least three times more computational time than the procedure with variance reduction. This experiment demonstrates that the simple and cheap estimation of $\partial u / \partial x$ by $\partial \hat{u} / \partial x$ works even in the case of discontinuous initial conditions.

5. Conclusions. Starting an MC simulation, first of all we have to estimate the number of trajectories required to reach a prescribed accuracy. Fortunately, we can easily do this because a reliable estimate of the variance can be obtained by a preliminary numerical experiment using a relatively small set of trajectories. If the required number of trajectories is too large, we run inevitably into the problem of variance reduction. The known variance reduction methods (the method of importance sampling, the method of control variates, and the combining method) are based on the assumption that approximations of the solution $u(t, x)$ of the considered problem and its spatial derivatives $\partial u(t, x) / \partial x^i$ are known. In this paper we proposed to construct such approximations as a part of the MC simulation using conditional probabilistic representations together with the regression method and thus make the variance reduction methods practical. The basis used in the regression method can be chosen using some a priori knowledge of the considered problems, as illustrated in the examples.

As is known (see, e.g., [16]), the variance reduction methods are applicable in the case of boundary value problems for parabolic and elliptic equations as well. Although here we illustrated the proposed implementation of these variance reduction methods for the Cauchy problems for parabolic equations, the approach is straightforwardly applicable to boundary value problems.

We also note that the proposed technique of conditional probabilistic representations together with regression can be used for evaluating different Greeks for American- and Bermudan-type options (see [1]).

REFERENCES

- [1] D. BELOMESTNY, G. N. MILSTEIN, AND J. G. M. SCHOENMAKERS, *Sensitivities for Bermudan Options by Regression Methods*, WIAS preprint 1247, WIAS, Berlin, 2007.
- [2] B. BOUCHARD, I. EKELAND, AND N. TOUZI, *On the Malliavin approach to Monte Carlo approximation of conditional expectations*, *Finance Stoch.*, 8 (2004), pp. 45–71.
- [3] J. FAN AND I. GIJBELS, *Local Polynomial Modelling and Its Applications*, Chapman & Hall, London, 1996.
- [4] E. FOURNIÉ, J.-M. LASRY, J. LEBUCHOUX, AND P.-L. LIONS, *Application of Malliavin calculus to Monte Carlo methods in finance II*, *Finance Stoch.*, 5 (2001), pp. 201–236.
- [5] E. FOURNIÉ, J.-M. LASRY, J. LEBUCHOUX, P.-L. LIONS, AND N. TOUZI, *Application of Malliavin calculus to Monte Carlo methods in finance*, *Finance Stoch.*, 3 (1999), pp. 391–412.
- [6] S. A. GLADYSHEV AND G. N. MILSTEIN, *The Runge-Kutta method for calculation of Wiener integrals of functionals of exponential type*, *Zh. Vychisl. Mat. i Mat. Fiz.*, 24 (1984), pp. 1136–1149.
- [7] P. GLASSERMAN, *Monte Carlo Methods in Financial Engineering*, Springer-Verlag, New York, 2004.
- [8] L. GYÖRFI, M. KOHLER, A. KRZYŻAK, AND H. WALK, *A Distribution-Free Theory of Nonparametric Regression*, Springer-Verlag, New York, 2002.
- [9] A. KEBAIER, *Statistical Romberg extrapolation: A new variance reduction method and applications to option pricing*, *Ann. Appl. Probab.*, 15 (2005), pp. 2681–2705.
- [10] A. KOHATSU-HIGA AND R. PETERSSON, *Variance reduction methods for simulation of densities on Wiener space*, *SIAM J. Numer. Anal.*, 40 (2002), pp. 431–450.
- [11] N. V. KRYLOV, *Controllable Processes of Diffusion Type*, Nauka, Moscow, 1977.
- [12] G. N. MILSTEIN, *Numerical Integration of Stochastic Differential Equations*, Ural State University, Sverdlovsk, 1988 (in Russian); English translation: Kluwer Academic, Dordrecht, The Netherlands, 1995.
- [13] G. N. MILSTEIN AND J. G. M. SCHOENMAKERS, *Monte Carlo construction of hedging strategies against multi-asset European claims*, *Stoch. Stoch. Rep.*, 73 (2002), pp. 125–157.
- [14] G. N. MILSTEIN, J. G. M. SCHOENMAKERS, AND V. SPOKOINY, *Transition density estimation for stochastic differential equations via forward-reverse representations*, *Bernoulli*, 10 (2004), pp. 281–312.
- [15] G. N. MILSTEIN AND M. V. TRETYAKOV, *Quasi-symplectic methods for Langevin-type equations*, *IMA J. Numer. Anal.*, 23 (2003), pp. 593–626.
- [16] G. N. MILSTEIN AND M. V. TRETYAKOV, *Stochastic Numerics for Mathematical Physics*, Springer-Verlag, Berlin, 2004.
- [17] G. N. MILSTEIN AND M. V. TRETYAKOV, *Numerical analysis of Monte Carlo evaluation of Greeks by finite differences*, *J. Comput. Finance*, 8 (2005), pp. 1–33.
- [18] G. N. MILSTEIN AND M. V. TRETYAKOV, *Numerical integration of stochastic differential equations with nonglobally Lipschitz coefficients*, *SIAM J. Numer. Anal.*, 43 (2005), pp. 1139–1154.
- [19] G. N. MILSTEIN AND M. V. TRETYAKOV, *Computing ergodic limits for Langevin equations*, *Phys. D*, 229 (2007), pp. 81–95.
- [20] N. NEWTON, *Continuous-time Monte Carlo methods and variance reduction*, in *Numerical Methods in Finance*, L. C. G. Rodgers and D. Talay, eds., Cambridge University Press, Cambridge, UK, 1997, pp. 22–42.
- [21] N. J. NEWTON, *Variance reduction for simulated diffusions*, *SIAM J. Appl. Math.*, 54 (1994), pp. 1780–1805.
- [22] D. NUALART, *The Malliavin Calculus and Related Topics*, Springer-Verlag, Berlin, 2006.
- [23] R. REIGADA, A. H. ROMERO, A. SARMIENTO, AND K. LINDENBERG, *One-dimensional arrays of oscillators: Energy localization in thermal equilibrium*, *J. Chem. Phys.*, 111 (1999), pp. 1373–1384.
- [24] W. WAGNER, *Monte Carlo evaluation of functionals of solutions of stochastic differential equations. Variance reduction and numerical examples*, *Stoch. Anal. Appl.*, 6 (1988), pp. 447–468.
- [25] G. ZOU AND R. D. SKEEL, *Robust variance reduction for random walk methods*, *SIAM J. Sci. Comput.*, 25 (2004), pp. 1964–1981.

A DOMAIN DECOMPOSITION METHOD FOR COMPUTING BIVARIATE SPLINE FITS OF SCATTERED DATA*

MING-JUN LAI[†] AND LARRY L. SCHUMAKER[‡]

Abstract. A domain decomposition method for solving large bivariate scattered data fitting problems with bivariate minimal energy, discrete least-squares, and penalized least-squares splines is described. The method is based on splitting the domain into smaller domains, solving the associated smaller fitting problems, and combining the coefficients to get a global fit. Explicit error bounds are established for how well our locally constructed spline fits approximate the global fits. Some numerical examples are given to illustrate the effectiveness of the method.

Key words. computation of bivariate splines, scattered data fitting

AMS subject classifications. 41A63, 41A15, 65D07

DOI. 10.1137/070710056

1. Introduction. Suppose f is a smooth function defined on a domain Ω in \mathbb{R}^2 with polygonal boundary. Given the values $\{f_i := f(x_i, y_i)\}_{i=1}^{n_d}$ of f at some set of scattered points in Ω , we consider the problem of computing a function s that interpolates the data, or in the case of noisy data or large sets of data, approximates rather than interpolates f . There are many methods for solving this problem, but here we will focus on three methods based on bivariate splines, namely,

- the minimal energy (ME) method,
- the discrete least-squares (DLS) method,
- the penalized least-squares (PLS) method.

These three variational methods have been extensively studied in the literature; see [1, 6, 7, 8, 12] and the references therein. It is well known that all three do a good job of fitting smooth functions. But they are global methods, which means that the coefficients of a fitting spline are computed from a single linear system of equations, which can be very large if the dimension of the spline space is large. This would appear to limit the applicability of variational spline methods to moderately sized problems. However, as we shall show in this paper, it is possible to efficiently compute ME-, DLS-, and PLS-splines, even with spline spaces of very large dimension.

Suppose that Δ is a triangulation of Ω , and that $\mathcal{S}(\Delta)$ is a spline space defined on Δ . Throughout this paper we assume that $\mathcal{S}(\Delta)$ has a stable local minimal determining set \mathcal{M} ; see section 4 or the book [10]. This means that each spline $s \in \mathcal{S}(\Delta)$ is uniquely determined by a set of coefficients $\{c_\xi\}_{\xi \in \mathcal{M}}$, where each c_ξ is associated with a unique (domain) point ξ of Δ .

The idea of our method is simple. Instead of finding all of the coefficients $\{c_\xi\}_{\xi \in \mathcal{M}}$ at once, this algorithm reduces the problem to a collection of smaller problems. To state our algorithm formally, we need some additional notation. If ω is a subset of Ω ,

*Received by the editors December 4, 2007; accepted for publication (in revised form) July 10, 2008; published electronically February 13, 2009.

<http://www.siam.org/journals/sinum/47-2/71005.html>

[†]Department of Mathematics, University of Georgia, Athens, GA 30602 (mjlai@math.uga.edu). This author's research was partially supported by the National Science Foundation under grant 0713807.

[‡]Department of Mathematics, Vanderbilt University, Nashville, TN 37240 (larry.schumaker@vanderbilt.edu).

we set $\text{star}^0(\omega) = \bar{\omega}$, and for all $\ell \geq 1$, recursively define

$$\text{star}^\ell(\omega) := \bigcup \{T \in \Delta : T \cap \text{star}^{\ell-1}(\omega) \neq \emptyset\}.$$

ALGORITHM 1.1 (domain decomposition method).

- 1) Choose a decomposition of Ω into disjoint connected sets $\{\Omega_i\}_{i=1}^m$.
- 2) Choose $k > 0$. For each $i = 1, \dots, m$, let $s_i^k \in \mathcal{S}(\Delta)|_{\Omega_i^k}$ be the spline fit based on data in $\Omega_i^k := \text{star}^k(\Omega_i)$. Let $\{c_{i,\xi}^k\}$ be the set of all coefficients of s_i^k .
- 3) For each $i = 1, \dots, m$, set

$$c_\xi = c_{i,\xi}^k \quad \text{for all } \xi \in \mathcal{M} \cap \Omega_i.$$

We call a spline s produced by this algorithm a *domain decomposition (DDC) spline*. We emphasize that this domain decomposition method is very different from domain decomposition methods used in classical numerical algorithms for partial differential equations and in the application of radial basis functions to scattered data fitting and meshless methods for PDE's; see Remark 1. As we shall see, our method

- is easy to implement,
- allows the solution of very large data fitting problems,
- with appropriately chosen m and k , produces a spline which is very close to the globally defined spline,
- is amenable to parallel processing,
- produces a spline s in the space $\mathcal{S}(\Delta)$, i.e., with the same smoothness as the global fit,
- does not make use of blending functions.

The paper is organized as follows. In section 2 we review the basics of minimal energy, discrete least-squares, and penalized least-squares spline fitting. Then in section 3 we present some numerical experiments to illustrate the performance of our domain decomposition method. There we also explore the following questions:

- How does the time required to compute a domain decomposition spline s compare with that required for finding a global spline fit s_g from $\mathcal{S}(\Delta)$?
- How does $\|s - s_g\|$ behave as we choose different decompositions and different values for the parameter k ?
- How well does the shape of s match that of s_g ?

In section 4 we review some Bernstein–Bézier tools needed to analyze our method and present two lemmas needed later. In section 5 we show that for the variational spline methods described in the following section, $\|s - s_g\| = \mathcal{O}(\sigma^k)$ for some $0 < \sigma < 1$. We conclude the paper with remarks and references.

2. Three variational spline fitting methods. Given $d > r \geq 1$ and a triangulation Δ of a domain $\Omega \in \mathbb{R}^2$, let

$$\mathcal{S}_d^r(\Delta) := \{s \in C^r(\Omega) : s|_T \in \mathcal{P}_d, \text{ all } T \in \Delta\}$$

be the associated space of bivariate splines of smoothness r and degree d . Here \mathcal{P}_d is the $\binom{d+2}{2}$ dimensional space of bivariate polynomials of degree d . Such spaces, along with various subspaces of so-called supersplines, have been intensely studied in the literature; see the book [10] and references therein. There are many spline-based methods for interpolation and approximation. Here we are interested in three particular methods.

2.1. Minimal energy interpolating splines. Suppose we are given values $\{f_i\}_{i=1}^{n_d}$ associated with a set of $n_d \geq 3$ abscissae $\mathcal{A} := \{(x_i, y_i)\}_{i=1}^{n_d}$ in the plane. The problem is to construct a smooth function s that interpolates this data in the sense that

$$s(x_i, y_i) = f_i, \quad i = 1, \dots, n_d.$$

To solve this problem, suppose Δ is a triangulation with vertices at the points of \mathcal{A} . Let $\mathcal{S}(\Delta)$ be a spline space defined on Δ with dimension $n \geq n_d$, and let

$$\Lambda(f) = \{s \in \mathcal{S}(\Delta) : s(x_i, y_i) = f_i, i = 1, \dots, n_d\}.$$

Let

$$(2.1) \quad E(s) = \int_{\Omega} [(s_{xx})^2 + 2(s_{xy})^2 + (s_{yy})^2] dx dy$$

be the well-known thin-plate energy of s . Then the *minimal energy (ME) interpolating spline* is the function s_E in Λ such that

$$(2.2) \quad E(s_E) = \min_{s \in \Lambda(f)} E(s).$$

Assuming $\Lambda(f)$ is nonempty, it is well known (see, e.g., [1, 6, 12]) that there exists a unique ME-spline characterized by the property

$$(2.3) \quad \langle s_E, g \rangle_E = 0, \quad \text{all } g \in \Lambda(0),$$

where

$$(2.4) \quad \langle \phi, \psi \rangle_E := \int_{\Omega} [\phi_{xx} \psi_{xx} + 2\phi_{xy} \psi_{xy} + \phi_{yy} \psi_{yy}] dx dy.$$

Moreover, its Bernstein–Bézier coefficients can be computed by solving an appropriate linear system of equations. For details on two different approaches to this computation, see [1] and [12].

Assuming the data come from a smooth function, i.e.,

$$(2.5) \quad f_i = f(x_i, y_i), \quad i = 1, \dots, n_d,$$

then it is possible to give an error bound for how well the corresponding minimal energy interpolating spline s_e approximates f . To state the result, suppose the triangulation Δ is β -uniform, i.e.,

$$(2.6) \quad \frac{|\Delta|}{\rho_{\Delta}} \leq \beta < \infty,$$

where $|\Delta|$ is the length of the longest edge in Δ , and ρ_{Δ} is the minimum of the inradii of the triangles of Δ . Let θ_{Δ} be the smallest angle in Δ . Then it was shown in Theorem 6.2 of [6] that for all $f \in W_{\infty}^2(\Omega)$,

$$(2.7) \quad \|f - s_E\|_{\Omega} \leq C |\Delta|^2 |f|_{2, \Omega},$$

where $\|\cdot\|_{\Omega}$ is the supremum norm on Ω , and $|\cdot|_{2, \Omega}$ is the corresponding Sobolev semi-norm. C is a constant depending only on d, ℓ, β , and θ_{Δ} if Ω is convex. If Ω is

nonconvex, the constant C may also depend on the Lipschitz constant of the boundary of Ω .

Now suppose s_E^k is a DDC ME spline computed using Algorithm 1.1 with parameter $k \geq \ell$. Then since the analog of (2.7) holds for each subdomain Ω_i of Ω , we have

$$(2.8) \quad \|s_E - s_E^k\|_\Omega \leq C|\Delta|^2|f|_{2,\Omega}.$$

This shows that the DDC ME spline s_E^k interpolating a given function f is close to the global minimal energy spline s_E whenever f is smooth and $|\Delta|$ is small. The estimate (2.8) does not depend on k , and so gives no information on how the difference behaves with increasing k . In section 5.1 we show that $\|s_E - s_E^k\|_\Omega = \mathcal{O}(\sigma^k)$ with $0 < \sigma < 1$.

2.2. Discrete least-squares splines. When the set of data is very large or the measurements $\{f_i\}_{i=1}^{n_d}$ are noisy, it is often better to construct an approximation from a spline space $\mathcal{S}(\Delta)$ of dimension $n < n_d$. Some or all of the vertices of Δ may be at points in $\mathcal{A} := \{(x_i, y_i)\}_{i=1}^{n_d}$, but they may also be completely different. The solution of the variational problem of minimizing

$$\|s - f\|_{\mathcal{A}}^2 := \sum_{j=1}^{n_d} [s(x_j, y_j) - f_j]^2$$

over all s in $\mathcal{S}(\Delta)$ is called the *discrete least-squares (DLS) spline* s_L . It is well known (see, e.g., [1, 12]) that if $\mathcal{S}(\Delta)$ satisfies the property

$$(2.9) \quad s(x_i, y_i) = 0, \quad i = 1, \dots, n_d, \quad \text{implies } s \equiv 0,$$

then there is a unique DLS spline s_L fitting the data. It is characterized by the property

$$(2.10) \quad \langle s_L - f, g \rangle_{\mathcal{A}} = 0, \quad \text{all } g \in \mathcal{S}(\Delta),$$

where

$$(2.11) \quad \langle \phi, \psi \rangle_{\mathcal{A}} := \sum_{i=1}^{n_d} \phi(x_i, y_i)\psi(x_i, y_i).$$

The Bernstein–Bézier coefficients of s_L can be computed by solving an appropriate linear system of equations. For details on two different approaches to this computation, see [1] and [12].

Assuming the data come from a smooth function, it is possible to give an error bound for how well the least-squares spline s_L approximates f . To state the result, suppose as before that the triangulation Δ is β -uniform. In addition, suppose that the data is sufficiently dense that for some constant $K_1 > 0$,

$$(2.12) \quad K_1 \|s\|_T \leq \left(\sum_{(x_j, y_j) \in T} s(x_j, y_j)^2 \right)^{1/2} \quad \text{for all } s \in \mathcal{S}(\Delta) \text{ and all } T \in \Delta.$$

Let

$$K_2 := \max_{T \in \Delta} \#(\mathcal{A} \cap T).$$

Then for all $f \in W_\infty^{m+1}(\Omega)$ with $0 \leq m \leq d$,

$$(2.13) \quad \|f - s_L\|_\Omega \leq C|\Delta|^{m+1}|f|_{m+1,\Omega};$$

see the remark following Theorem 8.1 in [7]. If Ω is convex, the constant C depends only on $d, \ell, \beta, K_2/K_1$, and θ_Δ . If Ω is nonconvex, C may also depend on the Lipschitz constant of the boundary of Ω .

Now suppose s_L^k is a DDC least-squares spline computed using Algorithm 1.1 with parameter $k \geq \ell$. Then the same error bound holds for each subdomain Ω_i of Ω , and combining with (2.13) gives

$$(2.14) \quad \|s_L - s_L^k\|_\Omega \leq C|\Delta|^{m+1}|f|_{m+1,\Omega}.$$

This shows that the DDC least-squares spline s_L^k fitting measurements of a given function f is close to the global least squares spline s_L whenever f is smooth and $|\Delta|$ is small. The estimate (2.14) does not depend on k , and so gives no information on how the difference behaves with increasing k . In section 5.2 we show that it is $\mathcal{O}(\sigma^k)$ with $0 < \sigma < 1$.

2.3. Penalized least-squares splines. Suppose $\mathcal{A} := \{x_i, y_i\}_{i=1}^{n_d}$ and $\mathcal{S}(\Delta)$ are as in the previous subsections. Fix $\lambda \geq 0$. Then given data values $\{f_i\}_{i=1}^{n_d}$, the corresponding *penalized least-squares (PLS) spline* is defined to be the spline s_λ in $\mathcal{S}(\Delta)$ that minimizes

$$E_\lambda(s) := \|s - f\|_{\mathcal{A}} + \lambda E(s),$$

where $E(s)$ is defined in (2.1). It is well known (cf. [1, 12]) that if \mathcal{S} is a spline space such that (2.9) holds, then there exists a unique PLS spline s_λ minimizing $E_\lambda(s)$ over $s \in \mathcal{S}(\Delta)$. Moreover, s_λ is characterized by

$$(2.15) \quad \langle s_\lambda - f, s \rangle_{\mathcal{A}} + \lambda \langle s_\lambda, s \rangle_E = 0, \quad \text{all } s \in \mathcal{S}(\Delta).$$

As with the other two methods, the Bernstein–Bézier coefficients of s_λ can be computed by solving an appropriate linear system of equations. For details on two different approaches to this computation, see [1] and [12]. It is known [8] that for all $f \in W_\infty^{m+1}$ with $0 \leq m \leq d$,

$$(2.16) \quad \|f - s_\lambda\|_\Omega \leq C(|\Delta|^{m+1}|f|_{m+1,\Omega} + \lambda|f|_{2,\Omega})$$

for λ sufficiently small compared to $|\Delta|$. The constant C depends only on $d, \ell, \beta, \theta_\Delta, K_2/K_1$, and the area of Ω . If Ω is nonconvex, C may also depend on the Lipschitz constant of the boundary of Ω .

Now suppose s_λ^k is a DDC PLS spline computed using Algorithm 1.1 with parameter $k \geq \ell$. Then since the analog of (2.16) holds for each subdomain Ω_i of Ω , we have

$$(2.17) \quad \|s_\lambda - s_\lambda^k\|_\Omega \leq C(|\Delta|^{m+1}|f|_{m+1,\Omega} + \lambda|f|_{2,\Omega}).$$

This shows that the DDC PLS spline s_λ^k fitting a given function f is close to the global PLS spline s_λ whenever f is smooth and $|\Delta|$ is small. The estimate (2.17) does not depend on k , and so gives no information on how the difference behaves with increasing k . In section 5.3 we show that it is $\mathcal{O}(\sigma^k)$ with $0 < \sigma < 1$.

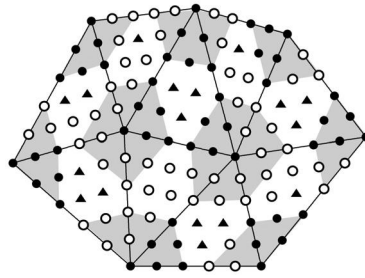


FIG. 1. A minimal determining set for $\mathcal{S}_5^{1,2}(\Delta)$.

3. Numerical examples. In this section we illustrate the domain decomposition method by applying it to compute minimal energy and discrete least-squares fits of scattered data. All of our examples are based on the superspline space

$$\mathcal{S}_5^{1,2}(\Delta) := \{s \in \mathcal{S}_5^1(\Delta) : s \in C^2(v) \text{ for all vertices } v \in \Delta\}.$$

Here $s \in C^2(v)$ means that all polynomial pieces of s on triangles sharing the vertex v have common derivatives up to order 2 at v . It is well known that the dimension of this space is $6n_V + n_E$, where n_V, n_E are the number of vertices and edges of Δ , respectively. The computations in this section are based on the algorithms in [12] which make use of a stable local minimal determining set \mathcal{M} for $\mathcal{S}_5^{1,2}(\Delta)$ and the associated stable local \mathcal{M} -bases defined in [10]. Figure 1 shows a minimal determining set for $\mathcal{S}_5^{1,2}(\Delta)$, where points in the set are marked with black dots and triangles.

3.1. Example 1. Let H be the unit square, and let

$$(3.1) \quad \begin{aligned} F(x, y) = & 0.75 \exp(-0.25(9x - 2)^2 - 0.25(9y - 2)^2) \\ & + 0.75 \exp(-(9x + 1)^2/49 - (9y + 1)/10) \\ & + 0.5 \exp(-0.25(9x - 7)^2 - 0.25(9y - 3)^2) \\ & - 0.2 \exp(-(9x - 4)^2 - (9y - 7)^2) \end{aligned}$$

be the well-known Franke function defined on H ; see Figure 2. Let Δ_{1087} be the triangulation shown in Figure 3. This triangulation has 1087 vertices, 3130 edges, and 2044 triangles. The dimension of the space $\mathcal{S}_5^{1,2}(\Delta_{1087})$ is 9652, and the total number of Bernstein–Bézier coefficients of a spline in this space is 25,871.

First we compute the minimal energy spline fit s_E of f from $\mathcal{S}_5^{1,2}(\Delta_{1087})$. This requires solving a linear system of 8565 equations with 322,989 nonzero entries. Although the largest element in the corresponding matrix is $\mathcal{O}(10^7)$, its condition number is of order $\mathcal{O}(10^4)$. For comparison purposes we computed the maximum error e_∞ over a 160×160 grid, along with the RMS error e_2 over the same grid. These errors are shown in the first line of Table 1, along with the computational time in seconds.

To explore the performance of our DDC technique, we computed approximations of s_E by decomposing Ω into squares $\{\Omega_i\}_{i=1}^{m^2}$ of width $1/m$. In Table 1 we list the results where k is the parameter controlling the size of the sets Ω_i^k in Algorithm 1.1. In addition to the errors e_∞ and e_2 measuring how well s_E fits f , we also tabulate the maximum difference e_∞^c between the coefficients of our DDC spline and the coefficients

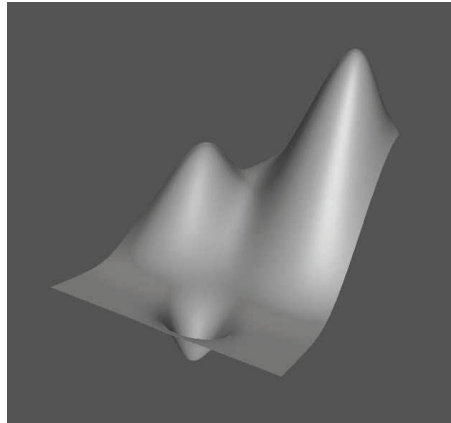


FIG. 2. *The Franke function.*

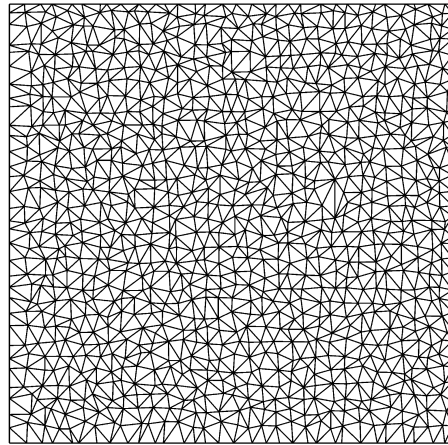
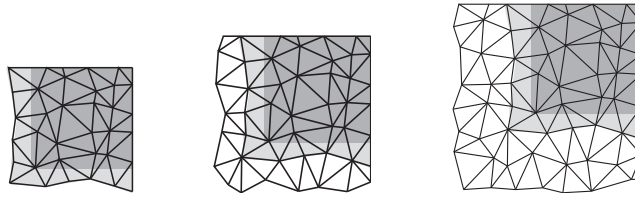


FIG. 3. *A triangulation of 1087 vertices.*

TABLE 1
DDC ME fits to Franke's function from $S_5^{1,2}(\Delta_{1087})$.

m	k	e_∞	e_2	e_∞^c	e_2^c	time
1	0	9.1(-4)	7.7(-5)			25
4	1	3.0(-3)	2.1(-4)	8.5(-3)	9.1(-5)	9
4	2	9.3(-4)	8.6(-5)	3.4(-3)	1.9(-5)	15
4	3	9.1(-4)	7.8(-5)	3.4(-4)	3.0(-6)	21
4	4	9.1(-4)	7.7(-5)	5.4(-5)	4.4(-7)	30
8	1	3.1(-3)	2.7(-4)	8.6(-3)	1.6(-4)	7
8	2	9.2(-4)	9.4(-5)	1.9(-3)	3.5(-5)	16
8	3	9.1(-4)	7.8(-5)	3.4(-4)	7.0(-6)	29
8	4	9.1(-4)	7.7(-5)	8.9(-5)	1.3(-7)	50

FIG. 4. $\text{star}^k(\Omega_{64})$ for $k = 1, 2, 3$.

of the global ME spline s_E . We also compute the RMS difference e_2^c for the coefficients, and list the computational time in seconds. We now comment on these results.

- Accuracy of fit: The table shows that in this experiment, the DDC splines with $k = 1$ do not fit f as well as the ME spline, but as soon as $k \geq 2$, the errors are virtually identical. From the standpoint of accuracy of fit, there is no need to use values of k larger than 2 or 3.
- Accuracy of coefficients: The table shows that the DDC fits also provide very good approximations of the coefficients of the global minimal energy spline s_E . Both e_∞^c and e_2^c decrease as k increases, as predicted by the theoretical results in section 5.1.
- Time: The main use of the DDC algorithm is to make it possible to solve large variational spline problems which could not be solved at all without using the method. For small problems, it often takes more time to solve for a DDC ME spline than for the global ME spline itself. For this moderately sized problem, we see that some of the DDC splines took less time to compute than the global fit, even for the same accuracy. For example, the DDC spline with $m = 8$ and $k = 2$ delivers virtually the same accuracy as the global ME spline, but in only about one half the computing time. For larger problems, the time required to compute DDC ME splines is substantially less than for the global splines; see Example 2.
- Condition numbers: Since the entries in the matrix of the linear systems depend on integrals of squares of second derivatives over triangles, when the triangles are of size $\mathcal{O}(h)$, the entries are of size $\mathcal{O}(h^{-4})$ and even larger if some triangles are very thin. In this example the largest entries are of the order $\mathcal{O}(10^7)$. For very regular triangulations (for example type-I triangulations), the condition numbers of the matrices are of size $\mathcal{O}(10^3)$, independent of how many triangles there are. For less regular triangulations, they can be much larger. However, for the matrices associated with the triangulations in Figure 4, they are of order $\mathcal{O}(10^4)$.
- Shape of star^k : Figure 4 shows $\text{star}^k(\Omega_{64})$ for $k = 1, 2, 3$, where $\Omega_{64} := [.875, 1] \times [.875, 1]$, shown in dark grey in the figure. The white triangles are the triangles added to form the stars.
- Shape of the surface: We have compared 3D plots of the global minimal energy fit of f with the DDC ME fits for the parameters in Table 1. For $k = 1$ we noticed slight deviations in shape, but for all higher values of k we got excellent shapes.

3.2. Example 2. We repeat Example 1 with a type-I triangulation of the unit square with 4225 vertices. This triangulation includes 12,416 edges and 8192 triangles. The dimension of the space $\mathcal{S}_5^{1,2}(\Delta_{4225})$ is 37,776, and the total number of Bernstein–

TABLE 2
 DDC ME fits to Franke's function from $\mathcal{S}_5^{1,2}(\Delta_{4225})$.

m	k	e_∞	e_2	e_∞^c	e_2^c	time
1	0	1.2(-4)	7.6(-6)			326
8	1	9.9(-4)	4.7(-5)	2.2(-3)	2.3(-5)	37
8	2	2.9(-4)	1.5(-5)	6.8(-4)	5.7(-6)	65
8	3	1.8(-4)	9.9(-6)	1.7(-4)	1.4(-6)	97
16	1	9.8(-4)	6.9(-5)	2.3(-3)	4.4(-5)	29
16	2	2.9(-4)	1.9(-5)	7.6(-4)	1.0(-5)	66
16	3	1.8(-4)	1.0(-5)	1.6(-4)	2.5(-6)	128

Bézier coefficients of a spline in this space is 103,041. We again fit the Franke function.

First we compute the minimal energy spline fit s_E of f from $\mathcal{S}_5^{1,2}(\Delta_{4225})$. This requires solving a linear system of 33,541 equations with 1,282,073 nonzero entries. Although the largest element in this matrix is $\mathcal{O}(10^7)$, its condition number is $\mathcal{O}(10^4)$. Our program took 326 seconds to compute s . For comparison purposes, we computed the maximum error e_∞ over a 160×160 grid, along with the RMS error e_2 over this grid. These errors are shown in the first line of Table 2, along with the computational time (in seconds).

We computed approximations of s_E using the same decompositions of Ω as in Example 1 based on m^2 squares of width $1/m$. In Table 2 we list the results. Here we see that using the DDC method results in substantial time savings. We also see that the errors e_∞^c and e_2^c behave like $\mathcal{O}(\sigma^k)$ with $\sigma \approx 1/4$, confirming the theoretical results in section 5.2.

3.3. Example 3. In this example we work with elevation heights measured at 15,585 points in the Black Forest of Germany. The corresponding DeLaunay triangulation Δ_{BF} is shown in Figure 5, although the triangulation is so fine in many areas that it is impossible to see the individual triangles without zooming in. This triangulation has 47,333 edges and 31,449 triangles. The dimension of the space $\mathcal{S}_5^{1,2}(\Delta_{BF})$ is 142,643, and the total number of Bernstein–Bézier coefficients of a spline in this space is 393,911.

The computation of the minimal energy spline fit s_E would require solving a linear system of 126,758 equations, which is beyond the capability of our software. So instead we computed a DDC approximation of the ME spline using the decomposition of Example 1 based on 100 squares. The computation took 288 seconds, and Figure 6 shows the resulting surface.

3.4. Example 4. In this example we again work on the unit square H . This time we approximate Franke's function by least squares based on measured data at 62,500 grid points in H . We approximate from the space $\mathcal{S}_5^{1,2}(\Delta_{1087})$, where Δ_{1087} is the same triangulation as in Example 1; see Figure 3. We choose this triangulation since it is big enough to illustrate how the DDC method works, but small enough so that we can compute the global least square spline for comparison purposes. This function can of course fit very well with much smaller spline spaces and much less data. For example, with a type-I triangulation with 81 vertices and 1089 grid data, the errors for the least-squares spline fit are $e_\infty = 5.2(-4)$ and $e_2 = 5.0(-5)$. The results of our experiments are shown in Table 3. Note that the times of computation for least-squares splines are significantly greater than for the ME splines reported in Table 1. This is due to the fact that a large part of the computation is taken up with finding the triangles containing the various data points. These times can be reduced

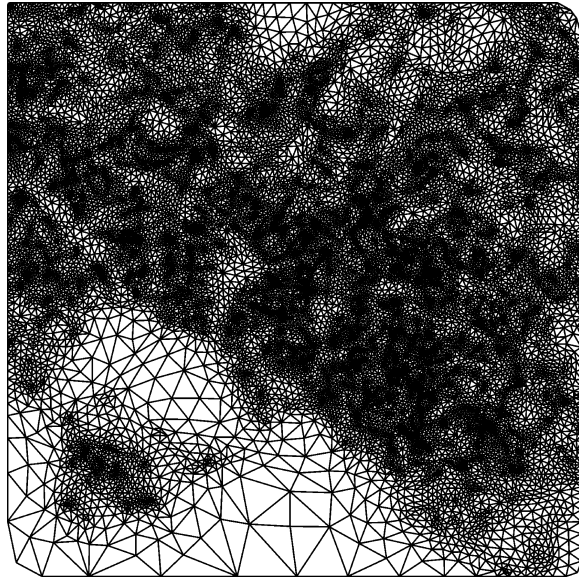


FIG. 5. *Triangulation of 15,585 points in the Black Forest.*

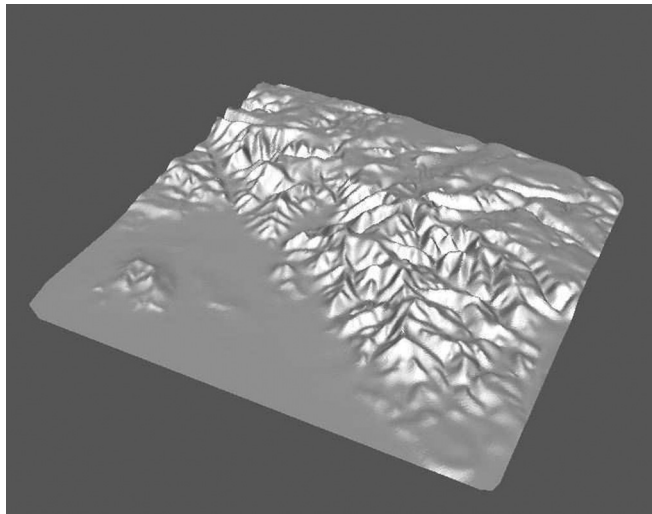


FIG. 6. *The minimal energy interpolant of the Black Forest data.*

TABLE 3
DDC least-squares fits to Franke’s function from $\mathcal{S}_5^{1,2}(\Delta_{BF})$.

m	k	e_∞	e_2	e_∞^c	e_2^c	time
1	0	4.5(-7)	2.3(-8)			42
4	1	4.7(-6)	7.1(-8)	1.9(-5)	2.1(-8)	44
4	2	3.8(-6)	5.3(-8)	5.6(-6)	1.0(-8)	62
4	3	9.9(-7)	3.2(-8)	1.7(-6)	5.5(-9)	82
8	1	5.5(-6)	1.1(-7)	2.0(-5)	4.3(-8)	48
8	2	3.8(-6)	8.0(-8)	1.1(-5)	2.2(-8)	93
8	3	1.7(-6)	6.8(-8)	3.9(-6)	1.7(-8)	151
10	2	2.5(-6)	9.8(-8)	5.3(-6)	2.8(-8)	113

by incorporating standard techniques for reducing the time required for these search operations.

- Accuracy of fit: Table 3 shows that in this experiment the DDC least-squares splines with $k = 1$ do not fit f quite as well as the global least-squares spline, but with increasing k they come very close. As with the minimal energy case, it appears that a good choice might be $k = 2$.
- Accuracy of coefficients: The table shows that the DDC fits also provide very good approximations of the coefficients of the global least-squares spline. Both e_∞^c and e_2^c decrease as k increases. Indeed, for $m = 4$, the error of e_∞^c behaves like $\mathcal{O}(\sigma^k)$ with $\sigma \approx 1/4$, while for $m = 8$, $\sigma \approx 1/2$. There is a similar effect for e_2 , confirming the theoretical results in section 5.2.
- Time: The main use of the DDC algorithm is to make it possible to solve large variational spline problems which could not be solved at all without using the method. For small problems, it can take more time to solve for a DDC least-squares spline than for the global least-squares spline itself. However, even for this moderately sized problem, we see that most of the DDC splines took less time to compute for nearly the same accuracy.
- Condition numbers: The condition numbers of the Gram matrix arising in DLS fitting with splines is dependent on a number of things. The size of β (which reflects whether there are skinny triangles in Δ) plays a role, but not as large a role as in the ME case (since here we are not working with second derivatives). What seems more critical in the least-squares case is the distribution of data over the triangles—if there are triangles with barely enough data to ensure a nonsingular system, the condition number tends to be high. For this particular example, the condition numbers of the matrices arising in the subproblems lie in the range of 10^5 to 10^6 .
- Shape of the surface: We have compared 3D plots of the global least-squares fit of f with the DDC least-squares fits for the parameters in Table 3. For $k = 1$ we noticed slight deviations in shape, but for all higher values of k we got excellent shapes.

4. Analytical tools. In this section we set the stage for the proofs in section 5 of our main results.

4.1. Bernstein–Bézier techniques. We make use of the Bernstein–Bézier representation of splines. Given d and Δ , let $\mathcal{D}_{d,\Delta} := \cup_{T \in \Delta} \mathcal{D}_{d,T}$ be the corresponding set of domain points, where for each $T := \langle v_1, v_2, v_3 \rangle$,

$$\mathcal{D}_{d,T} := \left\{ \xi_{ijk}^T := \frac{iv_1 + jv_2 + kv_3}{d} \right\}_{i+j+k=d}.$$

Then every spline $s \in \mathcal{S}_d^0(\Delta)$ is uniquely determined by its set of coefficients $\{c_\xi\}_{\xi \in \mathcal{D}_{d,\Delta}}$, and

$$s|_T := \sum_{\xi \in \mathcal{D}_{d,T}} c_\xi B_\xi^T,$$

where $\{B_\xi^T\}$ are the Bernstein basis polynomials associated with the triangle T .

Suppose now that $\mathcal{S}(\Delta)$ is a subspace of $\mathcal{S}_d^0(\Delta)$. Then a set $\mathcal{M} \subseteq \mathcal{D}_{d,\Delta}$ of domain points is called a *minimal determining set (MDS)* for $\mathcal{S}(\Delta)$ provided it is the smallest set of domain points such that the corresponding coefficients $\{c_\xi\}_{\xi \in \mathcal{M}}$ can be set independently, and all other coefficients of s can be consistently determined from smoothness conditions, i.e., in such a way that all smoothness conditions are satisfied (see p. 136 of [10]). The dimension of $\mathcal{S}(\Delta)$ is then equal to the cardinality of \mathcal{M} . Clearly, $\mathcal{M} = \mathcal{D}_{d,\Delta}$ is a minimal determining set for $\mathcal{S}_d^0(\Delta)$, and thus the dimension of $\mathcal{S}_d^0(\Delta)$ is $n_V + (d - 1)n_E + \binom{d-1}{2}n_T$, where n_V, n_E, n_T are the number of vertices, edges, and triangles of Δ .

For each $\eta \in \mathcal{D}_{d,\Delta} \setminus \mathcal{M}$, let Γ_η be the smallest subset of \mathcal{M} such that c_η can be computed from the coefficients $\{c_\xi\}_{\xi \in \Gamma_\eta}$ by smoothness conditions. Then \mathcal{M} is called ℓ -local provided there exists an integer ℓ not depending on Δ such that

$$(4.1) \quad \Gamma_\eta \subseteq \text{star}^\ell(T_\eta), \quad \text{all } \eta \in \mathcal{D}_{d,\Delta} \setminus \mathcal{M},$$

where T_η is a triangle containing η . \mathcal{M} is said to be *stable* provided there exists a constant K_3 depending only on ℓ and the smallest angle in the triangulation Δ such that

$$(4.2) \quad |c_\eta| \leq K_3 \max_{\xi \in \Gamma_\eta} |c_\xi|, \quad \text{all } \eta \in \mathcal{D}_{d,\Delta} \setminus \mathcal{M}.$$

Suppose \mathcal{M} is a stable local MDS for $\mathcal{S}(\Delta)$. For each $\xi \in \mathcal{M}$, let ψ_ξ be the spline in $\mathcal{S}(\Delta)$ such that $c_\xi = 1$ while $c_\eta = 0$ for all other $\eta \in \mathcal{M}$. Then the splines $\{\psi_\xi\}_{\xi \in \mathcal{M}}$ are clearly linearly independent and form a basis for $\mathcal{S}(\Delta)$. This basis is called the \mathcal{M} -basis for $\mathcal{S}(\Delta)$; see section 5.8 of [10]. It is stable and ℓ -local in the sense that for all $\xi \in \mathcal{M}$,

$$(4.3) \quad \|\psi_\xi\|_\Omega \leq K_4,$$

and

$$(4.4) \quad \text{supp } \psi_\xi \subseteq \text{star}^\ell(T_\xi),$$

where T_ξ is a triangle containing ξ . Here ℓ is the integer constant in (4.1), and the constant K_4 depends only on ℓ and the smallest angle in Δ .

There are many spaces with stable local bases. For example, the spaces $\mathcal{S}_d^0(\Delta)$ have stable local bases with $\ell = 1$. The same is true for the superspline spaces $\mathcal{S}_{4r+1}^{r,2r}(\Delta)$ for all $r \geq 1$. There are also several families of macroelement spaces defined for all $r \geq 1$ with the same property; see [10].

4.2. Two lemmas. For convenience we recall a lemma from [3].

LEMMA 4.1. *Suppose a_0, a_1, \dots , are nonnegative numbers such that*

$$(4.5) \quad \gamma \sum_{j \geq \nu} a_j \leq a_\nu \quad \text{for all } \nu = 0, 1, 2, \dots,$$

for some $0 < \gamma < 1$. Then $a_\nu \leq \frac{1}{\gamma} \sigma^\nu a_0$, where $\sigma := 1 - \gamma$.

We now establish a key lemma whose proof is modelled on the proof of Theorem 3.1 in [7]. Let \mathcal{W} be a space of spline functions defined on a triangulation Δ of Ω with inner product $\langle f, g \rangle_{\mathcal{W}}$ and norm $\|f\|_{\mathcal{W}}^2 := \langle f, f \rangle_{\mathcal{W}}$. Suppose that $\{B_\xi\}_{\xi \in \mathcal{M}}$ is a 1-local basis for \mathcal{W} such that for some constants C_1, C_2 ,

$$(4.6) \quad C_1 \sum_{\xi \in \mathcal{M}} |c_\xi|^2 \leq \left\| \sum_{\xi \in \mathcal{M}} c_\xi B_\xi \right\|_{\mathcal{W}}^2 \leq C_2 \sum_{\xi \in \mathcal{M}} |c_\xi|^2$$

for all coefficient vectors $c := \{c_\xi\}_{\xi \in \mathcal{M}}$.

LEMMA 4.2. *Let ω be a cluster of triangles in Δ , and let $T \in \omega$. Then there exists constants $0 < \sigma < 1$ and C depending only on the ratio C_2/C_1 such that if g is a function in \mathcal{W} with*

$$(4.7) \quad \langle g, w \rangle_{\mathcal{W}} = 0 \quad \text{for all } w \in \mathcal{W} \text{ with } \text{supp}(w) \subseteq \text{star}^k(\omega),$$

for some $k \geq 1$, then

$$(4.8) \quad \|g \cdot \chi_T\|_{\mathcal{W}} \leq C \sigma^k \|g\|_{\mathcal{W}}.$$

Proof. For each $\nu \geq 0$, let

$$\mathcal{M}_\nu^\omega := \{\xi \in \mathcal{M} : \text{supp}(B_\xi) \subseteq \text{star}^\nu(\mathbb{R}^2 \setminus \text{star}^k(\omega))\}.$$

Define $\mathcal{N}_0^\omega := \mathcal{M}_0^\omega$, and let $\mathcal{N}_\nu^\omega := \mathcal{M}_\nu^\omega \setminus \mathcal{M}_{\nu-1}^\omega$, for $\nu \geq 1$. Given $g := \sum_{\xi \in \mathcal{M}} c_\xi B_\xi$, let

$$g_\nu := \sum_{\xi \in \mathcal{M}_\nu^\omega} c_\xi B_\xi, \quad u_\nu := g - g_\nu, \quad a_\nu := \sum_{\xi \in \mathcal{N}_\nu^\omega} c_\xi^2.$$

By (4.6),

$$(4.9) \quad \sum_{j \geq \nu+1} a_j = \sum_{\xi \notin \mathcal{M}_\nu^\omega} c_\xi^2 \leq \frac{\|u_\nu\|_{\mathcal{W}}^2}{C_1},$$

while (4.7) implies $\langle g, u_\nu \rangle_{\mathcal{W}} = 0$. Since $\text{supp}(u_\nu) \cap \bigcup_{\xi \in \mathcal{M}_{\nu-1}^\omega} \text{supp}(B_\xi) = \emptyset$ for $\nu \geq 1$, it follows that

$$(4.10) \quad \begin{aligned} \|u_\nu\|_{\mathcal{W}}^2 &= \langle g - g_\nu, u_\nu \rangle_{\mathcal{W}} = -\langle g_\nu, u_\nu \rangle_{\mathcal{W}} \\ &= -\left\langle \sum_{\xi \in \mathcal{N}_\nu^\omega} c_\xi B_\xi, u_\nu \right\rangle_{\mathcal{W}} \leq \left\| \sum_{\xi \in \mathcal{N}_\nu^\omega} c_\xi B_\xi \right\|_{\mathcal{W}} \|u_\nu\|_{\mathcal{W}}. \end{aligned}$$

Dividing by $\|u_\nu\|_{\mathcal{W}}$ and squaring, then using (4.6), we get

$$\|u_\nu\|_{\mathcal{W}}^2 \leq \left\| \sum_{\xi \in \mathcal{N}_\nu^\omega} c_\xi B_\xi \right\|_{\mathcal{W}}^2 \leq C_2 a_\nu.$$

Combining (4.9) and (4.10) gives

$$(4.11) \quad \sum_{j \geq \nu} a_j \leq \frac{C_1 + C_2}{C_1} a_\nu, \quad \nu \geq 1.$$

Then applying Lemma 4.1 gives

$$a_\nu \leq \frac{(C_1 + C_2)}{C_1} \sigma^{\nu-1} a_1,$$

with $\sigma := C_2/(C_1 + C_2)$. On the other hand,

$$a_1 \leq \sum_{j \geq 0} a_j = \sum_{\xi \in \mathcal{M}} c_\xi^2 \leq \frac{1}{C_1} \|g\|_{\mathcal{W}}^2.$$

Now let q be the smallest integer such that there is a basis function B_ξ in \mathcal{M}_q^ω with $T \subseteq \text{supp}(B_\xi)$. Then by (4.6),

$$\begin{aligned} \|g \cdot \chi_T\|_{\mathcal{W}}^2 &= \left\| \sum_{B_\xi|_T \neq 0} c_\xi B_\xi \right\|_{\mathcal{W}}^2 \leq C_2 \sum_{\xi \notin \mathcal{M}_{q-1}^\omega} c_\xi^2 = C_2 \sum_{j \geq q} a_j \\ &\leq \frac{C_2}{C_1} \left(\frac{C_1 + C_2}{C_1} \right)^2 \sigma^{q-1} \|g\|_{\mathcal{W}}^2. \end{aligned}$$

Since $q \geq k + 1$, we have (4.8). □

5. Dependence of the errors on the parameter k . In this section we examine the difference between global splines and their DDC approximations as a function of the parameter k . We give separate results for ME, DLS, and PLS splines. Throughout the section we assume that Δ is a β -uniform triangulation, and that $\mathcal{S}(\Delta)$ is an associated spline space with a stable local \mathcal{M} -basis.

5.1. Minimal energy interpolating splines. Given a set of measurements $\{f_i\}_{i=1}^{n_d}$ of a function f at the vertices of a triangulation Δ , let s_E be the corresponding minimal energy interpolating spline. Let s_E^k be the DDC ME spline computed using Algorithm 1.1 with parameter k . In (2.8) we showed that if $f \in W_\infty^2(\Omega)$, then $\|s_E - s_E^k\|_\Omega = \mathcal{O}(|\Delta|^2)$. In this section we discuss the dependence of this difference on k .

THEOREM 5.1. *There exists $\sigma \in (0, 1)$ such that for all $f \in W_\infty^2(\Omega)$*

$$(5.1) \quad \|D_x^\alpha D_y^\beta (s_E - s_E^k)\|_\Omega \leq C \sigma^k |\Delta|^{1-\alpha-\beta} |f|_{2,\Omega}$$

for all $0 \leq \alpha + \beta \leq 1$. When Ω is convex, C is a constant depending only on $d, \ell, \beta, \theta_\Delta$, and the area of Ω . When Ω is nonconvex, C also depends on the Lipschitz constant of the boundary of Ω .

Proof. Let Ω_i be one of the subdomains in Algorithm 1.1. In view of the way in which s_E is defined, it suffices to estimate $\|s_E - s_E^k\|_{\Omega_i}$. Let Δ_i^k be the subtriangulation obtained by restricting Δ to $\Omega_i^k := \text{star}^k(\Omega_i)$. Fix $k \geq 1$. We make use of Lemma 4.2 applied to

$$\mathcal{W} = \{s \in \mathcal{S}(\Delta)|_{\Omega_i^k} : s(v) = 0 \text{ for all vertices } v \text{ of } \Delta_i^k\},$$

with the inner product

$$(5.2) \quad \langle \phi, \psi \rangle_{E, \Omega_i^k} := \int_{\Omega_i^k} [\phi_{xx} \psi_{xx} + 2\phi_{xy} \psi_{xy} + \phi_{yy} \psi_{yy}] dx dy.$$

Let $s_{E,\Omega_i^k} := s_E|_{\Omega_i^k}$ be the global ME interpolant of f restricted to Ω_i^k , and let $s_{E,i}^k$ be the ME interpolant of f in the space $\mathcal{S}(\Delta)|_{\Omega_i^k}$. Let $\{B_\xi\}_{\xi \in \mathcal{M}_i^k}$ be a stable 1-local basis for $\mathcal{S}(\Delta)|_{\Omega_i^k}$. It was shown in Corollary 5.3 of [6] that

$$(5.3) \quad C_1|\Delta|^{-2} \sum_{\xi \in \mathcal{M}_i^k} |c_\xi|^2 \leq \left\| \sum_{\xi \in \mathcal{M}_i^k} c_\xi B_\xi \right\|_{E,\Omega_i^k} \leq C_2|\Delta|^{-2} \sum_{\xi \in \mathcal{M}_i^k} |c_\xi|^2,$$

where C_1 and C_2 depend only on d, ℓ , and β . Writing $g := s_{E,\Omega_i^k} - s_{E,i}^k \in \mathcal{W}$, and using the characterization of ME splines, we have

$$(5.4) \quad \langle g, B_\xi \rangle_{E,\Omega_i^k} = 0, \quad \text{all } B_\xi \text{ with } \text{supp}(B_\xi) \subseteq \Omega_i^k.$$

Now suppose T is a triangle in Ω_i where $|g|$ takes its maximum. Since g is a polynomial on T , we can use Lemma 6.1 of [6] and Theorem 1.1 of [10] to get

$$(5.5) \quad \|g\|_{\Omega_i} = \|g\|_T \leq 12|T|^2|g|_{2,\infty,T} \leq C_3|\Delta||g|_{2,2,T} \leq C_3|\Delta|\|g \cdot \chi_T\|_{E,\Omega_i^k},$$

where C_3 depends only on d . In view of (5.3) and (5.4), we can apply Lemma 4.2 to get

$$(5.6) \quad \|g \cdot \chi_T\|_{E,\Omega_i^k} \leq C_4\sigma^k\|g\|_{E,\Omega_i^k} \leq C_4A^{1/2}\sigma^k|g|_{2,\infty,\Omega_i^k},$$

where A is the area of Ω_i^k . Note that C_4 does not depend on $|\Delta|$ since the constant in Lemma 4.2 depends on the ratio $C_2|\Delta|^{-2}/C_1|\Delta|^{-2}$. Now let τ be a triangle where $|g|_{2,\infty,\Omega_i}$ takes its maximum. Then using the Markov inequality, we have

$$(5.7) \quad |g|_{2,\infty,\Omega_i} = |g|_{2,\infty,\tau} \leq \frac{C_5}{|\tau|^2}\|g\|_\tau \leq \frac{C_5}{|\tau|^2}(\|f - s_E\|_\tau + \|f - s_{E,i}^k\|_\tau).$$

Combining the inequalities (5.5)–(5.7) with the error bound (2.7), we get (5.1) for $\alpha = \beta = 0$. To get the result for derivatives, we apply the Markov inequality on a triangle where $\|D_x^\alpha D_y^\beta g\|_\Omega$ takes its maximum value. \square

5.2. DLS splines. Given a set of measurements $\{f_i\}_{i=1}^{n_d}$ of a function f and a triangulation Δ , let s_L be the DLS spline fit of f from $\mathcal{S}(\Delta)$. Let s_L^k be the DDC least-squares spline produced by Algorithm 1.1 with parameter k . In (2.14) we showed that if $f \in W_\infty^{m+1}(\Omega)$, then $\|s_L - s_L^k\|_\Omega = \mathcal{O}(|\Delta|^{m+1})$. In this section we discuss the dependence of this difference on k . The following result gives results for the derivatives of the difference. As is customary in spline theory, the norm here is to be interpreted as the maximum of the supremum norms over the triangles in Δ since the splines s_L and s_L^k may not have derivatives at every point in Ω .

THEOREM 5.2. *There exists $\sigma \in (0, 1)$ such that if $f \in W_\infty^{m+1}(\Omega)$ with $0 \leq m \leq d$, then*

$$(5.8) \quad \|D_x^\alpha D_y^\beta (s_L - s_L^k)\|_\Omega \leq C\sigma^k|\Delta|^{m-\alpha-\beta}|f|_{m+1,\Omega}.$$

for all $0 \leq \alpha + \beta \leq m$. When Ω is convex, C is a constant depending only on d, ℓ, β, K_1, K_2 , and θ_Δ . When Ω is nonconvex, C also depends on the Lipschitz constant of the boundary of Ω .

Proof. Let Ω_i be one of the subdomains in Algorithm 1.1. In view of the way in which s_L is defined, it suffices to estimate the norm of $s_L - s_L^k$ on Ω_i . Let Δ_i^k be the

subtriangulation obtained by restricting Δ to $\Omega_i^k := \text{star}^k(\Omega_i)$. Fix $k \geq 1$. We make use of Lemma 4.2 applied to $\mathcal{W} = \mathcal{S}(\Delta)|_{\Omega_i^k}$ with the inner product

$$(5.9) \quad \langle \phi, \psi \rangle_{\mathcal{A}_i^k} := \sum_{(x_i, y_i) \in \Omega_i^k} \phi(x_i, y_i) \psi(x_i, y_i).$$

Let $s_{L, \Omega_i^k} := s_L|_{\Omega_i^k}$ be the restriction to Ω_i^k of the global least-squares spline fit s_L of f from $\mathcal{S}(\Delta)$, and let $s_{L, i}^k$ be the least-squares spline fit of f from the space $\mathcal{S}(\Delta)|_{\Omega_i^k}$. Let $\{B_\xi\}_{\xi \in \mathcal{M}_i^k}$ be a stable 1-local basis for $\mathcal{S}(\Delta)|_{\Omega_i^k}$. It was shown in Lemma 5.1 of [7] that

$$(5.10) \quad C_1 \sum_{\xi \in \mathcal{M}_i^k} |c_\xi|^2 \leq \left\| \sum_{\xi \in \mathcal{M}_i^k} c_\xi B_\xi \right\|_{\mathcal{A}_i^k} \leq C_2 \sum_{\xi \in \mathcal{M}_i^k} |c_\xi|^2.$$

Writing $g := s_{L, \Omega_i^k} - s_{L, i}^k \in \mathcal{W}$, and using the characterization of least-squares splines, we have

$$(5.11) \quad \langle g, B_\xi \rangle_{\mathcal{A}_i^k} = 0, \quad \text{all } B_\xi \text{ with } \text{supp}(B_\xi) \subseteq \Omega_i^k.$$

Now suppose T is a triangle in Ω_i where $|g|$ takes its maximum. Then using (2.12) and Lemma 4.2 we get

$$(5.12) \quad \|g\|_{\Omega_i} = \|g\|_T \leq \frac{1}{K_1} \|g \cdot \chi_T\|_{\mathcal{A}_i^k} \leq \frac{C_3}{K_1} \sigma^k \|g\|_{\mathcal{A}_i^k} \leq \frac{C_3 \sqrt{NK_2}}{K_1} \sigma^k \|g\|_{\Omega_i^k},$$

where N is the number of triangles in Ω_i^k . Note that $\sqrt{N} \leq C_4/|\Delta|$, where C_4 depends on the area of Ω_i^k and the constant β . On the other hand,

$$(5.13) \quad \|g\|_{\Omega_i^k} \leq \|f - s_L\|_{\Omega_i^k} + \|f - s_{L, i}^k\|_{\Omega_i^k}.$$

Combining the last two inequalities with the error bound (2.13), we get (5.8) for $\alpha = \beta = 0$. To get the result for the derivative $D_x^\alpha D_y^\beta$, we apply the Markov inequality to a triangle where $\|D_x^\alpha D_y^\beta g\|_\Omega$ takes its maximum. \square

5.3. PLS splines. Given a set of measurements $\{f_i\}_{i=1}^{n_d}$ of a function f and a triangulation Δ , let s_λ be the PLS spline fit of f from $\mathcal{S}(\Delta)$ with smoothing parameter $\lambda > 0$. Let s_λ^k be the DDC PLS spline produced by Algorithm 1.1 with parameter k . In (2.17) we showed that if $f \in W_\infty^{m+1}(\Omega)$, then $\|s_\lambda - s_\lambda^k\|_\Omega = \mathcal{O}(|\Delta|^{m+1}) + \mathcal{O}(\lambda)$. In this section we discuss the dependence of this difference on k .

THEOREM 5.3. *There exists $\sigma \in (0, 1)$ such that if $f \in W_\infty^{m+1}(\Omega)$ with $1 \leq m \leq d$, then*

$$(5.14) \quad \|s_\lambda - s_\lambda^k\|_\Omega \leq C \sigma^k \left(1 + \frac{\sqrt{\lambda}}{|\Delta|} \right) \left(|\Delta|^m |f|_{m+1, \Omega} + \frac{\lambda}{|\Delta|} |f|_{2, \Omega} \right)$$

if λ is sufficiently small compared to $|\Delta|$. When Ω is convex, C is a constant depending only on $d, \ell, \beta, K_1, K_2, \theta_\Delta$, and the area of Ω . When Ω is nonconvex, C also depends on the Lipschitz constant of the boundary of Ω .

Proof. Let Ω_i be one of the subdomains in Algorithm 1.1. In view of the way in which s_λ is defined, it suffices to estimate the norm of $s_\lambda - s_\lambda^k$ on Ω_i . Let Δ_i^k be the

subtriangulation obtained by restricting Δ to $\Omega_i^k := \text{star}^k(\Omega_i)$. Fix $k \geq 1$. We make use of Lemma 4.2 applied to $\mathcal{W} := \mathcal{S}(\Delta)|_{\Omega_i^k}$ with the inner product

$$(5.15) \quad \langle \phi, \psi \rangle_\lambda := \langle \phi, \psi \rangle_{\mathcal{A}_i^k} + \lambda \langle \phi, \psi \rangle_{E, \Omega_i^k},$$

where the inner-products in this definition are as in (5.2) and (5.9). Let $s_{\lambda, \Omega_i^k} := s_\lambda|_{\Omega_i^k}$ be the restriction to Ω_i^k of the global PLS spline fit s_λ of f from $\mathcal{S}(\Delta)$, and let $s_{\lambda, i}^k$ be the PLS spline fit of f from the space $\mathcal{S}(\Delta)|_{\Omega_i^k}$ using data in Ω_i^k . Let $\{B_\xi\}_{\xi \in \mathcal{M}_i^k}$ be a stable 1-local basis for $\mathcal{S}(\Delta)|_{\Omega_i^k}$ as in the proof of Theorem 5.2. Combining (5.3) and (5.10), we see that

$$(5.16) \quad C_1 \left(1 + \frac{\lambda}{|\Delta|^2}\right) \sum_{\xi \in \mathcal{M}_i^k} |c_\xi|^2 \leq \left\| \sum_{\xi \in \mathcal{M}_i^k} c_\xi B_\xi \right\|_\lambda \leq C_2 \left(1 + \frac{\lambda}{|\Delta|^2}\right) \sum_{\xi \in \mathcal{M}_i^k} |c_\xi|^2.$$

Writing $g := s_{\lambda, \Omega_i^k} - s_{\lambda, i}^k \in \mathcal{W}$, and using the characterization of PLS splines, we have

$$(5.17) \quad \langle g, B_\xi \rangle_\lambda = 0, \quad \text{all } B_\xi \text{ with } \text{supp}(B_\xi) \subseteq \Omega_i^k.$$

Now suppose T is a triangle in Ω_i where $|g|$ takes its maximum. Then by (2.12),

$$\|g\|_T \leq \frac{1}{K_1} \|g \cdot \chi_T\|_{\mathcal{A}_i^k} \leq \frac{1}{K_1} (\|g \cdot \chi_T\|_{\mathcal{A}_i^k}^2 + \lambda \|g \cdot \chi_T\|_{E, \Omega_i^k}^2)^{1/2} = \frac{1}{K_1} \|g \cdot \chi_T\|_\lambda.$$

Using Lemma 4.2, we get

$$\|g\|_T \leq \frac{C_3}{K_1} \sigma^k \|g\|_\lambda \leq \frac{C_3}{K_1} \sigma^k (\|g\|_{\mathcal{A}_i^k}^2 + \lambda \|g\|_{E, \Omega_i^k}^2)^{1/2} \leq \frac{C_3}{K_1} \sigma^k (\|g\|_{\mathcal{A}_i^k} + \sqrt{\lambda} \|g\|_{E, \Omega_i^k}),$$

where C_3 depends only on the ratio C_2/C_1 . Following the proofs of Theorems 5.1 and 5.2, we see that

$$\|g\|_{E, \Omega_i^k} \leq \frac{C_4}{|\Delta|^2} \|g\|_{\Omega_i^k}, \quad \|g\|_{\mathcal{A}_i^k} \leq \frac{C_5}{|\Delta|} \|g\|_{\Omega_i^k},$$

which gives

$$\|g\|_T \leq C_6 \sigma^k \left(\frac{1}{|\Delta|} + \frac{\sqrt{\lambda}}{|\Delta|^2} \right) \|g\|_{\Omega_i^k}.$$

Now

$$\|g\|_{\Omega_i^k} \leq \|f - s_\lambda\|_{\Omega_i^k} + \|f - s_{\lambda, i}^k\|_{\Omega_i^k},$$

and using (2.16) we get (5.14). \square

6. Remarks.

Remark 1. DDC methods have been studied for more than 150 years in the literature on the numerical solution of boundary value problems, going back at least to Schwarz’s alternating method; see, e.g., [11]. For a comprehensive treatment and an extensive list of references, see [13]. The idea of domain decomposition has recently been adapted to the problem of fitting scattered data with radial basis functions

(see [2]) as well as to meshless methods (based on radial basis functions) for solving boundary-value problems, see [4] and the book [5].

Remark 2. Many authors have tried to solve global fitting problems by dividing the domain into subdomains, computing fits on each subdomain, and then blending the resulting surface patches together with some kind of blending functions. In most of these methods the use of blending functions changes the form of the final approximant and produces a fit which may not be close to the global fit. Our DDC method is not based on blending functions, and our theorems ensure that the DDC-spline is close to the global fit.

Remark 3. As observed in [12], in computation with \mathcal{M} -bases it is important to exercise some care in choosing the MDS \mathcal{M} . Thus, for example in Figure 1, for each vertex v , the six black dots should be chosen in the triangle with largest angle at v . This means that the minimal determining sets for the subspaces $\mathcal{S}(\Delta)|_{\Omega_i^k}$ may not be subsets of the MDS for the full space.

Remark 4. For convenience, the results of section 5 assume that we are working with a spline space with a 1-local stable basis. However, the same analysis can be carried out with spline spaces with ℓ -local stable bases under the assumption that $k \geq \ell$.

Remark 5. The computations reported here were done on a Macintosh G5 computer using Fortran. The codes have not been optimized for storage or computational speed. We report computational times to give a feeling for how quickly DDC-spline fits can be computed, and to provide a basis for comparing various algorithms. Since the local fits in the DDC method can be computed independently, the actual run times can be greatly reduced by working on a multiprocessor machine (or on a cluster).

REFERENCES

- [1] G. AWANOU, M.-J. LAI, AND P. WENSTON, *The multivariate spline method for scattered data fitting and numerical solution of partial differential equations*, in Wavelets and Splines (Athens, 2005), G. Chen and M.-J. Lai, eds., Nashboro Press, Brentwood, TN, 2006, pp. 24–74.
- [2] R. K. BEATSON, W. A. LIGHT, AND S. BILLINGS, *Fast solution of the radial basis function interpolation equations: Domain decomposition methods*, SIAM J. Sci. Comput., 22 (2000), pp. 1717–1740.
- [3] C. DE BOOR, *A bound on the L_∞ -norm of L_2 -approximation by splines in terms of a global mesh ratio*, Math. Comp., 30 (1976), pp. 765–771.
- [4] Y. DUAN, *Meshless Galerkin method using radial basis functions based on domain decomposition*, Appl. Math. Comput., 179 (2006), pp. 750–762.
- [5] G. FASSHAUER, *Meshfree Approximation Methods with MATLAB*, World Scientific, Singapore, 2007.
- [6] M. VON GOLITSCHKEK, M.-J. LAI, AND L. L. SCHUMAKER, *Error bounds for minimal energy bivariate polynomial splines*, Numer. Math., 93 (2002), pp. 315–331.
- [7] M. VON GOLITSCHKEK AND L. L. SCHUMAKER, *Bounds on projections onto bivariate polynomial spline spaces with stable bases*, Constr. Approx., 18 (2002), pp. 241–254.
- [8] M.-J. LAI, *Multivariate splines for data fitting and approximation*, in Approximation Theory XII (San Antonio, 2007), M. Neamtu and L. L. Schumaker, eds., Nashboro Press, Brentwood, TN, 2008, pp. 210–228.
- [9] M.-J. LAI AND L. L. SCHUMAKER, *On the approximation power of bivariate splines*, Adv. Comput. Math., 9 (1998), pp. 251–279.
- [10] M.-J. LAI AND L. L. SCHUMAKER, *Spline Functions on Triangulations*, Cambridge University Press, Cambridge, UK, 2007.
- [11] M.-J. LAI AND P. WENSTON, *On Schwarz’s domain decomposition methods for elliptic boundary value problems*, Numer. Math., 84 (2000), pp. 475–495.
- [12] L. L. SCHUMAKER, *Computing bivariate splines in scattered data fitting and the finite-element method*, Numer. Algorithms, 48 (2008), pp. 237–260.
- [13] A. TOSELI AND O. WIDLUND, *Domain Decomposition Methods—Algorithms and Theory*, Springer-Verlag, Berlin, 2005.

COUPLED GENERALIZED NONLINEAR STOKES FLOW WITH FLOW THROUGH A POROUS MEDIUM*

V. J. ERVIN[†], E. W. JENKINS[†], AND S. SUN[†]

Abstract. In this article, we analyze the flow of a fluid through a coupled Stokes–Darcy domain. The fluid in each domain is non-Newtonian, modeled by the generalized nonlinear Stokes equation in the free flow region and the generalized nonlinear Darcy equation in the porous medium. A flow rate is specified along the inflow portion of the free flow boundary. We show existence and uniqueness of a variational solution to the problem. We propose and analyze an approximation algorithm and establish a priori error estimates for the approximation.

Key words. generalized nonlinear Stokes flow, coupled Stokes and Darcy flow, defective boundary condition

AMS subject classification. 65N30

DOI. 10.1137/070708354

1. Introduction. The coupling of Stokes and Darcy flow problems has received significant attention over the past several years due to its importance in modeling problems such as surface fluid flow coupled with flow in a porous media (see, for instance, [4, 9, 12, 14, 16, 20, 21]). As in [12], the investigation in this paper is motivated by industrial filtering applications where a non-Newtonian fluid passes through a filter to remove unwanted particulates. The lifetime of the filter is dictated by the increase in pressure drop across the porous medium. This pressure drop increase occurs as debris, transported into the filter by the free flowing fluid, deposits into the filter. Models of the coupled system are necessary to develop simulators that can aid in the design of filters with extended lifetimes and minimize release of debris into the downstream flow.

In these applications, flow rates are typically specified at the inflow of the filtering apparatus. Our first step in modeling the filtration problem is to consider the case of the coupled nonlinear Stokes–Darcy flow problem with defective boundary conditions. Namely, we assume that only flow rates are specified along the inflow boundary. In [12], the authors use the Darcy equation as a boundary condition for the Stokes problem in the free-flow region. We couple the flows across the internal boundary by using conservation of mass and balance of forces across the interface, as in [9, 14, 20, 21].

For Newtonian fluids the extra stress tensor, $\boldsymbol{\tau}$, is proportional to the deformation tensor, $\mathbf{d}(\mathbf{u})$, with the constant of proportionality being the value of the dynamic viscosity, ν . Our model problem uses generalized power law fluids, which are an extension of Newtonian fluids. Generalized power law fluids have a nonconstant viscosity that is a function of the magnitude of the deformation tensor. Models for such viscosity functions include the following [3, 17]:

*Received by the editors November 16, 2007; accepted for publication (in revised form) August 4, 2008; published electronically February 13, 2009.

<http://www.siam.org/journals/sinum/47-2/70835.html>

[†]Department of Mathematical Sciences, Clemson University, Clemson, SC 29634-0975 (vjervin@clemson.edu, lea@clemson.edu, shuyu@clemson.edu). The research of the first two authors was partially supported by the National Science Foundation under grant DMS-0410792.

Carreau model.

$$(1.1) \quad \nu(\mathbf{d}(\mathbf{u})) = \nu_\infty + (\nu_0 - \nu_\infty)/(1 + K|\mathbf{d}(\mathbf{u})|^2)^{(2-r)/2},$$

where $r > 1$, ν_0 , ν_∞ , and $K > 0$ are constants.

Cross model.

$$(1.2) \quad \nu(\mathbf{d}(\mathbf{u})) = \nu_\infty + (\nu_0 - \nu_\infty)/(1 + K|\mathbf{d}(\mathbf{u})|^{2-r}),$$

where $r > 1$, ν_0 , ν_∞ , and $K > 0$ are constants.

Power law model.

$$(1.3) \quad \nu(\mathbf{d}(\mathbf{u})) = K|\mathbf{d}(\mathbf{u})|^{r-2},$$

where $r > 1$ and $K > 0$ are constants.

Many generalized Newtonian fluids exhibit a shear thinning property; that is, the viscosity decreases as the magnitude of $\mathbf{d}(\mathbf{u})$ increases. For the above models this corresponds to a value for r between 1 and 2. Generalized power law viscosity models have been used in modeling the viscosity of biological fluids, lubricants, paints, and polymeric fluids. In the analysis below we assume a general function for $\nu(\mathbf{d}(\mathbf{u}))$ that satisfies particular continuity and monotonicity properties. (See (2.16), (2.17).)

For non-Newtonian fluid flow in a porous medium, various models for the effective viscosity ν_{eff} have been proposed in the literature. (See, for example, [15, 18] and the references cited therein.) Based upon dimensional analysis most models assume that ν_{eff} is a function of $|\mathbf{u}_p|/(\sqrt{\kappa} m_c)$, where κ denotes the permeability of the porous medium, \mathbf{u}_p the Darcy velocity, and m_c a constant related to the internal structure of the porous media. Models for ν_{eff} include the following [15, 18]:

Cross model.

$$(1.4) \quad \nu_{\text{eff}}(\mathbf{u}_p) = \nu_\infty + (\nu_0 - \nu_\infty)/(1 + K|\mathbf{u}_p|^{2-r}),$$

where $r > 1$, ν_0 , ν_∞ , and $K > 0$ are constants.

Power law model.

$$(1.5) \quad \nu_{\text{eff}}(\mathbf{u}_p) = K (|\mathbf{u}_p|/(\sqrt{\kappa} m_c))^{r-2},$$

where $r > 1$ and $K > 0$ are constants.

Again, in the analysis below we assume a general function for $\nu_{\text{eff}}(\mathbf{u}_p)$ that satisfies particular continuity and monotonicity properties. (See (2.16), (2.17).)

Remark. In this work we ignore the influence of pressure on viscosity.

The variational formulation presented below for the coupled nonlinear flow problem (ignoring the defective boundary conditions) is analogous to that for the linear coupled problem studied in [9, 14, 20, 21]. However, as the function setting for the linear problem is in Hilbert spaces ($H_1(\Omega)$, $L^2(\Omega)$) compared to Banach spaces ($W_{1,r}(\Omega)$, $L^{r'}(\Omega)$) for the nonlinear problem, the analysis used herein is considerably different than that in [9, 14, 20, 21].

2. Modeling equations. Let $\Omega \subset \mathbb{R}^n$, $n = 2$ or 3 , denote the flow domain of interest. Additionally, let Ω_f and Ω_p denote bounded Lipschitz domains for the nonlinear generalized Stokes flow and nonlinear generalized Darcy flow, respectively. The interface boundary between the domains we denote by $\Gamma := \partial\Omega_f \cap \partial\Omega_p$. Note that $\Omega := \Omega_f \cup \Omega_p \cup \Gamma$. The outward-pointing unit normal vectors to Ω_f and Ω_p are

denoted \mathbf{n}_f and \mathbf{n}_p , respectively. The tangent vectors on Γ are denoted by \mathbf{t}_1 (for $n = 2$), or $\mathbf{t}_l, l = 1, 2$ (for $n = 3$).

We assume that there is an inflow boundary Γ_{in} , a subset of $\partial\Omega_f \setminus \Gamma$, which is separated from Γ , and an outflow boundary Γ_{out} , a subset of $\partial\Omega_p \setminus \Gamma$, which is also separated from Γ . See Figure 2.1 for an illustration of the domain of the problem.

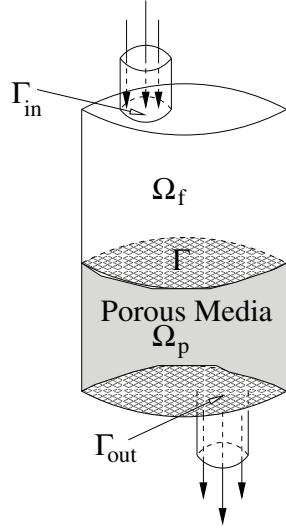


FIG. 2.1. Illustration of flow domain.

Define $\Gamma_f := \partial\Omega_f \setminus (\Gamma \cup \Gamma_{\text{in}})$, and $\Gamma_p := \partial\Omega_p \setminus (\Gamma \cup \Gamma_{\text{out}})$.

Velocities are denoted by $\mathbf{u}_j : \Omega_j \rightarrow \mathbb{R}^n, j = f, p$, and pressures are denoted by $p_j : \Omega_j \rightarrow \mathbb{R}, j = f, p$.

In Ω_f , we assume that the flow is governed by the nonlinear generalized Stokes flow, subject to a specified flow rate, $-fr$, across Γ_{in} and no-slip condition on Γ_f :

$$(2.1) \quad -\nabla \cdot (\boldsymbol{\sigma} - p_f \mathbf{I}) = \mathbf{f}_f \quad \text{in } \Omega_f,$$

$$(2.2) \quad \nabla \cdot \mathbf{u}_f = 0 \quad \text{in } \Omega_f,$$

$$(2.3) \quad \boldsymbol{\sigma} = g_f(\mathbf{d}(\mathbf{u}_f))\mathbf{d}(\mathbf{u}_f) \quad \text{in } \Omega_f,$$

$$(2.4) \quad \mathbf{u}_f = \mathbf{0} \quad \text{on } \Gamma_f,$$

$$(2.5) \quad \int_{\Gamma_{\text{in}}} \mathbf{u}_f \cdot \mathbf{n}_f \, ds = -fr,$$

where $\boldsymbol{\sigma}$ denotes the fluid's extra stress tensor and $\mathbf{d}(\mathbf{v}) := \frac{1}{2}(\nabla \mathbf{v} + \nabla^T \mathbf{v})$ is the deformation tensor. The particular form for the nonlinear viscosity function $g_f(\cdot)$ is discussed in section 2.2. For simplicity we consider here the case of a single inflow boundary Γ_{in} . Multiple inflow boundary segments with separately specified flow rates can also be modeled [6, 7, 11].

We assume that the flow in the porous domain Ω_p is governed by a generalized Darcy's equation subject to a specified flow rate, fr , across Γ_{out} and a nonpenetration

condition on Γ_p :

$$(2.6) \quad \mathbf{u}_p = -\frac{\kappa}{\nu_{\text{eff}}}\nabla p_p \quad \text{in } \Omega_p,$$

$$(2.7) \quad \nabla \cdot \mathbf{u}_p = 0 \quad \text{in } \Omega_p,$$

$$(2.8) \quad \mathbf{u}_p \cdot \mathbf{n}_p = 0 \quad \text{on } \Gamma_p,$$

$$(2.9) \quad \int_{\Gamma_{\text{out}}} \mathbf{u}_p \cdot \mathbf{n}_f \, ds = fr.$$

In general κ denotes a symmetric, positive definite tensor. For simplicity, we will assume κ is a positive (scalar) constant.

2.1. Interface conditions. The flows in Ω_f and Ω_p are coupled across the interface Γ . Conditions describing the coupling of the flows are discussed below.

Conservation of mass across Γ : The conservation of mass across Γ imposes the constraint

$$(2.10) \quad \mathbf{u}_f \cdot \mathbf{n}_f + \mathbf{u}_p \cdot \mathbf{n}_p = 0 \quad \text{on } \Gamma.$$

Balance of the normal forces across Γ : The balance of the normal forces across Γ imposes the constraint

$$(2.11) \quad p_f - (\boldsymbol{\sigma}\mathbf{n}_f) \cdot \mathbf{n}_f = p_p \quad \text{on } \Gamma.$$

Balance of the forces on Γ : For the tangential forces on Γ we use the Beavers–Joseph–Saffman condition [1, 13, 22]

$$(2.12) \quad \mathbf{u}_f \cdot \mathbf{t}_l = -csr_l (\boldsymbol{\sigma}\mathbf{n}_f) \cdot \mathbf{t}_l \quad \text{on } \Gamma, \quad l = 1, \dots, n - 1,$$

where csr_l , $l = 1, \dots, n - 1$, denote frictional constants that can be determined experimentally.

2.2. Variational formulations. Given $r \in \mathbb{R}$, $r > 1$, we denote its unitary conjugate by r' , satisfying $r^{-1} + (r')^{-1} = 1$.

For Ω_f , define

$$X_f := \{ \mathbf{v} : \mathbf{v} \in (W^{1,r}(\Omega_f))^n, \quad \mathbf{v}|_{\Gamma_f} = \mathbf{0} \} \quad \text{and} \quad M_f := L^{r'}(\Omega_f).$$

For $\mathbf{v} \in X_f$, $q \in M_f$, define $\|\mathbf{v}\|_{X_f} := \|\mathbf{v}\|_{(W^{1,r}(\Omega_f))^n}$, and $\|q\|_{M_f} := \|q\|_{L^{r'}(\Omega_f)}$.

For Ω_p , define

$$L^r(\text{div}, \Omega_p) := \{ \mathbf{v} : \mathbf{v} \in (L^r(\Omega_p))^n \text{ and } \nabla \cdot \mathbf{v} \in L^r(\Omega_p) \},$$

$$X_p := \{ \mathbf{v} : \mathbf{v} \in L^r(\text{div}, \Omega_p), \quad \mathbf{v} \cdot \mathbf{n}|_{\Gamma_p} = 0 \}, \quad \text{and} \quad M_p := L^{r'}(\Omega_p).$$

Similarly, for $\mathbf{v} \in X_p$, $q \in M_p$, define $\|\mathbf{v}\|_{X_p} := \|\mathbf{v}\|_{(L^r(\Omega_p))^n} + \|\nabla \cdot \mathbf{v}\|_{L^r(\Omega_p)}$ and $\|q\|_{M_p} := \|q\|_{L^{r'}(\Omega_p)}$.

We also use the spaces X and M defined on Ω by

$$X := X_f \times X_p \quad \text{and} \quad M := \left\{ q \in M_f \times M_p \mid \int_{\Omega} q \, dA = 0 \right\}$$

and denote the dual space of X by X^* .

For $\mathbf{v} = (\mathbf{v}_f, \mathbf{v}_p) \in X$ and $q = (q_f, q_p) \in M$,

$$\|\mathbf{v}\|_X := \|\mathbf{v}_f\|_{X_f} + \|\mathbf{v}_p\|_{X_p} \quad \text{and} \quad \|q\|_M := \left(\|q_f\|_{L^{r'}(\Omega_f)}^{r'} + \|q_p\|_{L^{r'}(\Omega_p)}^{r'} \right)^{1/r'}.$$

Also, for $f, k : \Omega \rightarrow \mathbb{R}^m$, $(f, k) := \int_{\Omega} f \cdot k \, dA$.

Let $g(\mathbf{x}) : \mathbb{R}^N \rightarrow \mathbb{R}^+ \cup \{0\}$ and $G(\mathbf{x}) : \mathbb{R}^N \rightarrow \mathbb{R}^N$ be given by $G(\mathbf{x}) := g(\mathbf{x})\mathbf{x}$. Further for $\mathbf{x}, \mathbf{h} \in \mathbb{R}^N$, let $G(\cdot)$ satisfy (for constants $C_1, C_2, C_3 > 0$, and $c \geq 0$)

$$(2.13) \quad \mathbf{A1:} \quad |G(\mathbf{x} + \mathbf{h}) - G(\mathbf{x})| |\mathbf{h}| \leq C_1 (G(\mathbf{x} + \mathbf{h}) - G(\mathbf{x})) \cdot \mathbf{h},$$

$$(2.14) \quad \mathbf{A2:} \quad \frac{|\mathbf{h}|^2}{c + |\mathbf{x}|^{2-r} + |\mathbf{x} + \mathbf{h}|^{2-r}} \leq C_2 (G(\mathbf{x} + \mathbf{h}) - G(\mathbf{x})) \cdot \mathbf{h},$$

$$(2.15) \quad \mathbf{A3:} \quad |G(\mathbf{x} + \mathbf{h}) - G(\mathbf{x})| \leq C_3 \frac{|\mathbf{h}|}{c + |\mathbf{x}|^{2-r} + |\mathbf{x} + \mathbf{h}|^{2-r}},$$

with the convention that $G(\mathbf{x}) = \mathbf{0}$ if $\mathbf{x} = \mathbf{0}$, and $|\mathbf{h}|/(c + |\mathbf{x}| + |\mathbf{h}|) = 0$ if $c = 0$ and $\mathbf{x} = \mathbf{h} = \mathbf{0}$.

From **A1**, **A2**, and **A3** it follows (see [23]) that there exist constants $C_4, C_5 > 0$ such that for $\mathbf{s}, \mathbf{t}, \mathbf{w} \in (L^r(\Omega))^N$

$$(2.16) \quad \int_{\Omega} (G(\mathbf{s}) - G(\mathbf{t})) \cdot (\mathbf{s} - \mathbf{t}) \, dA \geq C_4 \left(\int_{\Omega} |G(\mathbf{s}) - G(\mathbf{t})| |\mathbf{s} - \mathbf{t}| \, dA + \frac{\|\mathbf{s} - \mathbf{t}\|_{L^r(\Omega)}^2}{c + \|\mathbf{s}\|_{L^r(\Omega)}^{2-r} + \|\mathbf{t}\|_{L^r(\Omega)}^{2-r}} \right),$$

$$(2.17) \quad \int_{\Omega} (G(\mathbf{s}) - G(\mathbf{t})) \cdot \mathbf{w} \, dA \leq C_5 \frac{\|\mathbf{s} - \mathbf{t}\|}{c + \|\mathbf{s}\| + \|\mathbf{t}\|} \|\mathbf{w}\|_{L^r(\Omega)} \left(\int_{\Omega} |G(\mathbf{s}) - G(\mathbf{t})| |\mathbf{s} - \mathbf{t}| \, dA \right)^{1/r'}.$$

In Ω_p , with \mathbf{x}, \mathbf{h} in (2.13)–(2.15) denoting vectors in \mathbb{R}^n and \cdot the usual vector dot product, we assume that $g_p(\mathbf{u}_p) := \nu_{\text{eff}}/\kappa$, and let $G_p(\mathbf{v}) = g_p(\mathbf{v})\mathbf{v}$.

In Ω_f we assume that $\boldsymbol{\sigma} = g_f(\mathbf{d}(\mathbf{u}_f))\mathbf{d}(\mathbf{u}_f)$, and let $G_f(\boldsymbol{\tau}) := g_f(\boldsymbol{\tau})\boldsymbol{\tau}$, where we interpret \mathbf{x}, \mathbf{h} in (2.13)–(2.15) as tensors in $\mathbb{R}^{n \times n}$ and \cdot as the usual tensor scalar product.

Remark. For $\nu_{\infty} = 0$, conditions (2.13)–(2.15) are satisfied for $G_f(\boldsymbol{\tau})$ and $G_p(\mathbf{v})$, with $g_f(\mathbf{d}(\mathbf{u})) = 2\nu(\mathbf{d}(\mathbf{u}))$ described in (1.1)–(1.3) and $g_p(\mathbf{u}_p) = \nu_{\text{eff}}(\mathbf{u}_p)$ described in (1.4) and (1.5) (see [23]). Different functions spaces from the setting studied herein are required for $\nu_{\infty} > 0$.

Multiplying (2.1) through by $\mathbf{v}_1 \in X_f$, integrating over Ω_f , and using (2.3) and the fact that $\{\mathbf{n}_f, \mathbf{t}_l, l = 1, \dots, n-1\}$ form an orthonormal basis along Γ , we have

$$(2.18) \quad \begin{aligned} \int_{\Omega_f} \mathbf{f}_f \cdot \mathbf{v}_1 \, dA &= \int_{\Omega_f} \boldsymbol{\sigma} : \mathbf{d}(\mathbf{v}_1) \, dA - \int_{\Omega_f} p_f \nabla \cdot \mathbf{v}_1 \, dA - \int_{\Gamma \cup \Gamma_{\text{in}}} ((-p_f \mathbf{I} + \boldsymbol{\sigma})\mathbf{n}_f) \cdot \mathbf{v}_1 \, ds \\ &= \int_{\Omega_f} g_f(\mathbf{d}(\mathbf{u}_f))\mathbf{d}(\mathbf{u}_f) : \mathbf{d}(\mathbf{v}_1) \, dA - \int_{\Omega_f} p_f \nabla \cdot \mathbf{v}_1 \, dA \\ &\quad + \sum_{l=1}^{n-1} \int_{\Gamma} -\mathbf{n}_f^T \boldsymbol{\sigma} \mathbf{t}_l \, \mathbf{v}_1 \cdot \mathbf{t}_l \, ds \\ &\quad + \int_{\Gamma} (p_f - \mathbf{n}_f^T \boldsymbol{\sigma} \mathbf{n}_f) \, \mathbf{v}_1 \cdot \mathbf{n}_f \, ds - \int_{\Gamma_{\text{in}}} ((-p_f \mathbf{I} + \boldsymbol{\sigma})\mathbf{n}_f) \cdot \mathbf{v}_1 \, ds. \end{aligned}$$

Also, multiplying (2.6) through by $\mathbf{v}_2 \in X_p$ and integrating over Ω_p , we obtain

$$(2.19) \quad 0 = \int_{\Omega_p} g_p(\mathbf{u}_p)\mathbf{u}_p \cdot \mathbf{v}_2 \, dA - \int_{\Omega_p} p_p \nabla \cdot \mathbf{v}_2 \, dA + \int_{\Gamma_{\text{out}}} p_p \mathbf{v}_2 \cdot \mathbf{n}_p \, ds + \int_{\Gamma} p_p \mathbf{v}_2 \cdot \mathbf{n}_p \, ds.$$

The coupling of the Stokes and Darcy flows occurs through the interface conditions (2.10) and (2.11). Following [14], we introduce a new variable λ representing

$$(2.20) \quad \lambda := p_f - (\boldsymbol{\sigma}\mathbf{n}_f) \cdot \mathbf{n}_f = p_p$$

and incorporate (2.11) into (2.18) and (2.19). Equation (2.10) is imposed *weakly* in a separate equation. (See (2.32) below.)

Note that using the Beavers–Joseph–Saffman condition (2.12),

$$\sum_{l=1}^{n-1} \int_{\Gamma} -\mathbf{n}_f^T \boldsymbol{\sigma} \mathbf{t}_l \cdot \mathbf{v}_1 \cdot \mathbf{t}_l \, ds = \sum_{l=1}^{n-1} \int_{\Gamma} c_s r_l^{-1} (\mathbf{u}_f \cdot \mathbf{t}_l) (\mathbf{v}_1 \cdot \mathbf{t}_l) \, ds.$$

To incorporate the specified flow rate conditions into the mathematical formulation, we use a Lagrange multiplier approach. In (2.18) and (2.19)

$$(2.21) \quad \int_{\Gamma_{\text{in}}} ((-p_f \mathbf{I} + \boldsymbol{\sigma})\mathbf{n}_f) \cdot \mathbf{v}_1 \, ds \text{ is replaced by } \beta_{\text{in}} \int_{\Gamma_{\text{in}}} \mathbf{v}_1 \cdot \mathbf{n}_f \, ds$$

$$(2.22) \quad \int_{\Gamma_{\text{out}}} p_p \mathbf{v}_2 \cdot \mathbf{n}_p \, ds \text{ is replaced by } \beta_{\text{out}} \int_{\Gamma_{\text{out}}} \mathbf{v}_2 \cdot \mathbf{n}_p \, ds,$$

where $\beta_{\text{in}}, \beta_{\text{out}} \in \mathbb{R}$ are undetermined constants. We comment below on the implicit assumptions induced by using the Lagrange multiplier approach.

For $\mathbf{v} \in W^{0,r}(\text{div}, \Omega_p)$, we have that $\mathbf{v} \cdot \mathbf{n}_p \in W^{-1/r,r}(\partial\Omega_p)$ (see [8, p. 47]).

For $\mathbf{v} \in X_p$ and $\lambda \in W^{1/r,r'}(\Gamma)$ we define the operator $\mathbf{v} \cdot \mathbf{n}_p \in W^{-1/r,r}(\Gamma)$ as

$$(2.23) \quad \langle \mathbf{v} \cdot \mathbf{n}_p, \lambda \rangle_{\Gamma} := \langle \mathbf{v} \cdot \mathbf{n}_p, E_{\Gamma}^{r'} \lambda \rangle_{\partial\Omega_p},$$

with $E_{\Gamma}^{r'} \lambda$ defined as in Lemma A.1 in Appendix A (with the association $p = r'$, $\Omega = \Omega_p$, $\Gamma = \Gamma$, $\Gamma_b = \Gamma_p$, $\Gamma_d = \Gamma_{\text{out}}$).

Note that for $\mathbf{v} \in X_p$ sufficiently smooth,

$$\langle \mathbf{v} \cdot \mathbf{n}_p, \lambda \rangle_{\Gamma} = \langle \mathbf{v} \cdot \mathbf{n}_p, E_{\Gamma}^{r'} \lambda \rangle_{\partial\Omega_p} = \int_{\Gamma} \mathbf{v} \cdot \mathbf{n}_p \lambda \, ds.$$

For $\mathbf{v} \in (W^{1,r}(\Omega_f))^n$ we have that $\mathbf{v} \cdot \mathbf{n}_f \in W^{1/r',r}(\partial\Omega_f)$; hence $\int_{\Gamma} \mathbf{v} \cdot \mathbf{n}_f \lambda \, ds$ is well defined.

In order to compactly write the mathematical formulation, we introduce the following bilinear forms:

$$(2.24) \quad a_f(\mathbf{u}, \mathbf{v}) := \int_{\Omega_f} g_f(\mathbf{d}(\mathbf{u}))\mathbf{d}(\mathbf{u}) : \mathbf{d}(\mathbf{v}) \, dA + \sum_{l=1}^{n-1} \int_{\Gamma} c_s r_l^{-1} (\mathbf{u} \cdot \mathbf{t}_l) (\mathbf{v} \cdot \mathbf{t}_l) \, ds,$$

$$(2.25) \quad a_p(\mathbf{u}, \mathbf{v}) := \int_{\Omega_p} g_p(\mathbf{u})\mathbf{u} \cdot \mathbf{v} \, dA,$$

$$(2.26) \quad b_f(\mathbf{v}, q, \beta) := \int_{\Omega_f} q \nabla \cdot \mathbf{v} \, dA + \beta \int_{\Gamma_{\text{in}}} \mathbf{v} \cdot \mathbf{n}_f \, ds,$$

$$(2.27) \quad b_p(\mathbf{v}, q, \beta) := \int_{\Omega_p} q \nabla \cdot \mathbf{v} \, dA + \beta \int_{\Gamma_{\text{out}}} \mathbf{v} \cdot \mathbf{n}_p \, ds.$$

With the above notation, the modeling equations in Ω_f may be written as

$$(2.28) \quad a_f(\mathbf{u}_f, \mathbf{v}_1) - b_f(\mathbf{v}_1, p_f, \beta_{\text{in}}) + \int_{\Gamma} \mathbf{v}_1 \cdot \mathbf{n}_f \lambda \, ds = (\mathbf{f}_f, \mathbf{v}_1)_{\Omega_f} \quad \forall \mathbf{v}_1 \in X_f,$$

$$(2.29) \quad b_f(\mathbf{u}_f, q_1, \beta_1) = -\beta_1 fr \quad \forall (q_1 \times \beta_1) \in M_f \times \mathbb{R},$$

and in Ω_p as

$$(2.30) \quad a_p(\mathbf{u}_p, \mathbf{v}_2) - b_p(\mathbf{v}_2, p_p, \beta_{\text{out}}) + \langle \lambda, \mathbf{v}_2 \cdot \mathbf{n}_p \rangle_{\Gamma} = 0 \quad \forall \mathbf{v}_2 \in X_p,$$

$$(2.31) \quad b_p(\mathbf{u}_p, q_2, \beta_2) = \beta_2 fr \quad \forall (q_2 \times \beta_2) \in M_p \times \mathbb{R}.$$

Together with (2.28)–(2.31) we have the interface condition (2.10). We impose this constraint weakly using

$$(2.32) \quad \int_{\Gamma} \mathbf{u}_f \cdot \mathbf{n}_f \zeta \, ds + \langle \mathbf{u}_p \cdot \mathbf{n}_p, \zeta \rangle_{\Gamma} = 0 \quad \forall \zeta \in W^{1/r, r'}(\Gamma).$$

Introduce $\mathbf{f} := (\mathbf{f}_f, \mathbf{0})$, $b_I(\cdot, \cdot) : X \times W^{1/r, r'}(\Gamma) \rightarrow \mathbb{R}$ as

$$(2.33) \quad b_I(\mathbf{v}, \zeta) := \int_{\Gamma} \mathbf{v}_f \cdot \mathbf{n}_f \zeta \, ds + \langle \mathbf{v}_p \cdot \mathbf{n}_p, \zeta \rangle_{\Gamma},$$

and $a(\cdot, \cdot) : X \times X \rightarrow \mathbb{R}$, $b(\cdot, \cdot, \cdot) : X \times M \times \mathbb{R}^2 \rightarrow \mathbb{R}$ as

$$(2.34) \quad \begin{aligned} a(\mathbf{u}, \mathbf{v}) &:= a_f(\mathbf{u}_f, \mathbf{v}_f) + a_p(\mathbf{u}_p, \mathbf{v}_p) \quad \text{and} \\ b(\mathbf{v}, q, \boldsymbol{\gamma}) &:= b_f(\mathbf{v}_f, q_f, \gamma_1) + b_p(\mathbf{v}_p, q_p, \gamma_2). \end{aligned}$$

We then state the coupled fluid flow problem as follows: *Given $\mathbf{f} \in X^*$, $fr \in \mathbb{R}$, determine $(\mathbf{u}, p, \lambda, \boldsymbol{\beta}) \in X \times M \times W^{1/r, r'}(\Gamma) \times \mathbb{R}^2$ such that*

$$(2.35) \quad a(\mathbf{u}, \mathbf{v}) - b(\mathbf{v}, p, \boldsymbol{\beta}) + b_I(\mathbf{v}, \lambda) = (\mathbf{f}, \mathbf{v}) \quad \forall \mathbf{v} \in X,$$

$$(2.36) \quad b(\mathbf{u}, q, \boldsymbol{\gamma}) - b_I(\mathbf{u}, \zeta) = \boldsymbol{\gamma} \cdot \begin{bmatrix} -1 \\ 1 \end{bmatrix} fr \quad \forall (q, \zeta, \boldsymbol{\gamma}) \in M \times W^{1/r, r'}(\Gamma) \times \mathbb{R}^2.$$

The unique solvability of (2.35)–(2.36) hinges upon showing two inf-sup conditions: one for $b(\cdot, \cdot, \cdot)$ and the other for $b_I(\cdot, \cdot)$.

Equivalence of the differential equations and variational formulations.

As demonstrated above, the variational formulation (2.35)–(2.36) was obtained by multiplying the differential equations by sufficiently smooth functions, integrating over the domain, and, where appropriate, applying Green’s theorem. We also used (2.21)–(2.22) to impose the specified flow rate boundary conditions. For a smooth solution, the steps used in deriving the variational equations can be reversed to show that equations (2.1)–(2.5), (2.6)–(2.9) are satisfied. In addition we have that a smooth solution of (2.35)–(2.36) satisfies the following additional boundary conditions (see [7]).

For \mathbf{n}_f , the outward normal on Γ_{in} , express the extra stress vector on Γ_{in} , $\boldsymbol{\sigma}\mathbf{n}_f$, as

$$\boldsymbol{\sigma}\mathbf{n}_f = s_n\mathbf{n}_f + \mathbf{s}_T,$$

where $s_n = (\boldsymbol{\sigma}\mathbf{n}_f) \cdot \mathbf{n}_f$ and $\mathbf{s}_T = \boldsymbol{\sigma}\mathbf{n}_f - s_n\mathbf{n}_f$. The scalar s_n represents the magnitude of the extra stress in the outward normal direction to Γ_{in} , and \mathbf{s}_T the component of the extra stress vector which lies in the plane of Γ_{in} .

LEMMA 2.1. *Any smooth solution of (2.35), (2.36) satisfies the following boundary conditions:*

$$(2.37) \quad \text{on } \Gamma_{\text{in}}, \quad -p_f + s_n = -\beta_{\text{in}} \quad \text{and} \quad \mathbf{s}_T = \mathbf{0};$$

$$(2.38) \quad \text{on } \Gamma_{\text{out}}, \quad p_p = -\beta_{\text{out}}.$$

Proof. The proof follows as in [7]. □

Remark. Equations (2.1)–(2.5), (2.6)–(2.9), (2.10)–(2.12) do not uniquely define a solution, but rather a set of solutions. The variational formulation (2.35)–(2.36) chooses a solution from the solution set. Specifically, (2.35)–(2.36) chooses *the solution* which satisfies (2.37)–(2.38). A different variational formulation may result in the selection of a different solution from the solution set. (See, for example, [7].)

3. Existence and uniqueness of the variational formulation. In order to show the existence and uniqueness of the variational formulation, we introduce the following subspaces of X :

$$(3.1) \quad V := \{\mathbf{v} \in X : b_I(\mathbf{v}, \zeta) = 0 \quad \forall \zeta \in W^{1/r, r'}(\Gamma)\},$$

$$(3.2) \quad Z := \{\mathbf{v} \in V : b(\mathbf{v}, q, \boldsymbol{\gamma}) = 0 \quad \forall (q, \boldsymbol{\gamma}) \in M \times \mathbb{R}^2\}.$$

Consider $b(\cdot, \cdot, \cdot) : X \times M \times \mathbb{R}^2 \rightarrow \mathbb{R}$ defined in (2.34). Using Hölder’s inequality together with the definition (2.23), we have that $b(\cdot, \cdot, \cdot)$ is continuous. In addition, $b(\cdot, \cdot, \cdot)$ satisfies the following inf-sup condition.

LEMMA 3.1. *There exists $C_{MRV} > 0$ such that*

$$(3.3) \quad \inf_{(\mathbf{0}, \mathbf{0}) \neq (q, \boldsymbol{\gamma}) \in M \times \mathbb{R}^2} \sup_{\mathbf{u} \in V} \frac{b(\mathbf{u}, q, \boldsymbol{\gamma})}{\|\mathbf{u}\|_X \|(q, \boldsymbol{\gamma})\|_{M \times \mathbb{R}^2}} \geq C_{MRV},$$

where $\|(q, \boldsymbol{\gamma})\|_{M \times \mathbb{R}^2} := \|q\|_M + \|\boldsymbol{\gamma}\|_{\mathbb{R}^2}$.

Proof. Fix $(q, \boldsymbol{\gamma}) \in M \times \mathbb{R}^2$ and let

$$(3.4) \quad \hat{q} := \frac{|q|^{r'/r-1}q}{\|q\|_M^{r'-1}}, \quad \hat{\boldsymbol{\gamma}} := \frac{\boldsymbol{\gamma}}{\|\boldsymbol{\gamma}\|_{\mathbb{R}^2}}.$$

Note that $\int_{\Omega} q \hat{q} \, d\Omega = \|q\|_M$, $\|\hat{q}\|_{L^r(\Omega)} = 1$, and $\boldsymbol{\gamma} \cdot \hat{\boldsymbol{\gamma}} = \|\boldsymbol{\gamma}\|_{\mathbb{R}^2}$, $\|\hat{\boldsymbol{\gamma}}\|_{\mathbb{R}^2} = 1$.

Let $\Gamma_i^m \subset \Gamma_i$ such that $meas(\Gamma_i^m) > 0$ and $dist(\Gamma_i^m, \partial\Omega \setminus \Gamma_i) > 0$ for $i = in, out$.
 Let $h \in C(\partial\Omega) \subset W^{1/r',r}(\partial\Omega)$ be given by

$$h|_{\Gamma_i^m} := \hat{\gamma}_i/meas(\Gamma_i^m), \quad i = in, out,$$

$$h|_{\partial\Omega \setminus (\Gamma_{in} \cup \Gamma_{out})} := 0,$$

and on $\Gamma_i \setminus \Gamma_i^m$ h is either a strictly increasing or strictly decreasing function.

Also, let $\delta \in \mathbb{R}$ be given by

$$\delta := \left(\int_{\partial\Omega} h \, ds - \int_{\Omega} \hat{q} \, dA \right) / meas(\Omega).$$

From [8, p. 127], given $f \in L^r(\Omega)$, $\mathbf{a} \in W^{1-1/r,r}(\partial\Omega)$, $1 < r < \infty$, satisfying

$$(3.5) \quad \int_{\Omega} f \, dA = \int_{\partial\Omega} \mathbf{a} \cdot \mathbf{n} \, ds,$$

there exists $\mathbf{v} \in W^{1,r}(\Omega)$ such that

$$(3.6) \quad \nabla \cdot \mathbf{v} = f \quad \text{in } \Omega,$$

$$(3.7) \quad \mathbf{v} = \mathbf{a} \quad \text{on } \partial\Omega,$$

$$(3.8) \quad \text{with } \|\mathbf{v}\|_{W^{1,r}(\Omega)} \leq C (\|f\|_{L^r(\Omega)} + \|\mathbf{a}\|_{W^{1-1/r,r}(\partial\Omega)}).$$

Let $f = \hat{q} + \delta$, and for $\{\mathbf{n}, \mathbf{t}_i, i = 1, \dots, n-1\}$ denoting an orthonormal system on $\partial\Omega$, let \mathbf{a} be defined by

$$\begin{cases} \mathbf{a} \cdot \mathbf{n} = h, \\ \mathbf{a} \cdot \mathbf{t}_i = 0, \quad i = 1, \dots, n-1. \end{cases}$$

Remark. The choice of the constant δ guarantees that the compatibility condition $\int_{\Omega} f \, d\Omega = \int_{\partial\Omega} \mathbf{a} \cdot \mathbf{n} \, ds$ is satisfied.

Note that $\|\mathbf{a}\|_{W^{1/r',r}(\partial\Omega)} \leq C_1$ and $\|\hat{\gamma}\|_{\mathbb{R}^m} = C_1$. Also,

$$(3.9) \quad \int_{\Omega} \hat{q} \, dA \leq \|\hat{q}\|_{L^r(\Omega)} \|\mathbf{1}\|_{L^{r'}(\Omega)} = C_2,$$

$$(3.10) \quad \int_{\partial\Omega} h \, ds \leq \|\hat{\gamma}\|_{\mathbb{R}^2} \|\mathbf{1}\|_{\mathbb{R}^2} = C_3,$$

and thus $\|\delta\|_{L^r(\Omega)} \leq C_4$.

Let $\mathbf{v}_f = \mathbf{v}|_{\overline{\Omega}_f}$, $\mathbf{v}_p = \mathbf{v}|_{\overline{\Omega}_p}$, where \mathbf{v} denotes the solution of (3.6)–(3.7). From (3.8) we have

$$(3.11) \quad \|\mathbf{v}\|_X \leq C (1 + C_4 + C_1) \leq C_5.$$

Also, note that $\mathbf{v}_f \in W^{1/r',r}(\partial\Omega_f)$, $\mathbf{v}_p \in W^{1/r',r}(\partial\Omega_p)$, and $\mathbf{v}_f = \mathbf{v}_p$ on Γ . Thus, for $\lambda \in W^{1/r,r'}(\Gamma)$,

$$\int_{\Gamma} \mathbf{v}_f \cdot \mathbf{n}_f \lambda \, ds + \langle \mathbf{v}_p \cdot \mathbf{n}_p, \lambda \rangle_{\Gamma} = \int_{\Gamma} \mathbf{v}_f \cdot \mathbf{n}_f \lambda \, ds + \int_{\Gamma} \mathbf{v}_p \cdot \mathbf{n}_p \lambda \, ds = 0,$$

i.e., $\mathbf{v} \in V$.

Now,

$$\begin{aligned} b(\mathbf{v}, q, \gamma) &= \int_{\Omega} q \nabla \cdot \mathbf{v} dA + \gamma_1 \int_{\Gamma_{\text{in}}} \mathbf{v} \cdot \mathbf{n}_f ds + \gamma_2 \int_{\Gamma_{\text{out}}} \mathbf{v} \cdot \mathbf{n}_p ds \\ &\geq \int_{\Omega} q (\hat{q} + \delta) dA + \hat{\gamma} \cdot \gamma \\ &= \|q\|_M + \|\gamma\|_{\mathbb{R}^2} \\ &= \|(q, \gamma)\|_{M \times \mathbb{R}^2}, \end{aligned}$$

as $\int_{\Omega} q \delta dA = 0$ for $q \in M$. Thus,

$$\sup_{\mathbf{u} \in V} \frac{b(\mathbf{u}, (q, \gamma))}{\|(q, \beta)\|_{M \times \mathbb{R}^m} \|\mathbf{u}\|_X} \geq \frac{b(\mathbf{v}, (q, \gamma))}{\|(q, \beta)\|_{M \times \mathbb{R}^m} \|\mathbf{v}\|_X} \geq \frac{1}{C_5},$$

from which (3.3) directly follows. \square

The required inf-sup condition for $b_I(\cdot, \cdot)$ may be stated as follows.

LEMMA 3.2. *The bilinear form $b_I(\cdot, \cdot) : X \times W^{1/r, r'}(\Gamma) \rightarrow \mathbb{R}$ is continuous. Moreover, there exists $C_{X\Gamma} > 0$ such that*

$$(3.12) \quad \inf_{0 \neq \lambda \in W^{1/r, r'}(\Gamma)} \sup_{\mathbf{u} \in X} \frac{b_I(\mathbf{u}, \lambda)}{\|\mathbf{u}\|_X \|\lambda\|_{W^{1/r, r'}(\Gamma)}} \geq C_{X\Gamma}.$$

Proof. The continuity of $b_I(\cdot, \cdot)$ follows from the continuity of the trace operator and definition (2.23).

The proof of this inf-sup condition requires a suitable extension of a functional from $W^{-1/r, r}(\Gamma)$ to $W^{-1/r, r}(\partial\Omega_p)$ be defined. Some of the notation used in this proof is defined in the appendix, where suitable extension operators from Γ to $\partial\Omega_p$ are discussed.

To show (3.12), let $\lambda \in W^{1/r, r'}(\Gamma)$. Then, from the definition of the norm, there exists $f_{\Gamma} \in W^{-1/r, r}(\Gamma)$, $\|f_{\Gamma}\|_{W^{-1/r, r}(\Gamma)} = 1$, such that

$$(3.13) \quad \langle f_{\Gamma}, \lambda \rangle_{\Gamma} \geq \frac{1}{2} \|\lambda\|_{W^{1/r, r'}(\Gamma)}.$$

Given $f_{\Gamma} \in W^{-1/r, r}(\Gamma)$ we can extend it to a functional f in $W^{-1/r, r}(\partial\Omega_p)$ by

$$(3.14) \quad \langle f, \xi \rangle_{\partial\Omega_p} := \langle f_{\Gamma}, \xi|_{\Gamma} \rangle_{\Gamma} \text{ for } \xi \in W^{1/r, r'}(\partial\Omega_p).$$

Note that for $\eta \in W_{00}^{1/r, r'}(\partial\Omega_p \setminus \Gamma)$

$$\langle f, E_{00, \partial\Omega_p \setminus \Gamma}^{r'} \eta \rangle_{\partial\Omega_p} = \langle f_{\Gamma}, E_{00, \partial\Omega_p \setminus \Gamma}^{r'} \eta|_{\Gamma} \rangle_{\Gamma} = \langle f_{\Gamma}, 0 \rangle_{\Gamma} = 0.$$

Thus, from Definition A.3 (see Appendix A), $f|_{\partial\Omega_p \setminus \Gamma} = 0$.

Also,

$$\begin{aligned} \|f\|_{W^{-1/r, r}(\partial\Omega_p)} &= \sup_{\xi \in W^{1/r, r'}(\partial\Omega_p)} \frac{\langle f, \xi \rangle_{\partial\Omega_p}}{\|\xi\|_{W^{1/r, r'}(\partial\Omega_p)}} = \sup_{\xi \in W^{1/r, r'}(\partial\Omega_p)} \frac{\langle f_{\Gamma}, \xi|_{\Gamma} \rangle_{\Gamma}}{\|\xi\|_{W^{1/r, r'}(\partial\Omega_p)}} \\ &\leq \sup_{\xi \in W^{1/r, r'}(\partial\Omega_p)} \frac{\|f_{\Gamma}\|_{W^{-1/r, r}(\Gamma)} \|\xi|_{\Gamma}\|_{W^{1/r, r'}(\Gamma)}}{\|\xi\|_{W^{1/r, r'}(\partial\Omega_p)}} \\ (3.15) \quad &\leq \|f_{\Gamma}\|_{W^{-1/r, r}(\Gamma)} = 1. \end{aligned}$$

Let $\phi \in W^{1,r'}(\Omega_p)$ be given by the weak solution of

$$(3.16) \quad -\nabla \cdot |\nabla \phi|^{r'-2} \nabla \phi + |\phi|^{r'-2} \phi = 0 \quad \text{in } \Omega_p,$$

$$(3.17) \quad |\nabla \phi|^{r'-2} \nabla \phi \cdot \mathbf{n}_p = f \quad \text{on } \partial\Omega_p,$$

i.e., ϕ satisfies

$$(3.18) \quad \begin{aligned} (T(\phi), w) &:= \int_{\Omega_p} \left(|\nabla \phi|^{r'-2} \nabla \phi \cdot \nabla w + |\phi|^{r'-2} \phi w \right) dA \\ &= \int_{\partial\Omega_p} f w ds \quad \forall w \in W^{1,r'}(\Omega_p). \end{aligned}$$

Existence and uniqueness of ϕ follow from the strong monotonicity of $T : W^{1,r'}(\Omega_p) \rightarrow (W^{1,r'}(\Omega_p))^*$.

Note that

$$(3.19) \quad \begin{aligned} (T(\phi), \phi) &= \|\phi\|_{W^{1,r'}(\Omega_p)}^{r'} \leq \|f\|_{W^{-1/r,r}(\partial\Omega_p)} \|\phi\|_{W^{1,r'}(\partial\Omega_p)} \\ &\leq C_1 \|f\|_{W^{-1/r,r}(\partial\Omega_p)} \|\phi\|_{W^{1,r'}(\Omega_p)} \\ \implies \|\phi\|_{W^{1,r'}(\Omega_p)}^{r'} &\leq C_* \|f\|_{W^{-1/r,r}(\partial\Omega_p)}^r \leq C_*, \end{aligned}$$

as $\|f\|_{W^{-1/r,r}(\partial\Omega_p)} \leq 1$.

Now, let $\mathbf{v} := |\nabla \phi|^{r'-2} \nabla \phi$. Note from (3.16) that $\nabla \cdot \mathbf{v} = |\phi|^{r'-2} \phi$, and

$$(3.20) \quad \|\mathbf{v}\|_{W^{0,r}(\text{div}, \Omega_p)}^r = \|\phi\|_{W^{1,r'}(\Omega_p)}^{r'} \leq C_*,$$

i.e., $\mathbf{v} \in W^{0,r}(\text{div}, \Omega_p)$ and $\mathbf{v} \cdot \mathbf{n}_p \in W^{-1/r,r}(\partial\Omega_p)$.

Finally, let $\mathbf{w} = (\mathbf{0}, \mathbf{v}) \in X$. Then, in view of (2.23),

$$\begin{aligned} \sup_{\mathbf{u} \in X} \frac{b_I(\mathbf{u}, \lambda)}{\|\mathbf{u}\|_X} &\geq \frac{b_I(\mathbf{w}, \lambda)}{\|\mathbf{w}\|_X} = \frac{0 + \langle \mathbf{v} \cdot \mathbf{n}_p, \lambda \rangle_\Gamma}{\|\mathbf{v}\|_{W^{0,r}(\text{div}, \Omega_p)}} \\ &\geq \frac{\langle \mathbf{v} \cdot \mathbf{n}_p, E_\Gamma^{r'} \lambda \rangle_{\partial\Omega_p}}{C_*^{1/r}} \\ &= \frac{1}{C_*^{1/r}} \langle f, E_\Gamma^{r'} \lambda \rangle_{\partial\Omega_p} \\ &= \frac{1}{C_*^{1/r}} \langle f_\Gamma, \lambda \rangle_\Gamma \quad \text{as } f|_{\partial\Omega_p \setminus \Gamma} = 0 \quad (\text{see (A.7)}) \\ &\geq \frac{1}{2C_*^{1/r}} \|\lambda\|_{W^{1/r,r'}(\Gamma)} \quad \text{from (3.13)}. \quad \square \end{aligned}$$

We are now in a position to prove the existence and uniqueness of the solution.

THEOREM 3.3. *There exists a unique solution $(\mathbf{u}, p, \lambda, \beta) \in X \times M \times W^{1/r,r'}(\Gamma) \times \mathbb{R}^2$ satisfying (2.35)–(2.36). In addition, there exists a constant $C > 0$ such that*

$$(3.21) \quad \|\mathbf{u}\|_X \leq C \left(\|\mathbf{f}_f\|_{X_f^*} + |fr| \right).$$

Proof. For $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2) \in Z$, note that $\nabla \cdot \mathbf{v}_1 = 0$ a.e. in Ω_f and $\nabla \cdot \mathbf{v}_2 = 0$ a.e. in Ω_p . Hence, for $\mathbf{v} \in Z$, $\|\mathbf{v}_2\|_{X_p} = \|\mathbf{v}_2\|_{L^r(\Omega_p)}$ and $\|\mathbf{v}\|_X = \|\mathbf{v}_1\|_{X_f} + \|\mathbf{v}_2\|_{L^r(\Omega_p)}$.

From the continuity and inf-sup condition for $b(\cdot, \cdot, \cdot)$ [10, Remark 4.2, p. 61] there exists $\mathbf{u}_0 \in V$ such that

$$(3.22) \quad b(\mathbf{u}_0, q, \gamma) = \gamma \cdot \begin{bmatrix} -1 \\ 1 \end{bmatrix} fr \quad \forall (q, \gamma) \in M \times \mathbb{R}^2,$$

with $\|\mathbf{u}_0\|_X \leq C|fr|$.

Together with the continuity and inf-sup condition of $b_I(\cdot, \cdot)$, the existence and uniqueness of the solution to (2.35)–(2.36) can be equivalently stated as follows: *Given $\mathbf{f} \in X^*$, determine $\tilde{\mathbf{u}} \in Z$, $\mathbf{u} = \tilde{\mathbf{u}} + \mathbf{u}_0$, such that*

$$(3.23) \quad a(\tilde{\mathbf{u}} + \mathbf{u}_0, \mathbf{v}) = (\mathbf{f}, \mathbf{v}) \quad \forall \mathbf{v} \in Z.$$

The existence and uniqueness of the solution to (3.23) follows from the continuity and strict monotonicity of $a(\cdot, \cdot)$ on $Z \times Z$, which follows from assumptions (2.16)–(2.17) and the restriction that for $\Omega \subset \mathbb{R}^2$, $4/3 < r \leq 2$, and for $\Omega \subset \mathbb{R}^3$, $3/2 < r \leq 2$. This restriction arises in applying the Sobolev embedding theorem to verify the continuity of $a(\cdot, \cdot)$. Specifically,

$$\begin{aligned} & \sum_{l=1}^{n-1} \int_{\Gamma} csr_l^{-1} ((\mathbf{u}_f - \mathbf{w}_f) \cdot \mathbf{t}_l) (\mathbf{v}_f \cdot \mathbf{t}_l) ds \\ & \leq C \|\mathbf{u}_f - \mathbf{w}_f\|_{L^2(\Gamma)} \|\mathbf{v}_f\|_{L^2(\Gamma)} \\ & \leq C \|\mathbf{u}_f - \mathbf{w}_f\|_{W^{1-1/r, r}(\partial\Omega_f)} \|\mathbf{v}_f\|_{W^{1-1/r, r}(\partial\Omega_f)} \\ & \leq C \|\mathbf{u} - \mathbf{w}\|_X \|\mathbf{v}\|_X. \end{aligned}$$

Also, it follows from (2.16), (2.17), and (3.22) that

$$\|\tilde{\mathbf{u}}\|_X \leq C (\|\mathbf{f}\|_{X^*} + |fr|) = C \left(\|\mathbf{f}_f\|_{X_f^*} + |fr| \right),$$

and therefore the estimate

$$\|\mathbf{u}\|_X \leq C \left(\|\mathbf{f}_f\|_{X_f^*} + |fr| \right). \quad \square$$

4. Finite element approximation. In this section we discuss the finite element approximation to the coupled generalized nonlinear Stokes–Darcy system (2.35), (2.36). We focus our attention on the conforming approximating spaces

$$X_{f,h} \subset X_f, \quad M_{f,h} \subset M_f, \quad X_{p,h} \subset X_p, \quad M_{p,h} \subset M_p, \quad L_h \subset W^{1/r, r'}(\Gamma),$$

where $X_{f,h}, M_{f,h}$ denote velocity and pressure spaces typically used for fluid flow approximations, and $X_{p,h}, M_{p,h}$ denote velocity and pressure spaces typically used for (mixed formulation) Darcy flow approximations.

We begin by describing the finite element approximation framework used in the analysis. Let $\Omega_j \subset \mathbb{R}^n$ ($n = 2, 3$), $j = f, p$, be a polygonal domain and let $\mathcal{T}_{j,h}$ be a triangulation of $\overline{\Omega}_j$ made of triangles (in \mathbb{R}^2) or tetrahedra (in \mathbb{R}^3). Thus, the computational domain is defined by

$$\overline{\Omega} = \cup K; \quad K \in \mathcal{T}_{f,h} \cup \mathcal{T}_{p,h}.$$

We assume that there exist constants c_1, c_2 such that

$$c_1 h \leq h_K \leq c_2 \rho_K,$$

where h_K is the diameter of triangle (tetrahedron) K , ρ_K is the diameter of the greatest ball (sphere) included in K , and $h = \max_{K \in \mathcal{T}_{f,h} \cup \mathcal{T}_{p,h}} h_K$.

For simplicity, we assume that the triangulations on $\overline{\Omega}_f$ and $\overline{\Omega}_p$ induce the same partition on Γ , which we denote $\mathcal{T}_{\Gamma,h}$.

Let $P_k(A)$ denote the space of polynomials on A of degree no greater than k . Also, for $\mathbf{x} = [x_1, \dots, x_n]^T \in \mathbb{R}^n$, let $RT_k(A) := (P_k(A))^n + \mathbf{x}P_k(A)$ denote the k th order Raviart–Thomas elements. Then we define the finite element spaces as follows:

$$(4.1) \quad X_{f,h} := \{ \mathbf{v} \in X_f \cap C(\overline{\Omega}_f)^2 : \mathbf{v}|_K \in P_m(K) \quad \forall K \in \mathcal{T}_{f,h} \},$$

$$(4.2) \quad M_{f,h} := \{ q \in M_f \cap C(\overline{\Omega}_f) : q|_K \in P_{m-1}(K) \quad \forall K \in \mathcal{T}_{f,h} \},$$

$$(4.3) \quad X_{p,h} := \{ \mathbf{v} \in RT_k(K) \quad \forall K \in \mathcal{T}_{p,h} \},$$

$$(4.4) \quad M_{p,h} := \{ q \in M_p : q|_K \in P_k(K) \quad \forall K \in \mathcal{T}_{p,h} \},$$

$$(4.5) \quad L_h := \left\{ \zeta \in W^{1/r,r'}(\Gamma) \cap C(\Gamma) : \zeta|_K \in P_l(K) \quad \forall K \in \mathcal{T}_{\Gamma,h} \right\}.$$

Note that as we are assuming $1 < r < 2$, then $1/r > 1/2$, which implies that, for $\Omega \subset \mathbb{R}^2$, $\lambda \in W^{1/r,r'}(\Gamma)$ is continuous. For $m = 2$, $X_{f,h}$ and $M_{f,h}$ denote the Taylor–Hood spaces.

Below we assume that $m \geq 2$, $k \geq 1$, and $l \leq k$.

Let

$$X_{f,h}^0 := \{ \mathbf{v} \in X_{f,h} : \mathbf{v}|_{\partial\Omega_f \setminus \Gamma_{\text{in}}} = \mathbf{0} \} \quad \text{and} \quad X_{p,h}^0 := \{ \mathbf{v} \in X_{p,h} : \mathbf{v} \cdot \mathbf{n}_p|_{\partial\Omega_p \setminus \Gamma_{\text{out}}} = 0 \}.$$

LEMMA 4.1. *There exist constants $C_{f,h}, C_{p,h} > 0$, such that*

$$(4.6) \quad \inf_{0 \neq q_h \in M_{f,h}} \sup_{\mathbf{v}_h \in X_{f,h}^0} \frac{\int_{\Omega_f} q_h \nabla \cdot \mathbf{v}_h \, dA}{\|q_h\|_{M_f} \|\mathbf{v}_h\|_{X_f}} \geq C_{f,h},$$

$$(4.7) \quad \inf_{0 \neq q_h \in M_{p,h}} \sup_{\mathbf{v}_h \in X_{p,h}^0} \frac{\int_{\Omega_p} q_h \nabla \cdot \mathbf{v}_h \, dA}{\|q_h\|_{M_p} \|\mathbf{v}_h\|_{X_p}} \geq C_{p,h}.$$

Proof. For the case of the pressure spaces having mean value equal to zero, the inf-sup conditions (4.6) and (4.7) are well established. As mentioned in [14], one can extend the inf-sup conditions to the above pressure spaces via a local projector operator argument. (See [2, section VI.4].) \square

Remark. There are several other suitable choices of approximation spaces. (See discussions in [14, 9].)

Discrete approximation problem. Given $\mathbf{f} \in X^*$, $f_r \in \mathbb{R}$, determine $(\mathbf{u}_h, p_h, \lambda_h, \beta_h) \in X_h \times M_h \times L_h \times \mathbb{R}^2$ such that

$$(4.8) \quad a(\mathbf{u}_h, \mathbf{v}_h) - b(\mathbf{v}_h, p_h, \beta_h) + b_I(\mathbf{v}_h, \lambda_h) = (\mathbf{f}, \mathbf{v}_h) \quad \forall \mathbf{v}_h \in X_h,$$

$$(4.9)$$

$$b(\mathbf{u}_h, q_h, \gamma_h) - b_I(\mathbf{u}_h, \zeta_h) = \gamma_h \cdot \begin{bmatrix} -1 \\ 1 \end{bmatrix} f_r \quad \forall (q_h, \gamma_h, \zeta_h) \in M_h \times \mathbb{R}^2 \times L_h(\Gamma).$$

A more general inf-sup condition than that given by (4.6), (4.7) is needed for the analysis. This is established using the following two lemmas. (See also [24].)

Corresponding to V and Z as defined in (3.1) and (3.2), we have the discrete counterparts

$$(4.10) \quad V_h := \{\mathbf{v} \in X_h \mid b_I(\mathbf{v}_h, \zeta) = 0 \quad \forall \zeta \in L_h\},$$

$$(4.11) \quad Z_h := \{\mathbf{v} \in V_h \mid b(\mathbf{v}, q, \gamma) = 0 \quad \forall (q, \gamma) \in M_h \times \mathbb{R}^2\}.$$

LEMMA 4.2. *There exists $C_{RXh} > 0$ such that for h sufficiently small*

$$(4.12) \quad \inf_{\mathbf{0} \neq \boldsymbol{\beta} \in \mathbb{R}^2} \sup_{\mathbf{w}_h \in V_h} \frac{\int_{\Gamma_{\text{in}}} \beta_1 \mathbf{w}_{f,h} \cdot \mathbf{n}_f \, ds + \int_{\Gamma_{\text{out}}} \beta_2 \mathbf{w}_{p,h} \cdot \mathbf{n}_p \, ds}{\|\mathbf{w}_h\|_X \|\boldsymbol{\beta}\|_{\mathbb{R}^2}} \geq C_{RXh}.$$

Proof. We use (3.5)–(3.8) to construct a suitable function \mathbf{v} . Then using a linear interpolant for \mathbf{v} we obtain the stated result.

Assume $\boldsymbol{\beta} = [\beta_1, \beta_2]^T \in \mathbb{R}^2$ is given.

For $i \in \{\text{in}, \text{out}\}$, let $s_i(\mathbf{x})$ denote an arc length parameter on Γ_i , and define $\phi_i : \partial\Omega \rightarrow \mathbb{R}$ by

$$\phi_i(\mathbf{x}) = \begin{cases} \frac{2}{|\Gamma_i|} s_i(\mathbf{x}), & \mathbf{x} \in \Gamma_i, \quad 0 \leq s_i(\mathbf{x}) \leq \frac{|\Gamma_i|}{2}, \\ \frac{2}{|\Gamma_i|} (|\Gamma_i| - s_i(\mathbf{x})), & \mathbf{x} \in \Gamma_i, \quad \frac{|\Gamma_i|}{2} < s_i(\mathbf{x}) \leq |\Gamma_i|, \\ 0 & \text{otherwise.} \end{cases}$$

Further, let $\mathbf{a} \in W^{1-1/r, r}(\partial\Omega)$ and $f \in L^r(\Omega)$ be given by

$$(4.13) \quad \mathbf{a}(\mathbf{x}) = (\beta_1 \phi_{\text{in}}(\mathbf{x}) + \beta_2 \phi_{\text{out}}(\mathbf{x})) \mathbf{n}, \quad f(\mathbf{x}) = \frac{1}{|\Omega|^{1/r}} \int_{\partial\Omega} \mathbf{a} \cdot \mathbf{n} \, ds,$$

where \mathbf{n} denotes the outward-pointing unit normal to Ω . Note that

$$\|\mathbf{a}\|_{W^{1-1/r, r}(\partial\Omega)} \leq |\beta_1| \|\phi_{\text{in}} \mathbf{n}\|_{W^{1-1/r, r}(\partial\Omega)} + |\beta_2| \|\phi_{\text{out}} \mathbf{n}\|_{W^{1-1/r, r}(\partial\Omega)} \leq C \|\boldsymbol{\beta}\|_{\mathbb{R}^2}$$

and

$$\|f\|_{L^r(\Omega)} \leq (|\beta_1| |\Gamma_{\text{in}}| + |\beta_2| |\Gamma_{\text{out}}|) / 2 \leq C \|\boldsymbol{\beta}\|_{\mathbb{R}^2}.$$

With \mathbf{a} and f given by (4.13), let \mathbf{v} be given by (3.6), (3.7), and $\mathbf{v}_{f,h} = I_h(\mathbf{v})|_{\overline{\Omega_f}}$, $\mathbf{v}_{p,h} = I_h(\mathbf{v})|_{\overline{\Omega_p}}$, where $I_h(\mathbf{v})$ denotes a continuous linear interpolant of \mathbf{v} with respect to $\mathcal{T}_{f,h} \cup \mathcal{T}_{p,h}$.

Note that $\mathbf{v}_h = (\mathbf{v}_{f,h}, \mathbf{v}_{p,h}) \in V_h$ and

$$\|\mathbf{v} - \mathbf{v}_h\|_{W^{s,r}(\Omega)} \leq Ch^{1-s} \|\mathbf{v}\|_{W^{1,r}(\Omega)}, \quad s = 0, 1,$$

$$\|\mathbf{v} - \mathbf{v}_h\|_{W^{0,r}(\partial\Omega)} \leq Ch^{r'} \|\mathbf{v}\|_{W^{1,r}(\Omega)}.$$

Then, for h sufficiently small,

$$\begin{aligned} & \sup_{\mathbf{w}_h \in X_h} \frac{\int_{\Gamma_{\text{in}}} \beta_1 \mathbf{w}_{f,h} \cdot \mathbf{n}_f \, ds + \int_{\Gamma_{\text{out}}} \beta_2 \mathbf{w}_{p,h} \cdot \mathbf{n}_p \, ds}{\|\mathbf{w}_h\|_X} \\ & \geq \frac{\int_{\Gamma_{\text{in}}} \beta_1 \mathbf{v}_{f,h} \cdot \mathbf{n}_f \, ds + \int_{\Gamma_{\text{out}}} \beta_2 \mathbf{v}_{p,h} \cdot \mathbf{n}_p \, ds}{\|\mathbf{v}_h\|_X} \\ & \geq \frac{\int_{\Gamma_{\text{in}}} \beta_1 \mathbf{v}_f \cdot \mathbf{n}_f \, ds + \int_{\Gamma_{\text{out}}} \beta_2 \mathbf{v}_p \cdot \mathbf{n}_p \, ds + \int_{\Gamma_{\text{in}}} \beta_1 (\mathbf{v}_{f,h} - \mathbf{v}_f) \cdot \mathbf{n}_f \, ds + \int_{\Gamma_{\text{out}}} \beta_2 (\mathbf{v}_{p,h} - \mathbf{v}_p) \cdot \mathbf{n}_p \, ds}{C \|\mathbf{v}\|_X} \\ & \geq C_1 \|\boldsymbol{\beta}\|_{\mathbb{R}^2} - C_2 h^{r'} \|\boldsymbol{\beta}\|_{\mathbb{R}^2}, \end{aligned}$$

from which (4.12) follows. \square

LEMMA 4.3. For h sufficiently small, there exists $C_{bh} > 0$ such that

$$(4.14) \quad \inf_{(0,0) \neq (q_h, \boldsymbol{\beta}) \in M_h \times \mathbb{R}^2} \sup_{\mathbf{v}_h \in V_h} \frac{b(\mathbf{v}_h, (q_h, \boldsymbol{\beta}))}{\|\mathbf{v}_h\|_X \|(q, \boldsymbol{\beta})\|_{M \times \mathbb{R}^2}} \geq C_{bh}.$$

Proof. Let $(p_h, \boldsymbol{\beta}) \in M_h \times \mathbb{R}^2$. From Lemma 4.2, there exists $\hat{\mathbf{u}}_h \in X_h$ such that

$$(4.15) \quad \|\hat{\mathbf{u}}_h\|_X = \|\boldsymbol{\beta}\|_{\mathbb{R}^m} \quad \text{and} \quad \frac{\int_{\Gamma_{\text{in}}} \beta_1 \mathbf{v}_{f,h} \cdot \mathbf{n}_f \, ds + \int_{\Gamma_{\text{out}}} \beta_2 \mathbf{v}_{p,h} \cdot \mathbf{n}_p \, ds}{\|\hat{\mathbf{u}}_h\|_X} \geq C_{RXh} \|\boldsymbol{\beta}\|_{\mathbb{R}^2}.$$

Consider the following two problems.

Problem 1. Discrete power law problem in Ω_f . Determine $\tilde{\mathbf{u}}_{f,h} \in X_{f,h}^0$, $\tilde{p}_{f,h} \in M_{f,h}$ such that

$$(4.16) \quad (|\mathbf{d}(\tilde{\mathbf{u}}_{f,h})|^{r-2} \mathbf{d}(\tilde{\mathbf{u}}_{f,h}), \mathbf{d}(\mathbf{v})) - (\tilde{p}_{f,h}, \nabla \cdot \mathbf{v}) = 0 \quad \forall \mathbf{v} \in X_{f,h}^0,$$

$$(4.17) \quad (q, \nabla \cdot \tilde{\mathbf{u}}_{f,h}) = (q, \|p_{f,h}\|_{M_f}^{1-r'/r} |p_{f,h}|^{r'/r-1} p_{f,h} - \nabla \cdot \hat{\mathbf{u}}_{f,h}) \quad \forall q \in M_{f,h}.$$

Problem 2. Modified Darcy problem in Ω_p . Determine $\tilde{\mathbf{u}}_{p,h} \in X_{p,h}^0$, $\tilde{p}_{p,h} \in M_{p,h}$ such that

$$(4.18) \quad (|\tilde{\mathbf{u}}_{p,h}|^{r-2} \tilde{\mathbf{u}}_{p,h}, \mathbf{v}) - (\tilde{p}_{p,h}, \nabla \cdot \mathbf{v}) = 0 \quad \forall \mathbf{v} \in X_{p,h}^0,$$

$$(4.19) \quad (q, \nabla \cdot \tilde{\mathbf{u}}_{p,h}) = (q, \|p_{p,h}\|_{M_p}^{1-r'/r} |p_{p,h}|^{r'/r-1} p_{p,h} - \nabla \cdot \hat{\mathbf{u}}_{p,h}) \quad \forall q \in M_{p,h}.$$

Note that

$$\|p_{j,h}\|_{M_j}^{1-r'/r} |p_{j,h}|^{r'/r-1} p_{j,h} - \nabla \cdot \hat{\mathbf{u}}_{j,h} \in L^r(\Omega_j), \quad j = f, p.$$

Existence and uniqueness of $\tilde{\mathbf{u}}_{f,h} \in X_{f,h}^0$, $\tilde{p}_{f,h} \in P_{f,h}$ and $\tilde{\mathbf{u}}_{p,h} \in X_{p,h}^0$, $\tilde{p}_{p,h} \in P_{p,h}$ satisfying (4.16), (4.17) and (4.18), (4.19), respectively, follow from the inf-sup conditions (4.6), (4.7) and the strong monotonicity of $T : X \rightarrow X^*$, $(T(\phi), \psi) := \int |\phi|^{r-2} \phi \cdot \psi \, dA$.

From (4.16) and (4.17), choosing $\mathbf{v} = \tilde{\mathbf{u}}_{f,h}$ and $q = \tilde{p}_{f,h}$,

$$\begin{aligned}
\|\tilde{\mathbf{u}}_{f,h}\|_{X_f}^r &= (|\mathbf{d}(\tilde{\mathbf{u}}_{f,h})|^{r-2} \mathbf{d}(\tilde{\mathbf{u}}_{f,h}), \mathbf{d}(\tilde{\mathbf{u}}_{f,h})) \\
&= (\tilde{p}_{f,h}, \nabla \cdot \tilde{\mathbf{u}}_{f,h}) \\
&= (\tilde{p}_{f,h}, \|p_{f,h}\|_{M_f}^{1-r'/r} |p_{f,h}|^{r'/r-1} p_{f,h} - \nabla \cdot \hat{\mathbf{u}}_{f,h}) \\
&\leq \|\tilde{p}_{f,h}\|_{M_f} \left(\|p_{f,h}\|_{M_f}^{1-r'/r} \| |p_{f,h}|^{r'/r-1} p_{f,h} \|_{L^r} + \|\nabla \cdot \hat{\mathbf{u}}_{f,h}\|_{L^r} \right) \\
&\leq \|\tilde{p}_{f,h}\|_{M_f} (\|p_{f,h}\|_{M_f} + C \|\hat{\mathbf{u}}_{f,h}\|_{X_f}) \\
(4.20) \quad &\leq C \|\tilde{p}_{f,h}\|_{M_f} (\|p_{f,h}\|_{M_f} + \|\boldsymbol{\beta}\|_{\mathbb{R}^2}).
\end{aligned}$$

Also, from the inf-sup condition for spaces $X_{f,h}^0$ and $M_{f,h}$ we have

$$\begin{aligned}
c \|\tilde{p}_{f,h}\|_{M_f} &\leq \sup_{\mathbf{v} \in X_{f,h}^0} \frac{(\tilde{p}_{f,h}, \nabla \cdot \mathbf{v})}{\|\mathbf{v}\|_{X_f}} \\
&= \sup_{\mathbf{v} \in X_{f,h}^0} \frac{(|\mathbf{d}(\tilde{\mathbf{u}}_{f,h})|^{r-2} \mathbf{d}(\tilde{\mathbf{u}}_{f,h}), \mathbf{d}(\mathbf{v}))}{\|\mathbf{v}\|_{X_f}} \\
&\leq \sup_{\mathbf{v} \in X_{f,h}^0} \frac{(\| |\mathbf{d}(\tilde{\mathbf{u}}_{f,h})|^{r-2} \mathbf{d}(\tilde{\mathbf{u}}_{f,h}) \|_{L^{r'}} \|\mathbf{d}(\mathbf{v})\|_{L^r})}{\|\mathbf{v}\|_{X_f}} \\
&= \| |\mathbf{d}(\tilde{\mathbf{u}}_{f,h})|^{r-2} \mathbf{d}(\tilde{\mathbf{u}}_{f,h}) \|_{L^{r'}} \\
(4.21) \quad &= \|\tilde{\mathbf{u}}_{f,h}\|_{X_f}^{r/r'}.
\end{aligned}$$

Combining (4.20) and (4.21) we have the estimate

$$(4.22) \quad \|\tilde{\mathbf{u}}_{f,h}\|_{X_f} \leq C (\|p_{f,h}\|_{M_f} + \|\boldsymbol{\beta}\|_{\mathbb{R}^2}).$$

Proceeding in a similar fashion for $\tilde{\mathbf{u}}_{p,h}$ satisfying Problem 2 leads to the estimate

$$(4.23) \quad \|\tilde{\mathbf{u}}_{p,h}\|_{X_p} \leq C (\|p_{p,h}\|_{M_p} + \|\boldsymbol{\beta}\|_{\mathbb{R}^2}).$$

Let $\mathbf{u}_{j,h} = \tilde{\mathbf{u}}_{j,h} + \hat{\mathbf{u}}_{j,h}$, $j = f, p$. Note that as $\mathbf{u}_{f,h} = \mathbf{0}$ on Γ and $\mathbf{u}_{p,h} \cdot \mathbf{n}_p = 0$ on Γ , $\mathbf{u}_h \in V_h$.

Then, using (4.17), (4.19), and (4.12),

$$\begin{aligned}
b(\mathbf{u}_h, (p_h, \boldsymbol{\beta})) &= \int_{\Omega_f} p_{f,h} \nabla \cdot \mathbf{u}_{f,h} \, dA + \int_{\Omega_p} p_{p,h} \nabla \cdot \mathbf{u}_{p,h} \, dA + \beta_1 \int_{\Gamma_{\text{in}}} \mathbf{u}_{f,h} \cdot \mathbf{n}_f \, ds \\
&\quad + \beta_2 \int_{\Gamma_{\text{out}}} \mathbf{u}_{p,h} \cdot \mathbf{n}_p \, ds \\
&= \int_{\Omega_f} p_{f,h} \|p_{f,h}\|_{M_f}^{1-r'/r} |p_{f,h}|^{r'/r-1} p_{f,h} \, dA \\
&\quad + \int_{\Omega_p} p_{p,h} \|p_{p,h}\|_{M_p}^{1-r'/r} |p_{p,h}|^{r'/r-1} p_{p,h} \, dA \\
&\quad + \beta_1 \int_{\Gamma_{\text{in}}} \hat{\mathbf{u}}_{f,h} \cdot \mathbf{n}_f \, ds + \beta_2 \int_{\Gamma_{\text{out}}} \hat{\mathbf{u}}_{p,h} \cdot \mathbf{n}_p \, ds \\
(4.24) \quad &\geq c (\|p_h\|_M^2 + \|\boldsymbol{\beta}\|_{\mathbb{R}^2}^2).
\end{aligned}$$

Thus, using (4.24), (4.22), and (4.23), we have

$$\begin{aligned} \sup_{\mathbf{v}_h \in X_h} \frac{b(\mathbf{v}_h, (p_h, \boldsymbol{\beta}))}{\|\mathbf{v}_h\|_X} &\geq \frac{b(\mathbf{u}_h, (p_h, \boldsymbol{\beta}))}{\|\mathbf{u}_h\|_X} \\ &\geq C (\|p_h\|_P + \|\boldsymbol{\beta}\|_{\mathbb{R}^2}), \end{aligned}$$

from which (4.14) immediately follows. \square

The discrete inf-sup condition for $b_I(\cdot, \cdot)$ follows from the continuous inf-sup condition and the existence of a bounded interpolation operator $I_{p,h} : X_p \rightarrow X_{h,p}$ satisfying, for some $\alpha > 0$,

$$(4.25) \quad \|\mathbf{w} - I_{p,h}(\mathbf{w}) \cdot \mathbf{n}_p\|_{W^{-1/r,r}(\partial\Omega_p)} \leq C_{ap} h^\alpha \|\mathbf{w}\|_{X_p} \quad \text{and} \quad \|I_{p,h}(\mathbf{w})\|_{X_p} \leq C_{ip} \|\mathbf{w}\|_{X_p}.$$

LEMMA 4.4. *There exists $C_{X\Gamma h} > 0$ such that for h sufficiently small*

$$(4.26) \quad \inf_{0 \neq \lambda_h \in L_h} \sup_{\mathbf{u}_h \in X_h} \frac{b_I(\mathbf{u}_h, \lambda_h)}{\|\mathbf{u}_h\|_X \|\lambda_h\|_{W^{1/r,r'}(\Gamma)}} \geq C_{X\Gamma h}.$$

Proof. With $\lambda = \lambda_h$, let $\mathbf{v}_p \in W^{0,r}(\text{div}, \Omega_p)$ be as defined by (3.16)–(3.20), and let $\mathbf{v}_{p,h} = I_{R-T}(\mathbf{v}_p) \in X_{p,h}$ denote the Raviart–Thomas interpolant of \mathbf{v}_p . Further, let $\mathbf{v}_h = (\mathbf{0}, \mathbf{v}_{p,h}) \in X_h$. Then

$$\begin{aligned} \sup_{\mathbf{u}_h \in X_h} \frac{b_I(\mathbf{u}_h, \lambda_h)}{\|\mathbf{u}_h\|_X} &\geq \frac{b_I(\mathbf{v}_h, \lambda_h)}{\|\mathbf{v}_h\|_X} \\ &= \frac{0 + \langle \mathbf{v}_{p,h} \cdot \mathbf{n}_p, \lambda_h \rangle_\Gamma}{\|\mathbf{v}_{p,h}\|_{W^{0,r}(\text{div}, \Omega_p)}} \\ &= \frac{\langle \mathbf{v}_p \cdot \mathbf{n}_p, \lambda_h \rangle_\Gamma}{\|\mathbf{v}_{p,h}\|_{W^{0,r}(\text{div}, \Omega_p)}} + \frac{\langle (\mathbf{v}_{p,h} - \mathbf{v}_p) \cdot \mathbf{n}_p, \lambda_h \rangle_\Gamma}{\|\mathbf{v}_{p,h}\|_{W^{0,r}(\text{div}, \Omega_p)}} \\ &\geq \frac{\langle \mathbf{v}_p \cdot \mathbf{n}_p, \lambda_h \rangle_\Gamma}{C \|\mathbf{v}_p\|_{W^{0,r}(\text{div}, \Omega_p)}} + \frac{\langle (\mathbf{v}_{p,h} - \mathbf{v}_p) \cdot \mathbf{n}_p, E_\Gamma^{r'} \lambda_h \rangle_{\partial\Omega_p}}{\|\mathbf{v}_{p,h}\|_{W^{0,r}(\text{div}, \Omega_p)}} \\ &\geq \frac{1}{2C} \|\lambda\|_{W^{1/r,r'}(\Gamma)} + \frac{\langle (\mathbf{v}_{p,h} - \mathbf{v}_p) \cdot \mathbf{n}_p, E_\Gamma^{r'} \lambda_h \rangle_{\partial\Omega_p}}{\|\mathbf{v}_{p,h}\|_{W^{0,r}(\text{div}, \Omega_p)}}. \end{aligned}$$

With $\lambda = \lambda_h$ let φ be given by (A.1)–(A.3), and let $\varphi_h = I(\varphi)$ denote a continuous linear interpolant of φ with respect to $\mathcal{T}_{p,h}$. Note that $\lambda_h = \varphi_h$ on Γ and Γ_{out} .

Now,

$$\begin{aligned} \langle (\mathbf{v}_{p,h} - \mathbf{v}_p) \cdot \mathbf{n}_p, E_\Gamma^{r'} \lambda_h \rangle_{\partial\Omega_p} &= \langle (\mathbf{v}_{p,h} - \mathbf{v}_p) \cdot \mathbf{n}_p, \varphi_h \rangle_{\partial\Omega_p} \\ &\quad + \langle (\mathbf{v}_{p,h} - \mathbf{v}_p) \cdot \mathbf{n}_p, (E_\Gamma^{r'} \lambda_h - \varphi_h) \rangle_{\partial\Omega_p} \\ &= 0 + \langle \mathbf{v}_{p,h} \cdot \mathbf{n}_p, (E_\Gamma^{r'} \lambda_h - \varphi_h) \rangle_{\partial\Omega_p} \\ &\quad - \langle \mathbf{v}_p \cdot \mathbf{n}_p, (E_\Gamma^{r'} \lambda_h - \varphi_h) \rangle_{\partial\Omega_p}. \end{aligned}$$

As $E_\Gamma^{r'} \lambda_h - \varphi_h = 0$ on $\partial\Omega_p \setminus \Gamma_p$ and $\mathbf{v}_p \cdot \mathbf{n}_p|_{\Gamma_p} = 0$, then $\langle \mathbf{v}_p \cdot \mathbf{n}_p, (E_\Gamma^{r'} \lambda_h - \varphi_h) \rangle_{\partial\Omega_p} = 0$. Further, as $\mathbf{v}_{p,h} \cdot \mathbf{n}_p = 0$ on Γ_p , $\langle \mathbf{v}_{p,h} \cdot \mathbf{n}_p, (E_\Gamma^{r'} \lambda_h - \varphi_h) \rangle_{\partial\Omega_p} = 0$, from which (4.26) then follows. \square

We now state and prove the existence and uniqueness of solutions to (4.8)–(4.9).

THEOREM 4.5. *There exists a unique solution $(\mathbf{u}_h, p_h, \lambda_h, \beta_h) \in X_h \times M_h \times L_h \times \mathbb{R}^2$ satisfying (4.8)–(4.9). In addition, there exists a constant $C > 0$ such that*

$$(4.27) \quad \|\mathbf{u}_h\|_X \leq C \left(\|\mathbf{f}_f\|_{X_f^*} + |fr| \right).$$

Proof. With the inf-sup conditions given in (4.14) and (4.26), the existence and uniqueness follows exactly as for the continuous problem in Theorem 3.3. The norm estimate for \mathbf{u}_h follows in a similar manner to that for \mathbf{u} and uses the property that $\nabla \cdot X_{p,h} \subset M_{p,h}$. \square

4.1. A priori error estimate. Next we investigate the error between the solution of the continuous variational formulation and its discrete counterpart.

THEOREM 4.6. *Let*

$$\begin{aligned} \mathcal{E}(\mathbf{u}, \mathbf{u}_h) &= \left\| \frac{|\mathbf{d}(\mathbf{u}_f) - \mathbf{d}(\mathbf{u}_{f,h})|}{c + |\mathbf{d}(\mathbf{u}_f)| + |\mathbf{d}(\mathbf{u}_{f,h})|} \right\|_{L^\infty(\Omega_f)}^{\frac{2-r}{r}} + \left\| \frac{|\mathbf{u}_p - \mathbf{u}_{p,h}|}{c + |\mathbf{u}_p| + |\mathbf{u}_{p,h}|} \right\|_{L^\infty(\Omega_f)}^{\frac{2-r}{r}} \quad \text{and} \\ \mathcal{G}(\mathbf{u}, \mathbf{u}_h) &= \int_{\Omega_f} |g_f(d(\mathbf{u}_f))d(\mathbf{u}_f) - g_f(d(\mathbf{u}_{f,h}))d(\mathbf{u}_{f,h})| |d(\mathbf{u}_f) - d(\mathbf{u}_{f,h})| \, dA \\ &\quad + \int_{\Omega_p} |g_p(\mathbf{u}_p)\mathbf{u}_p - g_p(\mathbf{u}_{p,h})\mathbf{u}_{p,h}| |\mathbf{u}_p - \mathbf{u}_{p,h}| \, dA. \end{aligned}$$

Then for $(\mathbf{u}, p, \lambda, \beta)$ satisfying (2.35)–(2.36) and $(\mathbf{u}_h, p_h, \lambda_h, \beta_h)$ satisfying (4.8)–(4.9), and h sufficiently small, there exists a constant $C > 0$ such that

$$(4.28) \quad \begin{aligned} \|\mathbf{u} - \mathbf{u}_h\|_X^2 + \mathcal{G}(\mathbf{u}, \mathbf{u}_h) &\leq C \left\{ \inf_{\mathbf{v}_h \in X_h} (\|\mathbf{u} - \mathbf{v}_h\|_X^2 + \mathcal{E}(\mathbf{u}, \mathbf{u}_h)^r \|\mathbf{u} - \mathbf{v}_h\|_X^r) \right. \\ &\quad \left. + \inf_{q_h \in M_h} \|p - q_h\|_M^2 + \inf_{\zeta_h \in L_h} \|\lambda - \zeta_h\|_{W^{1/r,r'}(\Gamma)} \right\}, \\ (4.29) \quad &\|p - p_h\|_M + \|\beta - \beta_h\|_{\mathbb{R}^2} + \|\lambda - \lambda_h\|_{W^{1/r,r'}(\Gamma)} \\ &\leq C \left\{ \mathcal{E}(\mathbf{u}, \mathbf{u}_h) \mathcal{G}(\mathbf{u}, \mathbf{u}_h)^{1/r'} + \inf_{q_h \in M_h} \|p - q_h\|_M + \inf_{\zeta_h \in L_h} \|\lambda - \zeta_h\|_{W^{1/r,r'}(\Gamma)} \right\}. \end{aligned}$$

Note that the constant C in Theorem 4.6 may depend upon $\|\mathbf{u}\|_X$.

The following *combined* inf-sup condition is used in the proof of Theorem 4.6.

LEMMA 4.7. *There exists a constant $C_c > 0$ such that*

$$(4.30) \quad \inf_{(0,0,0) \neq (q_h, \zeta_h, \gamma_h) \in M_h \times L_h \times \mathbb{R}^2} \sup_{\mathbf{v}_h \in X_h} \frac{b(\mathbf{v}_h, q_h, \gamma_h) - b_I(\mathbf{v}_h, \zeta_h)}{(\|q_h\|_M + \|\zeta_h\|_{W^{1/r,r'}(\Gamma)} + \|\gamma_h\|_{\mathbb{R}^2}) \|\mathbf{v}_h\|_X} \geq C_c.$$

Proof. As $b(\cdot, \cdot, \cdot)$ and $b_I(\cdot, \cdot)$ are continuous and satisfy inf-sup conditions (4.14) and (4.26), the inf-sup condition (4.30) follows immediately. (See Theorem B.1 in Appendix B.) \square

Proof of Theorem 4.6. Introduce the affine subspace \tilde{Z}_h defined by

$$\begin{aligned} \tilde{Z}_h &:= \{(q_h, \zeta_h, \gamma_h) \in M_h \times L_h \times \mathbb{R}^2 : -b(\mathbf{v}_h, q_h, \gamma_h) + b_I(\mathbf{v}_h, \zeta_h) \\ &\quad = (\mathbf{f}, \mathbf{v}_h) - a(\mathbf{u}_h, \mathbf{v}_h) \quad \forall \mathbf{v}_h \in X_h\}. \end{aligned}$$

Note that $(p_h, \lambda_h, \beta_h) \in \tilde{Z}_h$.

For $\mathbf{u}_{f,h}$, from (2.16)

$$\begin{aligned} & \frac{\|\mathbf{d}(\mathbf{u}_f) - \mathbf{d}(\mathbf{u}_{f,h})\|_{L^r(\Omega_f)}^2}{c + \|\mathbf{d}(\mathbf{u}_f)\|_{L^r(\Omega_f)}^{2-r} + \|\mathbf{d}(\mathbf{u}_{f,h})\|_{L^r(\Omega_f)}^{2-r}} \\ & + \int_{\Omega_f} |g_f(\mathbf{d}(\mathbf{u}_f))\mathbf{d}(\mathbf{u}_f) - g_f(\mathbf{d}(\mathbf{u}_{f,h}))\mathbf{d}(\mathbf{u}_{f,h})| |\mathbf{d}(\mathbf{u}_f) - \mathbf{d}(\mathbf{u}_{f,h})| \, dA \\ & \leq C \int_{\Omega_f} (g_f(\mathbf{d}(\mathbf{u}_f))\mathbf{d}(\mathbf{u}_f) - g_f(\mathbf{d}(\mathbf{u}_{f,h}))\mathbf{d}(\mathbf{u}_{f,h})) : (\mathbf{d}(\mathbf{u}_f) - \mathbf{d}(\mathbf{u}_{f,h})) \, dA \\ & = C \int_{\Omega_f} (g_f(\mathbf{d}(\mathbf{u}_f))\mathbf{d}(\mathbf{u}_f) - g_f(\mathbf{d}(\mathbf{u}_{f,h}))\mathbf{d}(\mathbf{u}_{f,h})) : (\mathbf{d}(\mathbf{u}_f) - \mathbf{d}(\mathbf{v}_{f,h})) \, dA \\ & \quad + C \int_{\Omega_f} (g_f(\mathbf{d}(\mathbf{u}_f))\mathbf{d}(\mathbf{u}_f) - g_f(\mathbf{d}(\mathbf{u}_{f,h}))\mathbf{d}(\mathbf{u}_{f,h})) : (\mathbf{d}(\mathbf{v}_{f,h}) - \mathbf{d}(\mathbf{u}_{f,h})) \, dA \\ & = I_1 + I_2. \end{aligned}$$

To estimate I_1 we use (2.17).

$$\begin{aligned} & \int_{\Omega_f} (g_f(\mathbf{d}(\mathbf{u}_f))\mathbf{d}(\mathbf{u}_f) - g_f(\mathbf{d}(\mathbf{u}_{f,h}))\mathbf{d}(\mathbf{u}_{f,h})) : (\mathbf{d}(\mathbf{u}_f) - \mathbf{d}(\mathbf{v}_{f,h})) \, dA \\ & \leq C \left(\int_{\Omega_f} |g_f(\mathbf{d}(\mathbf{u}_f))\mathbf{d}(\mathbf{u}_f) - g_f(\mathbf{d}(\mathbf{u}_{f,h}))\mathbf{d}(\mathbf{u}_{f,h})| |\mathbf{d}(\mathbf{u}_f) - \mathbf{d}(\mathbf{u}_{f,h})| \, dA \right)^{1/r'} \\ & \quad \cdot \left\| \frac{|\mathbf{d}(\mathbf{u}_f) - \mathbf{d}(\mathbf{u}_{f,h})|}{c + |\mathbf{d}(\mathbf{u}_f)| + |\mathbf{d}(\mathbf{u}_{f,h})|} \right\|_{\infty}^{\frac{2-r}{r}} \|\mathbf{d}(\mathbf{u}_f) - \mathbf{d}(\mathbf{v}_{f,h})\|_{L^r(\Omega_f)} \\ & \leq \epsilon_1 \int_{\Omega_f} |g_f(\mathbf{d}(\mathbf{u}_f))\mathbf{d}(\mathbf{u}_f) - g_f(\mathbf{d}(\mathbf{u}_{f,h}))\mathbf{d}(\mathbf{u}_{f,h})| |\mathbf{d}(\mathbf{u}_f) - \mathbf{d}(\mathbf{u}_{f,h})| \, dA \\ & \quad + C \left\| \frac{|\mathbf{d}(\mathbf{u}_f) - \mathbf{d}(\mathbf{u}_{f,h})|}{c + |\mathbf{d}(\mathbf{u}_f)| + |\mathbf{d}(\mathbf{u}_{f,h})|} \right\|_{\infty}^{\frac{2-r}{r}r} \|\mathbf{d}(\mathbf{u}_f) - \mathbf{d}(\mathbf{v}_{f,h})\|_{L^r(\Omega_f)}^r. \end{aligned}$$

Thus we have that

$$\begin{aligned} & \frac{\|\mathbf{d}(\mathbf{u}_f) - \mathbf{d}(\mathbf{u}_{f,h})\|_{L^r(\Omega_f)}^2}{c + \|\mathbf{d}(\mathbf{u}_f)\|_{L^r(\Omega_f)}^{2-r} + \|\mathbf{d}(\mathbf{u}_{f,h})\|_{L^r(\Omega_f)}^{2-r}} \\ & \quad + \int_{\Omega_f} |g_f(\mathbf{d}(\mathbf{u}_f))\mathbf{d}(\mathbf{u}_f) - g_f(\mathbf{d}(\mathbf{u}_{f,h}))\mathbf{d}(\mathbf{u}_{f,h})| |\mathbf{d}(\mathbf{u}_f) - \mathbf{d}(\mathbf{u}_{f,h})| \, dA \\ (4.31) \quad & \leq C \left\| \frac{|\mathbf{d}(\mathbf{u}_f) - \mathbf{d}(\mathbf{u}_{f,h})|}{c + |\mathbf{d}(\mathbf{u}_f)| + |\mathbf{d}(\mathbf{u}_{f,h})|} \right\|_{\infty}^{2-r} \|\mathbf{d}(\mathbf{u}_f) - \mathbf{d}(\mathbf{v}_{f,h})\|_{L^r(\Omega_f)}^r + I_2. \end{aligned}$$

Similarly, we obtain that for $\mathbf{v}_{p,h} \in X_{p,h}$

$$\begin{aligned} & \frac{\|\mathbf{u}_p - \mathbf{u}_{p,h}\|_{L^r(\Omega_p)}^2}{c + \|\mathbf{u}_p\|_{L^r(\Omega_p)}^{2-r} + \|\mathbf{u}_{p,h}\|_{L^r(\Omega_p)}^{2-r}} + \int_{\Omega_p} |g_p(\mathbf{u}_p)\mathbf{u}_p - g_p(\mathbf{u}_{p,h})\mathbf{u}_{p,h}| |\mathbf{u}_p - \mathbf{u}_{p,h}| \, dA \\ (4.32) \quad & \leq C \left\| \frac{|\mathbf{u}_p - \mathbf{u}_{p,h}|}{c + |\mathbf{u}_p| + |\mathbf{u}_{p,h}|} \right\|_{\infty}^{2-r} \|\mathbf{u}_p - \mathbf{v}_{p,h}\|_{L^r(\Omega_p)}^r + I_4, \end{aligned}$$

where I_4 is given by

$$I_4 := C \int_{\Omega_p} (g_p(\mathbf{u}_p)\mathbf{u}_p - g_p(\mathbf{u}_{p,h})\mathbf{u}_{p,h}) : (\mathbf{v}_{p,h} - \mathbf{u}_{p,h}) \, dA.$$

Note that with $\mathbf{v}_h = (\mathbf{v}_{f,h}, \mathbf{v}_{p,h})$, $I_2 + I_4 = a(\mathbf{u}, \mathbf{v}_h - \mathbf{u}_h) - a(\mathbf{u}_h, \mathbf{v}_h - \mathbf{u}_h)$, and for $(q_h, \zeta_h, \gamma_h) \in \tilde{Z}_h$,

$$\begin{aligned} & a(\mathbf{u}, \mathbf{v}_h - \mathbf{u}_h) - a(\mathbf{u}_h, \mathbf{v}_h - \mathbf{u}_h) \\ &= b(\mathbf{v}_h - \mathbf{u}_h, p, \boldsymbol{\beta}) - b_I(\mathbf{v}_h - \mathbf{u}_h, \lambda) - b(\mathbf{v}_h - \mathbf{u}_h, p_h, \boldsymbol{\beta}_h) \\ &\quad + b_I(\mathbf{v}_h - \mathbf{u}_h, \lambda_h) \\ &= b(\mathbf{v}_h - \mathbf{u}_h, p, \boldsymbol{\beta}) - b_I(\mathbf{v}_h - \mathbf{u}_h, \lambda) \quad (\text{as } (p_h, \lambda_h, \boldsymbol{\beta}_h) \in \tilde{Z}_h) \\ &= b(\mathbf{v}_h - \mathbf{u}_h, p - q_h, \boldsymbol{\beta} - \boldsymbol{\gamma}_h) - b_I(\mathbf{v}_h - \mathbf{u}_h, \lambda - \zeta_h) \\ &= b(\mathbf{u} - \mathbf{u}_h, p - q_h, \boldsymbol{\beta} - \boldsymbol{\gamma}_h) - b(\mathbf{u} - \mathbf{v}_h, p - q_h, \boldsymbol{\beta} - \boldsymbol{\gamma}_h) \\ &\quad - b_I(\mathbf{u} - \mathbf{u}_h, \lambda - \zeta_h) + b_I(\mathbf{u} - \mathbf{v}_h, \lambda - \zeta_h) \\ &\leq \epsilon \|\mathbf{u} - \mathbf{u}_h\|_X^2 \\ (4.33) \quad &+ C \left(\|\mathbf{u} - \mathbf{v}_h\|_X^2 + \|p - q_h\|_M^2 + \|\lambda - \zeta_h\|_{W^{1/r, r'}(\Gamma)}^2 \right). \end{aligned}$$

In the last step of (4.33) we use the continuity of the operators $b(\cdot, \cdot, \cdot)$ and $b_I(\cdot, \cdot)$.

Combining (4.31)–(4.33) and the fact that $\nabla \cdot X_{p,h} \subset M_{p,h}$, we obtain the estimate (4.28) for $(q_h, \zeta_h, \gamma_h) \in \tilde{Z}_h$. The inf-sup condition (4.30) then enables (q_h, ζ_h, γ_h) to be lifted from \tilde{Z}_h to $M_h \times L_h \times \mathbb{R}^2$. (See [5] for details.)

To establish (4.29) we begin with the inf-sup condition (4.30).

$$\begin{aligned} & \|p_h - q_h\|_M + \|\boldsymbol{\beta}_h - \boldsymbol{\gamma}_h\|_{\mathbb{R}^2} + \|\lambda_h - \zeta_h\|_{W^{1/r, r'}(\Gamma)} \\ &\leq C \frac{b(\mathbf{v}_h, (p_h - q_h), (\boldsymbol{\beta}_h - \boldsymbol{\gamma}_h)) - b_I(\mathbf{v}_h, \lambda_h - \zeta_h)}{\|\mathbf{v}_h\|_X} \\ &\leq C \left(\frac{b(\mathbf{v}_h, (p - q_h), (\boldsymbol{\beta} - \boldsymbol{\gamma}_h)) - b_I(\mathbf{v}_h, \lambda - \lambda_h)}{\|\mathbf{v}_h\|_X} \right. \\ &\quad \left. - \frac{b(\mathbf{v}_h, (p - p_h), (\boldsymbol{\beta} - \boldsymbol{\beta}_h)) - b_I(\mathbf{v}_h, \lambda - \zeta_h)}{\|\mathbf{v}_h\|_X} \right) \\ &\leq C \left(\|p - q_h\|_M + \|\boldsymbol{\beta} - \boldsymbol{\gamma}_h\|_{\mathbb{R}^2} + \|\lambda_h - \zeta_h\|_{W^{1/r, r'}(\Gamma)} - \frac{a(\mathbf{u}, \mathbf{v}_h) - a(\mathbf{u}_h, \mathbf{v}_h)}{\|\mathbf{v}_h\|_X} \right) \\ &\leq C \left(\|p - q_h\|_M + \|\boldsymbol{\beta} - \boldsymbol{\gamma}_h\|_{\mathbb{R}^2} + \|\lambda_h - \zeta_h\|_{W^{1/r, r'}(\Gamma)} + \mathcal{E}(\mathbf{u}, \mathbf{u}_h) \mathcal{G}(\mathbf{u}, \mathbf{u}_h)^{1/r'} \right). \end{aligned} \tag{4.34}$$

Combining (4.34) with the triangle inequality, we obtain (4.29). □

Appendix A. Extension operator from Γ to $\partial\Omega$. Let Ω be a bounded Lipschitz domain in \mathbb{R}^n ($n = 2$ or 3), and let $\partial\Omega = \bar{\Gamma} \cup \bar{\Gamma}_b \cup \bar{\Gamma}_d$, where Γ , Γ_b , and Γ_d are pairwise disjoint and $\text{dist}(\Gamma, \Gamma_d) > 0$. Additionally, let $\Gamma^c = \partial\Omega \setminus \Gamma$.

We use standard notation to denote the function spaces used, for example, $W^{s,p}(\Omega)$, $W^{1,p}(\partial\Omega)$, etc., with $W_0^{-l,q}(\partial\Omega)$ denoting the dual space of $W_0^{l,p}(\partial\Omega)$, where q is the unitary conjugate of p , i.e., $1/q := 1 - 1/p$.

The expression $A \preceq B$ is used to denote the inequality $A \leq (\text{constant}) \cdot B$.

Next we investigate a suitable extension of a function λ defined on Γ to a function defined on $\partial\Omega$.

Assume that $p \geq 2$.

LEMMA A.1. *Given $\lambda \in W^{1/q,p}(\Gamma)$ define $E_\Gamma^p \lambda := \gamma_0 \varphi$, where γ_0 is the trace operator from $W^{1,p}(\Omega)$ to $W^{1/q,p}(\partial\Omega)$, and $\varphi \in W^{1,p}(\Omega)$ is the weak solution to*

$$(A.1) \quad -\nabla \cdot |\nabla \varphi|^{p-2} \nabla \varphi = 0 \quad \text{in } \Omega,$$

$$(A.2) \quad \varphi = \begin{cases} \lambda & \text{on } \Gamma, \\ 0 & \text{on } \Gamma_d, \end{cases}$$

$$(A.3) \quad |\nabla \varphi|^{p-2} \partial_{\mathbf{n}} \varphi = 0 \quad \text{on } \Gamma_b.$$

Then $E_\Gamma^p \lambda \in W^{1/q,p}(\partial\Omega)$, and $\|E_\Gamma^p \lambda\|_{W^{1/q,p}(\partial\Omega)} \preceq \|\lambda\|_{W^{1/q,p}(\Gamma)}$.

Proof. The proof follows from the strong monotonicity [19] of the operator $\mathcal{L} : X \rightarrow X^*$, $\mathcal{L}(u) := -\nabla \cdot |\nabla u|^{p-2} \nabla u$, where $X = \{f \in W^{1,p}(\Omega) : f|_{\Gamma \cup \Gamma_d} = 0\}$ [23]. \square

For $\lambda \in W^{1/q,p}(\Gamma)$, let $E_{00,\Gamma}^p \lambda$ denote the extension of λ by zero on Γ^c .

Remark. Note that $E_{00,\Gamma}^p \lambda \in W^{1/q,p}(\partial\Omega)$ if and only if $\lambda \in W_0^{1/q,p}(\Gamma)$.

LEMMA A.2 (see [9]). *For $\zeta \in W^{1/q,p}(\partial\Omega)$ there exist $\zeta_\Gamma \in W^{1/q,p}(\Gamma)$ and $\zeta_{\Gamma^c} \in W_0^{1/q,p}(\Gamma^c)$ such that $\zeta = E_\Gamma^p \zeta_\Gamma + E_{00,\Gamma^c}^p \zeta_{\Gamma^c}$. Moreover, this decomposition is unique.*

Proof. Let $\zeta \in W^{1/q,p}(\partial\Omega)$. Define, $\zeta_\Gamma := \zeta|_\Gamma$ and $\zeta_{\Gamma^c} := \xi|_{\Gamma^c}$, where $\xi := \zeta - E_\Gamma^p \zeta_\Gamma$. Note that $\zeta|_\Gamma \in W^{1/q,p}(\Gamma)$ and

$$\|E_\Gamma^p \zeta_\Gamma\|_{W^{1/q,p}(\partial\Omega)} \preceq \|\zeta_\Gamma\|_{W^{1/q,p}(\Gamma)} \leq \|\zeta\|_{W^{1/q,p}(\partial\Omega)},$$

and hence $\xi \in W^{1/q,p}(\partial\Omega)$. Also, $E_{00,\Gamma^c}^p \zeta_{\Gamma^c} = \xi$ as ζ and $E_\Gamma^p \zeta_\Gamma$ agree on Γ . Thus, from the remark above, $\zeta_{\Gamma^c} \in W_0^{1/q,p}(\Gamma^c)$.

To show uniqueness of the decomposition, observe that if $0 = E_\Gamma^p \zeta_\Gamma + E_{00,\Gamma^c}^p \zeta_{\Gamma^c}$, then ζ_Γ is the trace of the weak solution of (A.1)–(A.3) for $\lambda = 0$. Hence $\zeta_\Gamma = 0$. \square

Next we introduce the concept of *the restriction of an operator* in $W^{-1/q,q}(\partial\Omega)$ to be equal to zero.

DEFINITION A.3 (see [9]). *If $f \in W^{-1/q,q}(\partial\Omega)$, then $f|_{\Gamma^c} = 0$ means by definition that*

$$(A.4) \quad \langle f, E_{00,\Gamma^c}^p \xi \rangle_{\partial\Omega} = 0 \quad \forall \xi \in W_0^{1/q,p}(\Gamma^c).$$

The following lemma describes how an operator in $W^{-1/q,q}(\partial\Omega)$ can be decomposed into an operator in $W^{-1/q,q}(\Gamma)$ and an operator in $W^{-1/q,q}(\Gamma^c)$.

LEMMA A.4 (see [9]). *For $f \in W^{-1/q,q}(\partial\Omega)$ there exists $f_\Gamma \in W^{-1/q,q}(\Gamma)$ and $f_{\Gamma^c} \in W_0^{-1/q,q}(\Gamma^c)$ such that for $\zeta \in W^{1/q,p}(\partial\Omega)$, with $\zeta = E_\Gamma^p \zeta_\Gamma + E_{00,\Gamma^c}^p \zeta_{\Gamma^c}$, as defined in Lemma A.2, we have*

$$(A.5) \quad \langle f, \zeta \rangle_{\partial\Omega} = \langle f_\Gamma, \zeta_\Gamma \rangle_\Gamma + \langle f_{\Gamma^c}, \zeta_{\Gamma^c} \rangle_{\Gamma^c}.$$

Proof. For $\zeta_\Gamma \in W^{1/q,p}(\Gamma)$ and $\zeta_{\Gamma^c} \in W_0^{1/q,p}(\Gamma^c)$, define

$$(A.6) \quad \langle f_\Gamma, \zeta_\Gamma \rangle_\Gamma := \langle f, E_\Gamma^p \zeta_\Gamma \rangle_{\partial\Omega} \quad \text{and} \quad \langle f_{\Gamma^c}, \zeta_{\Gamma^c} \rangle_{\Gamma^c} := \langle f, E_{00,\Gamma^c}^p \zeta_{\Gamma^c} \rangle_{\partial\Omega}.$$

Then

$$\langle f_\Gamma, \zeta_\Gamma \rangle_\Gamma \leq \|f\|_{W^{-1/q,q}(\partial\Omega)} \|E_\Gamma^p \zeta_\Gamma\|_{W^{1/q,p}(\partial\Omega)} \preceq \|f\|_{W^{-1/q,q}(\partial\Omega)} \|\zeta_\Gamma\|_{W^{1/q,p}(\Gamma)},$$

and thus $f_\Gamma \in W^{-1/q,q}(\Gamma)$. Analogously, $f_{\Gamma^c} \in W_0^{-1/q,q}(\Gamma^c)$. Additionally,

$$\langle f_\Gamma, \zeta_\Gamma \rangle_\Gamma + \langle f_{\Gamma^c}, \zeta_{\Gamma^c} \rangle_{\Gamma^c} = \langle f, E_\Gamma^p \zeta_\Gamma \rangle_{\partial\Omega} + \langle f, E_{00,\Gamma^c}^p \zeta_{\Gamma^c} \rangle_{\partial\Omega} = \langle f, \zeta \rangle_{\partial\Omega}. \quad \square$$

Note that for $f \in W^{-1/q,q}(\partial\Omega)$ with $f|_{\Gamma^c} = 0$ (see Definition A.3), from (A.6),

$$(A.7) \quad \langle f, \zeta \rangle_{\partial\Omega} = \langle f_\Gamma, \zeta_\Gamma \rangle_\Gamma \quad \forall \zeta \in W^{1/q,p}(\partial\Omega).$$

Thus functionals in $W^{-1/q,q}(\partial\Omega)$ which are zero when restricted to $\partial\Omega \setminus \Gamma$ can be identified with functionals in $W^{-1/q,q}(\Gamma)$.

Appendix B. Combined inf-sup conditions. In deriving a priori error estimates for mixed methods, whose analysis relies on several inf-sup conditions, combined inf-sup conditions are needed. In this section we show that the required inf-sup conditions follow readily from the continuity of the bilinear forms and the individual inf-sup conditions.

THEOREM B.1. *Let V, Q_1, Q_2 be Banach spaces, and let $b_1(\cdot, \cdot) : V \times Q_1 \rightarrow \mathbb{R}$, $b_2(\cdot, \cdot) : V \times Q_2 \rightarrow \mathbb{R}$, and $Z_1 := \{v \in V \mid b_1(v, q) = 0 \ \forall q \in Q_1\}$. Assume that $b_2(\cdot, \cdot)$ is continuous and there exist $\beta_1, \beta_2 > 0$ such that*

$$\begin{aligned} \sup_{v \in V, \|v\|_V=1} b_1(v, q_1) &\geq \beta_1 \|q_1\|_{Q_1} \quad \forall q_1 \in Q_1, \\ \sup_{v \in Z_1, \|v\|_V=1} b_2(v, q_2) &\geq \beta_2 \|q_2\|_{Q_2} \quad \forall q_2 \in Q_2. \end{aligned}$$

Then there exists $\beta > 0$ such that

$$\sup_{v \in V, \|v\|_V=1} (b_1(v, q_1) + b_2(v, q_2)) \geq \beta (\|q_1\|_{Q_1} + \|q_2\|_{Q_2}) \quad \forall (q_1, q_2) \in Q_1 \times Q_2.$$

Proof. By the continuity of $b_2(\cdot, \cdot)$, there exists $C_2 > 0$ such that

$$b_2(v, q_2) \leq C_2 \|v\|_V \|q_2\|_{Q_2} \quad \forall (v, q_2) \in V \times Q_2.$$

Let $(q_1, q_2) \in Q_1 \times Q_2$ be given, and choose $v_1 \in V$ with $\|v_1\|_V = 1$ and $v_2 \in Z_1$ with $\|v_2\|_V = 1$, satisfying

$$b_1(v_1, q_1) \geq \frac{\beta_1}{2} \|q_1\|_{Q_1}, \quad b_2(v_2, q_2) \geq \frac{\beta_2}{2} \|q_2\|_{Q_2}.$$

Then for $u = v_1 + (1 + 2C_2/\beta_2)v_2$ we have

$$\begin{aligned} b_1(u, q_1) &= b_1(v_1, q_1) \geq \frac{\beta_1}{2} \|q_1\|_{Q_1}, \\ b_2(u, q_2) &= b_2(v_1, q_2) + \left(1 + \frac{2C_2}{\beta_2}\right) b_2(v_2, q_2) \geq \frac{\beta_2}{2} \|q_2\|_{Q_2}. \end{aligned}$$

Finally, as $\|u\|_V \leq 2(1 + 2C_2/\beta_2)$, with $u_0 = u/\|u\|_V$

$$b_1(u_0, q_1) + b_2(u_0, q_2) \geq \beta (\|q_1\|_{Q_1} + \|q_2\|_{Q_2}),$$

where $\beta = \min\{\beta_1, \beta_2\}/(4(1 + C_2/\beta_2))$. \square

COROLLARY B.2. *Let $Z_0, Q_i, i = 1, \dots, n$, be Banach spaces, and let $b_i(\cdot, \cdot) : Z_0 \times Q_i \rightarrow \mathbb{R}, i = 1, \dots, n$, and $Z_i := \{v \in Z_{i-1} \mid b_i(v, q) = 0 \forall q \in Q_i\}, i = 1, \dots, n-1$. Assume that $b_i(\cdot, \cdot)$ is continuous and there exist β_i such that*

$$\sup_{v \in Z_{i-1}, \|v\|_{Z_0}=1} b_i(v, q) \geq \beta_i \|q\|_{Q_i} \quad \forall q \in Q_i, i = 1, \dots, n.$$

Then there exists $\beta > 0$ such that

(B.1)

$$\sup_{v \in Z_0, \|v\|_{Z_0}=1} \sum_{i=1}^n b_i(v, q_i) \geq \beta (\|q_1\|_{Q_1} + \dots + \|q_n\|_{Q_n}) \quad \forall (q_1, \dots, q_n) \in Q_1 \times \dots \times Q_n.$$

Proof. The proof of (B.1) follows from Theorem B.1 and by induction. \square

Acknowledgment. The authors would like to thank the referees for their helpful suggestions.

REFERENCES

- [1] G. BEAVERS AND D. JOSEPH, *Boundary conditions at a naturally impermeable wall*, J. Fluid Mech., 30 (1967), pp. 197–207.
- [2] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
- [3] S.-S. CHOW AND G. CAREY, *Numerical approximation of generalized Newtonian fluids using Powell–Sabin–Heindl elements: I. Theoretical elements*, Internat. J. Numer. Methods Fluids, 41 (2003), pp. 1085–1118.
- [4] M. DISCACCIATI, E. MIGLIO, AND A. QUATERONI, *Mathematical and numerical models for coupling surface and groundwater flows*, Appl. Numer. Math., 43 (2002), pp. 57–74.
- [5] V. J. ERVIN AND H. LEE, *Numerical approximation of a quasi-Newtonian Stokes flow problem with defective boundary conditions*, SIAM J. Numer. Anal., 45 (2007), pp. 2120–2140.
- [6] V. ERVIN AND T. PHILLIPS, *Residual a posteriori error estimator for a three-field model of a non-linear generalized Stokes problem*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 2599–2610.
- [7] L. FORMAGGIA, J.-F. GERBEAU, F. NOBILE, AND A. QUATERONI, *Numerical treatment of defective boundary conditions for the Navier–Stokes equations*, SIAM J. Numer. Anal., 40 (2002), pp. 376–401.
- [8] G. GALDI, *An Introduction to the Mathematical Theory of the Navier–Stokes Equations*, Vol. 1, Springer-Verlag, New York, 1994.
- [9] J. GALVIS AND M. SARKIS, *Non-matching mortar discretization analysis for the coupling Stokes–Darcy equations*, Electron. Trans. Numer. Anal., 26 (2007), pp. 350–384.
- [10] V. GIRAULT AND P. RAVIART, *Finite Element Methods for Navier–Stokes Equations*, Springer-Verlag, Berlin, 1986.
- [11] M. D. GUNZBURGER AND S. L. HOU, *Treating inhomogeneous essential boundary conditions in finite element methods and the calculation of the boundary stresses*, SIAM J. Numer. Anal., 29 (1992), pp. 390–424.
- [12] N. HANSPAL, A. WAGHODE, V. NASSEHI, AND R. WAKEMAN, *Numerical analysis of coupled Stokes/Darcy flow in industrial filtrations*, Transp. Porous Media, 64 (2006), pp. 73–101.
- [13] W. JÄGER AND A. MIKELIĆ, *On the interface boundary condition of Beavers, Joseph, and Saffman*, SIAM J. Appl. Math., 60 (2000), pp. 1111–1127.
- [14] W. J. LAYTON, F. SCHIEWECK, AND I. YOTOV, *Coupling fluid flow with porous media flow*, SIAM J. Numer. Anal., 40 (2003), pp. 2195–2218.
- [15] X. LOPEZ, P. VALVATNE, AND M. BLUNT, *Predictive network modeling of single-phase non-Newtonian flow in a porous media*, J. Colloid Interface Sci., 264 (2003), pp. 256–265.

- [16] M. MU AND J. XU, *A two-grid method of a mixed Stokes–Darcy model for coupling fluid flow with porous media flow*, SIAM J. Numer. Anal., 45 (2007), pp. 1801–1813.
- [17] R. OWENS AND T. PHILLIPS, *Computational Rheology*, Imperial College Press, London, 2002.
- [18] J. PEARSON AND P. TARDY, *Models for flow of non-Newtonian and complex fluids through porous media*, J. Non-Newtonian Fluid Mech., 102 (2002), pp. 447–473.
- [19] M. RENARDY AND R. ROGERS, *An Introduction to Partial Differential Equations*, Springer-Verlag, New York, 1993.
- [20] B. RIVIÈRE, *Analysis of a discontinuous finite element method for the coupled Stokes and Darcy problems*, J. Sci. Comput., 22/23 (2005), pp. 479–500.
- [21] B. RIVIÈRE AND I. YOTOV, *Locally conservative coupling of Stokes and Darcy flows*, SIAM J. Numer. Anal., 42 (2005), pp. 1959–1977.
- [22] P. SAFFMAN, *On the boundary condition at the surface of a porous media*, Stud. Appl. Math., 50 (1971), pp. 93–101.
- [23] D. SANDRI, *On the numerical approximation of quasi-Newtonian flows whose viscosity obeys a power law or the Carreau law*, RAIRO Modél. Math. Anal. Numér., 27 (1993), pp. 131–155.
- [24] R. VERFÜRTH, *Finite element approximation of incompressible Navier-Stokes equations with slip boundary condition*, Numer. Math., 50 (1987), pp. 697–721.

ON OPTIMAL CONVERGENCE RATE OF THE RATIONAL KRYLOV SUBSPACE REDUCTION FOR ELECTROMAGNETIC PROBLEMS IN UNBOUNDED DOMAINS*

LEONID KNIZHNERMAN[†], VLADIMIR DRUSKIN[‡], AND MIKHAIL ZASLAVSKY[‡]

Abstract. We solve an electromagnetic frequency domain induction problem in \mathbf{R}^3 for a frequency interval using rational Krylov subspace (RKS) approximation. The RKS is constructed by spanning on the solutions for a certain a priori chosen set of frequencies. We reduce the problem of the optimal choice of these frequencies to the third Zolotaryov problem in the complex plane, having an approximate closed form solution, and determine the best Cauchy–Hadamard convergence rate. The theory is illustrated with numerical examples for Maxwell’s equations arising in 3D magnetotelluric geophysical exploration.

Key words. frequency domain problems, Galerkin method, third Zolotaryov problem in complex plane

AMS subject classifications. 30C85, 30E10, 41A05, 41A20, 65M60, 86-08

DOI. 10.1137/080715159

1. Introduction. Many boundary value problems can be reduced to computation of

$$u = f(A)\varphi,$$

where A is an operator in a Hilbert space, and u and φ are elements of the same space. In practice A can be a large ill-conditioned matrix obtained after discretization of a PDE operator, which is why it is convenient to consider A as an unbounded operator.

The resolvent

$$f(\lambda) = \frac{1}{\lambda + s}$$

is one of the most commonly used functions appearing in the solution of linear non-stationary equations in the frequency domain.

As an important practical application, we consider the direct problem of electromagnetic frequency sounding arising in geophysical prospecting. It can be reduced to the magnetic field formulation of the frequency-domain Maxwell equations in \mathbf{R}^3 in the low frequency regime (displacement currents are assumed to be negligible)

$$(1.1) \quad \nabla \times (\mu\sigma)^{-1} \nabla \times H + i\omega H = \nabla \times \sigma^{-1} J$$

with zero boundary conditions at infinity. Here H is the vector magnetic field induced by an external current J , ω is a frequency, μ is the magnetic permeability (which is assumed to be constant throughout the whole domain), and $c_1 \leq \sigma \leq c_2$ is variable electrical conductivity distribution, where c_1 and c_2 are positive constants. We solve

*Received by the editors February 6, 2008; accepted for publication (in revised form) August 7, 2008; published electronically February 13, 2009.

<http://www.siam.org/journals/sinum/47-2/71515.html>

[†]Central Geophysical Expedition, House 38, Building 3, Narodnogo opolcheniya St., Moscow, 123298 Russia (mmd@cge.ru).

[‡]Schlumberger Doll Research, 1 Hampshire St., Cambridge, MA 02139 (druskin1@slb.com, mzaslavsky@slb.com).

the resolvent problem with $s = i\omega$, $A = A^* = \nabla \times (\mu\sigma)^{-1}\nabla \times$ and $\varphi = \nabla \times \sigma^{-1}J$. Maxwell's operator $\nabla \times (\mu\sigma)^{-1}\nabla \times$ in unbounded domains has a continuum (without holes) spectrum supported on the entire $\mathbf{R}_+ = [0, +\infty]$ [33, section 9].

Usually, the electromagnetic field is measured on for $\omega \in [\omega_{\min}, \omega_{\max}]$; i.e., the resolvent must be computed for multiple values of s corresponding to this interval.

Two of the authors solved these problems using the so-called spectral Lanczos decomposition method (SLDM), which is Galerkin method on a Krylov subspace $K_m(A, \varphi)$ [6]. Similar approaches (with different names) were used in, e.g., [27, 26, 35, 9, 18]; however, the basic idea first appeared in the classical work of Hestenes and Stiefel [17]. The SLDM allows one to compute the resolvent for many frequencies with the cost of a single frequency problem using unpreconditioned conjugate gradients, and the time domain solution converges even asymptotically faster than the frequency domain solution [6]. However, the SLDM convergence was strongly affected by the condition number of the discrete problem and frequency range.

Spectral adaptation of Krylov methods and efficiency of rational approximation can be combined in the so-called rational Krylov subspaces (RKS) [30]. The approximate solution is projected onto an RKS, which is a span of different rational functions of A applied to φ . Let us consider a subdiagonal RKS in the generic form:

$$(1.2) \quad U_n = \text{span}\{b, Ab, \dots, A^{n-1}b\}, \quad b = \prod_{j=1}^n (A + s_j I)^{-1} \varphi.$$

Obviously, $(A + s_j I)^{-1} \varphi \in U_n$; i.e., the solution of the resolvent problem with $s = s_j$ is exactly approximated on U_n , so the shifts s_j are also called interpolating points. We assume that the RKS is computed using iterative methods for which there are no computational advantages to solving multiple linear systems with the same shifts (because of extensive memory requirements for the discretization of large scale electromagnetic problems in geophysics); i.e., we assume that s_j do not coincide. The RKS is widely used in model reduction, in particular for computation of transfer functions of linear problems; see reviews [3, 8] for details.

The question is, what is the optimal convergence rate with such an approach, and how do we choose s_j to achieve it? For unbounded frequency intervals the interpolating frequencies can be obtained using the H_2 -optimality conditions [23] by computing a sequence of Krylov subspaces [15]. In this work we consider bounded intervals, for which we compute optimal rates and corresponding interpolating points using the L_∞ -optimality condition.

The key of our approach is presenting the Galerkin solution as a particular case of the so-called skeleton approximation $f_{\text{skel}}(A, s)\varphi$, where $f_{\text{skel}}(\lambda, s)$ is a rational function of λ and s introduced in [34, 28]. The optimization of the error of the skeleton approximation can be reduced to the famous third Zolotaryov problem with asymptotically optimal s_j computed in terms of elliptic integrals. Given a bounded positive frequency interval, the computed interpolation points provide convergence with the optimal Cauchy–Hadamard rate for the class of operators with continuum spectrum supported on entire \mathbf{R}_+ and with a regular enough spectral measure.

2. Formulation of the problem. RKS Galerkin method. We compute action of the resolvent operator

$$(2.1) \quad u = (A + sI)^{-1} \varphi, \quad A \geq 0,$$

where A is a self-adjoint nonnegative definite operator acting in a Hilbert space H equipped with an inner product $\langle \cdot, \cdot \rangle$, and φ is a normalized vector from this space.

We assume that A has a continuum (without holes) spectrum supported on the entire \mathbf{R}_+ .

We assume that $s \in S$, where S is a compact subset of the complex plain not intersecting the real negative semiaxis. Should we have a solution u_s for a complex parameter s , we automatically also have the solution for the conjugate parameter \bar{s} as \bar{u}_s , so without loss of generality we can assume that S is symmetric with respect to the real axis.

Choose noncoinciding parameters $s_j \in S$, symmetric with respect to the real axis, $1 \leq j \leq n$, and construct RKS (1.2). Due to the continuity of the A 's spectrum the corresponding spectral measure has infinite number of increase points, so $\dim U_n = n$. To approximately solve (2.1), we will use Galerkin approximation on U_n . The Galerkin solution $\tilde{u} \in U_n$ satisfies the equalities

$$(2.2) \quad \langle (A + sI)\tilde{u}, v \rangle = \langle \varphi, v \rangle \quad \forall v \in U_n.$$

We construct a well-conditioned basis $G^n = \{g_1, \dots, g_n\}$ of U_n with the help of a recursive algorithm. There are many ways to construct G^n . They are known generically by the name rational Arnoldi method (see, e.g., [30, 14]). In our numerical experiments we implement the following well-known simple variant of rational Arnoldi. Set

$$g_1 = \frac{(A + s_1I)^{-1}\varphi}{\|(A + s_1I)^{-1}\varphi\|}.$$

Let $2 \leq l \leq n$ and g_1, \dots, g_{l-1} have been calculated. Then the vector g_l is obtained by the Gram–Schmidt orthogonalization of $(A + s_lI)^{-1}g_{l-1}$ to $g_j, j = 1, \dots, l - 1$. Usually, the most computationally expensive part of rational Arnoldi is the solution of shifted linear systems.

3. RKS Galerkin method and the third Zolotaryov problem in the complex plane.

3.1. RKS Galerkin method and skeleton approximants. Let $\mu(\lambda)$ be the spectral measure, associated with the couple (A, φ) . Using Parseval’s identity, we obtain $\langle f(A)\varphi, g(A)\varphi \rangle = \langle f, g \rangle_\mu$, where

$$\langle f, g \rangle_\mu = \int_0^{+\infty} \overline{g(\lambda)}f(\lambda) d\mu(\lambda).$$

Scalarizing the problem, i.e., considering it in the spectral coordinates, we will seek the Galerkin approximant $\tilde{w} \in V_n$ to the function

$$\frac{1}{\lambda + s}, \quad \lambda \in \mathbf{R}, \quad \lambda \geq 0, \quad s \in S,$$

where V_n is the spectral counterpart of U_n from (1.2) defined as

$$V_n = \text{span} \left\{ \frac{1}{q_n}, \frac{\lambda}{q_n}, \dots, \frac{\lambda^{n-1}}{q_n} \right\}, \quad q_n(\lambda) = \prod_{l=1}^n (\lambda + s_l).$$

The Galerkin solution $\tilde{v} \in V_n$ satisfies the equation

$$(3.1) \quad \langle v, (\lambda + s)\tilde{v} - 1 \rangle_\mu = 0 \quad \forall v \in V_n.$$

Problem (3.1) has a unique solution. Obviously, $(\lambda + s_l I)^{-1} \varphi \in V_n$, so they are the solutions of (3.1) for $s = s_l$, the points s_l being the interpolation ones of \tilde{v} as a function of s .

Let θ_j and $Z_j \in V_n$, $j = 1, \dots, n$, be, respectively, the Ritz values and (normalized) Ritz “vectors” (which are actually functions of λ) satisfying

$$(3.2) \quad \langle v, (\lambda - \theta_j) Z_j \rangle_\mu = 0 \quad \forall v \in V_n.$$

This problem (for the operator of multiplication by λ in $L_{2,\mu}$ and the trial subspace V_n) is Hermitian, so θ_j are positive and Z_j are orthonormal. The Galerkin solution can be presented via spectral decomposition as

$$(3.3) \quad \tilde{v} = \sum_{j=1}^n (\theta_j + s)^{-1} \langle Z_j, 1 \rangle_\mu Z_j.$$

By construction s_l are either real positive or have a complex conjugate counterpart in S , and thus $q_n(\lambda) > 0$ for $\lambda \in \mathbf{R}_+$, i.e., on the A ’s spectrum. So (3.1), (3.2), (3.3) can be equivalently considered as the polynomial problem with respect to $q_n \tilde{v}$ instead of \tilde{v} on the subspace $K_n = \text{span}\{1, \lambda, \dots, \lambda^{n-1}\}$ instead of V_n and spectral measure ρ instead of μ , where $d\rho(\lambda) = q_n(\lambda)^{-2} d\mu(\lambda)$. This allows us to apply to our rational approximant the known results from the theory of orthogonal polynomials (see [5]). First, we note that θ_j are the nodes of a Gaussian quadrature, and as such they don’t coincide. Also, (3.3) can be viewed as the Lagrange polynomial interpolating $\frac{q_n}{\lambda+s}$ at θ_j (with respect to λ).

So, we can summarize the interpolation properties of \tilde{v} as a function of λ and s in the following lemma.

LEMMA 3.1. *We have*

$$\left(\tilde{v} - \frac{1}{\lambda + s} \right) \Big|_{s=s_l} = 0, \quad \lambda \geq 0, \quad l = 1, \dots, n,$$

and

$$\left(\tilde{v} - \frac{1}{\lambda + s} \right) \Big|_{\lambda=\theta_l} = 0, \quad s \in S, \quad l = 1, \dots, n.$$

The so-called skeleton approximation of functions of two variables was introduced in [34] and then used in [12, 16]. This approximation for the function $1/(x + y)$ was investigated in [28]. This function is defined as

$$(3.4) \quad f_{\text{skel}}(\lambda, s) = \left(\frac{1}{\lambda + s_1}, \dots, \frac{1}{\lambda + s_n} \right) M^{-1} \begin{pmatrix} \frac{1}{s + \lambda_1} \\ \vdots \\ \frac{1}{s + \lambda_n} \end{pmatrix},$$

where $M = (M_{kl})$ is the $n \times n$ matrix with the entries $M_{kl} = 1/(\lambda_k + s_l)$.

Theorem 3 from [28] for our case can be written as

$$(3.5) \quad \delta = \left[\frac{1}{\lambda + s} - f_{\text{skel}}(\lambda, s) \right] / \frac{1}{\lambda + s} = \prod_{j=1}^n \frac{\lambda - \lambda_j}{\lambda + s_j} \cdot \prod_{j=1}^n \frac{s - s_j}{s + \lambda_j};$$

i.e., λ_j and s_j are interpolating points. Both \tilde{v} and f_{skel} are $(n-1)/n$ rational functions of λ and of s , so from Lemma 3.1 and (3.5) we obtain the following proposition.

PROPOSITION 3.2. *If $\theta_j = \lambda_j, j = 1, \dots, n$, then*

$$\tilde{v} \equiv f_{\text{skel}}.$$

The relative interpolation error, i.e., the left-hand side of (3.5), can be written as

$$\delta = \frac{r(\lambda)}{r(-s)}, \quad r(z) = \prod_{j=1}^n \frac{z - \lambda_j}{z + s_j}.$$

Introduce the quantity

$$(3.6) \quad \sigma_n(\mathbf{R}_+, -S) \equiv \min_{\lambda_1, \dots, \lambda_n, s_1, \dots, s_n} \frac{\max_{\lambda \geq 0} |r(\lambda)|}{\min_{z \in -S} |r(z)|}.$$

As will be discussed in detail later, minimization problem (3.6) is a partial case of the third Zolotaryov problem in the complex plane, and it has an asymptotically (in the Cauchy–Hadamard sense) best solution with $\lambda_j \in \mathbf{R}_+$ and $s_j \in S$, such that $\lambda_l \neq \lambda_j$ and $s_l \neq s_j$ if $l \neq j$.

We will use s_j obtained from (3.6) to construct the Galerkin subspace U . Optimal λ_j may differ from the Ritz values θ_j , but the Galerkin error can still be estimated via $\sigma_n(\mathbf{R}_+, -S)$.

PROPOSITION 3.3. *We have an estimate*

$$(3.7) \quad \left\| \frac{1}{\lambda + s} - \tilde{v} \right\|_{\mu} \leq 2 \frac{\sigma_n(\mathbf{R}_+, S)}{\text{dist}(\mathbf{R}_+, S)}.$$

Proof. For any λ_j and s_j ($j = 1, \dots, n$) obtained from the solution of Zolotaryov problem (3.6), $f_{\text{skel}}(\lambda, s) \in V$ and

$$(\lambda + s)f_{\text{skel}}(\lambda, s) = 1 - \delta(\lambda, s),$$

so $f_{\text{skel}}(\lambda, s)$ is the solution of the modified Galerkin problem

$$\langle v, (\lambda + s)f_{\text{skel}}(\lambda, s) - 1 + \delta(\lambda, s) \rangle_{\mu} = 0 \quad \forall v \in V.$$

Obviously,

$$f_{\text{skel}}(\lambda, s) = (\lambda + s)^{-1}[1 - \delta(\lambda, s)],$$

so

$$\left\| \frac{1}{\lambda + s} - f_{\text{skel}} \right\|_{\mu} = \|(\lambda + s)^{-1}\delta(\lambda, s)\|_{\mu} \leq \|(\lambda + s)^{-1}\|_{\mu} \|\delta(\lambda, s)\|_{\mu}.$$

From the identities $\|\varphi\| = \|1\|_{\mu} = \int_0^{\infty} d\mu = 1$ we get

$$\|\delta(\lambda, s)\|_{\mu} \leq \max_{\lambda \in \mathbf{R}_+} |\delta(\lambda, s)|.$$

For the optimal δ obtained with the help of (3.6) we obtain

$$(3.8) \quad \|\delta(\lambda, s)\|_{\mu} \leq \sigma_n(\mathbf{R}_+, -S)$$

and

$$(3.9) \quad \left\| \frac{1}{\lambda + s} - f_{\text{skel}} \right\|_{\mu} \leq \frac{\sigma_n(\mathbf{R}_+, -S)}{\text{dist}(\mathbf{R}_+, S)}.$$

Again, for any λ_j and s_j the spectral decomposition gives

$$\|f_{\text{skel}} - \tilde{v}\|_{\mu} = \left\| \sum_{j=1}^n (\theta_j + s)^{-1} \langle Z_j, \delta \rangle_{\mu} Z_j \right\| = \sqrt{\sum_{j=1}^n |(\theta_j + s)^{-2} \langle Z_j, \delta \rangle_{\mu}^2|}.$$

So, with the optimal δ obtained with the help of (3.6), using (3.8) and real positivity of θ_j , we infer

$$\|f_{\text{skel}} - \tilde{v}\|_{\mu} \leq \frac{\sigma_n(\mathbf{R}_+, -S)}{\text{dist}(\mathbf{R}_+, S)}.$$

Using this estimate, (3.9), and the triangle inequality, we obtain (3.7). \square

Obviously, the error of the Galerkin approximate cannot be smaller than the optimal error measured in the same norm, so we have a lower bound for the relative error of the Galerkin approximant in spectral coordinates as

$$\|(\lambda + s)\tilde{v} - 1\|_{L_{\infty}(\mathbf{R}_+)} \geq \sigma_n(\mathbf{R}_+, -S).$$

Thus, we have both the upper L_2 and lower L_{∞} error norms of order $\sigma_n(\mathbf{R}_+, -S)$. So it is natural to expect that Proposition 3.3 gives a sharp bound in the Cauchy–Hadamard sense and that ω_j are close to optimal in the same sense.

It follows from Parseval’s identity that the Galerkin error in the L_2 norm can be computed as

$$\|u - \tilde{u}\| = \left\| \frac{1}{\lambda + s} - \tilde{v} \right\|_{\mu} = \sqrt{\int_0^{\infty} |\tilde{v} - (\lambda + s)^{-1}|^2 d\mu(\lambda)}.$$

The Galerkin method can improve the convergence speed due to adaptation to the nonuniformity of μ . However, for the class of operators with regular enough spectral measures, supported on the entire \mathbf{R}_+ , the spectral adaptation cannot improve the Cauchy–Hadamard convergence rate.

3.2. The third Zolotaryov problem in the complex plane. Minimization problem (3.6) is a partial case of the third Zolotaryov problem in the complex plane (see [10] or [36, section 8.7]). This problem in relation to the alternating direction implicit (ADI) method was investigated in [22, 7, 20, 32]. Generally this problem can be solved numerically with the use of the Remez algorithm. In particular, we are interested in cases when

$$S = -S = D = i[\omega_{\min}, \omega_{\max}] \cup (-i)[\omega_{\min}, \omega_{\max}].$$

Such a problem arises in geophysical prospecting with low frequency electromagnetic sources (see the numerical examples). For these cases we shall calculate the asymptotical convergence factor and give a closed form approximate solution.

Let $\frac{\omega_{\min}}{\omega_{\max}} = 1 - \kappa^2$, $0 < \kappa < 1$.

Introduce the full elliptic integral of modulus \varkappa ,

$$K(\varkappa) = \int_0^1 \frac{dt}{\sqrt{(1-t^2)(1-\varkappa^2 t^2)}}.$$

THEOREM 3.4. *With the number*

$$(3.10) \quad \rho = \exp \left[-\frac{\pi K(\sqrt{1 - \kappa^2})}{2K(\kappa)} \right]$$

the following assertions are valid:

$$(3.11) \quad \sigma_n(\mathbf{R}_+, D) \geq \rho^n, \quad n \in \mathbf{N},$$

$$(3.12) \quad \lim_{n \rightarrow \infty} \sqrt[n]{\sigma_n(\mathbf{R}_+, D)} = \rho.$$

We shall give a proof of Theorem 3.4 in the appendix.

Later on we assume that the number of frequencies n is even. In practice, we work with functional spaces over \mathbf{C} , the operator A , and the right-hand-side vector φ being real. In such a situation, should we obtain the solution u for a frequency ω , the solution for the frequency $-\omega$ is just \bar{u} . Thus we can reckon that frequencies ω and $-\omega$ belong to the compact D simultaneously. In this case D is symmetric with respect to \mathbf{R} .

The proof of [36, section 8.7, Theorem 9] in conjunction with the maxim from [10, section 5, paragraph 1] says how parameters ω_j and λ_j should be asymptotically distributed on D and \mathbf{R}_+ , respectively, for approximation (3.4) to be optimal in the Cauchy–Hadamard sense. Since the measure β (see (A.12)) is equilibrium on D to Ω , we have taken

$$(3.13) \quad \frac{\omega_j}{\omega_{\max}} = 1 - (1 - \kappa^2) \operatorname{sn} \left(\frac{2j - 1}{n}, \kappa \right)^2, \quad \omega_{\frac{n}{2}+j} = -\omega_j, \quad j = 1, \dots, \frac{n}{2},$$

so on each connected component of D the parameters ω_j are asymptotically distributed as interpolation nodes of corresponding Zolotaryov approximants.

Remark 1. Optimal (in the Cauchy–Hadamard sense) parameters λ_j/ω_{\max} can be found as the roots U of the equations

$$(3.14) \quad \frac{1}{2K(\kappa)} \int_{1-\kappa^2}^1 \left[\arctan \left(\sqrt{\frac{2U}{v}} - 1 \right) + \arctan \left(\sqrt{\frac{2U}{v}} + 1 \right) \right] \frac{dv}{\sqrt{(v - 1 + \kappa^2)v(1 - v)}} = \frac{(j - 0.5)\pi}{n}, \quad j = 1, \dots, n.$$

But these parameters are not exploited in our reduced order models since we use Galerkin formulation (2.2) and its Ritz values may differ from optimal λ_j .

CONJECTURE 1. *Given (3.13) and (3.14), one can explicitly (in the Zolotaryov style) present the quantities $\max_{z \in D} |r(z)^{-1}|$ and $\max_{\lambda \geq 0} |r(\lambda)|$ in terms of elliptic functions and obtain the upper bound*

$$\sigma_n(\mathbf{R}_+, D) = O(\rho^n),$$

where ρ is defined by formula (3.10).

For the case when $\kappa \rightarrow 1 - 0$ it is possible to obtain an asymptotical formula for ρ containing only elementary functions. In fact, in this case $\kappa < 1$ tends to 1 and the formulae

$$K(\kappa) = \frac{1}{2} \log \frac{16}{1 - \kappa^2} + o(1)$$

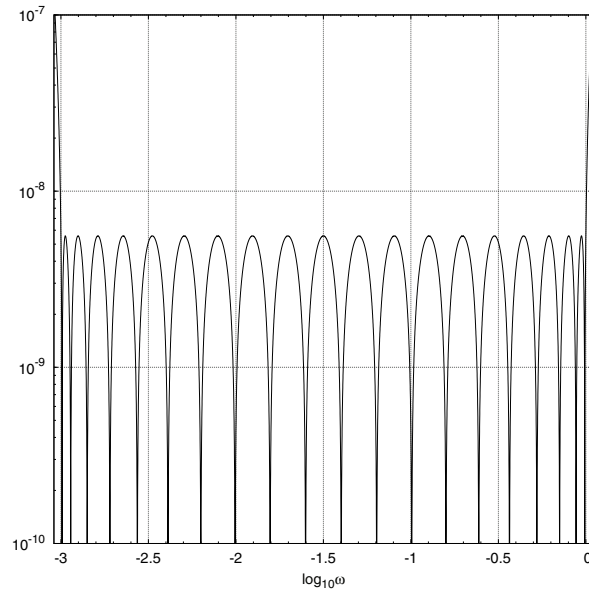


FIG. 1. $0.001 \leq \omega \leq 1$, $n = 40$, the error $\frac{\max_{\lambda \geq 0} |r(\lambda)|}{|r(i\omega)|}$.

(see [1, (17.3.26)]) and

$$K(\sqrt{1 - \kappa^2}) = \frac{\pi}{2} + o(1)$$

enable us to transform (3.10) into the expression

$$(3.15) \quad \rho = \exp \left[-\frac{\frac{\pi^2}{2} + o(1)}{\log \frac{\omega_{\max}}{\omega_{\min}} + \log 16} \right].$$

In Figures 1 and 2 we show the plots of the error $\frac{\max_{\lambda \geq 0} |r(\lambda)|}{|r(i\omega)|}$ for Zolotaryov approximants as functions of ω for $n = 40$ and 60 , respectively. The error graphs show almost equal ripples on the prescribed spectral interval, which, by analogy with the Chebyshev real approximation theory, enables us to conjecture that our approximants are almost the best.

4. Numerical experiments. We consider the direct problem of magnetotelluric geophysical exploration. The electromagnetic field excited by the Sun propagates into the Earth. Using the Fourier transform (transfer function) of the measured field, geophysicists determine underground distribution of conductivity σ , and the direct problem constitutes in the solution of (1.1) for a given frequency interval. In the geophysical exploration the problem is considered in the conductive inhomogeneous half-space with horizontal plane source at $+\infty$. We deal with the plane electric wave polarized along a horizontal (x) direction for the frequency interval from 0.01 Hz to 15 Hz. The measurements are the ratios of x -component of electric and y -component of magnetic fields (impedances) taken at the plane $z = 300$ m. In our experiments we estimated the relative L_2 norm of the error on the plane.

As was already mentioned, the most computationally expensive part of rational

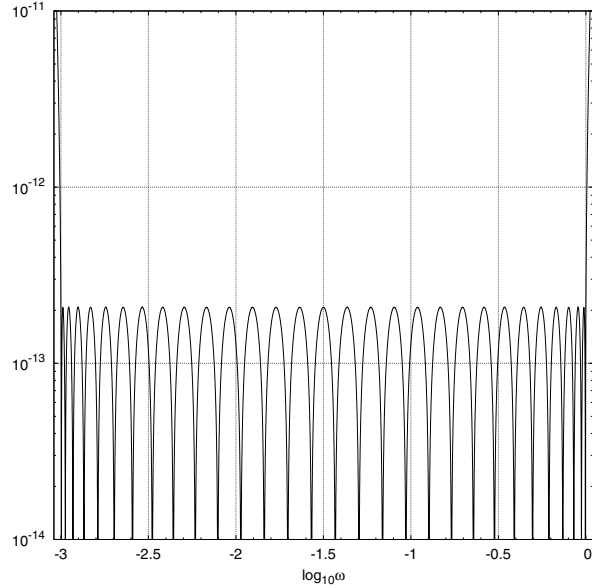


FIG. 2. $0.001 \leq \omega \leq 1$, $n = 60$, the error $\frac{\max_{\lambda \geq 0} |r(\lambda)|}{|r(i\omega)|}$.

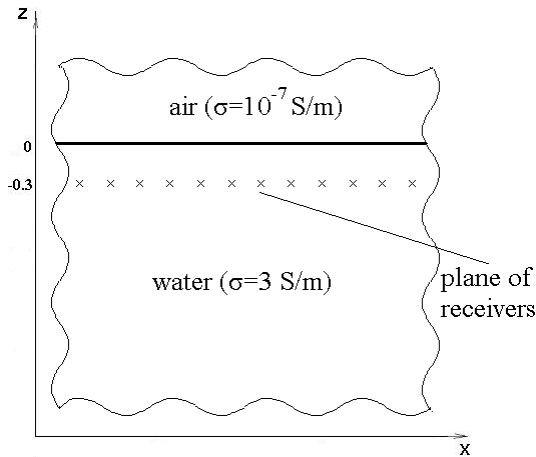


FIG. 3. Medium for test 1: A homogeneous conductive half-space.

Arnoldi is the solution of shifted linear systems. We used for this purpose a preconditioned Krylov subspace (QMR) solver [37].

In the first test we consider the homogeneous half-space shown in Figure 3. Figure 4 shows the comparison of frequency distribution of the errors for geometric and Zolotaryov grids for test 1 with $n = 16$. The geometric grid is the most common ad hoc grid used in applications. Indeed, Zolotaryov's grids are superior. However, for large $\omega_{\max}/\omega_{\min} \gg n$ the zeros of a Zolotaryov approximant's error are visually close to a geometric progression, and the convergence rate of the approximant, based on the

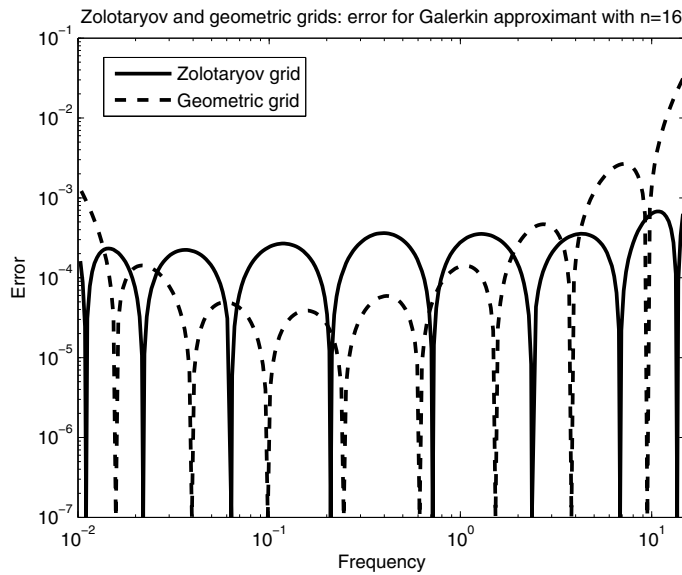


FIG. 4. Test 1: Error distribution for the geometric and Zolotaryov grids, $\omega_{\min} = 0.01, \omega_{\max} = 15$.

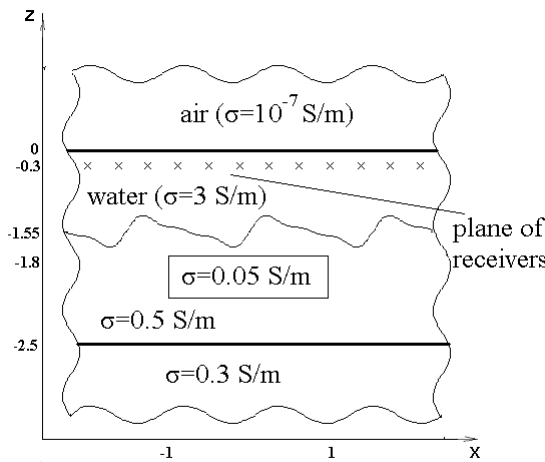


FIG. 5. Medium for test 2.

geometric progression grids, approaches that of the optimal (Zolotaryov's) one [19]. However, as we see from the graphs, the error distribution for the Zolotaryov grid is more uniform than the one for the geometric grid on $[\omega_{\min}, \omega_{\max}]$, which results in slightly better accuracy in the $L_{\infty}[\omega_{\min}, \omega_{\max}]$ norm.

In test 2 we consider a more complicated medium consisting of a resistive target (oil reservoir) embedded under the sea bottom of variable depth (see Figure 5). The spectral distribution for this problem varies more than for the previous one (though still without holes in the spectral measure's support), so both Zolotaryov and geomet-

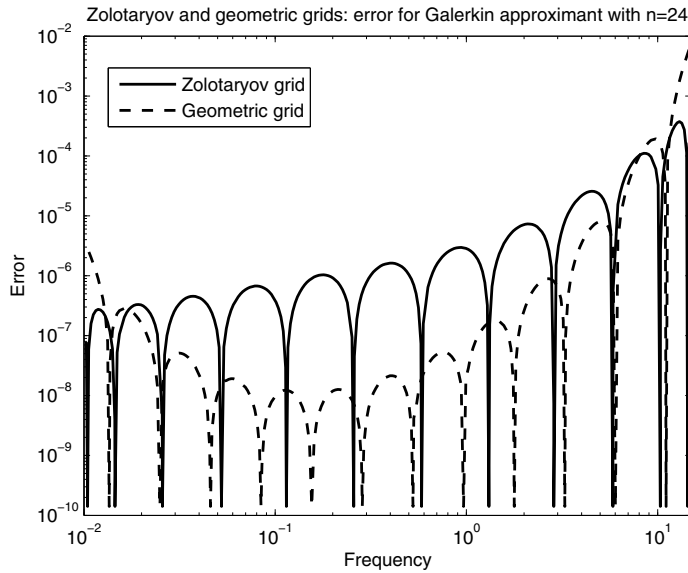


FIG. 6. Test 2: Error distribution for the geometric and Zolotaryov grids, $\omega_{\min} = 0.01, \omega_{\max} = 15$.

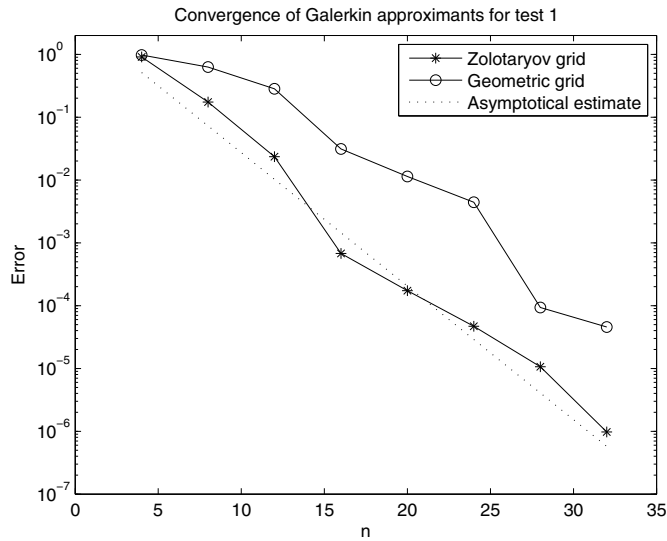


FIG. 7. Convergence for Zolotaryov and geometric grids (test 1) and comparison with theoretical results.

ric progression exhibit more nonuniform error distribution, but the Zolotaryov error remains more uniform and smaller in the $L_\infty[\omega_{\min}, \omega_{\max}]$ norm (see Figure 6).

In Figures 7 and 8 we show the errors (for both the grids) in the $L_\infty[\omega_{\min}, \omega_{\max}]$ norm as functions of n for tests 1 and 2, respectively. For both tests the Zolotaryov grid slightly overperforms the geometric one, and the average slopes of the Zolotaryov error curves are in good agreement with the asymptotic estimate determined by (3.15).

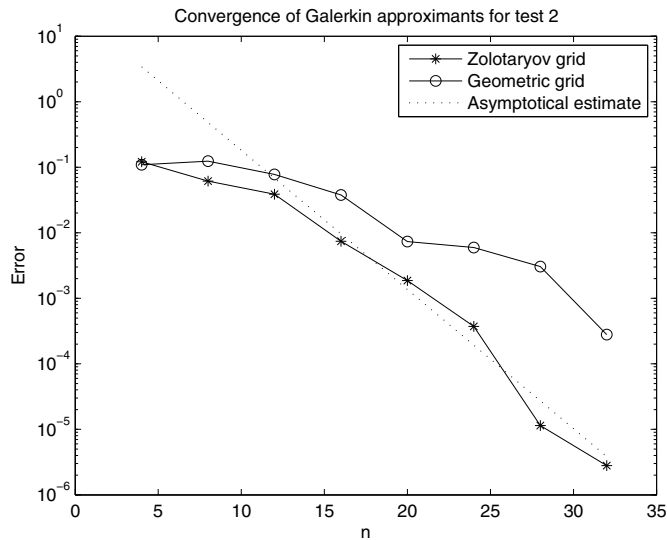


FIG. 8. Convergence for Zolotaryov and geometric grids (test 2) and comparison with theoretical results.

The asymptotic estimate is computed as

$$c \exp \left(-n \frac{\frac{\pi^2}{2}}{\log \frac{\omega_{\max}}{\omega_{\min}} + \log 16} \right),$$

with a constant c chosen to fit the actual Zolotaryov error.

For $n = 32$ it took 35 minutes of computer time on a PC with a Pentium IV 2 GHz processor to solve the problem from test 2 (our preconditioner allows us to obtain the exact solution after just one QMR iteration for test 1) with 6 digits of accuracy. For comparison, the same task took 32450 steps and 252 minutes of computer time for the SLDM. So the RKS reduction significantly overperforms the SLDM, but not without drawbacks. The RKS reduction requires additional memory to store G^n and a priori knowledge of the Krylov subspace dimension n .

5. Conclusive remarks.

- The problem of optimization of rational Krylov subspaces (RKS) for computation of the resolvent of self-adjoint operators can be reduced to the third Zolotaryov problem in the complex plane.
- This problem can be asymptotically solved in a closed form for a bounded positive frequency interval.
- The numerical experiments confirm the theoretical results for the models from geophysical applications.
- We are looking into possibilities of extension of the developed approach to non-Hermitian operators and the computation of exponentials and other functions of operators.
- A drawback of the developed approach is that the dimension of the rational Krylov subspace should be known a priori. We are planning to address this issue in our future research.

Appendix. Proof of Theorem 3.4 and auxiliary assertions. In subsection A.1 we shall establish properties of the Green function for the domain $\mathbf{C} \setminus \mathbf{R}_-$; the relation between values on \mathbf{R}_+ and on $i\mathbf{R}$ is the key point. In subsection A.2 we shall compare the corresponding potentials of two measures supported, respectively, on \mathbf{R} and $i\mathbf{R}$. This will enable us to express the asymptotical convergence factor of our (complex) third Zolotaryov problem through that of the classical (real) problem studied by Zolotaryov himself.

A.1. Green’s function.

Remark 2. Due to technical reasons, we prefer to handle the condenser (\mathbf{R}_-, D) instead of (\mathbf{R}_+, D) . Of course, $\sigma_n(\mathbf{R}_-, D) = \sigma_n(\mathbf{R}_+, D)$ because of the symmetry.

Removing from the complex plane the support \mathbf{R}_- of the measure, generating the Markov function

$$(A.1) \quad z^{-1/2} = \int_{-\infty}^0 \frac{1}{\pi\sqrt{-x}}(z-x)^{-1} dx, \quad z \notin \mathbf{R}_-$$

(see [4, part 1, section 2.2, p. 47]), we obtain the domain $\Omega = \mathbf{C} \setminus \mathbf{R}_-$.

According to a definition from [25, Chapter 5, section 5] or [31, section A.V], Green’s function (of two variables) for Ω

$$g_\Omega(z, x), \quad z, x \in \Omega,$$

is the one satisfying the following conditions: (1) the function $g_\Omega(z, x)$ as a function of z is harmonic in the domain $\Omega \setminus \{x\}$; (2) the function

$$g_\Omega(z, x) - \log \frac{1}{z-x}$$

is bounded in some vicinity of a point x ; (3) the limit value of $g_\Omega(z, x)$ as z tends to a point from \mathbf{R}_- is zero.

LEMMA A.1. *Green’s function (of two variables) for the domain Ω is expressed by the formula*

$$(A.2) \quad g_\Omega(z, x) = \log \left| \frac{\sqrt{z} + \sqrt{x}}{\sqrt{z} - \sqrt{x}} \right|, \quad z, x \in \Omega.$$

Proof. It is known [25, Chapter 5, section 5] that

$$(A.3) \quad g_\Omega(z, x) = \log |\phi(z, x)|, \quad z, x \in \Omega,$$

where with a fixed argument x the slice $z \mapsto \phi(z, x)$ conformally maps $\Omega \cup \{\infty\}$ onto the exterior to the unit circle in $\overline{\mathbf{C}}$ in such a way that $\phi(x, x) = \infty$. We shall build ϕ as a composition of the following conformal mappings:

$$(A.4) \quad z \mapsto \frac{\sqrt{z} - 1}{\sqrt{z} + 1}$$

transforms [21, p. 428] Ω into the open unit circle;

$$(A.5) \quad z \mapsto \frac{z-a}{1-\bar{a}z}, \quad |a| < 1,$$

transforms [24, p. 104] the open unit circle into itself; the inversion

$$(A.6) \quad z \mapsto \frac{1}{z}$$

transforms the open unit circle into the exterior to the open unit circle. We shall choose the parameter value

$$a = \frac{\sqrt{x} - 1}{\sqrt{x} + 1},$$

so that

$$(A.7) \quad \frac{1 + a}{1 - a} = \frac{\sqrt{x} + 1 + \sqrt{x} - 1}{\sqrt{x} + 1 - \sqrt{x} + 1} = \sqrt{x}.$$

Composing the mappings (A.4)–(A.6) and accounting (A.7), we obtain

$$\begin{aligned} \phi(z, x) &= \frac{1 - \bar{a} \frac{\sqrt{z}-1}{\sqrt{z}+1}}{\frac{\sqrt{z}-1}{\sqrt{z}+1} - a} = \frac{\sqrt{z} + 1 - \bar{a}(\sqrt{z} - 1)}{\sqrt{z} - 1 - a(\sqrt{z} + 1)} \\ &= \frac{(1 - \bar{a})\sqrt{z} + (1 + \bar{a})}{(1 - a)\sqrt{z} - (1 + a)} = \frac{1 - \bar{a}}{1 - a} \cdot \frac{\sqrt{z} + \sqrt{x}}{\sqrt{z} - \sqrt{x}}, \end{aligned}$$

which in conjunction with (A.3) follows (A.2). \square

Remark 2. Notwithstanding that representation (A.2) is unsymmetric, it is easy to see that the symmetry property

$$g_{\Omega}(z, x) = g_{\Omega}(x, z), \quad z, x \in \Omega,$$

holds.

LEMMA A.2. *If $u, v \in \mathbf{R}$, $u, v > 0$, then*

$$(A.8) \quad g_{\Omega}(ui, vi) + g_{\Omega}(ui, -vi) = g_{\Omega}(u, v).$$

Proof. Indeed, we derive from (A.2)

$$\begin{aligned} g_{\Omega}(ui, vi) + g_{\Omega}(ui, -vi) &= \log \left| \frac{\sqrt{ui} + \sqrt{-vi}}{\sqrt{ui} - \sqrt{vi}} \right| + \log \left| \frac{\sqrt{ui} + \sqrt{vi}}{\sqrt{ui} - \sqrt{-vi}} \right| \\ &= \log \left| \frac{\sqrt{u} + i\sqrt{v}}{\sqrt{u} - \sqrt{v}} \right| + \log \left| \frac{\sqrt{u} + \sqrt{v}}{\sqrt{u} - i\sqrt{v}} \right| = \log \left| \frac{\sqrt{u} + \sqrt{v}}{\sqrt{u} - \sqrt{v}} \right| = g_{\Omega}(u, v). \quad \square \end{aligned}$$

LEMMA A.3. *The following differential relations hold:*

$$(A.9) \quad \left. \frac{\partial g_{\Omega}(-u + \epsilon i, v)}{\partial \epsilon} \right|_{\epsilon=+0} = \frac{\sqrt{v}}{\sqrt{u}(u+v)},$$

$$(A.10) \quad \left. \frac{\partial g_{\Omega}(-u + \epsilon i, vi)}{\partial \epsilon} \right|_{\epsilon=+0} = \frac{\sqrt{v}}{\sqrt{2u} \left[\frac{v}{2} + (\sqrt{u} - \sqrt{\frac{v}{2}})^2 \right]},$$

$$(A.11) \quad \left. \frac{\partial g_{\Omega}(-u + \epsilon i, -vi)}{\partial \epsilon} \right|_{\epsilon=+0} = \frac{\sqrt{v}}{\sqrt{2u} \left[\frac{v}{2} + (\sqrt{u} + \sqrt{\frac{v}{2}})^2 \right]},$$

$$u, v \in \mathbf{R}, \quad u, v > 0.$$

Proof. The symbol \doteq will denote an equality up to an $o(\epsilon)$ addend.

The limit values of $g_\Omega(z, x)$ are zero, when z or x tends to a point from \mathbf{R}_- .

First, we have

$$\begin{aligned} \log \left| \frac{\sqrt{-u + \epsilon i} + \sqrt{v}}{\sqrt{-u + \epsilon i} - \sqrt{v}} \right| &= \log \left| \frac{\sqrt{-1 + \frac{\epsilon}{u} i} + \sqrt{\frac{v}{u}}}{\sqrt{-1 + \frac{\epsilon}{u} i} - \sqrt{\frac{v}{u}}} \right| \doteq \log \left| \frac{i + \frac{\epsilon}{2u} + \sqrt{\frac{v}{u}}}{i + \frac{\epsilon}{2u} - \sqrt{\frac{v}{u}}} \right| \\ &\doteq \log \sqrt{\frac{1 + \frac{v}{u} + \frac{\epsilon}{u} \sqrt{\frac{v}{u}}}{1 + \frac{v}{u} - \frac{\epsilon}{u} \sqrt{\frac{v}{u}}}} \doteq \frac{1}{2} \log \left(1 + 2 \frac{\frac{\epsilon}{u} \sqrt{\frac{v}{u}}}{1 + \frac{v}{u}} \right) \doteq \frac{\frac{\epsilon}{u} \sqrt{\frac{v}{u}}}{1 + \frac{v}{u}}, \end{aligned}$$

which gives (A.9).

Second, we obtain

$$\begin{aligned} \log \left| \frac{\sqrt{-u + \epsilon i} + \sqrt{-vi}}{\sqrt{-u + \epsilon i} - \sqrt{-vi}} \right| &\doteq \log \left| \frac{\sqrt{u} i \left(1 - \frac{\epsilon i}{2u}\right) + \sqrt{v} \frac{1-i}{\sqrt{2}}}{\sqrt{u} i \left(1 - \frac{\epsilon i}{2u}\right) - \sqrt{v} \frac{1+i}{\sqrt{2}}} \right| \\ &= \log \left| \frac{\left(\frac{\epsilon}{2\sqrt{u}} + \sqrt{\frac{v}{2}}\right) + (\sqrt{u} - \sqrt{\frac{v}{2}}) i}{\left(\frac{\epsilon}{2\sqrt{u}} - \sqrt{\frac{v}{2}}\right) + (\sqrt{u} - \sqrt{\frac{v}{2}}) i} \right| \\ &= \frac{1}{2} \log \frac{\left(\frac{\epsilon}{2\sqrt{u}} - \sqrt{\frac{v}{2}}\right)^2 + 4 \frac{\epsilon}{2\sqrt{u}} \sqrt{\frac{v}{2}} + (\sqrt{u} - \sqrt{\frac{v}{2}})^2}{\left(\frac{\epsilon}{2\sqrt{u}} + \sqrt{\frac{v}{2}}\right)^2 + (\sqrt{u} - \sqrt{\frac{v}{2}})^2} \\ &\doteq \frac{1}{2} \log \left[1 + \frac{\frac{2\epsilon}{\sqrt{u}} \sqrt{\frac{v}{2}}}{\frac{v}{2} + (\sqrt{u} - \sqrt{\frac{v}{2}})^2} \right] \doteq \frac{\epsilon \sqrt{\frac{v}{2u}}}{\frac{v}{2} + (\sqrt{u} - \sqrt{\frac{v}{2}})^2}; \end{aligned}$$

this leads to (A.10).

Third, we analogously derive

$$\begin{aligned} \log \left| \frac{\sqrt{-u + \epsilon i} + \sqrt{vi}}{\sqrt{-u + \epsilon i} - \sqrt{vi}} \right| &\doteq \log \left| \frac{\sqrt{u} i \left(1 - \frac{\epsilon i}{2u}\right) + \sqrt{v} \frac{1+i}{\sqrt{2}}}{\sqrt{u} i \left(1 - \frac{\epsilon i}{2u}\right) - \sqrt{v} \frac{1-i}{\sqrt{2}}} \right| \\ &= \log \left| \frac{\left(\frac{\epsilon}{2\sqrt{u}} + \sqrt{\frac{v}{2}}\right) + (\sqrt{u} + \sqrt{\frac{v}{2}}) i}{\left(\frac{\epsilon}{2\sqrt{u}} - \sqrt{\frac{v}{2}}\right) + (\sqrt{u} + \sqrt{\frac{v}{2}}) i} \right| \\ &= \frac{1}{2} \log \frac{\left(\frac{\epsilon}{2\sqrt{u}} - \sqrt{\frac{v}{2}}\right)^2 + 4 \frac{\epsilon}{2\sqrt{u}} \sqrt{\frac{v}{2}} + (\sqrt{u} + \sqrt{\frac{v}{2}})^2}{\left(\frac{\epsilon}{2\sqrt{u}} + \sqrt{\frac{v}{2}}\right)^2 + (\sqrt{u} + \sqrt{\frac{v}{2}})^2} \\ &\doteq \frac{1}{2} \log \left[1 + \frac{\frac{2\epsilon}{\sqrt{u}} \sqrt{\frac{v}{2}}}{\frac{v}{2} + (\sqrt{u} + \sqrt{\frac{v}{2}})^2} \right] \doteq \frac{\epsilon \sqrt{\frac{v}{2u}}}{\frac{v}{2} + (\sqrt{u} + \sqrt{\frac{v}{2}})^2}; \end{aligned}$$

this justifies (A.11). \square

A.2. Two measures and their potentials. It follows from the explicit formulae [2, section 39] for the extremal error points of diagonal Zolotaryov approximants to the function $z^{-1/2}$ on the segment $[1 - \kappa^2, 1]$ that, as the approximant's degree tends to infinity, the interpolation points are, in the limit, distributed according to

the probability measure α on $[1 - \kappa^2, 1]$, defined by the equality

$$\alpha'(x) = \frac{1}{2K(\kappa)\sqrt{(x + \kappa^2 - 1)x(1 - x)}}.$$

Since Zolotaryov approximants are optimal (though with a weight), the measure α is equilibrium with respect to Ω .

Without loss of generality we can assume that $\omega_{\max} = 1$. Now introduce the following probability measure β on the compact D :

$$(A.12) \quad \beta(iX) = \beta(-iX) = \frac{\alpha(X)}{2}, \quad X \text{ a measurable subset of } [1 - \kappa^2, 1].$$

Define the two potentials

$$(A.13) \quad g(\alpha, \Omega; z) = \int_{1-\kappa^2}^1 g_\Omega(z, x) d\alpha(x), \quad g(\beta, \Omega; z) = \int_D g_\Omega(z, x) d\beta(x).$$

PROPOSITION A.4. *The measure β is the equilibrium one for the compact D with respect to the domain Ω . The (common) value of $g(\alpha, \Omega; z)$ on D equals the half (common) value of $g(\alpha, \Omega; z)$ on $[1 - \kappa^2, 1]$.*

Proof. The two potentials on the corresponding supports owing to (A.8), (A.12), and (A.13) are related by

$$\begin{aligned} g(\beta, \Omega; ui) &= \frac{1}{2} \int_{1-\kappa^2}^1 g_\Omega(ui, vi) d\alpha(v) + \frac{1}{2} \int_{1-\kappa^2}^1 g_\Omega(ui, -vi) d\alpha(v) \\ &= \frac{1}{2} \int_{1-\kappa^2}^1 g_\Omega(u, v) d\alpha(v) = \frac{1}{2} g(\alpha, \Omega; u), \\ & \quad u \in \mathbf{R}, \quad 1 - \kappa^2 \leq u \leq 1. \end{aligned}$$

It remains to recall that the potential $g(\alpha, \Omega; u)$ is constant on $[1 - \kappa^2, 1]$. □

LEMMA A.5. *The two potentials satisfy the equality*

$$(A.14) \quad \int_0^{+\infty} \frac{\partial g(\alpha, \Omega; -u)}{\partial \nu_u} du = \int_0^{+\infty} \frac{\partial g(\beta, \Omega; -u)}{\partial \nu_u} du \quad (= \pi),$$

where ν is the upward (or, which is the same due to the symmetry, downward) normal.

Proof. On the one hand, in view of (A.1) and (A.9)

$$\begin{aligned} & \int_0^{+\infty} \frac{\partial g(\alpha, \Omega; -u)}{\partial \nu_u} du = \int_0^{+\infty} \int_{1-\kappa^2}^1 \frac{\partial g_\Omega(-u, v)}{\partial \nu_u} du d\alpha(v) \\ (A.15) \quad & = \int_{1-\kappa^2}^1 \int_0^{+\infty} \frac{du}{\sqrt{u}(u+v)} \sqrt{v} d\alpha(v) = \pi \int_{1-\kappa^2}^1 d\alpha(v) = \pi. \end{aligned}$$

On the other hand, making at a suitable moment the change of variables $u = vt^2$ and

exploiting formulae (A.10), (A.11) and [13, item 2.172], derive

$$\begin{aligned}
 & \int_0^{+\infty} \frac{\partial g(\beta, \Omega; -u)}{\partial \nu_u} du = \int_0^{+\infty} \int_{1-\kappa^2}^1 \frac{1}{2} \left[\frac{\partial g_\Omega(-u, vi)}{\partial \nu_u} + \frac{\partial g_\Omega(-u, -vi)}{\partial \nu_u} \right] du d\alpha(v) \\
 &= \frac{1}{2} \int_{1-\kappa^2}^1 \int_0^{+\infty} \left[\frac{1}{\sqrt{2u}(u+v-\sqrt{2uv})} + \frac{1}{\sqrt{2u}(u+v+\sqrt{2uv})} \right] du \sqrt{v} d\alpha(v) \\
 &= \frac{1}{2} \int_{1-\kappa^2}^1 \left[\int_0^{+\infty} \frac{2vt dt}{\sqrt{2vt^2}(vt^2+v-\sqrt{2v^2t^2})} + \int_0^{+\infty} \frac{2vt dt}{\sqrt{2vt^2}(vt^2+v+\sqrt{2v^2t^2})} \right] \sqrt{v} d\alpha(v) \\
 &= \frac{1}{\sqrt{2}} \int_{1-\kappa^2}^1 \left[\int_0^{+\infty} \frac{dt}{t^2-\sqrt{2}t+1} + \int_0^{+\infty} \frac{dt}{t^2+\sqrt{2}t+1} \right] d\alpha(v) \\
 &= \int_{1-\kappa^2}^1 \left(\arctan \frac{2t-\sqrt{2}}{\sqrt{2}} \Big|_{t=0}^{t=+\infty} + \arctan \frac{2t+\sqrt{2}}{\sqrt{2}} \Big|_{t=0}^{t=+\infty} \right) d\alpha(v) \\
 &= \frac{\pi}{2} + \frac{\pi}{4} + \frac{\pi}{2} - \frac{\pi}{4} = \pi.
 \end{aligned}$$

(A.16)

Comparing (A.15) and (A.16), we get (A.14). \square

A.3. Proof of Theorem 3.4.

Proof. It follows from [11, section 1] that the Riemann modulus of the condenser $(\mathbf{R}_-, [1-\kappa^2, 1])$ equals ρ^2 . This implies (see [10, section 3]) that

$$\lim_{n \rightarrow \infty} \sqrt[n]{\sigma_n(\mathbf{R}_-, [1-\kappa^2, 1])} = \rho^2.$$

Take into account that potentials (A.13), divided by their values on the compacts $[1-\kappa^2, 1]$ and D , respectively, solve the Dirichlet problems with the zero boundary condition on \mathbf{R}_- and unity boundary condition on $[1-\kappa^2, 1]$ or D (these harmonic functions are called harmonic measures; see [29, section 4.3]). Formula (27) from [36, section 8.7, Theorem 9] and the definition of the quantity τ from that theorem’s proof show how the quantities

$$\lim_{n \rightarrow \infty} \sqrt[n]{\sigma_n(\mathbf{R}_-, D)} \quad \text{and} \quad \lim_{n \rightarrow \infty} \sqrt[n]{\sigma_n(\mathbf{R}_-, [1-\kappa^2, 1])}$$

are expressed in terms of the harmonic measures: the asymptotic convergence factors’ logarithms are inversely proportional to the integral over \mathbf{R}_- of the normal derivative of harmonic measures (it is sufficient to know the integrals over one of the two edges of the slit \mathbf{R}_-). Assertion (3.12) is a consequence of Lemma A.5 and Proposition A.4.

Assertion (3.11) then follows from [10, Theorem 1]. \square

Remark 3. The proof of the mentioned Theorem 9 from [36, section 8.7] shows that solutions that are optimal in the Cauchy–Hadamard sense can be taken with $\omega_j \in D$ and $\lambda_j \in \mathbf{R}_-$.

Acknowledgments. The authors are thankful to A. B. Bogatyryov, M. Botchev, V. I. Lebedev, S. P. Suetin, and E. E. Tyrtshnikov for bibliographical support. The authors are grateful to B. Beckermann for pointing out a stable and simple variant of the rational Arnoldi method.

REFERENCES

- [1] M. ABRAMOWITZ AND J. STEGAN, EDs., *Handbook of Mathematical Functions*, Appl. Math. 55, National Bureau of Standards, Washington, D.C., 1964.
- [2] N. I. AKHIEZER, *Theory of Approximation*, Dover, New York, 1992.
- [3] Z. BAI, *Krylov subspace techniques for reduced-order modeling of large-scale dynamical systems*, Appl. Numer. Math., 43 (2002), pp. 9–44.
- [4] G. A. BAKER AND P. GRAVES-MORRIS, *Padé Approximants*, Addison–Wesley, London, 1996.
- [5] A. BULTHEEL, P. GONZALEZ-VERA, E. HENDRIKSEN, AND O. NJASTAD, *Orthogonal Rational Functions*, Cambridge University Press, Cambridge, UK, 1999.
- [6] V. L. DRUSKIN AND L. A. KNIZHNERMAN, *A spectral semi-discrete method for the numerical solution of three-dimensional non-stationary electrical prospecting problems*, Izv. Akad. Nauk SSSR Ser. Fiz. Zemli, 8 (1988), pp. 63–74 (in Russian; translated into English).
- [7] N. S. ELLNER AND E. L. WACHSPRESS, *Alternating direction implicit iteration for systems with complex spectra*, SIAM J. Numer. Anal., 28 (1991), pp. 859–870.
- [8] R. W. FREUND, *Model reduction methods based on Krylov subspaces*, Acta Numer., 12 (2003), pp. 267–319.
- [9] E. GALLOPOULOS AND Y. SAAD, *Efficient solution of parabolic equations by Krylov approximation method*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 1236–1264.
- [10] A. A. GONCHAR, *Zolotarev problems connected with rational functions*, Mat. Sb., 7 (1969), pp. 623–635 (in Russian; translated into English).
- [11] A. A. GONCHAR, *On the speed of rational approximation of some analytic functions*, Mat. Sb., 34 (1978), pp. 131–145 (in Russian; translated into English).
- [12] S. A. GOREINOV, *Mosaic-skeleton approximations of matrices generated by asymptotically smooth and oscillative kernels*, in Matrix Methods and Computations, Inst. Numer. Math. RAS, Moscow, 1999, pp. 42–76 (in Russian).
- [13] I. S. GRADSHTEYN AND I. M. RYZHIK, *Tables of Integrals, Series, and Products*, Academic Press, New York, 2000.
- [14] E. J. GRIMME, *Krylov Projection Methods for Model Reduction*, Ph.D. thesis, The University of Illinois at Urbana-Champaign, 1997.
- [15] S. GUGERCIN, A. ANTOULAS, AND C. BEATTIE, *A rational Krylov iteration for optimal H_2 model reduction*, in Proceedings of the 17th International Symposium on Mathematical Theory of Networks and Systems, Kyoto, Japan, 2006, pp. 1665–1667.
- [16] W. HACKBUSCH, B. N. KHOROMSKII, AND E. E. TYRTYSHNIKOV, *Hierarchical Kronecker tensor-product approximations*, J. Numer. Math., 13 (2005), pp. 119–156.
- [17] M. R. HESTENES AND E. STIEFEL, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Stand., 49 (1952), pp. 409–436.
- [18] M. HOCHBRUCK AND C. LUBICH, *On Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 34 (1997), pp. 1911–1925.
- [19] D. INGERMAN, V. DRUSKIN, AND L. KNIZHNERMAN, *Optimal finite difference grids and rational approximations of the square root. I. Elliptic functions*, Commun. Pure Appl. Math., 53 (2000), pp. 1039–1066.
- [20] M.-P. ISTANCE AND J.-P. THIRAN, *On the third and fourth Zolotarev problems in the complex plane*, SIAM J. Numer. Anal., 32 (1995), pp. 249–259.
- [21] P. K. KYTHE, *Computational Conformal Mapping*, Birkhäuser, Boston, 1998.
- [22] V. I. LEBEDEV, *On Zolotarev problems in the alternating direction method. II*, in Trudy Semin. S. L. Sobolev 1, Novosibirsk, Nauka, 1976, pp. 51–59 (in Russian).
- [23] L. MEIER AND D. LUENBERGER, *Approximation of linear constant systems*, IEEE Trans. Automat. Control, 12 (1967), pp. 585–588.
- [24] Z. NEHARI, *Conformal Mapping*, Dover, New York, 1975.
- [25] E. M. NIKISHIN AND V. N. SOROKIN, *Rational Approximations and Orthogonality*, Nauka, Moscow, 1988 (in Russian); English translation in Transl. Math. Monogr., AMS, Providence, RI, 1991.
- [26] B. NOUR-OMID, *Lanczos method for heat conduction analysis*, Internat. J. Numer. Methods Engrg., 24 (1987), pp. 251–262.
- [27] B. NOUR-OMID AND R. W. CLOUGH, *Dynamic analysis of structure using Lanczos co-ordinates*, Earthquake Eng. and Struct. Dynamics, 12 (1984), pp. 565–577.
- [28] I. V. OSELEDETS, *Lower bounds for separable approximations of the Hilbert kernel*, Mat. Sb., 198 (2007), pp. 425–432 (in Russian; translated into English).
- [29] T. RANSFORD, *Potential Theory in the Complex Plane*, London Math. Soc. Stud. Texts 28, Cambridge University Press, Cambridge, UK, 1995.
- [30] A. RUHE, *The rational Krylov algorithm for nonsymmetric eigenvalue problems. III: Complex shifts for real matrices*, BIT, 34 (1994), pp. 165–176.

- [31] H. STAHL AND V. TOTIK, *General Orthogonal Polynomials*, Encyclopedia Math. Appl. 43, Cambridge University Press, Cambridge, UK, 1992.
- [32] G. STARKE, *Optimal alternating direction implicit parameters for nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., 28 (1991), pp. 1431–1445.
- [33] M. E. TAYLOR, *Partial Differential Equations II: Qualitative Studies of Linear Equations*, Springer, New York, 1991.
- [34] E. E. TYRITYSHNIKOV, *Mosaic-skeleton approximations*, Calcolo, 33 (1996), pp. 47–57.
- [35] H. A. VAN DER VORST, *An iterative solution method for solving $f(A)x = b$ using Krylov subspace information obtained for the symmetric positive definite matrix*, J. Comput. Appl. Math., 18 (1987), pp. 249–263.
- [36] J. L. WALSH, *Interpolation and Approximation by Rational Functions in the Complex Domain*, AMS, Providence, RI, 1960.
- [37] M. ZASLAVSKY, S. DAVYDYCHEVA, V. DRUSKIN, A. ABUBAKAR, T. HABASHY, AND L. KNIZHNERMAN, *Finite-difference solution of the 3D electromagnetic problem using divergence-free preconditioners*, in Proceedings of SEG Annual Meeting, New Orleans, 2006, pp. 775–778.

HARDY SPACE INFINITE ELEMENTS FOR SCATTERING AND RESONANCE PROBLEMS*

THORSTEN HOHAGE[†] AND LOTHAR NANNEN[†]

Abstract. This paper introduces a new type of infinite element for scattering and resonance problems that is derived from a variant of the pole condition as radiation condition. This condition states that a certain transform of the exterior solution belongs to the Hardy space of L^2 boundary values of holomorphic functions on the unit disc if and only if the solution is outgoing. We obtain a symmetric variational formulation of the problem in this Hardy space. Our infinite elements correspond to a Galerkin discretization with respect to the standard monomial orthogonal basis of this Hardy space and lead to simple element matrices. Hardy space infinite elements are particularly well suited for solving resonance problems since they preserve the eigenvalue structure of the problem. We prove superalgebraic convergence for a separated problem. Numerical experiments exhibit fast convergence over a wide range of wave numbers.

Key words. transparent boundary conditions, radiation conditions, pole condition, infinite elements, Hardy spaces, Helmholtz equation

AMS subject classifications. 65N30, 65N12, 35B34, 35J20, 44A10

DOI. 10.1137/070708044

1. Introduction. For solving a time-harmonic wave equation on an unbounded domain by finite element methods, appropriate boundary conditions have to be imposed on the artificial boundary of the necessarily finite computational domain. These boundary conditions should be chosen in such a way that the solution of the boundary value problem on the computational domain is a good approximation to the restriction of the solution of the wave equation posed on the unbounded domain. Such conditions are called *transparent boundary conditions* and replace the radiation condition at infinity.

The method proposed in this paper works well for scattering problems, but a particular advantage over numerous competing transparent boundary conditions is the ability to easily treat resonance problems. Such problems appear in molecular physics, acoustics, lasers, and numerous other areas of engineering, natural sciences, and mathematics (cf. [22, 14, 13, 7, 25]). A typical resonance problem for the Neumann–Laplacian in the complement of a smooth, compact domain $K \subset \mathbb{R}^d$ such that $\mathbb{R}^d \setminus K$ is connected consists in finding a nontrivial eigenpair $(u, \lambda) \in H_{\text{loc}}^2(\mathbb{R}^d \setminus K) \times \mathbb{C}$ such that

$$(1.1a) \quad -\Delta u = \lambda u \quad \text{in } \mathbb{R}^d \setminus K,$$

$$(1.1b) \quad \frac{\partial u}{\partial \nu} = 0 \quad \text{on } \partial K,$$

$$(1.1c) \quad u \text{ satisfies a radiation condition.}$$

$\frac{\partial u}{\partial \nu}$ denotes the outward normal derivative. For other equivalent definitions of resonances we refer to [23, 25]. In the scattering problem corresponding to (1.1), the

*Received by the editors November 20, 2007; accepted for publication (in revised form) August 8, 2008; published electronically February 13, 2009. This work was supported by the Deutsche Forschungsgemeinschaft (DFG), grant Ho 2551/2-1.

<http://www.siam.org/journals/sinum/47-2/70804.html>

[†]Institute of Numerical and Applied Mathematics, University of Göttingen, D-37083 Göttingen, Germany (hohage@math.uni-goettingen.de, nannen@math.uni-goettingen.de).

number $\lambda \in (0, \infty)$ is given and the homogeneous boundary condition (1.1b) is replaced by an inhomogeneous boundary condition. In the following let $\lambda = \kappa^2$ with $\Re(\kappa) > 0$ and assume that K is contained in the ball $B_a := \{x : \|x\| < a\}$ of radius $a > 0$. One of a several equivalent formulations of the radiation condition (1.1c) is that u has an expansion in terms of Hankel functions $H_n^{(1)}$ of the first kind,

$$(1.2) \quad u(x) = \sum_{l=0}^{\infty} \sum_{m=0}^{M_l} \alpha_{l,m} (\kappa|x|)^{1-d/2} H_{l-1+d/2}^{(1)}(\kappa|x|) Y_{l,m} \left(\frac{x}{|x|} \right), \quad |x| > a,$$

where $\{Y_{l,0}, \dots, Y_{l,M_l}\}$ is an orthonormal basis of the l -th eigenspace of the Laplace–Beltrami operator on S^{d-1} . ($Y_{l,m}$ are spherical harmonics $d = 3$ and trigonometric monomials for $d = 2$.) A solution u to (1.1a) satisfying (1.2) is called *outgoing*, whereas a solution with a corresponding expansion in terms of Hankel functions of the second kind is called *incoming*. It can be shown that all resonances $\kappa = \sqrt{\lambda}$, $\Re\kappa > 0$ of (1.1) satisfy $\Im(\kappa) < 0$ (cf. [23]). For such values of κ , it follows from the asymptotic behavior of Hankel functions,

$$(1.3) \quad |H_l^{(1)}(z)| = \frac{|e^{iz}|}{\sqrt{|z|}} \left(1 + \mathcal{O}\left(\frac{1}{|z|}\right) \right), \quad |H_l^{(2)}(z)| = \frac{|e^{-iz}|}{\sqrt{|z|}} \left(1 + \mathcal{O}\left(\frac{1}{|z|}\right) \right), \quad |z| \rightarrow \infty,$$

that outgoing solutions are exponentially increasing at infinity, and incoming solutions are exponentially decreasing. This implies in particular that incoming, but not outgoing, solutions satisfy the Sommerfeld radiation condition

$$(1.4) \quad r^{(d-1)/2} \left(\frac{\partial u}{\partial r} - i\kappa u \right) \rightarrow 0 \quad \text{as } r = |x| \rightarrow \infty$$

for $\Im(\kappa) < 0$ since condition (1.4) (as well as the conjugate condition with $-i$ replaced by i) selects exponentially decaying solutions. Hence the Sommerfeld condition does not characterize outgoing waves for $\Im(\kappa) < 0$.

The fact that (1.4) is not valid for $\Im(\kappa) < 0$ rules out the simple transparent boundary condition $\partial u / \partial r = i\kappa u$ on ∂B_a for resonance problems as well as higher order local conditions [11, 6]. Standard infinite elements are based on the series expansion (1.2) or the Wilcox expansion [3, 4]. Since κ appears in (1.2) in a very nonlinear way inside the argument of the Hankel functions, standard infinite elements destroy the eigenvalue structure of problem (1.1). The same holds true for boundary element methods. On the other hand, the perfectly matched layer (PML) method preserves the eigenvalue structure, and has been used under the name *complex scaling* for the theoretical study and the numerical computation of resonances in molecular physics since the 1970s [14, 22]. Despite the name, Hardy space infinite elements are actually closer to PML than to classical infinite elements (cf. [10]).

In this paper we will use the *pole condition* as radiation condition (cf. [18, 9, 10]). The formulation used in this paper states that a function u is outgoing if and only if a certain transform of u in a radial direction belongs to the Hardy space $H^+(S^1)$ on the complex unit circle S^1 . Analogously u is incoming if and only if the same transform of u belongs to the orthogonal complement of $H^+(S^1)$ in $L^2(S^1)$. Therefore, we apply the above transform to the variational formulation of the exterior Helmholtz equation and incorporate the radiation condition by restricting $L^2(S^1)$ to the correct Hardy space. Hardy space infinite elements correspond to the Galerkin method applied to

this variational problem using the standard monomial orthogonal basis of the Hardy space $H^+(S^1)$. For one-dimensional time-dependent problems a similar approach has been studied in [16].

The rest of this paper is organized as follows: We first present a complete treatment of Hardy space infinite elements for one-dimensional problems in section 2. In the following section 3 we derive analogous Hardy space infinite elements in arbitrary space dimensions. Then the convergence of this method is analyzed using separation arguments in section 4. Numerical results are described in section 5 before we end this paper by some conclusions, including a discussion of pros and cons of the proposed method.

2. One-dimensional Helmholtz equation. In this section we will consider the one-dimensional time-harmonic wave equation

$$(2.1a) \quad -u''(r) - \kappa^2 p(r)u(r) = 0, \quad r \geq 0,$$

$$(2.1b) \quad u'(0) = f'_0,$$

$$(2.1c) \quad u \quad \text{outgoing},$$

with a given complex wave number $\kappa \in \mathbb{C}$ with positive real part, a boundary value $f'_0 \in \mathbb{C}$, and a positive potential $p \in L^\infty((0, \infty))$ satisfying $p(r) = 1$ for $r \geq a$. We will split u into an interior part $u_{\text{int}} := u|_{[0,a]}$ and an exterior part $u_{\text{ext}}(r) := u(r+a)$, $r > 0$. Actually, in one space dimension the Sommerfeld-type transparent boundary condition $u'(a) = i\kappa u(a)$ is exact even for $\Im(\kappa) < 0$, and (2.1) reduces to the simple boundary value problem

$$(2.2) \quad -u''_{\text{int}} - p\kappa^2 u_{\text{int}} = 0, \quad u'_{\text{int}}(0) = f'_0, u'_{\text{int}}(a) = i\kappa u_{\text{int}}(a).$$

To explain the basic ideas, we will apply Hardy space infinite elements to problem (2.1) even though this is more complicated than solving (2.2) and requires more degrees of freedom. Note, however, that for the corresponding resonance problem, (2.2) leads to a quadratic eigenvalue problem, whereas Hardy space infinite elements will lead to a linear eigenvalue problem.

2.1. Pole condition and Hardy spaces. Since we assumed $p \equiv 1$ on $[a, \infty)$, the exterior part of all solutions to (2.1a) is of the form

$$(2.3) \quad u_{\text{ext}}(r) = C_1 e^{i\kappa r} + C_2 e^{-i\kappa r}, \quad r \geq 0.$$

The term $C_1 e^{i\kappa r}$ corresponds to an outgoing wave, and $C_2 e^{-i\kappa r}$ to an incoming wave. The pole condition distinguishes these two solutions with the help of the Laplace transform $(\mathcal{L}f)(s) := \int_0^\infty e^{-sr} f(r) dr$. Due to the explicit form (2.3), $\hat{u} := \mathcal{L}u_{\text{ext}}$ is given by

$$(2.4) \quad \hat{u}(s) = \frac{C_1}{s - i\kappa} + \frac{C_2}{s + i\kappa}, \quad \Re(s) > |\Im(\kappa)|.$$

This function has a holomorphic extension to $\mathbb{C} \setminus \{i\kappa, -i\kappa\}$. u is outgoing if and only if \hat{u} has no pole in the lower complex half-plane and incoming if and only if \hat{u} has no pole in the upper complex half-plane. This motivates the use of the following Hardy spaces.

DEFINITION 2.1 ($H^-(\mathbb{R})$ and $H^+(\mathbb{R})$). *The Hardy space $H^\pm(\mathbb{R})$ is the set of all functions $f \in L^2(\mathbb{R})$ that are L^2 boundary values of a function v , which is holomorphic*

in $\mathbb{C}^\pm := \{s \in \mathbb{C} : \Im(\pm s) > 0\}$ and for which the integrals $\int_{\mathbb{R}} |v(x \pm i\epsilon)|^2 dx$ are uniformly bounded for $\epsilon > 0$.

u is outgoing if and only if $\widehat{u}|_{\mathbb{R}} \in H^-(\mathbb{R})$ and incoming if and only if $\widehat{u}|_{\mathbb{R}} \in H^+(\mathbb{R})$.

Equipped with the standard L^2 inner product, $H^\pm(\mathbb{R})$ are Hilbert spaces (cf. [5]). Moreover, by the Paley–Wiener theorem these spaces are characterized by

$$(2.5) \quad H^\pm(\mathbb{R}) = \{\widehat{u} \in L^2(\mathbb{R}) : \mathcal{F}^{-1}\widehat{u}(\pm t) = 0 \text{ for almost all } t > 0\}$$

in terms of the inverse Fourier transform $(\mathcal{F}^{-1}f)(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ist} f(s) ds$. This yields the orthogonal decomposition $L^2(\mathbb{R}) = H^+(\mathbb{R}) \oplus H^-(\mathbb{R})$. The function v in Definition 2.1 is uniquely determined by f and can be recovered by the Cauchy integral

$$(2.6) \quad v(s) = \frac{1}{2\pi i} \int_{\mathbb{R}} \frac{f(\bar{s})}{\bar{s} - s} d\bar{s}, \quad s \in \mathbb{C}^\pm.$$

Since we are interested in outgoing solutions, we will mainly deal with the space $H^-(\mathbb{R})$. Because of the lack of a convenient orthonormal basis of $H^-(\mathbb{R})$ we will apply a further transform to another closely related Hardy space.

DEFINITION 2.2 ($H^+(S^1)$). *The Hardy space $H^+(S^1)$ is the set of all functions $F \in L^2(S^1)$ that are L^2 boundary values of a function V , which is holomorphic in the unit disk $D := \{z \in \mathbb{C} : |z| < 1\}$ and for which the integrals $\int_0^{2\pi} |V(re^{i\theta})|^2 d\theta$ are uniformly bounded for $r \in [0, 1)$.*

Equipped with the L^2 scalar product, $H^+(S^1)$ is a Hilbert space, and a simple complete orthogonal system of $H^+(S^1)$ is given by the monomials z^k , $k = 0, 1, \dots$

A family of unitary operators identifying the Hilbert spaces $H^-(\mathbb{R})$ and $H^+(S^1)$ can be defined with the help of the Möbius transformations $\varphi_{\kappa_0}(z) := i\kappa_0 \frac{z+1}{z-1}$, $\kappa_0 > 0$, which map the unit disc D to the half-space \mathbb{C}^- . The parameter κ_0 will act as a tuning parameter in the algorithms to be discussed below. Since $\int_{-\infty}^{\infty} |f(t)|^2 dt = \int_0^{2\pi} |(f \circ \varphi_{\kappa_0})(e^{i\theta}) \sqrt{\varphi'_{\kappa_0}(e^{i\theta})}|^2 d\theta$ and $\varphi'_{\kappa_0}(z) = \frac{-2i\kappa_0}{(z-1)^2}$, the mappings

$$(2.7) \quad (\mathcal{M}_{\kappa_0} f)(z) := (f \circ \varphi_{\kappa_0})(z) \frac{1}{z-1}$$

are isometric from $L^2(\mathbb{R})$ to $L^2(S^1)$ up to the factor $\sqrt{-2i\kappa_0}$, and it can be shown that $\mathcal{M}_{\kappa_0}(H^-(\mathbb{R})) = H^+(S^1)$ (see [5]). Hence, $\sqrt{-2i\kappa_0} \mathcal{M}_{\kappa_0} : H^-(\mathbb{R}) \rightarrow H^+(S^1)$ is unitary.

Many of the operators on $H^+(S^1)$ which will appear in our analysis are of the following form.

DEFINITION 2.3 (Toeplitz operator). *Let $f \in L^\infty(S^1)$ be a complex-valued function and let $P : L^2(S^1) \rightarrow H^+(S^1)$ denote the orthogonal projection. Then the Toeplitz operator $T_f : H^+(S^1) \rightarrow H^+(S^1)$ with symbol f is defined by $T_f U := P(fU)$.*

We will need the following classical results on Toeplitz operators: If $f : S^1 \rightarrow \mathbb{C}$ is continuous and has no zeros, then T_f is a Fredholm operator, and $\text{ind}(T_f) = -\text{wn}(f)$ where $\text{wn}(f)$ denotes the winding number of f around 0 [1, Theorem 2.42]. Moreover, if $\text{ind}(T_f) = 0$, then T_f is injective and hence boundedly invertible [1, Corollary 2.40].

Let us consider the explicit form of the transform $\widehat{U} := \mathcal{M}_{\kappa_0} \widehat{u}$ of the outgoing solution u of (2.1). With $u_0 := u(a)$ we have

$$(2.8) \quad u_{\text{ext}}(r) = u_0 e^{i\kappa r} \xrightarrow{\mathcal{L}|_{\mathbb{R}}} \widehat{u}(s) = \frac{u_0}{s - i\kappa} \xrightarrow{\mathcal{M}_{\kappa_0}} \widehat{U}(z) = \frac{u_0}{i\kappa_0(z+1) - i\kappa(z-1)}.$$

Note that $\widehat{U}(1) = u_0/(2i\kappa_0)$. This will be convenient for coupling the transformed exterior to the interior problem. To take advantage of this fact we decompose

$$(2.9) \quad \widehat{U}(z) = \frac{1}{2i\kappa_0}(u_0 + (z - 1)U(z)) \quad \text{with} \quad U(z) := \frac{2i\kappa_0\widehat{U}(z) - u_0}{z - 1}.$$

Since the only singularities of the holomorphic extensions of \widehat{U} and U are simple poles at $\frac{\kappa_0 + \kappa}{\kappa_0 - \kappa}$ and since $\frac{\kappa_0 + \kappa}{\kappa_0 - \kappa} \notin \overline{D}$ for $\Re(\kappa/\kappa_0) > 0$, both \widehat{U} and U are analytic on S^1 and belong to $H^+(S^1)$.

2.2. Variational formulation. The formal variational formulation of the differential equation (2.1a) is

$$(2.10) \quad \int_0^a (u'_{\text{int}}v'_{\text{int}} - \kappa^2 p u_{\text{int}}v_{\text{int}})dr + \int_0^\infty (u'_{\text{ext}}v'_{\text{ext}} - \kappa^2 u_{\text{ext}}v_{\text{ext}})dr = -f'_0 v_{\text{int}}(0).$$

The basic identities for transforming the exterior variational problem to the Hardy space are

$$(2.11) \quad \int_0^\infty f(r)g(r)dr = -\frac{i}{2\pi} \int_{-\infty}^\infty \widehat{f}(-s)\widehat{g}(s)ds = \frac{-i\kappa_0}{\pi} \int_{S^1} \widehat{F}(\overline{z})\widehat{G}(z)|dz|,$$

with $\widehat{f} = (\mathcal{L}f)|_{\mathbb{R}}$, $\widehat{g} = (\mathcal{L}g)|_{\mathbb{R}}$, $\widehat{F} = \mathcal{M}_{\kappa_0}\widehat{f}$, and $\widehat{G} = \mathcal{M}_{\kappa_0}\widehat{g}$. They will be derived in Lemma A.1 for the more general case $\kappa_0 \in \mathbb{C}$ (cf. Remark 2.8 below). Introducing the bilinear form

$$(2.12) \quad A(F, G) := \int_{S^1} G(\overline{z})F(z)|dz|, \quad F, G \in H^+(S^1),$$

we have in particular that $\int_0^\infty fgdr = \frac{-i\kappa_0}{\pi} A(\widehat{F}, \widehat{G})$.

THEOREM 2.4. *Let $\kappa_0, \Re(\kappa) > 0$ and $X := H^1([0, a]) \oplus H^+(S^1)$. If $u \in H^2_{\text{loc}}([0, \infty))$ is a solution to (2.1), then $(u_{\text{int}}, U)^\top$ with U defined in (2.9) belongs to X and satisfies the variational equation*

$$(2.13) \quad B\left(\begin{pmatrix} u_{\text{int}} \\ U \end{pmatrix}, \begin{pmatrix} v_{\text{int}} \\ V \end{pmatrix}\right) = -f'_0 v_{\text{int}}(0),$$

with

$$B\left(\begin{pmatrix} u_{\text{int}} \\ U \end{pmatrix}, \begin{pmatrix} v_{\text{int}} \\ V \end{pmatrix}\right) := \int_0^a (u'_{\text{int}}v'_{\text{int}} - \kappa^2 p u_{\text{int}}v_{\text{int}})dr - \frac{i\kappa_0}{4\pi} A(u_0 + (z + 1)U, v_0 + (z + 1)V) - \frac{i\kappa^2}{4\pi\kappa_0} A(u_0 + (z - 1)U, v_0 + (z - 1)V)$$

for all $(v_{\text{int}}, V) \in X$ and $v_0 := v_{\text{int}}(a)$. Conversely, if $(u_{\text{int}}, U)^\top \in X$ is a solution of (2.13), then u_{int} belongs to $H^2([0, a])$ and is the restriction of a solution u to (2.1).

Proof. Assume first that u is a solution to (2.1). It suffices to show that (2.13) holds for all (v_{int}, V) in a dense subset of X . Hence, we start with a test function $v \in C([0, \infty)) \cap H^1([0, a])$ for which v_{ext} has the form

$$v_{\text{ext}}(r) = v_0 e^{ikr}, \quad \Im(k) > -\Im(\kappa), \quad \Re(k) > 0.$$

For such test functions, the product $u \cdot v$ and products of derivatives decay exponentially, and (2.10) can be derived by partial integration. Moreover, for these test functions the identity (2.11) holds both for $f = u_{\text{ext}}, g = v_{\text{ext}}$ and for $f = u'_{\text{ext}}, g = v'_{\text{ext}}$. In the second case we apply the identities

$$(2.14) \quad (\mathcal{L}f')(s) = s(\mathcal{L}f)(s) - f_0,$$

$$(\mathcal{M}_{\kappa_0}\mathcal{L}|_{\mathbb{R}}f')(z) = i\kappa_0 \frac{z+1}{z-1} \frac{f_0 + (z-1)F(z)}{2i\kappa_0} - \frac{f_0}{z-1} = \frac{1}{2} (f_0 + (z+1)F(z)),$$

where f_0 and F are defined in analogy to u_0 and U , to finally arrive at (2.13) with

$$(2.15) \quad V(z) = \frac{2i\kappa_0(\mathcal{M}_{\kappa_0}\mathcal{L}|_{\mathbb{R}}v_{\text{ext}})(z) - v_0}{z-1} = v_0 \frac{k - \kappa_0}{(\kappa_0 - k)z + (\kappa_0 + k)}.$$

Since by virtue of Lemma A.2 the span of such functions is dense in $H^+(S^1)$ and B is continuous on $X \times X$, (2.13) holds for all $(v_{\text{int}}, V)^\top \in X$.

Conversely, let $(u_{\text{int}}, U)^\top \in X$ be a solution to (2.13). For $v_{\text{int}} = 0$ it follows after multiplication by $-4\pi i\kappa_0$ that

$$(2.16) \quad \int_{S^1} V(\bar{z}) \left\{ -\kappa_0^2 \overline{(z+1)} [u_0 + (z+1)U(z)] - \kappa^2 \overline{(z-1)} [u_0 + (z-1)U(z)] \right\} |dz| = 0$$

for all $V \in H^+(S^1)$. Due to (2.20) below, the orthogonal projection $P : L^2(S^1) \rightarrow H^+(S^1)$ applied to the expression in braces vanishes. Since $P\bar{z} = 0$, we obtain

$$(2.17) \quad P\{mU\} = P\{(\kappa_0^2 - \kappa^2) + (\kappa_0^2 + \kappa^2)\bar{z}\} u_0 = (\kappa_0^2 - \kappa^2)u_0,$$

with $m(z) := -\kappa_0^2|z+1|^2 - \kappa^2|z-1|^2$.

The left-hand side of (2.17) is the Toeplitz operator T_m with symbol m applied to U . Since $m(z) = -2(\kappa^2 + \kappa_0^2) + 2(\kappa^2 - \kappa_0^2)\Re(z)$, the graph of m is the straight line connecting $-4\kappa^2$ and $-4\kappa_0^2$. Therefore, T_m is boundedly invertible by the results quoted after Definition 2.3. Hence, (2.17) has a unique solution. By the derivation of (2.13), this solution is given by (2.8) and (2.9), or explicitly $U(z) = u_0 \frac{\kappa - \kappa_0}{(\kappa_0 - \kappa)z + (\kappa_0 + \kappa)}$. Plugging this into (2.13) and using (2.16), we obtain the variational formulation of the boundary value problem (2.2):

$$(2.18) \quad \int_0^a (v'_{\text{int}} u'_{\text{int}} - \kappa^2 p v_{\text{int}} u_{\text{int}}) dr = i\kappa u_0 v_0 - v_{\text{int}}(0) f'_0.$$

By elliptic regularity results u_{int} belongs to $H^2([0, a])$ and solves (2.2). Hence, it is also part of a solution to (2.1). \square

2.3. Gårding-type inequality. It is obvious that the bilinear form B in Theorem 2.4 is bounded and symmetric. Moreover, the interior part $B_{\text{int}}(u_{\text{int}}, v_{\text{int}}) := \int_0^a (u'_{\text{int}} v'_{\text{int}} - \kappa^2 p u_{\text{int}} v_{\text{int}}) dr$ satisfies the standard Gårding inequality

$$(2.19) \quad \Re \{ B_{\text{int}}(u_{\text{int}}, \overline{u_{\text{int}}}) \} + \beta \|u_{\text{int}}\|_{L^2}^2 \geq \|u_{\text{int}}\|_{H^1}^2,$$

with $\beta := (|\kappa|^2 + 1)\|p\|_{L^\infty} \geq 0$. We want to derive a similar inequality for the whole bilinear form B . Note that we cannot simply choose $V = \overline{U}$ since $\overline{U} \notin H^+(S^1)$ for

$U \in H^+(S^1)$ in general. However, a useful conjugation on the Hilbert space $H^+(S^1)$ is given by the mapping $\mathcal{C} : H^+(S^1) \rightarrow H^+(S^1)$ defined by

$$(\mathcal{C}F)(z) := \overline{F(\bar{z})}.$$

It is easy to check that \mathcal{C} is well-defined, antilinear, and isometric, $\mathcal{C}^2 = I$; i.e., \mathcal{C} is indeed a conjugation. Moreover, it has the useful property that

$$(2.20) \quad A(F, \mathcal{C}G) = \langle F, G \rangle_{L^2(S^1)}.$$

THEOREM 2.5. *Let $\Re(\kappa^2), \kappa_0 > 0$. Then there exist constants $\alpha, \beta, \gamma > 0$, such that*

$$\Re \left\{ (i + \gamma)B \left(\begin{pmatrix} u_{\text{int}} \\ U \end{pmatrix}, \begin{pmatrix} \bar{u}_{\text{int}} \\ \mathcal{C}U \end{pmatrix} \right) \right\} + \beta \|u_{\text{int}}\|_{L^2}^2 \geq \alpha \left\| \begin{pmatrix} u_{\text{int}} \\ U \end{pmatrix} \right\|_X^2.$$

Proof. For the exterior part of the bilinear form $B_{\text{ext}} := B - B_{\text{int}}$ we obtain from the identity (2.20) that

$$\begin{aligned} \Re \left\{ (i + \gamma)B_{\text{ext}} \left(\begin{pmatrix} u_{\text{int}} \\ U \end{pmatrix}, \begin{pmatrix} \bar{u}_{\text{int}} \\ \mathcal{C}U \end{pmatrix} \right) \right\} &= \Re \left(\frac{\kappa_0(1 - \gamma i)}{4\pi} \right) \|u_0 + (z + 1)U\|_{L^2(S^1)}^2 \\ &\quad + \Re \left(\frac{\kappa^2(1 - \gamma i)}{4\pi\kappa_0} \right) \|u_0 + (z - 1)U\|_{L^2(S^1)}^2 \end{aligned}$$

for any $\gamma \in \mathbb{R}$. Due to the assumption $\Re(\kappa^2) > 0$, we may choose a $\gamma > 0$ such that $\Re(\kappa^2(1 - \gamma i)) > 0$. Using the inequality $\|x\|^2 + \|y\|^2 \geq \frac{1}{2}\|x - y\|^2$ with $x := u_0 + (z + 1)U$ and $y := u_0 + (z - 1)U$ we obtain

$$(2.21) \quad \Re \left\{ (i + \gamma)B_{\text{ext}} \left(\begin{pmatrix} u_{\text{int}} \\ U \end{pmatrix}, \begin{pmatrix} \bar{u}_{\text{int}} \\ \mathcal{C}U \end{pmatrix} \right) \right\} \geq \tilde{\alpha} \|U\|_{L^2}^2,$$

with $\tilde{\alpha} := \min \left(\Re \left(\frac{\kappa_0(1 - \gamma i)}{2\pi} \right), \Re \left(\frac{\kappa^2(1 - \gamma i)}{2\pi\kappa_0} \right) \right)$. This together with (2.19) yields the assertion with $\beta := \gamma(|\kappa|^2 + 1)\|p\|_{L^\infty} > 0$ and $\alpha := \min(\tilde{\alpha}, \gamma)$. \square

Using standard arguments, we obtain the following corollary.

COROLLARY 2.6. *If the variational equation (2.13) has only the trivial solution for $f'_0 = 0$, then it has a unique solution for all $f'_0 \in \mathbb{R}$, and the solution depends continuously on f'_0 .*

By virtue of Theorem 2.4, the variational equation (2.13) is uniquely solvable if and only if κ is not a resonance.

2.4. Galerkin approximation. In the following we will consider the Galerkin approximations to (2.13) using a finite element subspace V_h of $H^1([0, a])$ and the subspace $\Pi_N := \text{span}\{1, z, \dots, z^N\}$ of $H^+(S^1)$. This leads to the discrete variational problems

$$(2.22) \quad B \left(\begin{pmatrix} u_h \\ U_N \end{pmatrix}, \begin{pmatrix} v_h \\ V_N \end{pmatrix} \right) = -f'_0 v_h(0), \quad \begin{pmatrix} v_h \\ V_N \end{pmatrix} \in X_{h,N} := V_h \oplus \Pi_N.$$

Using Theorem 2.5 and the compactness of the embedding $H^1([0, a]) \hookrightarrow L^2([0, a])$, we obtain the following convergence result (cf. [12, Theorem 13.7]).

THEOREM 2.7. *Let $\Re(\kappa^2), \kappa_0 > 0$, and assume that κ is not a resonance. Let $(u_{\text{int}}, U)^\top \in X$ denote the unique solution to (2.13). Then there exist constants $C, N_0, h_0 > 0$ such that the variational problems (2.22) have a unique solution $(u_h, U_N)^\top \in X_{h,N}$ for $N \geq N_0$ and $h \leq h_0$, and*

$$\|u - u_h\|_{H^1}^2 + \|U - U_N\|_{L^2(S^1)}^2 \leq C \inf_{(v_h, V_N)^\top \in X_{h,N}} \left(\|u - v_h\|_{H^1}^2 + \|U - V_N\|_{L^2(S^1)}^2 \right).$$

Since U is analytic, we have exponential convergence in N , i.e., for some constants $c, \tilde{C} > 0$

$$\inf_{V_N \in \Pi_N} \|U - V_N\|_{L^2(S^1)} \leq \tilde{C} e^{-cN}.$$

Although the derivation of the exterior part of (2.13) is nonstandard, its implementation is rather simple: For $F(z) = \sum_{j=0}^\infty \alpha_j z^j$ and $G(z) = \sum_{j=0}^\infty \beta_j z^j$, we have $A(F, G) = 2\pi \sum_{j=0}^\infty \alpha_j \beta_j$. With respect to the monomial basis of Π_N the operators

$$(2.23) \quad \mathcal{T}_\pm : \mathbb{C} \oplus H^+(S^1) \rightarrow H^+(S^1), \quad \begin{pmatrix} f_0 \\ F \end{pmatrix} \mapsto \frac{1}{2} (f_0 + (\bullet \pm 1)F)$$

occurring in (2.13) are approximately represented by the bidiagonal matrices

$$(2.24) \quad \mathcal{T}_{N,\pm} := \frac{1}{2} \begin{pmatrix} 1 & \pm 1 & & & \\ & 1 & \pm 1 & & \\ & & \ddots & \ddots & \\ & & & & 1 & \pm 1 \end{pmatrix} \in \mathbb{R}^{(N+1) \times (N+2)}.$$

The Galerkin approximation (2.22) corresponds to the introduction of an “infinite element” with $N + 2$ degrees of freedom, which couples to the interior domain via the unknown u_0 . The local element matrix of this infinite element is given by $-2i\kappa_0 \{ \mathcal{T}_{N,+}^\top \mathcal{T}_{N,+} + (\kappa/\kappa_0)^2 \mathcal{T}_{N,-}^\top \mathcal{T}_{N,-} \}$.

In the space domain the monomial basis functions correspond to the functions $u_j := (\mathcal{L}|_{\mathbb{R}})^{-1} \{ \mathcal{M}_{\kappa_0}^{-1} \mathcal{T}_-(u_0, z^j) \}$, which are given by

$$(2.25) \quad u_j(r) = e^{i\kappa_0 r} \left\{ u_0 + \sum_{n=0}^j \binom{j}{n} \frac{(2i\kappa_0 r)^{n+1}}{(n+1)!} \right\}.$$

From this formula it is clear that if the sum over the u_j converges at some points in the exterior domain, the convergence will be slow, in particular far away from the coupling boundary. If the exterior solution is of interest, it can be computed from u_0 by Green’s formula, which for one space dimension reduces to $u(r) = u_0 \exp(i\kappa(r-a))$. For inhomogeneous exterior domains without explicitly known Green’s function other numerical realizations of the pole condition can be used to compute the exterior solution (see [19]).

Remark 2.8 (choice of κ_0). It follows from (2.8) and (2.9) or from (2.25) that for scattering problems the optimal choice of κ_0 is $\kappa_0 = \kappa$ since in this case $U \equiv 0$, and we obtain the exact transparent boundary condition even with no degrees of freedom in $H^+(S^1)$. For resonance problems, κ_0 should be chosen in the region of the complex plane where resonances are of interest. In this case it is advantageous to choose κ_0

as a complex number with $\Im(\kappa_0) < 0$ and $\Re(\kappa_0) > 0$. All results of this section can be generalized to this case: u is outgoing if and only if $\mathcal{L}u|_{\kappa_0\mathbb{R}}$ belongs to the space $H^-(\kappa_0\mathbb{R}) := \{f(\kappa_0^{-1}\bullet) : f \in H^-(\mathbb{R})\}$. \mathcal{M}_{κ_0} maps $H^-(\kappa_0\mathbb{R})$ bijectively to $H^+(S^1)$. In Theorems 2.4, 2.5, 2.7 and Corollary 2.6 we have to replace the conditions on κ and κ_0 by $\Re(\kappa/\kappa_0) > 0$ and $\Re(\kappa^2/\kappa_0) > 0$. These are reasonable assumptions, since κ_0 should be chosen close to the resonances κ of interest anyway.

3. Helmholtz equation in higher dimensions. In this section we will treat the Helmholtz equation in higher dimensions in a manner similar to that in the previous section for one dimension. Besides the resonance problem (1.1) we will also study the scattering problem

$$(3.1a) \quad -\Delta u - \kappa^2 u = 0 \quad \text{in } \mathbb{R}^d \setminus K,$$

$$(3.1b) \quad \frac{\partial u}{\partial \nu} = f \quad \text{on } \partial K,$$

$$(3.1c) \quad u \text{ satisfies a radiation condition}$$

for given $\kappa \in \mathbb{C}$ with $\Re(\kappa) > 0$ and $f \in H^{-1/2}(\partial K)$. This will be done by considering the Laplace transform of the scaled exterior solution

$$(3.2) \quad u_{\text{ext}}(r, \hat{x}) := (r + 1)^{(d-1)/2} u((r + 1)\hat{x}), \quad r > 0, \hat{x} \in \Gamma := \partial B_a,$$

with respect to the radial variable r , i.e.,

$$(3.3) \quad (\mathcal{L}u_{\text{ext}})(s, \hat{x}) := \int_0^\infty e^{-sr} u_{\text{ext}}(r, \hat{x}) dr, \quad \Re(s) > |\Im(\kappa)|, \hat{x} \in \Gamma.$$

The radial variable is scaled such that $u_{\text{ext}}(r, \hat{x}) \sim \exp(ikar)u_\infty(\hat{x})$ as $r \rightarrow \infty$. This scaling is not essential, but simplifies the computations. In particular, we will be able to use part of the analysis of the previous section.

3.1. Pole condition in terms of Hardy spaces. Recall that for Riemannian manifolds A, B the spaces

$$L^2(A; L^2(B)) \sim L^2(A \times B) \sim L^2(A) \otimes L^2(B)$$

are isometrically isomorphic. Consequently, $H^-(\mathbb{R}) \otimes L^2(\Gamma)$ can be considered as a closed subspace of $L^2(\mathbb{R} \times \Gamma)$. It consists of all functions $f \in L^2(\mathbb{R} \times \Gamma)$ for which there exists a measurable function $v : \mathbb{C}^- \times \Gamma \rightarrow \mathbb{C}$, which is holomorphic in the first variable such that $\sup_{\epsilon > 0} \int_{\mathbb{R}} \int_{\Gamma} |v(s - i\epsilon, \hat{x})|^2 d\hat{x} ds < \infty$ and

$$\int_{\mathbb{R}} \int_{\Gamma} |f(s, \hat{x}) - v(s - i\epsilon, \hat{x})|^2 d\hat{x} ds \xrightarrow{\epsilon \rightarrow 0} 0.$$

If $v = \mathcal{L}u_{\text{ext}}$, we will shorten this to $\mathcal{L}|_{\mathbb{R}}u_{\text{ext}} := f$. Again, v can be recovered from f by a Cauchy integral as in (2.6).

DEFINITION 3.1. *Let u be a complex-valued function on $\mathbb{R}^d \setminus K$, and assume that the Laplace transform $(\mathcal{L}u_{\text{ext}})(s, \bullet)$ is well defined by (3.2) and (3.3) for all s in some open region $D \subset \mathbb{C}$ and belongs to $L^2(\Gamma)$. We say that u satisfies the pole condition if the function $D \rightarrow L^2(\Gamma)$, $s \mapsto (\mathcal{L}u_{\text{ext}})(s, \bullet)$ has a holomorphic extension to \mathbb{C}^- , and $\mathcal{L}|_{\mathbb{R}}u_{\text{ext}}$ belongs to $H^-(\mathbb{R}) \otimes L^2(\Gamma)$.*

Remark 3.2. It is easy to see that Definition 3.1 without the condition $\mathcal{L}|_{\mathbb{R}}u_{\text{ext}} \in H^-(\mathbb{R}) \otimes L^2(\Gamma)$ is equivalent to the formulation in [9, Definition 2.1]. Moreover, it was

shown in [9, section 9] that the pole condition is equivalent to Sommerfeld’s radiation condition for solutions to the Helmholtz equation with $\kappa > 0$. From the results in that section in [9], in particular (9.14) and (9.9b), it can be seen that the condition $\mathcal{L}|_{\mathbb{R}}u_{\text{ext}} \in H^-(\mathbb{R}) \otimes L^2(\Gamma)$ is also satisfied at least for sufficiently large a .

Remark 3.3. In [9] only the case $\kappa > 0$ was considered. However, the pole condition is also a valid radiation condition for $\Im(\kappa) \neq 0$. The singularity of the Laplace transform $\mathcal{L}u_{\text{ext}}$ of an outgoing wave is still a singularity with a branch cut located at $i\kappa a$, and hence in the upper half-plane. As mentioned in the introduction, Sommerfeld’s radiation condition is not valid for $\Im(\kappa) < 0$, and hence no equivalence result holds true in this case. However, it is actually much simpler to prove equivalence of the pole condition and the radiation condition (1.2) since the Hankel function can be recovered from the pole condition approach (see [9, section 7]).

Note that the pole condition is independent of the differential equation. Solutions to the Helmholtz equation will belong to spaces of higher regularity with respect to the second variable.

In analogy to the previous section we consider the Möbius transform $\mathcal{M}_{\kappa_0} \otimes I_{L^2(\Gamma)}$ from $H^-(\kappa_0\mathbb{R}) \otimes L^2(\Gamma)$ to $H^+(S^1) \otimes L^2(\Gamma)$ and write $\widehat{U} := (\mathcal{M}_{\kappa_0} \otimes I_{L^2(\Gamma)})\mathcal{L}|_{\kappa_0\mathbb{R}}u_{\text{ext}}$. Moreover, we define $u_0 := u|_{\Gamma}$ and

$$(3.4) \quad U(z, \hat{x}) := \frac{2i\kappa_0\widehat{U}(z, \hat{x}) - u_0(\hat{x})}{z - 1}, \quad z \in S^1, \hat{x} \in \Gamma,$$

in analogy to (2.9).

3.2. Variational formulation. Assume that u is a solution to the scattering problem (3.1) and define $u_{\text{int}} := u|_{\Omega_{\text{int}}}$ with $\Omega_{\text{int}} := B_a \setminus K$ and u_{ext} by (3.2). Then for smooth, rapidly decaying test functions v a straightforward computation yields

$$\begin{aligned} & \int_{\Omega_{\text{int}}} \{ \nabla u_{\text{int}} \cdot \nabla v_{\text{int}} - \kappa^2 u_{\text{int}} v_{\text{int}} \} dx + \frac{d-1}{2a} \int_{\Gamma} u_0 v_0 d\hat{x} + \frac{1}{a} \int_{\Gamma} \int_0^{\infty} \partial_r u_{\text{ext}} \partial_r v_{\text{ext}} dr d\hat{x} \\ & + a \int_{\Gamma} \int_0^{\infty} \left\{ \frac{\nabla_{\hat{x}} u_{\text{ext}} \cdot \nabla_{\hat{x}} v_{\text{ext}}}{(r+1)^2} - \kappa^2 u_{\text{ext}} v_{\text{ext}} - \frac{C_d}{a^2} \frac{u_{\text{ext}} v_{\text{ext}}}{(r+1)^2} \right\} dr d\hat{x} = - \int_{\partial K} f v_{\text{int}} ds, \end{aligned}$$

with $C_d := \frac{(d-1)(3-d)}{4}$ and the surface gradient $\nabla_{\hat{x}}$ on Γ .

We first derive the transformation to the Hardy space formally. Due to (2.9), (2.14), and (2.23) we have

$$i\kappa_0(\mathcal{M}_{\kappa_0} \otimes I)\mathcal{L}|_{\kappa_0\mathbb{R}}u_{\text{ext}} = (\mathcal{T}_- \otimes I) \begin{pmatrix} u_0 \\ U \end{pmatrix}, \quad (\mathcal{M}_{\kappa_0} \otimes I)\mathcal{L}|_{\kappa_0\mathbb{R}}\partial_r u_{\text{ext}} = (\mathcal{T}_+ \otimes I) \begin{pmatrix} u_0 \\ U \end{pmatrix}.$$

By [9, Theorem 9.3] $(I \otimes \nabla_{\hat{x}})\mathcal{L}|_{\kappa_0\mathbb{R}}u_{\text{ext}}$ is also analytic with respect to the first variable s in \mathbb{C}^- and decays like $|s|^{-1}$ as $|s| \rightarrow \infty$. In addition we need to recall the identity

$$(3.5) \quad \mathcal{L} \left(\frac{f}{\bullet + 1} \right) (s) = (\widehat{J}\mathcal{L}f)(s) \quad \text{with} \quad (\widehat{J}f)(s) := \int_s^{\infty} e^{-(\sigma-s)} \widehat{f}(\sigma) d\sigma.$$

The inverse operator $\widehat{D} := \widehat{J}^{-1}$ arises from a multiplication with a factor $r + 1$, i.e., $(\widehat{D}\mathcal{L}f)(s) = \mathcal{L}\{(\bullet + 1)f\}(s) = (-\partial_s + 1)\mathcal{L}f(s)$. The Möbius transformed operators are defined by $D := \mathcal{M}_{\kappa_0}\widehat{D}\mathcal{M}_{\kappa_0}^{-1}$ and $J := \mathcal{M}_{\kappa_0}\widehat{J}\mathcal{M}_{\kappa_0}^{-1}$. As

$$\int_{\Gamma} \int_0^{\infty} f_1 f_2 dr d\hat{x} = \frac{-i\kappa_0}{\pi} A^{\#}(\widehat{F}_1, \widehat{F}_2),$$

with $A^\#(F_1, F_2) := \int_\Gamma \int_{S^1} F_1(\bar{z}, \hat{x}) F_2(z, \hat{x}) \, d|z| \, d\hat{x}$ for $\widehat{F}_j = (\mathcal{M}_{\kappa_0} \otimes I)\mathcal{L}|_{\kappa_0\mathbb{R}} f_j$, we obtain

$$\begin{aligned}
 & \int_{\Omega_{\text{int}}} \{ \nabla u_{\text{int}} \nabla v_{\text{int}} - \kappa^2 u_{\text{int}} v_{\text{int}} \} \, dx + \frac{d-1}{2a} \int_\Gamma u_0 v_0 \, d\hat{x} \\
 & - \frac{i\kappa_0}{a\pi} A^\# \left((\mathcal{T}_+ \otimes I) \begin{pmatrix} u_0 \\ U \end{pmatrix}, (\mathcal{T}_+ \otimes I) \begin{pmatrix} v_0 \\ V \end{pmatrix} \right) \\
 (3.6) \quad & - \frac{ai\kappa^2}{\pi\kappa_0} A^\# \left((\mathcal{T}_- \otimes I) \begin{pmatrix} u_0 \\ U \end{pmatrix}, (\mathcal{T}_- \otimes I) \begin{pmatrix} v_0 \\ V \end{pmatrix} \right) \\
 & + \frac{ai}{\pi\kappa_0} A^\#_{\text{tan}} \left((J\mathcal{T}_- \otimes \nabla_{\hat{x}}) \begin{pmatrix} u_0 \\ U \end{pmatrix}, (J\mathcal{T}_- \otimes \nabla_{\hat{x}}) \begin{pmatrix} v_0 \\ V \end{pmatrix} \right) \\
 & - \frac{iC_d}{\pi\kappa_0 a} A^\# \left((J\mathcal{T}_- \otimes I) \begin{pmatrix} u_0 \\ U \end{pmatrix}, (J\mathcal{T}_- \otimes I) \begin{pmatrix} v_0 \\ V \end{pmatrix} \right) = - \int_{\partial K} f v_{\text{int}}|_{\partial K} \, ds.
 \end{aligned}$$

If $L^2_{\text{tan}}(\Gamma)$ denotes the space of square integrable tangential vector fields on Γ , we define $A^\#_{\text{tan}}(F_1, F_2) := \int_\Gamma \int_{S^1} F_1(\bar{z}, \hat{x}) \cdot F_2(z, \hat{x}) \, d|z| \, d\hat{x}$.

This bilinear form suggests introducing the space

(3.7a)

$$X^\# := \left\{ \begin{pmatrix} u_{\text{int}} \\ U \end{pmatrix} \in H^1(\Omega_{\text{int}}) \oplus H^+(S^1) \otimes L^2(\Gamma) : (J\mathcal{T}_- \otimes \nabla_{\hat{x}}) \begin{pmatrix} u_0 \\ U \end{pmatrix} \in H^+(S^1) \otimes L^2_{\text{tan}}(\Gamma) \right\},$$

with the inner product

(3.7b)

$$\begin{aligned}
 \left\langle \begin{pmatrix} u_{\text{int}} \\ U \end{pmatrix}, \begin{pmatrix} v_{\text{int}} \\ V \end{pmatrix} \right\rangle_{X^\#} & := \langle u_{\text{int}}, v_{\text{int}} \rangle_{H^1(\Omega_{\text{int}})} + \langle U, V \rangle_{H^+(S^1) \otimes L^2(\Gamma)} \\
 & + \left\langle (J\mathcal{T}_- \otimes \nabla_{\hat{x}}) \begin{pmatrix} u_0 \\ U \end{pmatrix}, (J\mathcal{T}_- \otimes \nabla_{\hat{x}}) \begin{pmatrix} v_0 \\ V \end{pmatrix} \right\rangle_{H^+(S^1) \otimes L^2_{\text{tan}}(\Gamma)}.
 \end{aligned}$$

It is easy to see that the bilinear form in (3.6) is bounded with respect to the norm of $X^\#$. It is shown in Lemma A.3 that $X^\#$ with this inner product is a Hilbert space, and for each $v_{\text{int}} \in H^1(\Omega)$ there exists a vector in $X^\#$ containing $v_{\text{int}} \in H^1(\Omega)$ as first component (note that the surface gradient $\nabla_{\hat{x}}$ is not applied to the $H^{1/2}(\Gamma)$ -function v_0 , but to a sum with other functions). Moreover, it is shown in Lemma A.3 that there exists a dense subset of test functions $(v_{\text{int}}, V)^\top \in X^\#$ for which the transforms above are justified. Therefore, we obtain the following result.

THEOREM 3.4. *If u is a solution to the scattering problem (3.1), then (u_{int}, U) belongs to the space $X^\#$ and satisfies the symmetric variational equation (3.6).*

The converse result will be shown later in Corollary 4.3 using a separation argument.

3.3. Galerkin discretization. Let $V_h \subset H^1(\Omega_{\text{int}})$ be a finite element subspace on the computational domain Ω_{int} , and let $V_h|_\Gamma$ denote the set of traces of functions in V_h on the artificial boundary Γ . Moreover, we use the polynomial subspace $\Pi_N \subset H^+(S^1)$ as in section 2. We will use a Galerkin method where the space $X^\#$ in Theorem (3.4) is approximated by the finite-dimensional subspace

$$(3.8) \quad X^\#_{h,N} := V_h \oplus \Pi_N \otimes V_h|_\Gamma.$$

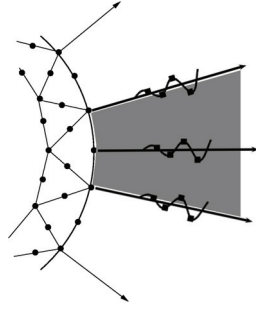


FIG. 3.1. Hardy space infinite element corresponding to quadratic Lagrange elements.

For a given finite element basis of V_h let $\{w_j : j = 0, \dots, N_\Gamma\}$ denote the corresponding set of nonvanishing traces on Γ . Then we choose the functions $(z, \hat{x}) \mapsto z^n w_j(\hat{x})$ ($j = 0, \dots, N_\Gamma, n = 0, \dots, N$) as the basis of $\Pi_N \otimes V_h|_\Gamma$. The system matrix with respect to this basis can be assembled elementwise in a finite element fashion as illustrated in Figure 3.1. Each infinite element couples with the interior finite elements via common degrees of freedom for the Dirichlet values on Γ . Moreover, there is a coupling between neighboring infinite elements. Due to the structure of the bilinear form (3.6), the local element matrices are sums of Kronecker products of matrices. Let M_{el}^Γ and S_{el}^Γ denote the element mass and stiffness matrix on Γ corresponding to the bilinear forms $\int_\Gamma u_0 v_0 d\hat{x}$ and $\int_\Gamma \nabla_{\hat{x}} u_0 \cdot \nabla_{\hat{x}} v_0 d\hat{x}$, respectively. The discrete representation of the operators \mathcal{T}_\pm has already been described in section 2; see (2.24). It remains to discuss the discretization of the operator J . Recall that J is the inverse of a differential operator D , which is given explicitly by

$$(3.9) \quad (DF)(z) = \frac{(z-1)^2}{2i\kappa_0} F'(z) + \left(\frac{z-1}{2i\kappa_0} + 1 \right) F(z), \quad F \in H^+(S^1).$$

To avoid numerical integrations, we use the inverse of the discretization of D

$$(3.10) \quad D_N := \text{id}_{(N+1) \times (N+1)} + \frac{1}{2i\kappa_0} \begin{pmatrix} -1 & 1 & & & & \\ 1 & -3 & 2 & & & \\ & 2 & -5 & 3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & & N & -2N-1 \end{pmatrix}$$

as the discretization of J . Hence, the element matrix of a Hardy space infinite element is given by

$$(3.11) \quad L_1 \otimes M_\Gamma^{el} + L_2 \otimes S_\Gamma^{el} - \kappa^2 L_3 \otimes M_\Gamma^{el},$$

with

$$L_1 = \frac{d-1}{2a} \begin{pmatrix} 1 & \\ & \mathbf{0} \end{pmatrix} - \frac{2i\kappa_0}{a} \mathcal{T}_{N,+}^\top \mathcal{T}_{N,+} - \frac{2C_{di}}{\kappa_0 a} \mathcal{T}_{N,-}^\top D_N^{-2} \mathcal{T}_{N,-},$$

$$L_2 = \frac{2ai}{\kappa_0} \mathcal{T}_{N,-}^\top D_N^{-2} \mathcal{T}_{N,-}, \quad \text{and} \quad L_3 = \frac{2ai}{\kappa_0} \mathcal{T}_{N,-}^\top \mathcal{T}_{N,-}.$$

Note that the eigenvalue structure with respect to κ^2 is preserved for the discretization with Hardy space infinite elements.

Remark 3.5. The Hardy space infinite element method is not restricted to the case of spherical artificial boundaries $\Gamma = \partial B_a$. We have applied the method also to boundaries $\Gamma = \partial P$ with convex polyhedrons P using the segmentation of the exterior domain $\Omega_{\text{ext}} := \mathbb{R}^d \setminus P$ presented in [17, 24]. Although the variational formulation becomes more complicated, the method still seems to converge superalgebraically (see [15]).

4. Convergence analysis for the separated problems. In this section we analyze the convergence of Hardy space infinite elements in the exterior domain (i.e., for the special case $K = B_a$) after a Fourier separation. Implications for the full problem are discussed in section 4.4.

4.1. The separated equations. For this end, we choose an orthonormal basis of eigenfunctions $\Phi_n \in L^2(\Gamma)$, $n \in \mathbb{N}_0$, such that $-\Delta_{\hat{x}}\Phi_n = \lambda_n\Phi_n$ for the Laplace–Beltrami operator $\Delta_{\hat{x}}$ on Γ . The functions u_0 and U have expansions with respect to this basis of the form $u_0(\hat{x}) = \sum_{n=0}^{\infty} u_{0,n}\Phi_n(\hat{x})$, $U(z, \hat{x}) = \sum_{n=0}^{\infty} U_n(z)\Phi_n(\hat{x})$, and similarly for v_0 and V . Moreover, the Neumann data on $\partial K = \Gamma$, which will be denoted by g instead of f in this section, can be decomposed into the Fourier series $g(\hat{x}) = \sum_{n=0}^{\infty} g_n\Phi_n(\hat{x})$. Then the variational problem (3.6) decouples into a series of variational problems in $\tilde{X} := \mathbb{C} \oplus H^+(S^1)$:

$$(4.1) \quad B_1\left(\begin{pmatrix} u_{0,n} \\ U_n \end{pmatrix}, \begin{pmatrix} v_0 \\ V \end{pmatrix}\right) + \frac{C_d - a^2\lambda_n}{\kappa_0^2 a} B_2\left(\begin{pmatrix} u_{0,n} \\ U_n \end{pmatrix}, \begin{pmatrix} v_0 \\ V \end{pmatrix}\right) = -g_n v_0, \quad \begin{pmatrix} v_0 \\ V \end{pmatrix} \in \tilde{X},$$

for the Fourier coefficients, where the bilinear forms B_1, B_2 on \tilde{X} are given by

$$\begin{aligned} B_1\left(\begin{pmatrix} u_0 \\ U \end{pmatrix}, \begin{pmatrix} v_0 \\ V \end{pmatrix}\right) &:= \frac{d-1}{2a} u_0 v_0 \\ &\quad - \frac{i\kappa_0}{a\pi} A\left(\mathcal{T}_+\begin{pmatrix} u_0 \\ U_n \end{pmatrix}, \mathcal{T}_+\begin{pmatrix} v_0 \\ V_n \end{pmatrix}\right) - \frac{ai\kappa^2}{\pi\kappa_0} A\left(\mathcal{T}_-\begin{pmatrix} u_0 \\ U \end{pmatrix}, \mathcal{T}_-\begin{pmatrix} v_0 \\ V \end{pmatrix}\right), \\ B_2\left(\begin{pmatrix} u_0 \\ U \end{pmatrix}, \begin{pmatrix} v_0 \\ V \end{pmatrix}\right) &:= -\frac{i\kappa_0}{\pi} A\left(J\mathcal{T}_-\begin{pmatrix} u_0 \\ U \end{pmatrix}, J\mathcal{T}_-\begin{pmatrix} v_0 \\ V \end{pmatrix}\right). \end{aligned}$$

We use the canonical inner product on \tilde{X} given by the sum of the inner products on \mathbb{C} and $H^+(S^1)$. Defining the operators $K_j : \tilde{X} \rightarrow \tilde{X}$ ($j = 1, 2$) implicitly by

$$\left\langle K_j \begin{pmatrix} u_0 \\ U \end{pmatrix}, \begin{pmatrix} v_0 \\ V \end{pmatrix} \right\rangle_{\tilde{X}} = B_j\left(\begin{pmatrix} u_0 \\ U \end{pmatrix}, \begin{pmatrix} \overline{v_0} \\ \mathcal{C}V \end{pmatrix}\right), \quad \begin{pmatrix} u_0 \\ U \end{pmatrix}, \begin{pmatrix} v_0 \\ V \end{pmatrix} \in \tilde{X},$$

the variational equations (4.1) can be reformulated as operator equations

$$(4.2) \quad K_1 \begin{pmatrix} u_{0,n} \\ U_n \end{pmatrix} + \frac{C_d - a^2\lambda_n}{a\kappa_0^2} K_2 \begin{pmatrix} u_{0,n} \\ U_n \end{pmatrix} = \begin{pmatrix} -g_n \\ 0 \end{pmatrix}.$$

4.2. Uniqueness and smoothness of solutions. Motivated by the Paley–Wiener theorem (2.5) we introduce a transform $\mathcal{Q} : \tilde{X} \rightarrow L^2(\mathbb{R}_+)$ by

$$(4.3) \quad \left(\mathcal{Q} \begin{pmatrix} f_0 \\ F \end{pmatrix}\right)(t) := \frac{-1}{2\pi} \int_{-\infty}^{\infty} e^{ist} \left(\mathcal{M}_{\kappa_0}^{-1} \mathcal{T}_-\begin{pmatrix} f_0 \\ F \end{pmatrix}\right)(\kappa_0 s) ds, \quad t \geq 0.$$

The following result will be used to show uniqueness, but may also be of independent interest.

LEMMA 4.1. \mathcal{Q} is a norm isomorphism from \tilde{X} to the Sobolev space $H^1(\mathbb{R}_+)$, and $f := \mathcal{Q}(f_0, F)^\top$ satisfies $f(0) = f_0$ and

$$(4.4) \quad f'(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ist} \left(\mathcal{M}_{\kappa_0}^{-1} \mathcal{T}_+ \begin{pmatrix} f_0 \\ F \end{pmatrix} \right) (\kappa_0 s) ds, \quad t \geq 0.$$

Proof. Let us first show that the range of \mathcal{Q} is contained in $H^1(\mathbb{R}_+)$. Due to (2.5) we have $f(t) = 0$ for $t < 0$ if we use definition (4.3) also for $t < 0$. Therefore we get $f \in H^1(\mathbb{R}^+)$ if we can show that $w(t) := f(t) + f(-t)$, $t \in \mathbb{R}$, belongs to $H^1(\mathbb{R})$. Introducing $\tilde{f} := (i\kappa_0)^{-1} \mathcal{M}_{\kappa_0}^{-1} \mathcal{T}_-(f_0, F)^\top$ we have $\tilde{f}(\kappa_0 s) = (-i\kappa_0)^{-1} (\mathcal{F}f)(s)$ and $-i\kappa_0 (\mathcal{F}w)(s) = \tilde{f}(\kappa_0 s) + \tilde{f}(-\kappa_0 s)$. Due to (2.14) and the definition (2.23) of \mathcal{T}_+ , the function $\bullet \tilde{f} - f_0 = \mathcal{M}_{\kappa_0}^{-1} \mathcal{T}_+(f_0, F)^\top$ belongs to $H^-(\kappa_0 \mathbb{R})$. Hence, the function

$$-i\kappa_0^2 s (\mathcal{F}w)(s) = \left(\kappa_0 s \tilde{f}(\kappa_0 s) - f_0 \right) - \left(-\kappa_0 s \tilde{f}(-\kappa_0 s) - f_0 \right), \quad s \in \mathbb{R},$$

is square integrable, and therefore $\int_{-\infty}^{\infty} (1 + s^2) |(\mathcal{F}w)(s)|^2 ds < \infty$. This implies that $w \in H^1(\mathbb{R})$. To prove the second assertion first note that

$$(4.5) \quad \int_0^{\infty} e^{-ist} f'(t) dt = -f(0) + is \int_0^{\infty} e^{-ist} f(t) dt, \quad s \in \mathbb{R}.$$

Since we have already shown that $f' \in L^2(\mathbb{R}_+)$, the right-hand side is a square integrable function of s by Plancherel's theorem. As $\bullet \tilde{f} - f_0 = \mathcal{M}_{\kappa_0}^{-1} \mathcal{T}_+(f_0, F)^\top$ also belongs to $L^2(\kappa_0 \mathbb{R})$, the constant function $f(0) - f_0$ is square integrable and hence 0. Therefore, $f(0) = f_0$, and applying the inverse Fourier transform to (4.5) yields (4.4).

\mathcal{Q} is injective as a composition of injective operators. To prove that \mathcal{Q} is onto, choose an arbitrary $v \in H^1(\mathbb{R}^+)$ and extend it by zero on the negative real axis. Then (2.5) implies that $\mathcal{F}v \in H^-(\mathbb{R})$, and hence $\widehat{V} := (-i\kappa_0)^{-1} \mathcal{M}_{\kappa_0}(\mathcal{F}v)(\kappa_0^{-1} \bullet)$ belongs to $H^+(S^1)$. Moreover, $\{\mathcal{M}_{\kappa_0}(\mathcal{F}v')(\kappa_0^{-1} \bullet)\}(z) = i\kappa_0 \widehat{V}(z) + \frac{2i\kappa_0 \widehat{V}(z) - v_0}{z-1}$ with $v_0 := v(0)$ is an element of $H^+(S^1)$. Hence, the function $V(z) := \frac{2i\kappa_0 \widehat{V}(z) - v_0}{z-1}$ (cf. (2.9)) belongs to $H^+(S^1)$, and we have $(\mathcal{M}_{\kappa_0} \mathcal{T}_-(v_0, V)^\top)(\kappa_0 s) = -(\mathcal{F}v)(s)$, so $\mathcal{Q}(v_0, V)^\top = v$. The boundedness of \mathcal{Q}^{-1} follows either directly from the construction above or the open mapping theorem. \square

Note that the separation index n is the index of an enumeration of the double indices $(l, m) = (l(n), m(n))$ in (1.2). Hence, solutions to (4.2) are given by modified (due to the scaling in (3.2)) and Laplace and Möbius transformed Hankel functions $\mathcal{H}_n^{(1/2)}(r) := r^{1-d/2} H_{l(n)-1+d/2}^{(1/2)}(r)$.

PROPOSITION 4.2. Let $\Re(\kappa/\kappa_0) > 0$. If $\mathcal{H}_n^{(1)'}(\kappa a) \neq 0$, then (4.1) has a unique solution $(u_{0,n}, U_n)^\top \in \tilde{X}$ and $u_{0,n} = \frac{\mathcal{H}_n^{(1)}(\kappa a)}{\kappa \mathcal{H}_n^{(1)'}(\kappa a)} g_n$. If $\mathcal{H}_n^{(1)'}(\kappa a) = 0$, then (4.1) has a solution if and only if $g_n = 0$.

Proof. Using Lemmas 4.1 and A.1 and the Fourier convolution theorem, it can be shown that (4.1) is equivalent to the variational problem to find $u_n \in H^1(\mathbb{R}_+)$ such

that

$$\begin{aligned} & \frac{i}{a\kappa_0} \int_0^\infty \left(-\kappa_0^2 u_n'(t)v'(t) - (\kappa a)^2 u_n(t)v(t) + \frac{a^2 \lambda_n - C_d}{\left(\frac{it}{\kappa_0} + 1\right)^2} u_n(t)v(t) \right) dt \\ & + \frac{d-1}{2a} u_n(0)v(0) = -g_n v(0) \end{aligned}$$

for all $v \in H^1(\mathbb{R}_+)$. This is the variational formulation of the exterior boundary value problem

$$(4.6a) \quad \kappa_0^2 u_n''(t) - \left((\kappa a)^2 + \frac{C_d - a^2 \lambda_n}{\left(\frac{it}{\kappa_0} + 1\right)^2} \right) u_n(t) = 0, \quad t \geq 0, \hat{x} \in \Gamma,$$

$$(4.6b) \quad u_n'(0) = \frac{i}{\kappa_0} \left(ag_n + \frac{d-1}{2} u_n(0) \right),$$

$$(4.6c) \quad u_n \in L^2(\mathbb{R}^+).$$

The general solution of the differential equation (4.6a) is given by

$$u_n(t) = \left(\frac{it}{\kappa_0} + 1 \right)^{(d-1)/2} \left(c_n^{(1)} \mathcal{H}_n^{(1)} \left(\kappa a \left(\frac{it}{\kappa_0} + 1 \right) \right) + c_n^{(2)} \mathcal{H}_n^{(2)} \left(\kappa a \left(\frac{it}{\kappa_0} + 1 \right) \right) \right).$$

Due to the asymptotic behavior (1.3) of the Hankel functions and the assumption $\Re(\kappa/\kappa_0) > 0$, (4.6c) implies that $c_n^{(2)} = 0$. If $\mathcal{H}_n^{(1)'(\kappa a)} \neq 0$, then the boundary condition (4.6b) implies $u_{0,n} = (\mathcal{H}_n^{(1)}(\kappa a)/(\kappa \mathcal{H}_n^{(1)'(\kappa a)}))g_n$. Otherwise (4.6b) is satisfied if and only if $g_n = 0$. \square

As a corollary we obtain the converse of Theorem 3.4.

COROLLARY 4.3. *If $(u_{\text{int}}, U)^\top \in X^\#$ is a solution to the variation problem (3.6) and $\mathcal{H}_n^{(1)'(\kappa a)} \neq 0$, then u_{int} is the restriction of a solution to (3.1).*

Proof. Let $(u_{\text{int}}, U)^\top$ be a solution to (3.6) and let $\partial_\nu u \in H^{-1/2}(\Gamma)$ denote the Neumann trace. We rearrange the terms in (3.6) such that only the integrals over Ω_{int} and ∂K are on the left-hand side to obtain

$$\int_{\partial K} f v_{\text{int}}|_{\partial K} ds + B_{\text{int}}(u_{\text{int}}, v_{\text{int}}) = B_{\text{ext}} \left(\begin{pmatrix} u_0 \\ U \end{pmatrix}, \begin{pmatrix} v_0 \\ V \end{pmatrix} \right).$$

It follows that $B_{\text{ext}}((u_0, U)^\top, (v_0, V)^\top) = \int_\Gamma \partial_\nu u v_0 ds$ for all $(v_0, V)^\top$. Now we can apply a Fourier separation on Γ and use Proposition 4.2 to obtain the relation $\mathcal{H}_n^{(1)'(\kappa a)} u_{0,n} = \mathcal{H}_n^{(1)}(\kappa a) (\partial_\nu u)_n$ for the Fourier coefficients $(\partial_\nu u)_n := \int_\Gamma \partial_\nu u \overline{\Phi_n} ds$. Therefore, we can define an outgoing exterior solution by (1.2) with the constants $\alpha_{l(n),m(n)} = \frac{\mathcal{H}_n^{(1)}(\kappa a)}{\kappa \mathcal{H}_n^{(1)'(\kappa a)}} (\partial_\nu u)_n$, which has the same Cauchy data on Γ as u_{int} . \square

LEMMA 4.4. *We have $U_n \in H^+(S^1) \cap C^\infty(S^1)$.*

Proof. It follows from [9, Proposition 6.6 and Lemma 6.3] that the Fourier coefficients of the Laplace transform, $\hat{u}_n(s) := \langle \mathcal{L}u_{\text{ext}}(x, \cdot), \Phi_n \rangle_{L^2(\Gamma)}$, have an integral representation of the form

$$\hat{u}_n(s) = -\frac{c_n}{i\kappa a - s} - \int_0^\infty \frac{c_n \psi_n(t)}{i\kappa a - t - s} dt, \quad s \in \mathbb{C} \setminus \{i\kappa a - t : t \geq 0\},$$

with a constant $c_n \in \mathbb{C}$ and a function $\psi_n(t)$ decaying exponentially as $t \rightarrow \infty$. This implies that $\hat{u}_n|_{\mathbb{R}} \in H^-(\mathbb{R}) \cap C^\infty(\mathbb{R})$. Hence, $\widehat{U}_n := \mathcal{M}_{\kappa_0}(\hat{u}_n)$ belongs to $H^+(S^1) \cap C^\infty(S^1 \setminus \{1\})$. It remains to study the asymptotic behavior of \hat{u}_n at infinity, or equivalently the behavior of \widehat{U}_n at 1. Expanding the integral kernel in powers of $1/(s - i\kappa_0)$ and using the exponential decay of ψ_n , it can be shown that \hat{u}_n has an asymptotic expansion

$$\hat{u}_n(s) = \sum_{j=1}^J \frac{\alpha_j^{(n)}}{(s - i\kappa_0)^j} + o(|s - i\kappa_0|^{-J}), \quad |s| \rightarrow \infty,$$

for any $J \in \mathbb{N}$. By well-known asymptotic formulas for the Laplace transform we have $u_{0,n} = \alpha_1^{(n)}$. Since $(\mathcal{M}_{\kappa_0}((\bullet - i\kappa_0)^{-j}))(z) = (z - 1)^{j-1}/(2i\kappa_0)^j$, it follows that \widehat{U}_n satisfies

$$\widehat{U}_n(z) = \sum_{j=1}^J \frac{\alpha_j^{(n)}}{(2i\kappa_0)^j} (z - 1)^{j-1} + o(|z - 1|^{J-1}), \quad \text{as } |z - 1| \rightarrow 0.$$

Therefore,

$$U_n(z) = \frac{2i\kappa_0 \widehat{U}_n(z) - \alpha_1^{(n)}}{z - 1} = \sum_{j=2}^J \frac{\alpha_j^{(n)}}{(2i\kappa_0)^{j-1}} (z - 1)^{j-2} + o(|z - 1|^{J-2}), \quad \text{as } |z - 1| \rightarrow 0.$$

This implies that U_n is $J - 2$ times differentiable at 1. Since J was arbitrary, this together with the properties of \widehat{U}_n shows that $U_n \in H^+(S^1) \cap C^\infty(S^1)$. \square

4.3. Convergence. The bilinear form aB_1 essentially coincides with the exterior part B_{ext} of the bilinear form from the one-dimensional case. As in (2.21) we have

$$(4.7) \quad \Re \{ (i + \gamma)B_1((u_0, U), (\bar{u}_0, \mathcal{C}U)) \} \geq \alpha \|U\|_X^2$$

for some $\alpha, \gamma > 0$ if $\Re(\kappa_0), \Re(\kappa^2/\kappa_0) > 0$. Therefore, K_1 is boundedly invertible.

LEMMA 4.5. *The operator K_2 is compact.*

Proof. K_2 is a rank-1 perturbation of the operator $K_3 : H^+(S^1) \rightarrow H^+(S^1)$ given implicitly by

$$(4.8) \quad (K_3U, V)_{H^+(S^1)} = -\frac{i\kappa_0}{\pi} \int_{S^1} (\bar{z} - 1)J^2(z - 1)U(z)\overline{V(z)}|dz|.$$

Here we have used the boundedness of $J : H^+(S^1) \rightarrow H^+(S^1)$ (see (4.9a)) and the symmetry property $A(U, JV) = A(JU, V)$, which follows from the representation of $D = J^{-1}$ with respect to the monomial basis. Since the orthogonal projection $P : L^2(S^1) \rightarrow H^+(S^1)$ and the operator $H^+(S^1) \rightarrow H^+(S^1)$, $U \mapsto J((\bullet - 1)U)$ are bounded, it suffices to show the compactness of $\tilde{K}_4 : H^+(S^1) \rightarrow L^2(S^1)$, $(\tilde{K}_4U)(z) = (\bar{z} - 1)(JU)(z)$, or equivalently the compactness of $K_4 := H^-(\mathbb{R}) \rightarrow L^2(\mathbb{R})$, $(K_4f)(s) := \frac{2i\kappa_0}{s + i\kappa_0}(\tilde{J}f)(s)$. The following inequalities hold for some constants $C > 0$, $f \in H^-(\mathbb{R})$, and $s, s_1, s_2 \in \mathbb{R}$:

$$(4.9a) \quad \|\widehat{J}f\|_2 \leq C\|f\|_2,$$

$$(4.9b) \quad |(K_4f)(s)| \leq \frac{C}{|s + i\kappa_0|}\|f\|_2,$$

$$(4.9c) \quad |(K_4f)(s_1) - (K_4f)(s_2)| \leq C\sqrt{|s_1 - s_2|}\|f\|_2.$$

The first inequality is a consequence of Plancherel’s theorem, since $\widehat{J}f = g * f$ with $g(t) := e^{-t}$ for $t \geq 0$ and $g(t) \equiv 0$ for $t < 0$:

$$\|\widehat{J}f\|_2 = \|g * f\|_2 = 2\pi\|\mathcal{F}(g * f)\|_2 = \sqrt{2\pi}\|\mathcal{F}g \mathcal{F}f\|_2 \leq \sqrt{2\pi}\|\mathcal{F}g\|_\infty\|\mathcal{F}f\|_2 \leq C\|f\|_2.$$

For the third inequality we assume without loss of generality that $s_2 > s_1$ and write

$$\left| \frac{(\widehat{J}f)(s_1)}{s_1 + i\kappa_0} - \frac{(\widehat{J}f)(s_2)}{s_2 + i\kappa_0} \right| = \left| \int_{s_1}^{s_2} \frac{e^{(s_1-\sigma)}}{s_1 + i\kappa_0} f(\sigma) d\sigma + \int_{s_2}^\infty \left(\frac{e^{(s_1-\sigma)}}{s_1 + i\kappa_0} - \frac{e^{(s_2-\sigma)}}{s_2 + i\kappa_0} \right) f(\sigma) d\sigma \right|.$$

The first integral can be estimated with the Cauchy–Schwarz inequality by

$$|I_1| \leq \sqrt{|s_2 - s_1|} \sup_{\sigma \in [s_1, s_2]} \left| \frac{e^{(s_1-\sigma)}}{s_1 + i\kappa_0} \right|^2 \left(\int_{s_1}^{s_2} |f(\sigma)|^2 d\sigma \right)^{1/2} \leq C\sqrt{|s_1 - s_2|}\|f\|_2.$$

For I_2 the mean value theorem and the Cauchy–Schwarz inequality yield

$$|I_2| \leq \widetilde{C}|s_2 - s_1| \sup_{t \in [0,1]} \left| e^{(t-1)(s_2-s_1)} \right| \left(\int_{s_2}^\infty |e^{(s_2-\sigma)}|^2 d\sigma \right)^{1/2} \|f\|_2,$$

and we have shown (4.9c). Inequality (4.9b) can be proven in an analogous manner.

In order to show the compactness of K_4 we use the Arzelà–Ascoli theorem. Thus let $(w_n)_{n \in \mathbb{N}}$ be a sequence in $H^-(\mathbb{R})$ with $\|w_n\|_2 \leq 1$ for all $n \in \mathbb{N}$ and $v_n := K_4 w_n$. Due to the Arzelà–Ascoli theorem, there exists a subsequence of (v_n) which converges in the supremum norm of a compact subset I of \mathbb{R} , since (v_n) is equicontinuous and bounded in I by (4.9b) and (4.9c). Let $I_j := [-j, j] \subset \mathbb{R}$, $v_{n_0(l)} := v_l$. Moreover, for every $j \in \mathbb{N}$ let $(v_{n_j(l)})$ be a subsequence of $(v_{n_{j-1}(l)})$ converging in the supremum norm of I_j . Thus the diagonal subsequence $v_{n(l)} := v_{n_l(l)}$ converges pointwise in \mathbb{R} and for each I_j in the supremum norm of I_j to a function v . For given $\epsilon > 0$ it remains to show that there exists a $l_0(\epsilon) \in \mathbb{N}$ such that $\|v_{n(l)} - v\|_2 < \epsilon$ for all $l \geq l_0$. This can be seen with (4.9b) since there exists a $j_0(\epsilon) \in \mathbb{N}$ such that

$$\int_{\mathbb{R} \setminus I_{j_0}} |v_{n(l)}(s) - v(s)|^2 ds \leq 2C \int_{\mathbb{R} \setminus I_{j_0}} \frac{1}{|s + i\kappa_0|^2} ds \leq \frac{\epsilon}{2}.$$

Because of the uniform convergence of $(v_{n(l)})$ in I_{j_0} , the subsequence $(v_{n(l)})$ of the image sequence $v_n = K_4 w_n$ converges in $L^2(\mathbb{R})$ and the proof is done. \square

With these preparations we easily obtain the following superalgebraic convergence result.

THEOREM 4.6. *Assume that $\kappa_0, \kappa/\kappa_0$, and κ^2/κ_0 have positive real part and that $\mathcal{H}_n^{(1)}(\kappa a) \neq 0$; i.e., κ is not a resonance of (4.1). Then there exist constants $N_0, C_l > 0$ such that for $N \geq N_0$ there exists a unique solution $(u_{0,n}^{(N)}, U_n^{(N)})^\top$ in the space $X_N := \mathbb{C} \oplus \Pi_N$ to the variational equation*

$$(4.10) \quad B_1 \left(\begin{pmatrix} u_{0,n}^{(N)} \\ U_n^{(N)} \end{pmatrix}, \begin{pmatrix} v_0^{(N)} \\ V^{(N)} \end{pmatrix} \right) + \frac{C_d - a^2 \lambda_n}{a\kappa_0^2} B_2 \left(\begin{pmatrix} u_{0,n}^{(N)} \\ U_n^{(N)} \end{pmatrix}, \begin{pmatrix} v_0^{(N)} \\ V^{(N)} \end{pmatrix} \right) = -g_n v_0^{(N)}$$

for $(v_0^{(N)}, V^{(N)})^\top \in X_N$. Moreover, for any $l \in \mathbb{N}$ the error estimate

$$(4.11) \quad \left\| \begin{pmatrix} u_{0,n}^{(N)} \\ U_n^{(N)} \end{pmatrix} - \begin{pmatrix} u_{0,n} \\ U_n \end{pmatrix} \right\|_{\widetilde{X}} \leq \frac{C}{N^l}$$

holds for some constant C depending on l, n , and κ .

Proof. Due to the coercivity estimate (4.7) the method converges for the bilinear form B_1 . Using [12, Theorem 13.7], Proposition 4.2, and Lemma 4.5, it follows that the whole method (4.10) is stable and convergent. From the approximation properties of trigonometric polynomials and Lemma 4.4, it follows that the speed of convergence is superalgebraic. \square

Since the operators on the left-hand side of (4.2) are compact perturbations of Toeplitz operators, we could have appealed to more general convergence results for the finite section method (cf. [1, Chapter 7]) for an alternative proof of Theorem 4.6.

4.4. Discussion. For a fixed finite element subspace of $H^{1/2}(\Gamma)$, a separation argument in this subspace and Theorem 4.6 yield superalgebraic convergence to a transformed outgoing solution as $N \rightarrow \infty$. However, our results do not exclude the possibility that the constants in the convergence estimate explode as the mesh size tends to 0. To our knowledge this is also the state of the art for usual infinite elements in the space domain (cf. [3, 4]). Numerical evidence presented in Figure 5.1 suggests that both the discrete bilinear forms are bounded from above, and their inf-sup constants are bounded from below, both uniformly in the Hardy dimension N and the separation index n . We have not been able to prove this for the inf-sup constants so far. With such uniform estimates one would obtain convergence of the Neumann-to-Dirichlet (or equivalently the Dirichlet-to-Neumann) operators in the natural operator norms, which easily yields a convergence result for the scattering problem (3.1) (cf. [11, 10]).

5. Numerical results. We first study the separated equations and decompose the norm $\|\bullet\|_{X^\#} := \sqrt{\langle \bullet, \bullet \rangle_{X^\#}}$ into the norms

$$(5.1) \quad \left\| \begin{pmatrix} u_{0,n} \\ U_n \end{pmatrix} \right\|_{X_n}^2 := \sqrt{1 + \lambda_n} |u_{0,n}|^2 + \|U_n\|_{H^+(S^1)}^2 + \lambda_n \left\| \mathcal{J}\mathcal{T}_- \begin{pmatrix} u_{0,n} \\ U_n \end{pmatrix} \right\|_{H^+(S^1)}^2$$

for each Fourier coefficient $(u_{0,n}, U_n)^\top$ such that $\| \begin{pmatrix} u_0 \\ U \end{pmatrix} \|_{X^\#}^2 = \sum_n \| \begin{pmatrix} u_{0,n} \\ U_n \end{pmatrix} \|_{X_n}^2$. If $\underline{U}_n^{(N)} \in \mathbb{C}^N$ denotes the vector of the first N Fourier coefficients of U_n , the discrete counterpart on $X_{N,n} := \mathbb{C}^{N+1}$ is the norm

$$(5.2) \quad \left\| \begin{pmatrix} u_{0,n} \\ \underline{U}_n^{(N)} \end{pmatrix} \right\|_{X_{N,n}}^2 := \begin{pmatrix} u_{0,n} \\ \underline{U}_n^{(N)} \end{pmatrix}^* \left(\begin{pmatrix} \sqrt{1 + \lambda_n} & \\ & \mathbf{1} \end{pmatrix} + \lambda_n \mathcal{T}_{N,-}^\top D_{N,-}^{-*} D_{N,-}^{-1} \mathcal{T}_{N,-} \right) \begin{pmatrix} u_{0,n} \\ \underline{U}_n^{(N)} \end{pmatrix}.$$

Figure 5.1 show the norms and inf-sup constants with respect to the norm in (5.2) of the bilinear form in (4.10), which is represented by the matrix $T_n^{(N)} := L_1 + \lambda_n L_2 - \kappa^2 L_3$ (see (3.11)). They were computed using a Cholesky factorization $G = L^* L$ of the Gramian matrix G in (5.2) as $\|(L^\top)^{-1} T_n^{(N)} L^{-1}\|_2$ and $\|[(L^\top)^{-1} T_n^{(N)} L^{-1}]^{-1}\|_2$, respectively. Here $\|A\|_2$ denotes the spectral norm, i.e., the largest singular value of a matrix A . The results suggest that the norms are bounded from above and the inf-sup constants are bounded from below, both uniformly in N and n .

Figure 5.2 shows the convergence of the relative errors of the numerical approximations to the Neumann-to-Dirichlet numbers $\text{NtD}(n, \kappa, a) := H_n^{(1)}(\kappa a) / \kappa H_n^{(1)'}(\kappa a)$. These numerical approximations are computed by solving (4.10) with $g_n = 1$; they are given by the negative upper left entry of the matrices $[T_n^{(N)}]^{-1}$ defined above. The results exhibit a fast, almost exponential convergence as $N \rightarrow \infty$ for each Fourier

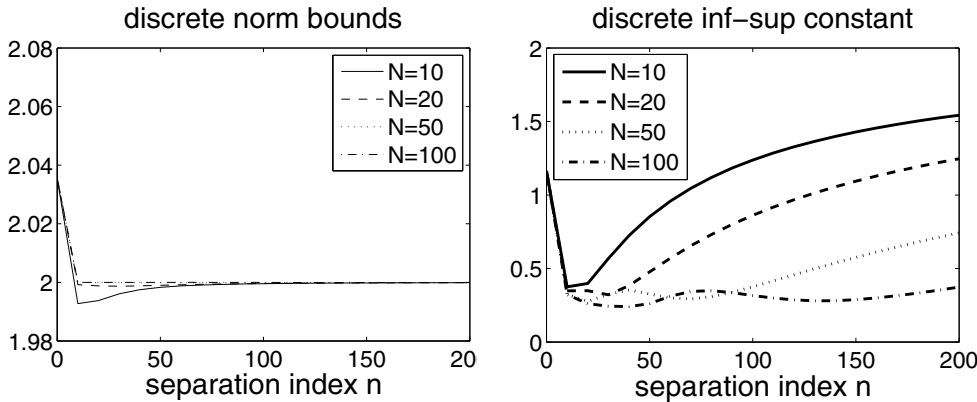


FIG. 5.1. Norms and *inf-sup* constants of the separated bilinear forms in (4.10) with respect to the norms defined in (5.2) for $\kappa = \kappa_0 = a = 1$ and $d = 2$.

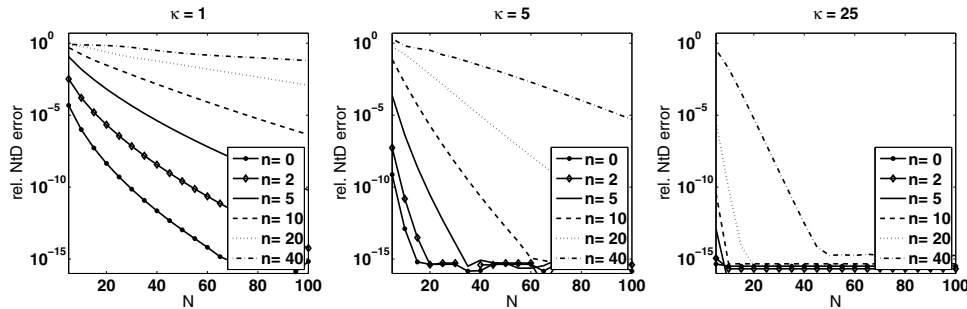


FIG. 5.2. Relative error of the Neumann-to-Dirichlet numbers for different Fourier modes n , different wave numbers κ , $a = 1$, $\kappa_0 = \kappa$, and $d = 2$.

mode n . The constants deteriorate as n grows, but improve as κ grows. Due to the stability shown in Figure 5.1, this must be due to the approximation properties of polynomial subspaces for the transformed Hankel functions.

The error for the full unseparated problem is mainly determined by the convergence behavior of the first Fourier modes as the size $|u_{0,n}|$ of the Fourier coefficients decays exponentially with n since u_0 is analytic. Figure 5.3 shows results for the scattering of plane incident waves with different wave numbers κ by a kite-shaped domain. As a reference solution we computed a pair of Cauchy data on Γ by a Nystrom integral equation method (cf. [2, section 3.5]). We used the reference Neumann data on spheres of radius 2 and 3.5 as initial data for the Hardy space method (HSM) and compared the Dirichlet data computed by the HSM to the reference Dirichlet data. As basis functions on Γ we used so-called hierarchic shape functions of high polynomial degrees (see [21, section 3.1.4]) such that the finite element error could be neglected. The error plot in Figure 5.3 clearly exhibits fast convergence with respect to N both for the wave number $\kappa = 5$ and $\kappa = 25$. As for other methods (e.g., PML or standard infinite elements), the error for a fixed number of degrees of freedom in the exterior domain grows smaller as the distance of the coupling boundary to the scatterer increases.

Since a crucial advantage of the HSM is the applicability of the method to resonance problems, we computed as a second example the resonances of a square with

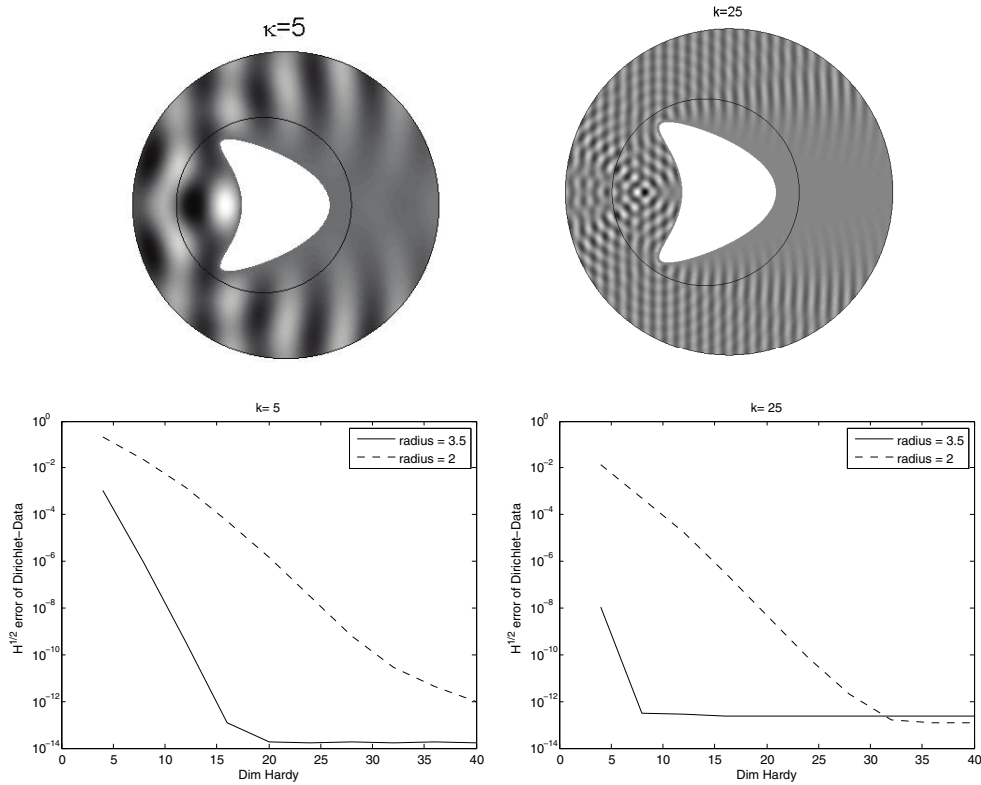


FIG. 5.3. $H^{1/2}(\Gamma)$ -error in the Dirichlet data for different wave numbers and radii as a function of the number N of degrees of freedom in the Hardy space $H^+(S^1)$.

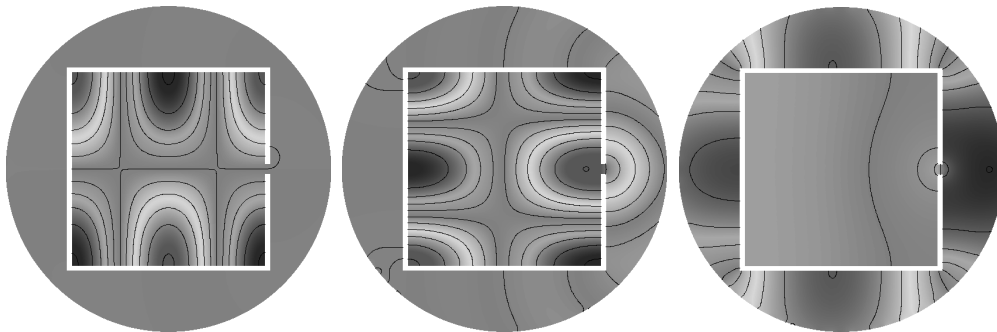


FIG. 5.4. Eigenfunctions of an open square.

a small opening. This was done using the finite element solver NGSOLVE, which is an add-on of the mesh generator NETGEN [20]. In Figure 5.4 three different eigenfunctions are plotted. Two of them correspond to the real valued eigenvalues of the Laplace operator in a closed square and the third to an exterior surface resonance, the location of which depends mainly on the circumference of the obstacle (cf. [25] and the references therein). In Figure 5.5 the exterior resonances of the sphere were computed as roots of the Hankel functions of the first kind. Additionally we used

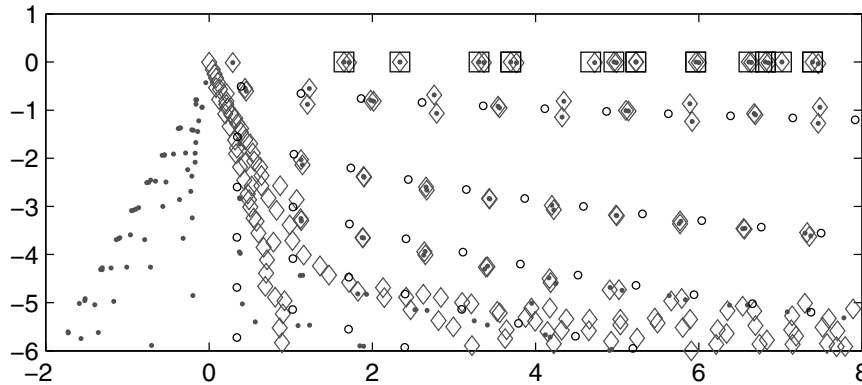


FIG. 5.5. *Resonances of an open square (●: HSM for open square; ◇: PML for open square; □: eigenvalues for closed square; ○: exterior resonances of a sphere with the same circumference as the square).*

PML (◇) as reference solution. The HSM resonances in the third quadrant and the PML resonances in the lower part of the plot are computational artifacts.

6. Conclusions. We have presented a new type of infinite elements based on the pole condition which are derived by transforming the exterior variational formulation of the Helmholtz equation to a Hardy space. They can be coupled with finite elements of arbitrary order in the interior domain and have simple, symmetric element matrices with a tensor product structure. The convergence with respect to the number of degrees of freedom in the transformed radial direction is superalgebraic. Moreover, they are particularly well suited for resonance problems since they preserve the eigenvalue structure. As opposed to other numerical realizations of the pole condition (cf. [8, 19]) it is not possible to recover the exterior solution directly by the HSM.

Let us compare Hardy space infinite elements with PML from a practical perspective: The PML method has the advantage of being easy to implement in standard software package, whereas the HSM requires the implementation of a new (in)finite element. The HSM has the advantage that it is a high order method which can easily be combined with low order codes. Moreover, the only tuning parameter in the HSM is κ_0 , and the rule $\kappa_0 \approx \kappa$ yields good results, whereas for PML at least the slope of the path in the complex plane, the width of the layer, and the polynomial degree have to be chosen. Our preliminary numerical experiments suggest that the HSM performs at least as good as PML, but for a definite conclusion more thorough numerical studies optimizing the various PML parameters will be necessary.

The HSM is not restricted to the situation studied in this paper, but can be extended to other differential equations and other coupling boundaries, which may be subject of future research.

Appendix. In this appendix we prove the lemmas needed for the transformation to the Hardy space.

LEMMA A.1. *Let $M \geq 0$ and $\kappa_0 \in \mathbb{C}$ be given constants with $\Re(\kappa_0) > 0$, and let $f, g : \mathbb{R}_+ \rightarrow \mathbb{C}$ be two measurable functions such that $f \exp(-M\bullet)$ and $g \exp(M\bullet)$ belong to $L^1([0, \infty)) \cap L^2([0, \infty))$. Moreover, assume that the Laplace transformed functions $\hat{f} := \mathcal{L}f$ and $\hat{g} := \mathcal{L}g$ have holomorphic extensions to the regions sketched*

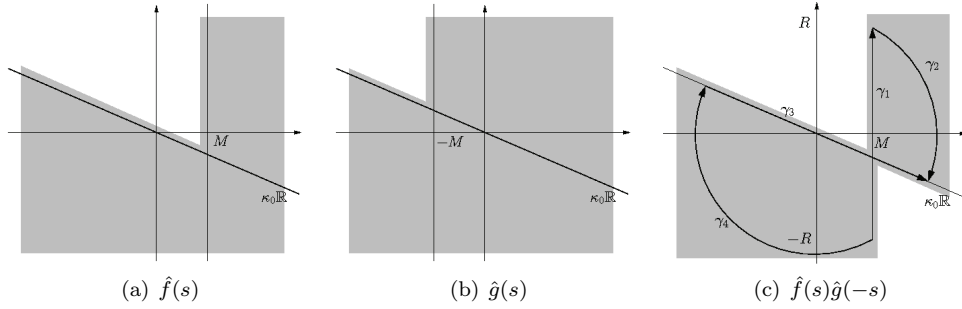


FIG. A.1. Regions to which the functions in Lemma A.1 have holomorphic extensions.

in Figure A.1 and that $|\hat{f}(s)s|, |\hat{g}(s)s|$ are uniformly bounded in these regions. Then

$$(A.1) \quad \int_0^\infty f(r)g(r)dr = -\frac{i}{2\pi} \int_{\kappa_0\mathbb{R}} \hat{f}(s)\hat{g}(-s)ds = \frac{-i\kappa_0}{\pi} \int_{S^1} F(z)G(\bar{z})|dz|,$$

with $F := \mathcal{M}_{\kappa_0}(\hat{f}|_{\kappa_0\mathbb{R}})$ and $G := \mathcal{M}_{\kappa_0}(\hat{g}|_{\kappa_0\mathbb{R}})$. (The orientation of the contour $\kappa_0\mathbb{R}$ is from left to right.)

Proof. We extend f, g by zero to $f^*, g^* : \mathbb{R} \rightarrow \mathbb{C}$ and write the integral as a Fourier transform $(\mathcal{F}\varphi)(s) := \int_{-\infty}^\infty e^{-ist}\varphi(t)dt$ evaluated at $s = 0$:

$$\int_0^\infty f(r)g(r)dr = \mathcal{F}\{f^*g^*\}(0) = \frac{1}{2\pi} \int_{-\infty}^\infty \mathcal{F}\{f^*e^{-M\bullet}\}(t)\mathcal{F}\{g^*e^{M\bullet}\}(-t)dt.$$

Here $\mathcal{F}\{f^*e^{-M\bullet}\}(t) = \hat{f}(it + M)$ and $\mathcal{F}\{g^*e^{M\bullet}\}(-t) = \hat{g}(-it + M)$ exist due to our assumptions. The first equation in (A.1) follows by Cauchy’s integral theorem for the closed contour $\gamma_1 + \gamma_2 - \gamma_3 + \gamma_4$ shown in Figure A.1(c), using the fact that the integrals over γ_2 and γ_4 vanish as $R \rightarrow \infty$ due to the assumed decay of \hat{f} and \hat{g} :

$$\int_0^\infty f(r)g(r)dr = -\frac{i}{2\pi} \lim_{R \rightarrow \infty} \int_{\gamma_1} \hat{f}(s)\hat{g}(-s)ds = -\frac{i}{2\pi} \lim_{R \rightarrow \infty} \int_{\gamma_3} \hat{f}(s)\hat{g}(-s)ds.$$

To prove the second equation we use the substitution of variables $s = \varphi_{\kappa_0}(z)$ and the identities $\varphi'_{\kappa_0}(z) = \frac{-2i\kappa_0}{(z-1)^2}$ and $-\varphi_{\kappa_0}(z) = \varphi_{\kappa_0}(\bar{z})$ for $z \in S^1$ to obtain

$$\begin{aligned} -\frac{i}{2\pi} \lim_{R \rightarrow \infty} \int_{\gamma_3} \hat{f}(s)\hat{g}(-s)ds &= \frac{-\kappa_0}{\pi} \int_{S^1, \circlearrowleft} \frac{\hat{f}(\varphi_{\kappa_0}(z))\hat{g}(-\varphi_{\kappa_0}(z))}{z-1} dz \\ &= \frac{-\kappa_0}{\pi} \int_{S^1, \circlearrowleft} F(z)G(\bar{z})\frac{\bar{z}-1}{z-1} dz. \end{aligned}$$

The symbol \circlearrowleft indicates clockwise orientation of the contour S^1 . Since $\frac{\bar{z}-1}{z-1} = \frac{1/z-1}{z-1} = -\frac{1}{z}$ for $z \in S^1$, and $dz = -iz|dz|$, we obtain the second equation in (A.1). \square

LEMMA A.2. Let $\kappa_0 \in \mathbb{C} \setminus \{0\}$, let E be an open subset of $\{k \in \mathbb{C} : \Re(k/\kappa_0) > 0\}$, and define $V_k(z) := \frac{k-\kappa_0}{(\kappa_0-k)z+(\kappa_0+k)}$ for $k \in E$. Then $\text{span}\{V_k : k \in E\}$ is dense in $H^+(S^1)$.

Proof. A straightforward computation shows that $(\mathcal{M}_1^{-1}V_k)(z) = \frac{i(k-\kappa_0)}{\kappa_0} \frac{1}{s-ik/\kappa_0}$, with the transform \mathcal{M}_1 defined in (2.7) (with $\kappa_0 = 1$, not the κ_0 given in the lemma).

Since $\mathcal{M}_1 : H^-(\mathbb{R}) \rightarrow H^+(S^1)$ is unitary, the statement is equivalent to the density of $Y := \text{span}\{1/(\bullet - ik/\kappa_0) : k \in E\}$ in $H^-(\mathbb{R})$. Assume that $f \in Y^\perp$, i.e., $\int_{\mathbb{R}} f(\bar{s})/(\bar{s} - ik/\kappa_0) d\bar{s} = 0$ for all $k \in E$. Then the holomorphic function

$$w(z) := \frac{1}{2\pi i} \int_{\mathbb{R}} \frac{f(\bar{s})}{\bar{s} - z} d\bar{s}, \quad z \in \mathbb{C}^-,$$

vanishes on $\{\overline{ik/\kappa_0} : k \in E\}$, which is an open subset of \mathbb{C}^- . Therefore, w vanishes identically in \mathbb{C}^- . Due to Definition 2.1 and (2.6), f are the boundary values of w on \mathbb{R} , and hence $f = 0$. This shows that $Y^\perp = \{0\}$, i.e., Y is dense in $H^-(\mathbb{R})$. \square

LEMMA A.3. Consider the set $X^\#$ and the inner product defined in (3.7), and let $\Re(\kappa), \Re(\kappa_0) > 0$.

- (1) $X^\#$ is a Hilbert space.
- (2) For each $v_{\text{int}} \in H^1(\Omega_{\text{int}})$ there exists $V \in H^+(S^1) \otimes L^2(\Gamma)$, such that $(v_{\text{int}}, V)^\top \in X^\#$.
- (3) There exists a dense subset $\tilde{X}^\# \subset X^\#$, such that for all $(v_{\text{int}}, V)^\top \in \tilde{X}^\#$ we have $v_{\text{int}} \in C^\infty(\overline{\Omega_{\text{int}}})$ and there exists a function $v_{\text{ext}} \in C^\infty([0, \infty) \times \Gamma)$ such that $(i\kappa_0)^{-1}(\mathcal{M}_{\kappa_0}^{-1} \mathcal{T}_- \otimes I)(v_0, V)^\top = \mathcal{L}|_{\kappa_0 \mathbb{R}} v_{\text{ext}}$ and the assumptions of Lemma A.1 are fulfilled with $f(r) := \exp(ikr)$ and $g(r) := v_{\text{ext}}(r, \hat{x})$ for all $\hat{x} \in \Gamma$ as well as with the first derivatives of v_{ext} .

Proof. (1) A straightforward argument using the closedness of the surface gradient ∇_x shows that $X^\#$ is complete.

(2) Let $v_{\text{int}} \in H^1(\Omega_{\text{in}})$ and define $v_0 := u_{\text{int}}|_\Gamma$. Since $v_0 \in H^{1/2}(\Gamma)$, the Fourier coefficients of v_0 satisfy $\sum_{n=0}^\infty (1 + \lambda_n)^{1/2} |v_{0,n}|^2 < \infty$. Here and in the following we use the notation of section 4. Define $V(z, \hat{x}) := \sum_{n=0}^\infty v_{0,n} V_{k_n}(z) \Phi_n(\hat{x})$ with a sequence (k_n) to be specified later. Since the functions V_k in Lemma A.2 satisfy $V_k(z) = (\frac{k/\kappa_0 + 1}{k/\kappa_0 - 1} - z)^{-1}$, it follows by radial symmetry that $\|V_k\|_{L^2(S^1)}^2 = \Xi(\frac{|k/\kappa_0 + 1|}{|k/\kappa_0 - 1|} - 1)$, with $\Xi(t) := \int_{S^1} |1 + t - z|^{-2} |dz|$ for $t > 0$. Setting $c := \int_{\pi/6}^{11\pi/6} |1 - \exp(i\theta)|^{-2} d\theta$, we obtain

$$\Xi(t) - c \leq \int_{-\pi/6}^{\pi/6} \frac{d\theta}{|1 + t - \exp(i\theta)|^2} \leq \int_{-\pi/6}^{\pi/6} \frac{d\theta}{t^2 + \theta^2/4} = \frac{4 \operatorname{atan}(\pi/12t)}{t} \leq \frac{2\pi}{t},$$

so $\Xi(t) = O(t^{-1})$ as $t \searrow 0$. From the identity $\mathcal{T}_-(1, V_k)^\top = \frac{\kappa_0}{\kappa_0 - k} V_k$ it follows that

$$\frac{|k - \kappa_0|^2}{|\kappa_0|^2} \left\| \mathcal{T}_- \begin{pmatrix} 1 \\ V_k \end{pmatrix} \right\|_{L^2(S^1)}^2 = \|V_k\|_{L^2(S^1)}^2 = \Xi\left(\frac{|k/\kappa_0 + 1|}{|k/\kappa_0 - 1|} - 1\right) = O(k)$$

as $\Re(k) \rightarrow \infty$. Now choose k_0 such that $\Re(k_0/\kappa_0) > 0$ and $k_n := k_0 + \sqrt{\lambda_n}$ for $n = 1, 2, \dots$. Then

$$\begin{aligned} \left\| \begin{pmatrix} v_{\text{int}} \\ V \end{pmatrix} \right\|_{X^\#}^2 - \|v_{\text{int}}\|_{H^1}^2 &= \sum_{n=0}^\infty |v_{0,n}|^2 \left\{ \|V_{k_n}\|_{L^2(S^1)}^2 + \lambda_n \left\| \mathcal{J} \mathcal{T}_- \begin{pmatrix} 1 \\ V_{k_n} \end{pmatrix} \right\|_{L^2(S^1)}^2 \right\} \\ &\leq C \sum_{n=0}^\infty |v_{0,n}|^2 |k_n| \left\{ 1 + \|\mathcal{J}\|^2 \frac{\lambda_n |\kappa_0|^2}{|k_n - \kappa_0|^2} \right\} \\ &\leq C \sum_{n=0}^\infty |v_{0,n}|^2 |k_n| \leq C \sum_{n=0}^\infty |v_{0,n}|^2 (1 + \lambda_n)^{1/2} < \infty, \end{aligned}$$

with a generic constant C . Hence, $(v_{\text{int}}, V)^\top \in X^\#$.

(3) With V as constructed above we have $v_{\text{ext}}(r, \hat{x}) = \sum_{n=0}^{\infty} v_{0,n} \exp(ik_n r) \Phi_n(\hat{x})$ (cf. (2.8), (2.9), (2.15)). If $v_{\text{int}} \in C^\infty(\overline{\Omega_{\text{int}}})$, then the Fourier coefficients $v_{0,n}$ decay superalgebraically, and the series together with its term-by-term derivatives converges uniformly on compact subsets. Moreover, $r \mapsto e^{i\kappa r} v_{\text{ext}}(r, \hat{x})$ decays exponentially if $\Im(k_n + \kappa) = \Im(k_0 + \kappa) > 0$. This can be arranged by an appropriate choice of k_0 . Hence, Lemma A.1 can be applied to $v_{\text{ext}}(r, \hat{x})$ and also to its first derivatives. Since everything above remains valid if k_n is chosen in a small ball around $k_0 + \sqrt{\lambda_n}$, the density property follows from Lemma A.2 and the density of $C^\infty(\overline{\Omega_{\text{int}}})$ in $H^1(\Omega_{\text{int}})$. \square

Acknowledgments. The idea to use a transform to the Hardy space $H^+(S^1)$ arose from discussions with Frank Schmidt and his group at Zuse Institut in Berlin within this project.

REFERENCES

- [1] A. BÖTTCHER AND B. SILBERMANN, *Analysis of Toeplitz Operators*, 2nd ed., Springer Monogr. Math., Springer-Verlag, Berlin, 2006.
- [2] D. COLTON AND R. KRESS, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Appl. Math. Sci. 93, Springer-Verlag, Berlin, 1998.
- [3] L. DEMKOWICZ AND K. GERDES, *Convergence of the infinite element methods for the Helmholtz equation in separable domains*, Numer. Math., 79 (1998), pp. 11–42.
- [4] L. DEMKOWICZ AND F. IHLENBURG, *Analysis of a coupled finite-infinite element method for exterior Helmholtz problems*, Numer. Math., 88 (2001), pp. 43–73.
- [5] P. L. DUREN, *Theory of H^p spaces*, Pure Appl. Math. 38, Academic Press, New York, 1970.
- [6] D. GIVOLI, *High-order nonreflecting boundary conditions without high-order derivatives*, J. Comput. Phys., 170 (2001), pp. 849–870.
- [7] S. HEIN, T. HOHAGE, W. KOCH, AND J. SCHÖBERL, *Acoustic resonances in high lift configuration*, J. Fluid Mech., 582 (2007), pp. 179–202.
- [8] T. HOHAGE, F. SCHMIDT, AND L. ZSCHIEDRICH, *A new method for the solution of scattering problems*, in Proceedings of the JEE'02 Symposium, B. Michielsen and F. Decavèle, eds., ONERA, Toulouse, France, 2002, pp. 251–256.
- [9] T. HOHAGE, F. SCHMIDT, AND L. ZSCHIEDRICH, *Solving time-harmonic scattering problems based on the pole condition I: Theory*, SIAM J. Math. Anal., 35 (2003), pp. 183–210.
- [10] T. HOHAGE, F. SCHMIDT, AND L. ZSCHIEDRICH, *Solving time-harmonic scattering problems based on the pole condition II: Convergence of the PML method*, SIAM J. Math. Anal., 35 (2003), pp. 547–560.
- [11] F. IHLENBURG, *Finite Element Analysis of Acoustic Scattering*, Appl. Math. Sci. 132, Springer-Verlag, New York, 1998.
- [12] R. KRESS, *Linear Integral Equations*, 2nd ed., Appl. Math. Sci. 82, Springer-Verlag, New York, 1999.
- [13] M. LENOIR, M. VULLIERME-LEDARD, AND C. HAZARD, *Variational formulations for the determination of resonant states in scattering problems*, SIAM J. Math. Anal., 23 (1992), pp. 579–608.
- [14] N. MOISEYEV, *Quantum theory of resonances: Calculating energies, width and cross-sections by complex scaling*, Phys. Rep., 302 (1998), pp. 211–293.
- [15] L. NANNEN, *Hardy-Raum Methoden zur numerischen Lösung von Streu- und Resonanzproblemen auf unbeschränkten Gebieten*, Ph.D. thesis, University of Göttingen, Tönning, 2008.
- [16] D. RUPRECHT, A. SCHÄDLE, F. SCHMIDT, AND L. ZSCHIEDRICH, *Transparent boundary conditions for time-dependent problems*, SIAM J. Sci. Comput., 30 (2008), pp. 2358–2385.
- [17] F. SCHMIDT, *A New Approach to Coupled Interior-Exterior Helmholtz-Type Problems: Theory and Algorithms*, habilitation, Freie Universität Berlin, 2002.
- [18] F. SCHMIDT AND P. DEUFLHARD, *Discrete transparent boundary conditions for the numerical solution of Fresnel's equation*, Comput. Math. Appl., 29 (1995), pp. 53–76.
- [19] F. SCHMIDT, T. HOHAGE, R. KLOSE, A. SCHÄDLE, AND L. ZSCHIEDRICH, *Pole condition: A numerical method for Helmholtz-type scattering problems with inhomogeneous exterior domain*, J. Comput. Appl. Math., 218 (2008), pp. 61–69.
- [20] J. SCHÖBERL, *Netgen—an advancing front 2d/3d-mesh generator based on abstract rules*, Comput. Visual. Sci., 1 (1997), pp. 41–52.

- [21] C. SCHWAB, *p- and hp-Finite Element Methods: Theory and Applications in Solid and Fluid Mechanics*, Numer. Math. Sci. Comput., The Clarendon Press, Oxford University Press, New York, 1998.
- [22] B. SIMON, *The theory of resonances for dilation analytic potentials and the foundations of time dependent perturbation theory*, Ann. Math., 97 (1973), pp. 247–274.
- [23] M. TAYLOR, *Partial Differential Equations: Qualitative Studies of Linear Equations*, Vol. 2, Springer-Verlag, New York, 1996.
- [24] L. ZSCHIEDRICH, R. KLOSE, A. SCHÄDLE, AND F. SCHMIDT, *A new finite element realization of the perfectly matched layer method for Helmholtz scattering problems on polygonal domains in two dimensions*, J. Comput. Appl. Math., 188 (2006), pp. 12–32.
- [25] M. ZWORSKI, *Resonances in physics and geometry*, Notices Amer. Math. Soc., 46 (1999), pp. 319–328.

ACCELERATED LINE-SEARCH AND TRUST-REGION METHODS*

P.-A. ABSIL[†] AND K. A. GALLIVAN[‡]

Abstract. In numerical optimization, line-search and trust-region methods are two important classes of descent schemes, with well-understood global convergence properties. We say that these methods are “accelerated” when the conventional iterate is replaced by any point that produces at least as much of a decrease in the cost function as a fixed fraction of the decrease produced by the conventional iterate. A detailed convergence analysis reveals that global convergence properties of line-search and trust-region methods still hold when the methods are accelerated. The analysis is performed in the general context of optimization on manifolds, of which optimization in \mathbb{R}^n is a particular case. This general convergence analysis sheds new light on the behavior of several existing algorithms.

Key words. line search, trust region, subspace acceleration, sequential subspace method, Riemannian manifold, optimization on manifolds, Riemannian optimization, Arnoldi, Jacobi–Davidson, locally optimal block preconditioned conjugate gradient (LOBPCG)

AMS subject classifications. 65B99, 65K05, 65J05, 65F15, 90C30

DOI. 10.1137/08072019X

1. Introduction. Let f be a real-valued function defined on a domain M , and let $\{x_k\}$ be a sequence of iterates generated as follows: for every k , some $x_{k+1/2} \in \mathcal{M}$ is generated (possibly implicitly) using a descent method that has global convergence to stationary points of f ; then x_{k+1} is chosen arbitrarily in the sublevel set $\{x \in M : f(x) \leq f(x_{k+1/2})\}$. We term “acceleration” the fact of choosing x_{k+1} rather than $x_{k+1/2}$ as the new iterate. The question addressed in this paper is whether the inclusion of the acceleration step preserves global convergence, i.e., whether $\{x_k\}$ converges to stationary points. We prove that the answer is positive for a wide class of methods.

The initial motivation for engaging in this general convergence analysis was to obtain a unifying convergence theory for several well-known eigenvalue algorithms. For example, the Jacobi–Davidson approach [38] is a popular technique for computing an eigenpair (eigenvalue and eigenvector) of a matrix A . It is an iterative method where the computation of the next iterate x_{k+1} from the current iterate x_k can be decomposed into two steps. The Jacobi step consists of solving (usually, approximately) a Newton-like equation to obtain an update vector η_k . Whereas in a classical Newton method the new iterate x_{k+1} is defined as $x_k + \eta_k$, the Davidson step uses the update vector η_k to expand a low-dimensional subspace and selects x_{k+1} as the “best” approximation (in some sense) of the sought eigenvector of A within the subspace. A key to the success of this approach is that the problem of computing x_{k+1} within the

*Received by the editors April 3, 2008; accepted for publication (in revised form) September 16, 2008; published electronically February 13, 2009. This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its authors. This work was supported in part by the US National Science Foundation under grant OCI0324944 and by the School of Computational Science of Florida State University.

<http://www.siam.org/journals/sinum/47-2/72019.html>

[†]Département d’ingénierie mathématique, Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgium (absil@inma.ucl.ac.be, <http://www.inma.ucl.ac.be/~absil>).

[‡]Department of Mathematics, Florida State University, Tallahassee, FL 32306-4510 (kgallivan@fsu.edu, <http://www.math.fsu.edu/~gallivan>).

subspace can be viewed as a reduced-dimensional eigenvalue problem, which can be solved efficiently when the dimension of the subspace is small. In certain situations, notably when x_{k+1} is chosen as the Ritz vector associated with an extreme Ritz value, the Davidson step can be interpreted as an acceleration step in the sense given above.

The reader primarily interested in eigenvalue algorithms can thus think of the purpose of this paper as formulating and analyzing this Jacobi–Davidson concept in the broad context of smooth optimization, i.e., the minimization of a smooth real-valued cost function over a smooth domain. The “Jacobi” step, instead of being restricted to (inexact) Newton methods, is expanded to cover general line-search and trust-region techniques. The “Davidson” step, or acceleration step, is also made more general: any iterate x_{k+1} is accepted provided that it produces a decrease in the cost function that is at least equal to a prescribed fraction of the decrease produced by the Jacobi update; minimizing the cost function over a subspace that contains the Jacobi update is just one way of achieving this goal.

This new analysis, while requiring only rather straightforward modifications of classical proofs found in the optimization literature, is very general and powerful. In particular, our global convergence analysis yields novel global convergence results for some well-known eigenvalue methods. Moreover, the proof technique is less *ad hoc* than the proofs and derivations usually found in the numerical linear algebra literature, since it simply relies on showing that the methods fit in the broad optimization framework.

What we mean by a smooth domain is a (smooth) manifold. Since the work of Gabay [17], there has been a growing interest for the optimization of smooth cost functions defined on manifolds. Major references include [22, 40, 34, 14, 3]. These differential-geometric techniques have found applications in various areas, such as signal processing, neural networks, computer vision, and econometrics (see, e.g., [6]). The concept of a manifold generalizes the notion of a smooth surface in a Euclidean space. It can thus be thought of as a natural setting for smooth optimization. Roughly speaking, a manifold is a set that is locally smoothly identified with open subsets of \mathbb{R}^d , where d is the dimension of the manifold. When the manifold is given to us as a subset of \mathbb{R}^n described by equality constraints, the differential-geometric approach can be viewed as an “informed way” of doing constrained optimization. The resulting algorithms have the property of being feasible (i.e., the iterates satisfy the constraints). In several important cases, however, the manifold is not available as a subset of \mathbb{R}^n but rather as a quotient space. Usually, the fundamental reason why the quotient structure appears is in order to take into account an inherent invariance in the problem. Smooth real-valued functions on quotient manifolds lend themselves as well to differential-geometric optimization techniques. We refer the reader to [6] for a recent overview of this area of research.

The reader solely interested in unconstrained optimization in \mathbb{R}^n should bear in mind that this situation is merely a particular case of the differential-geometric optimization framework considered here. We frequently mention in the text how unconstrained optimization in \mathbb{R}^n is subsumed.

Line-search and trust-region methods are two major techniques for unconstrained optimization in \mathbb{R}^n (see, e.g., [30]). Line-search techniques were proposed and analyzed on manifolds by several authors; see, e.g., [33, 34, 22, 40, 41, 6]. A trust-region framework, based on a systematic use of the concept of retraction, for optimizing functions defined on abstract Riemannian manifolds was proposed more recently [2, 6, 9]. Under reasonable conditions, which hold in particular for smooth cost functions on compact Riemannian manifolds, the trust-region method was shown to converge

to stationary points of the cost function (this is an extension of a well-known result for trust-region methods in \mathbb{R}^n). Furthermore, if the trust-region subproblems are (approximately) solved using a truncated conjugate gradient (CG) method with a well-chosen stopping criterion, then the method converges locally superlinearly to the nondegenerate local minima of the cost function. However, these favorable global and local convergence properties do not yield any information on the number of iterates needed, from a given initial point, to reach the local superlinear regime; and, indeed, problems can be crafted where this number of iterates is prohibitively high. The same can be said about the retraction-based line-search approach considered here. Acceleration techniques can be viewed as a way of improving the speed of convergence of those methods.

The acceleration idea is closely related to the subspace expansion concept in Davidson's method for the eigenvalue problem [12] (see also the more recent results in [38, 16, 15]), but the constraints we impose on the acceleration step are weaker than in Davidson-type algorithms. Our approach is also reminiscent of the *sequential subspace method* (SSM) of Hager [20, 25]. Whereas the latter uses subspace acceleration for the purpose of approximately solving trust-region subproblems, we use it as an outermost iteration wrapped around line-search and trust-region methods. The sequential subspace optimization algorithm of Narkiss and Zibulevsky [31] fits in the same framework.

The paper is organized as follows. In section 2, we define the concept of acceleration. The background in optimization on manifolds is recalled in section 3, with a particular emphasis on the case where the manifold is simply \mathbb{R}^n . We show global convergence properties for accelerated line-search (section 4) and trust-region (section 5) methods on Riemannian manifolds (of which the classical \mathbb{R}^n is a particular case). Section 6 gives a local convergence result. In section 7, these results are exploited to show global convergence properties of subspace acceleration methods. In particular, a conceptually simple accelerated conjugate gradient method, inspired from the work of Knyazev [26] for the symmetric eigenvalue problem, is proposed, and its global convergence is analyzed. Applications are mentioned in section 8, and conclusions are drawn in section 9.

A preliminary version of this paper appeared in the technical report [4], where the retraction-based line-search scheme and the acceleration concept were introduced.

2. Accelerated optimization methods. In this section, we define the concept of acceleration and briefly discuss acceleration strategies. An important acceleration technique, which consists of minimizing the cost function over an adequately chosen subspace, will be further discussed in section 7.

2.1. Definition. Let f be a cost function defined on an optimization domain M . Given a current iterate $x_k \in M$, line-search and trust-region methods generate a new iterate in M ; call it $x_{k+1/2}$. *Accelerating* the method consists of picking a new iterate $x_{k+1} \in M$ that produces at least as much of a decrease in the cost function as a fixed fraction of the decrease produced by $x_{k+1/2}$. In other words, x_{k+1} must satisfy

$$(1) \quad f(x_k) - f(x_{k+1}) \geq c (f(x_k) - f(x_{k+1/2}))$$

for some constant $c > 0$ independent of k .

2.2. Acceleration strategies. This relaxation on the choice of the new iterate introduces leeway for exploiting information that may improve the behavior of the method. For example, x_{k+1} can be determined by minimizing f over some well-

chosen subset of the domain M , built using information gained over the iterations. This idea is developed in section 7.

Moreover, a wide variety of “hybrid” optimization methods fit in the framework of (1). For example, let \mathcal{A} be a line-search or trust-region algorithm, and let \mathcal{B} be any descent method. If, for all k , $x_{k+1/2}$ is obtained from x_k by \mathcal{A} and x_{k+1} is obtained from $x_{k+1/2}$ by \mathcal{B} , then the sequence $\{x_k\}$ is generated by an accelerated line-search or trust-region algorithm. Likewise, for all k , let $x_{k+1/2}$ be obtained from x_k by \mathcal{A} , let $\tilde{x}_{k+1/2}$ be obtained from x_k by \mathcal{B} , and let $x_{k+1} = x_{k+1/2}$ if $f(x_{k+1/2}) \leq f(\tilde{x}_{k+1/2})$ and $x_{k+1} = \tilde{x}_{k+1/2}$ otherwise; then the sequence $\{x_k\}$ is again generated by an accelerated line-search or trust-region method.

Note that, until we reach section 7 on subspace acceleration, we make no assumption other than (1) on how x_{k+1} is chosen from $x_{k+1/2}$. We also point out that values of c in the open interval $(0, 1)$ do not correspond to acceleration in the intuitive sense of the term since $f(x_{k+1})$ is possibly greater than $f(x_{k+1/2})$. Actually, all practical accelerated methods considered in section 8 satisfy (1) with $c = 1$. However, we consider the general case $c > 0$ because it may be useful in some situations and the global convergence analysis for $c > 0$ is not significantly more complicated than for $c = 1$.

3. Preliminaries on Euclidean and Riemannian optimization. In this paper, we assume that the optimization domain M is a (finite-dimensional) Riemannian manifold. The particularization to unconstrained optimization in \mathbb{R}^n is made explicit whenever we feel that it improves readability.

Loosely speaking, a manifold is a topological set covered by mutually compatible local parameterizations. We refer, e.g., to [13, 6] for details. An important type of manifolds are those subsets of \mathbb{R}^n with a tangent space of constant dimension defined at each point (simple examples are spheres and \mathbb{R}^n itself). If the tangent spaces $T_x M$ are equipped with an inner product $\langle \cdot, \cdot \rangle_x$ that varies smoothly with x , then the manifold is called *Riemannian*. In this paper, we consider the problem of minimizing a real function f (the *cost function*) defined on a Riemannian manifold M .

Classical unconstrained optimization in \mathbb{R}^n corresponds to the case $M = \mathbb{R}^n$. The tangent space to \mathbb{R}^n at any point $x \in \mathbb{R}^n$ is canonically identified with \mathbb{R}^n itself: $T_x \mathbb{R}^n \simeq \mathbb{R}^n$. The canonical Riemannian structure on \mathbb{R}^n is its usual Euclidean vector space structure, where the inner product at $x \in \mathbb{R}^n$ defined by $\langle \xi, \zeta \rangle := \xi^T \zeta$ for all $\xi, \zeta \in T_x \mathbb{R}^n \simeq \mathbb{R}^n$.

The major problem to overcome is that manifolds are in general not flat so that the sum of two elements of M or their multiplication by scalars is not defined. A remedy advocated in [2] is to locally “flatten” the manifold onto the tangent space $T_{x_k} M$ at the current iterate x_k . This is done by means of a *retraction*, a concept proposed by Shub [32, 3].

DEFINITION 3.1 (retraction). *A retraction on a manifold M is a mapping R from the tangent bundle TM into M with the following properties (let R_x denote the restriction of R to $T_x M$):*

1. R is continuously differentiable.
2. $R_x(\xi) = x$ if and only if $\xi = 0_x$, the zero element of $T_x M$.
3. $DR_x(0_x) = \text{id}_{T_x M}$, where $DR_x(0_x)$ denotes the differential of $R_x(\cdot)$ at 0_x and $\text{id}_{T_x M}$ denotes the identity mapping on $T_x M$, with the canonical identification $T_{0_x}(T_x M) \simeq T_x M$.

Instead of the third condition, it is equivalent to require that $\frac{d}{dt} R_x(t\xi_x)|_{t=0} = \xi_x$ for all $\xi_x \in T_x M$.

We do not necessarily assume that R is defined on the whole tangent bundle TM , but we make the blanket assumption that its evaluation never fails in the algorithms. Note that the third condition implies that R_x is defined on a neighborhood of the origin of T_xM for all $x \in M$; this guarantees that, given $\eta_x \in T_xM$, $R_x(t\eta_x)$ is well-defined at least on some nonempty interval $-\epsilon < t < \epsilon$.

On a Riemannian manifold, it is always possible to choose the retraction R as the exponential mapping (which is defined everywhere when the manifold is *complete*). Using the exponential, however, may not be computationally sensible. The concept of retraction gives the possibility of choosing more efficient substitutes (see [3, 6]). Given a cost function f on a manifold M equipped with a retraction R , we define the lifted cost function at $x \in M$ as

$$(2) \quad \hat{f}_x : T_xM \rightarrow \mathbb{R} : \xi \mapsto f(R_x(\xi)).$$

When $M = \mathbb{R}^n$, the natural retraction is given by

$$(3) \quad R_x(\xi) := x + \xi,$$

and \hat{f}_x satisfies $\hat{f}_x(\xi) = f(x + \xi)$ for all $x \in \mathbb{R}^n$ and all $\xi \in T_x\mathbb{R}^n \simeq \mathbb{R}^n$.

Given a current iterate x_k on M , any line-search or trust-region method applied to \hat{f}_{x_k} produces a vector η_k in $T_{x_k}M$. In a line-search method, η_k is used as a search direction: a point is sought on the curve $t \mapsto R_{x_k}(t\eta_k)$ that satisfies some conditions on the cost function (e.g., a line minimizer or the Armijo condition). In a trust-region method [2], η_k defines a proposed new iterate $R_{x_k}(\eta_k)$. In both cases, the optimization method yields a proposed new iterate $x_{k+1/2}$ in M . Below we study the convergence properties of such schemes when they are accelerated in the sense of (1).

4. Accelerated line-search methods. Line-search methods (without acceleration) on a manifold M endowed with a retraction R are based on the update formula

$$x_{k+1} = R_{x_k}(t_k\eta_k),$$

where η_k is in $T_{x_k}M$ and t_k is a scalar. The two issues are to select the search direction η_k and then the step length t_k . To obtain global convergence results, some restrictions have to be imposed on η_k and t_k . The following definition concerning η_k is adapted from [10].

DEFINITION 4.1 (gradient-related). *A sequence $\{\eta_k\}$, $\eta_k \in T_{x_k}M$, is gradient-related if, for any subsequence $\{x_k\}_{k \in \mathcal{K}}$ in M that converges to a nonstationary point, the corresponding subsequence $\{\eta_k\}_{k \in \mathcal{K}}$ is bounded and satisfies*

$$\limsup_{k \rightarrow \infty, k \in \mathcal{K}} \langle \text{grad } f(x_k), \eta_k \rangle_{x_k} < 0.$$

When $M = \mathbb{R}^n$ with its canonical Euclidean structure, we have $\text{grad } f(x) = [\partial_1 f(x) \ \cdots \ \partial_n f(x)]^T$ and $\langle \text{grad } f(x), \eta \rangle = \eta^T \text{grad } f(x)$, where we used the canonical identification $T_x\mathbb{R}^n \simeq \mathbb{R}^n$. (One must bear in mind that when we use the identification $T_x\mathbb{R}^n \simeq \mathbb{R}^n$, we lose the information on the foot x of the tangent vector. In order to specify the foot, we say that $\{\eta_k\} \subseteq \mathbb{R}^n$ is gradient-related to $\{x_k\}$.)

There is a relation between the gradient relatedness of $\{\eta_k\}$ and the angle between η_k and the steepest-descent direction. Let $\angle(-\text{grad } f(x_k), \eta_k) = \arccos \frac{\langle -\text{grad } f(x_k), \eta_k \rangle_{x_k}}{\|\text{grad } f(x_k)\|_{x_k} \|\eta_k\|_{x_k}}$ denote the angle between η_k and the steepest-descent direction $-\text{grad } f(x_k)$. Let $\{\eta_k\}$ be such that $c_1 \leq \|\eta_k\|_{x_k} \leq c_2$ for some $0 < c_1 < c_2 < \infty$ and all k . Then the condition $\angle(-\text{grad } f(x_k), \eta_k) \geq \theta$ for some fixed $\theta > \frac{\pi}{2}$ and all k is sufficient for the

sequence $\{\eta_k\}$ to be gradient-related to $\{x_k\}$. In particular, assume that η_k is obtained by solving a linear system $\mathcal{A}_k \eta_k = -\text{grad } f(x_k)$, where \mathcal{A}_k is a linear symmetric positive-definite transformation of $T_{x_k}M$. Then $\cos \angle(-\text{grad } f(x_k), \eta_k) \geq \kappa^{-1}(\mathcal{A}_k)$, where $\kappa(\mathcal{A}_k)$ denotes the condition number of \mathcal{A}_k . Hence if the smallest eigenvalue of \mathcal{A}_k is bounded away from zero and the largest eigenvalue of \mathcal{A}_k is bounded, then $\{\eta_k\}$ is bounded away from zero and infinity and the condition number of \mathcal{A}_k is bounded, and thus $\{\eta_k\}$ is gradient-related. (Note that the condition that the linear operator $\mathcal{A} : T_x M \rightarrow T_x M$ is symmetric positive-definite means that $\langle u, \mathcal{A}v \rangle_x = \langle \mathcal{A}u, v \rangle_x$ for all $u, v \in T_x M$, and $\langle u, \mathcal{A}u \rangle_x > 0$ for all nonzero $u \in T_x M$. In the case of \mathbb{R}^n endowed with its canonical inner product, this corresponds to the classical definitions of symmetry and positive definiteness for the matrix representing the operator \mathcal{A} .)

The next definition, related to the choice of the step length t_k , relies on Armijo's backtracking procedure [7] (or see [10]) to find a point at which there is sufficient decrease of the cost function.

DEFINITION 4.2 (Armijo point). *Given a differentiable cost function f on a Riemannian manifold M with retraction R , a point $x \in M$, a nonzero descent vector $\eta \in T_x M$ (i.e., $\langle \text{grad } f(x), \eta \rangle_x < 0$), a scalar $\bar{\alpha} > 0$ such that the segment $[0, \bar{\alpha}] \eta \subseteq T_x M$ is included in the domain of R , and scalars $\beta \in (0, 1)$ and $\sigma \in (0, 1)$, the Armijo vector is defined as $\eta^A = \beta^m \bar{\alpha} \eta$, where m is the first nonnegative integer such that*

$$(4) \quad f(x) - f(R_x(\beta^m \bar{\alpha} \eta)) \geq -\sigma \langle \text{grad } f(x), \beta^m \bar{\alpha} \eta \rangle_x.$$

The Armijo point is $R_x(\beta^m \bar{\alpha} \eta) \in M$.

It can be shown, using the classical Armijo theory for the lifted cost function \hat{f}_x , that there is always an m such that (4) holds, and hence the definition is legitimate. A similar definition was proposed in [41] for the particular case where the retraction is the exponential mapping. When $M = \mathbb{R}^n$ with its canonical Euclidean structure, the definition reduces to the classical situation described, e.g., in [10].

We propose the following accelerated Riemannian line-search algorithm.

ALGORITHM 1. ACCELERATED LINE SEARCH (ALS)

Require: Riemannian manifold M ; continuously differentiable scalar field f on M ; retraction R from TM to M as in Definition 3.1; scalars $\bar{\alpha} > 0$, $c, \beta, \sigma \in (0, 1)$.

Input: Initial iterate $x_0 \in M$.

Output: Sequence of iterates $\{x_k\} \subseteq M$ and search directions $\{\eta_k\} \subseteq TM$.

- 1: **for** $k = 0, 1, 2, \dots$ **do**
- 2: Pick a descent vector η_k in $T_{x_k}M$ such that $t\eta_k$ is in the domain of R for all $t \in [0, \bar{\alpha}]$.
- 3: Select $x_{k+1} \in M$ such that

$$(5) \quad f(x_k) - f(x_{k+1}) \geq c (f(x_k) - f(R_{x_k}(\eta^A))),$$

where η^A is the Armijo vector (Definition 4.2 with $x := x_k$ and $\eta := \eta_k$).

- 4: **end for**
-

Observe that Algorithm 1, as well as most other algorithms in this paper, describes a *class* of numerical algorithms; one could call it an *algorithm model*. The purpose of this analysis paper is to give (strong) convergence results for (broad) classes of algorithms. For Algorithm 1, we have the following convergence result, whose proof closely follows [10, Proposition 1.2.1]. The result is, however, more general in three

aspects. (1) Even when the optimization domain is \mathbb{R}^n , the line search is not necessarily done on a straight line, because the choice of the retraction is not restricted to the natural retraction (3) in \mathbb{R}^n . (2) Even in the case of \mathbb{R}^n , points other than the Armijo point can be selected, as long as they satisfy the acceleration condition (5). (3) Finally, the optimization domain can be any Riemannian manifold.

THEOREM 4.3. *Let $\{x_k\}$ be an infinite sequence of iterates generated by Algorithm 1 (ALS), and assume that the generated sequence $\{\eta_k\}$ of search directions is gradient-related (Definition 4.1). Then every limit point of $\{x_k\}$ is a stationary point of f .*

Proof. The proof is by contradiction. Suppose that there is a subsequence $\{x_k\}_{k \in \mathcal{K}}$ converging to some x^* with $\text{grad } f(x^*) \neq 0$. Since $\{f(x_k)\}$ is nonincreasing, it follows that $\{f(x_k)\}$ converges to $f(x^*)$. Hence $f(x_k) - f(x_{k+1})$ goes to zero. By the construction of the algorithm,

$$f(x_k) - f(x_{k+1}) \geq -c\sigma\alpha_k \langle \text{grad } f(x_k), \eta_k \rangle_{x_k},$$

where $\alpha_k \eta_k$ is the Armijo vector. Since $\{\eta_k\}$ is gradient-related, it follows that $\{\alpha_k\}_{k \in \mathcal{K}} \rightarrow 0$. It follows that for all k greater than some \bar{k} , $\alpha_k < \bar{\alpha}$, which means that $\alpha_k = \beta^m \bar{\alpha}$ for some $m \geq 1$, which implies that the previously tried step size $\beta^{m-1} \bar{\alpha} = \alpha_k / \beta$ did not satisfy the Armijo condition. In other words,

$$f(x_k) - f(R_{x_k}(\alpha_k / \beta) \eta_k) < -\sigma(\alpha_k / \beta) \langle \text{grad } f(x_k), \eta_k \rangle_{x_k} \quad \forall k \in \mathcal{K}, k \geq \bar{k}.$$

Denoting

$$(6) \quad \tilde{\eta}_k = \frac{\eta_k}{\|\eta_k\|} \quad \text{and} \quad \tilde{\alpha}_k = \frac{\alpha_k \|\eta_k\|}{\beta},$$

the inequality above reads

$$\frac{\hat{f}_{x_k}(0) - \hat{f}_{x_k}(\tilde{\alpha}_k \tilde{\eta}_k)}{\tilde{\alpha}_k} < -\sigma \langle \text{grad } f(x_k), \eta_k \rangle_{x_k} \quad \forall k \in \mathcal{K}, k \geq \bar{k},$$

where \hat{f} is defined as in (2). The mean value theorem yields

$$(7) \quad -\langle \text{grad } \hat{f}_{x_k}(t \tilde{\eta}_k), \tilde{\eta}_k \rangle_{x_k} < -\sigma \langle \text{grad } f(x_k), \eta_k \rangle_{x_k} \quad \forall k \in \mathcal{K}, k \geq \bar{k},$$

where t is in the interval $[0, \tilde{\alpha}_k]$. Since $\{\alpha_k\}_{k \in \mathcal{K}} \rightarrow 0$ and since η_k is gradient-related, hence bounded, it follows that $\{\tilde{\alpha}_k\}_{k \in \mathcal{K}} \rightarrow 0$. Moreover, since $\tilde{\eta}_k$ has unit norm and its foot x_k converges on the index set \mathcal{K} , it follows that $\{\eta_k\}_{k \in \mathcal{K}}$ is included in some compact subset of the tangent bundle TM , and therefore there exists an index set $\tilde{\mathcal{K}} \subseteq \mathcal{K}$ such that $\{\tilde{\eta}_k\}_{k \in \tilde{\mathcal{K}}} \rightarrow \tilde{\eta}^*$ for some $\tilde{\eta}^* \in T_{x^*}M$ with $\|\tilde{\eta}^*\| = 1$. We now take the limit in (7) over $\tilde{\mathcal{K}}$. Since the Riemannian metric is continuous (by definition), $f \in C^1$, and $\text{grad } \hat{f}_{x_k}(0) = \text{grad } f(x_k)$ (because of point 3 in Definition 3.1, see [6, equation (4.4)]), we obtain

$$-\langle \text{grad } f(x^*), \tilde{\eta}^* \rangle_{x^*} \leq -\sigma \langle \text{grad } f(x^*), \tilde{\eta}^* \rangle_{x^*}.$$

Since $0 < \sigma < 1$, it follows that $\langle \text{grad } f(x^*), \tilde{\eta}^* \rangle_{x^*} \geq 0$. On the other hand, from the fact that $\{\eta_k\}$ is gradient-related, one obtains that $\langle \text{grad } f(x^*), \tilde{\eta}^* \rangle_{x^*} < 0$, a contradiction. \square

More can be said under compactness assumptions, using a standard topological argument. (The purpose of the compactness assumption is to ensure that every subsequence of $\{x_k\}$ has at least one limit point.)

COROLLARY 4.4. *Let $\{x_k\}$ be an infinite sequence of iterates generated by Algorithm 1 (ALS), and assume that the generated sequence $\{\eta_k\}$ of search directions is gradient-related (Definition 4.1). Assume that there is a compact set \mathcal{C} such that $\{x_k\} \subseteq \mathcal{C}$. (This assumption holds in particular when the sublevel set $\mathcal{L} = \{x \in M : f(x) \leq f(x_0)\}$ is compact: the iterates all belong to the sublevel set since f is nonincreasing. It also holds when M itself is compact.) Then $\lim_{k \rightarrow \infty} \|\text{grad } f(x_k)\| = 0$.*

Proof. The proof is by contradiction. Assume the contrary; i.e., there is a subsequence $\{x_k\}_{k \in \mathcal{K}}$ and $\epsilon > 0$ such that $\|\text{grad } f(x_k)\| > \epsilon$ for all $k \in \mathcal{K}$. Since $\{x_k\} \subseteq \mathcal{C}$, with \mathcal{C} compact, it follows that $\{x_k\}_{k \in \mathcal{K}}$ has a limit point x^* in \mathcal{C} (Bolzano–Weierstrass theorem). By continuity of $\text{grad } f$, one has $\|\text{grad } f(x^*)\| \geq \epsilon$, i.e., x^* is not stationary, a contradiction with Theorem 4.3. \square

5. Accelerated trust-region algorithm. We first briefly recall the basics of the Riemannian trust-region scheme (RTR) proposed in [2]. Let M be a Riemannian manifold with retraction R . Given a cost function $f : M \rightarrow \mathbb{R}$ and a current iterate $x_k \in M$, we use R_{x_k} to locally map the minimization problem for f on M into a minimization problem for the cost function \hat{f}_{x_k} defined as in (2). The Riemannian metric g turns $T_{x_k}M$ into a Euclidean space endowed with the inner product $g_{x_k}(\cdot, \cdot)$, which makes it possible to consider the following *trust-region subproblem* in the Euclidean space $T_{x_k}M$:

$$(8a) \quad \min_{\eta \in T_{x_k}M} m_{x_k}(\eta) \quad \text{subject to } \langle \eta, \eta \rangle_{x_k} \leq \Delta_k^2,$$

where

$$(8b) \quad m_{x_k}(\eta) \equiv f(x_k) + \langle \text{grad } f(x_k), \eta \rangle_{x_k} + \frac{1}{2} \langle \mathcal{H}_{x_k} \eta, \eta \rangle_{x_k},$$

Δ_k is the *trust-region radius*, and $\mathcal{H}_{x_k} : T_{x_k}M \rightarrow T_{x_k}M$ is some symmetric linear operator, i.e., $\langle \mathcal{H}_{x_k} \xi, \chi \rangle_{x_k} = \langle \xi, \mathcal{H}_{x_k} \chi \rangle_{x_k}$, $\xi, \chi \in T_{x_k}M$. Note that m_{x_k} need not be the exact quadratic Taylor expansion of \hat{f}_{x_k} about zero, since \mathcal{H}_k is freely chosen.

Next, an approximate solution η_k to the trust-region subproblem (8) is produced. For the purpose of obtaining global convergence results, the η_k need not be the exact solution provided it produces a sufficient decrease of the model, as specified later. The decision to accept or not the candidate $R_{x_k}(\eta_k)$ and to update the trust-region radius is based on the quotient

$$(9) \quad \rho_k = \frac{f(x_k) - f(R_{x_k}(\eta_k))}{m_{x_k}(0_{x_k}) - m_{x_k}(\eta_k)} = \frac{\hat{f}_{x_k}(0_{x_k}) - \hat{f}_{x_k}(\eta_k)}{m_{x_k}(0_{x_k}) - m_{x_k}(\eta_k)}$$

measuring the agreement between the model decrease and the function decrease at the proposed iterate.

The following algorithm differs from the RTR algorithm of [2] only below the line “if $\rho_k > \rho'$.” (The specific rules for accepting the proposed new iterate and updating the trust-region radius come from [30]; they form a particular instance of the rules given in [11].)

Next, we study the global convergence of Algorithm 2. We show that, under some assumptions on the cost function, the model and the quality of η_k , it holds

ALGORITHM 2. ACCELERATED TRUST REGION (ATR)

Require: Riemannian manifold M ; scalar field f on M ; retraction R from TM to M as in Definition 3.1. Parameters: $\bar{\Delta} > 0$, $\Delta_0 \in (0, \bar{\Delta})$, and $\rho' \in [0, \frac{1}{4})$, $c \in (0, 1)$, $c_1 > 0$.

Input: Initial iterate $x_0 \in M$.

Output: Sequence of iterates $\{x_k\}$.

```

1: for  $k = 0, 1, 2, \dots$  do
2:   Obtain  $\eta_k$  by (approximately) solving (8).
3:   Evaluate  $\rho_k$  from (9);
4:   if  $\rho_k < \frac{1}{4}$  then
5:      $\Delta_{k+1} = \frac{1}{4}\Delta_k$ 
6:   else if  $\rho_k > \frac{3}{4}$  and  $\|\eta_k\| = \Delta_k$  then
7:      $\Delta_{k+1} = \min(2\Delta_k, \bar{\Delta})$ 
8:   else
9:      $\Delta_{k+1} = \Delta_k$ ;
10:  end if
11:  if  $\rho_k > \rho'$  then
12:    Select  $x_{k+1} \in M$  such that
      (10)           $f(x_k) - f(x_{k+1}) \geq c (f(x_k) - f(R_{x_k}(\eta_k)))$ ;
13:  else
14:    Select  $x_{k+1} \in M$  such that
      (11)           $f(x_k) - f(x_{k+1}) \geq 0$ ;
15:  end if
16: end for

```

that the gradient of the cost function goes to zero at least on a subsequence of $\{x_k\}$. This is done by slightly modifying the corresponding development given in [2] to take acceleration into account.

We need the following definition.

DEFINITION 5.1 (radially L - C^1 function). *Let $\hat{f} : TM \rightarrow \mathbb{R}$ be as in (2). We say that \hat{f} is radially Lipschitz continuously differentiable if there exist reals $\beta_{RL} > 0$ and $\delta_{RL} > 0$ such that, for all $x \in M$, for all $\xi \in TM$ with $\|\xi\| = 1$, and for all $t < \delta_{RL}$, it holds that*

$$(12) \quad \left| \frac{d}{d\tau} \hat{f}_x(\tau\xi)|_{\tau=t} - \frac{d}{d\tau} \hat{f}_x(\tau\xi)|_{\tau=0} \right| \leq \beta_{RL} t.$$

For the purpose of Algorithm 2, which is a descent algorithm, this condition needs only to be imposed in the level set

$$(13) \quad \{x \in M : f(x) \leq f(x_0)\}.$$

We also require the approximate solution η_k of the trust-region subproblem (8) to produce a sufficient decrease in the model. More precisely, η_k must produce at least as much of a decrease in the model function as a fixed fraction of the so-called Cauchy decrease; see [30, section 4.3]. Since the trust-region subproblem (8) is expressed on

a Euclidean space, the definition of the Cauchy point is adapted from \mathbb{R}^n without difficulty, and the bound

$$(14) \quad m_k(0) - m_k(\eta_k) \geq c_1 \|\text{grad}f(x_k)\| \min \left(\Delta_k, \frac{\|\text{grad}f(x_k)\|}{\|\mathcal{H}_k\|} \right),$$

for some constant $c_1 > 0$, is readily obtained from the \mathbb{R}^n case, where $\|\mathcal{H}_k\|$ is defined as

$$(15) \quad \|\mathcal{H}_k\| := \sup\{\|\mathcal{H}_k\zeta\| : \zeta \in T_{x_k}M, \|\zeta\| = 1\}.$$

In particular, the Steihaug–Toint truncated CG method (see, e.g., [37, 30, 11]) satisfies this bound (with $c_1 = \frac{1}{2}$, see [30, Lemma 4.5]) since it first computes the Cauchy point and then attempts to improve the model decrease.

With these things in place, we can state and prove the following global convergence result.

THEOREM 5.2. *Let $\{x_k\}$ be a sequence of iterates generated by Algorithm 2 (ATR) with $\rho' \in [0, \frac{1}{4})$. Suppose that f is C^1 and bounded below on the level set (13), that \hat{f} is radially L - C^1 (Definition 5.1), and that $\|\mathcal{H}_k\| \leq \beta$ for some constant β . Further suppose that all approximate solutions η_k of (8) satisfy the Cauchy decrease inequality (14) for some positive constant c_1 . We then have*

$$\liminf_{k \rightarrow \infty} \|\text{grad}f(x_k)\| = 0.$$

Proof. Here is a brief outline of the proof for the reader’s convenience. We will assume for contradiction that the norm of the gradient is bounded away from zero. Then a key to reaching a contradiction is that the trust-region does not shrink to zero (21). This is ensured by showing that ρ_k is greater than $\frac{1}{2}$ whenever Δ_k is smaller than a global value (20). This result itself is obtained by imposing that the discrepancy between the model and the cost function is uniformly quadratic (17) and that the denominator of ρ_k is bounded below by a ramp function of Δ_k (14).

We now turn to the detailed proof. First, we perform some manipulation of ρ_k from (9):

$$(16) \quad \begin{aligned} |\rho_k - 1| &= \left| \frac{(f(x_k) - \hat{f}_{x_k}(\eta_k)) - (m_k(0) - m_k(\eta_k))}{m_k(0) - m_k(\eta_k)} \right| \\ &= \left| \frac{m_k(\eta_k) - \hat{f}_{x_k}(\eta_k)}{m_k(0) - m_k(\eta_k)} \right|. \end{aligned}$$

Direct manipulations on the function $t \mapsto \hat{f}_{x_k}(t \frac{\eta_k}{\|\eta_k\|})$ yield

$$\begin{aligned} \hat{f}_{x_k}(\eta_k) &= \hat{f}_{x_k}(0_{x_k}) + \|\eta_k\| \frac{d}{d\tau} \hat{f}_{x_k} \left(\tau \frac{\eta_k}{\|\eta_k\|} \right) \Big|_{\tau=0} \\ &\quad + \int_0^{\|\eta_k\|} \left(\frac{d}{d\tau} \hat{f}_{x_k} \left(\tau \frac{\eta_k}{\|\eta_k\|} \right) \Big|_{\tau=t} - \frac{d}{d\tau} \hat{f}_{x_k} \left(\tau \frac{\eta_k}{\|\eta_k\|} \right) \Big|_{\tau=0} \right) dt \\ &= f(x_k) + \langle \text{grad}f(x_k), \eta_k \rangle_{x_k} + \epsilon', \end{aligned}$$

where $|\epsilon'| < \int_0^{\|\eta_k\|} \beta_{RL} t \, dt = \frac{1}{2} \beta_{RL} \|\eta_k\|^2$ whenever $\|\eta_k\| < \delta_{RL}$, and β_{RL} and δ_{RL} are the constants in the radially L - C^1 property (12). Therefore, it follows from the

definition (8b) of m_k that

$$(17) \quad \begin{aligned} \left| m_k(\eta_k) - \hat{f}_{x_k}(\eta_k) \right| &= \left| \frac{1}{2} \langle \mathcal{H}_{x_k} \eta_k, \eta_k \rangle_{x_k} - \epsilon' \right| \\ &\leq \frac{1}{2} \beta \|\eta_k\|^2 + \frac{1}{2} \beta_{RL} \|\eta_k\|^2 \leq \beta' \|\eta_k\|^2 \end{aligned}$$

whenever $\|\eta_k\| < \delta_{RL}$, where $\beta' = \max(\beta, \beta_{RL})$. Assume for the purpose of contradiction that $\liminf_{k \rightarrow \infty} \|\text{grad} f(x_k)\| \neq 0$; that is, assume that there exist $\epsilon > 0$ and a positive index K such that

$$(18) \quad \|\text{grad} f(x_k)\| \geq \epsilon \quad \forall k \geq K.$$

From (14) for $k \geq K$, we have

$$(19) \quad m_k(0) - m_k(\eta_k) \geq c_1 \|\text{grad} f(x_k)\| \min \left(\Delta_k, \frac{\|\text{grad} f(x_k)\|}{\|\mathcal{H}_k\|} \right) \geq c_1 \epsilon \min \left(\Delta_k, \frac{\epsilon}{\beta'} \right).$$

Substituting (17) and (19) into (16), we have that

$$(20) \quad |\rho_k - 1| \leq \frac{\beta' \|\eta_k\|^2}{c_1 \epsilon \min \left(\Delta_k, \frac{\epsilon}{\beta'} \right)} \leq \frac{\beta' \Delta_k^2}{c_1 \epsilon \min \left(\Delta_k, \frac{\epsilon}{\beta'} \right)}$$

whenever $\|\eta_k\| < \delta_{RL}$. We can choose a value of $\hat{\Delta}$ that allows us to bound the right-hand side of the inequality (20) when $\Delta_k \leq \hat{\Delta}$. Choose $\hat{\Delta}$ as follows:

$$\hat{\Delta} \leq \min \left(\frac{c_1 \epsilon}{2\beta'}, \frac{\epsilon}{\beta'}, \delta_{RL} \right).$$

This gives us $\min(\Delta_k, \frac{\epsilon}{\beta'}) = \Delta_k$. We can now write (20) as follows:

$$|\rho_k - 1| \leq \frac{\beta' \hat{\Delta} \Delta_k}{c_1 \epsilon \min \left(\Delta_k, \frac{\epsilon}{\beta'} \right)} \leq \frac{\Delta_k}{2 \min \left(\Delta_k, \frac{\epsilon}{\beta'} \right)} = \frac{1}{2}.$$

Therefore, $\rho_k \geq \frac{1}{2} > \frac{1}{4}$ whenever $\Delta_k \leq \hat{\Delta}$ so that, by the workings of Algorithm 2, it follows that $\Delta_{k+1} \geq \Delta_k$ whenever $\Delta_k \leq \hat{\Delta}$. It follows that a reduction of Δ_k (by a factor of $\frac{1}{4}$) can occur in Algorithm 2 only when $\Delta_k > \hat{\Delta}$. Therefore, we conclude that

$$(21) \quad \Delta_k \geq \min \left(\Delta_K, \hat{\Delta}/4 \right) \quad \forall k \geq K.$$

Consequently, $\rho_k \geq \frac{1}{4}$ must hold infinitely many times (otherwise $\{\Delta_k\}$ would go to zero by the workings of the algorithm). So there exists an infinite subsequence \mathcal{K} such that $\rho_k \geq \frac{1}{4} > \rho'$ for $k \in \mathcal{K}$. If $k \in \mathcal{K}$ and $k \geq K$, it follows from (19) and (10) that

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq c \left(f_{x_k} - \hat{f}_{x_k}(\eta_k) \right) \\ &\geq c \frac{1}{4} (m_k(0) - m_k(\eta_k)) \\ &\geq c \frac{1}{4} c_1 \epsilon \min \left(\Delta_k, \frac{\epsilon}{\beta'} \right) \\ &\geq c \frac{1}{4} c_1 \epsilon \min \left(\Delta_K, \frac{\hat{\Delta}}{4}, \frac{\epsilon}{\beta'} \right). \end{aligned}$$

Since, moreover, $f(x_k) - f(x_{k+1}) \geq 0$ for all $k \notin \mathcal{K}$, it follows that $f(x_k) \rightarrow -\infty$, a contradiction since f is bounded below on the level set containing $\{x_k\}$. \square

The convergence result of Theorem 5.2 is essentially identical to the corresponding result for the non-accelerated Riemannian trust-region method (see [2] or [6]), which itself is a natural generalization of a convergence result of the classical (non-accelerated) trust-region method in \mathbb{R}^n . In the classical convergence theory of trust-region methods in \mathbb{R}^n (see, e.g., [30, 11]), this result is followed by another theorem stating that, under further assumptions, $\lim_{k \rightarrow \infty} \|\text{grad } f(x_k)\| = 0$; i.e., the gradient of the cost function goes to zero on the *whole* sequence of iterates. This result also has a natural generalization for the non-accelerated Riemannian trust-region method (see [2, Theorem 4.4] or [6, Theorem 7.4.4]). It is an open question whether this result extends verbatim to the accelerated case. At least we can say that the proof cannot be adapted in a simple way: the condition that there exist $\mu > 0$ and $\delta_\mu > 0$ such that

$$(22) \quad \|\xi\| \geq \mu \text{dist}(x, R_x(\xi)) \quad \text{for all } x \in \mathcal{M}, \text{ for all } \xi \in T_x \mathcal{M}, \|\xi\| \leq \delta_\mu,$$

no longer implies that $\|\eta_k\| \geq \mu \text{dist}(x_k, x_{k+1})$ when acceleration comes into play. A simple fix is to require that there exists $\mu > 0$ such that the iterates satisfy

$$(23) \quad \|\eta_k\| \geq \mu \text{dist}(x_k, x_{k+1}) \quad \text{for all } k.$$

We then obtain the following result. (We refer to [2, 6] for the concept of Lipschitz continuous differentiability of f on the Riemannian manifold M ; the definition reduces to the classical one when the manifold is \mathbb{R}^n . The extension of the proof of [6, Theorem 7.4.4] to a proof of Theorem 5.3 is left to the reader.)

THEOREM 5.3. *Let $\{x_k\}$ be a sequence of iterates generated by Algorithm 2 (ATR). Suppose that all of the assumptions of Theorem 5.2 are satisfied. Further suppose that $\rho' \in (0, \frac{1}{4})$, that f is Lipschitz continuously differentiable, and that (23) is satisfied for some $\mu > 0$. It then follows that*

$$\lim_{k \rightarrow \infty} \text{grad } f(x_k) = 0.$$

6. Local convergence. We now briefly comment on how accelerating an optimization method may affect its order of convergence. Consider an algorithm that converges locally with order q to a local minimum v of the cost function f ; that is,

$$\text{dist}(x_+, v) \leq c_0 (\text{dist}(x, v))^q$$

for some $c_0 > 0$ and all x in some neighborhood of v , where x_+ stands for the next iterate computed from the current iterate x . If the algorithm is accelerated in the sense of (1), then local convergence to v is no longer guaranteed without further hypotheses; i.e., the algorithm may converge to stationary points other than v . However, for sequences of iterates of the accelerated algorithm that converge to v , we have the following result.

PROPOSITION 6.1. *Let v be a nondegenerate minimizer of $f \in C^3(M)$, where M is a Riemannian manifold. Consider a descent algorithm that converges locally with order $q > 1$ to v . If $\{x_k\}$ is a sequence of iterates of an accelerated version of the descent algorithm, in the sense of (1) with $c = 1$, and $\{x_k\}$ converges to v , then it does so with order q .*

Proof. We work in a coordinate system around v . Abusing notation, we use the same symbols for points of M and their coordinate representations. There is a neighborhood \mathcal{U} of v such that, for all $x \in \mathcal{U}$, we have

$$\frac{1}{2} \lambda_m \|x - v\|^2 \leq f(x) - f(v) \leq 2 \lambda_M \|x - v\|^2,$$

where $\lambda_M \geq \lambda_m > 0$ denote the largest and smallest eigenvalues, respectively, of the Hessian of f at v (they are positive since v is a nondegenerate minimizer). Since $c = 1$, it follows from (1) that $f(x_{k+1}) \leq f(x_{k+1/2})$. Moreover, by the equivalence of norms, there is a neighborhood \mathcal{U}_1 of v and constants c_1 and c_2 such that, for all $x \in \mathcal{U}_1$, $\frac{1}{c_1} \text{dist}(x, v) \leq \|x - v\| \leq c_2 \text{dist}(x, v)$. Since the original descent algorithm converges locally with order q to v , there exists a nonempty open ball $B_\epsilon(v)$ such that, whenever $x_k \in B_\epsilon(v)$, it holds that $x_{k+1/2} \in B_\epsilon(v)$ with $\text{dist}(x_{k+1/2}, v) \leq c_0 (\text{dist}(x_k, v))^q$. Moreover, ϵ can be chosen such that $B_\epsilon(v) \subseteq \mathcal{U} \cap \mathcal{U}_1$. Since $\{x_k\}$ converges to v , there is K such that, for all $k > K$, x_k belongs to $B_\epsilon(v)$. We have, for all $k > K$,

$$\begin{aligned} (\text{dist}(x_{k+1}, v))^2 &\leq c_1^2 \|x_{k+1} - v\|^2 \\ &\leq c_1^2 \frac{2}{\lambda_m} (f(x_{k+1}) - f(v)) \leq c_1^2 \frac{2}{\lambda_m} (f(x_{k+1/2}) - f(v)) \\ &\leq c_1^2 \frac{4}{\lambda_m} \lambda_M \|x_{k+1/2} - v\|^2 \leq c_1^2 \frac{4}{\lambda_m} \lambda_M c_0^2 (\text{dist}(x_{k+1/2}, v))^2 \\ &\leq c_1^2 \frac{4}{\lambda_m} \lambda_M c_0^2 c_2^2 (\text{dist}(x_k, v))^{2q}. \quad \square \end{aligned}$$

7. Sequential subspace optimization methods. We consider sequential subspace optimization methods in the form given in Algorithm 3 below. It generalizes the sequential subspace optimization (SESOP) algorithm of [31] to Riemannian manifolds.

ALGORITHM 3. SESOP

Require: Riemannian manifold M ; continuously differentiable scalar field f on M ; retraction R from TM to M as in Definition 3.1.

Input: Initial iterate $x_0 \in M$.

Output: Sequence of iterates $\{x_k\} \subseteq M$

- 1: **for** $k = 0, 1, 2, \dots$ **do**
 - 2: Select a subspace $\mathcal{S}_k \subseteq T_{x_k}M$.
 - 3: Find $\xi_k = \arg \min_{\xi \in \mathcal{S}_k} f(R_{x_k}(\xi))$.
 - 4: Set $x_{k+1} = R_{x_k}(\xi_k)$.
 - 5: **end for**
-

If \mathcal{S}_k is chosen in step 2 such that \mathcal{S}_k contains η_k , where η_k is as in Algorithm 1 (ALS) (resp., Algorithm 2 (ATR)), then Algorithm SESOP becomes an instance of Algorithm 1 (resp., Algorithm 2), with $c = 1$. The SESOP framework thus provides a strategy for accelerating line-search and trust-region methods.

When $M = \mathbb{R}^n$ with its natural retraction, Algorithm 3 becomes Algorithm 4 below, which can be found in [31] in an almost identical formulation. Observe that

ALGORITHM 4. \mathbb{R}^n -SESOP

Require: Continuously differentiable scalar field f on \mathbb{R}^n .

Input: Initial iterate $x_0 \in \mathbb{R}^n$.

Output: Sequence of iterates $\{x_k\} \subseteq \mathbb{R}^n$

- 1: **for** $k = 0, 1, 2, \dots$ **do**
 - 2: Select a real matrix W_k with n rows.
 - 3: Find $y^* = \arg \min_y f(x + W_k y)$.
 - 4: Set $x_{k+1} = x_k + W_k y^*$.
 - 5: **end for**
-

if $x_k \in \text{col}(W_k)$, where $\text{col}(W)$ denotes the subspace spanned by the columns of W , then x_{k+1} admits the expression

$$(24) \quad x_{k+1} = \arg \min_{x \in \text{col}(W_k)} f(x).$$

DEFINITION 7.1 (gradient-related sequence of subspaces). *A sequence $\{\mathcal{S}_k\}$ of subspaces of $T_{x_k}M$ is gradient-related if there exists a gradient-related sequence $\{\eta_k\}$ such that $\eta_k \in \mathcal{S}_k$ for all k ; equivalently, for any subsequence $\{x_k\}_{k \in K}$ that converges to a nonstationary point, we have*

$$\limsup_{k \rightarrow \infty, k \in K} \left(\inf_{\eta \in \mathcal{S}_k, \|\eta\|=1} \langle \text{grad } f(x_k), \eta \rangle \right) < 0.$$

When $M = \mathbb{R}^n$, the condition that \mathcal{S}_k be a subspace of $T_{x_k}M$ reduces to \mathcal{S}_k being a subspace of \mathbb{R}^n (in view of the canonical identification $T_x\mathbb{R}^n \simeq \mathbb{R}^n$).

PROPOSITION 7.2. *Let $\{x_k\}$ be an infinite sequence of iterates generated by Algorithm 3 (SESOP). Assume that the sequence $\{\mathcal{S}_k\}$ produced by Algorithm 3 is gradient-related (Definition 7.1). Then every limit point of $\{x_k\}$ is a stationary point of f . Assume further that $\{x_k\}$ is included in some compact set \mathcal{C} . Then $\lim_{k \rightarrow \infty} \|\text{grad } f(x_k)\| = 0$.*

Proof. The proof is a direct consequence of the convergence analysis of Algorithm 1 (ALS). \square

We now discuss a detailed procedure for selecting \mathcal{S}_k in Algorithm 3 (SESOP). It generalizes an idea in [26], which can be traced back to [39]. We denote by $P_\gamma^{t \leftarrow t_0} \zeta$ the vector of $T_{\gamma(t)}M$ obtained by parallel transporting a vector $\zeta \in T_{\gamma(t_0)}M$ along a curve γ . We refer, e.g., to [13, 6] for details on parallel translation. In \mathbb{R}^n , the natural parallel translation is simply given by $P_\gamma^{t \leftarrow t_0} \zeta = \zeta$ (where the ζ on the left-hand side is viewed as an element of $T_{\gamma(t)}M$ and the ζ on the right-hand side is viewed as an element of $T_{\gamma(t_0)}M$).

The name *conjugate gradient* is justified by the following property. Let M be the Euclidean space \mathbb{R}^n with retraction $R_x(\xi) := x + \xi$. Let f be given by $f(x) = \frac{1}{2}x^T A x$, where A is a symmetric positive-definite matrix. Then Algorithm 5 reduces to the classical linear CG method. This result is a consequence of the minimizing properties of the CG method. Again in the Euclidean case, but for general cost functions, Algorithm 5 can be viewed as a “locally optimal” nonlinear CG method: instead of computing a search direction ξ_k as a correction of $-\text{grad } f(x_k)$ along ξ_{k-1} (as is done in classical CG methods), the vector ξ_k is computed as a minimizer over the space spanned by $\{-\text{grad } f(x_k), \xi_{k-1}\}$. For the general Riemannian case, assuming that the retraction is chosen as the Riemannian exponential, Algorithm 5 can be thought of as a locally optimal version of the Riemannian CG algorithms proposed by Smith [34] (see also [14]).

By construction, the sequence $\{\mathcal{S}_k\}$ in Algorithm 5 is gradient-related. The following result thus follows from Proposition 7.2.

PROPOSITION 7.3. *Let $\{x_k\}$ be an infinite sequence of iterates generated by Algorithm 5. Then every limit point of $\{x_k\}$ is a stationary point of f . Assume further that $\{x_k\} \subseteq \mathcal{C}$ for some compact set \mathcal{C} . Then $\lim_{k \rightarrow \infty} \|\text{grad } f(x_k)\| = 0$.*

This result still holds if the parallel transport in Algorithm 5 is replaced by any *vector transport* as defined in [6]; indeed, the sequence $\{\mathcal{S}_k\}$ is still gradient-related by construction. Moreover, we point out that since Algorithm 5 is based on CG, it tends to display fast local convergence.

ALGORITHM 5. ACCELERATED CONJUGATE GRADIENT (ACG)

Require: Riemannian manifold M ; continuously differentiable scalar field f on M ; retraction R from TM to M as in Definition 3.1.**Input:** Initial iterate $x_0 \in M$.**Output:** Sequence of iterates $\{x_k\}$.

- 1: $\xi_0 := 0$; $x_1 := x_0$;
 - 2: **for** $k = 1, 2, \dots$ **do**
 - 3: Compute ξ_k as a minimizer of \hat{f}_{x_k} over $\mathcal{S}_k := \text{span} \{P_\gamma^{1 \leftarrow 0} \xi_{k-1}, \text{grad } f(x_k)\}$ where $\gamma(t) := R_{x_{k-1}}(t \xi_{k-1})$;
 - 4: Compute $x_{k+1} = R_{x_k}(\xi_k)$;
 - 5: **end for**
-

8. Applications. Several occurrences of Algorithms 1 (ALS), 2 (ATR), and 3 (SESOP) appear in the literature, e.g., in [20], [31], and in several eigenvalue algorithms. Indeed, it is well-known that subspace acceleration can remarkably improve the efficiency of eigensolvers; see, for example, the numerical comparison in [6, Figure 4.3] between a steepest descent algorithm and an accelerated version thereof, equivalent to locally optimal block preconditioned conjugate gradient (LOBPCG). Since, moreover, subspace acceleration is easy to perform for the eigenvalue problem, there are few methods that do not exploit it.

In the context of this analysis paper, we will focus on showing that the theory developed in the previous sections leads to convergence results for certain well-known algorithms. Some of these convergence results are new, to the best of our knowledge. In other cases, we recover results that have already been established, but the acceleration-based proof technique is novel and arguably more streamlined.

8.1. Lanczos algorithm. In a Ritz-restarted Lanczos algorithm for computing the leftmost eigenpair of a symmetric matrix A , the next iterate x_{k+1} is chosen as a minimizer of the Rayleigh quotient over the subspace $\mathcal{K}_m(x_k) := \text{span}\{x_k, Ax_k, A^2x_k, \dots, A^m x_k\}$, $m \geq 1$. Recall that the Rayleigh quotient of A is the function

$$f : \mathbb{R}_0^n \rightarrow \mathbb{R} : x \mapsto \frac{x^T A x}{x^T x}.$$

Its stationary points are the eigenvectors of A , and at those points it takes the value of the corresponding eigenvalue. (Note, however, that $f(x) = \lambda_i$, where λ_i is an eigenvalue of A , does not imply that x is an eigenvector of A , unless λ_i is an extreme eigenvalue of A .) Since x_k belongs to $\mathcal{K}_m(x_k)$, we are in the situation (24), and thus the Ritz-restarted Lanczos algorithm is an instance of Algorithm 3 (SESOP) (specifically, of Algorithm 4 (\mathbb{R}^n -SESOP)). The gradient of the Rayleigh quotient at x_k is collinear with $Ax_k - f(x_k)x_k$, which belongs to $\mathcal{K}_m(x_k)$, and hence $\{\mathcal{K}_m(x_k)\}$ is gradient-related to $\{x_k\}$. It follows from Theorem 7.2 that every limit point of $\{x_k\}$ is an eigenvector of A , regardless of x_0 . Taking into account the properties of the Rayleigh quotient f along with the fact that $\{x_k\}$ is a descent sequence for f , it follows that $\{x_k\}$ converges to the eigenspace associated to an eigenvalue of A . The same conclusion holds for the Ritz-restarted Krylov method proposed by Golub and Ye [19] for the symmetric definite generalized eigenvalue problem. In other words, we recovered [19, Theorem 3.2].

8.2. LOBPCG. Knyazev's LOBPCG method [26], in combination with a symmetric positive-definite preconditioner, is a popular algorithm for computing approx-

imations to the smallest eigenvalues and eigenvectors of the eigenproblem

$$Au = Bu\lambda,$$

where A and B are real symmetric positive-definite matrices of order n . Here we consider LOBPCG as formulated in [21, Algorithm 1] (with some changes in the notation), and we show, using Theorem 4.3, that the limit points of $\{\text{col}(X_k)\}$ are invariant subspaces of the pencil (A, B) . Moreover, invariant subspaces that do not correspond to the smallest eigenvalues are “unstable,” in the sense explained below.

The LOBPCG algorithm is described in Algorithm 6. In the algorithm, $(Y, \Theta) = \text{RR}(S, p)$ performs a Rayleigh–Ritz analysis where the pencil $(S^T AS, S^T BS)$ has eigenvectors Y and eigenvalues Θ , i.e.,

$$S^T ASY = S^T BSY\Theta \quad \text{and} \quad Y^T S^T BSY = I_{b \times b},$$

where $I_{b \times b}$ is the identity matrix of size $b \times b$. The first p pairs with smallest Ritz values are returned in Y and in the diagonal matrix Θ in a nondecreasing order. Note that we consider the formulation [21, Algorithm 1] because it is simple to state and comprehend. However, it should be kept in mind that the matrix $[X_k, H_k, P_k]$ may become singular or ill-conditioned [21]. Therefore, in practical implementations, it is recommended to rely on the robust representation given in [21, Algorithm 2]. The convergence results obtained below also hold in this case.

ALGORITHM 6. LOBPCG [26, 21] WITHOUT SOFT-LOCKING

Require: Symmetric positive-definite matrices A and B of order n ; symmetric positive-definite preconditioner N ; block-size p .

- 1: Select an initial guess $\tilde{X} \in \mathbb{R}^{n \times p}$.
 - 2: $X_0 = \tilde{X}Y$ where $(Y, \Theta_0) = \text{RR}(\tilde{X}, p)$.
 - 3: $R_k = AX_0 - MX_0\Theta_0$.
 - 4: $P_k = []$.
 - 5: **for** $k = 0, 1, 2, \dots$ **do**
 - 6: Solve the preconditioned linear system $NH_k = R_k$.
 - 7: Let $S = [X_k, H_k, P_k]$ and compute $(Y_k, \Theta_{k+1}) = \text{RR}(S, p)$.
 - 8: $X_{k+1} = [X_k, H_k, P_k]Y_k$.
 - 9: $R_{k+1} = AX_{k+1} - MX_{k+1}\Theta_{k+1}$.
 - 10: $P_{k+1} = [0, H_k, P_k]Y_k$.
 - 11: **end for**
-

In the case $p = 1$, it takes routine manipulations to check, using Proposition 7.2 with the Rayleigh quotient as the cost function, that all of the limit points of $\{X_k\}$ are eigenvectors of the pencil (A, B) . We now consider the general case $p \geq 1$ in detail.

Let $\mathbb{R}_*^{n \times p}$ denote the set of all full-rank $n \times p$ real matrices. Observe that $\mathbb{R}_*^{n \times p}$ is an open subset of $\mathbb{R}^{n \times p}$ (it is thus an open submanifold of the linear manifold $\mathbb{R}^{n \times p}$, see [6]) and that $T_X \mathbb{R}_*^{n \times p} \simeq \mathbb{R}^{n \times p}$ for all $X \in \mathbb{R}_*^{n \times p}$. In $\mathbb{R}_*^{n \times p}$, consider the inner product defined by

$$(25) \quad \langle Z_1, Z_2 \rangle_X = 2 \text{trace} \left((X^T BX)^{-1} Z_1^T Z_2 \right), \quad X \in \mathbb{R}_*^{n \times p}, \quad Z_1, Z_2 \in T_X \mathbb{R}_*^{n \times p}.$$

(The factor of 2 is included here to prevent factors of 2 from appearing in the formula of the gradient below. This is still a valid inner product, and it turns $\mathbb{R}_*^{n \times p}$ into a

Riemannian manifold.) Consider the cost function

$$(26) \quad f : \mathbb{R}_*^{n \times p} \rightarrow \mathbb{R} : X \mapsto \text{trace} \left((X^T B X)^{-1} X^T A X \right).$$

This generalized Rayleigh quotient was studied, e.g., in [6] (when $B = I$, it reduces to the *extended Rayleigh quotient* of [22]). It satisfies the property $f(XW) = f(X)$ for all $X \in \mathbb{R}_*^{n \times p}$ and all W invertible of size $p \times p$. A matrix $X \in \mathbb{R}_*^{n \times p}$ is a stationary point of f if and only if its column space is an invariant subspace of the pencil (A, B) . The value of f at an invariant subspace is the sum of the corresponding eigenvalues. The stationary points whose column space is the rightmost invariant subspace of (A, B) (i.e., the one corresponding to the largest eigenvalues) are global maximizers of f . The stationary points whose column space is the leftmost invariant subspace of (A, B) (i.e., the one corresponding to the smallest eigenvalues) are global minimizers of f . All of the other stationary points are saddle points.

The fact that $\mathbb{R}_*^{n \times p}$ is $\mathbb{R}^{n \times p}$ with infinitely many elements excerpeted makes it difficult to view LOBPCG as an instance of Algorithm 3 (SESOP). Instead, we view it as an instance of Algorithm 1 (ALS). The gradient of f with respect to the Riemannian metric (25) is

$$\text{grad } f(X) = AX - BX (X^T B X)^{-1} X^T A X;$$

see, e.g., [6, equation (6.37)]. Referring to Algorithm 6, we have $H_k = N^{-1} \text{grad } f(X_k)$ and

$$\langle \text{grad } f(X_k), -H_k \rangle_{X_k} = \left\| N^{-\frac{1}{2}} \text{grad } f(X_k) \right\|_F^2,$$

from which it follows that $\{-H_k\}$ is gradient-related to $\{X_k\}$ (Definition 4.1). We consider the retraction given by $R_X(Z) = X + Z$, $X \in \mathbb{R}_*^{n \times p}$, $Z \in T_X \mathbb{R}_*^{n \times p} \simeq \mathbb{R}^{n \times p}$. The Armijo point along $-H_k$ takes the form

$$X_{k+1/2} = X_k - \alpha_k H_k$$

for some $\alpha_k > 0$. Hence

$$X_{k+1/2} = [X_k, H_k, P_k] Y$$

for some Y . Without preconditioning ($N = I$), $X_{k+1/2}$ is full-rank (i.e., it belongs to $\mathbb{R}_*^{n \times p}$) for any α_k . Indeed, we have that $X_k^T X_{k+1/2} = X_k^T (I - \alpha_k A) X_k + \alpha_k X_k^T A X_k = X_k^T X_k$ is full-rank. (Observe that all iterates are B -orthogonal, hence of full rank.) With the preconditioner, however, this property is no longer guaranteed. Nevertheless, given A, B and N symmetric positive-definite matrices of order n , it is possible to find $\bar{\alpha}$ such that $X - \alpha N^{-1} \text{grad } f(X)$ has full rank for all B -orthonormal X and all $\alpha \in [0, \bar{\alpha}]$. (This is because $\{X \in \mathbb{R}^{n \times p} : X^T B X = I\}$ is a compact subset of $\mathbb{R}^{n \times p}$ and $\mathbb{R}^{n \times p} \setminus \mathbb{R}_*^{n \times p}$ is a closed subset of $\mathbb{R}^{n \times p}$ that do not intersect, and hence their distance does not vanish.) With this $\bar{\alpha}$, LOBPCG becomes an instance of Algorithm 1 (ALS), provided we show that the acceleration bound (5) holds for some $c > 0$. It does hold for $c = 1$, as a consequence of the following result.

LEMMA 8.1. *In the context of Algorithm 6, we have*

$$\begin{aligned} f(X_{k+1}) &= \min \{ f([X_k, H_k, P_k] Y) : Y \in \mathbb{R}^{3p \times p}, Y^T [X_k, H_k, P_k]^T B [X_k, H_k, P_k] Y = I \} \\ &= \min \{ f([X_k, H_k, P_k] Y) : Y \in \mathbb{R}^{3p \times p}, [X_k, H_k, P_k] Y \text{ full rank} \}, \end{aligned}$$

where f denotes the Rayleigh quotient (26).

Proof. The three expressions are equal to the sum of the p leftmost eigenvalues of the pencil $(U^T AU, U^T BU)$, where U is a full-rank matrix with $\text{col}(U) = \text{col}([X_k, H_k, P_k])$. \square

This yields the following result.

PROPOSITION 8.2. *Let $\{X_k\}$ be a sequence of iterates generated by Algorithm 6 (LOBPCG). Then the following holds.*

- (a) *Every limit point X_* of $\{X_k\}$ is a stationary point of f ; i.e., $\text{col}(X_*)$ is an invariant subspace of (A, B) ;*
- (b) *$\lim_{k \rightarrow \infty} \|AX_k - BX_k \Theta_k\| = 0$, where Θ_k is as in Algorithm 6 (LOBPCG);*
- (c) *The limit points of $\{\text{col}(X_k)\}$ are p -dimensional invariant subspaces of (A, B) ;*
- (d) *$\lim_{k \rightarrow \infty} f(X_k)$ exists (where f is the generalized Rayleigh quotient (26)), and thus f takes the same value at all limit points of $\{X_k\}$.*
- (e) *Let \mathcal{V} be a limit point of $\{\text{col}(X_k)\}$ that is not a leftmost invariant subspace of (A, B) (“leftmost” means related to the smallest eigenvalues). Then \mathcal{V} is unstable in the following sense: there is $\epsilon > 0$ such that for all $\delta > 0$ there exists $K > 0$ and $Z \in \mathbb{R}^{n \times p}$, with $\|Z\| < \delta$, such that if X_K is perturbed to $X_K + Z$ and the algorithm is pursued from this new iterate, then the new sequence satisfies $\angle(\text{col}(X_k), \mathcal{V}) > \epsilon$ for all but finitely many iterates.*

Proof. Point (a) follows from Proposition 4.3 as explained above. Point (b) follows from Corollary 4.4 since all iterates belong to the compact set $\{X \in \mathbb{R}^{n \times p} : X^T B X = I\}$. Note that $\text{grad } f(X_k) = AX_k - BX_k \Theta_k$. Point (c) involves the topology of the quotient manifold. The result follows from the fact that the col mapping is continuous from $\mathbb{R}_*^{n \times p}$ to the Grassmann manifold of p -planes in \mathbb{R}^n . (The topology of the Grassmann manifolds is precisely the one that makes the col mapping continuous; see, e.g., [6] for details.) Point (d) holds because LOBPCG is a descent method for f . Point (e) can be deduced from the fact that the non-leftmost invariant subspaces of (A, B) are saddle points or maxima for f and from the fact that LOBPCG is a descent method for f . \square

8.3. Jacobi–Davidson methods. The Jacobi–Davidson algorithm for computing the smallest eigenvalue and eigenvector of an $n \times n$ symmetric matrix A , as described in [38, Algorithm 1], clearly fits within Algorithm 3 (SESOP). However, without further assumptions, it is not guaranteed that $\{\mathcal{S}_k\}$ be gradient-related: it all depends on how the Jacobi correction equation is “approximately” solved. If the approximate solution can be guaranteed to be gradient-related, then it follows from Proposition 7.2 that all limit points are stationary points of the Rayleigh quotient; i.e., they are eigenvectors.

For example, consider, as in [28], the Jacobi equation in the form

$$(27) \quad (I - x_k x_k^T)(A - \tau I)(I - x_k x_k^T) \eta_k = -(I - x_k x_k^T) A x_k, \quad x_k^T \eta_k = 0,$$

where τ is some target less than the smallest eigenvalue λ_1 of A , and assume that the approximate solution η_k is obtained with m_k steps of the CG iteration ($1 \leq m_k < n$ for all k). We show that the sequence $\{\eta_k\}$ is gradient-related to $\{x_k\}$, and thus $\{\mathcal{S}_k\}$ is gradient-related to $\{x_k\}$ when \mathcal{S}_k contains η_k for all k . By the workings of CG (with zero initial condition), η_k is equal to $V_{m_k} y_k$, where V_{m_k} is an orthonormal basis of the Krylov subspace \mathcal{K}_{m_k} generated from $-(I - x_k x_k^T) A x_k$ using the operator $(I - x_k x_k^T)(A - \tau I)(I - x_k x_k^T)$ and where y_k solves

$$(28) \quad V_{m_k}^T (A - \tau I) V_{m_k} y_k = -V_{m_k}^T A x_k.$$

Notice that the Krylov subspace is orthogonal to x_k and contains the gradient $(I -$

$x_k x_k^T)Ax_k$, and hence we have the identities $(I - x_k x_k^T)V_{m_k} = V_{m_k}$ and $V_{m_k} V_{m_k}^T Ax_k = V_{m_k} V_{m_k}^T (I - x_k x_k^T)Ax_k = (I - x_k x_k^T)Ax_k$. Since $A - \tau I$ is positive-definite, it follows that the condition number of the projected matrix $V_{m_k}^T (A - \tau I)V_{m_k}$ is bounded, and hence in view of (28) the angle between y_k and $-V_{m_k}^T Ax_k$ is bounded away from $\frac{\pi}{2}$, and so is the angle between $V_{m_k} y_k = \eta_k$ and $-V_{m_k} V_{m_k}^T Ax_k = -(I - x_k x_k^T)Ax_k$ because V_{m_k} is an orthonormal basis. Moreover, $\{y_k\}$ is bounded away from zero and infinity, and so is $\{\eta_k\}$. We have thus shown that the sequence $\{\eta_k\}$ is gradient-related to $\{x_k\}$ (see the discussion that follows Definition 4.1). Thus Proposition 8.2 holds, mutatis mutandis, for the Jacobi–Davidson method [38, Algorithm 1] when the Jacobi equation (27) is defined and solved approximately with CG as in [28].

The result still holds when the CG iteration for (approximately) solving (27) is preconditioned with a positive-definite preconditioner N_k . Indeed, the preconditioned CG for solving a linear system $B\eta = -g$ amounts to applying the “regular” CG method to the transformed system $\tilde{B}\tilde{\eta} = -\tilde{g}$, where $\tilde{B} = N^{-1}BN^{-1}$, $\tilde{\eta} = N\eta$, and $\tilde{g} = N^{-1}g$ (see, e.g., [18, section 10.3]). If $\tilde{\eta}_j$ is an iterate of the regular CG applied to $\tilde{B}\tilde{\eta} = -\tilde{g}$ and thus $\eta_j = N^{-1}\tilde{\eta}_j$ is the iterate of the preconditioned CG, then we have $\langle \tilde{\eta}_j, \tilde{g} \rangle = \langle N\eta_j, N^{-1}g \rangle = \langle \eta_j, g \rangle$. Thus the sequence $\{\eta_k\}$, where η_k is the approximate solution of (27) returned by the preconditioned CG, is gradient-related.

Note that the choice of τ to make $(A - \tau I)$ positive-definite in (27) is crucial in the development above. In the frequently encountered case where τ is selected as the Rayleigh quotient θ_k at x_k , it seems difficult to provide a theoretical guarantee that the approximate solution η_k of (27) is gradient-related, unless we assume that the iteration starts close enough to the minor eigenvector so that $(I - x_k x_k^T)(A - \theta_k I)(I - x_k x_k^T)$ is positive definite as a linear transformation of the orthogonal complement of x_k . (An example of the requirement that the iteration start sufficiently close to the minor eigenvector is the condition $\theta_k < \frac{\lambda_1 + \lambda_2}{2}$ in [29, Theorem 4.3].) However, in practice, it is quite clear that a solver producing a sequence $\{\eta_k\}$ that is not gradient-related would have to be particularly odd. It is thus not surprising that the global convergence properties stated in Proposition 8.2 have been empirically observed in general for eigenvalue algorithms that fit in the Jacobi–Davidson framework.

Another example (which does not fit, strictly speaking, in the Jacobi–Davidson framework, but is closely related) is when, as in [1], the Jacobi equation is solved approximately using a truncated CG algorithm and the approximate solution is accepted or rejected using a trust-region mechanism. The method becomes an instance of Algorithm 2 applied to the Rayleigh quotient cost function, and Proposition 8.2 holds, mutatis mutandis.

8.4. Sequential subspace method. All of the algorithms thus far in this section are concerned with the eigenvalue problem; however, the area of application of the convergence theory developed in this paper is not restricted to eigenvalue solvers. An example is the SSM of Hager [20] for minimizing an arbitrary quadratic function over a sphere. This algorithm is an instance of Algorithm 3 (SESOP). In [20], $\{\mathcal{S}_k\}$ is required to contain $\text{grad} f(x_k)$; therefore, all limit points are stationary by Proposition 7.2. This was proven in [25], where stronger global convergence results are obtained by making additional assumptions on $\{\mathcal{S}_k\}$.

9. Concluding remarks. If we accelerate, in the sense of (1), an optimization algorithm that converges globally to stationary points of the cost function, do we preserve the global convergence result? We have answered this question positively for a wide class of line-search and trust-region methods. The global convergence of several eigenvalue algorithms follows from this result, under mild conditions, as shown

in section 8. We suspect that several other existing methods satisfy the conditions of the global convergence theorems proven in this paper.

An important practical issue in the design of accelerated algorithms is to strike a good balance of the workload between the “Jacobi-like” step (i.e., the computation of an update vector η_k) and the “Davidson-like” step (i.e., the improvement on the Jacobi update, for example, via a minimization within a subspace containing η_k). For example, at one extreme, the simplified Jacobi–Davidson in [28] simply turns off the Davidson step. Note that the algorithm in [8], where the “Jacobi” step consists of solving approximately a certain trust-region-like problem, shows promising numerical results even without using a “Davidson” step. At the other extreme, the workings of the the Jacobi–Davidson approach [38] can be exploited to let the Davidson step compensate for a crude approximation of the Jacobi update. In LOBPCG, the balance of the workload between the Jacobi-like step (computation of H_k) and the Davidson-like step (computation of X_{k+1} from $[X_k, H_k, P_k]$ by a Ritz process) depends much on the complexity of the chosen preconditioner; we refer, e.g., to [5, 27] for more information on preconditioners in LOBPCG. Note that in an eigenvalue method for a matrix A , the structure of A and the nature of the preconditioner will affect the computational burden on the Jacobi-like step, whereas the Davidson-like step, if implemented efficiently, should require only some orthogonalization routines and be largely independent of the cost of the operators. Hence, when the operators are inexpensive, it becomes more affordable to require a higher accuracy in the Jacobi-like step. We refer to [35, 24, 23, 36] for further work along these lines.

Finally, we point out that there is not necessarily a unique way of separating the instructions of an iterative loop into a Jacobi-like step and a Davidson-like step that satisfy the conditions for the global convergence analysis. For example, the application of a preconditioner can be considered as part of the Jacobi-like step or as part of the acceleration step if the preconditioning leads to an acceleration bound (1).

Acknowledgments. This work benefited in particular from discussions with Chris Baker, Bill Hager, Ekkehard Sachs, and Gerard Sleijpen. Special thanks to Chris Baker for his helpful comments on the manuscript.

REFERENCES

- [1] P.-A. ABSIL, C. G. BAKER, AND K. A. GALLIVAN, *A truncated-CG style method for symmetric generalized eigenvalue problems*, J. Comput. Appl. Math., 189 (2006), pp. 274–285.
- [2] P.-A. ABSIL, C. G. BAKER, AND K. A. GALLIVAN, *Trust-region methods on Riemannian manifolds*, Found. Comput. Math., 7 (2007), pp. 303–330.
- [3] R. L. ADLER, J.-P. DEDIEU, J. Y. MARGULIES, M. MARTENS, AND M. SHUB, *Newton’s method on Riemannian manifolds and a geometric model for the human spine*, IMA J. Numer. Anal., 22 (2002), pp. 359–390.
- [4] P.-A. ABSIL AND K. A. GALLIVAN, *Accelerated Line-search and Trust-region Methods*, Technical report FSU-SCS-2005-095, School of Computational Science, Florida State University, Tallahassee, FL, 2005.
- [5] P. ARBENZ, U. L. HETMANIUK, R. B. LEHOUCQ, AND R. S. TUMINARO, *A comparison of eigen-solvers for large-scale 3D modal analysis using AMG-preconditioned iterative methods*, Internat. J. Numer. Methods Engrg., 64 (2005), pp. 204–236.
- [6] P.-A. ABSIL, R. MAHONY, AND R. SEPULCHRE, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ, 2008.
- [7] L. ARMIJO, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific J. Math., 16 (1966), pp. 1–3.
- [8] C. G. BAKER, P.-A. ABSIL, AND K. A. GALLIVAN, *An implicit trust-region method on Riemannian manifolds*, IMA J. Numer. Anal., to appear.

- [9] C. G. BAKER, *Riemannian Manifold Trust-region Methods with Applications to Eigenproblems*, Ph.D. thesis, School of Computational Science, Florida State University, Tallahassee, FL, 2008.
- [10] D. P. BERTSEKAS, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1995.
- [11] A. R. CONN, N. I. M. GOULD, AND P. L. TOINT, *Trust-Region Methods*, MPS/SIAM Ser. Optim. 1, SIAM, Philadelphia, 2000.
- [12] E. R. DAVIDSON, *The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices*, J. Comput. Phys., 17 (1975), pp. 87–94.
- [13] M. P. DO CARMO, *Riemannian geometry*, Math. Theory Appl., Birkhäuser Boston, Boston, MA, 1992. Translated from the second Portuguese edition by Francis Flaherty.
- [14] A. EDELMAN, T. A. ARIAS, AND S. T. SMITH, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 303–353.
- [15] D. R. FOKKEMA, G. L. G. SLEIJPEN, AND H. A. VAN DER VORST, *Accelerated inexact Newton schemes for large systems of nonlinear equations*, SIAM J. Sci. Comput., 19 (1998), pp. 657–674.
- [16] D. R. FOKKEMA, G. L. G. SLEIJPEN, AND H. A. VAN DER VORST, *Jacobi–Davidson style QR and QZ algorithms for the reduction of matrix pencils*, SIAM J. Sci. Comput., 20 (1998), pp. 94–125.
- [17] D. GABAY, *Minimizing a differentiable function over a differential manifold*, J. Optim. Theory Appl., 37 (1982), pp. 177–219.
- [18] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 3rd ed., Johns Hopkins S. Math. Sci., Johns Hopkins University Press, Baltimore, MD, 1996.
- [19] G. H. GOLUB AND Q. YE, *An inverse free preconditioned Krylov subspace method for symmetric generalized eigenvalue problems*, SIAM J. Sci. Comput., 24 (2002), pp. 312–334.
- [20] W. W. HAGER, *Minimizing a quadratic over a sphere*, SIAM J. Optim., 12 (2001), pp. 188–208.
- [21] U. HETMANIUK AND R. LEHOUCQ, *Basis selection in LOBPCG*, J. Comput. Phys., 218 (2006), pp. 324–332.
- [22] U. HELMKE AND J. B. MOORE, *Optimization and Dynamical Systems*, Comm. Control Engrg. Ser., Springer-Verlag, London, 1994.
- [23] M. E. HOCHSTENBACH AND Y. NOTAY, *Controlling Inner Iterations in the Jacobi–Davidson Method*, SIAM J. Matrix Anal. Appl., to appear.
- [24] M. E. HOCHSTENBACH AND Y. NOTAY, *The Jacobi–Davidson method*, GAMM Mitt. Ges. Angew. Math. Mech., 29 (2006), pp. 368–382.
- [25] W. W. HAGER AND S. PARK, *Global convergence of SSM for minimizing a quadratic over a sphere*, Math. Comp., 74 (2005), pp. 1413–1423.
- [26] A. V. KNYAZEV, *Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method*, SIAM J. Sci. Comput., 23 (2001), pp. 517–541.
- [27] I. LASHUK, M. ARGENTI, E. OVTCHINNIKOV, AND A. KNYAZEV, *Preconditioned eigensolver LOBPCG in Hypra and PETSc*, in Domain Decomposition Methods in Science and Engineering XVI, Lect. Notes Comput. Sci. Eng. 55, Springer-Verlag, Berlin, 2007.
- [28] Y. NOTAY, *Combination of Jacobi–Davidson and conjugate gradients for the partial symmetric eigenproblem*, Numer. Linear Algebra Appl., 9 (2004), pp. 21–44.
- [29] Y. NOTAY, *Is Jacobi–Davidson faster than Davidson?*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 522–543.
- [30] J. NOCEDAL AND S. J. WRIGHT, *Numerical Optimization*, Springer Ser. Oper. Res., Springer-Verlag, New York, 1999.
- [31] G. NARKISS AND M. ZIBULEVSKY, *Sequential Subspace Optimization Method for Large-Scale Unconstrained Problems*, Technical report CCIT 559, EE Dept., Technion, Haifa, Israel, 2005.
- [32] M. SHUB, *Some remarks on dynamical systems and numerical analysis*, in Dynamical Systems and Partial Differential Equations, Proceedings of the VII ELAM, L. Lara-Carrero and J. Lewowicz, eds., Equinoccio, Universidad Simón Bolívar, Caracas, 1986, pp. 69–91.
- [33] S. T. SMITH, *Geometric Optimization Methods for Adaptive Filtering*, Ph.D. thesis, Division of Applied Sciences, Harvard University, Cambridge, MA, 1993.
- [34] S. T. SMITH, *Optimization techniques on Riemannian manifolds*, in Hamiltonian and Gradient Flows, Algorithms and Control, Fields Inst. Commun. 3, American Mathematical Society, Providence, RI, 1994, pp. 113–136.
- [35] A. STATHOPOULOS AND Y. SAAD, *Restarting techniques for the (Jacobi-)Davidson symmetric eigenvalue methods*, Electron. Trans. Numer. Anal., 7 (1998), pp. 163–181.
- [36] A. STATHOPOULOS, *Nearly optimal preconditioned methods for Hermitian eigenproblems under limited memory. Part I: Seeking one eigenvalue*, SIAM J. Sci. Comput., 29 (2007), pp. 481–514.

- [37] T. STEIHAUG, *The conjugate gradient method and trust regions in large scale optimization*, SIAM J. Numer. Anal., 20 (1983), pp. 626–637.
- [38] G. L. G. SLEIJPEN AND H. A. VAN DER VORST, *A Jacobi–Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 401–425.
- [39] I. TAKAHASHI, *A note on the conjugate gradient method*, Inform. Process. Japan, 5 (1965), pp. 45–49.
- [40] C. UDRIȘTE, *Convex Functions and Optimization Methods on Riemannian Manifolds*, Math. Appl. 297, Kluwer Academic, Dordrecht, the Netherlands, 1994.
- [41] Y. YANG, *Globally convergent optimization algorithms on Riemannian manifolds: Uniform framework for unconstrained and constrained optimization*, J. Optim. Theory Appl., 132 (2007), pp. 245–265.

ON PRECONDITIONED ITERATIVE METHODS FOR CERTAIN TIME-DEPENDENT PARTIAL DIFFERENTIAL EQUATIONS*

ZHONG-ZHI BAI[†], YU-MEI HUANG[‡], AND MICHAEL K. NG[§]

Abstract. When the Newton method or the fixed-point method is employed to solve the systems of nonlinear equations arising in the sinc-Galerkin discretization of certain time-dependent partial differential equations, in each iteration step we need to solve a structured subsystem of linear equations iteratively by, for example, a Krylov subspace method such as the preconditioned GMRES. In this paper, based on the tensor and the Toeplitz structures of the linear subsystems we construct structured preconditioners for their coefficient matrices and estimate the eigenvalue bounds of the preconditioned matrices under certain assumptions. Numerical examples are given to illustrate the effectiveness of the proposed preconditioning methods. It has been shown that a combination of the Newton/fixed-point iteration with the preconditioned GMRES method is efficient and robust for solving the systems of nonlinear equations arising from the sinc-Galerkin discretization of the time-dependent partial differential equations.

Key words. time-dependent partial differential equation, sinc-Galerkin discretization, Toeplitz-like matrix, preconditioning, eigenvalue bound, GMRES method

AMS subject classifications. 65F10, 65F15, 65T10; CR: G1.3

DOI. 10.1137/080718176

1. Introduction. We consider the numerical solution of time-dependent partial differential equations of the form

$$(1.1) \quad \begin{cases} p_t(t) \frac{\partial u}{\partial t}(x, t) + p_x(x) u(x, t) \frac{\partial u}{\partial x}(x, t) - \varepsilon \frac{\partial^2 u}{\partial x^2}(x, t) = f(x, t), & a < x < b, \quad t \geq 0, \\ u(a, t) = \gamma(t) \quad \text{and} \quad u(b, t) = \delta(t), & t \geq 0, \\ u(x, 0) = g(x), & a \leq x \leq b, \end{cases}$$

where $p_z(z)$, $z \in \{x, t\}$, are given continuously differentiable functions, $f(x, t)$, $\gamma(t)$, $\delta(t)$, and $g(x)$ are given bounded functions, and ε is a prescribed small positive parameter. Note that when $p_z(z) \equiv 1$, $z \in \{x, t\}$, the partial differential equation (1.1) reduces to the Burgers equation; see [16] for more details.

When the time-dependent partial differential equation (1.1) is discretized by the sinc-Galerkin method, in an analogous approach to [5] we can obtain systems of nonlinear equations of the form

$$(1.2) \quad \mathbf{F}(\mathbf{u}) := B\mathbf{u} + C\Psi(\mathbf{u}) - \mathbf{b} = 0,$$

*Received by the editors March 11, 2008; accepted for publication (in revised form) October 13, 2008; published electronically February 13, 2009.

<http://www.siam.org/journals/sinum/47-2/71817.html>

[†]State Key Laboratory of Scientific/Engineering Computing, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, P.O. Box 2719, Beijing 100080, People's Republic of China (bzz@lsec.cc.ac.cn). This author's research was supported by The National Basic Research Program (2005CB321702) and The National Outstanding Young Scientist Foundation (10525102), People's Republic of China.

[‡]School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, People's Republic of China.

[§]Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong (mng@math.hkbu.edu.hk). This author's research was supported in part by RGC grants 7046/03P, 7035/04P, and 7035/05P and FRG/04-05/II-51.

where B and C are known n -by- n matrices, \mathbf{b} is a given n -vector, and $\Psi : \mathbb{R}^n \rightarrow \mathbb{R}^n$, with

$$\Psi(\mathbf{u}) = (\psi_1(u_1), \psi_2(u_2), \dots, \psi_n(u_n))^T \quad \text{and} \quad \mathbf{u} = (u_1, u_2, \dots, u_n)^T,$$

is a continuous diagonal mapping defined on the open ball

$$\mathcal{U}_\delta := \{u \in \mathbb{R}^n \mid \|\mathbf{u}\| < \delta\}.$$

Here, δ is a positive constant. The matrices B and C are given by

$$\begin{aligned} B = \varepsilon & \left(T_x^{(2)} + D_x^{(1)} T_x^{(1)} + T_x^{(1)} D_x^{(1)} + D_x^{(2)} \right) \otimes Q_t \\ & + Q_x \otimes \left(D_t^{(3)} T_t^{(1)} + T_t^{(1)} D_t^{(3)} + D_t^{(4)} \right) \end{aligned} \tag{1.3}$$

and

$$C = \left(D_x^{(3)} T_x^{(1)} + T_x^{(1)} D_x^{(3)} + D_x^{(4)} \right) \otimes Q_t, \tag{1.4}$$

and the mapping Ψ is given by

$$\Psi(\mathbf{u}) = (u_1^2, u_2^2, \dots, u_n^2)^T, \tag{1.5}$$

where $T_z^{(i)}$ ($i = 1, 2$ and $z \in \{x, t\}$) are $(m_z + n_z + 1)$ -by- $(m_z + n_z + 1)$ Toeplitz matrices, with

$$T_z^{(1)} = \begin{bmatrix} 0 & -1 & \frac{1}{2} & \cdots & \frac{(-1)^{m_z+n_z}}{m_z+n_z} \\ 1 & & & & \vdots \\ -\frac{1}{2} & & \ddots & & \frac{1}{2} \\ \vdots & & & & -1 \\ -\frac{(-1)^{m_z+n_z}}{m_z+n_z} & \cdots & -\frac{1}{2} & 1 & 0 \end{bmatrix}, \tag{1.6}$$

$$T_z^{(2)} = \begin{bmatrix} \frac{\pi^2}{3} & -2 & \frac{2}{2^2} & \cdots & \frac{(-1)^{m_z+n_z} 2}{(m_z+n_z)^2} \\ -2 & & & & \vdots \\ \frac{2}{2^2} & & \ddots & & \frac{2}{2^2} \\ \vdots & & & & -2 \\ \frac{(-1)^{m_z+n_z} 2}{(m_z+n_z)^2} & \cdots & \frac{2}{2^2} & -2 & \frac{\pi^2}{3} \end{bmatrix}, \tag{1.7}$$

and $D_z^{(i)}$ and Q_z ($i = 1, 2, 3, 4$ and $z \in \{x, t\}$) are $(m_z + n_z + 1)$ -by- $(m_z + n_z + 1)$ diagonal matrices, with

$$D_z^{(1)} = \frac{h_z}{2} \cdot \text{diag} \left[\left\{ -\frac{\phi_z''(z)}{(\phi_z'(z))^2} - \frac{2\omega_z'(z)}{\phi_z'(z)\omega_z(z)} \right\}_{z=-m_z}^{n_z} \right], \tag{1.8}$$

$$D_z^{(2)} = \frac{h_z^2}{2} \cdot \text{diag} \left[\left\{ -\frac{\omega_z''(z)}{(\phi_z'(z))^2 \omega_z(z)} \right\}_{z=-m_z}^{n_z} \right], \tag{1.9}$$

$$(1.10) \quad D_z^{(3)} = \frac{h_z}{2} \cdot \text{diag} \left[\left\{ -p_z(z)\omega_z(z) \right\}_{z=-m_z}^{n_z} \right],$$

$$(1.11) \quad D_z^{(4)} = \frac{h_z^2}{2} \cdot \text{diag} \left[\left\{ -\frac{(p_z(z)\omega_z(z))'}{\phi_z'(z)} \right\}_{z=-m_z}^{n_z} \right],$$

and

$$(1.12) \quad Q_z = \text{diag} \left[\left\{ \frac{\omega_z(z)}{\phi_z'(z)} \right\}_{z=-m_z}^{n_z} \right].$$

Here, m_x, n_x and m_t, n_t are positive integers representing the numbers of the bases used in the spatial and the temporal spaces, respectively, $\phi_x(x)$ and $\phi_t(t)$ are the restrictions of the conformal mapping $\phi_z(z)$ onto the real intervals (a, b) and $(0, +\infty)$, respectively, with $\phi_z(z)$ a mapping from a simply connected domain \mathcal{D} onto

$$\mathcal{D}_d := \{z \mid z = x + iy, |y| < d, d > 0\},$$

with ι the imaginary unit; and $\omega_x(x)$ and $\omega_t(t)$ are two weighting functions with respect to the spatial and the temporal variables, respectively. See [16, 5] for a detailed description about the sinc-Galerkin discretization. We remark that the first and the second derivatives of $\phi_z(z)$ and $\omega_z(z)$ with respect to the variable z will be denoted as $\phi_z'(z), \omega_z'(z)$ and $\phi_z''(z), \omega_z''(z)$, respectively, and the matrices $T_z^{(1)}, z \in \{x, t\}$, defined in (1.6) are skew-symmetric, while the matrices $T_z^{(2)}, z \in \{x, t\}$, defined in (1.7) are symmetric positive definite; see Lemmas 2.1 and 2.2.

The system of nonlinear equations (1.2) is usually termed as a mildly nonlinear system in literature; see [19, 21] for general backgrounds and applications, [2, 5] for the basic existence and uniqueness theory about the solution, and [1, 2, 7, 8, 21, 22] for several splitting iteration methods in the sequential and parallel computing senses. When the system of mildly nonlinear equations (1.2) is solved by the Newton or the fixed-point iteration method, at each step we need to solve a subsystem of linear equations of the form

$$(1.13) \quad (B + CD)\mathbf{z} = \mathbf{r},$$

where D is a diagonal matrix approximating the Jacobian matrix of the mapping $\Psi : \mathbb{R}^n \rightarrow \mathbb{R}^n$ and \mathbf{r} is the current residual vector. Unfortunately, direct methods such as the Gaussian elimination or the fast Toeplitz algorithms [15, 14] are not applicable to effectively solve this class of diagonally scaled Toeplitz-plus-diagonal linear systems due to the considerably high computational complexity; see [9, 10, 11, 12, 13]. However, noticing that the matrix-vector product $(B+CD)\mathbf{q}$ can be computed in $\mathcal{O}(n \log n)$ operations for any vector $\mathbf{q} \in \mathbb{R}^n$, we can employ Krylov subspace iteration methods such as GMRES [20] to iteratively solve the linear subsystem (1.13) in an economical cost. Usually, in order to accelerate the convergence speeds of the Krylov subspace iteration methods, we need to precondition the linear subsystem (1.13) by a good approximating matrix with respect to the coefficient matrix $A := B + CD$. Therefore, in order to solve the original linear subsystem, we turn to solving the corresponding preconditioned linear subsystem instead; see [6, 5] and the references therein.

In this paper, we construct a structured preconditioner M for the matrix A by making use of the tensor-product structure of the original matrix A and the diagonally

scaled Toeplitz-plus-diagonal structure of the matrix blocks involved. The positive definiteness of both matrices A and M are discussed in detail, and the eigenvalue bounds about the preconditioned matrix $M^{-1}A$ are estimated precisely by utilizing the generalized Bendixson theorem [6]. Theoretical analysis shows that the eigenvalues of the matrix $M^{-1}A$ are tightly and uniformly bounded in a rectangle on the complex plane independent of the size of the matrix. Numerical implementations show that the Newton-GMRES and the fixed-point-GMRES iteration methods, when incorporated with the structured preconditioner M , are effective and robust nonlinear solvers for the systems of mildly nonlinear equations arising from the sinc-Galerkin discretization of the referred time-dependent partial differential equations.

The organization of the paper is as follows. In section 2, we construct a structured preconditioner for the coefficient matrix of the linear subsystem (1.13) and analyze basic properties of the original and the preconditioning matrices. In section 3, we demonstrate several preliminary results associated with the spectral analysis of the preconditioned matrix. The eigenvalue bounds of the preconditioned matrix are estimated in section 4, and numerical examples are given in section 5 to show the effectiveness of the proposed preconditioning and the corresponding preconditioned iteration methods. Finally, in section 6, we end this paper with some concluding remarks.

2. The structured preconditioners. Consider the system of mildly nonlinear equations (1.2), with the function $\Psi(\mathbf{u})$ being given in (1.5) and the matrices B and C being given in (1.3) and (1.4), respectively, where $T_z^{(i)}$ ($i = 1, 2, z \in \{x, t\}$), $D_z^{(i)}$ ($i = 1, 2, 3, 4$ and $z \in \{x, t\}$) and Q_z ($z \in \{x, t\}$) are defined in (1.6)–(1.12). Denote by I the identity matrix. Let Ω be a positive definite diagonal matrix such that $D := I \otimes \Omega$ is an approximation to the Jacobian matrix of $\Psi(\mathbf{u})$. Then the target matrix under consideration is

$$\begin{aligned} A &= B + CD \\ &= \varepsilon \left(T_x^{(2)} + D_x^{(1)} T_x^{(1)} + T_x^{(1)} D_x^{(1)} + D_x^{(2)} \right) \otimes Q_t \\ &\quad + Q_x \otimes \left(D_t^{(3)} T_t^{(1)} + T_t^{(1)} D_t^{(3)} + D_t^{(4)} \right) \\ (2.1) \quad &\quad + \left(D_x^{(3)} T_x^{(1)} + T_x^{(1)} D_x^{(3)} + D_x^{(4)} \right) \otimes (Q_t \Omega). \end{aligned}$$

By utilizing the special structure of the matrix A , we can construct its preconditioner M as

$$\begin{aligned} M &= \widehat{B} + \widehat{C}D \\ &= \varepsilon \left(B_x^{(2)} + D_x^{(1)} B_x^{(1)} + B_x^{(1)} D_x^{(1)} + D_x^{(2)} \right) \otimes Q_t \\ &\quad + Q_x \otimes \left(D_t^{(3)} B_t^{(1)} + B_t^{(1)} D_t^{(3)} + D_t^{(4)} \right) \\ (2.2) \quad &\quad + \left(D_x^{(3)} B_x^{(1)} + B_x^{(1)} D_x^{(3)} + D_x^{(4)} \right) \otimes (Q_t \Omega), \end{aligned}$$

where

$$\begin{aligned} \widehat{B} &= \varepsilon \left(B_x^{(2)} + D_x^{(1)} B_x^{(1)} + B_x^{(1)} D_x^{(1)} + D_x^{(2)} \right) \otimes Q_t \\ &\quad + Q_x \otimes \left(D_t^{(3)} B_t^{(1)} + B_t^{(1)} D_t^{(3)} + D_t^{(4)} \right) \end{aligned}$$

and

$$\widehat{C} = \left(D_x^{(3)} B_x^{(1)} + B_x^{(1)} D_x^{(3)} + D_x^{(4)} \right) \otimes Q_t,$$

and, for $z \in \{x, t\}$,

$$(2.3) \quad B_z^{(1)} = \text{tridiag}[1, 0, -1] \quad \text{and} \quad B_z^{(2)} = \text{tridiag}[-1, 2, -1]$$

are tridiagonal approximations to $T_z^{(1)}$ and $T_z^{(2)}$, respectively. Note that the preconditioning matrix M is obtained by replacing only $T_z^{(i)}$ ($i = 1, 2, z \in \{x, t\}$) in the matrix A by $B_z^{(i)}$ ($i = 1, 2, z \in \{x, t\}$), correspondingly.

We remark that the preconditioner M is a block tridiagonal matrix and is usually of mild size as, compared with the finite-difference system, the sinc-Galerkin system needs not be very large and is of mild size in order to achieve the same discretization accuracy [17, 18, 5]. Therefore, for any given vector \mathbf{r} , the generalized residual equation $M\mathbf{w} = \mathbf{r}$ involved in the preconditioned GMRES iteration method can be solved in $\mathcal{O}(N_x N_t^2)$ or $\mathcal{O}(N_x^2 N_t)$ operations by using a variety of linear solvers such as the sparse direct methods, where $N_z = m_z + n_z + 1$, with $z \in \{x, t\}$.

It was proved in [16] that the Toeplitz matrix $T_x^{(2)}$ is symmetric positive definite and its eigenvalues are located in a positive interval. This result, together with some eigenproperties of the Toeplitz matrices $T_z^{(1)}$ ($z \in \{x, t\}$), is precisely described in the following lemma.

LEMMA 2.1 (see [16, Theorems 4.18 and 4.19]). *Let the matrices $T_z^{(1)}$ ($z \in \{x, t\}$) and $T_x^{(2)}$ be defined as in (1.6) and (1.7), respectively. Then*

- (i) *for $z \in \{x, t\}$, $T_z^{(1)}$ is a skew-symmetric matrix and its eigenvalues $\{i\lambda_j^{(1)}\}_{j=-m_z}^{n_z}$ satisfy $\lambda_j^{(1)} \in [-\pi, \pi]$, $-m_z \leq j \leq n_z$;*
- (ii) *$T_x^{(2)}$ is a symmetric positive definite matrix and its eigenvalues $\{\lambda_j^{(2)}\}_{j=-m_x}^{n_x}$ satisfy $\lambda_j^{(2)} \in [4 \sin^2(\frac{\pi}{2(N_x+1)}), \pi^2]$, where $N_x = m_x + n_x + 1$.*

Analogously, the structural properties and the eigenvalue locations about the matrices $B_z^{(1)}$ ($z \in \{x, t\}$) and $B_x^{(2)}$ are precisely described in the following lemma; see [4].

LEMMA 2.2 (see [4, Lemma A.1]). *Let the matrices $B_z^{(1)}$ ($z \in \{x, t\}$) and $B_x^{(2)}$ be defined as in (2.3). Then*

- (i) *for $z \in \{x, t\}$, $B_z^{(1)}$ is a skew-symmetric matrix and its eigenvalues $\{i\lambda_j^{(1)}\}_{j=-m_z}^{n_z}$ satisfy $\lambda_j^{(1)} \in [-\cos(\frac{\pi}{N_z+1}), \cos(\frac{\pi}{N_z+1})]$, $-m_z \leq j \leq n_z$, where $N_z = m_z + n_z + 1$;*
- (ii) *$B_x^{(2)}$ is a symmetric positive definite matrix and its eigenvalues $\{\lambda_j^{(2)}\}_{j=-m_x}^{n_x}$ satisfy $\lambda_j^{(2)} \in [4 \sin^2(\frac{\pi}{2(N_x+1)}), 4 \cos^2(\frac{\pi}{2(N_x+1)})]$, where $N_x = m_x + n_x + 1$.*

Based on these two lemmas, we now demonstrate the positive definiteness of the matrix A defined in (2.1) and its preconditioning matrix M defined in (2.2).

To this end, in what follows we use $(\cdot)^*$ to denote the conjugate transpose of either a vector or a square matrix. For a given square matrix X , we use $\mathcal{H}(X)$ and $\mathcal{S}(X)$ to denote, respectively, its Hermitian and skew-Hermitian parts [4] and $\lambda(X)$ its spectral set.

THEOREM 2.1. *Assume that $D_x^{(2)}$, $D_x^{(4)}$, and $D_t^{(4)}$ are positive semidefinite diagonal matrices and Q_z ($z \in \{x, t\}$) and Ω are positive definite diagonal matrices. Then both $\mathcal{H}(A)$ and $\mathcal{H}(M)$ are symmetric positive definite matrices. Hence, A and M are positive definite ¹ and, thus, are nonsingular.*

¹A matrix is positive definite if its Hermitian part is positive definite. Note that a positive definite matrix is not necessarily Hermitian; see [4, 3].

Proof. The Hermitian and the skew-Hermitian parts of A and M are

$$\begin{aligned}\mathcal{H}(A) &= \frac{1}{2}(A + A^*) \\ &= \varepsilon \left(T_x^{(2)} + D_x^{(2)} \right) \otimes Q_t + Q_x \otimes D_t^{(4)} + D_x^{(4)} \otimes (Q_t \Omega), \\ \mathcal{S}(A) &= \frac{1}{2}(A - A^*) \\ &= \varepsilon \left(D_x^{(1)} T_x^{(1)} + T_x^{(1)} D_x^{(1)} \right) \otimes Q_t + Q_x \otimes \left(D_t^{(3)} T_t^{(1)} + T_t^{(1)} D_t^{(3)} \right) \\ &\quad + \left(D_x^{(3)} T_x^{(1)} + T_x^{(1)} D_x^{(3)} \right) \otimes (Q_t \Omega)\end{aligned}$$

and

$$\begin{aligned}\mathcal{H}(M) &= \frac{1}{2}(M + M^*) \\ &= \varepsilon \left(B_x^{(2)} + D_x^{(2)} \right) \otimes Q_t + Q_x \otimes D_t^{(4)} + D_x^{(4)} \otimes (Q_t \Omega), \\ \mathcal{S}(M) &= \frac{1}{2}(M - M^*) \\ &= \varepsilon \left(D_x^{(1)} B_x^{(1)} + B_x^{(1)} D_x^{(1)} \right) \otimes Q_t + Q_x \otimes \left(D_t^{(3)} B_t^{(1)} + B_t^{(1)} D_t^{(3)} \right) \\ &\quad + \left(D_x^{(3)} B_x^{(1)} + B_x^{(1)} D_x^{(3)} \right) \otimes (Q_t \Omega).\end{aligned}$$

Because the diagonal matrices $D_x^{(2)}$, $D_x^{(4)}$, and $D_t^{(4)}$ are positive semidefinite, the diagonal matrices Q_z ($z \in \{x, t\}$) and Ω are positive definite, and from Lemma 2.1 the Toeplitz matrices $T_x^{(2)}$ are symmetric positive definite, so we know that $\mathcal{H}(A)$ is symmetric positive definite. Therefore, A is a positive definite matrix and, thus, is nonsingular.

From Lemma 2.2 the matrix $B_x^{(2)}$ is symmetric positive definite. By applying the same arguments to the preconditioning matrix M , we can immediately show that M is positive definite and nonsingular, too. \square

3. Several preliminary lemmas. In this section, we are going to demonstrate several lemmas that are indispensable for estimating the eigenvalue bounds of the preconditioned matrix $M^{-1}A$.

LEMMA 3.1. *Let $\Delta = \text{diag}(\delta_1, \delta_2, \dots, \delta_n)$ be an n -by- n positive diagonal matrix and $H \in \mathbb{C}^{n \times n}$ be a Hermitian positive definite matrix. Then it holds that*

$$\frac{v^*(\Delta \otimes H)v}{v^*(H \otimes \Delta)v} \leq \kappa(\Delta)\kappa(H) \quad \forall v \in \mathbb{C}^n \setminus \{0\},$$

where $\kappa(\cdot)$ denotes the Euclidean condition number of the corresponding matrix.

Proof. Because $H \in \mathbb{C}^{n \times n}$ is a Hermitian positive definite matrix, there exist a unitary matrix $U \in \mathbb{C}^{n \times n}$ and a positive diagonal matrix $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n) \in$

$\mathbb{R}^{n \times n}$ such that $H = U^* \Lambda U$. Therefore, for all $v \in \mathbb{C}^n \setminus \{0\}$ we have

$$\begin{aligned} \frac{v^*(\Delta \otimes H)v}{v^*(H \otimes \Delta)v} &= \frac{v^*[\Delta \otimes (U^* \Lambda U)]v}{v^*[(U^* \Lambda U) \otimes \Delta]v} \\ &= \frac{v^*[(I \otimes U)^*(\Delta \otimes \Lambda)(I \otimes U)]v}{v^*[(U \otimes I)^*(\Lambda \otimes \Delta)(U \otimes I)]v} \\ &\leq \frac{\max_{1 \leq \ell, j \leq n} \{\delta_\ell \lambda_j\}}{\min_{1 \leq \ell, j \leq n} \{\delta_j \lambda_\ell\}} \\ &= \frac{\max_{1 \leq \ell \leq n} \delta_\ell}{\min_{1 \leq \ell \leq n} \delta_\ell} \cdot \frac{\max_{1 \leq \ell \leq n} \lambda_\ell}{\min_{1 \leq \ell \leq n} \lambda_\ell} \\ &= \kappa(\Delta) \kappa(H). \quad \square \end{aligned}$$

While Lemma 3.1 gives an upper bound about the generalized Rayleigh quotient with respect to the Hermitian positive definite matrix H , the following lemma presents an estimate about the generalized Rayleigh quotient with respect to the Hermitian and the skew-Hermitian matrices H and S .

LEMMA 3.2. Let $\Gamma = \text{diag}(\gamma_1, \gamma_2, \dots, \gamma_n)$ and $\Delta = \text{diag}(\delta_1, \delta_2, \dots, \delta_n)$ be n -by- n positive diagonal matrices, $H \in \mathbb{C}^{n \times n}$ be a Hermitian positive definite matrix, and $S \in \mathbb{C}^{n \times n}$ be a skew-Hermitian matrix. Then it holds that

$$\left| \frac{v^*(S \otimes \Gamma)v}{v^*(H \otimes \Delta)v} \right| \leq \tau \left| \frac{v^*(S \otimes \Gamma)v}{v^*(H \otimes \Gamma)v} \right| \quad \forall v \in \mathbb{C}^n \setminus \{0\},$$

where $\tau = \max_{1 \leq \ell \leq n} \{\frac{\gamma_\ell}{\delta_\ell}\}$.

Proof. Because $H \in \mathbb{C}^{n \times n}$ is Hermitian positive definite, there exist a unitary matrix $U \in \mathbb{C}^{n \times n}$ and a positive diagonal matrix $\Lambda \in \mathbb{R}^{n \times n}$ such that $H = U^* \Lambda U$. Therefore, for all $v \in \mathbb{C}^n \setminus \{0\}$ we have

$$\begin{aligned} v^*(H \otimes \Delta)v &= v^*(U^* \Lambda U \otimes \Delta)v = v^*((U^* \otimes I)(\Lambda \otimes \Delta)(U \otimes I))v \\ &\geq \frac{1}{\tau} (v^*((U^* \otimes I)(\Lambda \otimes \Gamma)(U \otimes I))v) = \frac{1}{\tau} (v^*(H \otimes \Gamma)v). \end{aligned}$$

It then follows that

$$\left| \frac{v^*(S \otimes \Gamma)v}{v^*(H \otimes \Delta)v} \right| \leq \tau \left| \frac{v^*(S \otimes \Gamma)v}{v^*(H \otimes \Gamma)v} \right|. \quad \square$$

The following *generalized Bendixson theorem*, established in [6], is essential for us to derive a rectangular domain for bounding the eigenvalues of the preconditioned matrix $M^{-1}A$.

THEOREM 3.1 (see [6, Theorem 2.4]). Let $A, M \in \mathbb{C}^{n \times n}$ be n -by- n complex matrices, and, for $\forall v \in \mathbb{C}^n \setminus \{0\}$, it holds that $v^* \mathcal{H}(A)v \neq 0$ and $v^* \mathcal{H}(M)v \neq 0$. Let the functions $h(v)$, $f_A(v)$, and $f_M(v)$ be defined as

$$h(v) = \frac{v^* \mathcal{H}(A)v}{v^* \mathcal{H}(M)v}, \quad f_A(v) = \frac{1}{i} \cdot \frac{v^* \mathcal{S}(A)v}{v^* \mathcal{H}(A)v}, \quad \text{and} \quad f_M(v) = \frac{1}{i} \cdot \frac{v^* \mathcal{S}(M)v}{v^* \mathcal{H}(M)v},$$

respectively. Assume that there exist positive constants γ_1 and γ_2 such that

$$\gamma_1 \leq h(v) \leq \gamma_2 \quad \forall v \in \mathbb{C}^n \setminus \{0\}$$

and nonnegative constants η and μ such that

$$-\mu \leq f_A(v) \leq \mu \quad \text{and} \quad -\eta \leq f_M(v) \leq \eta \quad \forall v \in \mathbb{C}^n \setminus \{0\}.$$

Then, when $\eta\mu \leq 1$, we have

$$\begin{cases} \frac{(1 - \eta\mu)\gamma_1}{1 + \eta^2} \leq \operatorname{Re}(\lambda(M^{-1}A)) \leq (1 + \eta\mu)\gamma_2, \\ -(\eta + \mu)\gamma_2 \leq \operatorname{Im}(\lambda(M^{-1}A)) \leq (\eta + \mu)\gamma_2. \end{cases}$$

Here, $\operatorname{Re}(\cdot)$ and $\operatorname{Im}(\cdot)$ represent the real and the imaginary parts of the corresponding complex, respectively.

In order to derive the bounded domain about the eigenvalues of the matrix $M^{-1}A$ by making use of the generalized Bendixson theorem, we essentially need the bounds of several generalized Rayleigh quotients with respect to certain parts of the matrices A and M defined in (2.1) and (2.2). These bounds are precisely stated in the following two lemmas.

LEMMA 3.3 (see [6, Lemma 4.2]). Assume that $D_x^{(2)}$ defined in (1.9) is a positive semidefinite diagonal matrix. Let $T_x^{(2)}$ be the Toeplitz matrix defined in (1.7) and $B_x^{(2)}$ the tridiagonal matrix defined in (2.3), respectively. Then it holds that

$$1 \leq \frac{v^* (T_x^{(2)} + D_x^{(2)}) v}{v^* (B_x^{(2)} + D_x^{(2)}) v} \leq \frac{\pi^2}{4} \quad \forall v \in \mathbb{C}^n \setminus \{0\}.$$

LEMMA 3.4. Assume that $D_x^{(2)}$ defined in (1.9) is a positive semidefinite diagonal matrix, Q_t defined in (1.12) is a positive definite diagonal matrix, and $D_z^{(j)}$ ($j = 1, 3$, $z \in \{x, t\}$) are the diagonal matrices defined in (1.8) and (1.10). Let $T_z^{(1)}$ ($z \in \{x, t\}$) and $T_x^{(2)}$ be the Toeplitz matrices defined in (1.6) and (1.7) and $B_z^{(1)}$ ($z \in \{x, t\}$) and $B_x^{(2)}$ be the tridiagonal matrices defined in (2.3), respectively. Denote $c_x^{(2)} = 4 \sin^2(\frac{\pi}{2(N_x+1)})$. For $z \in \{x, t\}$, let $N_z = m_z + n_z + 1$ and assume $N := N_x = N_t$. Define

$$\bar{d}_z^{(j)} = \max_{1 \leq \ell \leq N} \left\{ \left[D_z^{(j)} \right]_{\ell\ell} \right\} \quad (j = 1, 3, \quad z \in \{x, t\}), \quad d_x^{(2)} = \min_{1 \leq \ell \leq N} \left\{ \left[D_x^{(2)} \right]_{\ell\ell} \right\}$$

and

$$\begin{aligned} \mu_z^{(j)} &= \frac{2\pi \bar{d}_z^{(j)}}{\sqrt{(c_x^{(2)} + d_x^{(2)}) (\pi^2 + d_x^{(2)})}}, \\ \eta_z^{(j)} &= \frac{\bar{d}_z^{(j)} \left(\sqrt{d_x^{(2)} + 4} - \sqrt{d_x^{(2)}} \right)}{\sqrt{c_x^{(2)} + d_x^{(2)}}}, \quad j = 1, 3, \quad z \in \{x, t\}. \end{aligned}$$

Then, for $j = 1, 3$, $z \in \{x, t\}$, and all $v \in \mathbb{C}^n \setminus \{0\}$, it holds that

$$\max \left\{ \left| \frac{v^* \left[\left(D_z^{(j)} T_z^{(1)} + T_z^{(1)} D_z^{(j)} \right) \otimes Q_t \right] v}{v^* \left[\left(T_x^{(2)} + D_x^{(2)} \right) \otimes Q_t \right] v} \right|, \left| \frac{v^* \left[Q_t \otimes \left(D_z^{(j)} T_z^{(1)} + T_z^{(1)} D_z^{(j)} \right) \right] v}{v^* \left[Q_t \otimes \left(T_x^{(2)} + D_x^{(2)} \right) \right] v} \right| \right\} \leq \mu_z^{(j)}$$

and

$$\max \left\{ \left| \frac{v^* \left[\left(D_z^{(j)} B_z^{(1)} + B_z^{(1)} D_z^{(j)} \right) \otimes Q_t \right] v}{v^* \left[\left(B_x^{(2)} + D_x^{(2)} \right) \otimes Q_t \right] v} \right|, \left| \frac{v^* \left[Q_t \otimes \left(D_z^{(j)} B_z^{(1)} + B_z^{(1)} D_z^{(j)} \right) \right] v}{v^* \left[Q_t \otimes \left(B_x^{(2)} + D_x^{(2)} \right) \right] v} \right| \right\} \leq \eta_z^{(j)}.$$

Proof. By making use of Lemma 2.1, following the same arguments as in the proof of [6, Lemma 4.3] we can obtain these estimates. \square

4. The spectral analysis. In this section, we will derive precise bounds for the eigenvalues of the preconditioned matrix $M^{-1}A$, where the matrices A and M are defined in (2.1) and (2.2), respectively. To this end, we first estimate the bounds of the function $h(v)$ defined in Theorem 3.1.

LEMMA 4.1. Assume that $D_x^{(2)}$ and $D_z^{(4)}$ ($z \in \{x, t\}$) defined in (1.9) and (1.11) are positive semidefinite diagonal matrices and Q_z ($z \in \{x, t\}$) defined in (1.12) and Ω are positive definite diagonal matrices. Let $T_x^{(2)}$ be the Toeplitz matrix defined in (1.7) and $B_x^{(2)}$ be the tridiagonal matrix defined in (2.3). Then

$$(4.1) \quad 1 \leq \frac{v^* \mathcal{H}(A)v}{v^* \mathcal{H}(M)v} \leq \frac{\pi^2}{4} \quad \forall v \in \mathbb{C}^n \setminus \{0\}.$$

Proof. For notational simplicity we denote

$$D_\delta = Q_x \otimes D_t^{(4)} + D_x^{(4)} \otimes (Q_t \Omega) + \delta I,$$

where $\delta > 0$ is arbitrary. Evidently, D_δ is a positive definite diagonal matrix. Therefore, for any $v \in \mathbb{C}^n \setminus \{0\}$, according to the proof of Theorem 2.1 we have

$$\begin{aligned} \frac{v^* [\mathcal{H}(A) + \delta I]v}{v^* [\mathcal{H}(M) + \delta I]v} &= \frac{v^* \left[\varepsilon \left(T_x^{(2)} + D_x^{(2)} \right) \otimes Q_t + D_\delta \right] v}{v^* \left[\varepsilon \left(B_x^{(2)} + D_x^{(2)} \right) \otimes Q_t + D_\delta \right] v} \\ &\leq \max \left\{ \frac{v^* \left[\varepsilon \left(T_x^{(2)} + D_x^{(2)} \right) \otimes Q_t \right] v}{v^* \left[\varepsilon \left(B_x^{(2)} + D_x^{(2)} \right) \otimes Q_t \right] v}, \frac{v^* D_\delta v}{v^* D_\delta v} \right\} \\ &= \max \left\{ \frac{v^* \left[\left(T_x^{(2)} + D_x^{(2)} \right) \otimes Q_t \right] v}{v^* \left[\left(B_x^{(2)} + D_x^{(2)} \right) \otimes Q_t \right] v}, 1 \right\}. \end{aligned}$$

The above inequality follows from the basic inequality:

$$\frac{\beta_1 + \beta_2}{\alpha_1 + \alpha_2} \leq \max \left\{ \frac{\beta_1}{\alpha_1}, \frac{\beta_2}{\alpha_2} \right\} \quad \forall \alpha_j, \beta_j > 0, \quad j = 1, 2.$$

Based on Lemma 3.3, we can demonstrate the validity of the estimate

$$\frac{v^* [\mathcal{H}(A) + \delta I]v}{v^* [\mathcal{H}(M) + \delta I]v} \leq \frac{\pi^2}{4}$$

in an analogous fashion to [6, Lemma 4.2]. Moreover, as $\delta > 0$ is arbitrary, it then follows that

$$\frac{v^* \mathcal{H}(A)v}{v^* \mathcal{H}(M)v} \leq \frac{\pi^2}{4}.$$

Similarly, the left-hand side of the inequality (4.1) can be verified. \square

For the bounds of the functions $f_A(v)$ and $f_M(v)$ defined in Theorem 3.1, we can give the following estimates.

LEMMA 4.2. Assume that $D_x^{(2)}$ and $D_z^{(4)}$ ($z \in \{x, t\}$) defined in (1.9) and (1.11) are positive semidefinite diagonal matrices, Q_z ($z \in \{x, t\}$) defined in (1.12) and Ω are positive definite diagonal matrices, and $D_z^{(j)}$ ($j = 1, 3, z \in \{x, t\}$) are the diagonal matrices defined in (1.8) and (1.10). Let $T_z^{(1)}$ ($z \in \{x, t\}$) and $T_x^{(2)}$ be the Toeplitz matrices defined in (1.6) and (1.7) and $B_z^{(1)}$ ($z \in \{x, t\}$) and $B_x^{(2)}$ be the tridiagonal matrices defined in (2.3), respectively. Denote $c_x^{(2)} = 4 \sin^2(\frac{\pi}{2(N_x+1)})$. For $z \in \{x, t\}$, let $N_z = m_z + n_z + 1$ and assume $N := N_x = N_t$. Define

$$\bar{d}_z^{(j)} = \max_{1 \leq \ell \leq N} \left\{ \left[D_z^{(j)} \right]_{\ell\ell} \right\} \quad (j = 1, 2, 3), \quad d_x^{(2)} = \min_{1 \leq \ell \leq N} \left\{ \left[D_x^{(2)} \right]_{\ell\ell} \right\}$$

and

$$\begin{aligned} \mu_z^{(j)} &= \frac{2\pi \bar{d}_z^{(j)}}{\sqrt{(c_x^{(2)} + d_x^{(2)}) (\pi^2 + d_x^{(2)})}}, \\ \eta_z^{(j)} &= \frac{\bar{d}_z^{(j)} \left(\sqrt{d_x^{(2)} + 4} - \sqrt{d_x^{(2)}} \right)}{\sqrt{c_x^{(2)} + d_x^{(2)}}}, \quad j = 1, 3, \quad z \in \{x, t\}. \end{aligned}$$

Let

$$\begin{cases} \mu = \mu_x^{(1)} + \frac{\varepsilon \left(\pi^2 + \bar{d}_x^{(2)} \right) \kappa(Q_t) \max_{1 \leq \ell \leq N} \left\{ \left[Q_t^{-1} Q_x \right]_{\ell\ell} \right\}}{c_x^{(2)} + d_x^{(2)}} \mu_t^{(3)} + \max_{1 \leq \ell \leq N} \left\{ \left[\Omega \right]_{\ell\ell} \right\} \mu_x^{(3)}, \\ \eta = \eta_x^{(1)} + \frac{\varepsilon \left(4 - c_x^{(2)} + \bar{d}_x^{(2)} \right) \kappa(Q_t) \max_{1 \leq \ell \leq N} \left\{ \left[Q_t^{-1} Q_x \right]_{\ell\ell} \right\}}{c_x^{(2)} + d_x^{(2)}} \eta_t^{(3)} + \max_{1 \leq \ell \leq N} \left\{ \left[\Omega \right]_{\ell\ell} \right\} \eta_x^{(3)}. \end{cases}$$

Then it holds that

$$\left| \frac{v^* \mathcal{S}(A)v}{v^* \mathcal{H}(A)v} \right| \leq \mu \quad \text{and} \quad \left| \frac{v^* \mathcal{S}(M)v}{v^* \mathcal{H}(M)v} \right| \leq \eta \quad \forall v \in \mathbb{C}^n \setminus \{0\}.$$

Proof. For notational simplicity we denote

$$D^{(4)} = Q_x \otimes D_t^{(4)} + D_x^{(4)} \otimes (Q_t \Omega).$$

Because $D_z^{(4)}$ ($z \in \{x, t\}$) are positive semidefinite diagonal matrices and Q_z ($z \in \{x, t\}$) and Ω are positive definite diagonal matrices, we see that $D^{(4)}$ is a positive semidefinite diagonal matrix.

For any $v \in \mathbb{C}^n \setminus \{0\}$, according to the proof of Theorem 2.1 we have

$$\begin{aligned}
 \left| \frac{v^* \mathcal{S}(A)v}{v^* \mathcal{H}(A)v} \right| &\leq \left| \frac{v^* \left[\varepsilon \left(D_x^{(1)} T_x^{(1)} + T_x^{(1)} D_x^{(1)} \right) \otimes Q_t \right] v}{v^* \left[\varepsilon \left(T_x^{(2)} + D_x^{(2)} \right) \otimes Q_t + D^{(4)} \right] v} \right| \\
 &\quad + \left| \frac{v^* \left[Q_x \otimes \left(D_t^{(3)} T_t^{(1)} + T_t^{(1)} D_t^{(3)} \right) \right] v}{v^* \left[\varepsilon \left(T_x^{(2)} + D_x^{(2)} \right) \otimes Q_t + D^{(4)} \right] v} \right| \\
 &\quad + \left| \frac{v^* \left[\left(D_x^{(3)} T_x^{(1)} + T_x^{(1)} D_x^{(3)} \right) \otimes (Q_t \Omega) \right] v}{v^* \left[\varepsilon \left(T_x^{(2)} + D_x^{(2)} \right) \otimes Q_t + D^{(4)} \right] v} \right| \\
 &\leq \left| \frac{v^* \left[\left(D_x^{(1)} T_x^{(1)} + T_x^{(1)} D_x^{(1)} \right) \otimes Q_t \right] v}{v^* \left[\left(T_x^{(2)} + D_x^{(2)} \right) \otimes Q_t \right] v} \right| \\
 &\quad + \left| \frac{v^* \left[Q_x \otimes \left(D_t^{(3)} T_t^{(1)} + T_t^{(1)} D_t^{(3)} \right) \right] v}{v^* \left[\varepsilon \left(T_x^{(2)} + D_x^{(2)} \right) \otimes Q_t \right] v} \right| \\
 &\quad + \left| \frac{v^* \left[\left(D_x^{(3)} T_x^{(1)} + T_x^{(1)} D_x^{(3)} \right) \otimes (Q_t \Omega) \right] v}{v^* \left[\varepsilon \left(T_x^{(2)} + D_x^{(2)} \right) \otimes Q_t \right] v} \right|
 \end{aligned} \tag{4.2}$$

and

$$\begin{aligned}
 \left| \frac{v^* \mathcal{S}(M)v}{v^* \mathcal{H}(M)v} \right| &\leq \left| \frac{v^* \left[\varepsilon \left(D_x^{(1)} B_x^{(1)} + B_x^{(1)} D_x^{(1)} \right) \otimes Q_t \right] v}{v^* \left[\varepsilon \left(B_x^{(2)} + D_x^{(2)} \right) \otimes Q_t + D^{(4)} \right] v} \right| \\
 &\quad + \left| \frac{v^* \left[Q_x \otimes \left(D_t^{(3)} B_t^{(1)} + B_t^{(1)} D_t^{(3)} \right) \right] v}{v^* \left[\varepsilon \left(B_x^{(2)} + D_x^{(2)} \right) \otimes Q_t + D^{(4)} \right] v} \right| \\
 &\quad + \left| \frac{v^* \left[\left(D_x^{(3)} B_x^{(1)} + B_x^{(1)} D_x^{(3)} \right) \otimes (Q_t \Omega) \right] v}{v^* \left[\varepsilon \left(B_x^{(2)} + D_x^{(2)} \right) \otimes Q_t + D^{(4)} \right] v} \right| \\
 &\leq \left| \frac{v^* \left[\left(D_x^{(1)} B_x^{(1)} + B_x^{(1)} D_x^{(1)} \right) \otimes Q_t \right] v}{v^* \left[\left(B_x^{(2)} + D_x^{(2)} \right) \otimes Q_t \right] v} \right| \\
 &\quad + \left| \frac{v^* \left[Q_x \otimes \left(D_t^{(3)} B_t^{(1)} + B_t^{(1)} D_t^{(3)} \right) \right] v}{v^* \left[\varepsilon \left(B_x^{(2)} + D_x^{(2)} \right) \otimes Q_t \right] v} \right| \\
 &\quad + \left| \frac{v^* \left[\left(D_x^{(3)} B_x^{(1)} + B_x^{(1)} D_x^{(3)} \right) \otimes (Q_t \Omega) \right] v}{v^* \left[\varepsilon \left(B_x^{(2)} + D_x^{(2)} \right) \otimes Q_t \right] v} \right|.
 \end{aligned} \tag{4.3}$$

Here in both estimates we have technically split the nominators into three parts and then used the triangular inequality to obtain the first inequalities. The second inequalities are directly obtained by using the positive semidefiniteness of the diagonal

matrix $D^{(4)}$. In addition, we have used the facts that $D_x^{(2)}$ is a positive semidefinite diagonal matrix and both $T_x^{(2)}$ and $B_x^{(2)}$ are positive definite Toeplitz matrices; see Lemma 2.1.

From Lemma 3.4 we easily see that

$$(4.4) \quad \left| \frac{v^* \left[\left(D_x^{(1)} T_x^{(1)} + T_x^{(1)} D_x^{(1)} \right) \otimes Q_t \right] v}{v^* \left[\left(T_x^{(2)} + D_x^{(2)} \right) \otimes Q_t \right] v} \right| \leq \mu_x^{(1)}$$

and

$$(4.5) \quad \left| \frac{v^* \left[\left(D_x^{(1)} B_x^{(1)} + B_x^{(1)} D_x^{(1)} \right) \otimes Q_t \right] v}{v^* \left[\left(B_x^{(2)} + D_x^{(2)} \right) \otimes Q_t \right] v} \right| \leq \eta_x^{(1)}$$

hold true. It follows from Lemmas 2.1 and 2.2 that

$$(4.6) \quad \kappa \left(T_x^{(2)} + D_x^{(2)} \right) \leq \frac{\pi^2 + \bar{d}_x^{(2)}}{c_x^{(2)} + d_x^{(2)}} \quad \text{and} \quad \kappa \left(B_x^{(2)} + D_x^{(2)} \right) \leq \frac{4 - c_x^{(2)} + \bar{d}_x^{(2)}}{c_x^{(2)} + d_x^{(2)}}.$$

By making use of Lemma 3.1 and (4.6), we have

$$(4.7) \quad \begin{aligned} v^* \left[\varepsilon \left(T_x^{(2)} + D_x^{(2)} \right) \otimes Q_t \right] v &\geq \frac{v^* \left[Q_t \otimes \varepsilon \left(T_x^{(2)} + D_x^{(2)} \right) \right] v}{\kappa \left(\varepsilon \left(T_x^{(2)} + D_x^{(2)} \right) \right) \kappa(Q_t)} \\ &\geq \frac{c_x^{(2)} + d_x^{(2)}}{\varepsilon \left(\pi^2 + \bar{d}_x^{(2)} \right) \kappa(Q_t)} \cdot v^* \left[Q_t \otimes \varepsilon \left(T_x^{(2)} + D_x^{(2)} \right) \right] v \\ &= \frac{1}{\sigma_T} \cdot v^* \left[Q_t \otimes \varepsilon \left(T_x^{(2)} + D_x^{(2)} \right) \right] v \end{aligned}$$

and

$$(4.8) \quad \begin{aligned} v^* \left[\varepsilon \left(B_x^{(2)} + D_x^{(2)} \right) \otimes Q_t \right] v &\geq \frac{v^* \left[Q_t \otimes \varepsilon \left(B_x^{(2)} + D_x^{(2)} \right) \right] v}{\kappa \left(\varepsilon \left(B_x^{(2)} + D_x^{(2)} \right) \right) \kappa(Q_t)} \\ &\geq \frac{c_x^{(2)} + d_x^{(2)}}{\varepsilon \left(4 - c_x^{(2)} + \bar{d}_x^{(2)} \right) \kappa(Q_t)} \cdot v^* \left[Q_t \otimes \varepsilon \left(B_x^{(2)} + D_x^{(2)} \right) \right] v \\ &= \frac{1}{\sigma_B} \cdot v^* \left[Q_t \otimes \varepsilon \left(B_x^{(2)} + D_x^{(2)} \right) \right] v, \end{aligned}$$

where

$$\sigma_T = \frac{\varepsilon \left(\pi^2 + \bar{d}_x^{(2)} \right) \kappa(Q_t)}{c_x^{(2)} + d_x^{(2)}} \quad \text{and} \quad \sigma_B = \frac{\varepsilon \left(4 - c_x^{(2)} + \bar{d}_x^{(2)} \right) \kappa(Q_t)}{c_x^{(2)} + d_x^{(2)}}.$$

Therefore, according to Lemmas 3.2 and 3.4, as well as (4.7)–(4.8), it holds that

$$\begin{aligned}
 \left| \frac{v^* \left[Q_x \otimes \left(D_t^{(3)} T_t^{(1)} + T_t^{(1)} D_t^{(3)} \right) \right] v}{v^* \left[\varepsilon \left(T_x^{(2)} + D_x^{(2)} \right) \otimes Q_t \right] v} \right| &\leq \sigma_T \left| \frac{v^* \left[Q_x \otimes \left(D_t^{(3)} T_t^{(1)} + T_t^{(1)} D_t^{(3)} \right) \right] v}{v^* \left[Q_t \otimes \varepsilon \left(T_x^{(2)} + D_x^{(2)} \right) \right] v} \right| \\
 &\leq \sigma_T \tau_Q \left| \frac{v^* \left[Q_x \otimes \left(D_t^{(3)} T_t^{(1)} + T_t^{(1)} D_t^{(3)} \right) \right] v}{v^* \left[Q_x \otimes \varepsilon \left(T_x^{(2)} + D_x^{(2)} \right) \right] v} \right| \\
 (4.9) \qquad \qquad \qquad &\leq \sigma_T \tau_Q \mu_t^{(3)}
 \end{aligned}$$

and

$$\begin{aligned}
 \left| \frac{v^* \left[Q_x \otimes \left(D_t^{(3)} B_t^{(1)} + B_t^{(1)} D_t^{(3)} \right) \right] v}{v^* \left[\varepsilon \left(B_x^{(2)} + D_x^{(2)} \right) \otimes Q_t \right] v} \right| &\leq \sigma_B \left| \frac{v^* \left[Q_x \otimes \left(D_t^{(3)} B_t^{(1)} + B_t^{(1)} D_t^{(3)} \right) \right] v}{v^* \left[Q_t \otimes \varepsilon \left(B_x^{(2)} + D_x^{(2)} \right) \right] v} \right| \\
 &\leq \sigma_B \tau_Q \left| \frac{v^* \left[Q_x \otimes \left(D_t^{(3)} B_t^{(1)} + B_t^{(1)} D_t^{(3)} \right) \right] v}{v^* \left[Q_x \otimes \varepsilon \left(B_x^{(2)} + D_x^{(2)} \right) \right] v} \right| \\
 (4.10) \qquad \qquad \qquad &\leq \sigma_B \tau_Q \eta_t^{(3)},
 \end{aligned}$$

where $\tau_Q = \max_{1 \leq \ell \leq N} \{ [Q_t^{-1} Q_x]_{\ell\ell} \}$. In addition, according to Lemmas 3.2 and 3.4 it holds that

$$\begin{aligned}
 \left| \frac{v^* \left[\left(D_x^{(3)} T_x^{(1)} + T_x^{(1)} D_x^{(3)} \right) \otimes (Q_t \Omega) \right] v}{v^* \left[\varepsilon \left(T_x^{(2)} + D_x^{(2)} \right) \otimes Q_t \right] v} \right| &\leq \tau_\Omega \left| \frac{v^* \left[\left(D_x^{(3)} T_x^{(1)} + T_x^{(1)} D_x^{(3)} \right) \otimes (Q_t \Omega) \right] v}{v^* \left[\varepsilon \left(T_x^{(2)} + D_x^{(2)} \right) \otimes (Q_t \Omega) \right] v} \right| \\
 (4.11) \qquad \qquad \qquad &\leq \tau_\Omega \mu_x^{(3)}
 \end{aligned}$$

and

$$\begin{aligned}
 \left| \frac{v^* \left[\left(D_x^{(3)} B_x^{(1)} + B_x^{(1)} D_x^{(3)} \right) \otimes (Q_t \Omega) \right] v}{v^* \left[\varepsilon \left(B_x^{(2)} + D_x^{(2)} \right) \otimes Q_t \right] v} \right| &\leq \tau_\Omega \left| \frac{v^* \left[\left(D_x^{(3)} B_x^{(1)} + B_x^{(1)} D_x^{(3)} \right) \otimes (Q_t \Omega) \right] v}{v^* \left[\varepsilon \left(B_x^{(2)} + D_x^{(2)} \right) \otimes (Q_t \Omega) \right] v} \right| \\
 (4.12) \qquad \qquad \qquad &\leq \tau_\Omega \eta_x^{(3)},
 \end{aligned}$$

where $\tau_Q = \max_{1 \leq \ell \leq N} \{ [\Omega]_{\ell\ell} \}$.

Now, by substituting the inequalities (4.4), (4.5), (4.9), (4.10), (4.11), and (4.12) into (4.2) and (4.3), we immediately obtain the estimates that we are deriving. \square

By using Theorem 3.1 and Lemmas 4.1 and 4.2, we can straightforwardly obtain the main theorem of this paper.

THEOREM 4.1. *Let the conditions of Lemma 4.2 be satisfied. Without loss of generality, we make use of scaling on the original system of linear equations such that $\mu\eta < 1$. Then it holds that*

$$\frac{1 - \mu\eta}{1 + \eta^2} \leq \operatorname{Re} (\lambda (M^{-1}A)) \leq \frac{\pi^2(1 + \mu\eta)}{4}$$

and

$$-\frac{\pi^2(\mu + \eta)}{4} \leq \operatorname{Im} (\lambda (M^{-1}A)) \leq \frac{\pi^2(\mu + \eta)}{4}.$$

Based on Theorem 4.1, we can immediately obtain a theoretical estimate about the asymptotic convergence rate of the preconditioned GMRES method with the preconditioner M in (2.2) for solving the system of linear equations (1.13). Here, we should suitably scale the partial differential equation (1.1) and appropriately choose the weighting functions $\omega_x(x)$ and $\omega_t(t)$ and the conformal mappings $\phi_x(x)$ and $\phi_t(t)$ such that $\mu\eta < 1$. For details, we refer to [20, 6].

We remark that, when Theorem 4.1 is specialized to the matrices A and M , arising from the sinc-Galerkin discretization of the Burgers equation, much sharper bounds than those given in [5] about the eigenvalues of the preconditioned matrix $M^{-1}A$ can be straightforwardly obtained under weaker restrictions. This is one of the theoretical advantages of our new result.

5. Numerical experiments. In this section, we use two examples of the time-dependent partial differential equation (1.1) to demonstrate the effectiveness of the preconditioning and the corresponding preconditioned GMRES iteration method. Here, both Newton and fixed-point methods are applied to solve the discretized system of nonlinear equations (1.2).

In our computations, the initial guess is set to be the zero vector and the outer nonlinear iteration is stopped once the current residual satisfies the criteria

$$\frac{\|r^{(k)}\|_2}{\|r^{(0)}\|_2} \leq 10^{-6}.$$

In each outer iteration step, a preconditioned linear system

$$(5.1) \quad M^{-1}Az = M^{-1}\mathbf{r}, \quad \text{with } A = B + CD \quad \text{and} \quad M = \widehat{B} + \widehat{C}D,$$

is solved, which forms the inner iteration process for solving the linear subsystems involved in each step of the Newton or the fixed-point method; see (1.13) and (2.2). Here, the stopping criteria for the inner iteration, i.e., the preconditioned GMRES method, is that the relative reduction on the norm of the residual is less than 10^{-6} . Besides, all codes are written in MATLAB 7.01 and all experiments are implemented on a personal computer with 2.66GHz central processing unit and 0.99G memory.

For the positive diagonal matrix $\Omega = \text{diag}([\Omega]_{11}, [\Omega]_{22}, \dots, [\Omega]_{N_t N_t})$, we can construct it according to a certain approximating rule. With respect to the Newton iteration method, we may minimize $\|I \otimes \Omega - \Psi'(\mathbf{u}^{(c)})\|_2$ to obtain the Ω , where $\mathbf{u}^{(c)} = (u_1^{(c)}, u_2^{(c)}, \dots, u_n^{(c)})^T$ is the current Newton iterate. As now $\Psi'(\mathbf{u}) = 2 \cdot \text{diag}(u_1, u_2, \dots, u_n)$, with $\mathbf{u} = (u_1, u_2, \dots, u_n)^T$ and $n = N_x N_t$, by direct computations we can obtain the formulas for the diagonal elements of Ω as follows:

$$[\Omega]_{jj} = \frac{2}{N_x} \sum_{k=0}^{N_x-1} u_{kN_t+j}^{(c)}, \quad j = 1, 2, \dots, N_t.$$

Analogously, with respect to the fixed-point iteration method, we can choose

$$[\Omega]_{jj} = \frac{1}{N_x} \sum_{k=0}^{N_x-1} u_{kN_t+j}^{(c)}, \quad j = 1, 2, \dots, N_t,$$

where $\mathbf{u}^{(c)} = (u_1^{(c)}, u_2^{(c)}, \dots, u_n^{(c)})^T$ denotes the current fixed-point iterate. Note that the difference between these two Ω 's is just a factor of 2.

The following two equations in the form of (1.1) are used to examine the numerical performance of the new preconditioner M defined in (2.2) and to show the accuracy of the computed solution.

Example 5.1. The time-dependent partial differential equation

$$\left\{ \begin{aligned} &-\frac{\partial u}{\partial t}(x, t) + \frac{u(x, t)}{x} \frac{\partial u}{\partial x}(x, t) - \varepsilon \frac{\partial^2 u}{\partial x^2}(x, t) \\ &= e^{-\pi^2 t} \sin(\pi x) \\ &\quad \cdot \left(\pi^2 t - 1 + \frac{\pi t e^{-\pi^2 t} \cos(\pi x)}{x} + \varepsilon \pi^2 t \right), \quad 0 < x < 1 \quad \text{and} \quad t \geq 0, \\ &u(0, t) = 0 \quad \text{and} \quad u(1, t) = 0, \quad t \geq 0, \\ &u(x, 0) = 0, \quad 0 \leq x \leq 1, \end{aligned} \right.$$

with the exact solution being $u(x, t) = te^{-\pi^2 t} \sin(\pi x)$.

Example 5.2. The time-dependent partial differential equation

$$\left\{ \begin{aligned} &-\frac{\partial u}{\partial t}(x, t) + \frac{u(x, t)}{x} \frac{\partial u}{\partial x}(x, t) - \varepsilon \frac{\partial^2 u}{\partial x^2}(x, t) \\ &= -xe^{-t}(1-x)(1-t) \\ &\quad + t^2 e^{-2t}(1-x)(1-2x) - 2\varepsilon t e^{-t}, \quad 0 < x < 1 \quad \text{and} \quad t \geq 0, \\ &u(0, t) = 0 \quad \text{and} \quad u(1, t) = 0, \quad t \geq 0, \\ &u(x, 0) = 0, \quad 0 \leq x \leq 1, \end{aligned} \right.$$

with the exact solution being $u(x, t) = x(1-x)te^{-t}$.

The conformal mappings are chosen as $\phi(z) = \ln(\frac{z}{1-z})$ and $\psi(z) = \ln(\sinh(z))$ so that their restrictions onto the real intervals $(0, 1)$ and $(0, +\infty)$ are $\phi_x(x) := \phi(x) = \ln(\frac{x}{1-x})$ and $\phi_t(t) := \psi(t) = \ln(\sinh(t))$, which are used for the discretizations of x and t variables, respectively. And the weighting functions are chosen to be $\omega_x(x) = 1/\phi'_x(x)$ and $\omega_t(t) = 1/\phi'_t(t)$.

In the numerical tables, the symbol I means that no preconditioner is used when solving the linear subsystems involved in the nonlinear iterations, while M represents that the preconditioner M defined in (2.2) is used. We use N_{IT} to denote the number of the Newton iteration steps, F_{IT} that of the fixed-point iteration steps, G_{IT} the average number of GMRES iteration steps in each Newton or fixed-point iteration, CPU the total computing timings, and Se the maximum absolute discretization error at the sinc grid points and Ue that on the corresponding uniform grid points, while we use “average Se ” and “average Ue ” to represent the average absolute errors at all of the sinc grid points and at all of the uniform grid points, respectively. In addition, the symbol $*$ is used to denote that the iteration does not satisfy the terminating criterion within 50 steps of the Newton or the fixed-point iteration while $+$ that the inner iteration does not satisfy the GMRES terminating criterion within 1000 iteration steps.

We solve Example 5.1 when $\varepsilon = 10^{-3}$ and $\varepsilon = 10^{-4}$. Tables 5.1–5.2 list the numbers of iteration steps and the CPU timings required for the convergence of the Newton iteration, and Tables 5.3–5.4 list those required for the convergence of the fixed-point iteration, respectively, when they are applied to solve the system of nonlinear equations (1.2) resulting from the sinc-Galerkin discretization of Example 5.1. Tables 5.5 and 5.6 list iteration numbers and CPU timings when the Newton and the fixed-point methods are applied, respectively, to Example 5.2, with $\varepsilon = 10^{-3}$. In all tables, some errors for reflecting the accuracy of the computed solutions are also shown.

TABLE 5.1
Results for Example 5.1. $\varepsilon = 10^{-3}$, and the Newton method is applied.

n	I			M						
	N_{IT}	G_{IT}	CPU	N_{IT}	G_{IT}	Se	average Se	Ue	average Ue	CPU
81	4	80	0.33	4	32	2.22×10^{-3}	7.73×10^{-4}	2.14×10^{-3}	6.54×10^{-4}	0.33
289	4	282	3.00	4	58	1.21×10^{-3}	1.72×10^{-4}	1.08×10^{-3}	8.70×10^{-5}	0.98
1089	4	977	62.36	4	111	1.55×10^{-3}	1.60×10^{-4}	1.30×10^{-3}	1.41×10^{-5}	6.48
4225	*	+	—	4	246	1.69×10^{-3}	1.86×10^{-4}	1.48×10^{-3}	1.05×10^{-5}	78.59

TABLE 5.2
Results for Example 5.1. $\varepsilon = 10^{-4}$, and the Newton method is applied.

n	I			M						
	N_{IT}	G_{IT}	CPU	N_{IT}	G_{IT}	Se	average Se	Ue	average Ue	CPU
81	4	80	0.33	4	35	2.36×10^{-3}	7.31×10^{-4}	2.23×10^{-3}	6.22×10^{-4}	0.25
289	4	283	3.02	4	68	5.11×10^{-4}	8.08×10^{-5}	2.61×10^{-4}	7.05×10^{-5}	1.13
1089	4	963	61.25	4	148	3.70×10^{-4}	3.03×10^{-5}	1.71×10^{-4}	4.31×10^{-6}	8.72
4225	*	+	—	5	359	4.62×10^{-4}	2.91×10^{-5}	1.85×10^{-4}	1.43×10^{-6}	170.20

TABLE 5.3
Results for Example 5.1. $\varepsilon = 10^{-3}$, and the fixed-point method is applied.

n	I			M						
	F_{IT}	G_{IT}	CPU	F_{IT}	G_{IT}	Se	average Se	Ue	average Ue	CPU
81	5	65	0.33	5	25	2.22×10^{-3}	7.73×10^{-4}	2.14×10^{-3}	6.54×10^{-4}	0.23
289	4	210	2.25	4	40	1.21×10^{-3}	1.72×10^{-4}	1.08×10^{-3}	8.70×10^{-5}	0.64
1089	6	824	83.67	6	76	1.54×10^{-3}	1.62×10^{-4}	1.30×10^{-3}	1.41×10^{-5}	6.32
4225	*	+	—	12	140	1.69×10^{-3}	1.90×10^{-4}	1.47×10^{-3}	1.05×10^{-5}	111.63

TABLE 5.4
Results for Example 5.1. $\varepsilon = 10^{-4}$, and the fixed-point method is applied.

n	I			M						
	F_{IT}	G_{IT}	CPU	F_{IT}	G_{IT}	Se	average Se	Ue	average Ue	CPU
81	6	68	0.34	6	27	2.36×10^{-3}	7.31×10^{-4}	2.23×10^{-3}	6.22×10^{-4}	0.30
289	4	211	2.27	4	46	5.11×10^{-4}	8.08×10^{-5}	2.60×10^{-4}	7.06×10^{-5}	0.75
1089	4	711	47.06	3	72	2.04×10^{-4}	2.49×10^{-5}	1.74×10^{-4}	4.30×10^{-6}	3.23
4225	*	+	—	3	123	2.44×10^{-4}	2.77×10^{-5}	1.99×10^{-4}	1.47×10^{-6}	25.67

TABLE 5.5
Results for Example 5.2. $\varepsilon = 10^{-3}$, and the Newton method is applied.

n	I			M						
	N_{IT}	G_{IT}	CPU	N_{IT}	G_{IT}	Se	average Se	Ue	average Ue	CPU
289	9	285	7.08	9	87	4.23×10^{-3}	1.30×10^{-3}	4.27×10^{-3}	1.41×10^{-3}	3.31
1089	9	996	149.11	9	179	1.82×10^{-3}	6.59×10^{-4}	1.80×10^{-3}	4.87×10^{-4}	25.27
4225	*	+	—	10	508	2.35×10^{-3}	5.77×10^{-4}	2.17×10^{-3}	3.17×10^{-4}	653.31

TABLE 5.6
Results for Example 5.2. $\varepsilon = 10^{-3}$, and the fixed-point method is applied.

n	I			M						
	F_{IT}	G_{IT}	CPU	F_{IT}	G_{IT}	Se	average Se	Ue	average Ue	CPU
289	12	246	7.64	11	53	4.23×10^{-3}	1.30×10^{-3}	4.27×10^{-3}	1.41×10^{-3}	2.38
1089	16	808	204.30	14	93	1.82×10^{-3}	6.60×10^{-4}	1.80×10^{-3}	4.87×10^{-4}	17.98
4225	*	+	—	33	168	2.35×10^{-3}	5.77×10^{-4}	2.17×10^{-3}	3.17×10^{-4}	377.16

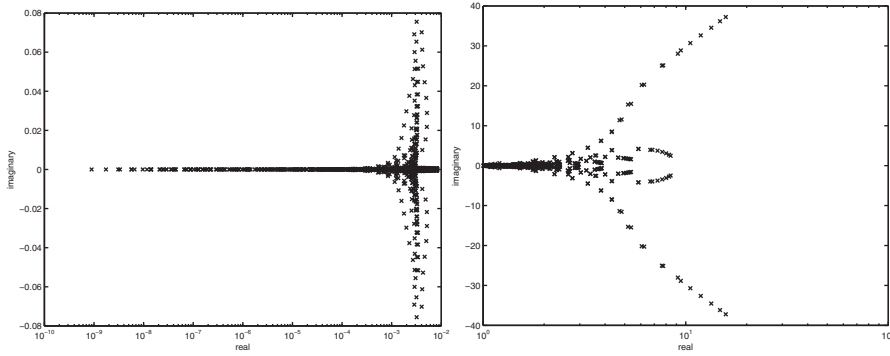


FIG. 5.1. Spectral distribution of Example 5.1. $\varepsilon = 10^{-3}$ and $n = 1089$; without preconditioning (left), with the preconditioner M (right); and the Newton method is applied.

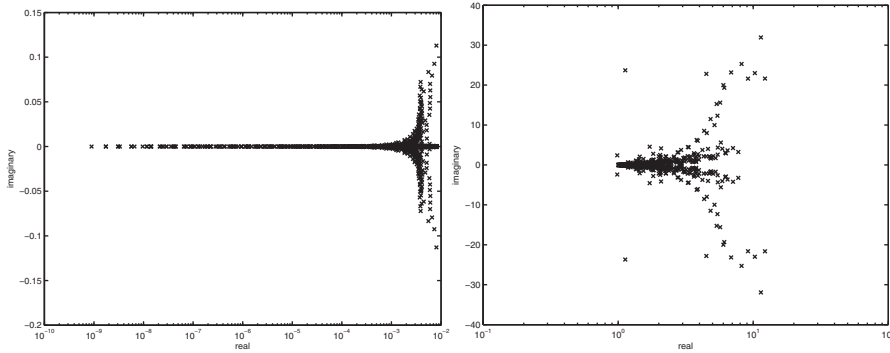


FIG. 5.2. Spectral distribution of Example 5.2. $\varepsilon = 10^{-3}$ and $n = 1089$; without preconditioning (left), with the preconditioner M (right); and the fixed-point method is applied.

From these tables, we see that the new preconditioner can considerably improve the convergence properties of both Newton and fixed-point iteration methods and greatly reduce the running times. Moreover, with increasing of the problem size n , the number of the Newton or the fixed-point iteration steps keeps almost the same or increases slowly if the inner iteration solver, i.e., GMRES, is preconditioned by the new preconditioner while GMRES cannot achieve the prescribed tolerance within 1000 iteration steps and, therefore, the Newton or the fixed-point iteration cannot achieve the prescribed tolerance within 50 iteration steps if GMRES without using a preconditioner is employed as the inner iteration solver. Therefore, the new preconditioning method can substantially improve the convergence behaviors of both Newton and fixed-point iterations and, consequently, lead to fast convergent nonlinear solvers for the systems of nonlinear equations (1.2) arising in the sinc-Galerkin discretization of the time-dependent partial differential equation (1.1).

Figures 5.1 and 5.2 depict the spectral distributions of the original coefficient matrix A and the preconditioned matrix $M^{-1}A$ when the Newton method is applied to Example 5.1 and the fixed-point method is applied to Example 5.2, respectively. The figures clearly show that the matrices without preconditioning are very ill-conditioned and, therefore, the corresponding GMRES method may be convergent very slowly or even divergent, while the matrices with preconditioning are well-conditioned as they

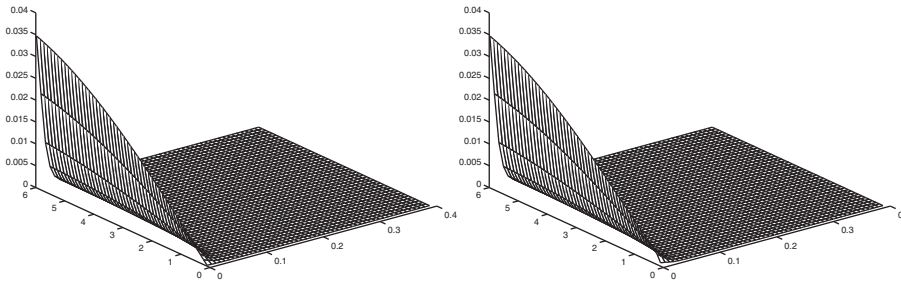


FIG. 5.3. Solutions of Example 5.1. $\varepsilon = 10^{-3}$ and $n = 1089$; exact solution (left), computed solution (right); and the Newton method is applied.

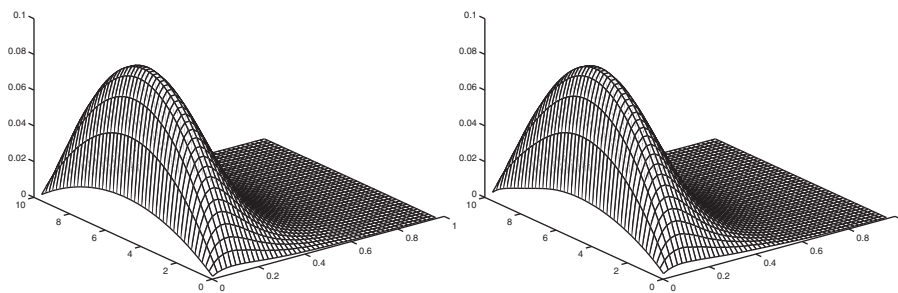


FIG. 5.4. Solutions of Example 5.2. $\varepsilon = 10^{-3}$ and $n = 1089$; exact solution (left), computed solution (right); and the fixed-point method is applied.

have tightly clustered eigenvalues and, thus, the corresponding preconditioned GMRES method may be convergent very quickly to the exact solutions of the subsystems of linear equations. As a result, the preconditioned GMRES method used as the inner linear solver may lead to a fast convergent Newton or fixed-point method for solving the sinc-Galerkin nonlinear systems of the form (1.2).

In Figures 5.3 and 5.4, we plot the exact and the computed solutions of Examples 5.1 and 5.2 corresponding to the cases shown in Figures 5.1 and 5.2, respectively, where the computed solution is obtained by using either the Newton or the fixed-point method. It is clear from Figures 5.3 and 5.4 that the new preconditioned iteration methods can compute reasonably accurate results.

6. Concluding remarks. We have constructed a structured preconditioner that can efficiently improve the convergence property of the GMRES iteration employed to inexactly solve the subsystem of linear equations involved in each Newton or fixed-point iteration for solving the system of nonlinear equations resulting from the sinc-Galerkin discretization of the time-dependent partial differential equation (1.1). The bounds of the eigenvalues of the preconditioned matrix were precisely estimated by making use of the generalized Bendixson theorem, which, in particular, can lead to sharper eigenvalue bounds than those derived in [5] for the preconditioned matrix arising from the sinc-Galerkin discretization of the Burgers equation. Numerical experiments have shown the effectiveness of this new preconditioning method.

REFERENCES

- [1] Z.-Z. BAI, *Parallel multisplitting AOR method for solving a class of system of nonlinear algebraic equations*, Appl. Math. Mech., 16 (1995), pp. 675–682.
- [2] Z.-Z. BAI, *Parallel nonlinear AOR method and its convergence*, Comput. Math. Appl., 31 (1996), pp. 21–31.
- [3] Z.-Z. BAI, G.H. GOLUB, L.-Z. LU, AND J.-F. YIN, *Block triangular and skew-Hermitian splitting methods for positive-definite linear systems*, SIAM J. Sci. Comput., 26 (2005), pp. 844–863.
- [4] Z.-Z. BAI, G.H. GOLUB, AND M.K. NG, *Hermitian and skew-Hermitian splitting methods for non-Hermitian positive definite linear systems*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 603–626.
- [5] Z.-Z. BAI, Y.-M. HUANG, AND M.K. NG, *On preconditioned iterative methods for Burgers equations*, SIAM J. Sci. Comput., 29 (2007), pp. 415–439.
- [6] Z.-Z. BAI AND M.K. NG, *Preconditioners for nonsymmetric block Toeplitz-like-plus-diagonal linear systems*, Numer. Math., 96 (2003), pp. 197–220.
- [7] Z.-Z. BAI AND D.-R. WANG, *Asynchronous multisplitting nonlinear Gauss-Seidel type method*, Appl. Math. J. Chinese Univ. Ser. B, 9 (1994), pp. 189–194.
- [8] Z.-Z. BAI AND D.-R. WANG, *Asynchronous parallel multisplitting nonlinear Gauss-Seidel iteration*, Appl. Math. Chinese Univ. Ser. B, 12 (1997), pp. 179–194.
- [9] R.H. CHAN AND X.-Q. JIN, *A family of block preconditioners for block systems*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 1218–1235.
- [10] R.H. CHAN, W.-F. NG, AND H.-W. SUN, *Fast construction of optimal circulant preconditioners for matrices from the fast dense matrix method*, BIT, 40 (2000), pp. 24–40.
- [11] X.-Q. JIN, *A note on preconditioned block Toeplitz matrices*, SIAM J. Sci. Comput., 16 (1995), pp. 951–955.
- [12] X.-Q. JIN, *Band Toeplitz preconditioners for block Toeplitz systems*, J. Comput. Appl. Math., 70 (1996), pp. 225–230.
- [13] X.-Q. JIN, *Developments and Applications of Block Toeplitz Iterative Solvers*, Kluwer Academic Publishers, Dordrecht Science Press, Beijing, 2002.
- [14] T. KAILATH AND A.H. SAYED, *Displacement structure: Theory and applications*, SIAM Rev., 37 (1995), pp. 297–386.
- [15] N. LEVINSON, *The Wiener RMS (root mean square) error criterion in filter design and prediction*, J. Math. Phys. Mass. Inst. Tech., 25 (1947), pp. 261–278.
- [16] J. LUND AND K.L. BOWERS, *Sinc Methods for Quadrature and Differential Equations*, SIAM, Philadelphia, 1992.
- [17] M.K. NG, *Fast iterative methods for symmetric sinc-Galerkin systems*, IMA J. Numer. Anal., 19 (1999), pp. 357–373.
- [18] M.K. NG AND D. POTTS, *Fast iterative methods for sinc systems*, SIAM J. Matrix Anal. Appl., 24 (2002), pp. 581–598.
- [19] J.M. ORTEGA AND W.C. RHEINOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, London, 1970.
- [20] Y. SAAD, *Iterative Methods for Sparse Linear Systems*, PWS, Boston, 1996.
- [21] D.-R. WANG, Z.-Z. BAI, AND D.J. EVANS, *Asynchronous multisplitting relaxed iterations for weakly nonlinear systems*, Int. J. Comput. Math., 54 (1994), pp. 57–76.
- [22] D.-R. WANG, Z.-Z. BAI, AND D.J. EVANS, *On the monotone convergence of multisplitting method for a class of system of weakly nonlinear equations*, Int. J. Comput. Math., 60 (1996), pp. 229–242.

CONVERGENCE ANALYSIS OF A DISCONTINUOUS GALERKIN METHOD WITH PLANE WAVES AND LAGRANGE MULTIPLIERS FOR THE SOLUTION OF HELMHOLTZ PROBLEMS*

MOHAMED AMARA[†], RABIA DJELLOULI[‡], AND CHARBEL FARHAT[§]

Abstract. We analyze the convergence of a discontinuous Galerkin method (DGM) with plane waves and Lagrange multipliers that was recently proposed by Farhat, Harari, and Hetmaniuk [*Comput. Methods Appl. Mech. Engrg.*, 192 (2003), pp. 1389–1419] for solving two-dimensional Helmholtz problems at relatively high wavenumbers. We prove that the underlying hybrid variational formulation is well-posed. We also present various a priori error estimates that establish the convergence and order of accuracy of the simplest element associated with this method. We prove that, for $k(kh)^{\frac{2}{3}}$ sufficiently small, the *relative* error in the L^2 -norm (resp. in the H^1 seminorm) is of order $k(kh)^{\frac{4}{3}}$ (resp. of order $(kh)^{\frac{2}{3}}$) for a solution being in $H^{\frac{5}{3}}(\Omega)$. In addition, we establish an a posteriori error estimate that can be used as a practical error indicator when refining the partition of the computational domain.

Key words. acoustic scattering, discontinuous Galerkin, Helmholtz problems, hybrid finite element, inf-sup condition, plane waves

AMS subject classifications. 65N12, 65N15, 35J05, 65N30, 74J20, 35Q60, 39A12, 78A45

DOI. 10.1137/060673230

Introduction. The discontinuous enrichment method (DEM) was developed in [1, 2] for the solution of multiscale boundary value problems (BVPs) with sharp gradients and rapid oscillations. These are problems for which the standard finite element method (FEM) can become prohibitively expensive. DEM can be described as a discontinuous Galerkin method (DGM) with Lagrange multiplier degrees of freedom (DOFs), in which the standard finite element polynomial field is enriched within each element by free-space solutions of the homogeneous partial differential equation to be solved. Usually, these are easily obtained in *analytical* form and are discontinuous across the element interfaces. The Lagrange multiplier DOFs are introduced at these interfaces to enforce a weak continuity of the solution. For the Helmholtz equation, the enrichment field can be constructed with plane waves, as these are free-space solutions of this equation. In [3], it was shown that for a large class of Helmholtz problems, the polynomial field is not necessary for efficiently capturing the solution. Hence, for these applications, the polynomial field was dropped, and the DEM was transformed into a

*Received by the editors November 13, 2006; accepted for publication (in revised form) October 16, 2008; published electronically February 13, 2009.

<http://www.siam.org/journals/sinum/47-2/67323.html>

[†]Laboratoire de Mathématiques Appliquées, Université de Pau et des Pays de l'Adour et CNRS-UMR5142, BP 1155, 64013 Pau cedex, France (Mohamed.Amara@univ-pau.fr).

[‡]Corresponding author. Department of Mathematics, California State University Northridge, Northridge, CA 91330-8313 (rabia.djellouli@csun.edu). This author's research was partially supported by the National Science Foundation (NSF) under grant DMS-0406617 and by the Office of Naval Research (ONR) under grant N-00014-01-1-0356. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF or the ONR.

[§]Department of Mechanical Engineering and Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305 (cfarhat@stanford.edu). This author's research was partially supported by the National Science Foundation (NSF) under grant DMS-0406617 and by the Office of Naval Research (ONR) under grant N-00014-01-1-0356. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF or the ONR.

DGM with plane wave basis functions. Similar exponential functions were previously introduced in the weak element method (WEM) [4], the partition of unity method (PUM) [5], the ultra weak variational method [6], and the least-squares method (LSM) presented in [7] for the solution of the Helmholtz equation. However, unlike WEM, the DGM proposed in [3] is based on a variational framework, and unlike PUM, it is discontinuous. Furthermore, in contrast to LSM, the continuity of the solution at the interelement boundaries is enforced in DEM by Lagrange multipliers rather than penalty parameters, which increases the robustness and accuracy of the underlying framework of approximation.

In [3], two lower-order rectangular DGM elements with four and eight plane waves, respectively, were constructed and applied to the solution of two-dimensional waveguide problems with $10 \leq kl \leq 100$, where k denotes the wavenumber and l is a characteristic length of the waveguide. The discretization by these elements of such Helmholtz problems was found to require five to seven times fewer DOFs than their discretization by the standard $Q2$ element, depending on the desired level of accuracy. In [8], this DGM was extended to exterior Helmholtz problems and was coupled with a second-order absorbing boundary condition. A lower-order quadrilateral element with eight Lagrange multiplier DOFs was designed and highlighted with the solution on unstructured meshes of sample acoustic scattering problems with $20 \leq kl \leq 40$, where l denotes a characteristic length of the scatterer. This element was shown to deliver significant improvement over the performance of the standard and comparable $Q2$ element. In [9], two higher-order quadrilateral DGM elements with 16 and 32 plane waves, respectively, were presented. The DGM element with 16 plane waves has a computational complexity that is comparable to that of the standard $Q4$ element and was shown numerically to have the same convergence rate with respect to the mesh size. However, this DGM element was also shown numerically in [8] to deliver the same level of accuracy as $Q4$ using six times fewer DOFs. All of these performance results highlight the potential of the DGM introduced in [3] and expanded in [8] and [9].

However, no mathematical analysis of this method has been performed yet. The objective of this paper is to fill this gap in the specific context of the two-dimensional low-order element with four plane waves in order to set this DGM method on a firm theoretical basis. The proposed study assumes that the computational domain Ω is a polygonal-shaped domain that can be partitioned into rectangular elements. Note that the computational domain Ω may have reentrant corners, and therefore, the considered acoustic scattered field is in $H^{\frac{2}{3}}(\Omega)$ only. We partition the computational domain into *rectangular*-shaped elements and consider the case of the so-called R-4-1 element, that is, we approximate *locally* the primal variable by four plane waves and the dual variable by constants on the edges of interior elements. We must point out that this study cannot be extended—at this time—to higher-order elements because it assumes that the normal derivative of the primal variable is constant along the interior edges. This *crucial* property is valid *only* in the case of the R-4-1 element. We prove that for $k(kh)^{\frac{2}{3}}$ small enough, the *relative* error in the L^2 -norm (resp. in the H^1 seminorm) is of order $k(kh)^{\frac{4}{3}}$ (resp. $(kh)^{\frac{2}{3}}$). We recall that in the case of the standard FEM using P_1 element (see [10, 11]), it has been established that for k^2h small enough, the relative error in the L^2 -norm (resp. in the H^1 seminorm) is of order k^3h^2 (resp. kh). Moreover, if we assume that kh is small enough, it has been established in [11] that the relative error for both the L^2 -norm and the H^1 seminorm are bounded by $k(kh)^2$. However, all these error estimates have been established assuming that the scattered field is in $H^2(\Omega)$, which is not a realistic assumption for most applications. We must also point out that, to the best of our knowledge, no

error estimates have been derived yet in the particular case of the $Q4$ finite element when applied to Helmholtz problems.

We also derive a posteriori error estimate that can be used as a practical error indicator when refining the partition of the computational domain. This error estimate reveals that the relative error in the L^2 -norm depends on the errors in the approximation of the interior and exterior boundary conditions as well as on the jump across the elements of the partition.

The remainder of this paper is organized as follows. In section 1, we specify the notations and assumptions used in this paper, state the formulation of a two-dimensional acoustic scattering problem in a bounded domain, and prove that the hybrid problem obtained by applying the DGM introduced above to the solution of the focus Helmholtz problem is well-posed in the sense of Hadamard [12]. More specifically, we introduce Theorem 1 to address the issues of existence, uniqueness, and stability of the DGM formulation. Next, we devote section 3 to the analysis of the discrete solution obtained with a DGM element with four plane waves. More specifically, we recall in section 3.2 the discrete DGM formulation and announce the main results of this paper. These are existence and uniqueness results, a priori error estimates that are stated in Theorem 2, and an a posteriori estimate that is stated in Theorem 3. The proofs of these three sets of fundamental results are detailed in sections 3.3 and 3.4. Finally, section 4 concludes this paper.

1. Preliminaries. We consider throughout this paper the acoustic scattering problem by a *sound-hard* scatterer [13] formulated in a bounded domain as follows:

$$(1.1) \quad (\text{BVP}) \left\{ \begin{array}{ll} \text{Find } u \in H^1(\Omega) \text{ such that} & \\ \Delta u + k^2 u = 0 & \text{in } \Omega, \\ \partial_n u = -\partial_n e^{ik\mathbf{x} \cdot \mathbf{d}} & \text{on } \Gamma, \\ \partial_n u = iku & \text{on } \Sigma, \end{array} \right.$$

where u is the scattered field and Ω is the computational domain. Ω is a *bounded* polygonal-shaped domain that can be partitioned into rectangular elements. Γ is its interior boundary, and Σ is the exterior boundary. \mathbf{n} is the unitary outward normal vector to the boundaries Γ and Σ , and ∂_n is the normal derivative. k is a positive number representing the wavenumber. \mathbf{d} is a unit vector representing the direction of the incident plane wave. The equation on Γ is the Neumann boundary condition that characterizes the sound-hard property of the scatterer. We must point out that the interior Neumann boundary condition on Γ and the exterior condition on Σ are used only for simplicity. The results presented herein apply to all types of admissible boundary conditions. In addition, as it is well-known, one should use *higher-order* local absorbing boundary conditions for solving practical problems.

2. The continuous hybrid variational formulation.

2.1. Nomenclature and properties. We use throughout this paper the following notations and properties.

- K is a *rectangular*-shaped element of Ω and ∂K is its boundary. $\partial K = \bigcup_{j=1}^4 T_K^j$, where T_K^j is the j th edge of K with vertices $(\mathbf{s}_j^K, \mathbf{s}_{j+1}^K)$ and \mathbf{n}_j^K its outward unitary normal vector.

- h_j^K is the length of the edge T_K^j , and $h_K = \max_{1 \leq j \leq 4} h_j^K$.
- $(\mathcal{T}_h)_h$ is a regular triangulation of the computational domain $\bar{\Omega}$ into elements K , i.e.,

$$\exists \hat{c} > 0 / \forall h, \forall K \in \mathcal{T}_h ; h_K^2 \leq \hat{c}|K|,$$

where $|K|$ denotes the area of the element K [14]. Note that $(\mathcal{T}_h)_h$ is a quasi-uniform triangulation, since its elements K are rectangles.

- $h = \max_{K \in \mathcal{T}_h} h_K$. We also assume that $kh \leq \pi$. This condition means that there is at least two elements per wavelength.
- X is the space of the primal variable. X is given by

$$X = \{v \in L^2(\Omega); \forall K \in \mathcal{T}_h, v_K = v|_K \in H^1(K)\} \approx \prod_{K \in \mathcal{T}_h} H^1(K)$$

and is equipped with the following norm:

$$\|v\|_X = \left(\sum_{K \in \mathcal{T}_h} \|v_K\|_{X(K)}^2 \right)^{\frac{1}{2}} \quad \forall v \in X,$$

where

$$\|v_K\|_{X(K)} = \left(|v_K|_{1,K}^2 + \frac{1}{|K|} \|v_K\|_{0,K}^2 \right)^{\frac{1}{2}}.$$

$\|\cdot\|_{0,K}$ (resp. $|\cdot|_{1,K}$) is the L^2 -norm (resp. seminorm) on the element K .

- $|\cdot|_{1,\mathcal{T}_h}$ is the seminorm in the space X defined by

$$|v|_{1,\mathcal{T}_h} = \left(\sum_{K \in \mathcal{T}_h} |v_K|_{1,K}^2 \right)^{\frac{1}{2}} \quad \forall v \in X.$$

- $H^{\frac{1}{2}}(\partial K)$ is the space of the traces of elements of $H^1(K)$, and $H^{-\frac{1}{2}}(\partial K)$ is the dual space of $H^{\frac{1}{2}}(\partial K)$. $H^{\frac{1}{2}}(\partial K)$ is equipped with the following norm:

$$(2.1) \quad \|\lambda\|_{\frac{1}{2},\partial K} = \inf_{w \in W(\lambda)} \|w\|_{X(K)} = \|\Lambda\|_{X(K)},$$

where $W(\lambda) = \{w \in H^1(K) ; w|_{\partial K} = \lambda\}$ and Λ is the unique element in $W(\lambda)$ satisfying

$$-\Delta\Lambda + \frac{1}{|K|}\Lambda = 0 \quad \text{a.e. in } K.$$

It follows from the definition of the norm $\|\cdot\|_X$ and (2.1) that

$$(2.2) \quad \|v\|_{\frac{1}{2},\partial K} \leq \|v\|_{X(K)} \quad \forall v \in H^1(K).$$

- M is the space of the dual variable defined by

$$M = \left\{ \mu \in \prod_{K \in \mathcal{T}_h} H^{-\frac{1}{2}}(\partial K) ; \forall \lambda \in T, \sum_{K \in \mathcal{T}_h} \langle \mu^K, \lambda^K \rangle_{-\frac{1}{2} \times \frac{1}{2}, \partial K} = 0 \right\},$$

where $\mu^K = \mu|_{\partial K}$ and the space T is given by

$$T = \left\{ \lambda \in \prod_{K \in \mathcal{T}_h} H^{\frac{1}{2}}(\partial K); \forall K \neq K' \in \mathcal{T}_h, \lambda^K = \lambda^{K'} \text{ on } \partial K \cap \partial K' \right\}.$$

The space M is equipped with the following norm:

$$\|\mu\|_M = \left(\sum_{K \in \mathcal{T}_h} \|\mu^K\|_{-\frac{1}{2}, \partial K}^2 \right)^{\frac{1}{2}} \quad \forall \mu \in M,$$

where

$$\begin{aligned} \|\mu^K\|_{-\frac{1}{2}, \partial K} &= \sup_{\lambda \in H^{\frac{1}{2}}(\partial K)} \frac{|\langle \mu^K, \lambda \rangle_{-\frac{1}{2} \times \frac{1}{2}, \partial K}|}{\|\lambda\|_{\frac{1}{2}, \partial K}} \\ (2.3) \qquad &= \sup_{v \in H^1(K)} \frac{|\langle \mu^K, v \rangle_{-\frac{1}{2} \times \frac{1}{2}, \partial K}|}{\|v\|_{X(K)}} \end{aligned}$$

and $\langle \cdot, \cdot \rangle_{-\frac{1}{2} \times \frac{1}{2}, \partial K}$ is the duality product between $H^{-\frac{1}{2}}(\partial K)$ and $H^{\frac{1}{2}}(\partial K)$ [15].

- \mathcal{M} is a subspace of M defined by

$$\mathcal{M} = \left\{ \mu \in \prod_{K \in \mathcal{T}_h} L^2(\partial K); \mu = 0 \text{ on } \partial\Omega \text{ and } \forall K \neq K' \in \mathcal{T}_h, \mu^K + \mu^{K'} = 0 \text{ on } \partial K \cap \partial K' \right\}.$$

Therefore, we have

$$\mathcal{M} = M \cap \prod_{K \in \mathcal{T}_h} L^2(\partial K).$$

2.2. Formulation and mathematical results. We adopt the following hybrid-type variational formulation (VP) for solving the BVP. Note that the VP is equivalent to BVP as indicated in Remark 1.

$$(2.4) \quad (\text{VP}) \begin{cases} \text{Find } (u, \lambda) \in X \times M \text{ such that} \\ a(u, v) + b(v, \lambda) = F(v) & \forall v \in X, \\ b(u, \mu) = 0 & \forall \mu \in M, \end{cases}$$

where the *bilinear* forms $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ and the function F are given by

$$\begin{aligned} a(u, v) &= \sum_{K \in \mathcal{T}_h} \left(\int_K \nabla u \cdot \nabla \bar{v} \, dx - k^2 \int_K u \bar{v} \, dx - ik \int_{\partial K \cap \Omega} u \bar{v} \, dt \right) \quad \forall u, v \in X, \\ b(v, \mu) &= \sum_{K \in \mathcal{T}_h} \langle \mu^K, \bar{v} \rangle_{-\frac{1}{2} \times \frac{1}{2}, \partial K} \quad \forall (v, \mu) \in X \times M, \\ F(v) &= - \sum_{K \in \mathcal{T}_h} \int_{\partial K \cap \Gamma} \bar{v} \partial_n e^{ik\mathbf{x} \cdot \mathbf{d}} \, dt \quad \forall v \in X. \end{aligned}$$

Note that the bilinear form $b(\cdot, \cdot)$ also satisfies

$$b(v, \mu) = \sum_{K \in \mathcal{T}_h} \int_{\partial K} \mu^K \bar{v} \, dt \quad \forall (v, \mu) \in X \times \mathcal{M}.$$

In addition, the bilinear forms $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ satisfy the following important properties.

PROPERTY 1. *The bilinear forms $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ are continuous on $X \times X$ and $X \times M$, respectively. Furthermore, we have the following:*

(i) *$a(\cdot, \cdot)$ satisfies the Gårding inequality in $H^1(\Omega)$*

$$(2.5) \quad \Re a(v, v) + k^2 \|v\|_{0,\Omega}^2 = |v|_{1,\mathcal{T}_h}^2 \quad \forall v \in X,$$

where \Re designates the real part.

(ii) *The null space \mathcal{N} corresponding to the bilinear form $b(\cdot, \cdot)$ is given by*

$$(2.6) \quad \mathcal{N} = \{v \in X ; \quad b(v, \mu) = 0 \quad \forall \mu \in M\} = H^1(\Omega).$$

(iii) *The bilinear form $b(\cdot, \cdot)$ satisfies the so-called inf-sup condition [21]:*

$$(2.7) \quad \forall \mu \in M, \quad \exists \phi \in X : \quad \sup_{v \in X} \frac{|b(v, \mu)|}{\|v\|_X} = \frac{|b(\phi, \mu)|}{\|\phi\|_X} = \|\mu\|_M.$$

Proof of Property 1. We prove only the third point, since the proof of (2.5) and (2.6) is straightforward. From the continuity of the bilinear form $b(\cdot, \cdot)$, we deduce that

$$(2.8) \quad \sup_{v \in X} \frac{|b(v, \mu)|}{\|v\|_X} \leq \|\mu\|_M \quad \forall \mu \in M.$$

Next, for a fixed $\mu \in M$, we consider the function $\phi \in X$ such that, for every $K \in \mathcal{T}_h$, $\phi|_K = \phi_K$ is the *unique* solution of the following variational problem:

$$(2.9) \quad \int_K \nabla \phi_K \cdot \nabla \bar{v} \, dx + \frac{1}{|K|} \int_K \phi_K \bar{v} \, dx = \langle \mu^K, \bar{v} \rangle_{-\frac{1}{2} \times \frac{1}{2}, \partial K} \quad \forall v \in H^1(K).$$

Hence, using (2.2) and (2.9), we have

$$\|\phi_K\|_{X(K)}^2 = \langle \mu^K, \overline{\phi_K} \rangle_{-\frac{1}{2} \times \frac{1}{2}, \partial K} \leq \|\mu^K\|_{-\frac{1}{2}, \partial K} \|\phi_K\|_{\frac{1}{2}, \partial K} \leq \|\mu^K\|_{-\frac{1}{2}, \partial K} \|\phi_K\|_{X(K)}.$$

Thus, we deduce that $\|\phi_K\|_{X(K)} \leq \|\mu^K\|_{-\frac{1}{2}, \partial K}$ and then $\|\phi\|_X \leq \|\mu\|_M$.

Moreover, from (2.3) and (2.9), we have $\|\mu^K\|_{-\frac{1}{2}, \partial K} \leq \|\phi_K\|_{X(K)}$.

Therefore, it follows that $\|\phi\|_X = \|\mu\|_M$.

On the other hand, from (2.9) and the definition of the bilinear form $b(\cdot, \cdot)$, we also have

$$b(\phi, \mu) = \sum_{K \in \mathcal{T}_h} \|\phi_K\|_{X(K)}^2 = \|\phi\|_X^2 = \|\phi\|_X \|\mu\|_M,$$

which concludes the proof of the *inf-sup* condition given by (2.7). □

Remark 1. The problems BVP and VP are equivalent in the following sense:

- (i) If the pair (u, λ) is a solution of VP, then it follows from the second equation of VP that u is in $H^1(\Omega)$. Moreover, using the first equation of VP with test functions $v \in \mathcal{D}(\Omega)$, we deduce that u is the solution of the first equation of BVP. Last, the use of test functions $v \in H^1(\Omega)$ allows us to verify that u satisfies the boundary conditions on Γ and Σ .
- (ii) If u is the solution of BVP, then from the standard regularity results for Laplace’s operator [22] and due to the possible reentrant corners (with a measure angle of $\frac{3\pi}{2}$), it follows that $u \in H^{\frac{5}{3}}(\Omega)$. Thus, $\partial_n u^K \in L^2(\partial K) \forall K \in \mathcal{T}_h$ ($\partial_n u^K$ is even in $H^{\frac{1}{6}}(\partial K)$). Then we set

$$(2.10) \quad \lambda^K = \begin{cases} -\partial_n u & \text{on } \partial K \setminus \partial\Omega, \\ 0 & \text{on } \partial K \cap \partial\Omega. \end{cases}$$

Therefore, the dual variable λ satisfies (2.10) in the $L^2(\partial K)$ sense, which is the *classical* sense. Having that in mind, one can multiply BPV by test functions $v \in X$ and deduce that the pair (u, λ) satisfies VP.

Next, we prove that the variational problem (VP) is well-posed in the sense of Hadamard [12]. This is main result of this section. It is stated in the following theorem.

THEOREM 1. *The variational problem (VP) admits a unique solution $(u, \lambda) \in X \times M$. In addition, u belongs to $H^{\frac{5}{3}}(\Omega)$ and for all $\theta \in [0, \frac{5}{3}]$, there is a positive constant C (C depends on Ω and θ only) such that*

$$|u|_{\theta, \Omega} \leq C(1+k)^\theta.$$

The proof of this theorem is based on the following intermediate stability result.

LEMMA 1. *Let f be in $L^2(\Omega)$. Then, the following BVP*

$$(2.11) \quad \begin{cases} \Delta U + k^2 U = f & \text{in } \Omega, \\ \partial_n U = 0 & \text{on } \Gamma, \\ \partial_n U = ikU & \text{on } \Sigma, \end{cases}$$

has one and only one solution U in $H^{\frac{5}{3}}(\Omega)$. Moreover, for all $\theta \in [0, \frac{5}{3}]$, there is a positive constant C (C depends on Ω and θ only) such that

$$(2.12) \quad |U|_{\theta, \Omega} \leq C(1+k)^{\theta-1} \|f\|_{0, \Omega}.$$

Proof of Lemma 1. First, observe that the variational formulation corresponding to the BVP (2.11) is given by

$$(2.13) \quad \begin{cases} \text{Find } U \in H^1(\Omega) \text{ such that} \\ a(U, v) = - \int_{\Omega} f \bar{v} dx \quad \forall v \in H^1(\Omega). \end{cases}$$

From (2.5), it follows that the bilinear form $a(\cdot, \cdot)$ satisfies the Fredholm alternative on $H^1(\Omega)$. Hence, the uniqueness ensures the existence of the solution U in $H^1(\Omega)$.

Therefore, we need only to prove the uniqueness of the solution of the BVP (2.11). Let w be the solution of the corresponding homogeneous BVP. The function w satisfies

$$a(w, w) = 0 \quad \text{then } w = 0 \quad \text{on } \Sigma,$$

and we deduce that

$$\partial_n w = 0 \quad \text{on } \Gamma \quad \text{and} \quad w = \partial_n w = 0 \quad \text{on } \Sigma.$$

Therefore, using the continuation theorem [16, 17], we obtain that $w = 0$ in Ω .

From the standard regularity results for second-order elliptic BVPs [22] and due to the possible reentrant corners (with a measure angle of $\frac{3\pi}{2}$), it follows that the solution of problem (2.11) satisfies $U \in H^{\frac{5}{3}}(\Omega)$, and there is a positive constant C (C depends on Ω only) such that

$$(2.14) \quad \|U\|_{\frac{5}{3},\Omega} \leq C \left(\|\Delta U\|_{-\frac{1}{3},\Omega} + \|\partial_n U\|_{\frac{1}{6},\partial\Omega} \right).$$

Moreover, using the results established in [18] and [19], we deduce the existence of a positive constant C (C depends on Ω only) such that

$$(2.15) \quad \|U\|_{0,\Omega} \leq \frac{C}{1+k} \|f\|_{0,\Omega} \quad \text{and} \quad |U|_{1,\Omega} \leq C \|f\|_{0,\Omega}.$$

Next, we establish the estimate (2.12). To do this, we will use the space interpolation results in [20]. First, using boundary conditions in BVP (2.11), we deduce that there is a positive constant C (C depends on Ω only) such that

$$\|\partial_n U\|_{\frac{1}{6},\partial\Omega} = \|\partial_n U\|_{\frac{1}{6},\Sigma} = k \|U\|_{\frac{1}{6},\Sigma} \leq C k \|U\|_{\frac{2}{3},\Omega}.$$

Therefore, it follows from the space interpolation results in [20] that there is a positive constant C (C depends on Ω only) such that

$$\|\partial_n U\|_{\frac{1}{6},\partial\Omega} \leq C k \|U\|_{0,\Omega}^{\frac{1}{3}} |U|_{1,\Omega}^{\frac{2}{3}}.$$

Finally, it follows from (2.15) that there exists a positive constant C (C depends on Ω only) such that

$$(2.16) \quad \|\partial_n U\|_{\frac{1}{6},\partial\Omega} \leq C (1+k)^{\frac{2}{3}} \|f\|_{0,\Omega}.$$

Furthermore, from the first equation of BVP (2.11), we deduce that

$$\|\Delta U\|_{0,\Omega} \leq k^2 \|U\|_{0,\Omega} + \|f\|_{0,\Omega}.$$

Hence, it follows from (2.15) that there is a positive C (C depends on Ω only) such that

$$\|\Delta U\|_{0,\Omega} \leq C (1+k) \|f\|_{0,\Omega}.$$

In addition, from the norms properties and (2.15), there is a positive C (C depends on Ω only) such that

$$\|\Delta U\|_{-1,\Omega} \leq |U|_{1,\Omega} \leq \|U\|_{1,\Omega} \leq C \|f\|_{0,\Omega}.$$

Consequently, it follows from these equations and the interpolation space results theorem (see [20]) that there is a positive constant C (C depends on the domain Ω only) such that

$$(2.17) \quad \|\Delta U\|_{-\frac{1}{3},\Omega} \leq C (1+k)^{\frac{2}{3}} \|f\|_{0,\Omega}.$$

Estimate (2.12) is then a direct consequence of (2.14), (2.16), and (2.17). □

Proof of Theorem 1. Since $H^1(\Omega)$ is the null space of the bilinear form $b(\cdot, \cdot)$ (see (2.6)), the VP is reduced to the variational problem

$$a(u, v) = F(v) \quad \forall v \in H^1(\Omega).$$

From (2.5), it follows that the bilinear form $a(\cdot, \cdot)$ satisfies the Fredholm alternative on $H^1(\Omega)$. Hence, the uniqueness ensures the existence of the solution u in $H^1(\Omega)$. On the other hand, the uniqueness results readily from the solution of BVP (2.11). Therefore, the solution u of the reduced variational problem in the null space $H^1(\Omega)$ of the bilinear form $b(\cdot, \cdot)$ exists and is unique. Therefore, both existence and uniqueness of the solution of the complete variational problem VP are standard consequences (see, for example, [21]) of the inf-sup condition given by (2.7).

To prove the stability estimates, we first observe that the pair (u, λ) solution of the variational formulation (VP) satisfies the following *mixed* BVP:

$$\begin{cases} \Delta u + k^2 u = 0 & \text{in } \Omega, \\ \partial_n u = -\partial_n e^{ik\mathbf{x}\cdot\mathbf{d}} & \text{on } \Gamma, \\ \partial_n u = ik u & \text{on } \Sigma, \end{cases}$$

and $\forall K \in \mathcal{T}_h$, we have

$$\lambda^K = \begin{cases} -\partial_n u & \text{on } \partial K \setminus \partial\Omega, \\ 0 & \text{on } \partial K \cap \partial\Omega. \end{cases}$$

Consequently, if we set

$$(2.18) \quad U = u + e^{ik\mathbf{x}\cdot\mathbf{d}} \phi$$

where $\phi \in \mathcal{D}(\overline{\Omega})$ satisfies

$$\phi = 1 \text{ on } \Gamma, \quad \partial_n \phi = 0 \text{ on } \Gamma, \quad \phi = \partial_n \phi = 0 \text{ on } \Sigma,$$

then it is easy to verify that U is the unique solution of BVP (2.11) with the right-hand side f given by

$$f = (2ik \mathbf{d} \cdot \nabla \phi + \Delta \phi) e^{ik\mathbf{x}\cdot\mathbf{d}},$$

and there is a positive constant C (C depends on Ω only) such that

$$\|f\|_{0,\Omega} \leq C(1+k).$$

Therefore, the proof of Theorem 1's estimate is an immediate consequence of estimate (2.12) in Lemma 1, which concludes the proof of Theorem 1. \square

3. The discrete formulation.

3.1. Assumptions, notations, and properties. We adopt, throughout this section, the following notations and properties.

- $\forall K \in \mathcal{T}_h, \phi_j^K = e^{ik \mathbf{n}_j^K \cdot (\mathbf{x} - \mathbf{s}_j^K)}$; $1 \leq j \leq 4$.
- X_h is the discrete space for the *primal* variable. X_h is given by

$$X_h = \{v_h \in X; \forall K \in \mathcal{T}_h, v_h|_K \in X_h(K)\},$$

where

$$X_h(K) = \left\{ v_h^K \in H^1(K) ; v_h^K = \sum_{j=1}^4 \alpha_j^K \phi_j^K, \quad \text{where } \alpha_j^K \in \mathbb{C} \right\}.$$

Note that $X_h \subseteq X$, and therefore, X_h is also equipped with the norm $\|\cdot\|_X$.

- M_h is the discrete space of the *dual* variable. M_h is defined as follows:

$$M_h = \left\{ \mu_h \in \mathcal{M}; \forall K \in \mathcal{T}_h \text{ and } \forall T_j^K \subset \partial K : \mu_j^K = \mu|_{T_j^K} \in \mathbb{C}, 1 \leq j \leq 4 \right\}.$$

- For every $K \in \mathcal{T}_h$, the matrix $B^K = (B_{lj}^K)_{1 \leq l, j \leq 4}$ represents the elementary matrix corresponding to the bilinear form $b(\cdot, \cdot)$. Hence, the entries of the matrix B^K are given by

$$(3.1) \quad B_{lj}^K = \frac{1}{h_l^K} \int_{T_l^K} \phi_j^K dt, \quad 1 \leq l, j \leq 4.$$

- \hat{C} designates a generic positive constant. \hat{C} is independent of k , Ω , and the triangulation \mathcal{T}_h .
- For a given $K \in \mathcal{T}_h$ and $\forall v^K \in H^1(K)$, we have the following two classical inequalities [14]:

$$(3.2) \quad \|v^K\|_{0,\partial K} \leq \hat{C} \left(\frac{1}{h_K} \|v^K\|_{0,K}^2 + h_K |v^K|_{1,K}^2 \right)^{\frac{1}{2}},$$

$$(3.3) \quad \left\| v^K - \frac{1}{|K|} \int_K v^K dx \right\|_{0,K} \leq \hat{C} h_K |v^K|_{1,K}.$$

In addition, it follows from combining (3.2) (when applied to $v^K - \frac{1}{|K|} \int_K v^K dx$) and (3.3) that

$$(3.4) \quad \left\| v^K - \frac{1}{|K|} \int_K v^K dx \right\|_{0,\partial K} \leq \hat{C} h_K^{\frac{1}{2}} |v^K|_{1,K}.$$

3.2. Discrete formulation and announcement of the main results. The discrete variational problem (DVP) corresponding to the variational formulation (VP) can be formulated as follows:

$$(3.5) \quad \text{(DVP)} \quad \begin{cases} \text{Find } (u_h, \lambda_h) \in X_h \times M_h \text{ such that} \\ a(u_h, v_h) + b(v_h, \lambda_h) = F(v_h) & \forall v_h \in X_h, \\ b(u_h, \mu_h) = 0 & \forall \mu_h \in M_h. \end{cases}$$

The next two theorems summarize the main results of this section.

THEOREM 2. *The DVP admits a unique solution $(u_h, \lambda_h) \in X_h \times M_h$.*

Moreover, for $h_0 > 0$ such that $k(1+k)^{\frac{2}{3}} h_0^{\frac{2}{3}}$ is “sufficiently small” and $kh_0 \leq \pi$, there is a positive constant C (C depends on Ω only) such that for all $h \leq h_0$, we have

$$(3.6) \quad \begin{aligned} \|u - u_h\|_{0,\Omega} &\leq C(1+k)^{\frac{7}{3}} h^{\frac{4}{3}}, \\ |u - u_h|_{1,\mathcal{T}_h} + \|\lambda - \lambda_h\|_M &\leq C(1+k)^{\frac{5}{3}} h^{\frac{2}{3}}, \end{aligned}$$

where (u, λ) is the solution of the continuous variational problem VP (2.4).

THEOREM 3. *Let u be the solution of the continuous variational problem VP (2.4) and u_h be the solution of the DVP. We assume that $kh \leq \pi$, then there exists a constant $C > 0$ (C depends on Ω only) such that*

$$(3.7) \quad \|u - u_h\|_{0,\Omega} \leq \hat{C} \left(\left(\sum_{e \in \Sigma} h_e \|\partial_n u_h - ik u_h\|_{0,e}^2 \right)^{\frac{1}{2}} + \left(\sum_{e \in \Gamma} h_e \|\partial_n u_h + \partial_n e^{ikx \cdot d}\|_{0,e}^2 \right)^{\frac{1}{2}} + \left(\sum_{e \text{ interior}} h_e^{-1} \|[u_h]\|_{0,e}^2 \right)^{\frac{1}{2}} \right),$$

where e is an edge of \mathcal{T}_h , $[u_h]$ is the jump of u_h across the edge e , and h_e is the length of e .

Remark 2. We must point out that it has been reported in [10, 11] that for a high-frequency regime, the use of P_1 FEM leads to the following estimates: $|u - u_h|_{1,\Omega} \leq C k^2 h$ and $\|u - u_h\|_{0,\Omega} \leq C k^3 h^2$ when $k^2 h$ is small enough. These estimates were derived assuming that $u \in H^2(\Omega)$, which is not, however, valid for most problems.

The a posteriori estimate given by (3.7) is a practical tool for a mesh adaptive strategy. This estimate reveals that the L^2 error depends on how well the jump of the primal variable as well as the interior and exterior boundary conditions are approximated at the element level. In order to prove Theorems 2 and 3, we need first to establish intermediate interpolation results. This is accomplished in section 3.3. Then, we prove in section 3.4.1 the existence and the uniqueness of the solution of the DVP. This result is established as a direct consequence of Proposition 1 and Proposition 2. Section 3.4.2 is devoted to the proof of (3.6) and (3.7). The error estimate given by (3.6) is established in four steps, each step is formulated as a lemma (see Lemma 7 to Lemma 10). The a posteriori error estimate given by (3.7) is established at the end of section 3.4.2.

The next result, that can be easily established, shows why the existence and the uniqueness of the solution of (DVP) is not a direct consequence of the existence and the uniqueness of the solution of (VP).

LEMMA 2. *The null space \mathcal{N}_h corresponding to the bilinear form $b(\cdot, \cdot)$ defined by*

$$\mathcal{N}_h = \{v_h \in X_h : b(v_h, \mu_h) = 0 ; \quad \forall \mu_h \in M_h\}$$

satisfies

$$(3.8) \quad \mathcal{N}_h = \left\{ v_h \in X_h ; \quad \int_{\partial K \cap \partial K'} v_h^K dt = \int_{\partial K \cap \partial K'} v_h^{K'} dt, \quad \forall K \neq K' \in \mathcal{T}_h \right\}.$$

Remark 3. Lemma 2 states that \mathcal{N}_h is not a subspace of $\mathcal{N} = H^1(\Omega)$, which is the null space of the bilinear form $b(\cdot, \cdot)$. Indeed, the trace of an element of \mathcal{N}_h on an edge of an element K is *weakly* continuous in the sense given by (3.8), while the trace of an element of \mathcal{N} on an edge of an element K is “continuous” almost everywhere. Therefore, the inf-sup condition given by (2.7) and then Theorem 1 are no longer valid if we simply replace X and M by X_h and M_h , respectively.

3.3. Mathematical analysis of the interpolation operators. We establish in this section intermediate interpolation results that summarize the main properties of the *projection operator* Π_h from X onto X_h and the *projection operator* P_h from

\mathcal{M} onto M_h . These results are obtained in the case of a *rectangular*-shaped partition of the computational domain Ω .

3.3.1. Interpolation operator in X_h .

LEMMA 3. For a fixed $K \in \mathcal{T}_h$, we have the following two properties:

- (i) The normal derivative $\partial_n \phi_j^K$ is constant on every edge T_l^K ($1 \leq l, j \leq 4$).
- (ii) If $kh_K \leq \pi$, then the matrix B^K is invertible and there is a positive constant \hat{C} such that

$$(3.9) \quad \left\| (B^K)^{-1} \right\|_2 \leq \frac{\hat{C}}{k^2 h_K^2}.$$

Proof of Lemma 3. It follows from the definition of ϕ_j^K (see section 3.1) that

$$\partial_n \phi_j^K = ik \mathbf{n}_j^K \cdot \mathbf{n}_l^K \phi_j^K \quad \text{on } T_l^K \quad (1 \leq l, j \leq 4).$$

Therefore, since K is a *rectangular*-shaped element, a simple calculation shows that

$$\partial_n \phi_j^K = ik \text{ on } T_j^K, \quad \partial_n \phi_j^K = -ik \text{ on } T_{j+2}^K, \quad \text{and} \quad \partial_n \phi_j^K = 0 \text{ on } T_{j+1}^K \cup T_{j+3}^K.$$

In addition, it follows from the definition of the elementary matrix B^K (see (3.1)) that

$$B^K = \begin{bmatrix} 1 & b_1 & a_2 & b_1 \\ b_2 & 1 & b_2 & a_1 \\ a_2 & b_1 & 1 & b_1 \\ b_2 & a_1 & b_2 & 1 \end{bmatrix},$$

where $a_j = e^{-ikh_j^K}$ and $b_j = \frac{1 - e^{-ikh_j^K}}{ikh_j^K}$, $1 \leq j \leq 4$.

We set $\Delta = (1 + a_1)(1 + a_2) - 4b_1b_2$. Then, it is easy to verify that $\Delta \neq 0$ for $kh_K \leq \pi$ (which is, in fact, a sufficient but not necessary condition). This ensures that the matrix B^K is invertible, and we have

$$[B^K]^{-1} = \frac{1}{2} \begin{bmatrix} \frac{1+a_1}{\Delta} + \frac{1}{1-a_2} & -2\frac{b_1}{\Delta} & \frac{1+a_1}{\Delta} - \frac{1}{1-a_2} & -2\frac{b_1}{\Delta} \\ -2\frac{b_2}{\Delta} & \frac{1+a_2}{\Delta} + \frac{1}{1-a_1} & -2\frac{b_2}{\Delta} & \frac{1+a_2}{\Delta} - \frac{1}{1-a_1} \\ \frac{1+a_1}{\Delta} - \frac{1}{1-a_2} & -2\frac{b_1}{\Delta} & \frac{1+a_1}{\Delta} + \frac{1}{1-a_2} & -2\frac{b_1}{\Delta} \\ -2\frac{b_2}{\Delta} & \frac{1+a_2}{\Delta} - \frac{1}{1-a_1} & -2\frac{b_2}{\Delta} & \frac{1+a_2}{\Delta} + \frac{1}{1-a_1} \end{bmatrix}.$$

Finally, one can verify that there is a positive constant \hat{C} and k such that

$$\left\| [B^K]^{-1} \right\|_2 \leq \frac{\hat{C}}{k^2 h_K^2}. \quad \square$$

Next, we introduce the sequence of linear operators $(\pi_K)_{K \in \mathcal{T}_h}$ defined as follows:

$$\left| \begin{array}{l} \pi_K : H^1(K) \longrightarrow \mathbb{C}^4 \\ v^K \longmapsto \pi_K v^K, \end{array} \right.$$

where

$$(3.10) \quad (\pi_K v^K)_j = \frac{1}{h_j^K} \int_{T_j^K} v^K dt, \quad 1 \leq j \leq 4.$$

Then, it follows from (3.2) that, for any h_K independent vectorial norm $|||\cdot|||$ in \mathbb{C}^4 , there is a positive constant \hat{C} such that

$$(3.11) \quad |||\pi_K v^K||| \leq \hat{C} \|v^K\|_{X(K)} \quad \forall v^K \in H^1(K).$$

In addition, we have

$$(3.12) \quad \forall v_h^K \in X_h(K), \quad v_h^K = \sum_{j=1}^4 \alpha_j^K \phi_j^K, \quad \text{where } \alpha_j^K = \left([B^K]^{-1} \pi_K v_h^K \right)_j, \quad 1 \leq j \leq 4.$$

The next result states that, for a given $K \in \mathcal{T}_h$, the set of DOFs associated to the planar waves $(\phi_j^K)_{j=1}^4$ is *unisolvent*.

LEMMA 4. *For a given $K \in \mathcal{T}_h$ and for any $v_h^K \in X_h(K)$, we have the following equivalence:*

$$\left(\int_{T_l^K} v_h^K dt = 0, \quad 1 \leq l \leq 4 \right) \iff (v_h^K = 0 \text{ on } K).$$

Proof of Lemma 4. Using (3.10) and (3.12), it follows that for a given $K \in \mathcal{T}_h$, we have

$$\int_{T_l^K} v_h^K dt = 0, \quad 1 \leq l \leq 4 \iff \pi_K v_h^K = 0 \iff v_h^K = 0,$$

which proves Lemma 4. \square

Consequently, one can construct a sequence of *local* linear operator $(\Pi_K)_{K \in \mathcal{T}_h}$ as follows:

$$\begin{cases} \Pi_K & : & H^1(K) & \longrightarrow & X_h(K), \\ & & v^K & \longmapsto & \Pi_K v^K, \end{cases}$$

with

$$(3.13) \quad \int_{T_j^K} v^K dt = \int_{T_j^K} \Pi_K v^K dt, \quad 1 \leq j \leq 4.$$

Next, we state three properties of the operator Π_K . These properties are immediate consequences of the definition of Π_K , the inequalities (3.2)–(3.3), property (3.13) of the operator Π_K , and the characterization of elements of $X_h(K)$ with the elementary matrix B^K (see (3.12)). Note that the second identity of (3.14) is obtained by Green’s formula using the rectangular shape of K .

PROPERTY 2. *The operator Π_K satisfies the following three properties:*

(i) $\forall K \in \mathcal{T}_h$ and $\forall v \in H^1(K)$, we have

$$(3.14) \quad \int_{\partial K} (v^K - \Pi_K v^K) dt = 0, \quad \int_K \nabla (v^K - \Pi_K v^K) dx = 0.$$

(ii) *There is a positive constant \hat{C} such that*

$$(3.15) \quad \forall K \in \mathcal{T}_h, \quad \|v^K - \Pi_K v^K\|_{0,\partial K} \leq \hat{C} h_K^{\frac{1}{2}} |v^K - \Pi_K v^K|_{1,K} \quad \forall v^K \in H^1(K).$$

(iii) For a given $v^K \in H^1(K)$, we have

$$(3.16) \quad \pi_K v^K = \pi_K \circ \Pi_K v^K \quad \text{and} \\ \Pi_K v^K = \sum_{j=1}^4 \alpha_j^K \phi_j^K, \quad \text{with} \quad \alpha_j^K = \left([B^K]^{-1} \pi_K v^K \right)_j.$$

Proof of Property 2. We prove only the second property, since the two others are immediate. Using (3.14) and the definition of the norm $\|\cdot\|_{0,\partial K}$, we have

$$\begin{aligned} \|v^K - \Pi_K v^K\|_{0,\partial K} &= \left\| v^K - \Pi_K v^K - \frac{1}{|\partial K|} \int_{\partial K} (v^K - \Pi_K v^K) dt \right\|_{0,\partial K} \\ &\leq \inf_{\beta \in \mathbb{C}} \|v^K - \Pi_K v^K - \beta\|_{0,\partial K} \\ &\leq \|v^K - \Pi_K v^K - \frac{1}{|K|} \int_K (v - \Pi_K v^K) dt\|_{0,\partial K}. \end{aligned}$$

We then conclude using (3.4).

In the next two lemmas, we establish a priori estimates on the operator Π_K .

LEMMA 5. Assume $kh \leq \pi$. Then, there is a positive constant \hat{C} such that $\forall K \in \mathcal{T}_h$ and $\forall v^K \in H^1(K)$, we have

$$(3.17) \quad \|v^K - \Pi_K v^K\|_{0,K} \leq \hat{C} h_K |v^K - \Pi_K v^K|_{1,K},$$

$$(3.18) \quad k \|\Pi_K v^K\|_{0,K} + \|\Pi_K v^K\|_{X(K)} \leq \hat{C} \|v^K\|_{X(K)}.$$

Proof of Lemma 5. We establish the estimate given by (3.17) using the Aubin-Nitsche argument [23, 24, 25].

More specifically, consider the following auxiliary BVP:

$$\begin{cases} \text{Find } \varphi \in H_0^1(K) \text{ such that} \\ -\Delta \varphi = v^K - \Pi_K v^K \quad \text{on } K. \end{cases}$$

Since K is a *rectangular-shaped* element, then φ is, in fact, in $H^2(K) \cap H_0^1(K)$, and we have

$$|\varphi|_{2,K} = \|\Delta \varphi\|_{0,K} = \|v^K - \Pi_K v^K\|_{0,K}.$$

It follows that

$$\|v^K - \Pi_K v^K\|_{0,K}^2 = \int_K \nabla (v^K - \Pi_K v^K) \cdot \nabla \bar{\varphi} dx - \int_{\partial K} (v^K - \Pi_K v^K) \partial_n \bar{\varphi} dt.$$

Using (3.14), we deduce that

$$\left| \int_K \nabla (v^K - \Pi_K v^K) \cdot \nabla \bar{\varphi} dx \right| = \left| \int_K \nabla (v^K - \Pi_K v^K) \cdot \left(\nabla \bar{\varphi} - \frac{1}{|K|} \int_K \nabla \bar{\varphi} dx \right) dx \right|.$$

Then,

$$\left| \int_K \nabla (v^K - \Pi_K v^K) \cdot \nabla \bar{\varphi} dx \right| \leq |v^K - \Pi_K v^K|_{1,K} \left\| \nabla \varphi - \frac{1}{|K|} \int_K \nabla \varphi dx \right\|_{0,K}.$$

It follows from (3.3) that there is a positive constant \hat{C} such that

$$\left| \int_K \nabla (v^K - \Pi_K v^K) \cdot \nabla \bar{\varphi} \, dx \right| \leq \hat{C} h_K |v^K - \Pi_K v^K|_{1,K} |\varphi|_{2,K}.$$

Moreover, using (3.13) we obtain that

$$\left| \int_{\partial K} (v^K - \Pi_K v^K) \partial_n \bar{\varphi} \, dt \right| = \left| \int_{\partial K} (v^K - \Pi_K v^K) \left(\nabla \bar{\varphi} - \frac{1}{|K|} \int_K \nabla \bar{\varphi} \, dx \right) \cdot \mathbf{n}^K \, dt \right|.$$

Hence, we have

$$\left| \int_{\partial K} (v^K - \Pi_K v^K) \partial_n \bar{\varphi} \, dt \right| \leq \|v^K - \Pi_K v^K\|_{0,\partial K} \left\| \nabla \varphi - \frac{1}{|K|} \int_K \nabla \varphi \, dx \right\|_{0,\partial K}.$$

Finally, using inequality (3.4) and (3.15), it follows that there is positive constant \hat{C} such that

$$\left| \int_{\partial K} (v^K - \Pi_K v^K) \partial_n \bar{\varphi} \, dt \right| \leq \hat{C} h_K |v^K - \Pi_K v^K|_{1,K} |\varphi|_{2,K}.$$

Therefore, (3.17) results from

$$\begin{aligned} \|v^K - \Pi_K v^K\|_{0,K}^2 &\leq \hat{C} h_K |v^K - \Pi_K v^K|_{1,K} |\varphi|_{2,K} \\ &= \hat{C} h_K |v^K - \Pi_K v^K|_{1,K} \|v^K - \Pi_K v^K\|_{0,K}. \end{aligned}$$

Next, we establish the estimate given by (3.18). To do this, we first note that it follows from (3.16) that

$$\forall v^K \in H^1(K), \quad |||\Pi_K v^K||| \leq \sum_{j=1}^4 |\alpha_j^K| |||\phi_j^K|||,$$

where $|||\cdot|||$ is any norm in $X_h(K)$. Hence, using (3.12), (3.11), and (3.9), there is a positive constant \hat{C} such that

$$\forall v^K \in H^1(K), \quad |||\Pi_K v^K||| \leq \frac{\hat{C}}{k^2 h_K^2} \|v^K\|_{X(K)} \max_{1 \leq j \leq 4} |||\phi_j^K|||.$$

On the other hand, it is easy to verify that

$$\|\phi_j^K\|_{0,K} \leq h_K \quad \text{and} \quad |\phi_j^K|_{1,K} \leq k h_K.$$

Consequently, there is a positive constant \hat{C} such that

$$\|\Pi_K v^K\|_{0,K} \leq \frac{\hat{C}}{k^2 h_K} \|v^K\|_{X(K)} \quad \text{and} \quad |\Pi_K v^K|_{1,K} \leq \frac{\hat{C}}{k h_K} \|v^K\|_{X(K)}.$$

Furthermore, using (3.17), we deduce that

$$\begin{aligned} \|v^K - \Pi_K v^K\|_{0,K} &\leq \hat{C} \left(h_K |v^K|_{1,K} + h_K |\Pi_K v^K|_{1,K} \right) \\ &\leq \hat{C} \left(h_K |v^K|_{1,K} + \frac{C}{k} \|v^K\|_{X(K)} \right). \end{aligned}$$

Thus,

$$k \|v^K - \Pi_K v^K\|_{0,K} \leq \hat{C} \left(kh_K |v^K|_{1,K} + \|v^K\|_{X(K)} \right),$$

and therefore, using the definition of the norm $\|\cdot\|_{X(K)}$, it follows that

$$k \|\Pi_K v^K\|_{0,K} \leq \hat{C} \|v^K\|_{X(K)},$$

which concludes the proof of the first part of (3.18).

Finally, we establish the second part of the estimate given by (3.18). To do this, we observe that $\forall v^K \in H^1(K)$, we have

$$\begin{aligned} |v^K - \Pi_K v^K|_{1,K}^2 &= \int_K \nabla (v^K - \Pi_K v^K) \cdot \nabla \bar{v}^K \, dx + \int_K (v^K - \Pi_K v^K) \Delta \Pi_K \bar{v}^K \, dx \\ &= \int_K \nabla (v^K - \Pi_K v^K) \cdot \nabla \bar{v}^K \, dx - k^2 \int_K (v^K - \Pi_K v^K) \Pi_K \bar{v}^K \, dx \\ &\leq |v^K - \Pi_K v^K|_{1,K} |v^K|_{1,K} + k^2 \|v^K - \Pi_K v^K\|_{0,K} \|\Pi_K v^K\|_{0,K}. \end{aligned}$$

Note that there are no boundary terms in the previous equalities because of Lemma 3 and (3.13).

Using again (3.17), we deduce the existence of a positive constant \hat{C} such that

$$|v^K - \Pi_K v^K|_{1,K} \leq |v^K|_{1,K} + \hat{C} k^2 h_K \|\Pi_K v^K\|_{0,K}.$$

Therefore, using the first part of (3.18), we deduce that

$$|v^K - \Pi_K v^K|_{1,K} \leq |v^K|_{1,K} + \hat{C} kh_K \|v^K\|_{X(K)}.$$

Consequently, there is a positive constant \hat{c} such that

$$|\Pi_K v^K|_{1,K} \leq 2 |v^K|_{1,K} + \hat{C} kh_K \|v^K\|_{X(K)} \leq \hat{c} \|v^K\|_{X(K)}.$$

Moreover, using (3.17), we deduce that there is a positive constant \hat{C} such that

$$\|\Pi_K v^K\|_{0,K} \leq \|v^K\|_{0,K} + \hat{C} h_K |v^K - \Pi_K v^K|_{1,K},$$

and thus,

$$\|\Pi_K v^K\|_{0,K} \leq \hat{C} h_K \|v^K\|_{X(K)},$$

which concludes the proof of (3.18). \square

LEMMA 6. Assume $kh \leq \pi$. Then for every $s \in [0, 1]$, there is a positive constant \hat{C} such that for all $K \in \mathcal{T}_h$, we have

$$(3.19) \quad |v_K - \Pi_K v_K|_{1,K} \leq \hat{C}_2 \left(h_K^s |v_K|_{1+s,K} + k^2 h_K \|v_K\|_{0,K} + k^2 h_K^2 |v_K|_{1,K} \right) \quad \forall v_K \in H^{1+s}(K).$$

Proof of Lemma 6. First, let φ be in $P_1(K)$, where $P_1(K)$ is the space of the affine polynomial functions. Then, using first (3.14) and the fact that $\nabla \varphi$ is constant in

each triangle, next that functions in X_h satisfy the homogeneous Helmholtz equation in each triangle, we can write

$$\begin{aligned} |\varphi - \Pi_K \varphi|_{1,K}^2 &= \int_K \nabla(\varphi - \Pi_K \varphi) \cdot \nabla(\bar{\varphi} - \Pi_K \bar{\varphi}) \, d\mathbf{x} = - \int_K \nabla(\varphi - \Pi_K \varphi) \cdot \nabla \Pi_K \bar{\varphi} \, d\mathbf{x} \\ &= \int_K (\varphi - \Pi_K \varphi) \cdot \Delta \Pi_K \bar{\varphi} \, d\mathbf{x} - \int_{\partial K} (\varphi - \Pi_K \varphi) \cdot \partial_n \Pi_K \bar{\varphi} \, dt \\ &= \int_K (\varphi - \Pi_K \varphi) \cdot \Delta \Pi_K \bar{\varphi} \, d\mathbf{x} = -k^2 \int_K (\varphi - \Pi_K \varphi) \cdot \Pi_K \bar{\varphi} \, d\mathbf{x} \\ &\leq k^2 \|\varphi - \Pi_K \varphi\|_{0,K} \|\Pi_K \bar{\varphi}\|_{0,K}. \end{aligned}$$

From relation (3.17), we obtain

$$\|\varphi - \Pi_K \varphi\|_{0,K} \leq \hat{C} h_K |\varphi - \Pi_K \varphi|_{1,K}.$$

Moreover, (3.18) gives

$$\|\Pi_K \varphi\|_{0,K} \leq \hat{C} (\|\varphi\|_{0,K} + h_K |\varphi|_{1,K}).$$

Hence,

$$|\varphi - \Pi_K \varphi|_{1,K} \leq \hat{C} k^2 h_K (\|\varphi\|_{0,K} + h_K |\varphi|_{1,K}).$$

On the other hand, it follows from (3.18) that for $v_K \in H^1(K)$ and $\varphi \in P_1(K)$, we have

$$|\Pi_K(\varphi - v_K)|_{1,K} \leq \hat{C} \left(\frac{1}{h_K} \|v_K - \varphi\|_{0,K} + |v_K - \varphi|_{1,K} \right)$$

and then

$$\begin{aligned} |v_K - \Pi_K v_K|_{1,K} &\leq |v_K - \varphi|_{1,K} + |\varphi - \Pi_K \varphi|_{1,K} + |\Pi_K(\varphi - v_K)|_{1,K} \\ &\leq \hat{C} \left(\frac{1}{h_K} \|v_K - \varphi\|_{0,K} + |v_K - \varphi|_{1,K} + k^2 h_K (\|\varphi\|_{0,K} + h_K |\varphi|_{1,K}) \right). \end{aligned}$$

Furthermore, since $kh_K \leq \pi$, we deduce that

$$\begin{aligned} |v_K - \Pi_K v_K|_{1,K} &\leq \hat{C} \left(\frac{1}{h_K} \|v_K - \varphi\|_{0,K} + |v_K - \varphi|_{1,K} + k^2 h_K \|v_K\|_{0,K} + k^2 h_K^2 |v_K|_{1,K} \right). \end{aligned}$$

Since $v_K \in H^{1+s}(K)$ with $s \in [0, 1]$, we chose φ to be the P_1 -polynomial approximation (the Lagrange polynomial interpolation) of v on K if $s \neq 0$ and $\varphi = \frac{1}{|K|} \int_K v \, dx$ if $s = 0$. Therefore, it follows from the standard P_1 interpolation results on K (see [14]) that

$$|v_K - \Pi_K v_K|_{1,K} \leq \hat{C} (h_K^s |v_K|_{1+s,K} + k^2 h_K \|v_K\|_{0,K} + k^2 h_K^2 |v_K|_{1,K}). \quad \square$$

Next, we introduce the *global* interpolation linear operator Π_h as follows:

$$\begin{cases} \Pi_h & : X \longrightarrow X_h, \\ & v \longmapsto \Pi_h v, \end{cases}$$

with

$$(\Pi_h v)|_K = \Pi_K(v|_K) \in X_h(K) \quad \forall K \in \mathcal{T}_h.$$

PROPERTY 3. *The global interpolation operator $\Pi_h : X \rightarrow X_h$ satisfies the following four properties:*

(i) $\forall v \in H^{1+s}(\Omega)$ with $s \in [0, 1]$, we have

$$(3.20) \quad \|v - \Pi_h v\|_{0,\Omega} \leq \hat{C} (h^{1+s}|v|_{1+s,\Omega} + k^2 h^3 |v|_{1,\Omega} + k^2 h^2 \|v\|_{0,\Omega}),$$

$$(3.21) \quad |v - \Pi_h v|_{1,\mathcal{T}_h} \leq \hat{C} (h^s |v|_{1+s,\Omega} + k^2 h^2 |v|_{1,\Omega} + k^2 h \|v\|_{0,\Omega}).$$

(ii) $\forall v \in H^1(\Omega)$, $\Pi_h v \in \mathcal{N}_h$, where \mathcal{N}_h is the null space of $b(\cdot, \cdot)$.

(iii) $\forall v \in X$ and $\forall v_h \in X_h$, we have

$$(3.22) \quad \begin{aligned} a(v - \Pi_h v, v_h) &= -ik \sum_{K \in \mathcal{T}_h} \int_{\partial K \cap \Sigma} (v - \Pi_h v) \bar{v}_h dt, \\ a(v_h, v - \Pi_h v) &= -ik \sum_{K \in \mathcal{T}_h} \int_{\partial K \cap \Sigma} v_h (\bar{v} - \Pi_h \bar{v}) dt. \end{aligned}$$

(iv) $\forall v \in X$ and $\forall \mu_h \in M_h$, we have

$$(3.23) \quad b(v, \mu_h) = b(\Pi_h v, \mu_h).$$

Note that (3.20)–(3.21) are immediate consequences of Lemma 6, while the two equalities given by (3.22) are obtained by Green’s formula and using the fact that the plane waves are solutions of the Helmholtz equation.

3.3.2. Interpolation operator in M_h . We introduce here the projection operator P_h for the dual variable λ . P_h is defined as follows:

$$\begin{cases} P_h : \mathcal{M} & \longrightarrow M_h, \\ \mu & \longmapsto P_h \mu, \end{cases}$$

where

$$\forall K \in \mathcal{T}_h, \quad P_h \mu|_{T_j^K} = \frac{1}{h_j^K} \int_{T_j^K} \mu dt, \quad 1 \leq j \leq 4.$$

Then, the operator P_h satisfies

$$(3.24) \quad \forall K \in \mathcal{T}_h, \quad \forall \mu \in \mathcal{M}, \quad \int_{\partial K} \mu dt = \int_{\partial K} P_h \mu dt.$$

3.4. Proof of Theorem 2. We first prove that the DVP admits a unique solution (u_h, λ_h) in $X_h \times M_h$ and then we establish the error estimate given by (3.6).

3.4.1. Existence and uniqueness. First, we prove that the bilinear form $b(\cdot, \cdot)$ satisfies the inf-sup condition [21]. This result is stated in Proposition 1. Then, we prove in Proposition 2 the uniqueness of the solution of the *homogeneous* problem corresponding to the variational problem (DVP). The existence and uniqueness of the DVP is then a direct consequence of Proposition 1 and Proposition 2.

PROPOSITION 1. Assume $kh \leq \pi$. Then, there is a positive constant γ independent of k and h such that

$$\gamma \|\mu_h\|_M \leq \sup_{v_h \in X_h} \frac{|b(v_h, \mu_h)|}{\|v_h\|_X} \leq \|\mu_h\|_M \quad \forall \mu_h \in M_h.$$

Proof of Proposition 1. From (2.8), we deduce that

$$\forall \mu_h \in M_h, \quad \sup_{v_h \in X_h} \frac{|b(v_h, \mu_h)|}{\|v_h\|_X} \leq \|\mu_h\|_M.$$

In addition, it follows from (2.7) that

$$\forall \mu_h \in M_h, \exists \phi \in X, \quad \sup_{v \in X} \frac{|b(v, \mu_h)|}{\|v\|_X} = \frac{|b(\phi, \mu_h)|}{\|\phi\|_X} = \|\mu_h\|_M.$$

Therefore, it follows from (3.23) that

$$\|\mu_h\|_M = \frac{|b(\Pi_h \phi, \mu_h)|}{\|\Pi_h \phi\|_X} \frac{\|\Pi_h \phi\|_X}{\|\phi\|_X}.$$

Since $kh \leq \pi$, it follows from (3.18) that there is a positive constant \hat{C} such that

$$\|\mu_h\|_M \leq \hat{C} \sup_{v_h \in X_h} \frac{|b(v_h, \mu_h)|}{\|v_h\|_X},$$

which concludes the proof of Proposition 1. \square

PROPOSITION 2. Assume $kh \leq \pi$. Then, the only solution of the following homogeneous DVP

$$\begin{cases} \text{Find } u_h \in \mathcal{N}_h \text{ such that} \\ a(u_h, v_h) = 0 \quad \forall v_h \in \mathcal{N}_h, \end{cases}$$

is the trivial one.

Proof of Proposition 2. Let $u_h \in \mathcal{N}_h$ such that $a(u_h, v_h) = 0 \quad \forall v_h \in \mathcal{N}_h$, then $a(u_h, u_h) = 0$, which implies

$$u_h = 0 \quad \text{on } \Sigma \quad \text{and} \quad k \|u_h\|_{0,\Omega} = |u_h|_{1,\mathcal{T}_h}.$$

In addition, since $u_h \in X_h$, then $\Delta u_h + k^2 u_h = 0$ in every $K \in \mathcal{T}_h$. Therefore, using integration by parts, it follows that

$$a(u_h, v_h) = \sum_{K \in \mathcal{T}_h} \int_{\partial K} \bar{v}_h \partial_n u_h \, dt = 0 \quad \forall v_h \in \mathcal{N}_h.$$

Then, we also have $\partial_n u_h = 0$ on $\Gamma \cup \Sigma$ and $[\partial_n u_h] = 0$ on $\partial K \cap \partial K' \quad \forall K \neq K' \in \mathcal{T}_h$, where $[\partial_n u_h] = \partial_n u_h^K + \partial_n u_h^{K'}$ is the jump of the normal derivative of u_h across $\partial K \cap \partial K'$.

To conclude the proof of this proposition, we use a discrete continuation result. We consider first the following property (P).

Let $K \in \mathcal{T}_h$ and T_l^K and T_m^K be two adjacent edges of K such that

$$\partial_n u_h^K|_{T_l^K} = \partial_n u_h^K|_{T_m^K} = \int_{T_l^K} u_h dt = \int_{T_m^K} u_h dt = 0, \quad \text{then} \quad u_h = 0 \quad \text{in} \quad K.$$

Note that property (P) is easy to establish since $u_h \in X_h$ (a sum of four plane waves), and therefore, u_h satisfies the Helmholtz equation at the element level K .

Now, since there is at least one element $K \in \mathcal{T}_h$ with two adjacent edges belonging to the boundary Σ , then using property (P) leads to $u_h = 0$ in K . Then, we obtain sequentially that $u_h = 0$ in all the quadrilaterals belonging to the first layer adjacent to the boundary Σ . We repeat this process on the second layer of the quadrilaterals and so on, until the boundary Γ is reached, which proves the uniqueness of the solution u_h . \square

3.4.2. A priori error estimates. In the next lemmas, we establish a priori estimates in order to prove the error estimate (3.6) given in Theorem 2 between the exact solution (u, λ) and the discrete solution (u_h, λ_h) .

We consider the following notations:

$$(3.25) \quad \kappa_h = h(1 + k) \quad \text{and} \quad z_h = u_h - \Pi_h u.$$

LEMMA 7. *There is a positive constant \hat{C} independent of k and h such that the solution λ of the variational problem VP (2.4) satisfies*

$$\|\lambda - P_h \lambda\|_M \leq \hat{C} \kappa_h^{\frac{2}{3}} (1 + k).$$

Proof of Lemma 7. First, recall that

$$\lambda^K = \begin{cases} -\partial_n u & \text{on } \partial K \setminus \partial\Omega, \\ 0 & \text{on } \partial K \cap \partial\Omega. \end{cases}$$

Therefore, using the definition of the operator P_h along with the fact the normal unit vector \mathbf{n}^K is constant on each edge e of K , we deduce that $\forall K \in \mathcal{T}_h$, we have

$$\begin{aligned} \|\lambda - P_h \lambda\|_{0,\partial K}^2 &= \sum_{e \in K, e \text{ interior}} \left\| \nabla u \cdot \mathbf{n}^K - \frac{1}{|e|} \int_e \nabla u \cdot \mathbf{n}^K dt \right\|_{0,e}^2 \\ &\leq \sum_{e \in K, e \text{ interior}} \left\| \nabla u - \frac{1}{|e|} \int_e \nabla u dt \right\|_{0,e}^2 = \sum_{e \in K, e \text{ interior}} \inf_{\beta \in \mathbb{C}^2} \|\nabla u - \beta\|_{0,e}^2 \\ &\leq \sum_{e \in K, e \text{ interior}} \left\| \nabla u - \frac{1}{|K|} \int_K \nabla u dx \right\|_{0,e}^2 \leq \left\| \nabla u - \frac{1}{|K|} \int_K \nabla u dx \right\|_{0,\partial K}^2. \end{aligned}$$

Finally, using classical interpolation results [14], there is a positive constant \hat{C} such that

$$(3.26) \quad \forall K \in \mathcal{T}_h, \quad \|\lambda - P_h \lambda\|_{0,\partial K} \leq \hat{C} h_K^{\frac{1}{6}} |u|_{\frac{5}{3},K}.$$

In addition, we have from (2.3) that

$$\|\lambda - P_h \lambda\|_{H^{-\frac{1}{2}}(\partial K)} = \sup_{v \in H^1(K)} \frac{|\int_{\partial K} (\lambda - P_h \lambda) v dt|}{\|v\|_{X(K)}}.$$

On the other hand, from (3.24), we deduce that

$$\left| \int_{\partial K} (\lambda - P_h \lambda) v dt \right| = \left| \int_{\partial K} (\lambda - P_h \lambda) \left(v - \frac{1}{|K|} \int_K v dx \right) dt \right| \quad \forall v \in H^1(K).$$

Hence,

$$\left| \int_{\partial K} (\lambda - P_h \lambda) v dt \right| \leq \|\lambda - P_h \lambda\|_{0,\partial K} \left\| v - \frac{1}{|K|} \int_K v dx \right\|_{0,\partial K} \quad \forall v \in H^1(K).$$

Using the following classical interpolation results [14], it follows that there is a positive constant \hat{C} such that

$$\left\| v - \frac{1}{|K|} \int_K v dx \right\|_{0,\partial K} \leq \hat{C} h_K^{\frac{1}{2}} |v|_{1,K} \leq \hat{C} h_K^{\frac{1}{2}} \|v\|_{X(K)}.$$

We then deduce the existence of a positive constant \hat{C} such that

$$(3.27) \quad \forall K \in \mathcal{T}_h, \quad \|\lambda - P_h \lambda\|_{H^{-\frac{1}{2}}(\partial K)} \leq \hat{C} h_K^{\frac{1}{2}} \|\lambda - P_h \lambda\|_{0,\partial K} \quad \forall \mu \in \mathcal{M}.$$

Lemma 7 is the consequence of (3.26)–(3.27) and Theorem 1. \square

The next lemma can be viewed as a consistency result.

LEMMA 8. *Assume $kh \leq \pi$. Then, there is a positive constant \hat{C} independent of k and h such that $\forall v_h \in X_h$ and $\forall v \in H^1(\Omega)$,*

$$|a(z_h, v_h) + b(v_h, \lambda_h - P_h \lambda)| \leq \hat{C} (1 + k) \kappa_h^{\frac{2}{3}} [\kappa_h |v_h|_{1,\mathcal{T}_h} + |v - v_h|_{1,\mathcal{T}_h}].$$

Proof of Lemma 8. We have

$$a(z_h, v_h) = a(u_h - \Pi_h u, v_h) = a(u - \Pi_h u, v_h) - a(u - u_h, v_h).$$

Moreover, since u satisfies VP, we have

$$a(u, v_h) + b(v_h, \lambda) = F(v_h),$$

and since u_h satisfies DVP, we have

$$a(u_h, v_h) + b(v_h, \lambda_h) = F(v_h).$$

Consequently, we obtain

$$a(u - u_h, v_h) = -b(v_h, \lambda - \lambda_h),$$

which leads to

$$a(z_h, v_h) + b(v_h, \lambda_h - P_h \lambda) = a(u - \Pi_h u, v_h) + b(v_h, \lambda - P_h \lambda).$$

Hence, it follows from (3.22) that

$$(3.28) \quad a(u_h - \Pi_h u, v_h) + b(v_h, \lambda_h - P_h \lambda) = -ik \int_{\Sigma} (u - \Pi_h u) \bar{v}_h dt + b(v_h, \lambda - P_h \lambda) \quad \forall v_h \in X_h.$$

Next, using (3.13) and following the same proof of (3.26) in Lemma 7, we obtain

$$\begin{aligned} \left| \int_{\Sigma} (u - \Pi_h u) \bar{v}_h dt \right| &\leq \sum_{e \in \mathcal{C}\Sigma} \int_e |u - \Pi_h u| |\bar{v}_h - \frac{1}{|e|} \int_e \bar{v}_h dt| dt \\ &\leq \sum_{\partial K \subset \Sigma} \|u - \Pi_h u\|_{0, \partial K} \left\| v_h - \frac{1}{|K|} \int_K v_h dx \right\|_{0, \partial K}. \end{aligned}$$

Hence, using (3.4), it follows that there is a positive constant \hat{C} such that

$$\left| \int_{\Sigma} (u - \Pi_h u) \bar{v}_h dt \right| \leq \hat{C} \sum_{K \in \mathcal{T}_h} h_K |u - \Pi_h u|_{1, K} |v_h|_{1, K}.$$

Then, it follows from using Theorem 1 and Lemma 6 that there is a positive constant \hat{C} such that

$$\left| \int_{\Sigma} (u - \Pi_h u) \bar{v}_h dt \right| \leq \hat{C} \left(\kappa_h^{\frac{5}{3}} + \kappa_h^2 + \kappa_h^3 \right) |v_h|_{1, \mathcal{T}_h},$$

which implies (assuming $kh \leq \pi$) that

$$(3.29) \quad \left| \int_{\Sigma} (u - \Pi_h u) \bar{v}_h dt \right| \leq \hat{C} \kappa_h^{\frac{5}{3}} |v_h|_{1, \mathcal{T}_h}.$$

On the other hand, we have $\forall v \in H^1(\Omega)$,

$$\begin{aligned} |b(v_h, \lambda - P_h \lambda)| &= \left| \sum_{e \text{ interior}} \int_e [\bar{v}_h] (\lambda - P_h \lambda) dt \right| = \left| \sum_{e \text{ interior}} \int_e [\bar{v} - \bar{v}_h] (\lambda - P_h \lambda) dt \right| \\ &= \left| \sum_{e \text{ interior}} \int_e (\lambda - P_h \lambda) \cdot \left[(\bar{v} - \bar{v}_h) - \frac{1}{|e|} \int_e (\bar{v} - \bar{v}_h) \right] dt \right| \\ &\leq \sum_K \|\lambda - P_h \lambda\|_{0, \partial K} \left\| (v - v_h) - \frac{1}{|K|} \int_K (v - v_h) dx \right\|_{0, \partial K}. \end{aligned}$$

Therefore, it follows from using using (3.4) that there is a positive constant \hat{C} such that

$$|b(v_h, \lambda - P_h \lambda)| \leq \hat{C} \sum_{K \in \mathcal{T}_h} h_K^{\frac{1}{2}} |v - v_h|_{1, K} \|\lambda - P_h \lambda\|_{0, \partial K}.$$

Hence, from (3.26) and Theorem 1, we obtain that there is a positive constant \hat{C} such that

$$(3.30) \quad |b(v_h, \lambda - P_h \lambda)| \leq \hat{C} \kappa_h^{\frac{2}{3}} (1 + k) |v - v_h|_{1, \mathcal{T}_h}.$$

We conclude the proof of Lemma 8 by substituting (3.29) and (3.30) into (3.28). \square

Remark 4. We deduce from Lemma 8 that, when $kh \leq \pi$, there is a positive constant \hat{C} such that $\forall v_h \in \mathcal{N}_h$ and $\forall v \in H^1(\Omega)$,

$$(3.31) \quad |a(z_h, v_h)| \leq \hat{C} (1 + k) \kappa_h^{\frac{2}{3}} [\kappa_h |v_h|_{1, \mathcal{T}_h} + |v - v_h|_{1, \mathcal{T}_h}].$$

LEMMA 9. Assume $kh \leq \pi$. Then, there is a positive constant C (C depends on Ω only) such that

$$(3.32) \quad \|z_h\|_{0, \Omega} \leq C \kappa_h^{\frac{2}{3}} \left[(1 + k) \kappa_h^{\frac{2}{3}} + |z_h|_{1, \mathcal{T}_h} \right].$$

Proof of Lemma 9. First, observe that z_h belongs to \mathcal{N}_h and let ϕ be the solution of the following BVP (see Lemma 1):

$$-\Delta \bar{\phi} - k^2 \bar{\phi} = \bar{z}_h \quad \text{in } \Omega,$$

and

$$\partial_n \bar{\phi} = 0 \quad \text{on } \Gamma, \quad \partial_n \bar{\phi} = ik \bar{\phi} \quad \text{on } \Sigma.$$

Hence, it follows from Lemma 1 that $\phi \in H^{\frac{5}{3}}(\Omega)$ and (see (2.12)) there is constant $C > 0$ (C depends on Ω only) such that, for every $s \in [0, \frac{5}{3}]$, we have

$$(3.33) \quad |\phi|_{s, \Omega} \leq C (1 + k)^{s-1} \|z_h\|_{0, \Omega}.$$

In addition, we have

$$(3.34) \quad \|z_h\|_{0, \Omega}^2 = a(z_h, \phi) - \sum_{e \text{ interior}} \int_e [z_h] \partial_n \bar{\phi} dt.$$

Equation (3.34) results from multiplying the BVP introduced in Lemma 9, integrating by parts on Ω , and using the definition of the bilinear form a . The second term of this equality is due to the discontinuity of z_h along the interior edges. Recall that the jump $[\phi]$ along $e \in \partial K \cap \partial K'$ is given by $[\phi] = \phi^K - \phi^{K'}$.

On the other hand, we have

$$|a(z_h, \phi)| \leq |a(z_h, \Pi_h \phi)| + |a(z_h, \phi - \Pi_h \phi)|.$$

It follows from (3.22) that

$$(3.35) \quad |a(z_h, \phi)| \leq |a(z_h, \Pi_h \phi)| + k \left| \int_{\Sigma} z_h (\bar{\phi} - \Pi_h \bar{\phi}) dt \right|.$$

Since $\Pi_h \phi \in \mathcal{N}_h$ (see property (ii) in Property 3), then it follows from Remark 4 that there is a positive constant \hat{C} such that

$$|a(z_h, \Pi_h \phi)| \leq \hat{C} (1 + k) \kappa_h^{\frac{2}{3}} [\kappa_h |\Pi_h \phi|_{1, \mathcal{T}_h} + |\phi - \Pi_h \phi|_{1, \mathcal{T}_h}].$$

Moreover, it follows from Lemma 6 that there is a positive constant \hat{C} such that

$$|\phi - \Pi_h \phi|_{1, \mathcal{T}_h} \leq \hat{C} \left\{ h^{\frac{2}{3}} |\phi|_{\frac{5}{3}, \Omega} + k^2 h \|\phi\|_{0, \Omega} + k^2 h^2 |\phi|_{1, \Omega} \right\}.$$

Then, using relation (3.33) and the assumption $kh \leq \pi$, we obtain

$$|\phi - \Pi_h \phi|_{1, \mathcal{T}_h} \leq \hat{C} \kappa_h^{\frac{2}{3}} \|z_h\|_{0, \Omega} \quad \text{and} \quad |\Pi_h \phi|_{1, \mathcal{T}_h} \leq \hat{C} \|z_h\|_{0, \Omega}.$$

We then obtain

$$|a(z_h, \Pi_h \phi)| \leq \hat{C} (1+k) \kappa_h^{\frac{4}{3}} \|z_h\|_{0, \Omega}.$$

For the second part of (3.35), we have

$$\left| \int_{\Sigma} z_h (\bar{\phi} - \Pi_h \bar{\phi}) dt \right| \leq \hat{C} h |\phi - \Pi_h \phi|_{1, \mathcal{T}_h} |z_h|_{1, \mathcal{T}_h} \leq \hat{C} h \kappa_h^{\frac{2}{3}} |z_h|_{1, \mathcal{T}_h} \|z_h\|_{0, \Omega}.$$

Note that the previous inequality was obtained using the same methodology to prove Lemma 5. Hence, first we use (3.13) when we add the constant $(-\frac{1}{|K|} \int_K z_h dt)$ to z_h . Then, we apply Cauchy–Schwarz along with inequalities (3.2) and (3.4).

Finally, it follows that there is a positive constant C (C depends on Ω only) such that

$$(3.36) \quad |a(z_h, \phi)| \leq C \left[(1+k) \kappa_h^{\frac{4}{3}} + \kappa_h^{\frac{5}{3}} |z_h|_{1, \mathcal{T}_h} \right] \|z_h\|_{0, \Omega}.$$

Next, we estimate the term $|\sum_e \text{interior} \int_e [z_h] \partial_n \bar{\phi} dt|$ in (3.34). First, observe that

$$\int_e z_h^K dt = \int_e z_h^{K'} dt \quad \forall e \in \partial K \cap \partial K' \text{ and } K \neq K' \in \mathcal{T}_h$$

and

$$\begin{aligned} & \int_{e \in \partial K \cap \partial K'} (z_h^K - z_h^{K'}) \partial_n \bar{\phi} dt \\ &= \int_e \left(z_h^K - \frac{1}{|e|} \int_e z_h^K dt \right) \left(\nabla \phi - \frac{1}{|K|} \int_K \nabla \phi dx \right) \cdot \mathbf{n}^K dt \\ &+ \int_e \left(z_h^{K'} - \frac{1}{|e|} \int_e z_h^{K'} dt \right) \left(\nabla \phi - \frac{1}{|K'|} \int_{K'} \nabla \phi dx \right) \cdot \mathbf{n}^{K'} dt. \end{aligned}$$

Therefore,

$$\left| \sum_e \text{interior} \int_e [z_h] \partial_n \bar{\phi} dt \right| \leq \sum_{K \in \mathcal{T}_h} \sum_{e \subset K} \int_e \left| z_h - \frac{1}{|e|} \int_e z_h dt \right| \left| \nabla \phi - \frac{1}{|K|} \int_K \nabla \phi dx \right| dt.$$

Hence, it follows that

$$(3.37) \quad \left| \sum_e \text{interior} \int_e [z_h] \partial_n \bar{\phi} dt \right| \leq \hat{C} h^{\frac{2}{3}} |z_h|_{1, \mathcal{T}_h} |\phi|_{\frac{5}{3}, \Omega} \leq C \kappa_h^{\frac{2}{3}} |z_h|_{1, \mathcal{T}_h} \|z_h\|_{0, \Omega}.$$

We conclude the proof of Lemma 9 by substituting (3.36) and (3.37) into equation (3.34). \square

LEMMA 10. *Let h_0 be a positive number such that $kh_0^{\frac{2}{3}}(1+k)^{\frac{2}{3}}$ is “sufficiently small.” Then, there is a positive constant C (C depends on Ω only) such that, for all $h \leq h_0$, we have*

$$\|u_h - \Pi_h u\|_{0, \Omega} \leq \hat{C} (1+k) \kappa_h^{\frac{4}{3}} \quad \text{and} \quad |u_h - \Pi_h u|_{1, \mathcal{T}_h} \leq \hat{C} (1+k) \kappa_h^{\frac{2}{3}}.$$

Proof of Lemma 10. It follows from the definition of the bilinear form $a(.,.)$ that

$$|a(z_h, z_h)|^2 = \left| |z_h|_{1, \mathcal{T}_h}^2 - k^2 \|z_h\|_{0, \Omega}^2 \right|^2 + k^2 \|z_h\|_{0, \Gamma}^4.$$

Moreover, using Remark 4 with $v_h = z_h$ and $v = 0$ along with the fact that $kh \leq \pi$, we obtain

$$|a(z_h, z_h)| \leq \hat{C} (1 + k) \kappa_h^{\frac{2}{3}} |z_h|_{1, \mathcal{T}_h}.$$

Therefore, we deduce that

$$|z_h|_{1, \mathcal{T}_h}^2 \leq k^2 \|z_h\|_{0, \Omega}^2 + \hat{C} (1 + k) \kappa_h^{\frac{2}{3}} |z_h|_{1, \mathcal{T}_h}.$$

Then, using (3.32) along with Young's inequality, we obtain

$$|z_h|_{1, \mathcal{T}_h}^2 \leq C \left[k^2 (1 + k)^2 \kappa_h^{\frac{8}{3}} + k^2 \kappa_h^{\frac{4}{3}} |z_h|_{1, \mathcal{T}_h}^2 + (1 + k) \kappa_h^{\frac{2}{3}} |z_h|_{1, \mathcal{T}_h} \right].$$

Consequently, we have

$$|z_h|_{1, \mathcal{T}_h}^2 \leq C \left[k^2 (1 + k)^2 \kappa_h^{\frac{8}{3}} + k^2 \kappa_h^{\frac{4}{3}} |z_h|_{1, \mathcal{T}_h}^2 + (1 + k)^2 \kappa_h^{\frac{4}{3}} \right].$$

Let us consider h_0 such that $Ck^2(1+k)^{\frac{4}{3}}h_0^{\frac{4}{3}} \leq \frac{1}{2}$, then for every $h \leq h_0$, we have $Ck^2\kappa_h^{\frac{4}{3}} \leq \frac{1}{2}$. We deduce that

$$|z_h|_{1, \mathcal{T}_h}^2 \leq C \left[k^2 (1 + k)^2 \kappa_h^{\frac{8}{3}} + (1 + k)^2 \kappa_h^{\frac{4}{3}} \right], \quad \text{then} \quad |z_h|_{1, \mathcal{T}_h} \leq \hat{C} (1 + k) \kappa_h^{\frac{2}{3}}.$$

In addition, we obtain, from using (3.32), that

$$\|z_h\|_{0, \Omega} \leq \hat{C} (1 + k) \kappa_h^{\frac{4}{3}},$$

which concludes the proof of Lemma 10. \square

Proof of the a priori error estimate of Theorem 2. We are now ready to prove the estimate given by (3.6).

- From Lemmas 6 and 10, it follows that there is a positive constant C (C depends on Ω only) such that

$$\|u - u_h\|_{0, \Omega} \leq \|u - \Pi_h u\|_{0, \Omega} + \|u_h - \Pi_h u\|_{0, \Omega} \leq C \left[\kappa_h^{\frac{4}{3}} + (1 + k) \kappa_h^{\frac{4}{3}} \right]$$

and

$$|u - u_h|_{1, \mathcal{T}_h} \leq |u - \Pi_h u|_{1, \mathcal{T}_h} + |u_h - \Pi_h u|_{1, \mathcal{T}_h} \leq C \left[\kappa_h^{\frac{2}{3}} + k\kappa_h + (1 + k) \kappa_h^{\frac{2}{3}} \right].$$

Hence, we deduce that

$$\|u - u_h\|_{0, \Omega} \leq C (1 + k) \kappa_h^{\frac{4}{3}} \quad \text{and} \quad |u - u_h|_{1, \mathcal{T}_h} \leq C (1 + k) \kappa_h^{\frac{2}{3}}.$$

- Moreover, we deduce from Lemma 8 that there is a positive constant \hat{C} such that

$$|b(v_h, \lambda_h - P_h \lambda)| \leq \hat{C} (1 + k) \kappa_h^{\frac{2}{3}} |v_h|_{1, \mathcal{T}_h} + |a(z_h, v_h)| \quad \forall v_h \in X_h.$$

On the other hand, it follows from the definition of the bilinear form $a(\cdot, \cdot)$ that

$$|a(z_h, v_h)| \leq |z_h|_{1, \mathcal{T}_h} |v_h|_{1, \mathcal{T}_h} + k^2 \left| \int_{\Omega} z_h \cdot \bar{v}_h dx \right| + k \|z_h\|_{0, \Sigma} \|v_h\|_{0, \Sigma} \quad \forall v_h \in X_h.$$

Therefore, using the definition of the norm $\|\cdot\|_X$ and inverse inequality results, we deduce that there is a positive constant \hat{C} such that

$$|a(z_h, v_h)| \leq (|z_h|_{1, \mathcal{T}_h}^2 + k^2 h^2 \|z_h\|_{0, \Omega}^2)^{\frac{1}{2}} \|v_h\|_X + \hat{C} k \|z_h\|_{0, \Sigma} h^{\frac{1}{2}} \|v_h\|_X \quad \forall v_h \in X_h.$$

In addition, it follows from the definition of the bilinear form $a(\cdot, \cdot)$ and from using (3.31) with $v_h = z_h$ and $v = 0$ (see Remark 4) that there is a positive constant \hat{C} such that

$$k \|z_h\|_{0, \Sigma}^2 \leq |a(z_h, z_h)| \leq \hat{C} (1 + k) \kappa_h^{\frac{2}{3}} |z_h|_{1, \mathcal{T}_h}.$$

Therefore, using Lemma 10, we deduce that there is a positive constant C (C depends on Ω only) such that

$$k^{\frac{1}{2}} \|z_h\|_{0, \Sigma} \leq (1 + k) \kappa_h^{\frac{2}{3}}.$$

Hence, we deduce that there is a positive constant C (C depends on Ω only) such that

$$|a(z_h, v_h)| \leq C (1 + k) \kappa_h^{\frac{2}{3}} \|v_h\|_X \quad \forall v_h \in X_h.$$

Consequently, it follows from Proposition 1 that there is a positive constant C (C depends on Ω only) such that

$$\|\lambda_h - P_h \lambda\|_M \leq C (1 + k) \kappa_h^{\frac{2}{3}}.$$

Finally, we deduce from Lemma 7 that there is a positive constant C (C depends on Ω only) such that

$$\|\lambda - \lambda_h\|_M \leq C (1 + k) \kappa_h^{\frac{2}{3}},$$

which concludes the proof of the error estimate of Theorem 2. \square

Proof of the a posteriori error estimate (3.7) in Theorem 3. Let ϕ be the solution of the BVP (2.11) (see Lemma 1) with $f = u - u_h$. Then, this solution ϕ belongs to $H^{\frac{5}{3}}(\Omega)$ and for every $s \in [0, \frac{5}{3}]$, there exists a constant $C > 0$ depending only on s and Ω such that

$$|\phi|_{s, \Omega} \leq C(1 + k)^{s-1} \|u - u_h\|_{0, \Omega}.$$

Using integration by parts, one can easily verify that

$$\begin{aligned} \|u - u_h\|_{0, \Omega}^2 &= \sum_{K \in \mathcal{T}_h} \int_{\partial K \cap \Sigma} \bar{\phi} (\partial_n u_h - ik u_h) dt + \sum_{K \in \mathcal{T}_h} \int_{\partial K \cap \Gamma} \bar{\phi} (\partial_n u_h + \partial_n e^{ikx \cdot d}) dt \\ &\quad + \sum_{e \text{ interior}} \int_e [\partial_n u_h] \bar{\phi} dt - \sum_{e \text{ interior}} \int_e [u_h] \partial_n \bar{\phi} dt. \end{aligned}$$

On the other hand, we also have

$$a(u_h, \Pi_h \phi) = - \sum_{K \in \mathcal{T}_h} \int_{\partial K \cap \Gamma} \partial_n e^{ikx \cdot \mathbf{d}} \Pi_h \bar{\phi} dt.$$

Therefore, using integration by parts along with the fact that u_h satisfies the Helmholtz equation at the element level, we have

$$\begin{aligned} \sum_{K \in \mathcal{T}_h} \int_{\partial K \cap \Gamma} \partial_n u_h \Pi_h \bar{\phi} dt + \sum_{K \in \mathcal{T}_h} \int_{\partial K \cap \Sigma} (\partial_n u_h - ik u_h) \Pi_h \bar{\phi} dt \\ + \sum_{e \text{ interior}} \int_e [\partial_n u_h] \Pi_h \bar{\phi} dt \\ = - \sum_{K \in \mathcal{T}_h} \int_{\partial K \cap \Gamma} \partial_n e^{ikx \cdot \mathbf{d}} \Pi_h \bar{\phi} dt. \end{aligned}$$

Consequently, using the fact that for every interior edge e , we have $\int_e [\partial_n u_h] \bar{\phi} dt = \int_e [\partial_n u_h] \Pi_h \bar{\phi} dt$, we deduce that

(3.38)

$$\begin{aligned} \|u - u_h\|_{0,\Omega}^2 &= \sum_{K \in \mathcal{T}_h} \int_{\partial K \cap \Sigma} (\bar{\phi} - \Pi_h \bar{\phi}) (\partial_n u_h - ik u_h) dt \\ &\quad + \sum_{K \in \mathcal{T}_h} \int_{\partial K \cap \Gamma} (\bar{\phi} - \Pi_h \bar{\phi}) (\partial_n u_h + \partial_n e^{ikx \cdot \mathbf{d}}) dt - \sum_{e \text{ interior}} \int_e [u_h] \partial_n \bar{\phi} dt. \end{aligned}$$

Next, we estimate each integral in the right-hand side of (3.38) to deduce the a posteriori estimate given by (3.7) in Theorem 3.

- First, we estimate: $I_1 = \left| \sum_{K \in \mathcal{T}_h} \int_{\partial K \cap \Sigma} (\bar{\phi} - \Pi_h \bar{\phi}) (\partial_n u_h - ik u_h) dt \right|.$

We have

$$\begin{aligned} I_1 &\leq \left(\sum_{e \subset \Sigma} h_e \|\partial_n u_h - ik u_h\|_{0,e}^2 \right)^{\frac{1}{2}} \left(\sum_{e \subset \Sigma} h_e^{-1} \|\bar{\phi} - \Pi_h \bar{\phi}\|_{0,e}^2 \right)^{\frac{1}{2}} \\ &\leq \hat{C} \left(\sum_{e \subset \Sigma} h_e \|\partial_n u_h - ik u_h\|_{0,e}^2 \right)^{\frac{1}{2}} |\bar{\phi} - \Pi_h \bar{\phi}|_{1,\mathcal{T}_h}. \end{aligned}$$

Therefore, assuming that $kh \leq \pi$, it follows from the properties of the operator Π (see (3.21) in Property 3) that there is a positive constant \hat{C} such that

$$I_1 \leq \hat{C}_1 \left(\sum_{e \subset \Sigma} h_e \|\partial_n u_h - ik u_h\|_{0,e}^2 \right)^{\frac{1}{2}} \left(h^{\frac{2}{3}} |\bar{\phi}|_{\frac{5}{3},\Omega} + |\bar{\phi}|_{1,\Omega} + k \|\bar{\phi}\|_{0,\Omega} \right).$$

We deduce from the a priori estimate on $|\bar{\phi}|_{s,\Omega}$ that there is a positive constant \hat{C}_1 such that

$$I_1 \leq \hat{C}_1 \left(\sum_{e \subset \Sigma} h_e \|\partial_n u_h - ik u_h\|_{0,e}^2 \right)^{\frac{1}{2}} \|u - u_h\|_{0,\Omega}.$$

- Similarly, there is also a positive constant \hat{C}_2 such that

$$I_2 = \left| \sum_{K \in \mathcal{T}_h} \int_{\partial K \cap \Gamma} (\bar{\phi} - \Pi_h \bar{\phi}) (\partial_n u_h + \partial_n e^{ik\mathbf{x} \cdot \mathbf{d}}) dt \right| \leq \hat{C} \left(\sum_{e \in \Gamma} h_e \|\partial_n u_h + \partial_n e^{ik\mathbf{x} \cdot \mathbf{d}}\|_{0,e}^2 \right)^{\frac{1}{2}} |\phi - \Pi_h \phi|_{1, \mathcal{T}_h}.$$

Then, there is there is a positive constant denoted again by \hat{C}_2 such that

$$I_2 \leq \hat{C}_2 \left(\sum_{e \in \Gamma} h_e \|\partial_n u_h + \partial_n e^{ik\mathbf{x} \cdot \mathbf{d}}\|_{0,e}^2 \right)^{\frac{1}{2}} \|u - u_h\|_{0, \Omega}.$$

- Last, we estimate $I_3 = |\sum_e \text{interior} \int_e [u_h] \partial_n \bar{\phi} dt|$. Consider an interior edge $e = \partial K(e) \cap \partial K'(e)$, then

$$\int_e [u_h] \partial_n \bar{\phi} dt = \int_e [u_h] \nabla \bar{\phi} \cdot \mathbf{n} dt = \int_e [u_h] (\nabla \bar{\phi} - \beta) \cdot \mathbf{n} dt \quad \forall \beta \in \mathbb{C}^2.$$

We then obtain

$$\left| \int_e [u_h] \partial_n \bar{\phi} dt \right| \leq \|[u_h]\|_{0,e} \inf_{\beta \in \mathbb{C}^2} \|\nabla \bar{\phi} - \beta\|_{0,e}.$$

On the other hand, since there is a positive constant \hat{C} such that

$$\inf_{\beta \in \mathbb{C}^2} \|\nabla \bar{\phi} - \beta\|_{0,e} \leq \hat{C} h_e^{\frac{1}{3}} |\phi|_{\frac{5}{3}, K(e)},$$

it follows that

$$I_3 \leq \hat{C} \sum_{e \text{ interior}} h_e^{\frac{1}{6}} \|[u_h]\|_{0,e} |\phi|_{\frac{5}{3}, K(e)} \leq \hat{C} \left(\sum_{e \text{ interior}} h_e^{-1} \|[u_h]\|_{0,e}^2 \right)^{\frac{1}{2}} h^{\frac{2}{3}} |\phi|_{\frac{5}{3}, \Omega}.$$

Then, there is a positive constant \hat{C}_3 such that

$$I_3 \leq \hat{C}_3 \left(\sum_{e \text{ interior}} h_e^{-1} \|[u_h]\|_{0,e}^2 \right)^{\frac{1}{2}} \|u - u_h\|_{0, \Omega}.$$

4. Conclusion. A DGM with plane waves and Lagrange multipliers was recently proposed by Farhat, Harari, and Hetmaniuk [3] for solving two-dimensional Helmholtz problems at relatively high wavenumbers. In many previous papers, this method was shown numerically to offer a significant potential for wave propagation problems including acoustic scattering. However, it lacked a formal convergence theory. This paper is a first step toward filling this gap. Indeed, it is proved that the hybrid variational formulation underlying this DGM is well-posed in the sense of Hadamard. In addition, a priori error estimates proved for the so-called R-4-1 element, that is, the simplest two-dimensional element associated with this discretization method, establish the convergence of this element and reveal its formal order of accuracy. Furthermore, an a posteriori error estimate was derived that can be used as a practical error indicator when refining the partition of the computational domain. Higher-order elements will be analyzed in future research.

Acknowledgment. The authors are grateful to the referees for their constructive suggestions and remarks.

REFERENCES

- [1] C. FARHAT, I. HARARI, AND L. P. FRANCA, *The discontinuous enrichment method*, Comput. Methods Appl. Mech. Engrg, 190 (2001), pp. 6455–6479.
- [2] C. FARHAT, I. HARARI, AND U. HETMANIUK, *The discontinuous enrichment method for multiscale analysis*, Comput. Methods Appl. Mech. Engrg, 192 (2003), pp. 3195–3210.
- [3] C. FARHAT, I. HARARI, AND U. HETMANIUK, *A discontinuous Galerkin method with Lagrange multipliers for the solution of Helmholtz problems in the mid-frequency regime*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 1389–1419.
- [4] M. E. ROSE, *Weak element approximations to elliptic differential equations*, Numer. Math., 24 (1975), pp. 185–204.
- [5] I. BABUŠKA I, AND J. M. MELENK, *The partition of unity method*, Internat. J. Numer. Methods Engrg., 40 (1997), pp. 727–758.
- [6] O. CESSENAT AND B. DESPRES, *Application of an ultra weak variational formulation of elliptic PDEs to the two-dimensional Helmholtz problem*, SIAM J. Numer. Anal., 35 (1998), pp. 255–299.
- [7] P. MONK AND D. Q. WANG, *A least-squares method for the Helmholtz equation*, Comput. Methods Appl. Mech. Engrg., 175 (1999), pp. 121–136.
- [8] C. FARHAT, P. WEIDEMANN-GOIRAN, AND R. TEZAUER, *A discontinuous Galerkin method with plane waves and Lagrange multipliers for the solution of short wave exterior Helmholtz problems on unstructured meshes*, Wave Motion, 39 (2004), pp. 307–317.
- [9] C. FARHAT, R. TEZAUER, AND P. WIEDEMANN-GOIRAN, *Higher-order extensions of a discontinuous Galerkin method for mid-frequency Helmholtz problems*, Internat. J. Numer. Methods Engrg., 61 (2004), pp. 1938–1956.
- [10] A. BAYLISS, C. I. GOLDSTEIN, AND E. TURKEL, *On accuracy conditions for the numerical computations of waves*, J. Comput. Phys., 59 (1985), pp. 396–404.
- [11] F. IHLENBURG, *Finite Element Analysis of Acoustic Scattering*, Appl. Math. Sci. 132, Springer-Verlag, New York, 1998.
- [12] J. HADAMARD, *Lectures on Cauchy's Problem in Linear Partial Differential Equations*, Yale University Press, New Haven, 1923.
- [13] D. COLTON AND R. KRESS, *Inverse Acoustic and Electromagnetic Scattering Theory*, Appl. Math. Sci. 93, Springer-Verlag, New York, 1992.
- [14] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
- [15] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
- [16] L. HÖRMANDER, *The Analysis of Linear Partial Differential Operator*, Springer-Verlag, New York, 1985.
- [17] M. E. TAYLOR, *Partial Differential Equations I: Basic Theory*, Springer-Verlag, New York, 1997.
- [18] M. MELENK, *On Generalized Finite Element Methods*, Ph.D. thesis, University of Maryland, College Park, MD, 1995.
- [19] U. HETMANIUK, *Stability estimates for a class of Helmholtz problems*, Commun. Math., Sci., 5 (2007), pp. 665–678.
- [20] J. L. LIONS AND E. MAGENES, *Non-homogeneous Boundary Value Problems and Applications, Volume I*, Springer-Verlag, New York, 1972.
- [21] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
- [22] P. GRISVARD, *Elliptic Problems in Non Smooth Domains*, Pitman, Boston, 1985.
- [23] J. P. AUBIN, *Analyse Fonctionnelle Appliquée*, Presse Universitaire de France, Paris, 1987.
- [24] J. NITSCHKE, *Ein Kriterium für die quasi-optimalität des Ritzchen Verfahrens*, Numer. Math., 11 (1968), pp. 346–348.
- [25] J. CEA, *Approximation variationnelle des problèmes aux limites*, Ann. Inst. Fourier, 14 (1964), pp. 345–444.

A CONVERGENT ADAPTIVE METHOD FOR ELLIPTIC EIGENVALUE PROBLEMS*

S. GIANI[†] AND I. G. GRAHAM[‡]

Abstract. We prove the convergence of an adaptive linear finite element method for computing eigenvalues and eigenfunctions of second-order symmetric elliptic partial differential operators. The weak form is assumed to yield a bilinear form which is bounded and coercive in H^1 . Each step of the adaptive procedure refines elements in which a standard a posteriori error estimator is large and also refines elements in which the computed eigenfunction has high oscillation. The error analysis extends the theory of convergence of adaptive methods for linear elliptic source problems to elliptic eigenvalue problems, and in particular deals with various complications which arise essentially from the nonlinearity of the eigenvalue problem. Because of this nonlinearity, the convergence result holds under the assumption that the initial finite element mesh is sufficiently fine.

Key words. second-order elliptic problems, eigenvalues, adaptive finite element methods, convergence

AMS subject classifications. 65N12, 65N25, 65N30, 65N50

DOI. 10.1137/070697264

1. Introduction. In the last decades, mesh adaptivity has been widely used to improve the accuracy of numerical solutions to many scientific problems. The basic idea is to refine the mesh only where the error is high, with the aim of achieving an accurate solution using an optimal number of degrees of freedom. There is a large amount of numerical analysis literature on adaptivity, in particular on reliable and efficient a posteriori error estimates (e.g., [1]). Recently, the question of convergence of adaptive methods has received intensive interest and a number of convergence results for the adaptive solution of boundary value problems have appeared (e.g., [8, 18, 19, 7, 6, 23]).

We prove here the convergence of an adaptive linear finite element algorithm for computing eigenvalues and eigenvectors of scalar symmetric elliptic partial differential operators in bounded polygonal or polyhedral domains, subject to Dirichlet boundary data. Such problems arise in many applications, e.g., resonance problems, nuclear reactor criticality, and the modelling of photonic band gap materials, to name but three.

Our refinement procedure is based on two locally defined quantities, firstly, a standard a posteriori error estimator and secondly a measure of the variability (or “oscillation”) of the computed eigenfunction. (Measures of “data oscillation” appear in the theory of adaptivity for boundary value problems, e.g., [18]. In the eigenvalue problem the computed eigenvalue and eigenfunction on the present mesh plays the role of “data” for the next iteration of the adaptive procedure.) Our algorithm performs local refinement on all elements on which the minimum of these two local quantities is sufficiently large. We prove that the adaptive method converges provided the initial mesh is sufficiently fine. The latter condition, while absent for adaptive methods for

*Received by the editors July 16, 2007; accepted for publication (in revised form) October 20, 2008; published electronically February 13, 2009.

<http://www.siam.org/journals/sinum/47-2/69726.html>

[†]School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD, UK (Stefano.Giani@nottingham.ac.uk).

[‡]Department of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, UK (I.G.Graham@bath.ac.uk).

linear symmetric elliptic boundary value problems, commonly appears for nonlinear problems and can be thought of as a manifestation of the nonlinearity of the eigenvalue problem.

We believe that the present paper is the first contribution to the topic of convergence of adaptive methods for eigenvalue problems. Since writing this paper, substantial improvements in the theory have been made in [5], where the need to adapt on the oscillations of the eigenvalue is removed and, in addition, the general convergence of the adaptive scheme to a nonspurious eigenvalue of the continuous problem is established.

The outline of the paper is as follows. In section 2 we briefly describe the model elliptic eigenvalue problem and the numerical method and in section 3 we describe a priori estimates, most of which are classical. Section 4 describes the a posteriori estimates and the adaptive algorithm. Section 5 proves that proceeding from one mesh to another ensures error reduction (up to oscillation of the computed eigenfunction) while the convergence result is presented in section 6. Numerical experiments illustrating the theory are presented in section 7.

2. Eigenvalue problem and numerical method. Throughout, Ω will denote a bounded domain in \mathbb{R}^d ($d = 2$ or 3). In fact, Ω will be assumed to be a polygon ($d = 2$) or polyhedron ($d = 3$). We will be concerned with the problem of finding an eigenvalue $\lambda \in \mathbb{R}$ and eigenfunction $0 \neq u \in H_0^1(\Omega)$ satisfying

$$(2.1) \quad a(u, v) := \lambda b(u, v), \quad \text{for all } v \in H_0^1(\Omega),$$

where, for real valued functions u and v ,

$$(2.2) \quad a(u, v) = \int_{\Omega} \nabla u(x)^T \mathcal{A}(x) \nabla v(x) dx \quad \text{and} \quad b(u, v) = \int_{\Omega} \mathcal{B}(x) u(x) v(x) dx .$$

Here, the matrix-valued function \mathcal{A} is required to be uniformly positive definite, i.e.,

$$(2.3) \quad 0 < \underline{a} \leq \xi^T \mathcal{A}(x) \xi \leq \bar{a} \quad \text{for all } \xi \in \mathbb{R}^d \quad \text{with } |\xi| = 1 \quad \text{and all } x \in \Omega.$$

The scalar function \mathcal{B} is required to be bounded above and below by positive constants for all $x \in \Omega$, i.e.,

$$(2.4) \quad 0 < \underline{b} \leq \mathcal{B}(x) \leq \bar{b} \quad \text{for all } x \in \Omega.$$

We will assume that \mathcal{A} and \mathcal{B} are both piecewise constant on Ω and that any jumps in \mathcal{A} and \mathcal{B} are aligned with the meshes \mathcal{T}_n (introduced below), for all n .

Throughout the paper, for any polygonal (polyhedral) subdomain of $D \subset \Omega$, and any $s \in [0, 1]$, $\|\cdot\|_{s,D}$ and $|\cdot|_{s,D}$ will denote the standard norm and seminorm in the Sobolev space $H^s(D)$. Also $(\cdot, \cdot)_{0,D}$ denotes the $L_2(D)$ inner product. We also define the energy norm induced by the bilinear form a :

$$\|u\|_{\Omega}^2 := a(u, u) \quad \text{for all } u \in H_0^1(\Omega),$$

which, by (2.3), is equivalent to the $H^1(\Omega)$ seminorm. (The equivalence constant depends on the contrast \bar{a}/\underline{a} , but we are not concerned with this dependence in the present paper.) We also introduce the weighted L_2 norm:

$$\|u\|_{0,\mathcal{B},\Omega}^2 = b(u, u) = \int_{\Omega} \mathcal{B}(x) |u(x)|^2 dx,$$

and note the norm equivalence

$$(2.5) \quad \sqrt{\underline{b}}\|v\|_{0,\Omega} \leq \|v\|_{0,\mathcal{B},\Omega} \leq \sqrt{\bar{b}}\|v\|_{0,\Omega}.$$

Rewriting the eigenvalue problem (2.1) in standard normalized form, we seek $(\lambda, u) \in \mathbb{R} \times H_0^1(\Omega)$ such that

$$(2.6) \quad \left. \begin{aligned} a(u, v) &= \lambda b(u, v), \quad \text{for all } v \in H_0^1(\Omega) \\ \|u\|_{0,\mathcal{B},\Omega} &= 1 \end{aligned} \right\}.$$

By the continuity of a and b and the coercivity of a on $H_0^1(\Omega)$ it is a standard result that (2.6) has a countable sequence of nondecreasing positive eigenvalues λ_j , $j = 1, 2, \dots$ with corresponding eigenfunctions $u_j \in H_0^1(\Omega)$ [3, 12, 24].

In this paper we will need some additional regularity for the eigenfunctions u_j , which will be achieved by making the following regularity assumption for the elliptic problem induced by a .

Assumption 2.1. We assume that there exists a constant $C_{\text{ell}} > 0$ and $s \in [0, 1]$ with the following property. For $f \in L_2(\Omega)$, if $v \in H_0^1(\Omega)$ solves the problem $a(v, w) = (f, w)_{0,\Omega}$ for all $w \in H_0^1(\Omega)$, then $\|v\|_{1+s,\Omega} \leq C_{\text{ell}}\|f\|_{0,\Omega}$.

Assumption 2.1 is satisfied with $s = 1$ when \mathcal{A} is constant (or smooth) and Ω is has a smooth boundary or is a convex polygon. In a range of other practical cases $s \in (0, 1)$, for example, Ω nonconvex (see [4]), or \mathcal{A} having a discontinuity across an interior interface (see [2]). Under Assumption 2.1 it follows that the eigenfunctions u_j of the problem (2.6) satisfy $\|u_j\|_{1+s,\Omega} \leq C_{\text{ell}}\lambda_j\sqrt{\bar{b}}$.

To approximate problem (2.6) we use the piecewise linear finite element method. Accordingly, let $\mathcal{T}_n, n = 1, 2, \dots$ denote a family of conforming triangular ($d = 2$) or tetrahedral ($d = 3$) meshes on Ω . Each mesh consists of elements denoted $\tau \in \mathcal{T}_n$. We assume that for each n , \mathcal{T}_{n+1} is a refinement of \mathcal{T}_n . For a typical element τ of any mesh, its diameter is denoted H_τ and the diameter of its largest inscribed ball is denoted ρ_τ . For each n , let H_n denote the piecewise constant mesh function on Ω , whose value on each element $\tau \in \mathcal{T}_n$ is H_τ and let $H_n^{\text{max}} = \max_{\tau \in \mathcal{T}_n} H_\tau$. Throughout we will assume that the family of meshes \mathcal{T}_n is shape regular; i.e., there exists a constant C_{reg} such that

$$(2.7) \quad H_\tau \leq C_{\text{reg}}\rho_\tau, \quad \text{for all } \tau \in \mathcal{T}_n \quad \text{and all } n = 1, 2, \dots$$

In the later sections of the paper, the \mathcal{T}_n will be produced by an adaptive process which ensures shape regularity.

We let V_n denote the usual finite dimensional subspace of $H_0^1(\Omega)$, consisting of all continuous piecewise linear functions with respect to the mesh \mathcal{T}_n . Then the discrete formulation of problem (2.6) is to seek the eigenpairs $(\lambda_n, u_n) \in \mathbb{R} \times V_n$ such that

$$(2.8) \quad \left. \begin{aligned} a(u_n, v_n) &= \lambda_n b(u_n, v_n), \quad \text{for all } v_n \in V_n \\ \|u_n\|_{0,\mathcal{B},\Omega} &= 1. \end{aligned} \right\}$$

The problem (2.8) has $N = \dim V_n$ positive eigenvalues (counted according to multiplicity) which we denote in nondecreasing order as $\lambda_{n,1} \leq \lambda_{n,2} \leq \dots \leq \lambda_{n,N}$. It is well-known (see [24, section 6.3]) that for any j , $\lambda_{n,j} \rightarrow \lambda_j$ as $H_n^{\text{max}} \rightarrow 0$ and (by the minimax principle—see, e.g., [24, section 6.1]) the convergence of the $\lambda_{n,j}$ is monotone decreasing, i.e.,

$$(2.9) \quad \lambda_{n,j} \geq \lambda_{m,j} \geq \lambda_j, \quad \text{for all } j = 1, \dots, N, \quad \text{and all } m \geq n.$$

Thus, it is clear that there exists a *separation constant* $\rho > 0$ (depending on the spectrum of (2.6)) with the following property: If $\lambda_j = \lambda_{j+1} = \dots = \lambda_{j+R-1}$ is any eigenvalue of (2.6) of multiplicity $R \geq 1$, then

$$(2.10) \quad \frac{\lambda_j}{|\lambda_{n,\ell} - \lambda_j|} \leq \rho, \quad \ell \neq j, j+1, \dots, j+R-1,$$

provided H_n^{\max} is sufficiently small. (Note that for $\ell \neq j, j+1, \dots, j+R-1$, $\lambda_{n,\ell} \rightarrow \lambda_\ell \neq \lambda_j$.)

The a priori error analysis for our eigenvalue problem is classical (see, e.g., [3], [12], and [24]). In the next section, we briefly recall some of the main known results and also prove a nonclassical result (Theorem 3.2) which is essential to the proof of convergence of our adaptive scheme.

3. A priori analysis. In this section we shall assume that λ_j is an eigenvalue of (2.6) and $\lambda_{n,j}$ is its approximation as described above. Let u_j and $u_{n,j}$ be any corresponding normalized eigenvectors as defined in (2.6) and (2.8). From these we obtain the important basic identity:

$$(3.1) \quad \begin{aligned} a(u_j - u_{n,j}, u_j - u_{n,j}) &= a(u_j, u_j) + a(u_{n,j}, u_{n,j}) - 2a(u_j, u_{n,j}) \\ &= \lambda_j + \lambda_{n,j} - 2\lambda_j b(u_j, u_{n,j}) \\ &= \lambda_{n,j} - \lambda_j + \lambda_j (2 - 2b(u_j, u_{n,j})) \\ &= \lambda_{n,j} - \lambda_j + \lambda_j b(u_j - u_{n,j}, u_j - u_{n,j}). \end{aligned}$$

Using this and (2.9), we obtain

$$(3.2) \quad |||u_j - u_{n,j}|||_\Omega^2 = |\lambda_j - \lambda_{n,j}| + \lambda_j \|u_j - u_{n,j}\|_{0,\mathcal{B},\Omega}^2.$$

The following theorem investigates the convergence of discrete eigenpairs. Although parts of it are very well-known, we do not know a suitable reference for all the results given below, so a brief proof is given for completeness. In the proof we make use of the orthogonal projection Q_n of $H_0^1(\Omega)$ onto V_n with respect to the inner product induced by $a(\cdot, \cdot)$, which has the property:

$$(3.3) \quad a(Q_n u, v_n) = \lambda b(u, v_n) \quad \text{for all } v_n \in V_n.$$

In the main result of this paper we prove convergence for adaptive approximations to eigenvalues and eigenvectors *assuming for simplicity a simple eigenvalue*. The following preliminary theorem is stated for a simple eigenvalue. However, this result is known for multiple eigenvalues (see, e.g., [24]). More details are given in [10].

THEOREM 3.1. *Let λ_j be a simple eigenvalue of (2.6), let $\lambda_{n,j}$ be its associated approximation from solving (2.8), and let u_j and $u_{n,j}$ be any corresponding normalized eigenvectors. Then for all $1 \leq j \leq N$,*

(i)

$$(3.4) \quad |\lambda_j - \lambda_{n,j}| \leq |||u_j - u_{n,j}|||_\Omega^2;$$

(ii) *There are constants $C_1, C_2 > 0$ and scalars $\alpha_{n,j} \in \{\pm 1\}$ such that*

$$(3.5) \quad \begin{aligned} \|u_j - \alpha_{n,j} u_{n,j}\|_{0,\mathcal{B},\Omega} &\leq C_1 (H_n^{\max})^s |||u_j - Q_n u_j|||_\Omega \\ &\leq C_1 (H_n^{\max})^s |||u_j - \alpha_{n,j} u_{n,j}|||_\Omega, \end{aligned}$$

where s is as in Assumption 2.1.

(iii) For sufficiently small H_n^{\max} there is a constant C_2 such that

$$(3.6) \quad \| |u_j - \alpha_{n,j} u_{n,j}| \|_{\Omega} \leq C_2 (H_n^{\max})^s.$$

The constants C_1, C_2 depend on the spectral information $\lambda_\ell, u_\ell, \ell = 1, \dots, j$, the separation constant ρ , the constants $C_{\text{ell}}, C_{\text{reg}}$ in Assumption 2.1 and in (2.7) and on the bounds $\bar{a}, \underline{a}, \bar{b}, \underline{b}$ in (2.3), (2.4).

Proof. The estimate (3.4) follows directly from (3.2). Note that (3.4) holds even if $u_{n,j}$ is not close to u , which may occur due to the nonuniqueness of the eigenvectors.

The proof of (3.5) is obtained by a reworking of the results in [24]. By the symmetry of a and b there exists a basis $\{u_{n,\ell} : \ell = 1, \dots, N\}$ of V_n (containing $u_{n,j}$) which is orthonormal with respect to inner product b , and each $u_{n,\ell}$ is an eigenvector of (2.8) corresponding to eigenvalue $\lambda_{n,\ell}$. Then with $\beta_{n,j} := b(Q_n u_j, u_{n,j})$, Parseval's equality yields

$$(3.7) \quad \|Q_n u_j - \beta_{n,j} u_{n,j}\|_{0,\mathcal{B},\Omega}^2 = \sum_{\substack{\ell=1 \\ \ell \neq j}}^N b(Q_n u_j, u_{n,\ell})^2.$$

Then, since

$$\lambda_{n,\ell} b(Q_n u_j, u_{n,\ell}) = a(Q_n u_j, u_{n,\ell}) = a(u_j, u_{n,\ell}) = \lambda_j b(u_j, u_{n,\ell}),$$

we have $(\lambda_{n,\ell} - \lambda_j) b(Q_n u_j, u_{n,\ell}) = \lambda_j b(u_j - Q_n u_j, u_{n,\ell})$, and so

$$\begin{aligned} \|Q_n u_j - \beta_{n,j} u_{n,j}\|_{0,\mathcal{B},\Omega}^2 &= \sum_{\substack{\ell=1 \\ \ell \neq j}}^N \left(\frac{\lambda_j}{\lambda_{n,\ell} - \lambda_j} \right)^2 b(u_j - Q_n u_j, u_{n,\ell})^2 \\ &\leq \rho^2 \sum_{\substack{\ell=1 \\ \ell \neq j}}^N b(u_j - Q_n u_j, u_{n,\ell})^2 \leq \rho^2 \|u_j - Q_n u_j\|_{0,\mathcal{B},\Omega}^2, \end{aligned}$$

with the last step again by Parseval's equality. Hence,

$$(3.8) \quad \|u_j - \beta_{n,j} u_{n,j}\|_{0,\mathcal{B},\Omega} \leq (1 + \rho) \|u_j - Q_n u_j\|_{0,\mathcal{B},\Omega}.$$

Moreover,

$$\|u_j\|_{0,\mathcal{B},\Omega} - \|u_j - \beta_{n,j} u_{n,j}\|_{0,\mathcal{B},\Omega} \leq \|\beta_{n,j} u_{n,j}\|_{0,\mathcal{B},\Omega} \leq \|u_j\|_{0,\mathcal{B},\Omega} + \|u_j - \beta_{n,j} u_{n,j}\|_{0,\mathcal{B},\Omega}.$$

Since the u_j and the $u_{n,j}$ are normalized, this implies

$$1 - \|u_j - \beta_{n,j} u_{n,j}\|_{0,\mathcal{B},\Omega} \leq |\beta_{n,j}| \leq 1 + \|u_j - \beta_{n,j} u_{n,j}\|_{0,\mathcal{B},\Omega}$$

and, combining these with (3.8), we have

$$\| |\beta_{n,j}| - 1 | \leq (1 + \rho) \|u_j - Q_n u_j\|_{0,\mathcal{B},\Omega}.$$

Thus, with $\alpha_{n,j} := \text{sign}(\beta_{n,j})$, we have $|\beta_{n,j} - \alpha_{n,j}| \leq (1 + \rho) \|u_j - Q_n u_j\|_{0,\mathcal{B},\Omega}$, and

$$\|u_j - \alpha_{n,j} u_{n,j}\|_{0,\mathcal{B},\Omega} \leq 2(1 + \rho) \|u_j - Q_n u_j\|_{0,\mathcal{B},\Omega}.$$

The first inequality in (3.5) now follows from an application of the standard Aubin–Nitsche duality argument, while the second is just the best approximation of Q_n in the energy norm.

The proof of (3.6) is a slight modification of that given in [24, Theorem 6.2]. The argument consists of obtaining an $\mathcal{O}((H_n^{\max})^{2s})$ estimate for the eigenvalue error $|\lambda_j - \lambda_{n,j}|$ and then combining this with (3.2) and (3.5). \square

The next theorem is a generalization to eigenvalue problems of the standard monotone convergence property for linear symmetric elliptic PDEs, namely, that if one enriches the finite dimensional space, then the error is bound to decrease. This result fails to hold for eigenvalue problems (even for symmetric elliptic partial differential operators), because of the nonlinearity of such problems. The best that we can do is to show that if the finite dimensional space is enriched, then the error will not increase very much. This is the subject of Theorem 3.2.

THEOREM 3.2. *For any $1 \leq j \leq N$, there exists a constant $q > 1$ such that, for $m \geq n$, the corresponding computed eigenpair $(\lambda_{m,j}, u_{m,j})$ satisfies:*

$$(3.9) \quad \|u_j - \alpha_{m,j}u_{m,j}\|_{\Omega} \leq q \|u_j - \alpha_{n,j}u_{n,j}\|_{\Omega}.$$

Proof. From Theorem 3.1 (ii), we obtain

$$(3.10) \quad \|u_j - \alpha_{m,j}u_{m,j}\|_{0,\mathcal{B},\Omega} \leq C_1(H_m^{\max})^s \|u_j - Q_m u_j\|_{\Omega}.$$

Since \mathcal{T}_m is a refinement of \mathcal{T}_n , it follows that $V_n \subset V_m$ and so the best approximation property of Q_m ensures that

$$\|u_j - Q_m u_j\|_{\Omega} \leq \|u_j - Q_n u_j\|_{\Omega}.$$

Hence, from (3.10) and using the fact that $H_m^{\max} \leq H_n^{\max}$, we have

$$(3.11) \quad \|u_j - \alpha_{m,j}u_{m,j}\|_{0,\mathcal{B},\Omega} \leq C_1(H_n^{\max})^s \|u_j - Q_n u_j\|_{\Omega}.$$

Recalling that (3.2) holds for all eigenfunctions, and using (3.11) and then (2.9), we obtain

$$(3.12) \quad \begin{aligned} \|u_j - \alpha_{m,j}u_{m,j}\|_{\Omega}^2 &\leq |\lambda_j - \lambda_{m,j}| + \lambda_j \|u_j - \alpha_{m,j}u_{m,j}\|_{0,\mathcal{B},\Omega}^2 \\ &\leq |\lambda_j - \lambda_{m,j}| + \lambda_j C_1^2 (H_n^{\max})^{2s} \|u_j - Q_n u_j\|_{\Omega}^2 \\ &\leq |\lambda_j - \lambda_{n,j}| + \lambda_j C_1^2 (H_n^{\max})^{2s} \|u_j - Q_n u_j\|_{\Omega}^2. \end{aligned}$$

Hence, from (3.4) we obtain

$$(3.13) \quad \|u_j - \alpha_{m,j}u_{m,j}\|_{\Omega}^2 \leq \|u_j - \alpha_{n,j}u_{n,j}\|_{\Omega}^2 + \lambda_j C_1^2 (H_n^{\max})^{2s} \|u_j - Q_n u_j\|_{\Omega}^2.$$

But, since Q_n yields the best approximation from V_n in the energy norm, we have

$$(3.14) \quad \|u_j - \alpha_{m,j}u_{m,j}\|_{\Omega}^2 \leq (1 + \lambda_j C_1^2 (H_n^{\max})^{2s}) \|u_j - \alpha_{n,j}u_{n,j}\|_{\Omega}^2,$$

which is in the required form. \square

Remark 3.3. From now on we will be concerned with a true eigenpair (λ_j, u_j) and its computed approximation $(\lambda_{j,n}, u_{j,n})$ on the mesh \mathcal{T}_n . Theorem 3.1 tells us that a priori $\lambda_{n,j}$ is “close” to λ_j and that the spaces spanned by u_j and $u_{n,j}$ are close. From now on we drop the subscript j and we simply write (λ, u) for the eigenpair of (2.6) (λ_n, u_n) for a corresponding eigenpair of (2.8) and the scalar $\alpha_{n,j}$ is abbreviated α_n .

4. A posteriori analysis. This section contains our a posteriori error estimator and the definition of the adaptive algorithm for which convergence will be proved in the following sections.

Recalling the mesh sequence \mathcal{T}_n defined above, we let \mathcal{S}_n denote the set of all the interior edges (or the set of interior faces in 3D) of the elements of the mesh \mathcal{T}_n . For each $S \in \mathcal{S}_n$, we denote by $\tau_1(S)$ and $\tau_2(S)$ the elements sharing S (i.e., $\tau_1(S) \cap \tau_2(S) = S$) and we write $\Omega(S) = \tau_1(S) \cup \tau_2(S)$. We let \vec{n}_S denote the unit normal vector to S , orientated from $\tau_1(S)$ to $\tau_2(S)$. All elements, faces, and edges are considered to be closed sets. Furthermore, we denote the diameter of S by H_S . Note that, by mesh regularity, $\text{diam}(\Omega(S)) \sim H_{\tau_i(S)}$, $i = 1, 2$.

NOTATION 4.1. We write $A \lesssim B$ when A/B is bounded by a constant which may depend on the functions \mathcal{A} and \mathcal{B} in (2.2), on $\underline{a}, \bar{a}, \underline{b}$, and \bar{b} , on C_{ell} in Assumption 2.1, C_{reg} in (2.7). The notation $A \cong B$ means $A \lesssim B$ and $A \gtrsim B$.

All the constants depending on the spectrum, namely, ρ in (2.10), q in (3.9), C_1 and C_2 in (3.5) and (3.6), are handled explicitly. Similarly all mesh size dependencies are explicit. Note that all eigenvalues of (2.8) satisfy $\lambda_n \gtrsim 1$, since $\lambda_n \geq \lambda_1 = a(u_1, u_1) \gtrsim |u_1|_{1,\Omega}^2 \gtrsim \|u_1\|_{0,\Omega}^2 \gtrsim \|u_1\|_{0,\mathcal{B},\Omega}^2 = 1$.

Our error estimator is obtained by adapting standard estimates for source problems to the eigenvalue problem. Analogous eigenvalue estimates can be found in [9] (for the Laplace problem) and [25] (for linear elasticity) and related results are in [14].

For a function g , which is piecewise continuous on the mesh \mathcal{T}_n , we introduce its jump across an edge (face) $S \in \mathcal{S}_n$ by:

$$[g]_S(x) := \left(\lim_{\tilde{x} \rightarrow x} g(\tilde{x}) - \lim_{\tilde{x} \in \tau_2(S)} g(\tilde{x}) \right), \quad \text{for } x \in \text{int}(S).$$

Then for any function v with piecewise continuous gradient on \mathcal{T}_n we define, for $S \in \mathcal{S}_n$,

$$J_S(v)(x) := [\vec{n}_S \cdot \mathcal{A}\nabla v]_S(x), \quad \text{for } x \in \text{int}(S).$$

The error estimator η_n on the mesh \mathcal{T}_n is defined as

$$(4.1) \quad \eta_n^2 := \sum_{S \in \mathcal{S}_n} \eta_{S,n}^2,$$

where, for each $S \in \mathcal{S}_n$,

$$(4.2) \quad \eta_{S,n}^2 := \|H_n \lambda_n u_n\|_{0,\mathcal{B},\Omega(S)}^2 + \left\| H_S^{1/2} J_S(u_n) \right\|_{0,S}^2.$$

The following lemma is proved, in a standard way, by adapting the usual arguments for linear source problems. Note again that λ is an eigenvalue of (2.6), λ_n is a nearby eigenvalue of (2.8), and u, u_n are any corresponding normalized eigenfunctions which are only “near” in the sense of Theorem 3.1.

LEMMA 4.2 (reliability).

$$(4.3) \quad \| \|u - u_n \| \|_{\Omega} \lesssim \eta_n + G_n,$$

and

$$(4.4) \quad G_n := \frac{1}{2}(\lambda + \lambda_n) \frac{\|u - u_n\|_{0,\mathcal{B},\Omega}^2}{\| \|u - u_n \| \|_{\Omega}}.$$

Remark 4.3. Recalling Remark 3.3, u_n in Lemma 4.2 is any normalized eigenvector of (2.8) corresponding to the simple eigenvalue λ ; i.e., its sign is not unique. However, the error estimators $\eta_{S,n}$ are independent of the sign of u_n . This is not a contradiction: we shall see that only one choice of eigenfunction will guarantee that the second term on the right-hand side of (4.3) is small, and only in this case is the left-hand side also guaranteed to be small.

A similar result to Lemma 4.2 was proved in [25, Proposition 5].

Proof. To ease readability we set $e_n = u - u_n$ in the proof. Note first that, since (λ, u) and (λ_n, u_n) , respectively, solve the eigenvalue problems (2.1) and (2.8), we have, for all $w_n \in V_n$,

$$\begin{aligned}
 \|e_n\|_{\Omega}^2 &= a(e_n, e_n) \\
 &= a(e_n, e_n - w_n) + a(e_n, w_n) \\
 &= a(e_n, e_n - w_n) + a(u, w_n) - a(u_n, w_n) \\
 &= a(e_n, e_n - w_n) + b(\lambda u - \lambda_n u_n, w_n) \\
 (4.5) \quad &= a(e_n, e_n - w_n) - b(\lambda u - \lambda_n u_n, e_n - w_n) + b(\lambda u - \lambda_n u_n, e_n).
 \end{aligned}$$

To estimate the first two terms on the right-hand side of (4.5), first note that, for all $v \in H_0^1(\Omega)$,

$$a(e_n, v) - b(\lambda u - \lambda_n u_n, v) = -a(u_n, v) + \lambda_n b(u_n, v).$$

Hence, using elementwise integration by parts (and the fact that $\mathcal{A}\nabla u_n$ is constant on each element and v vanishes on $\partial\Omega$), we obtain

$$\begin{aligned}
 a(e_n, v) - b(\lambda u - \lambda_n u_n, v) &= - \sum_{\tau \in \mathcal{T}_n} \int_{\tau} (\mathcal{A}\nabla u_n) \cdot \nabla v + \lambda_n b(u_n, v) \\
 (4.6) \quad &= - \sum_{S \in \mathcal{S}_n} \int_S J_S(u_n) v + \lambda_n b(u_n, v),
 \end{aligned}$$

and hence, for all $w_n \in V_n$,

$$\begin{aligned}
 (4.7) \quad & a(e_n, e_n - w_n) - b(\lambda u - \lambda_n u_n, e_n - w_n) = - \sum_{S \in \mathcal{S}_n} \int_S J_S(u_n) (e_n - w_n) + \lambda_n b(u_n, e_n - w_n).
 \end{aligned}$$

Now recall the Scott–Zhang quasi-interpolation operator ([22]) which has the property that, for all $v \in H_0^1(\Omega)$, $I_n v \in V_n$ and

$$(4.8) \quad \|v - I_n v\|_{0,\tau} \lesssim H_{\tau} |v|_{1,\omega(\tau)}, \quad \|v - I_n v\|_{0,S} \lesssim H_S^{\frac{1}{2}} |v|_{1,\omega(S)},$$

where $\omega(\tau)$ is the union of all elements sharing at least a point with τ , and $\omega(S)$ is the union of all elements sharing at least a point with S . (Note $\Omega(S) \subseteq \omega(S)$.) Substituting $w_n = I_n e_n$ in (4.7) and using the Cauchy–Schwarz inequality, together with estimates (4.8), we obtain

$$(4.9) \quad a(e_n, e_n - w_n) - b(\lambda u - \lambda_n u_n, e_n - w_n) \lesssim \eta_n \|e_n\|_{\Omega}.$$

To estimate the third term on the right-hand side of (4.5), we simply observe that due to the normalization in each of the eigenvalue problems (2.1) and (2.8) we have

$$(4.10) \quad b(\lambda u - \lambda_n u_n, e_n) = (\lambda + \lambda_n)(1 - b(u, u_n)) = \frac{1}{2}(\lambda + \lambda_n) \|e_n\|_{0,\mathcal{B},\Omega}^2.$$

Now, combine (4.9) and (4.10) with (4.5) and divide by $\|e_n\|_\Omega$ to obtain the result. \square

Remark 4.4. We shall see below that G_n defined above constitutes a “higher order term”.

For mesh refinement based on the local contributions to η_n , we use the same marking strategy as in [8] and [18]. The idea is to refine a subset of the elements of \mathcal{T}_n whose side residuals sum up to a fixed proportion of the total residual η_n .

DEFINITION 4.5 (marking strategy 1). *Given a parameter $0 < \theta < 1$, the procedure is: mark the sides in a minimal subset $\hat{\mathcal{S}}_n$ of \mathcal{S}_n such that*

$$(4.11) \quad \left(\sum_{S \in \hat{\mathcal{S}}_n} \eta_{S,n}^2 \right)^{1/2} \geq \theta \eta_n.$$

To compute $\hat{\mathcal{S}}_n$, we compute all the “local residuals” $\eta_{S,n}$, then insert edges (faces) into $\hat{\mathcal{S}}_n$ in order of nonincreasing magnitude of $\eta_{S,n}$, until (4.11) is satisfied. A minimal subset $\hat{\mathcal{S}}_n$ may not be unique. After this is done, we construct another set $\hat{\mathcal{T}}_n$, containing all the elements of \mathcal{T}_n , which contain at least one edge (face) belonging to $\hat{\mathcal{S}}_n$.

In order to prove our convergence theory, we require an additional marking strategy based on oscillations (Definition 4.7 below). This also appears in some theories of adaptivity for source problems, e.g., [8], [18], [16], [7], and [6]), but to our knowledge has not yet been used in connection with eigenvalue problems.

The concept of “oscillation” is just a measure of how well a function may be approximated by piecewise constants on a particular mesh. For any function $v \in L_2(\Omega)$, and any mesh \mathcal{T}_n , we introduce its orthogonal projection $P_n v$ onto piecewise constants defined by

$$(4.12) \quad (P_n v)|_\tau = \frac{1}{|\tau|} \int_\tau v_n, \quad \text{for all } \tau \in \mathcal{T}_n.$$

Then we make the definition:

DEFINITION 4.6 (oscillations). *On a mesh \mathcal{T}_n , we define*

$$(4.13) \quad \text{osc}(v, \mathcal{T}_n) := \|H_n(v - P_n v)\|_{0,\mathcal{B},\Omega}.$$

Note that

$$\text{osc}(v, \mathcal{T}_n) = \left(\sum_{\tau \in \mathcal{T}_n} H_\tau^2 \|v - P_n v\|_{0,\mathcal{B},\tau}^2 \right)^{1/2},$$

and that (by standard approximation theory and the ellipticity of $a(\cdot, \cdot)$),

$$(4.14) \quad \text{osc}(v, \mathcal{T}_n) \lesssim (H_n^{\max})^2 \|v\|_\Omega, \quad \text{for all } v \in H_0^1(\Omega).$$

The second marking strategy (introduced below) aims to reduce the oscillations corresponding to a particular approximate eigenfunction u_n .

DEFINITION 4.7 (marking strategy 2). *Given a parameter $0 < \tilde{\theta} < 1$: mark the elements in a minimal subset $\tilde{\mathcal{T}}_n$ of \mathcal{T}_n such that*

$$(4.15) \quad \text{osc}(u_n, \tilde{\mathcal{T}}_n) \geq \tilde{\theta} \text{osc}(u_n, \mathcal{T}_n).$$

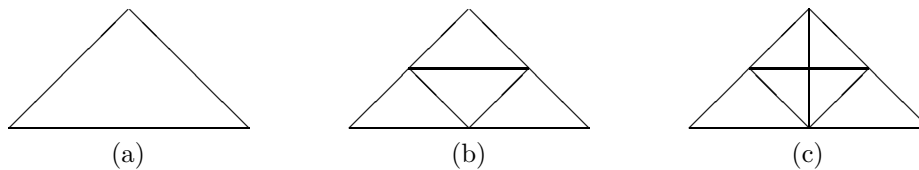


FIG. 4.1. The refinement procedure applied to an element of the mesh. In (a) the element before the refinement, in (b) after the three sides have been refined, and in (c) after the bisection of one of the three new segments.

Analogously to (4.11), we compute $\tilde{\mathcal{T}}_n$ by inserting elements τ into $\tilde{\mathcal{T}}_n$ according to nonincreasing order of their local contributions $H_\tau^2 \| (u_n - P_n u_n) \|_{0, \mathcal{B}, \tau}^2$ until (4.15) is satisfied.

Our adaptive algorithm can then be stated:

Algorithm 1 Converging algorithm

Require: $0 < \theta < 1$

Require: $0 < \tilde{\theta} < 1$

loop

Solve the Problem (2.8) for (λ_n, u_n)

Mark the elements using the first marking strategy (Definition 4.5)

Mark any additional unmarked elements using the second marking strategy (Definition 4.7)

Refine the mesh \mathcal{T}_n and construct \mathcal{T}_{n+1}

end loop

In 2D at the n th iteration in Algorithm 1 each element in the set $\hat{\mathcal{T}}_n \cup \tilde{\mathcal{T}}_n$ is refined using the algorithm illustrated in Figure 4.1. This consists of three recursive applications of the newest node algorithm [17] to each marked triangle, first creating two sons, then four grandsons, and finally bisecting two of the grandsons. This well-known algorithm is stated without name in [18, section 5.1]), is called “bisection5” in [7] and is called “full refinement” in [23]. This technique creates of a new node in the middle of each marked side in $\hat{\mathcal{S}}_n$ and also a new node in the interior of each marked element. It follows from [17] that this algorithm yields shape regular conforming meshes in 2D.

In the 3D case we use a suitable refinement that creates a new node on each marked face in $\hat{\mathcal{S}}_n$ and a node in the interior of each marked element.

In [18] and [16] it has been shown for linear source problems that the reduction of the error, as the mesh is refined, is triggered by the decay of oscillations of the source on the sequence of constructed meshes. For the eigenvalue problem (2.1) the quantity λu plays the role of data and in principle we have to ensure that oscillations of this quantity (or, more precisely, of its finite element approximation $\lambda_n u_n$) are sufficiently small. However, $\lambda_n u_n$ may change if the mesh changes and so the proof of error reduction for eigenvalue problems is not as simple as it is for linear source problems. This is the essence of the theoretical difficulty dealt with in this paper.

5. Error reduction. In this section we give the proof of error reduction for Algorithm 1. The proof has been inspired by the corresponding theory for source problems in [18]. However, the nonlinearity of the eigenvalue problem introduces new complications, and there are several lemmas before the main theorem (Theorem 5.6). For the rest of the section let (λ_n, u_n) be an approximate eigenpair on a mesh \mathcal{T}_n , let

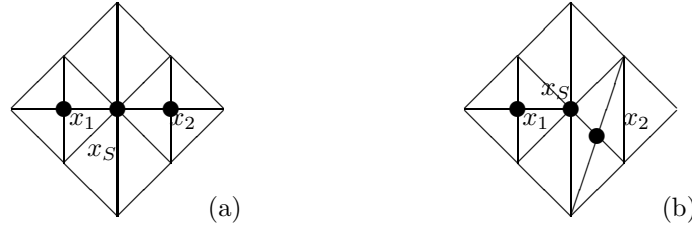


FIG. 5.1. Two cases of refined couples of elements.

\mathcal{T}_{n+1} be the mesh obtained by one iteration of Algorithm 1, and let (λ_{n+1}, u_{n+1}) be the corresponding eigenpair in the sense made precise in Remark 3.3.

The first lemma uses ideas from [18, Lemma 4.2] for the 2D case. The extension of this lemma to the 3D case is treated in Remark 5.2.

LEMMA 5.1. *Consider the 2D case. Let $\hat{\mathcal{S}}_n$ be as defined in Definition 4.5 and let P_n be as defined in (4.12). For any $S \in \hat{\mathcal{S}}_n$, there exists a function $\Phi_S \in V_{n+1}$ such that $\text{supp}(\Phi_S) = \Omega(S)$ and also*

$$(5.1) \quad \lambda_n \int_{\Omega(S)} \mathcal{B}(P_n u_n) \Phi_S - \int_S J_S(u_n) \Phi_S = \|H_n \lambda_n P_n u_n\|_{0,\mathcal{B},\Omega(S)}^2 + \left\| H_S^{1/2} J_S(u_n) \right\|_{0,S}^2,$$

and

$$(5.2) \quad \|\Phi_S\|_{\Omega(S)}^2 \lesssim \|H_n \lambda_n P_n u_n\|_{0,\mathcal{B},\Omega(S)}^2 + \left\| H_S^{1/2} J_S(u_n) \right\|_{0,S}^2,$$

where $\|v\|_{\Omega(S)}^2 := \int_{\Omega(S)} \nabla v^T \mathcal{A} \nabla v$.

Proof. Figure 5.1 illustrates two possible configurations of the domain $\Omega(S)$.

We then define

$$(5.3) \quad \Phi_S := \alpha_S \varphi_S + \beta_1 \varphi_1 + \beta_2 \varphi_2,$$

where φ_S and φ_i are the nodal basis functions associated with the points x_S and x_i on \mathcal{T}_{n+1} , and α_S, β_i are defined by

$$(5.4) \quad \alpha_S = \begin{cases} \frac{\left\| H_S^{1/2} J_S(u_n) \right\|_{0,S}^2}{\int_S J_S(u_n) \varphi_S} & \text{if } J_S(u_n) \neq 0, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$(5.5) \quad \beta_i = \begin{cases} \frac{\|H_n \lambda_n P_n u_n\|_{0,\mathcal{B},\tau_i(S)}^2 - \alpha_S \int_{\tau_i(S)} \mathcal{B} \lambda_n(P_n u_n) \varphi_S}{\int_{\tau_i(S)} \mathcal{B} \lambda_n(P_n u_n) \varphi_i} & \text{if } P_n u_n|_{\tau_i(S)} \neq 0, \\ 0 & \text{otherwise,} \end{cases}$$

for $i = 1, 2$.

Note that $J_S(u_n)$ and $P_n u_n$ are constant on each element τ . Using the fact that $\text{supp}(\varphi_i) = \tau_i(S)$, for $i = 1, 2$ we can easily see that the above formulae imply

$$(5.6) \quad \alpha_S \int_S J_S(u_n) \varphi_S = - \left\| H_S^{1/2} J_S(u_n) \right\|_{0,S}^2,$$

$$(5.7) \quad \int_{\Omega(S)} \mathcal{B} \lambda_n (P_n u_n) (\alpha_S \varphi_S + \beta_1 \varphi_1 + \beta_2 \varphi_2) = \|H_n \lambda_n P_n u_n\|_{0,\mathcal{B},\Omega(S)}^2$$

(and that these formulae remain true even if $J_S(u_n)$ or $P_n u_n|_{\tau_i(S)}$ vanish). Hence,

$$\begin{aligned} \lambda_n \int_{\Omega(S)} \mathcal{B}(P_n u_n) \Phi_S - \int_S J_S(u_n) \Phi_S \\ = \lambda_n \int_{\Omega(S)} \mathcal{B}(P_n u_n) (\alpha_S \varphi_S + \beta_1 \varphi_1 + \beta_2 \varphi_2) - \alpha_S \int_S J_S(u_n) \varphi_S \end{aligned}$$

and (5.1) follows immediately on using (5.6) and (5.7).

To proceed from here note that by the shape-regularity of the mesh and the standard inverse estimate,

$$\|\phi_S\|_{\Omega(S)} \lesssim H_S^{-1} \|\phi_S\|_{0,\Omega(S)}.$$

Also, for all elements $\tau \in \mathcal{T}_{n+1}$ with $\tau \subset \text{supp } \phi_S$, there exists an affine map $\chi : \hat{\tau} \rightarrow \tau$, where $\hat{\tau}$ is the unit simplex in \mathbb{R}^2 and $\hat{\phi}_S := \phi_S \circ \chi$ is a nodal basis function on $\hat{\tau}$. The Jacobian J_χ of χ is constant and is proportional to the area of τ . Hence,

$$\|\phi_S\|_{0,\tau}^2 = \int_\tau |\phi_S|^2 = \int_{\hat{\tau}} |\hat{\phi}_S|^2 J_\chi \sim H_S^2,$$

which ensures that $\|\phi_S\|_{\Omega(S)} \lesssim 1$ and, similarly, $\|\varphi_i\|_{\Omega(S)} \lesssim 1$. Combining these with (5.3), we obtain

$$(5.8) \quad \|\Phi_S\|_{\Omega(S)}^2 \lesssim |\alpha_S|^2 + |\beta_1|^2 + |\beta_2|^2.$$

Now, note that by a simple change of variable, $\int_S \varphi_S$ is the integral over $[-H_S/2, H_S/2]$ of the one-dimensional hat function centered on 0 and so $\int_S \varphi_S \sim H_S$. Since $J_S(u_n)$ is constant on S , we have

$$(5.9) \quad |\alpha_S| \lesssim \frac{|J_S(u_n)| \left\| H_S^{1/2} \right\|_{0,S}^2}{H_S} \lesssim |J_S(u_n)| H_S \sim \left\| H_S^{1/2} J_S(u_n) \right\|_{0,S}.$$

Also, since $P_n u_n$ is constant on each $\tau_i(S)$ and, since $\int_{\tau_i(S)} \mathcal{B} \phi_i \sim H_{\tau_i(S)}^2$, we have

$$\begin{aligned} |\beta_i| &\lesssim \frac{\lambda_n | (P_n u_n)|_{\tau_i(S)} | \|H_n\|_{0,\mathcal{B},\tau_i(S)}^2 + |\alpha_S| H_{\tau_i(S)}^2}{H_{\tau_i(S)}^2} \\ &\lesssim \lambda_n | (P_n u_n)|_{\tau_i(S)} | H_{\tau_i(S)}^2 + |\alpha_S| \sim \lambda_n \|H_n P_n u_n\|_{0,\mathcal{B},\tau_i(S)} + |\alpha_S|. \end{aligned}$$

This implies

$$(5.10) \quad |\beta_i|^2 \lesssim \|\lambda_n H_n P_n u_n\|_{0,\mathcal{B},\tau_i(S)}^2 + |\alpha_S|^2 \lesssim \|\lambda_n H_n P_n u_n\|_{0,\mathcal{B},\tau_i(S)}^2 + \left\| H_S^{1/2} J_S(u_n) \right\|_{0,S}^2,$$

and the proof is completed by combining (5.8) with (5.9) and (5.10). \square

Remark 5.2. To extend the results in Lemma 5.1 to the 3D case we need to use a refinement procedure for tetrahedra that creates a new node on each marked face in $\hat{\mathcal{S}}_n$ and a node in the interior of each marked element. The proof in the 3D case is similar to the proof in the 2D case: for each couple of refined elements we define

$$\Phi_S := \alpha_S \varphi_S + \beta_1 \varphi_1 + \beta_2 \varphi_2,$$

where φ_S is the nodal basis function associated to the new node on the shared face and φ_i are the nodal basis functions associated to the new nodes in the interior of the elements. The coefficients $\alpha_S, \beta_1,$ and β_2 can be chosen in the same way as in Lemma 5.1, and the rest of the proof proceeds similarly.

In the next lemma, we bound the local error estimator above by the local difference of two discrete solutions coming from consecutive meshes, plus higher order terms. This kind of result is called “discrete local efficiency” by many authors.

Recall that \mathcal{T}_{n+1} is the refinement of \mathcal{T}_n obtained by applying Algorithm 1.

LEMMA 5.3. *For any $S \in \hat{\mathcal{S}}_n$, we have*

$$(5.11) \quad \eta_{S,n}^2 \lesssim \|u_{n+1} - u_n\|_{\Omega(S)}^2 + \|H_n(\lambda_{n+1}u_{n+1} - \lambda_n P_n u_n)\|_{0,\mathcal{B},\Omega(S)}^2 + \|H_n \lambda_n (u_n - P_n u_n)\|_{0,\mathcal{B},\Omega(S)}^2.$$

Proof. Since the function Φ_S defined in Lemma 5.1 is in V_{n+1} and $\text{supp}(\Phi_S) = \Omega(S)$, we have

$$(5.12) \quad a(u_{n+1} - u_n, \Phi_S) = a(u_{n+1}, \Phi_S) - a(u_n, \Phi_S) = \lambda_{n+1} \int_{\Omega(S)} \mathcal{B}u_{n+1} \Phi_S - a(u_n, \Phi_S).$$

Now applying integration by parts to the last term on the right-hand side of (5.12), we obtain

$$(5.13) \quad a(u_{n+1} - u_n, \Phi_S) = \lambda_{n+1} \int_{\Omega(S)} \mathcal{B}u_{n+1} \Phi_S - \int_S J_S(u_n) \Phi_S.$$

Rewriting (5.13) and combining with (5.1), we obtain

$$(5.14) \quad \begin{aligned} a(u_{n+1} - u_n, \Phi_S) - \int_{\Omega(S)} \mathcal{B}(\lambda_{n+1}u_{n+1} - \lambda_n P_n u_n) \Phi_S &= \lambda_n \int_{\Omega(S)} \mathcal{B}(P_n u_n) \Phi_S - \int_S J_S(u_n) \Phi_S \\ &= \|H_n \lambda_n P_n u_n\|_{0,\mathcal{B},\Omega(S)}^2 + \|H_S^{1/2} J_S(u_n)\|_{0,S}^2. \end{aligned}$$

Rearranging this, and then applying the triangle and Cauchy–Schwarz inequalities, we obtain

$$(5.15) \quad \begin{aligned} &\|H_n \lambda_n P_n u_n\|_{0,\mathcal{B},\Omega(S)}^2 + \|H_S^{1/2} J_S(u_n)\|_{0,S}^2 \\ &\leq |a(u_{n+1} - u_n, \Phi_S)| + \left| \int_{\Omega(S)} \mathcal{B}(\lambda_{n+1}u_{n+1} - \lambda_n P_n u_n) \Phi_S \right| \\ &\leq \|u_{n+1} - u_n\|_{\Omega(S)} \|\Phi_S\|_{\Omega(S)} + \|\lambda_{n+1}u_{n+1} - \lambda_n P_n u_n\|_{0,\mathcal{B},\Omega(S)} \|\Phi_S\|_{0,\mathcal{B},\Omega(S)} \\ &\lesssim \left(\|u_{n+1} - u_n\|_{\Omega(S)} + \|H_n(\lambda_{n+1}u_{n+1} - \lambda_n P_n u_n)\|_{0,\mathcal{B},\Omega(S)} \right) \|\Phi_S\|_{\Omega(S)}. \end{aligned}$$

In the final step of (5.15) we made use of the Poincaré inequality $\|\Phi_S\|_{0,\mathcal{B},\Omega(S)} \lesssim H_S \|\Phi_S\|_{\Omega(S)}$ and also the shape-regularity of the meshes. In view of (5.2), we have

$$\begin{aligned}
 (5.16) \quad & \|H_n \lambda_n P_n u_n\|_{0,\mathcal{B},\Omega(S)}^2 + \left\| H_S^{1/2} J_S(u_n) \right\|_{0,S}^2 \\
 & \lesssim \left(\|u_{n+1} - u_n\|_{\Omega(S)} + \|H_n(\lambda_{n+1} u_{n+1} - \lambda_n P_n u_n)\|_{0,\mathcal{B},\Omega(S)} \right)^2 \\
 & \lesssim \|u_{n+1} - u_n\|_{\Omega(S)}^2 + \|H_n(\lambda_{n+1} u_{n+1} - \lambda_n P_n u_n)\|_{0,\mathcal{B},\Omega(S)}^2.
 \end{aligned}$$

Now, from the definition of $\eta_{S,n}$ in (4.2), and the triangle inequality, we have

$$(5.17) \quad \eta_{S,n}^2 \lesssim \|H_n \lambda_n P_n u_n\|_{0,\mathcal{B},\Omega(S)}^2 + \left\| H_S^{1/2} J_S(u_n) \right\|_{0,S}^2 + \|H_n \lambda_n (u_n - P_n u_n)\|_{0,\mathcal{B},\Omega(S)}^2.$$

The required inequality (5.11) now follows from (5.16) and (5.17). \square

In the main result of this section, Theorem 5.6 below, we will be interested in achieving an error reduction result of the form $\|u - \alpha_{n+1} u_{n+1}\|_{\Omega} \leq \rho \|u - \alpha_n u_n\|_{\Omega}$ for some $\rho < 1$. Note that we need to introduce the scalar α_n here to ensure nearness of the approximate eigenfunction to the true one.

To prove error reduction we exploit the identity

$$\begin{aligned}
 (5.18) \quad & \|u - \alpha_n u_n\|_{\Omega}^2 = \|u - \alpha_{n+1} u_{n+1} + \alpha_{n+1} u_{n+1} - \alpha_n u_n\|_{\Omega}^2 \\
 & = \|u - \alpha_{n+1} u_{n+1}\|_{\Omega}^2 + \|\alpha_{n+1} u_{n+1} - \alpha_n u_n\|_{\Omega}^2 \\
 & \quad + 2a(u - \alpha_{n+1} u_{n+1}, \alpha_{n+1} u_{n+1} - \alpha_n u_n).
 \end{aligned}$$

In the case of source problems (e.g., [18, 19]), the α_n is not needed and the last term on the right-hand side vanishes due to Galerkin orthogonality. However, this approach is not available to us in the eigenvalue problem. Therefore, a more technical approach is needed to bound the last two terms on the right-hand side of (5.18) from below. The main technical result is in the following lemma. Recall the convention in Notation 4.1.

LEMMA 5.4. *With u, u_n, α_n as in Remark 3.3,*

$$(5.19) \quad \|\alpha_{n+1} u_{n+1} - \alpha_n u_n\|_{\Omega}^2 \gtrsim \theta^2 \|u - \alpha_n u_n\|_{\Omega}^2 - \text{osc}(\lambda_n u_n, \mathcal{T}_n)^2 - L_n^2,$$

where θ is defined in the marking strategy in Definition 4.5 and L_n satisfies the estimate:

$$(5.20) \quad L_n \leq \hat{C} (H_n^{\max})^s \|u - \alpha_n u_n\|_{\Omega},$$

where \hat{C} depends on $\theta, \lambda, C_1, C_2,$ and q .

Remark 5.5. Note that the oscillation term in (5.19) is unaffected if we replace $\alpha_n u_n$ by u_n .

Proof. By Definition 4.5 and Lemma 5.3, we have

$$\begin{aligned}
 \theta^2 \eta_n^2 & \leq \sum_{S \in \hat{\mathcal{S}}_n} \eta_{S,n}^2 \\
 & \lesssim \|\alpha_{n+1} u_{n+1} - \alpha_n u_n\|_{\Omega}^2 \\
 & \quad + \|H_n(\lambda_{n+1} \alpha_{n+1} u_{n+1} - \lambda_n P_n \alpha_n u_n)\|_{0,\mathcal{B},\Omega}^2 + \text{osc}(\lambda_n u_n, \mathcal{T}_n)^2.
 \end{aligned}$$

Hence, rearranging and making use of Lemma 4.2 and Remark 4.3, we have

$$\begin{aligned}
 (5.21) \quad & \| |\alpha_{n+1}u_{n+1} - \alpha_n u_n| \|_{\Omega}^2 \gtrsim \theta^2 \eta_n^2 - \| H_n(\lambda_{n+1}\alpha_{n+1}u_{n+1} - \lambda_n P_n \alpha_n u_n) \|_{0,\mathcal{B},\Omega}^2 \\
 & \quad - \text{osc}(\lambda_n u_n \mathcal{T}_n)^2 \\
 & \gtrsim \theta^2 \| |u - \alpha_n u_n| \|_{\Omega}^2 - \text{osc}(\lambda_n u_n \mathcal{T}_n)^2 \\
 & \quad - \theta^2 \tilde{G}_n^2 - \| H_n(\lambda_{n+1}\alpha_{n+1}u_{n+1} - \lambda_n P_n \alpha_n u_n) \|_{0,\mathcal{B},\Omega}^2,
 \end{aligned}$$

where \tilde{G}_n is the same as G_n in Lemma 4.2, but with u_n replaced by $\alpha_n u_n$.

Note that (5.21) is of the required form (5.19) with

$$L_n := \left(\theta^2 \tilde{G}_n^2 + \| H_n(\lambda_{n+1}\alpha_{n+1}u_{n+1} - \lambda_n P_n \alpha_n u_n) \|_{0,\mathcal{B},\Omega}^2 \right)^{1/2}.$$

We now estimate the last two terms in (5.21) to obtain (5.20). To estimate \tilde{G}_n , we use Theorem 3.1(ii) to obtain

$$\begin{aligned}
 (5.22) \quad & \tilde{G}_n \lesssim \frac{1}{2} (\lambda + \lambda_n) C_1^2 (H_n^{\max})^{2s} \frac{\| |u - Q_n u| \|_{\Omega}^2}{\| |u - \alpha_n u_n| \|_{\Omega}} \\
 & \leq \frac{1}{2} (\lambda + \lambda_n) C_1^2 (H_n^{\max})^{2s} \| |u - \alpha_n u_n| \|_{\Omega}.
 \end{aligned}$$

To estimate the last term in (5.21), we first use the triangle inequality to obtain

$$\begin{aligned}
 (5.23) \quad & \| H_n(\lambda_{n+1}\alpha_{n+1}u_{n+1} - \lambda_n P_n \alpha_n u_n) \|_{0,\mathcal{B},\Omega} \leq \\
 & \quad \| H_n(\lambda_{n+1}\alpha_{n+1}u_{n+1} - \lambda_n \alpha_n u_n) \|_{0,\mathcal{B},\Omega} + \text{osc}(\lambda_n u_n, \mathcal{T}_n).
 \end{aligned}$$

For the first term on the right-hand side of (5.23), we have

$$\begin{aligned}
 (5.24) \quad & \| H_n(\lambda_{n+1}\alpha_{n+1}u_{n+1} - \lambda_n \alpha_n u_n) \|_{0,\mathcal{B},\Omega} \leq \\
 & \quad H_n^{\max} (\| \lambda u - \lambda_{n+1}\alpha_{n+1}u_{n+1} \|_{0,\mathcal{B},\Omega} + \| \lambda u - \lambda_n \alpha_n u_n \|_{0,\mathcal{B},\Omega}).
 \end{aligned}$$

Then, recalling (2.6) and Theorem 3.1, we obtain

$$\begin{aligned}
 (5.25) \quad & \| \lambda u - \lambda_{n+1}\alpha_{n+1}u_{n+1} \|_{0,\mathcal{B},\Omega} \leq |\lambda - \lambda_{n+1}| \| u \|_{0,\mathcal{B},\Omega} \\
 & \quad + \lambda_{n+1} \| u - \alpha_{n+1}u_{n+1} \|_{0,\mathcal{B},\Omega} \\
 & \leq \| |u - \alpha_{n+1}u_{n+1}| \|_{\Omega}^2 \\
 & \quad + \lambda_{n+1} C_1 (H_n^{\max})^s \| |u - \alpha_{n+1}u_{n+1}| \|_{\Omega}.
 \end{aligned}$$

Using Theorem 3.1 (iii) and then Theorem 3.2, this implies

$$\begin{aligned}
 (5.26) \quad & \| \lambda u - \lambda_{n+1}\alpha_{n+1}u_{n+1} \|_{0,\mathcal{B},\Omega} \lesssim (C_2 + \lambda_{n+1} C_1) (H_n^{\max})^s \| |u - \alpha_{n+1}u_{n+1}| \|_{\Omega} \\
 & \leq q (C_2 + \lambda_{n+1} C_1) (H_n^{\max})^s \| |u - \alpha_n u_n| \|_{\Omega}.
 \end{aligned}$$

An identical argument shows

$$(5.27) \quad \| \lambda u - \lambda_n \alpha_n u_n \|_{0,\mathcal{B},\Omega} \lesssim (C_2 + \lambda_n C_1) (H_n^{\max})^s \| |u - \alpha_n u_n| \|_{\Omega}.$$

Combining (5.26) and (5.27) with (5.24), and using (2.9), we obtain

$$\begin{aligned}
 (5.28) \quad & \| H_n(\lambda_{n+1}\alpha_{n+1}u_{n+1} - \lambda_n \alpha_n u_n) \|_{0,\mathcal{B},\Omega} \lesssim (1+q) (C_2 + \lambda_n C_1) (H_n^{\max})^{s+1} \| |u - \alpha_n u_n| \|_{\Omega}.
 \end{aligned}$$

Now combining (5.28) with (5.21), (5.22), and (5.23) we obtain the result. \square

The next theorem contains the main result of this section. It shows that, provided we start with a “fine enough” mesh \mathcal{T}_n , the mesh adaptivity algorithm will reduce the error in the energy norm.

THEOREM 5.6 (error reduction). *For each $\theta \in (0, 1)$, there exists a sufficiently fine mesh threshold H_n^{\max} and constants $\mu > 0$ and $\rho \in (0, 1)$ (all of which may depend on θ and on the eigenvalue λ), with the following property. For any $\varepsilon > 0$ the inequality*

$$(5.29) \quad \text{osc}(\lambda_n u_n, \mathcal{T}_n) \leq \mu \varepsilon$$

implies either $\|u - \alpha_n u_n\|_{\Omega} \leq \varepsilon$ or

$$\|u - \alpha_{n+1} u_{n+1}\|_{\Omega} \leq \rho \|u - \alpha_n u_n\|_{\Omega}.$$

Proof. In view of (5.18) and remembering that $\alpha_{n+1} u_{n+1} - \alpha_n u_n \in V_{n+1}$ we have

$$(5.30) \quad \begin{aligned} & \|u - \alpha_n u_n\|_{\Omega}^2 - \|u - \alpha_{n+1} u_{n+1}\|_{\Omega}^2 \\ &= \|\alpha_{n+1} u_{n+1} - \alpha_n u_n\|_{\Omega}^2 + 2a(u - \alpha_{n+1} u_{n+1}, \alpha_{n+1} u_{n+1} - \alpha_n u_n) \\ &= \|\alpha_{n+1} u_{n+1} - \alpha_n u_n\|_{\Omega}^2 + 2b(\lambda u - \lambda_{n+1} \alpha_{n+1} u_{n+1}, \alpha_{n+1} u_{n+1} - \alpha_n u_n). \end{aligned}$$

Before proceeding further, recall that by the assumptions (2.3) and (2.4), and the Poincaré inequality, there exists a constant C_P (depending on \mathcal{A} , \mathcal{B} and Ω) such that

$$\|v\|_{0,\mathcal{B},\Omega} \leq C_P \|v\|_{\Omega}, \quad \text{for all } v \in H_0^1(\Omega).$$

Now using Cauchy–Schwarz and then the Young inequality $2ab \leq \frac{1}{4C_P^2} a^2 + 4C_P^2 b^2$ on the second term on the right-hand side of (5.30), we get

$$(5.31) \quad \begin{aligned} & \|u - \alpha_n u_n\|_{\Omega}^2 - \|u - \alpha_{n+1} u_{n+1}\|_{\Omega}^2 \\ & \geq \|\alpha_{n+1} u_{n+1} - \alpha_n u_n\|_{\Omega}^2 - 2\|\lambda u - \lambda_{n+1} \alpha_{n+1} u_{n+1}\|_{0,\mathcal{B},\Omega} \|\alpha_{n+1} u_{n+1} - \alpha_n u_n\|_{0,\mathcal{B},\Omega} \\ & \geq \|\alpha_{n+1} u_{n+1} - \alpha_n u_n\|_{\Omega}^2 - \frac{1}{4C_P^2} \|\alpha_{n+1} u_{n+1} - \alpha_n u_n\|_{0,\mathcal{B},\Omega}^2 \\ & \quad - 4C_P^2 \|\lambda u - \lambda_{n+1} \alpha_{n+1} u_{n+1}\|_{0,\mathcal{B},\Omega}^2 \\ & \geq \frac{3}{4} \|\alpha_{n+1} u_{n+1} - \alpha_n u_n\|_{\Omega}^2 - 4C_P^2 \|\lambda u - \lambda_{n+1} \alpha_{n+1} u_{n+1}\|_{0,\mathcal{B},\Omega}^2. \end{aligned}$$

Hence

$$\begin{aligned} \|u - \alpha_{n+1} u_{n+1}\|_{\Omega}^2 & \leq \|u - \alpha_n u_n\|_{\Omega}^2 - \frac{3}{4} \|\alpha_{n+1} u_{n+1} - \alpha_n u_n\|_{\Omega}^2 \\ & \quad + 4C_P^2 \|\lambda u - \lambda_{n+1} \alpha_{n+1} u_{n+1}\|_{0,\mathcal{B},\Omega}^2. \end{aligned}$$

Applying Lemma 5.4, we see that there exist constants C, \hat{C} such that

$$\begin{aligned} \|u - \alpha_{n+1} u_{n+1}\|_{\Omega}^2 & \leq \left(1 - \frac{3}{4} C \theta^2 + \frac{3}{4} C \hat{C}^2 (H_n^{\max})^{2s}\right) \|u - \alpha_n u_n\|_{\Omega}^2 \\ & \quad + 4 C_P^2 \|\lambda u - \lambda_{n+1} \alpha_{n+1} u_{n+1}\|_{0,\mathcal{B},\Omega}^2 \\ & \quad + \frac{3}{4} C \text{osc}(\lambda_n u_n, \mathcal{T}_n)^2. \end{aligned}$$

Then, making use of (5.26) we have

$$(5.32) \quad |||u - \alpha_{n+1}u_{n+1}|||_{\Omega}^2 \leq \gamma_n |||u - \alpha_n u_n|||_{\Omega}^2 + \frac{3}{4} C \operatorname{osc}(\lambda_n u_n, \mathcal{T}_n)^2$$

with

$$(5.33) \quad \gamma_n := \left[1 - \frac{3}{4} C \theta^2 + C'(H_n^{\max})^{2s} \right],$$

where C' is another constant independent of n . Note that H_n^{\max} can be chosen sufficiently small so that $\gamma_m \leq \gamma$ for some $\gamma \in (0, 1)$ and all $m \geq n$.

Consider now the consequences of the inequality (5.29). If $|||u - \alpha_n u_n|||_{\Omega} > \varepsilon$, then (5.32) implies

$$|||u - \alpha_{n+1}u_{n+1}|||_{\Omega}^2 \leq \left[\gamma + \frac{3}{4} C \mu^2 \right] |||u - \alpha_n u_n|||_{\Omega}^2.$$

Now choose μ small enough so that

$$(5.34) \quad \rho := \left(\gamma + \frac{3}{4} C \mu^2 \right)^{1/2} < 1$$

to complete the proof. \square

6. Proof of convergence. The main result of this paper is Theorem 6.2 below, which proves convergence of the adaptive method and also demonstrates the decay of oscillations of the sequence of approximate eigenfunctions. Before proving this result we need a final lemma.

LEMMA 6.1. *There exists a constant $\tilde{\rho} \in (0, 1)$ such that*

$$(6.1) \quad \operatorname{osc}(u_{n+1}, \mathcal{T}_{n+1}) \leq \tilde{\rho} \operatorname{osc}(u_n, \mathcal{T}_n) + (1 + q)(H_n^{\max})^2 |||u - \alpha_n u_n|||_{\Omega}.$$

Proof. First, recall that one of the key results in [18], namely, [18, Lemma 3.8], is the proof that the oscillations of any fixed function $v \in H_0^1(\Omega)$ are reduced by applying one refinement based on Marking Strategy 2 (Definition 4.7). Thus, we have (in view of Algorithm 1):

$$(6.2) \quad \operatorname{osc}(u_n, \mathcal{T}_{n+1}) \leq \tilde{\rho} \operatorname{osc}(u_n, \mathcal{T}_n),$$

where $0 < \tilde{\rho} < 1$ is independent of u_n . Thus, a simple application of the triangle inequality combined with (6.2) yields

$$(6.3) \quad \begin{aligned} \operatorname{osc}(u_{n+1}, \mathcal{T}_{n+1}) &\leq \operatorname{osc}(u_n, \mathcal{T}_{n+1}) + \operatorname{osc}(\alpha_{n+1}u_{n+1} - \alpha_n u_n, \mathcal{T}_{n+1}) \\ &\leq \tilde{\rho} \operatorname{osc}(u_n, \mathcal{T}_n) + \operatorname{osc}(\alpha_{n+1}u_{n+1} - \alpha_n u_n, \mathcal{T}_{n+1}). \end{aligned}$$

(Recall, again, that $\operatorname{osc}(u_n, \mathcal{T}_n) = \operatorname{osc}(\alpha_n u_n, \mathcal{T}_n)$.) A further application of the triangle inequality and then (4.14) yields

$$(6.4) \quad \begin{aligned} \operatorname{osc}(\alpha_{n+1}u_{n+1} - \alpha_n u_n, \mathcal{T}_{n+1}) &\leq \operatorname{osc}(u - \alpha_{n+1}u_{n+1}, \mathcal{T}_{n+1}) + \operatorname{osc}(u - \alpha_n u_n, \mathcal{T}_{n+1}) \\ &\lesssim (H_n^{\max})^2 (|||u - \alpha_{n+1}u_{n+1}|||_{\Omega} + |||u - \alpha_n u_n|||_{\Omega}), \end{aligned}$$

and then combining (6.3) and (6.4) and applying Theorem 3.2 completes the proof. \square

THEOREM 6.2. *Provided the initial mesh \mathcal{T}_0 is chosen so that H_0^{\max} is small enough, there exists a constant $p \in (0, 1)$, such that the recursive application of Algorithm 1 yields a convergent sequence of approximate eigenvalues and eigenvectors, with the property:*

$$(6.5) \quad \| \|u - \alpha_n u_n \| \|_{\Omega} \leq B_0 q p^n,$$

and

$$(6.6) \quad \lambda_n \operatorname{osc}(u_n, \mathcal{T}_n) \leq B_1 p^n,$$

where B_0 and B_1 are constants and q is the constant defined in Theorem 3.2.

Remark 6.3. The initial mesh convergence threshold and the constants B_0 and B_1 may depend on θ , $\tilde{\theta}$, and λ .

Proof. The proof of this theorem is by induction and the induction step contains an application of Theorem 5.6. In order to ensure the reduction of the error, we have to assume that the starting mesh \mathcal{T}_0 is fine enough and μ in Theorem 5.6 is small enough such that, for the chosen value of θ , the quantity ρ in (5.34) satisfies $\rho < 1$.

Then with $\tilde{\rho}$ as in Lemma 6.1, choose p in the range

$$\max\{\rho, \tilde{\rho}\} < p < 1.$$

We also set

$$B_1 = \operatorname{osc}(\lambda_0 u_0, \mathcal{T}_0) \quad \text{and} \quad B_0 = \max\{\mu^{-1} p^{-1} B_1, \| \|u - \alpha_0 u_0 \| \|_{\Omega}\}.$$

To perform the inductive proof, first note that by the definition of B_0 and Theorem 3.2,

$$\| \|u - \alpha_0 u_0 \| \|_{\Omega} \leq B_0 \leq B_0 q,$$

since $q > 1$. Combined with the definition of B_1 we have shown the result for $n = 0$.

Now, suppose that, for some $n > 0$, the inequalities (6.5) and (6.6) hold.

Now let us consider the outcomes, depending on whether the inequality

$$(6.7) \quad \| \|u - \alpha_n u_n \| \|_{\Omega} \leq B_0 p^{n+1}$$

holds or not. If (6.7) holds, then we can apply Theorem 3.2 to conclude that

$$\| \|u - \alpha_{n+1} u_{n+1} \| \|_{\Omega} \leq q \| \|u - \alpha_n u_n \| \|_{\Omega} \leq q B_0 p^{n+1},$$

which proves (6.5) for $n + 1$.

On the other hand, if (6.7) does not hold, then, by definition of B_0 ,

$$(6.8) \quad \| \|u - \alpha_n u_n \| \|_{\Omega} > B_0 p^{n+1} \geq \mu^{-1} B_1 p^n.$$

Also, since we have assumed (6.6) for n , we have

$$(6.9) \quad \lambda_n \operatorname{osc}(u_n, \mathcal{T}_n) \leq \mu\varepsilon \quad \text{with} \quad \varepsilon := \mu^{-1}B_1p^n.$$

Then (6.8) and (6.9) combined with Theorem 5.6 yields

$$\| \|u - \alpha_{n+1}u_{n+1}\| \|_{\Omega} \leq \rho \| \|u - \alpha_n u_n\| \|_{\Omega},$$

and so, using the inductive hypothesis (6.5) combined with the definition of p , we have

$$\| \|u - \alpha_{n+1}u_{n+1}\| \|_{\Omega} \leq \rho B_0 q p^n \leq q B_0 p^{n+1},$$

which, again, proves (6.5) for $n + 1$.

To conclude the proof, we have to show that also (6.6) holds for $n + 1$. Using Lemma 6.1, (2.9), and the inductive hypothesis, we have

$$(6.10) \quad \begin{aligned} \lambda_{n+1} \operatorname{osc}(u_{n+1}, \mathcal{T}_{n+1}) &\leq \tilde{\rho} B_1 p^n + (1 + q)(H_n^{\max})^2 \lambda_n B_0 q p^n \\ &\leq (\tilde{\rho} B_1 + (1 + q)(H_0^{\max})^2 \lambda_0 B_0 q) p^n. \end{aligned}$$

Now, (recalling that $\tilde{\rho} < p$), in addition to the condition already imposed on H_0^{\max} , we can further require that

$$\tilde{\rho} B_1 + (1 + q)(H_0^{\max})^2 \lambda_0 B_0 q \leq p B_1.$$

This ensures that

$$\lambda_{n+1} \operatorname{osc}(u_{n+1}, \mathcal{T}_{n+1}) \leq B_1 p^{n+1},$$

thus concluding the proof. \square

7. Numerical experiments. We present numerical experiments to illustrate the convergence theory. Algorithm 1 has been implemented in FORTRAN95. The mesh refinement has been done using the toolbox ALBERTA [20]. We used the package ARPACK [15] to compute eigenpairs and the sparse direct linear solver ME27 from the HSL [21, 13] to carry out the shift-invert solves required by ARPACK. Additional numerical experiments on photonic crystal problems and on 3D problems are given in [10] and [11].

7.1. Example: Laplace operator. In the first set of simulations, we have solved the Laplace eigenvalue problem (i.e., $\mathcal{A} = I$ and $\mathcal{B} = 1$ in (2.2)) on a unit square with Dirichlet boundary conditions. The exact eigenvalues are known explicitly.

We compare different runs of Algorithm 1 using different values for θ and $\tilde{\theta}$ in Table 7.1. Since the problem is smooth, from Theorem 3.1 it follows that using uniform refinement the rate of convergence for eigenvalues should be $\mathcal{O}(H_n^{\max})^2$, or, equivalently, the rate of convergence in the number of degrees of freedom (DOFs) N should be $\mathcal{O}(N^{-1})$. We measure the rate of convergence by conjecturing that $|\lambda - \lambda_n| = CN^{-\beta}$ and estimating β for each pair of computations from the formula $\beta = -\log(|\lambda - \lambda_n|/|\lambda - \lambda_{n-1}|)/\log(\text{DOFs}_n/\text{DOFs}_{n-1})$. Similarly, Table 7.2 contains the same kind of information relative to the fourth smallest eigenvalue of the problem. Our results show a convergence rate close to $\mathcal{O}(N^{-1})$ for $\theta, \tilde{\theta}$ sufficiently large. However, the rate of convergence is sensitive to the values of θ and $\tilde{\theta}$.

TABLE 7.1

Comparison of the reduction of the error and DOFs of the adaptive method for the smallest eigenvalue for the Laplace problem on the unit square.

Iteration	$\theta = \tilde{\theta} = 0.2$			$\theta = \tilde{\theta} = 0.5$			$\theta = \tilde{\theta} = 0.8$		
	$ \lambda - \lambda_n $	DOFs	β	$ \lambda - \lambda_n $	DOFs	β	$ \lambda - \lambda_n $	DOFs	β
1	0.1350	400	-	0.1350	400	-	0.1350	400	-
2	0.1327	498	0.0802	0.1177	954	0.1581	0.0529	1989	0.5839
3	0.1293	613	0.1228	0.0779	1564	0.8349	0.0176	5205	1.1407
4	0.1256	731	0.1645	0.0501	1977	1.8788	0.0073	15980	0.7877
5	0.1215	854	0.2138	0.0351	2634	1.2383	0.0024	48434	0.9836
6	0.1165	970	0.3340	0.0176	4004	0.7885	0.0009	122699	1.0673
7	0.1069	1097	0.6962	0.0121	6588	0.7217	0.0003	312591	1.0083

TABLE 7.2

Comparison of the reduction of the error and DOFs of the adaptive method for the fourth smallest eigenvalue for the Laplace problem on the unit square.

Iteration	$\theta = \tilde{\theta} = 0.2$			$\theta = \tilde{\theta} = 0.5$			$\theta = \tilde{\theta} = 0.8$		
	$ \lambda - \lambda_n $	DOFs	β	$ \lambda - \lambda_n $	DOFs	β	$ \lambda - \lambda_n $	DOFs	β
1	2.1439	400	-	2.1439	400	-	2.1439	400	-
2	2.0997	505	0.0895	1.8280	1016	0.1658	0.7603	2039	0.6365
3	2.0549	626	0.1004	1.0850	1636	1.1662	0.2439	6793	0.9447
4	1.9945	759	0.1548	0.7792	2254	1.0331	0.0917	18717	0.9652
5	1.9164	883	0.2638	0.4936	3067	1.4826	0.0331	54113	0.9583
6	1.7717	1017	0.5557	0.3484	4681	0.8240	0.0120	146056	1.0181
7	1.6463	1131	0.6911	0.2578	7321	0.6730	0.0046	382024	0.9970

In the theory presented in [24], it is shown that the error for eigenvalues for smooth problems is bounded in terms of the square of the considered eigenvalue, i.e.,

$$(7.1) \quad |\lambda - \lambda_n| \leq C \lambda^2 (H_n^{\max})^2.$$

Also, we know that the first and the fourth eigenvalues are 19.7392089 and 78.9568352, and so, $\lambda_4 = 4\lambda_1$. Comparing errors in Table 7.2 with those in Table 7.1, we see that the errors are roughly multiplied by a factor of 16, as predicted by (7.1).

Often h-adaptivity uses only a marking strategy based on an estimation of the error, as in Marking Strategy 1 and avoids refining based on oscillations as in Marking Strategy 2. (Convergence of an adaptive scheme for eigenvalue problems which does not use marking strategy 2 is recently proved in [5].) To investigate the effects of refinement based on oscillations, in Table 7.3 we have computed the smallest eigenvalue for the Laplace problem keeping θ fixed and varying $\tilde{\theta}$ only. Reducing $\tilde{\theta}$ towards 0 has the effect of turning off the refinement arising from Marking Strategy 2. The results in Table 7.3 seem to suggest that the rate of convergence slightly increases as $\tilde{\theta}$ increases.

We investigate this further in Table 7.4, where we take iterations 5, 6, and 7 from Table 7.3, and we present the quantity $C^* := N \times |\lambda - \lambda_n|$, where N denotes the number of DOFs. Then C^* gives an indication of the size of the unknown constant in the optimal error estimate $|\lambda - \lambda_n| = \mathcal{O}(N^{-1})$. The results suggest that C^* stays fairly constant independent of $\tilde{\theta}$.

In Table 7.5, we have set $\tilde{\theta} = 0$. Although the convergence result given in this paper does not hold any more, the method is still clearly convergent. Comparing Table 7.1, Table 7.3, and Table 7.5, we see that with the second marking strategy the

TABLE 7.3

Comparison of the reduction of the error and DOFs of the adaptive method for the smallest eigenvalue for the Laplace problem on the unit square for a fixed value of θ and varying $\tilde{\theta}$.

Iteration	$\theta = 0.8, \tilde{\theta} = 0.1$			$\theta = 0.8, \tilde{\theta} = 0.3$			$\theta = 0.8, \tilde{\theta} = 0.5$		
	$ \lambda - \lambda_n $	DOFs	β	$ \lambda - \lambda_n $	DOFs	β	$ \lambda - \lambda_n $	DOFs	β
1	0.1350	400	-	0.1350	400	-	0.1350	400	-
2	0.0704	1269	0.5646	0.0698	1372	0.5353	0.0673	1555	0.5131
3	0.0307	2660	1.1215	0.0300	2821	1.1700	0.0285	3229	1.1757
4	0.0137	7492	0.7770	0.0133	7846	0.7980	0.0115	9140	0.8731
5	0.0056	18853	0.9699	0.0052	20189	0.9918	0.0046	22793	0.9913
6	0.0021	52247	0.9587	0.0020	55640	0.9382	0.0018	61582	0.9310
7	0.0008	140049	0.9834	0.0008	145773	1.0011	0.0007	161928	1.0238

TABLE 7.4

Values of C^* computed from Table 7.3.

Iteration	$\theta = 0.8, \tilde{\theta} = 0.1$	$\theta = 0.8, \tilde{\theta} = 0.3$	$\theta = 0.8, \tilde{\theta} = 0.5$
5	1.06×10^2	1.05×10^2	1.05×10^2
6	1.10×10^2	1.11×10^2	1.11×10^2
7	1.12×10^2	1.12×10^2	1.13×10^2

TABLE 7.5

Comparison of the reduction of the error and DOFs of the adaptive method for the smallest eigenvalue for the Laplace problem on the unit square using marking strategy 1 only.

Iteration	$\theta = 0.2$			$\theta = 0.5$			$\theta = 0.8$		
	$ \lambda - \lambda_n $	DOFs	β	$ \lambda - \lambda_n $	DOFs	β	$ \lambda - \lambda_n $	DOFs	β
1	0.1350	400	-	0.1350	400	-	0.1350	400	-
2	0.1328	447	0.1525	0.1209	648	0.2289	0.0704	1253	0.5704
3	0.1299	503	0.1824	0.0859	1036	0.7283	0.0307	2646	1.1125
4	0.1271	565	0.1958	0.0627	1455	0.9301	0.0138	7490	0.7697
5	0.1238	637	0.2157	0.0458	1965	1.0429	0.0056	18847	0.9734
6	0.1189	712	0.3650	0.0323	3031	0.8066	0.0021	52239	0.9585
7	0.1113	795	0.6014	0.0228	4372	0.9531	0.0008	140194	0.9828

TABLE 7.6

Comparison between the number of marked elements by strategy 1 (i.e., $\#\hat{\mathcal{T}}_n$) and the number of marked elements by strategy 2 only (i.e., $\#(\tilde{\mathcal{T}}_n \setminus \hat{\mathcal{T}}_n)$) for different values of θ and $\tilde{\theta}$ for the smallest eigenvalue of the Laplace problem on the unit square.

Iteration	$\theta = \tilde{\theta} = 0.2$		$\theta = \tilde{\theta} = 0.5$		$\theta = \tilde{\theta} = 0.8$	
	$\#\hat{\mathcal{T}}_n$	$\#(\tilde{\mathcal{T}}_n \setminus \hat{\mathcal{T}}_n)$	$\#\hat{\mathcal{T}}_n$	$\#(\tilde{\mathcal{T}}_n \setminus \hat{\mathcal{T}}_n)$	$\#\hat{\mathcal{T}}_n$	$\#(\tilde{\mathcal{T}}_n \setminus \hat{\mathcal{T}}_n)$
1	12	15	85	99	299	285
2	13	15	102	85	953	19
3	14	15	100	25	3069	198
4	14	14	173	7	7965	2053
5	15	13	310	48	22426	1486
6	15	12	552	184	58075	3005

number of degrees of freedom grows faster than without it. To illustrate this effect better, Table 7.6 tabulates the number of elements $\#\hat{\mathcal{T}}_n$ (marked by Marking Strategy 1) with the extra number of elements $\#(\tilde{\mathcal{T}}_n \setminus \hat{\mathcal{T}}_n)$ (marked by Marking Strategy 2 alone). Note that the new DOFs created by mesh refinement come only from the refinement of

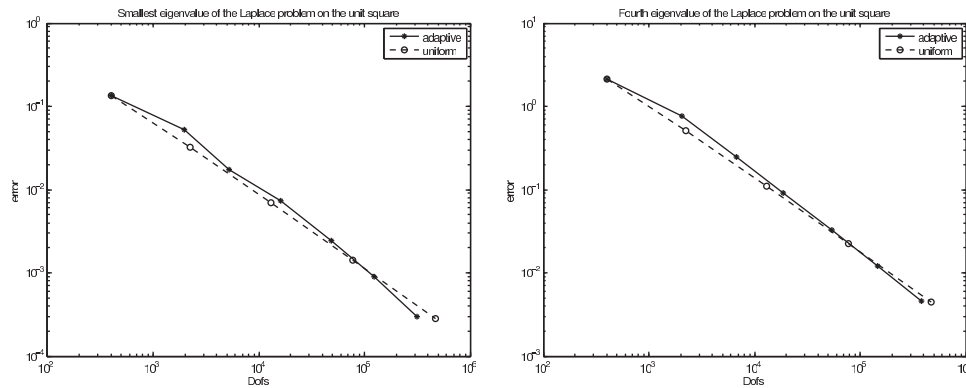


FIG. 7.1. Loglog plots of convergence of adaptive and uniform refinement for first eigenvalue of the Laplacian (left) and fourth eigenvalue of the Laplacian (right).

TABLE 7.7

Comparison of the reduction of the error and DOFs of the adaptive method for the second smallest eigenvalue for the Laplace problem on the unit square.

n	$\theta = \bar{\theta} = 0.2$			$\theta = \bar{\theta} = 0.5$			$\theta = \bar{\theta} = 0.8$		
	$ \lambda - \lambda_n $	N	β	$ \lambda - \lambda_n $	N	β	$ \lambda - \lambda_n $	N	β
1	0.5802	400	-	0.5802	400	-	0.5802	400	-
2	0.5678	478	0.1212	0.4935	811	0.2291	0.2447	1533	0.6427
3	0.5514	562	0.1816	0.3201	1275	0.9564	0.0959	3640	1.0826
4	0.5329	646	0.2449	0.2295	1728	1.0953	0.0368	11747	0.8169
5	0.5111	735	0.3237	0.1521	2374	1.2950	0.0136	32881	0.9651
6	0.4758	829	0.5942	0.1078	3498	0.8875	0.0050	82968	1.0778
7	0.4392	918	0.7856	0.0782	5555	0.6938	0.0020	221521	0.9574

the marked elements, but also from the closures used to keep the meshes conforming. It is clear that the number of elements marked as a result of the oscillations continues to rise as refinement proceeds, although much more slowly than the number marked by the residual-based criterion (Marking Strategy 1).

In Figure 7.1 we compare the performance of the adaptive algorithm with uniform bisection5 refinement (see Figure 4.1) for the first and fourth eigenvalues of the Laplace operator. We note that in this case both methods converge with a similar rate, as is expected since in this case the regularity of eigenfunctions is H^2 . To complete this section, we give in Table 7.7 an example of the performance of the adaptive method for computing nonsimple eigenvalues. In this case, we considered the second smallest eigenvalue of the Laplace operator on the unit square which has multiplicity 2. We see that, although the theory given above does not strictly hold, the method performs similarly to the case of simple eigenvalues.

7.2. Example: Elliptic operator with discontinuous coefficients. In this example, we investigate how our method copes with discontinuous coefficients. In order to do that, we modified the smooth problem from Example 7.1. We inserted a square subdomain of side 0.5 in the center of the unit square domain. In the bilinear form (2.2), we also chose the function \mathcal{A} to be the scalar piecewise constant function, which assumes the value 100 inside the inner subdomain and the value 1 outside it. As before, \mathcal{B} in (2.2) is chosen as $\mathcal{B} = 1$. The jump in the value of \mathcal{A} generally

TABLE 7.8

Comparison of the reduction of the error and DOFs of the adaptive method for the smallest eigenvalue for the 2D problem with discontinuous coefficient.

Iteration	$\theta = \tilde{\theta} = 0.2$			$\theta = \tilde{\theta} = 0.5$			$\theta = \tilde{\theta} = 0.8$		
	$ \lambda - \lambda_n $	DOFs	β	$ \lambda - \lambda_n $	DOFs	β	$ \lambda - \lambda_n $	DOFs	β
1	1.1071	81	-	1.1071	81	-	1.1071	81	-
2	1.0200	103	0.3410	0.8738	199	0.2632	0.4834	356	0.5597
3	1.0105	129	0.0416	0.5848	314	0.8805	0.2244	799	0.9494
4	1.0039	147	0.0498	0.3983	491	0.8591	0.0990	2235	0.7957
5	0.8968	167	0.8843	0.2766	673	1.1564	0.0401	4764	1.1932
6	0.8076	194	0.6996	0.1933	975	0.9665	0.0180	12375	0.8372
7	0.8008	217	0.0747	0.1346	1476	0.8722	0.0065	29148	1.1888
8	0.7502	237	0.7401	0.0948	2080	1.0237	0.0020	65387	1.4482

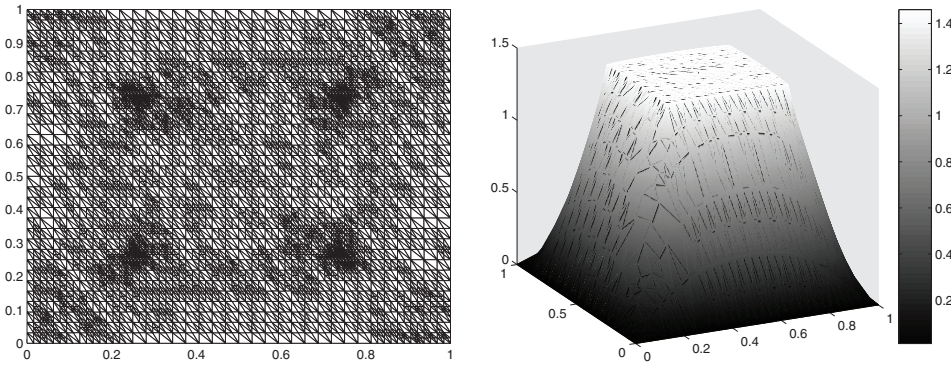


FIG. 7.2. A refined mesh from the adaptive method corresponding to the first eigenvalue of the 2D problem with discontinuous coefficient, and the corresponding eigenfunction.

produces a jump in the gradient of the eigenfunctions all along the boundary of the subdomain, and at the corners of the subdomain (from both inside and outside) the eigenfunction has infinite gradient, arising from the usual corner singularities. We choose our initial mesh to be aligned with the discontinuity in \mathcal{A} and so only the corner singularities are active here. We still have Assumption 2.1, but now $s < 1$ and, from Theorem 3.1, using uniform refinement, the rate of convergence for eigenvalues should be $\mathcal{O}(H_n^{\max})^{2s}$ or, equivalently, $\mathcal{O}(N^{-s})$, where N is the number of DOFs. The adaptive method yields the optimal order of $\mathcal{O}(N^{-1})$ (which holds for uniform meshes and smooth problems) for large enough θ and $\tilde{\theta}$. (See Table 7.8.) Here we compute the “exact” λ using a mesh with about half a million of DOFs.

In Figure 7.2, we depict the mesh coming from the fourth iteration of Algorithm 1 with $\theta = \tilde{\theta} = 0.8$ for the smallest eigenvalue of this problem. This mesh is the result of multiple refinements using both marking strategies 1 and 2 each time. As can be seen, the corners of the subdomain are much more refined than the rest of the mesh. This is clearly the effect of the first marking strategy, since the edge residuals have detected the singularity in the gradient of the eigenfunction at these points. In Figure 7.2, we also depict the corresponding eigenfunction.

In Figure 7.3, analogously to Figure 7.1, we compare the convergence of the adaptive method with uniform refinement for this example. Now, because of the lack of regularity, the superiority of the adaptive method is clearly visible.

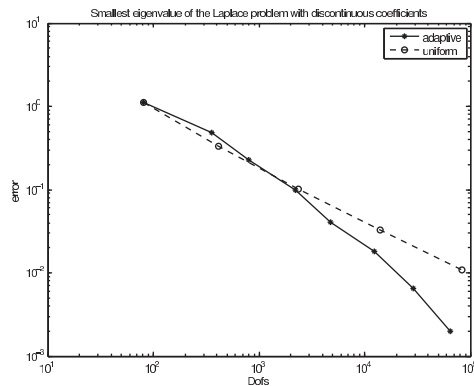


FIG. 7.3. Loglog plot of convergence of adaptive and uniform refinement for first eigenvalue of the problem with discontinuous coefficient.

Acknowledgment. We would like to thank Carsten Carstensen for his kind support and very useful discussions.

REFERENCES

- [1] M. AINSWORTH AND J.T. ODEN, *A Posteriori Error Estimation in Finite Element Analysis*, Wiley, New York, 2000.
- [2] I. BABUŠKA, *The finite element method for elliptic equations with discontinuous coefficients*, Computing, 5 (1970), pp. 207–213.
- [3] I. BABUŠKA AND J. OSBORN, *Eigenvalue problems*, in Handbook of Numerical Analysis Vol. II, P.G. Cairlet and J.L. Lions, eds., North Holland, 1991, pp. 641–787.
- [4] M. BOURLAND, M. DAUGE, M.-S. LUBUMA, AND S. NIÇAISE, *Coefficients of the singularities for elliptic boundary value problems on domains with conical points. III: Finite element methods on polygonal domains*, SIAM J. Numer. Anal., 29 (1992), pp. 136–155.
- [5] C. CARSTENSEN AND J. GEDICKE, *An oscillation-free adaptive FEM for symmetric eigenvalue problems*, preprint, 2008.
- [6] C. CARSTENSEN AND R.H.W. HOPPE, *Convergence analysis of an adaptive nonconforming finite element method*, Numer. Math., 103 (2006), pp. 251–266.
- [7] C. CARSTENSEN AND R.H.W. HOPPE, *Error reduction and convergence for an adaptive mixed finite element method*, Math. Comput., 75 (2006) pp. 1033–1042.
- [8] W. DÖRFLER, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.
- [9] R.G. DURÁN, C. PADRA, AND R. RODRÍGUEZ, *A posteriori estimates for the finite element approximation of eigenvalue problems*, Math. Models Methods Appl. Sci., 13 (2003), pp. 1219–1229.
- [10] S. GIANI, *Convergence of adaptive finite element methods for elliptic eigenvalue problems with application to photonic crystals*, Ph.D. Thesis, University of Bath, Bath, UK, 2008.
- [11] S. GIANI AND I.G. GRAHAM, *A convergent adaptive method for elliptic eigenvalue problems and numerical experiments*, Research Report 14/08, Bath Institute for Complex Systems, 2008. <http://www.bath.ac.uk/math-sci/BICS/>
- [12] W. HACKBUSCH, *Elliptic Differential Equations*, Springer, New York, 1992.
- [13] HSL archive, <http://hsl.rl.ac.uk/archive/hslarchive.html>
- [14] M.G. LARSON, *A posteriori and a priori analysis for finite element approximations of self-adjoint elliptic eigenvalue problems*, SIAM J. Numer. Anal., 38 (2000), pp. 608–625.
- [15] R.B. LEHOUCQ, D.C. SORENSEN, AND C. YANG, *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, SIAM, 1998
- [16] K. MEKCHAY AND R. H. NOCHETTO, *Convergence of adaptive finite element methods for general second order linear elliptic pdes*, SIAM J. Numer. Anal., 43 (2005), pp. 1803–1827.
- [17] W. MITCHELL, *Optimal multilevel iterative methods for adaptive grids*, SIAM J. Sci. Stat. Comput., 13 (1992), pp. 146–167.

- [18] P. MORIN, R.H. NOCHETTO, AND K.G. SIEBERT, *Data oscillation and convergence of adaptive fem*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.
- [19] P. MORIN, R.H. NOCHETTO, AND K.G. SIEBERT, *Convergence of adaptive finite element methods*, SIAM Rev., 44 (2002), pp. 631–658.
- [20] A. SCHMIDT AND K.G. SIEBERT, *ALBERT: An adaptive hierarchical finite element toolbox*, Manual, p. 244, preprint 06/2000 Freiburg.
- [21] J.A. SCOTT, *Sparse direct methods: An introduction*, Lecture Notes in Physics, 535, 401, 2000.
- [22] R.L. SCOTT AND S. ZHANG, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comput., 54 (1990), pp. 483–493.
- [23] R. STEVENSON, *Optimality of a standard adaptive finite element method*, Found. Comput. Math., 7 (2007), pp. 245–269.
- [24] G. STRANG AND G.J. FIX, *An Analysis of the Finite Element Method*, Prentice-Hall, 1973.
- [25] T.F. WALSH, G.M. REESE, AND U.L. HETMANIUK, *Explicit a posteriori error estimates for eigenvalue analysis of heterogeneous elastic structures*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 3614–3623.

THE DERIVATION OF HYBRIDIZABLE DISCONTINUOUS GALERKIN METHODS FOR STOKES FLOW*

BERNARDO COCKBURN[†] AND JAYADEEP GOPALAKRISHNAN[‡]

Abstract. In this paper, we introduce a new class of discontinuous Galerkin methods for the Stokes equations. The main feature of these methods is that they can be implemented in an efficient way through a hybridization procedure which reduces the globally coupled unknowns to certain approximations on the element boundaries. We present four ways of hybridizing the methods, which differ by the choice of the globally coupled unknowns. Classical methods for the Stokes equations can be thought of as limiting cases of these new methods.

Key words. Stokes equations, mixed methods, discontinuous Galerkin methods, hybridized methods, Lagrange multipliers

AMS subject classifications. 65N30, 65M60, 35L65

DOI. 10.1137/080726653

1. Introduction. This paper is devoted to the derivation of a new class of discontinuous Galerkin (DG) methods for the three-dimensional Stokes problem

$$\begin{aligned} -\Delta \mathbf{u} + \operatorname{grad} p &= \mathbf{f} && \text{in } \Omega, \\ \operatorname{div} \mathbf{u} &= 0 && \text{in } \Omega, \\ \mathbf{u} &= \mathbf{g} && \text{on } \partial\Omega. \end{aligned}$$

As usual, we assume that \mathbf{f} is in $L^2(\Omega)^3$, that $\mathbf{g} \in H^{1/2}(\partial\Omega)^3$, and that \mathbf{g} satisfies the compatibility condition

$$(1.1) \quad \int_{\partial\Omega} \mathbf{g} \cdot \mathbf{n} = 0,$$

where \mathbf{n} is the outward unit normal on $\partial\Omega$. We assume that Ω is a bounded simply connected domain with connected Lipschitz polyhedral boundary $\partial\Omega$.

The novelty in the class of DG methods derived here lies in the fact that they can be hybridized. Hybridized methods are primarily attractive due to the reduction in the number of globally coupled unknowns, especially in the high order case. Hybridization for conforming methods was traditionally thought of as a reformulation that moves the interelement continuity constraints of approximations from the finite element spaces to the system of equations. Such reformulations are now well known to possess various advantages [9] (in addition to the reduction in the number of unknowns). In adapting the hybridization idea to DG methods, we face the difficulty that DG methods have no interelement continuity constraints to begin with. Nonetheless, some DG methods realize interelement coupling through constraints on

*Received by the editors June 10, 2008; accepted for publication (in revised form) October 13, 2008; published electronically February 19, 2009.

<http://www.siam.org/journals/sinum/47-2/72665.html>

[†]School of Mathematics, University of Minnesota, Minneapolis, MN 55455 (cockburn@math.umn.edu). This author's research was supported in part by the National Science Foundation (grant DMS-0712955) and by the University of Minnesota Supercomputing Institute.

[‡]Department of Mathematics, University of Florida, Gainesville, FL 32611–8105 (jayg@math.ufl.edu). This author's research was supported in part by the National Science Foundation under grants DMS-0713833 and CREMS-0619080.

numerical traces, which can be used to perform hybridization. This idea was exploited in the context of the Poisson-like equations in [10]. It will feature again in this paper, manifesting in a more complicated form suited to the Stokes system.

Let us put this contribution in perspective. This paper can be considered part of a series of papers in which we study hybridization of finite element methods. The hybridization of classical mixed methods for second-order elliptic problems was considered in [5, 6]. Hybridization of a DG method for the two-dimensional Stokes system was carried out in [3], while hybridization of a mixed method for the three-dimensional Stokes system was developed in [7, 8]. A short review of the work done up to 2005 is provided in [9].

Recently in [10] it was shown how mixed, discontinuous, continuous, and even nonconforming Galerkin methods can be hybridized in a single, unifying framework. This was done for second-order elliptic problems. In this paper, we extend this approach to Galerkin methods for the Stokes problem. However, although the hybridization techniques we propose here provide a similar unifying framework, we prefer to sacrifice generality for the sake of clarity and concentrate our efforts on a particular, new class of methods we call the hybridizable discontinuous Galerkin (HDG) methods. Then, just as was done for second-order elliptic problems in [10], we show that this procedure also applies to mixed and other classic methods which can be obtained as particular or limiting cases of these HDG methods.

Our results are also an extension of previous work on hybridization of a DG [3] and a classical mixed method [7, 8] for the Stokes equations. For these two methods, hybridization was used to circumvent the difficult task of constructing a local basis for divergence-free spaces for velocity. Moreover, in [7, 8], it was shown that hybridization results in a *new* formulation of the method which *only* involves the tangential velocity and the pressure on the faces of the elements. In this paper, we show that such a formulation can also be obtained for the HDG methods. We also show that these methods can be hybridized in *three* additional ways differing in the choice of variables which are globally coupled.

The organization of the paper is as follows. In section 2, we present the HDG methods and show that their approximate solution is well defined. In section 3, we present the four hybridizations of the HDG methods in full detail. Proofs of the theorems therein are displayed in section 4. Finally, in section 5, we end with some concluding remarks.

2. The HDG methods.

2.1. Definition of the methods. Let us describe the HDG methods under consideration. We begin by introducing our notation. We denote by $\Omega_h = \{K\}$ a subdivision of the domain Ω into shape-regular tetrahedra K satisfying the usual assumptions of finite element meshes and set $\partial\Omega_h := \{\partial K : K \in \Omega_h\}$. We associate to this mesh the set of interior faces \mathcal{E}_h^o and the set of boundary faces \mathcal{E}_h^∂ . We say that $e \in \mathcal{E}_h^o$ if there are two tetrahedra K^+ and K^- in Ω_h such that $e = \partial K^+ \cap \partial K^-$, and we say that $e \in \mathcal{E}_h^\partial$ if there is a tetrahedra K in Ω_h such that $e = \partial K \cap \partial\Omega$. We set $\mathcal{E}_h := \mathcal{E}_h^o \cup \mathcal{E}_h^\partial$.

The HDG methods provide an approximate solution $(\omega_h, \mathbf{u}_h, p_h)$ in some finite-dimensional space $\mathbf{W}_h \times \mathbf{V}_h \times P_h$ of the form

$$\begin{aligned}\mathbf{W}_h &= \{\boldsymbol{\tau} \in \mathbf{L}^2(\Omega) : \boldsymbol{\tau}|_K \in \mathbf{W}(K) \quad \forall K \in \Omega_h\}, \\ \mathbf{V}_h &= \{\mathbf{v} \in \mathbf{L}^2(\Omega) : \mathbf{v}|_K \in \mathbf{V}(K) \quad \forall K \in \Omega_h\}, \\ P_h &= \{q \in L^2(\Omega) : q|_K \in P(K) \quad \forall K \in \Omega_h\},\end{aligned}$$

where the local spaces $\mathbf{W}(K), \mathbf{V}(K)$, and $P(K)$ are finite-dimensional polynomial spaces that we shall specify later.

To define the approximate solution, we use the following formulation of the Stokes equations:

$$\begin{aligned}
 (2.1a) \quad & \boldsymbol{\omega} - \mathbf{curl} \mathbf{u} = 0 && \text{in } \Omega, \\
 (2.1b) \quad & \mathbf{curl} \boldsymbol{\omega} + \text{grad} p = \mathbf{f} && \text{in } \Omega, \\
 (2.1c) \quad & \text{div} \mathbf{u} = 0 && \text{in } \Omega, \\
 (2.1d) \quad & \mathbf{u} = \mathbf{g} && \text{on } \partial\Omega.
 \end{aligned}$$

Multiplying the first three equations by test functions and integrating by parts, we arrive at the following formulation for determining an approximate solution $(\boldsymbol{\omega}_h, \mathbf{u}_h, p_h)$ in $\mathbf{W}_h \times \mathbf{V}_h \times P_h$:

$$\begin{aligned}
 (2.2a) \quad & (\boldsymbol{\omega}_h, \boldsymbol{\tau})_{\Omega_h} - (\mathbf{u}_h, \mathbf{curl} \boldsymbol{\tau})_{\Omega_h} + \langle \hat{\mathbf{u}}_h, \mathbf{n} \times \boldsymbol{\tau} \rangle_{\partial\Omega_h} = 0, \\
 (2.2b) \quad & (\boldsymbol{\omega}_h, \mathbf{curl} \mathbf{v})_{\Omega_h} + \langle \hat{\boldsymbol{\omega}}_h, \mathbf{v} \times \mathbf{n} \rangle_{\partial\Omega_h} \\
 & - (p_h, \text{div} \mathbf{v})_{\Omega_h} + \langle \hat{p}_h, \mathbf{v} \cdot \mathbf{n} \rangle_{\partial\Omega_h} = (\mathbf{f}, \mathbf{v})_{\Omega_h}, \\
 (2.2c) \quad & - (\mathbf{u}_h, \text{grad} q)_{\Omega_h} + \langle \hat{\mathbf{u}}_h \cdot \mathbf{n}, q \rangle_{\partial\Omega_h} = 0
 \end{aligned}$$

for all $(\boldsymbol{\tau}, \mathbf{v}, q) \in \mathbf{W}_h \times \mathbf{V}_h \times P_h$. The notation for volume innerproducts above is defined by

$$(\zeta, \omega)_{\Omega_h} := \sum_{K \in \Omega_h} \int_K \zeta(x) \omega(x) dx \quad \text{and} \quad (\boldsymbol{\sigma}, \mathbf{v})_{\Omega_h} := \sum_{i=1}^3 (\sigma_i, v_i)_{\Omega_h}$$

for all ζ, ω in $L^2(\Omega_h) := \{v : v|_K \in L^2(K) \text{ for all } K \text{ in } \Omega_h\}$, and all $\boldsymbol{\sigma}, \mathbf{v} \in \mathbf{L}^2(\Omega_h) := [L^2(\Omega_h)]^3$. More generally, our notation is such that if S represents the notation for any given space (e.g., S can be L^2, H^1 , etc.), the bold face notation $\mathbf{S}(\Omega_h)$ denotes $[S(\Omega_h)]^3$, and

$$\begin{aligned}
 \mathbf{S}(\Omega_h) &:= \{\omega : \Omega_h \mapsto \mathbb{R}, \omega|_K \in S(K) \forall K \in \Omega_h\}, \\
 \mathbf{S}(\partial\Omega_h) &:= \{\omega : \partial\Omega_h \mapsto \mathbb{R}, \omega|_{\partial K} \in S(\partial K) \forall K \in \Omega_h\}.
 \end{aligned}$$

The boundary innerproducts in (2.2) are defined by

$$\langle \mathbf{v} \odot \mathbf{n}, \mu \rangle_{\partial\Omega_h} := \sum_{K \in \Omega_h} \int_{\partial K} \mathbf{v}(\gamma) \odot \mathbf{n} \mu(\gamma) d\gamma,$$

where \odot is either \cdot (the dot product) or \times (the cross product) and \mathbf{n} denotes the unit outward normal vector on ∂K . Similarly, for any $\mathcal{F}_h \subseteq \mathcal{E}_h$, the notation $\langle \cdot, \cdot \rangle_{\mathcal{F}_h}$ indicates a sum of integrals over the faces in \mathcal{F}_h .

To complete the definition of the HDG methods, we need to specify the numerical traces, for which we need the following notation. For any vector-valued function \mathbf{v} we set

$$\begin{aligned}
 (2.3a) \quad & \mathbf{v}_t := \mathbf{n} \times (\mathbf{v} \times \mathbf{n}), \\
 (2.3b) \quad & \mathbf{v}_n := \mathbf{n} (\mathbf{v} \cdot \mathbf{n}).
 \end{aligned}$$

Note that we have that $\mathbf{v} = \mathbf{v}_n + \mathbf{v}_t$. In this paper we will often use double-valued functions on \mathcal{E}_h^o . One example is \mathbf{n} . Indeed, on each interior mesh face $e = \partial K^+ \cap \partial K^-$,

the unit normal \mathbf{n} is double valued with two branches, one from K^+ , which we denote by \mathbf{n}^+ , and another from K^- , which we denote by \mathbf{n}^- . Similarly, if \mathbf{v} is in $\mathbf{H}^1(\Omega_h)$, its full trace, as well as the tangential and normal traces in (2.3), are generally double valued on \mathcal{E}_h^o . We use \mathbf{v}^+ and \mathbf{v}^- to denote the full trace on e of \mathbf{v} from K^+ and K^- , respectively. On each $e = \partial K^+ \cap \partial K^-$, the jumps of double-valued functions \mathbf{v} in $\mathbf{H}^1(\Omega_h)$ and q in $H^1(\Omega_h)$ are defined by

$$(2.4a) \quad \llbracket q \mathbf{n} \rrbracket := q^+ \mathbf{n}^+ + q^- \mathbf{n}^-,$$

$$(2.4b) \quad \llbracket \mathbf{v} \odot \mathbf{n} \rrbracket := \mathbf{v}^+ \odot \mathbf{n}^+ + \mathbf{v}^- \odot \mathbf{n}^-,$$

where \odot is either \cdot or \times .

With these preparations we can now specify our definition of the numerical traces appearing in (2.2). On the interior faces \mathcal{E}_h^o , we set

$$(2.5a) \quad (\widehat{\boldsymbol{\omega}}_h)_t = \left(\frac{\tau_t^- (\boldsymbol{\omega}_h^+)_t + \tau_t^+ (\boldsymbol{\omega}_h^-)_t}{\tau_t^- + \tau_t^+} \right) + \left(\frac{\tau_t^+ \tau_t^-}{\tau_t^- + \tau_t^+} \right) \llbracket \mathbf{u}_h \times \mathbf{n} \rrbracket,$$

$$(2.5b) \quad (\widehat{\mathbf{u}}_h)_t = \left(\frac{\tau_t^+ (\mathbf{u}_h^+)_t + \tau_t^- (\mathbf{u}_h^-)_t}{\tau_t^- + \tau_t^+} \right) + \left(\frac{1}{\tau_t^- + \tau_t^+} \right) \llbracket \mathbf{n} \times \boldsymbol{\omega}_h \rrbracket,$$

$$(2.5c) \quad (\widehat{\mathbf{u}}_h)_n = \left(\frac{\tau_n^+ (\mathbf{u}_h^+)_n + \tau_n^- (\mathbf{u}_h^-)_n}{\tau_n^- + \tau_n^+} \right) + \left(\frac{1}{\tau_n^- + \tau_n^+} \right) \llbracket p_h \mathbf{n} \rrbracket,$$

$$(2.5d) \quad \widehat{p}_h = \left(\frac{\tau_n^- p_h^+ + \tau_n^+ p_h^-}{\tau_n^- + \tau_n^+} \right) + \left(\frac{\tau_n^+ \tau_n^-}{\tau_n^- + \tau_n^+} \right) \llbracket \mathbf{u}_h \cdot \mathbf{n} \rrbracket,$$

where the so-called *penalization* or *stabilization* parameters τ_t and τ_n are functions on \mathcal{E}_h that are constant on each e in \mathcal{E}_h and *double* valued on \mathcal{E}_h^o ; indeed, if $e = \partial K^+ \cap \partial K^-$, then τ_t^\pm and τ_n^\pm are the values on $e \cap \partial K^\pm$ of the stabilization parameters. Finally, on the boundary faces of \mathcal{E}_h^∂ , we set

$$(2.6a) \quad (\widehat{\mathbf{u}}_h)_t = \mathbf{g}_t,$$

$$(2.6b) \quad (\widehat{\mathbf{u}}_h)_n = \mathbf{g}_n,$$

$$(2.6c) \quad (\widehat{\boldsymbol{\omega}}_h)_t = (\boldsymbol{\omega}_h)_t + \tau_t (\mathbf{u}_h - \widehat{\mathbf{u}}_h) \times \mathbf{n},$$

$$(2.6d) \quad \widehat{p}_h = p_h + \tau_n (\mathbf{u}_h - \widehat{\mathbf{u}}_h) \cdot \mathbf{n}.$$

This completes the definition of the HDG method in (2.2), save the specification of the spaces on each element.

Let us briefly motivate the choice of the above numerical traces. First, we want them to be linear combinations of the traces of the approximate solution $(\boldsymbol{\omega}_h, \mathbf{u}_h, p_h)$. We also want them to be *consistent* and *conservative*; these are very important properties of the numerical traces as was shown in [1] in the context of second-order elliptic problems. They are consistent because when the approximate solution is continuous across interelement boundaries, or at the boundary of Ω , we have that

$$((\widehat{\boldsymbol{\omega}}_h)_t, (\widehat{\mathbf{u}}_h)_t, (\widehat{\mathbf{u}}_h)_n, \widehat{p}_h) = ((\boldsymbol{\omega}_h)_t, (\mathbf{u}_h)_t, (\mathbf{u}_h)_n, p_h).$$

They are conservative because they are single valued.

The above general considerations, however, are not enough to justify the specific expression of the numerical traces on the parameters τ_t and τ_n . We take this particular

expression because it allows the hybridization of the methods. Although this will become evident when we develop each of its four hybridizations, we can briefly argue why this is so. Suppose that we want the numerical trace of the velocity, $\widehat{\mathbf{u}}_h = (\widehat{\mathbf{u}}_h)_t + (\widehat{\mathbf{u}}_h)_n$, to be the globally coupled unknown. This means that, on each element $K \in \Omega_h$, we should be able to express *all* the remaining unknowns in terms of $\widehat{\mathbf{u}}_h$. If in the weak formulation defining the method, (2.2), we take test functions with support in the element K , we see that we can achieve this if we could write

$$(\widehat{\boldsymbol{\omega}}_h)_t = (\boldsymbol{\omega}_h)_t + \tau_t (\mathbf{u}_h - \widehat{\mathbf{u}}_h) \times \mathbf{n} \quad \text{and} \quad \widehat{p}_h = p_h + \tau_n (\mathbf{u}_h - \widehat{\mathbf{u}}_h) \cdot \mathbf{n},$$

where $(\boldsymbol{\omega}_h, \mathbf{u}_h, p_h)$ is the approximation on the element K , \mathbf{n} is the outward unit normal to K , and τ_t and τ_n take the values associated with K . Note that this is consistent with the choice of the corresponding numerical traces on the border of Ω , equations (2.6c) and (2.6d). Since the element K was arbitrary, we should then have

$$\begin{aligned} (\widehat{\boldsymbol{\omega}}_h)_t &= (\boldsymbol{\omega}_h)_t^+ + \tau_t^+ (\mathbf{u}_h^+ - \widehat{\mathbf{u}}_h) \times \mathbf{n}^+ = (\boldsymbol{\omega}_h)_t^- + \tau_t^- (\mathbf{u}_h^- - \widehat{\mathbf{u}}_h) \times \mathbf{n}^-, \\ \widehat{p}_h &= p_h^+ + \tau_n^+ (\mathbf{u}_h^+ - \widehat{\mathbf{u}}_h) \cdot \mathbf{n}^+ = p_h^- + \tau_n^- (\mathbf{u}_h^- - \widehat{\mathbf{u}}_h) \cdot \mathbf{n}^- \end{aligned}$$

on all interior faces. A simple algebraic manipulation shows that this is possible only if the numerical traces therein are taken as in (2.5).

Let us end this subsection by remarking that the choice of the penalization parameters τ_t and τ_n can be crucial since it can have an important effect on both the stability and the accuracy of the method. This constitutes ongoing work; see the last paragraph of section 5. In subsection 3.5, we show how, by taking special choices of these parameters, several already known methods for the Stokes system are recovered.

2.2. Other boundary conditions. The vorticity-velocity variational formulation admits imposition of boundary conditions other than (2.1d); see a short discussion in subsection 4.3 in [16]. In this paper, we consider the following types of boundary conditions:

$$(2.7a) \quad \left. \begin{aligned} \mathbf{u}_t &= \mathbf{g}_t \\ p &= r \end{aligned} \right\} \text{Type I boundary conditions,}$$

$$(2.7b) \quad \left. \begin{aligned} \mathbf{u}_t &= \mathbf{g}_t \\ \mathbf{u}_n &= \mathbf{g}_n \end{aligned} \right\} \text{Type II boundary conditions,}$$

$$(2.7c) \quad \left. \begin{aligned} \boldsymbol{\omega}_t &= \boldsymbol{\gamma}_t \\ \mathbf{u}_n &= \mathbf{g}_n \end{aligned} \right\} \text{Type III boundary conditions,}$$

$$(2.7d) \quad \left. \begin{aligned} \boldsymbol{\omega}_t &= \boldsymbol{\gamma}_t \\ p &= r \end{aligned} \right\} \text{Type IV boundary conditions.}$$

We have already defined the HDG method in the case of the Type II boundary conditions in the previous subsection. Neither the equations of the HDG method (2.2) nor the equations of the interior numerical traces (2.5a)–(2.5d) change when the other boundary conditions are considered. But the equations for the boundary numerical

traces, namely (2.6a)–(2.6d), must be changed as follows:

$$\begin{aligned}
 (2.8a) \quad & \left. \begin{aligned}
 (\hat{\mathbf{u}}_h)_t &= \mathbf{g}_t, \\
 (\hat{\mathbf{u}}_h)_n &= (\mathbf{u}_h)_n + \frac{1}{\tau_n}(p_h - \hat{p}_h)\mathbf{n}, \\
 \hat{\boldsymbol{\omega}}_h &= (\boldsymbol{\omega}_h)_t + \tau_t(\mathbf{u}_h - \hat{\mathbf{u}}_h) \times \mathbf{n}, \\
 \hat{p}_h &= r,
 \end{aligned} \right\} \text{for Type I,} \\
 (2.8b) \quad & \left. \begin{aligned}
 (\hat{\mathbf{u}}_h)_t &= (\mathbf{u}_h)_t + \frac{1}{\tau_t}\mathbf{n} \times (\boldsymbol{\omega}_h - \hat{\boldsymbol{\omega}}_h), \\
 (\hat{\mathbf{u}}_h)_n &= \mathbf{g}_n, \\
 (\hat{\boldsymbol{\omega}}_h)_t &= \boldsymbol{\gamma}_t, \\
 \hat{p}_h &= p_h + \tau_n(\mathbf{u}_h - \hat{\mathbf{u}}_h) \cdot \mathbf{n},
 \end{aligned} \right\} \text{for Type III,} \\
 (2.8c) \quad & \left. \begin{aligned}
 (\hat{\mathbf{u}}_h)_t &= (\mathbf{u}_h)_t + \frac{1}{\tau_t}\mathbf{n} \times (\boldsymbol{\omega}_h - \hat{\boldsymbol{\omega}}_h), \\
 (\hat{\mathbf{u}}_h)_n &= (\mathbf{u}_h)_n + \frac{1}{\tau_n}(p_h - \hat{p}_h)\mathbf{n}, \\
 (\hat{\boldsymbol{\omega}}_h)_t &= \boldsymbol{\gamma}_t, \\
 \hat{p}_h &= r,
 \end{aligned} \right\} \text{for Type IV.}
 \end{aligned}$$

When we do not have boundary conditions on pressure, the pressure variable in Stokes flow is determined only up to a constant. Therefore, for Type II and Type III boundary conditions, in order to obtain unique solvability we must change the pressure space from P_h to

$$P_h^0 = P_h \cap L_0^2(\Omega),$$

where $L_0^2(\Omega)$ is the set of functions in $L^2(\Omega)$ whose mean on Ω is zero. In the case of Type I and Type IV boundary conditions, the pressure space is simply P_h . Finally, let us point out that the Type IV boundary conditions are not particularly useful since they have to be complemented by additional conditions on the velocity. For this reason, we do not consider them as possible boundary conditions for the Stokes equations. However, we discuss them here because, as we are going to see, there is a one-to-one correspondence between the four types of boundary conditions just considered and the four hybridizations of the HDG method.

2.3. Existence and uniqueness of the HDG solution. With (strictly) positive penalty parameters, the HDG method is well defined, as we next show. When we say that a multivalued function τ is positive on $\partial\Omega_h$, we mean that both branches of τ are positive on all faces of \mathcal{E}_h^o and furthermore that the branch from within Ω is positive on the faces of $\partial\Omega$. Of course, the branch from outside Ω is zero.

To simplify our notation, we will use a symbol for averages of double-valued functions. On any interior face $e = \partial K^+ \cap \partial K^-$, let

$$\{\{v\}\}_\alpha = v^+ \alpha^+ + v^- \alpha^-$$

for any double-valued function α . The notation $\{\{v\}\}$ (without a subscript) denotes $\{\{v\}\}_\alpha$ with $\alpha^+ = \alpha^- = 1/2$. As a final note on our notation, we do not distinguish between functions and their extensions by zero. Accordingly, we use the previously defined notations like $[\cdot]$ and $\{\{\cdot\}\}$ even for boundary faces in \mathcal{E}_h^∂ with the understanding that one of the branches involved is zero (which is the case when the function is

extended by zero); e.g., on a boundary face e , the penalty function τ_n has only one nonzero branch, say τ_n^- , so $\{\{\tau_n\}\}$ on e equals $\tau_n^-/2$. With this notation it is easy to verify that the identities

$$(2.9a) \quad \langle \boldsymbol{\sigma}, \mathbf{v} \times \mathbf{n} \rangle_{\partial\Omega_h} = \langle \{\{\boldsymbol{\sigma}\}\}_\alpha, \llbracket \mathbf{v} \times \mathbf{n} \rrbracket \rangle_{\mathcal{E}_h} - \langle \llbracket \boldsymbol{\sigma} \times \mathbf{n} \rrbracket, \{\{\mathbf{v}\}\}_{1-\alpha} \rangle_{\mathcal{E}_h},$$

$$(2.9b) \quad \langle q, \mathbf{v} \cdot \mathbf{n} \rangle_{\partial\Omega_h} = \langle \{\{q\}\}_\alpha, \llbracket \mathbf{v} \cdot \mathbf{n} \rrbracket \rangle_{\mathcal{E}_h} + \langle \llbracket q \mathbf{n} \rrbracket, \{\{\mathbf{v}\}\}_{1-\alpha} \rangle_{\mathcal{E}_h}$$

hold for any α whose branches sum to one, i.e., $\alpha^+ + \alpha^- = 1$ on every face e in \mathcal{E}_h .

PROPOSITION 2.1. *Assume that τ_t and τ_n are positive on $\partial\Omega_h$. Assume also that*

$$\begin{aligned} \mathbf{curl} \mathbf{V}(K) &\subset \mathbf{W}(K), \\ \mathbf{grad} P(K) &\subset \mathbf{V}(K), \\ \mathbf{div} \mathbf{V}(K) &\subset P(K) \end{aligned}$$

for every element $K \in \Omega_h$. Then we have the following:

1. For the Type I boundary conditions, there is one and only one $(\boldsymbol{\omega}_h, \mathbf{u}_h, p_h)$ in the space $\mathbf{W}_h \times \mathbf{V}_h \times P_h$ satisfying (2.2), (2.5), and (2.8a).
2. For the Type II boundary conditions, there is a solution $(\boldsymbol{\omega}_h, \mathbf{u}_h, p_h)$ in the space $\mathbf{W}_h \times \mathbf{V}_h \times P_h$ satisfying (2.2), (2.5), and (2.6) if and only if \mathbf{g} satisfies (1.1). When a solution $(\boldsymbol{\omega}_h, \mathbf{u}_h, p_h)$ exists, all solutions are of the form $(\boldsymbol{\omega}_h, \mathbf{u}_h, p_h + \kappa)$ for some constant function κ . There is a unique solution if P_h is replaced by P_h^0 .
3. For Type III, the statements of the Type II case holds verbatim after replacing (2.6) with (2.8b).

Proof. The proof proceeds by setting all data to zero and finding the null space in each of the three cases. Taking $(\boldsymbol{\tau}, \mathbf{v}, q) := (\boldsymbol{\omega}_h, \mathbf{u}_h, p_h)$ in (2.2) and adding them, we obtain

$$(2.10) \quad (\boldsymbol{\omega}_h, \boldsymbol{\omega}_h)_{\Omega_h} + \Theta_h = 0,$$

where

$$\begin{aligned} \Theta_h := & \langle -\mathbf{u}_h, \mathbf{n} \times \boldsymbol{\omega}_h \rangle_{\partial\Omega_h} + \langle \widehat{\mathbf{u}}_h, \mathbf{n} \times \boldsymbol{\omega}_h \rangle_{\partial\Omega_h} - \langle \mathbf{u}_h, \mathbf{n} \times \widehat{\boldsymbol{\omega}}_h \rangle_{\partial\Omega_h} \\ & - \langle p_h, \mathbf{u}_h \cdot \mathbf{n} \rangle_{\partial\Omega_h} + \langle \widehat{p}_h, \mathbf{u}_h \cdot \mathbf{n} \rangle_{\partial\Omega_h} + \langle p_h, \widehat{\mathbf{u}}_h \cdot \mathbf{n} \rangle_{\partial\Omega_h}. \end{aligned}$$

Rewriting Θ_h using (2.9), we obtain

$$\begin{aligned} \Theta_h = & - \langle \widehat{\boldsymbol{\omega}}_h - \{\{\boldsymbol{\omega}_h\}\}_{1-\alpha}, \llbracket \mathbf{n} \times \mathbf{u}_h \rrbracket \rangle_{\mathcal{E}_h} + \langle \widehat{\mathbf{u}}_h - \{\{\mathbf{u}_h\}\}_\alpha, \llbracket \mathbf{n} \times \widehat{\boldsymbol{\omega}}_h \rrbracket \rangle_{\mathcal{E}_h} \\ & + \langle \widehat{p}_h - \{\{p_h\}\}_{1-\beta}, \llbracket \mathbf{u}_h \cdot \mathbf{n} \rrbracket \rangle_{\mathcal{E}_h} + \langle \widehat{\mathbf{u}}_h - \{\{\mathbf{u}_h\}\}_\beta, \llbracket p_h \mathbf{n} \rrbracket \rangle_{\mathcal{E}_h} \end{aligned}$$

for any α and β whose branches sum to one on every face of \mathcal{E}_h . We set $\alpha = \tau_t/2 \{\{\tau_t\}\}$ and $\beta = \tau_n/2 \{\{\tau_n\}\}$ on all the interior faces of \mathcal{E}_h^o . On the remaining boundary faces, we set α and β case by case as follows, letting $\alpha_{\partial\Omega}^-, \beta_{\partial\Omega}^-$ and $\alpha_{\partial\Omega}^+, \beta_{\partial\Omega}^+$ denote the branches of α, β from outside and inside Ω , respectively.

For the Type I case, we set $\alpha_{\partial\Omega}^+ = 0, \alpha_{\partial\Omega}^- = 1, \beta_{\partial\Omega}^+ = 1, \beta_{\partial\Omega}^- = 0$. Then, inserting the expressions for the interior and boundary numerical traces given by (2.5) and (2.8a), we obtain

$$\Theta_h = \Theta_h^o + \langle \tau_t, |\mathbf{u}_h \times \mathbf{n}|^2 \rangle_{\partial\Omega} + \langle \tau_n, |p_h \mathbf{n}|^2 \rangle_{\partial\Omega},$$

where

$$\begin{aligned} \Theta_h^o &= \left\langle \frac{2}{\{\{\tau_t\}\}}, |[\mathbf{n} \times \boldsymbol{\omega}_h]|^2 \right\rangle_{\mathcal{E}_h^o} + \left\langle \frac{2}{\{\{1/\tau_t\}\}}, |[\mathbf{u}_h \times \mathbf{n}]|^2 \right\rangle_{\mathcal{E}_h^o} \\ &+ \left\langle \frac{2}{\{\{1/\tau_n\}\}}, [\mathbf{u}_h \cdot \mathbf{n}]^2 \right\rangle_{\mathcal{E}_h^o} + \left\langle \frac{2}{\{\{\tau_n\}\}}, |[p_h \mathbf{n}]|^2 \right\rangle_{\mathcal{E}_h^o}. \end{aligned}$$

Hence (2.10) implies that $\boldsymbol{\omega}_h$ vanishes, \mathbf{u}_h and p_h are continuous on Ω , and $(\mathbf{u}_h)_t$ and p_h vanish on $\partial\Omega$. With this in mind, we integrate by parts the equations defining the method, namely (2.2), to obtain

$$\begin{aligned} (\mathbf{curl} \mathbf{u}_h, \boldsymbol{\tau})_{\Omega_h} &= 0, \\ (\mathbf{grad} p_h, \mathbf{v})_{\Omega_h} &= 0, \\ (\mathbf{div} \mathbf{u}_h, q)_{\Omega_h} &= 0 \end{aligned}$$

for all $(\boldsymbol{\tau}, \mathbf{v}, q) \in \mathbf{W}_h \times \mathbf{V}_h \times P_h$. By our assumptions on the local spaces, this implies that the following (global) distributional derivatives on Ω vanish:

$$(2.11) \quad \mathbf{grad} p_h = \mathbf{0}, \quad \mathbf{div} \mathbf{u}_h = 0, \quad \text{and} \quad \mathbf{curl} \mathbf{u}_h = \mathbf{0}.$$

The first equality implies that p_h vanishes since we already found p_h to vanish on $\partial\Omega$. Moreover, since $(\mathbf{u}_h)_t$ vanishes on the boundary $\partial\Omega$, and since we have assumed that $\partial\Omega$ consists of just one connected component, the last two equalities imply that $\mathbf{u}_h = 0$. Thus, the null space is trivial.

For the Type II case, we set $\alpha_{\partial\Omega}^+ = 0$, $\alpha_{\partial\Omega}^- = 1$, $\beta_{\partial\Omega}^+ = 0$, and $\beta_{\partial\Omega}^- = 1$ and simplify Θ_h using the interior and boundary numerical traces given by (2.5) and (2.6) to find that

$$\Theta_h = \Theta_h^o + \langle \tau_t, |\mathbf{u}_h \times \mathbf{n}|^2 \rangle_{\partial\Omega} + \langle \tau_n, |\mathbf{u}_h \cdot \mathbf{n}|^2 \rangle_{\partial\Omega}.$$

Hence (2.10) implies that $\boldsymbol{\omega}_h$ vanishes, \mathbf{u}_h is continuous on Ω , and \mathbf{u}_h is zero on $\partial\Omega$, and p_h is continuous on Ω . Proceeding as in the Type I case, we find that (2.11) holds, so \mathbf{u}_h vanishes. But unlike the Type I case, we can now conclude only that p_h is constant. Thus the null space consists of $(\boldsymbol{\omega}_h, \mathbf{u}_h, p_h) = (\mathbf{0}, \mathbf{0}, \kappa)$ for constant functions κ . Hence, all statements of the proposition on the Type II case follow.

The Type III case is proved similarly. \square

It is interesting to note that the proof of the Type II case required only minimal topological assumptions on Ω , namely, that Ω is connected. However, the proof of the other two cases used the further assumptions we placed on Ω . The mixed method presented in [8] without such topological assumptions dealt only with the Type II boundary conditions.

We can now give some possible choices for polynomial spaces that can be set within each element. Clearly, Proposition 2.1 gives the conditions that we must satisfy. Let \mathcal{P}_d denote the space of polynomials of degree at most d , and let \mathcal{P}_d denote the space of vector functions whose components are polynomials in \mathcal{P}_d . Let $d_P \geq 1$, $d_V \geq 0$, $d_W \geq 0$ be some integers satisfying

$$(2.12) \quad d_P - 1 \leq d_V \leq \min(d_P + 1, d_W + 1).$$

Then if we set

$$\mathbf{W}(K) = \mathcal{P}_{d_W}, \quad \mathbf{V}(K) = \mathcal{P}_{d_V}, \quad P(K) = \mathcal{P}_{d_P},$$

the conditions of Proposition 2.1 are satisfied. Some examples are

$$(d_W, d_V, d_P) = \begin{cases} (k-1, & k-1, & k), & (k, & k-1, & k), & (k+1, & k-1, & k), \\ (k-1, & k, & k), & (k, & k, & k), & (k+1, & k, & k), \\ (k, & k+1, & k), & (k+1, & k+1, & k) \end{cases}$$

for some integer $k \geq 1$. Clearly there is greater flexibility in the choice of spaces than, for instance, in the choice of spaces for mixed methods for the Stokes problem; e.g., from (2.12) it is clear that we can choose d_W to be as large as we wish and the method continues to be well defined.

Having established that the HDG methods are well defined, we show in the next section that they can be hybridized in different ways according to the choice of variables that are globally coupled.

3. Hybridizations of the HDG methods. In this section, we will restrict ourselves to considering the Stokes problem with the Type II boundary conditions. We hybridize the HDG method for this case. As we shall see, while hybridizing we can choose to set HDG methods with the other types of boundary conditions within mesh elements.

For constructing hybridized methods based on the vorticity-velocity formulation, let us recall the following four transmission conditions for the Stokes solution components:

$$(3.1) \quad \llbracket \boldsymbol{\omega} \times \mathbf{n} \rrbracket \Big|_{\mathcal{E}_h^o} = 0, \quad \llbracket \mathbf{u} \times \mathbf{n} \rrbracket \Big|_{\mathcal{E}_h^o} = 0, \quad \llbracket \mathbf{u} \cdot \mathbf{n} \rrbracket \Big|_{\mathcal{E}_h^o} = 0, \quad \llbracket p \mathbf{n} \rrbracket \Big|_{\mathcal{E}_h^o} = 0.$$

Corresponding to these four transmission conditions, there are four variables on which boundary conditions of the following form can be prescribed:

$$(3.2) \quad \boldsymbol{\omega}_t = \boldsymbol{\gamma}_t, \quad \mathbf{u}_t = \boldsymbol{\lambda}_t, \quad \mathbf{u}_n = \boldsymbol{\lambda}_n, \quad p = \rho.$$

With this correspondence in view, we can describe our approach for constructing hybridization techniques as follows. We *pick any two* of the variables in (3.2) as unknown boundary values on the boundary of *each* mesh element. (Once these values are known, the solution inside the element can be computed locally.) Then, we formulate a global system of equations for the chosen unknown variables, using the transmission conditions on the *other two* variables in (3.1). Of course, we must identify the proper discrete versions of these transmission conditions for this purpose. According to this strategy, there appears to be six possible cases. But two of the six cases yield underdetermined or overdetermined systems. For instance if we pick $\boldsymbol{\gamma}_t$ and $\boldsymbol{\lambda}_t$ as unknowns, counting their components, we would have a total of four scalar unknown functions. However, the transmission conditions (the last two in (3.1)) form only two scalar equations so will yield an underdetermined system. Similarly, if we pick $\boldsymbol{\lambda}_n$ and ρ as the unknowns, we get an overdetermined system. We discard these two possibilities. In the remainder, we now work out the specifics for the remaining four cases.

3.1. Hybridization of Type I.

A formulation with tangential velocity and pressure. Here, we choose the second and the last of the variables in (3.2), namely $(\mathbf{u})_t$ and p , as the unknowns on the mesh interfaces. Their discrete approximations will be denoted by $\boldsymbol{\lambda}_t$ and ρ , respectively. We shall then use the transmission conditions on the other two variables, namely,

$$(3.3) \quad \llbracket \boldsymbol{\omega} \times \mathbf{n} \rrbracket \Big|_{\mathcal{E}_h^o} = 0 \quad \text{and} \quad \llbracket \mathbf{u} \cdot \mathbf{n} \rrbracket \Big|_{\mathcal{E}_h^o} = 0,$$

to derive a hybridized formulation that will help us solve for the approximations $\boldsymbol{\lambda}_t$ and ρ .

The success of this approach relies on us being able to compute approximate solutions within each element locally, once the discrete approximations $\boldsymbol{\lambda}_t \approx \mathbf{u}_t$ and $\rho \approx p$ are found. In other words, we need a discretization of the following Stokes problem on *one* element:

$$\begin{aligned} \boldsymbol{\omega}_K - \mathbf{curl} \mathbf{u}_K &= 0 && \text{in } K, \\ \mathbf{curl} \boldsymbol{\omega}_K + \text{grad } p_K &= \mathbf{f} && \text{in } K, \\ \text{div } \mathbf{u}_K &= 0 && \text{in } K, \\ (\mathbf{u}_K)_t &= \boldsymbol{\lambda}_t && \text{on } \partial K, \\ p_K &= \rho && \text{on } \partial K. \end{aligned}$$

We use the HDG method (with Type I boundary conditions) applied to a single element as our discretization. Specifically, given $(\boldsymbol{\lambda}_t, \rho, \mathbf{f})$ in $\mathbf{L}^2(\partial\Omega_h) \times L^2(\partial\Omega_h) \times \mathbf{L}^2(\Omega)$, we define $(\mathbf{W}, \mathbf{u}, \mathcal{P})$ in $\mathbf{W}_h \times \mathbf{V}_h \times P_h$ on the element $K \in \Omega_h$ as the function in $\mathbf{W}(K) \times \mathbf{V}(K) \times P(K)$ satisfying

$$(3.4a) \quad (\mathbf{W}, \boldsymbol{\tau})_K - (\mathbf{u}, \mathbf{curl} \boldsymbol{\tau})_K = -\langle \boldsymbol{\lambda}_t, \mathbf{n} \times \boldsymbol{\tau} \rangle_{\partial K},$$

$$(3.4b) \quad (\mathbf{W}, \mathbf{curl} \mathbf{v})_K + \langle \widehat{\mathbf{W}}, \mathbf{v} \times \mathbf{n} \rangle_{\partial K} - (\mathcal{P}, \text{div } \mathbf{v})_K = (\mathbf{f}, \mathbf{v})_K - \langle \rho, \mathbf{v} \cdot \mathbf{n} \rangle_{\partial K},$$

$$(3.4c) \quad -(\mathbf{u}, \text{grad } q)_K + \langle \widehat{\mathbf{u}} \cdot \mathbf{n}, q \rangle_{\partial K} = 0$$

for all $(\boldsymbol{\tau}, \mathbf{v}, q) \in \mathbf{W}(K) \times \mathbf{V}(K) \times P(K)$, where

$$(3.4d) \quad \widehat{\mathbf{u}}_n = (\mathbf{u})_n + \frac{1}{\tau_n} (\mathcal{P} - \rho) \mathbf{n},$$

$$(3.4e) \quad \widehat{\mathbf{W}} = \mathbf{W} + \tau_t (\mathbf{u} - \boldsymbol{\lambda}_t) \times \mathbf{n}.$$

Note that the above system (3.4) is obtained from the HDG system (2.2) with Ω set to K and the numerical traces set by (2.8a) (and there are no interior faces). The above system of equations thus defines a linear map (the “local solver”)

$$(3.4f) \quad (\boldsymbol{\lambda}_t, \rho, \mathbf{f}) \xrightarrow{\mathcal{L}^1} (\mathbf{W}, \mathbf{u}, \mathcal{P})$$

due to the unique solvability of the HDG method on one element, as given by Proposition 2.1(1).

Next, we identify conditions on $\boldsymbol{\lambda}_t$ and ρ that make $(\mathbf{W}, \mathbf{u}, \mathcal{P})$ identical to the approximation $(\boldsymbol{\omega}_h, \mathbf{u}_h, p_h)$. We begin by restricting the function $(\boldsymbol{\lambda}_t, \rho)$ to the space $(\mathbf{M}_h)_t \times \Psi_h$, where

$$(3.5a) \quad (\mathbf{M}_h)_t := \{ \boldsymbol{\mu}_t \in \mathbf{L}^2(\mathcal{E}_h) : \boldsymbol{\mu}_t|_e \in \mathbf{M}(e) \quad \forall e \in \mathcal{E}_h^o \},$$

$$(3.5b) \quad \Psi_h := \{ \psi \in L^2(\mathcal{E}_h) : \psi|_e \in \Psi(e) \quad \forall e \in \mathcal{E}_h \},$$

where, on each face $e \in \mathcal{E}_h$, the finite-dimensional spaces $\mathbf{M}(e)$ and $\Psi(e)$ are such that

$$(3.5c) \quad \mathbf{M}(e) \supseteq \{ (\mathbf{v}_t + \mathbf{n} \times \boldsymbol{\tau})|_e : (\boldsymbol{\tau}, \mathbf{v}) \in \mathbf{W}(K) \times \mathbf{V}(K) \quad \forall K : e \subset \partial K \},$$

$$(3.5d) \quad \Psi(e) \supseteq \{ (q + \mathbf{v} \cdot \mathbf{n})|_e : (\mathbf{v}, q) \in \mathbf{V}(K) \times P(K) \quad \forall K : e \subset \partial K \}.$$

The next theorem identifies discrete analogues of the transmission conditions (3.3) as the requirements for recovering the discrete solution.

THEOREM 3.1 (conditions for Type I hybridization). *Suppose $(\boldsymbol{\omega}_h, \mathbf{u}_h, p_h)$ is the solution of the HDG method defined by (2.2), (2.5), and (2.6). Assume that $(\boldsymbol{\lambda}_t, \rho) \in (\mathbf{M}_h)_t \times \Psi_h$ is such that*

$$(3.6a) \quad \boldsymbol{\lambda}_t = \mathbf{g}_t \quad \text{on } \partial\Omega,$$

$$(3.6b) \quad \langle \llbracket \mathbf{n} \times \widehat{\mathbf{W}} \rrbracket, \boldsymbol{\mu}_t \rangle_{\mathcal{E}_h^o} = 0 \quad \forall \boldsymbol{\mu}_t \in \mathbf{M}_h,$$

$$(3.6c) \quad \langle \llbracket \widehat{\mathbf{U}} \cdot \mathbf{n} \rrbracket, \psi \rangle_{\mathcal{E}_h} = \langle \mathbf{g} \cdot \mathbf{n}, \psi \rangle_{\partial\Omega} \quad \forall \psi \in \Psi_h,$$

$$(3.6d) \quad (\mathcal{P}, 1)_\Omega = 0.$$

Then $(\mathbf{W}, \mathbf{u}, \mathcal{P}) = (\boldsymbol{\omega}_h, \mathbf{u}_h, p_h)$, $\boldsymbol{\lambda}_t = (\widehat{\mathbf{u}}_h)_t$, and $\rho = \widehat{p}_h$.

Proof. We begin by noting that $(\mathbf{W}, \mathbf{u}, \mathcal{P})$ is in the space $\mathbf{W}_h \times \mathbf{V}_h \times P_h$, by the definition of the local solvers. Moreover, by adding the equations defining the local solver, namely (3.4a)–(3.4c), we find that $(\mathbf{W}, \mathbf{u}, \mathcal{P})$ satisfies the equations of (2.2), with $(\widehat{\mathbf{W}})_t$ in place of $(\widehat{\boldsymbol{\omega}}_h)_t$, $\boldsymbol{\lambda}_t$ in place of $(\widehat{\mathbf{u}}_h)_t$, $(\widehat{\mathbf{U}})_n$ in place of $(\widehat{\mathbf{u}}_h)_t$, and ρ in place of \widehat{p}_h . Hence, if we show that $(\widehat{\mathbf{W}})_t$, $\boldsymbol{\lambda}_t$, $(\widehat{\mathbf{U}})_n$, and ρ can be related to $(\mathbf{W}, \mathbf{u}, \mathcal{P})$, as in the expressions for the numerical traces (2.5a)–(2.5d), then the proof will be complete because of the uniqueness result of Proposition 2.1(2) (which applies due to condition (3.6d)).

Therefore, let us first derive such expressions for $\boldsymbol{\lambda}_t$ and ρ . By the choice of the space $\mathbf{M}_h \times \Psi_h$, the jump conditions (3.6b) and (3.6c) imply that

$$\llbracket \mathbf{n} \times \widehat{\mathbf{W}} \rrbracket = 0 \quad \text{and} \quad \llbracket \widehat{\mathbf{U}} \cdot \mathbf{n} \rrbracket = 0 \quad \text{on } \mathcal{E}_h^o.$$

Inserting the definition of the numerical traces (3.4d) and (3.4e), we readily obtain that, on \mathcal{E}_h^o ,

$$\begin{aligned} \llbracket \mathbf{n} \times \mathbf{W} \rrbracket + \tau_t^+ (\mathbf{U}^+)_t + \tau_t^- (\mathbf{U}^-)_t - (\tau_t^+ + \tau_t^-) \boldsymbol{\lambda}_t &= 0, \\ \llbracket \mathbf{u} \cdot \mathbf{n} \rrbracket + \frac{1}{\tau_n^+} \mathcal{P}^+ + \frac{1}{\tau_n^-} \mathcal{P}^- - \left(\frac{1}{\tau_n^+} + \frac{1}{\tau_n^-} \right) \rho &= 0, \end{aligned}$$

or, equivalently,

$$\begin{aligned} \boldsymbol{\lambda}_t &= \left(\frac{\tau_t^- (\mathbf{U}^+)_t + \tau_t^+ (\mathbf{U}^-)_t}{\tau_t^- + \tau_t^+} \right) + \left(\frac{1}{\tau_t^- + \tau_t^+} \right) \llbracket \mathbf{n} \times \mathbf{W} \rrbracket, \\ \rho &= \left(\frac{\tau_n^- \mathcal{P}^+ + \tau_n^+ \mathcal{P}^-}{\tau_n^- + \tau_n^+} \right) + \left(\frac{\tau_n^- \tau_n^+}{\tau_n^- + \tau_n^+} \right) \llbracket \mathbf{u} \cdot \mathbf{n} \rrbracket. \end{aligned}$$

Substituting these expressions into (3.4d) and (3.4e), we obtain

$$\begin{aligned} (\widehat{\mathbf{W}})_t &= \left(\frac{\tau_t^- (\mathbf{W}^+)_t + \tau_t^+ (\mathbf{W}^-)_t}{\tau_t^- + \tau_t^+} \right) + \left(\frac{\tau_t^+ \tau_t^-}{\tau_t^- + \tau_t^+} \right) \llbracket \mathbf{u} \times \mathbf{n} \rrbracket, \\ (\widehat{\mathbf{U}})_n &= \left(\frac{\tau_n^+ (\mathbf{U}^+)_n + \tau_n^- (\mathbf{U}^-)_n}{\tau_n^- + \tau_n^+} \right) + \left(\frac{1}{\tau_n^- + \tau_n^+} \right) \llbracket \mathcal{P} \mathbf{n} \rrbracket. \end{aligned}$$

In other words, the numerical traces satisfy (2.5). The fact that they satisfy (2.6a) and (2.6b) follows from conditions (3.6a) and (3.6c), respectively. Finally, (2.6c) and

(2.6d) follow directly from the definition of the numerical traces of the local solvers (3.4e) and (3.4d), respectively.

Thus, by the uniqueness result of Proposition 2.1(2), we now conclude that $(\mathcal{W}, \mathbf{u}, \mathcal{P})$ coincides with $(\omega_h, \mathbf{u}_h, p_h)$, and consequently, $\lambda_t = (\widehat{\mathbf{u}}_h)_t$ and $\rho = \widehat{p}_h$. This completes the proof. \square

At this point, we can comment more on our strategy for construction of hybridized DG methods. Roughly speaking, the derivation of our hybridized methods proceeds by imposing discrete versions of all four transmission conditions in (3.1) through the four numerical traces of the HDG method. The two numerical traces we picked as unknowns in this case, namely λ_t and ρ , being single valued on \mathcal{E}_h^o , already satisfy a zero-jump transmission condition, so we have in some sense already imposed the second and the fourth of the conditions in (3.1). The discrete analogues of the remaining two (the first and the third) transmission conditions are (3.6b) and (3.6c), which requires the remaining two numerical traces to be single valued. Theorem 3.1 shows that once these conditions are imposed, the HDG solution is recovered.

Next, we give a characterization of unknown traces λ_t and ρ and the discrete HDG solution $(\omega_h, \mathbf{u}_h, p_h)$ in terms of the local solvers. In particular, we show that the jump conditions (3.6b) and (3.6c) define a mixed method for the tangential velocity and the pressure. To state the result, we need to introduce some notation. Letting $\lambda_t^o = \lambda_t|_{\mathcal{E}_h^o}$, and remembering our identification of functions with their zero extension, we can write

$$\lambda_t = \lambda_t^o + \mathbf{g}_t.$$

We denote by $(\mathbf{M}_h^o)_t$ the functions of $(\mathbf{M}_h)_t$ which are zero on $\partial\Omega$ (so λ_t^o is in $(\mathbf{M}_h^o)_t$). Finally, we use the following notation for certain specific local solutions:

$$(3.7a) \quad (\mathcal{W}_{\lambda_t}, \mathbf{u}_{\lambda_t}, \mathcal{P}_{\lambda_t}) := \mathcal{L}^i(\lambda_t, 0, \mathbf{0}),$$

$$(3.7b) \quad (\mathcal{W}_\rho, \mathbf{u}_\rho, \mathcal{P}_\rho) := \mathcal{L}^i(\mathbf{0}, \rho, \mathbf{0}),$$

$$(3.7c) \quad (\mathcal{W}_f, \mathbf{u}_f, \mathcal{P}_f) := \mathcal{L}^i(\mathbf{0}, 0, \mathbf{f}),$$

where \mathcal{L}^i is as in (3.4f). We are now ready to state our main result for this case.

THEOREM 3.2 (characterization of the approximate solution). *We have that*

$$\begin{aligned} \omega_h &= \mathcal{W}_{\lambda_t^o} + \mathcal{W}_\rho + \mathcal{W}_f + \mathcal{W}_{\mathbf{g}_t}, \\ \mathbf{u}_h &= \mathbf{u}_{\lambda_t^o} + \mathbf{u}_\rho + \mathbf{u}_f + \mathbf{u}_{\mathbf{g}_t}, \\ p_h &= \mathcal{P}_{\lambda_t^o} + \mathcal{P}_\rho + \mathcal{P}_f + \mathcal{P}_{\mathbf{g}_t}, \end{aligned}$$

where (λ_t^o, ρ) is the only element of $(\mathbf{M}_h^o)_t \times \Psi_h$ such that

$$(3.8a) \quad a_h(\lambda_t^o, \boldsymbol{\mu}_t) + b_h(\rho, \boldsymbol{\mu}_t) = \ell_1(\boldsymbol{\mu}_t),$$

$$(3.8b) \quad -b_h(\psi, \lambda_t^o) + c_h(\rho, \psi) = \ell_2(\psi)$$

for all $(\boldsymbol{\mu}_t, \psi) \in (\mathbf{M}_h^o)_t \times \Psi_h$, and

$$(3.8c) \quad (\mathcal{P}_{\lambda_t^o} + \mathcal{P}_\rho + \mathcal{P}_f + \mathcal{P}_{\mathbf{g}_t}, 1)_\Omega = 0.$$

Here

$$\begin{aligned}
 a_h(\boldsymbol{\lambda}_t, \boldsymbol{\mu}_t) &:= (\mathbf{W}_{\boldsymbol{\lambda}_t}, \mathbf{W}_{\boldsymbol{\mu}_t})_{\Omega_h} + \langle \tau_t(\boldsymbol{\lambda}_t - \mathbf{u}_{\boldsymbol{\lambda}_t})_t, (\boldsymbol{\mu}_t - \mathbf{u}_{\boldsymbol{\mu}_t})_t \rangle_{\partial\Omega_h} + \left\langle \frac{1}{\tau_n} \mathcal{P}_{\boldsymbol{\lambda}_t}, \mathcal{P}_{\boldsymbol{\mu}_t} \right\rangle_{\partial\Omega_h}, \\
 b_h(\rho, \boldsymbol{\mu}_t) &:= \left\langle \rho, \mathbf{n} \cdot \mathbf{u}_{\boldsymbol{\mu}_t} + \frac{1}{\tau_n} \mathcal{P}_{\boldsymbol{\mu}_t} \right\rangle_{\partial\Omega_h}, \\
 c_h(\rho, \psi) &:= (\mathbf{W}_\rho, \mathbf{W}_\psi)_{\Omega_h} + \langle \tau_t(\mathbf{u}_\rho)_t, (\mathbf{u}_\psi)_t \rangle_{\partial\Omega_h} + \left\langle \frac{1}{\tau_n} (\rho - \mathcal{P}_\rho), (\psi - \mathcal{P}_\psi) \right\rangle_{\partial\Omega_h},
 \end{aligned}$$

and

$$\begin{aligned}
 \ell_1(\boldsymbol{\mu}_t) &:= (\mathbf{f}, \mathbf{u}_{\boldsymbol{\mu}_t})_{\Omega_h} - a_h(\mathbf{g}, \boldsymbol{\mu}_t), \\
 \ell_2(\psi) &:= -(\mathbf{f}, \mathbf{u}_\psi)_{\Omega_h} - \langle \mathbf{g} \cdot \mathbf{n}, \psi \rangle_{\partial\Omega} + b_h(\psi, \mathbf{g}_t).
 \end{aligned}$$

The proof of this theorem is in section 4. In view of this theorem, we can obtain the HDG solution by first solving a symmetric global system that is smaller than (2.2) and then locally recovering all solution components (by applying \mathcal{L}). This is the main advantage brought about by hybridization. It makes this HDG method competitive in comparison with other existing DG methods for Stokes flow.

It is interesting to note that the space in which the trace variables lie, namely $(\mathbf{M}_h)_t$ and Ψ_h , can be arbitrarily large. While it is in the interest of efficiency to choose as small a space as possible (for a given accuracy), in mixed methods one also often require spaces to be not too large for stability reasons. In the HDG method, stability is guaranteed through the penalty parameters τ_n and τ_t . A consequence of this is that (3.8) is uniquely solvable, no matter how large $(\mathbf{M}_h)_t$ and Ψ_h are. For the analogous hybridized mixed method of [8], we needed the trace spaces corresponding to $(\mathbf{M}_h)_t$ and Ψ_h to be exactly equal to certain spaces of jumps, which created additional implementation issues such as construction of local basis functions for the spaces.

3.2. Hybridization of Type II.

A formulation with velocity and means of pressure. Recalling our scheme for construction of hybridized methods described in the beginning of this section, we now consider the case when \mathbf{u}_t and \mathbf{u}_n (i.e., all components of \mathbf{u}) are chosen as the unknowns in the mesh interfaces. Correspondingly, we should use the transmission conditions on the other two variables, namely,

$$(3.9) \quad \left[[\boldsymbol{\omega} \times \mathbf{n}] \right]_{\mathcal{E}_h^\circ} = 0 \quad \text{and} \quad \left[[p \mathbf{n}] \right]_{\mathcal{E}_h^\circ} = 0,$$

to derive a hybridized formulation. However, the success of this strategy relies on us being able to compute approximate Stokes solutions within each element locally, once a discrete approximation to \mathbf{u} , say $\boldsymbol{\lambda}$, is obtained on the boundary of every mesh element. Here we find a difficulty not encountered in the previous case, namely, that the HDG discretization (2.2) on one element with $\boldsymbol{\lambda}$ as boundary data (of Type II) is not solvable in general, unless

$$(3.10) \quad \int_{\partial K} \boldsymbol{\lambda}_n \cdot \mathbf{n} = 0,$$

as seen from Proposition 2.1(2). Thus we are led to modify our local solvers, which in turn necessitates the introduction of a new variable (\bar{p}) approximating the means of pressures on the element boundaries, as we shall see now.

The new local solver, denoted by \mathcal{L}^{I} , maps a given function $(\boldsymbol{\lambda}, \bar{\rho}, \mathbf{f})$ in $\mathbf{L}^2(\partial\Omega_h) \times \ell^2(\partial\Omega_h) \times \mathbf{L}^2(\Omega)$ to a triple $(\mathbf{W}, \mathbf{u}, \mathcal{P}) \in \mathbf{W}_h \times \mathbf{V}_h \times P_h$ defined below. Here, $\ell^2(\partial\Omega_h)$ denotes the set of functions in $L^2(\partial\Omega_h)$ that are constant on each ∂K for all mesh elements K . On any element $K \in \Omega_h$, the function $(\mathbf{W}, \mathbf{u}, \mathcal{P})$ restricted to K is in $\mathbf{W}(K) \times \mathbf{V}(K) \times P(K)$ and satisfies

$$(3.11a) \quad (\mathbf{W}, \boldsymbol{\tau})_K - (\mathbf{u}, \mathbf{curl} \boldsymbol{\tau})_K = -\langle \boldsymbol{\lambda}, \mathbf{n} \times \boldsymbol{\tau} \rangle_{\partial K},$$

$$(3.11b) \quad (\mathbf{W}, \mathbf{curl} \mathbf{v})_K + \langle \widehat{\mathbf{W}}, \mathbf{v} \times \mathbf{n} \rangle_{\partial K}$$

$$(3.11c) \quad - (\mathcal{P}, \mathbf{div} \mathbf{v})_K + \langle \widehat{\mathcal{P}}, \mathbf{v} \cdot \mathbf{n} \rangle_{\partial K} = (\mathbf{f}, \mathbf{v})_K,$$

$$(3.11d) \quad - (\mathbf{u}, \mathbf{grad} q)_K = \langle \boldsymbol{\lambda} \cdot \mathbf{n}, q - \bar{q} \rangle_{\partial K},$$

$$(3.11e) \quad \bar{\mathcal{P}} = \bar{\rho},$$

where

$$(3.11e) \quad \widehat{\mathbf{W}} = \mathbf{W} + \tau_t (\mathbf{u} - \boldsymbol{\lambda}) \times \mathbf{n},$$

$$(3.11f) \quad \widehat{\mathcal{P}} = \mathcal{P} + \tau_n (\mathbf{u} - \boldsymbol{\lambda}) \cdot \mathbf{n}.$$

Here, we use the convention that for a given function q (that may not be in $\ell^2(\partial\Omega_h)$), we understand \bar{q} to mean the function in $\ell^2(\partial\Omega_h)$ satisfying

$$(3.12) \quad \bar{q}|_{\partial K} = \frac{1}{|\partial K|} \int_{\partial K} q \, d\gamma.$$

Obviously, for functions ρ in $\ell^2(\partial\Omega_h)$, we have $\bar{\rho} = \rho$. Let $\boldsymbol{\lambda}_n^0$ be the function on $\partial\Omega_h$ defined by $\boldsymbol{\lambda}_n^0|_{\partial K} = \boldsymbol{\lambda}_n|_{\partial K} - \overline{\boldsymbol{\lambda} \cdot \mathbf{n}}|_{\partial K} \mathbf{n}$ for all mesh elements K . Then, we can rewrite the right-hand side of (3.11c) as $\langle \boldsymbol{\lambda}_n^0, q\mathbf{n} \rangle_{\partial K}$. Hence, the system (3.11) minus (3.11d) is the same as the HDG system (2.2) applied to one element with the data $\mathbf{g}_t = \boldsymbol{\lambda}_t$ and $\mathbf{g}_n = \boldsymbol{\lambda}_n^0$. Consequently, by Proposition 2.1(2), the system has a solution, and moreover, the solution is unique once (3.11d) is added to the system. Thus, the map \mathcal{L}^{I} is well defined.

Note that (3.11) is the HDG discretization of the exact Stokes problem

$$\begin{aligned} \boldsymbol{\omega}_K - \mathbf{curl} \mathbf{u}_K &= 0 && \text{in } K, \\ \mathbf{curl} \boldsymbol{\omega}_K + \mathbf{grad} p_K &= \mathbf{f} && \text{in } K, \\ \mathbf{div} \mathbf{u}_K &= 0 && \text{in } K, \\ \mathbf{u}_K &= \boldsymbol{\lambda}_t + \boldsymbol{\lambda}_n^0 && \text{on } \partial K, \\ \bar{p}_K &= \bar{\rho} \end{aligned}$$

on a single element K .

Next, we find conditions on $(\boldsymbol{\lambda}, \bar{\rho}, \mathbf{f})$ that make $(\mathbf{W}, \mathbf{u}, \mathcal{P}) \equiv \mathcal{L}^{\text{I}}(\boldsymbol{\lambda}, \bar{\rho}, \mathbf{f})$ equal to $(\boldsymbol{\omega}_h, \mathbf{u}_h, p_h)$. First, we restrict $\boldsymbol{\lambda}$ to the space \mathbf{M}_h defined by

$$(3.13a) \quad \mathbf{M}_h := \{ \boldsymbol{\mu} \in \mathbf{L}^2(\mathcal{E}_h) : \boldsymbol{\mu}|_e \in \mathbf{M}(e) \quad \forall e \in \mathcal{E}_h^o \},$$

$$(3.13b) \quad \bar{\Psi}_h := \ell^2(\partial\Omega_h),$$

where $\mathbf{M}(e)$ is a finite-dimensional space on the face $e \in \mathcal{E}_h$ such that

$$(3.13c) \quad \mathbf{M}(e) \supseteq \{ (\mathbf{v} + \mathbf{n} \times \boldsymbol{\tau} + \mathbf{n} q)|_e : (\boldsymbol{\tau}, \mathbf{v}, q) \in \mathbf{W}(K) \times \mathbf{V}(K) \times P(K) \quad \forall K : e \subset \partial K \}.$$

Then we have the following theorem, which identifies certain discrete analogues of (3.9) as sufficient conditions for the coincidence of the locally recovered solution with the HDG solution.

THEOREM 3.3 (conditions for Type II hybridization). *Suppose $(\boldsymbol{\omega}_h, \mathbf{u}_h, p_h)$ is the solution of the HDG method defined by (2.2), (2.5), and (2.6). Assume that $(\boldsymbol{\lambda}, \bar{\rho}) \in \mathbf{M}_h \times \bar{\Psi}_h$ is such that*

$$\begin{aligned}
 (3.14a) \quad & \boldsymbol{\lambda} = \mathbf{g} \quad \text{on } \partial\Omega, \\
 (3.14b) \quad & \langle \llbracket \mathbf{n} \times \widehat{\mathbf{W}} \rrbracket, \boldsymbol{\mu}_t \rangle_{\mathcal{E}_h^o} = 0 \quad \forall \boldsymbol{\mu} \in \mathbf{M}_h, \\
 (3.14c) \quad & \langle \llbracket \widehat{\mathcal{P}} \mathbf{n} \rrbracket, \boldsymbol{\mu}_n \rangle_{\mathcal{E}_h^o} = 0 \quad \forall \boldsymbol{\mu} \in \mathbf{M}_h, \\
 (3.14d) \quad & \langle \boldsymbol{\lambda} \cdot \mathbf{n}, \bar{q} \rangle_{\partial\Omega_h} = 0 \quad \forall \bar{q} \in \bar{\Psi}_h, \\
 (3.14e) \quad & (\mathcal{P}, 1)_\Omega = 0.
 \end{aligned}$$

Then $(\mathbf{W}, \mathbf{u}, \mathcal{P}) = (\boldsymbol{\omega}_h, \mathbf{u}_h, p_h)$, $\boldsymbol{\lambda}_t = (\widehat{\mathbf{u}}_h)_t$, and $\boldsymbol{\lambda}_n = (\widehat{\mathbf{u}}_h)_n$.

Proof. We will show that $(\mathbf{W}, \mathbf{u}, \mathcal{P})$ and $(\boldsymbol{\omega}_h, \mathbf{u}_h, p_h)$ satisfy the same set of equations. To do this, just as in the proof of Theorem 3.1, it suffices to show that the numerical traces $(\widehat{\mathbf{W}})_t$, $\boldsymbol{\lambda}_t$, $\boldsymbol{\lambda}_n$, and $\widehat{\mathcal{P}}$ can be related to $(\mathbf{W}, \mathbf{u}, \mathcal{P})$ through the expressions in (2.5).

We therefore derive expressions for $(\widehat{\mathbf{W}})_t$, $\boldsymbol{\lambda}_t$, $\boldsymbol{\lambda}_n$, and $\widehat{\mathcal{P}}$. By the choice of the space \mathbf{M}_h , the jump conditions (3.14b) and (3.14c) imply that

$$\llbracket \mathbf{n} \times \widehat{\mathbf{W}} \rrbracket = 0 \quad \text{and} \quad \llbracket \widehat{\mathcal{P}} \mathbf{n} \rrbracket = \mathbf{0} \quad \text{on } \mathcal{E}_h^o.$$

Inserting the definition of the numerical traces (3.11e) and (3.11f), we readily obtain that, on \mathcal{E}_h^o ,

$$\begin{aligned}
 \llbracket \mathbf{n} \times \mathbf{W} \rrbracket + \tau_t^+ (\mathbf{u}^+)_t + \tau_t^- (\mathbf{u}^-)_t - (\tau_t^+ + \tau_t^-) \boldsymbol{\lambda}_t &= 0, \\
 \llbracket \mathcal{P} \mathbf{n} \rrbracket + \tau_n^+ (\mathbf{u}^+)_n + \tau_n^- (\mathbf{u}^-)_n - (\tau_n^+ + \tau_n^-) \boldsymbol{\lambda}_n &= 0,
 \end{aligned}$$

or equivalently,

$$\begin{aligned}
 \boldsymbol{\lambda}_t &= \left(\frac{\tau_t^+ (\mathbf{u}^+)_t + \tau_t^- (\mathbf{u}^-)_t}{\tau_t^- + \tau_t^+} \right) + \left(\frac{1}{\tau_t^- + \tau_t^+} \right) \llbracket \mathbf{n} \times \mathbf{W} \rrbracket, \\
 \boldsymbol{\lambda}_n &= \left(\frac{\tau_n^+ (\mathbf{u}^+)_n + \tau_n^- (\mathbf{u}^-)_n}{\tau_n^- + \tau_n^+} \right) + \left(\frac{1}{\tau_n^- + \tau_n^+} \right) \llbracket \mathcal{P} \mathbf{n} \rrbracket.
 \end{aligned}$$

Hence,

$$\begin{aligned}
 (\widehat{\mathbf{W}})_t &= \left(\frac{\tau_t^- (\mathbf{W}^+)_t + \tau_t^+ (\mathbf{W}^-)_t}{\tau_t^- + \tau_t^+} \right) + \left(\frac{\tau_t^+ \tau_t^-}{\tau_t^- + \tau_t^+} \right) \llbracket \mathbf{u} \times \mathbf{n} \rrbracket, \\
 \widehat{\mathcal{P}} &= \left(\frac{\tau_n^+ \mathcal{P}^+ + \tau_n^- \mathcal{P}^-}{\tau_n^- + \tau_n^+} \right) + \left(\frac{\tau_n^+ \tau_n^-}{\tau_n^- + \tau_n^+} \right) \llbracket \mathbf{u} \cdot \mathbf{n} \rrbracket.
 \end{aligned}$$

In other words, the numerical traces satisfy (2.5a), (2.5b), (2.5c), and (2.5d). The fact that they also satisfy (2.6) follows from conditions (3.14a) and (3.14c) and the definition of the local solvers.

Consequently, by Proposition 2.1(2), we conclude that the difference between $(\mathbf{W}, \mathbf{u}, \mathcal{P})$ and $(\boldsymbol{\omega}_h, \mathbf{u}_h, p_h)$ is $(\mathbf{0}, \mathbf{0}, \kappa)$ for some constant function κ . Equation (3.14e) then completes the proof. \square

Next, we show that the jump conditions (3.14b) and (3.14c) define a mixed method for the velocity traces and pressure averages on element boundaries. We denote by \mathbf{M}_h^o the set of functions in \mathbf{M}_h that vanish on $\partial\Omega$ and split $\boldsymbol{\lambda} = \boldsymbol{\lambda}^o + \mathbf{g}$ with $\boldsymbol{\lambda}^o$ in \mathbf{M}_h^o . In analogy with (3.7) of the Type I hybridization, we now define the specific local solutions for this case by

$$(3.15a) \quad (\mathcal{W}_\lambda, \mathbf{u}_\lambda, \mathcal{P}_\lambda) := \mathcal{L}^{\text{II}}(\boldsymbol{\lambda}, 0, \mathbf{0}),$$

$$(3.15b) \quad (\mathcal{W}_{\bar{p}}, \mathbf{u}_{\bar{p}}, \mathcal{P}_{\bar{p}}) := \mathcal{L}^{\text{II}}(\mathbf{0}, \bar{p}, \mathbf{0}),$$

$$(3.15c) \quad (\mathcal{W}_f, \mathbf{u}_f, \mathcal{P}_f) := \mathcal{L}^{\text{II}}(\mathbf{0}, 0, \mathbf{f}),$$

but note that by Proposition 2.1(2),

$$(3.16) \quad (\mathcal{W}_{\bar{p}}, \mathbf{u}_{\bar{p}}, \mathcal{P}_{\bar{p}}) = (\mathbf{0}, \mathbf{0}, \bar{p}).$$

Our main result for the Type II hybridization is the following theorem.

THEOREM 3.4 (characterization of the approximate solution). *We have that*

$$\begin{aligned} \boldsymbol{\omega}_h &= \mathcal{W}_{\boldsymbol{\lambda}^o} + \mathcal{W}_f + \mathcal{W}_g, \\ \mathbf{u}_h &= \mathbf{u}_{\boldsymbol{\lambda}^o} + \mathbf{u}_f + \mathbf{u}_g, \\ p_h &= \mathcal{P}_{\boldsymbol{\lambda}^o} + \mathcal{P}_f + \mathcal{P}_g + \mathcal{P}_{\bar{p}}, \end{aligned}$$

where $(\boldsymbol{\lambda}^o, \bar{p})$ is the only element of $\mathbf{M}_h^o \times \bar{\Psi}_h$ such that

$$\begin{aligned} a_h(\boldsymbol{\lambda}^o, \boldsymbol{\mu}) + b_h(\bar{p}, \boldsymbol{\mu}) &= \ell(\boldsymbol{\mu}), \\ -b_h(\bar{p}, \boldsymbol{\lambda}^o) &= 0 \end{aligned}$$

for all $(\boldsymbol{\mu}, \bar{\psi}) \in \mathbf{M}_h^o \times \bar{\Psi}_h$, and

$$(\mathcal{P}_{\boldsymbol{\lambda}^o} + \mathcal{P}_{\bar{p}} + \mathcal{P}_f + \mathcal{P}_g, 1)_\Omega = 0.$$

Here

$$\begin{aligned} a_h(\boldsymbol{\lambda}, \boldsymbol{\mu}) &= (\mathcal{W}_\lambda, \mathcal{W}_\mu)_{\Omega_h} + \langle \tau_t(\boldsymbol{\lambda} - \mathbf{u}_\lambda)_t, (\boldsymbol{\mu} - \mathbf{u}_\mu)_t \rangle_{\partial\Omega_h} \\ &\quad + \langle \tau_n(\boldsymbol{\lambda} - \mathbf{u}_\lambda)_n, (\boldsymbol{\mu} - \mathbf{u}_\mu)_n \rangle_{\partial\Omega_h}, \\ b_h(\bar{p}, \boldsymbol{\mu}) &= -\langle \bar{p}, \boldsymbol{\mu} \cdot \mathbf{n} \rangle_{\partial\Omega_h}, \\ \ell(\boldsymbol{\mu}) &= (\mathbf{f}, \mathbf{u}_\mu)_{\Omega_h} - a_h(\mathbf{g}, \boldsymbol{\mu}). \end{aligned}$$

A proof can be found in section 4. For appropriate choice of polynomial spaces, as in the previous case, to satisfy the conditions of Proposition 2.1, we choose the degrees d_P, d_V , and d_W to be integers obeying (2.12). Then \mathbf{M}_h is fixed once we pick any $\mathbf{M}(e)$ satisfying (3.13c), e.g., $\mathbf{M}(e) = \mathcal{P}_{\max(d_V, d_W, d_P)}(e)$.

3.3. Hybridization of Type III.

A formulation with tangential vorticity, normal velocity, and pressure means. Next we hybridize the HDG methods by making another choice of two variables in (3.2), namely $\boldsymbol{\omega}_t$ and \mathbf{u}_n , as the unknowns on the mesh interfaces. Their discrete approximations will be denoted by $\boldsymbol{\gamma}_t$ and $\boldsymbol{\lambda}_n$, respectively. When we try to formulate a system for these unknowns using the transmission conditions on the other two variables, namely,

$$(3.17) \quad \left[\mathbf{u} \times \mathbf{n} \right] \Big|_{\mathcal{E}_h^o} = 0 \quad \text{and} \quad \left[p \mathbf{n} \right] \Big|_{\mathcal{E}_h^o} = 0,$$

we again face the same difficulty we faced in the Type II case. Consequently, as we shall see, we must introduce a new variable \bar{p} approximating the averages of pressure on element boundaries, just as in the Type II case.

To hybridize the HDG method, we begin as in the previous cases by introducing discrete local solutions. These will be obtained using the HDG discretization of the Stokes problem

$$\begin{aligned} \boldsymbol{\omega}_K - \mathbf{curl} \mathbf{u}_K &= 0 && \text{in } K, \\ \mathbf{curl} \boldsymbol{\omega}_K + \mathbf{grad} p_K &= \mathbf{f} && \text{in } K, \\ \mathbf{div} \mathbf{u}_K &= 0 && \text{in } K, \\ (\boldsymbol{\omega}_K)_t &= \boldsymbol{\gamma}_t && \text{on } \partial K, \\ (\mathbf{u}_K)_n &= \boldsymbol{\lambda}_n^0 && \text{on } \partial K, \\ \bar{p}_K &= \bar{p} \end{aligned}$$

on a single element K . Given the function $(\boldsymbol{\gamma}_t, \boldsymbol{\lambda}_n, \bar{p}, \mathbf{f})$ in $\mathbf{L}^2(\partial\Omega_h) \times L^2(\partial\Omega_h) \times \ell^2(\partial\Omega_h) \times \mathbf{L}^2(\Omega)$, we define $(\mathbf{W}, \mathbf{u}, \mathcal{P})$ in $\mathbf{W}_h \times \mathbf{V}_h \times P_h$ on the element $K \in \Omega_h$ as the function in $\mathbf{W}(K) \times \mathbf{V}(K) \times P(K)$ such that

$$(3.18a) \quad (\mathbf{W}, \boldsymbol{\tau})_K - (\mathbf{u}, \mathbf{curl} \boldsymbol{\tau})_K + \langle \hat{\mathbf{u}}, \mathbf{n} \times \boldsymbol{\tau} \rangle_{\partial K} = 0,$$

$$(3.18b) \quad (\mathbf{W}, \mathbf{curl} \mathbf{v})_K - (\mathcal{P}, \mathbf{div} \mathbf{v})_K + \langle \hat{\mathcal{P}}, \mathbf{v} \cdot \mathbf{n} \rangle_{\partial K} = (\mathbf{f}, \mathbf{v})_K - \langle \boldsymbol{\gamma}_t, \mathbf{v} \times \mathbf{n} \rangle_{\partial K},$$

$$(3.18c) \quad -(\mathbf{u}, \mathbf{grad} q)_K = -\langle \boldsymbol{\lambda}_n \cdot \mathbf{n}, q - \bar{q} \rangle_{\partial K},$$

$$(3.18d) \quad \bar{\mathcal{P}} = \bar{p},$$

where

$$(3.18e) \quad \hat{\mathbf{u}} = \mathbf{u} + \frac{1}{\tau_t} \mathbf{n} \times (\mathbf{W} - \boldsymbol{\gamma}_t),$$

$$(3.18f) \quad \hat{\mathcal{P}} = \mathcal{P} + \tau_n (\mathbf{u} - \boldsymbol{\lambda}_n) \cdot \mathbf{n}.$$

By Proposition 2.1(3), there is a unique solution to (3.18) on each mesh element K . In other words, the local solver $\mathcal{L}^{\text{III}}(\boldsymbol{\gamma}_t, \boldsymbol{\lambda}_n, \bar{p}, \mathbf{f}) := (\mathbf{W}, \mathbf{u}, \mathcal{P})$ is well defined.

As in the previous cases, we now proceed to identify the discrete analogues of (3.17) that make $\mathcal{L}^{\text{III}}(\boldsymbol{\gamma}_t, \boldsymbol{\lambda}_n, \bar{p}, \mathbf{f})$ identical to $(\boldsymbol{\omega}_h, \mathbf{u}_h, p_h)$. This will yield a mixed method for $(\boldsymbol{\gamma}_t, \boldsymbol{\lambda}_n, \bar{p}, \mathbf{f})$. To do this, we begin by restricting the function $(\boldsymbol{\gamma}_t, \boldsymbol{\lambda}_n, \bar{p})$ to the space $(\mathbf{G}_h)_t \times (\mathbf{M}_h)_n \times \bar{\Psi}_h$, where

$$(3.19a) \quad (\mathbf{G}_h)_t := \{\boldsymbol{\delta}_t \in L^2(\mathcal{E}_h) : \boldsymbol{\delta}_t|_e \in \mathbf{G}(e) \quad \forall e \in \mathcal{E}_h\},$$

$$(3.19b) \quad (\mathbf{M}_h)_n := \{\boldsymbol{\mu}_n \in L^2(\mathcal{E}_h) : \boldsymbol{\mu}_n|_e \in \mathbf{M}(e) \quad \forall e \in \mathcal{E}_h^o\},$$

$$(3.19c) \quad \bar{\Psi}_h := \{\bar{\psi} \in L^2(\partial\Omega_h) : \bar{\psi}|_{\partial K} \in \mathbb{R} \quad \forall K \in \Omega_h\} \equiv \ell^2(\partial\Omega_h),$$

where $\mathbf{G}(e)$ and $\mathbf{M}(e)$ for each face $e \in \mathcal{E}_h$ are finite-dimensional spaces satisfying

$$(3.19d) \quad \mathbf{G}(e) \supseteq \{(\mathbf{v}_t + \mathbf{n} \times \boldsymbol{\tau})|_e : (\boldsymbol{\tau}, \mathbf{v}) \in \mathbf{W}(K) \times \mathbf{U}(K) \quad \forall K : e \subset \partial K\},$$

$$(3.19e) \quad \mathbf{M}(e) \supseteq \{(\mathbf{v}_n + \mathbf{n} q)|_e : (\mathbf{v}, q) \in \mathbf{U}(K) \times P(K) \quad \forall K : e \subset \partial K\}.$$

THEOREM 3.5 (conditions for Type III hybridization). *Suppose $(\boldsymbol{\omega}_h, \mathbf{u}_h, p_h)$ is the solution of the HDG method defined by (2.2), (2.5), and (2.6). Assume that*

$(\gamma_t, \lambda_n, \bar{\rho}) \in (\mathbf{G}_h)_t \times (\mathbf{M}_h)_n \times \bar{\Psi}_h$ is such that

$$(3.20a) \quad \lambda_n = \mathbf{g}_n \quad \text{on } \partial\Omega,$$

$$(3.20b) \quad \langle [\widehat{\mathbf{U}} \times \mathbf{n}], \delta_t \rangle_{\mathcal{E}_h} = \langle \mathbf{g}_t \times \mathbf{n}, \delta_t \rangle_{\partial\Omega} \quad \forall \delta_t \in (\mathbf{G}_h)_t,$$

$$(3.20c) \quad \langle [\widehat{\mathcal{P}} \mathbf{n}], \mu_n \rangle_{\mathcal{E}_h^\circ} = 0 \quad \forall \mu_n \in (\mathbf{M}_h)_n,$$

$$(3.20d) \quad \langle \lambda_n \cdot \mathbf{n}, \bar{q} \rangle_{\partial\Omega_h} = 0 \quad \forall \bar{q} \in \bar{\Psi}_h,$$

$$(3.20e) \quad (\mathcal{P}, 1)_\Omega = 0.$$

Then $(\mathbf{W}, \mathbf{U}, \mathcal{P}) = (\omega_h, \mathbf{u}_h, p_h)$, $\lambda_n = (\widehat{\mathbf{u}}_h)_n$, and $\gamma_t = (\widehat{\omega}_h)_t$.

Proof. We begin by noting that $(\mathbf{W}, \mathbf{U}, \mathcal{P})$ is in the space $\mathbf{W}_h \times \mathbf{V}_h \times P_h$. Moreover, $(\mathbf{W}, \mathbf{U}, \mathcal{P})$ satisfies the weak formulation (2.2) by the definition of the local solver (3.18).

Next, we note that, by the choice of the space $(\mathbf{G}_h)_t \times (\mathbf{M}_h)_n$, the jump conditions (3.20b) and (3.20c) imply that

$$[[\widehat{\mathbf{U}} \times \mathbf{n}]] = 0 \quad \text{and} \quad [[\widehat{\mathcal{P}} \mathbf{n}]] = 0 \quad \text{on } \mathcal{E}_h^\circ.$$

Inserting the definition of the numerical traces (3.18e) and (3.18f), we readily obtain that, on \mathcal{E}_h° ,

$$\begin{aligned} [[\mathbf{U} \times \mathbf{n}]] + \frac{1}{\tau_t^+} (\mathbf{W}^+)_t + \frac{1}{\tau_t^-} (\mathbf{W}^-)_t - \left(\frac{1}{\tau_t^+} + \frac{1}{\tau_t^-} \right) \gamma_t &= 0, \\ [[\mathcal{P} \mathbf{n}]] + \tau_n^+ (\mathbf{U}_h^+)_n + \tau_n^- (\mathbf{U}_h^-)_n - (\tau_n^+ + \tau_n^-) \lambda_n &= 0, \end{aligned}$$

or, equivalently,

$$\begin{aligned} \gamma_t &= \left(\frac{\tau_t^- (\mathbf{W}^+)_t + \tau_t^+ (\mathbf{W}^-)_t}{\tau_t^- + \tau_t^+} \right) + \left(\frac{\tau_t^- \tau_t^+}{\tau_t^- + \tau_t^+} \right) [[\mathbf{U} \times \mathbf{n}]], \\ \lambda_n &= \left(\frac{\tau_n^+ (\mathbf{U}^+)_n + \tau_n^- (\mathbf{U}^-)_n}{\tau_n^- + \tau_n^+} \right) + \left(\frac{1}{\tau_n^- + \tau_n^+} \right) [[\mathcal{P} \mathbf{n}]]. \end{aligned}$$

Hence,

$$\begin{aligned} (\widehat{\mathbf{u}})_t &= \left(\frac{\tau_t^+ (\mathbf{U}^+)_t + \tau_t^- (\mathbf{U}^-)_t}{\tau_t^- + \tau_t^+} \right) + \left(\frac{1}{\tau_t^- + \tau_t^+} \right) [[\mathbf{n} \times \mathbf{W}]], \\ \widehat{\mathcal{P}} &= \left(\frac{\tau_n^- \mathcal{P}^+ + \tau_n^+ \mathcal{P}^-}{\tau_n^- + \tau_n^+} \right) + \left(\frac{\tau_n^- \tau_n^+}{\tau_n^- + \tau_n^+} \right) [[\mathbf{U} \cdot \mathbf{n}]]. \end{aligned}$$

In other words, the numerical traces satisfy (2.5a), (2.5b), (2.5c), and (2.5d). The fact that they also satisfy (2.6) follows from conditions (3.20a) and (3.20c). They also satisfy (2.6c) and (2.6d) by definition of the local solvers.

By the uniqueness result of Proposition 2.1(2), we can now conclude that the approximation $(\mathbf{W}, \mathbf{U}, \mathcal{P})$ coincides with $(\omega_h, \mathbf{u}_h, p_h)$. Moreover, we also have $\gamma_t = (\widehat{\omega}_h)_t$ and $\lambda_n = (\widehat{\mathbf{u}}_h)_n$. This completes the proof. \square

We now proceed to formulate a mixed method for the numerical traces. Define specific local solutions by

$$\begin{aligned} (\mathbf{W}_{\gamma_t}, \mathbf{U}_{\gamma_t}, \mathcal{P}_{\gamma_t}) &:= \mathcal{L}^{\text{III}}(\gamma_t, \mathbf{0}, 0, \mathbf{0}), & (\mathbf{W}_{\lambda_n}, \mathbf{U}_{\lambda_n}, \mathcal{P}_{\lambda_n}) &:= \mathcal{L}^{\text{III}}(\mathbf{0}, \lambda_n, 0, \mathbf{0}), \\ (\mathbf{W}_{\bar{\rho}}, \mathbf{U}_{\bar{\rho}}, \mathcal{P}_{\bar{\rho}}) &:= \mathcal{L}^{\text{III}}(\mathbf{0}, \mathbf{0}, \bar{\rho}, \mathbf{0}), & (\mathbf{W}_{\mathbf{f}}, \mathbf{U}_{\mathbf{f}}, \mathcal{P}_{\mathbf{f}}) &:= \mathcal{L}^{\text{III}}(\mathbf{0}, \mathbf{0}, 0, \mathbf{f}), \end{aligned}$$

and observe that by Proposition 2.1(2), $(\mathbf{W}_{\bar{\rho}}, \mathbf{U}_{\bar{\rho}}, \mathcal{P}_{\bar{\rho}}) = (\mathbf{0}, \mathbf{0}, \bar{\rho})$. We additionally denote by $(\mathbf{M}_h^o)_n$ the functions of $(\mathbf{M}_h)_n$ which are zero on $\partial\Omega$, and we write $\boldsymbol{\lambda}_n$ as the sum of $\boldsymbol{\lambda}_n^o$ and \mathbf{g}_n , where $\boldsymbol{\lambda}_n^o$ is in $(\mathbf{M}_h^o)_n$. We are now ready to state our main result.

THEOREM 3.6 (characterization of the approximate solution). *We have that*

$$\begin{aligned} \boldsymbol{\omega}_h &= \mathbf{W}_{\boldsymbol{\gamma}_t} + \mathbf{W}_{\boldsymbol{\lambda}_n^o} + \mathbf{W}_{\mathbf{f}} + \mathbf{W}_{\mathbf{g}_n}, \\ \mathbf{u}_h &= \mathbf{U}_{\boldsymbol{\gamma}_t} + \mathbf{U}_{\boldsymbol{\lambda}_n^o} + \mathbf{U}_{\mathbf{f}} + \mathbf{U}_{\mathbf{g}_n}, \\ p_h &= \mathcal{P}_{\boldsymbol{\gamma}_t} + \mathcal{P}_{\boldsymbol{\lambda}_n^o} + \mathcal{P}_{\mathbf{f}} + \mathcal{P}_{\mathbf{g}_n} + \mathcal{P}_{\bar{\rho}}, \end{aligned}$$

where $(\boldsymbol{\gamma}_t, \boldsymbol{\lambda}_n^o, \bar{\rho})$ is the only element of $(\mathbf{G}_h)_t \times (\mathbf{M}_h^o)_n \times \bar{\Psi}_h$ such that

$$\begin{aligned} a_h(\boldsymbol{\gamma}_t, \boldsymbol{\delta}_t) + b_h(\boldsymbol{\lambda}_n, \boldsymbol{\delta}_t) &= \ell_1(\boldsymbol{\delta}_t), \\ -b_h(\boldsymbol{\mu}_n, \boldsymbol{\gamma}_t) + c_h(\boldsymbol{\lambda}_n, \boldsymbol{\mu}_n) + d_h(\bar{\rho}, \boldsymbol{\mu}_n) &= \ell_2(\boldsymbol{\mu}_n), \\ -d_h(\bar{\rho}, \boldsymbol{\lambda}_n) &= 0 \end{aligned}$$

for all $(\boldsymbol{\delta}_t, \boldsymbol{\mu}_n, \bar{\rho}) \in \mathbf{G}_h \times (\mathbf{M}_h^o)_n \times \bar{\Psi}_h$, and

$$(\mathcal{P}_{\boldsymbol{\lambda}_n^o} + \mathcal{P}_{\rho} + \mathcal{P}_{\mathbf{f}} + \mathcal{P}_{\mathbf{g}_t}, 1)_{\Omega} = 0.$$

Here

$$\begin{aligned} a_h(\boldsymbol{\gamma}_t, \boldsymbol{\delta}_t) &:= (\mathbf{W}_{\boldsymbol{\gamma}_t}, \mathbf{W}_{\boldsymbol{\delta}_t})_{\Omega_h} \\ &\quad + \left\langle \frac{1}{\tau_t} \mathbf{n} \times (\boldsymbol{\gamma}_t - \mathbf{W}_{\boldsymbol{\gamma}_t}), \mathbf{n} \times (\boldsymbol{\delta}_t - \mathbf{W}_{\boldsymbol{\delta}_t}) \right\rangle_{\partial\Omega_h} + \langle \tau_n (\mathbf{U}_{\boldsymbol{\gamma}_t})_n, (\mathbf{U}_{\boldsymbol{\delta}_t})_n \rangle_{\partial\Omega_h}, \\ b_h(\boldsymbol{\lambda}_n, \boldsymbol{\delta}_t) &:= \langle \boldsymbol{\lambda}_n, \mathcal{P}_{\boldsymbol{\delta}_t} + \tau_n (\mathbf{U}_{\boldsymbol{\lambda}_t})_n \rangle_{\partial\Omega_h}, \\ c_h(\boldsymbol{\lambda}_n, \boldsymbol{\mu}_n) &:= (\mathbf{W}_{\boldsymbol{\lambda}_n}, \mathbf{W}_{\boldsymbol{\mu}_n})_{\Omega_h} \\ &\quad + \left\langle \frac{1}{\tau_t} \mathbf{n} \times \mathbf{W}_{\boldsymbol{\mu}_n}, \mathbf{n} \times \mathbf{W}_{\boldsymbol{\lambda}_n} \right\rangle_{\partial\Omega_h} + \langle \tau_n (\boldsymbol{\mu}_n - \mathbf{U}_{\boldsymbol{\mu}_n})_n, (\boldsymbol{\lambda}_n - \mathbf{U}_{\boldsymbol{\lambda}_n})_n \rangle_{\partial\Omega_h}, \\ d_h(\bar{\rho}, \boldsymbol{\mu}_n) &:= -\langle \bar{\rho}, \boldsymbol{\mu}_n \cdot \mathbf{n} \rangle_{\partial\Omega_h}, \end{aligned}$$

and

$$\begin{aligned} \ell_1(\boldsymbol{\delta}_t) &:= -(\mathbf{f}, \mathbf{U}_{\boldsymbol{\delta}_t})_{\Omega_h} - b_h(\mathbf{g}_n, \boldsymbol{\delta}_t) - \langle \mathbf{g}_t \times \mathbf{n}, \boldsymbol{\delta}_t \rangle_{\partial\Omega}, \\ \ell_2(\boldsymbol{\mu}_n) &:= (\mathbf{f}, \mathbf{U}_{\boldsymbol{\mu}_n})_{\Omega_h} - c_h(\mathbf{g}_n, \boldsymbol{\mu}_n). \end{aligned}$$

3.4. Hybridization of Type IV.

A formulation with tangential vorticity, pressure, and harmonic velocity potentials. There is now only one more remaining choice of two variables from in (3.2), namely $\boldsymbol{\omega}_t$ and p , that we have not yet investigated. This is the Type IV case. This case presents additional complications not found in the previous three cases. The complications are rooted in the same reason for which we did not consider ‘‘Type IV boundary conditions’’ in section 2.

To explain the difficulty, suppose we are given an approximation $(\boldsymbol{\gamma}_t, \rho)$ to $(\boldsymbol{\omega}_t, p)$ on $\partial\Omega_h$. To obtain an approximate solution inside the mesh elements, let us try to define a local solution $(\mathbf{W}, \mathbf{U}, \mathcal{P})$ generated by data $(\boldsymbol{\gamma}_t, \rho, \mathbf{f})$ in $\mathbf{L}^2(\partial\Omega_h) \times L^2(\partial\Omega_h) \times L^2(\Omega)$. For this, we would like to use the HDG method applied to one element K , with

boundary conditions on tangential vorticity and pressure (which would be discrete versions of boundary conditions $\boldsymbol{\omega}_t = \boldsymbol{\gamma}_t$ and $p = \rho$ on ∂K). Thus we are led to take $(\mathbf{W}, \mathbf{u}, \mathcal{P})$ as the function in $\mathbf{W}(K) \times \mathbf{V}(K) \times P(K)$ such that

$$\begin{aligned} (\mathbf{W}, \boldsymbol{\tau})_K - (\mathbf{u}, \mathbf{curl} \boldsymbol{\tau})_K + \langle \widehat{\mathbf{u}}, \mathbf{n} \times \boldsymbol{\tau} \rangle_{\partial K} &= 0, \\ (\mathbf{W}, \mathbf{curl} \mathbf{v})_K - (\mathcal{P}, \operatorname{div} \mathbf{v})_K &= (\mathbf{f}, \mathbf{v})_K \\ &\quad - \langle \mathbf{n} \times \boldsymbol{\gamma}_t + \rho \mathbf{n}, \mathbf{v} \rangle_{\partial K}, \\ -(\mathbf{u}, \operatorname{grad} q)_K + \langle \widehat{\mathbf{u}} \cdot \mathbf{n}, q \rangle_{\partial K} &= 0, \end{aligned}$$

with $(\widehat{\mathbf{u}})_t = (\mathbf{u})_t + \tau_t^{-1} \mathbf{n} \times (\mathbf{W} - \boldsymbol{\gamma}_t)$ and $(\widehat{\mathbf{u}})_n = (\mathbf{u})_n + \tau_n^{-1} (\mathcal{P} - \rho) \mathbf{n}$. Unfortunately this problem is not solvable in general, which is the same reason we omitted this type of boundary condition in Proposition 2.1.

Nonetheless, upon reviewing the proof of Proposition 2.1 in the case of one element, we find that the null space of the above system is of the form $(\mathbf{W}, \mathbf{u}, \mathcal{P}) = (\mathbf{0}, \operatorname{grad} \phi, 0)$, where ϕ is in the following local space *harmonic velocity potentials*:

$$\Phi(K) = \{ \xi : \operatorname{grad} \xi \in \mathbf{V}(K) : \Delta \xi = 0 \text{ and } (\xi, 1)_K = 0 \}.$$

Hence we can recover unique solvability if the velocity is kept orthogonal to $\Phi(K)$. Keeping this in mind, we are motivated to reformulate the local problems to give a consistent system of equations as follows. Denote the L^2 -projection of $\mathbf{v} \in \mathbf{V}(K)$ into $\operatorname{grad} \Phi(K)$ by $\operatorname{grad} \phi_{\mathbf{v}}$. Given the function $(\boldsymbol{\gamma}_t, \rho, \phi, \mathbf{f})$ in $\mathbf{L}^2(\partial\Omega_h) \times L^2(\partial\Omega_h) \times H^1(\Omega_h) \times \mathbf{L}^2(\Omega)$, we define $(\mathbf{W}, \mathbf{u}, \mathcal{P})$ in $\mathbf{W}_h \times \mathbf{V}_h \times P_h$ on the element $K \in \Omega_h$ as the function in $\mathbf{W}(K) \times \mathbf{V}(K) \times P(K)$ such that

(3.21a)

$$(\mathbf{W}, \boldsymbol{\tau})_K - (\mathbf{u}, \mathbf{curl} \boldsymbol{\tau})_K + \langle \widehat{\mathbf{u}}, \mathbf{n} \times \boldsymbol{\tau} \rangle_{\partial K} = 0,$$

(3.21b)

$$\begin{aligned} (\mathbf{W}, \mathbf{curl} \mathbf{v})_K - (\mathcal{P}, \operatorname{div} \mathbf{v})_K &= (\mathbf{f}, \mathbf{v} - \operatorname{grad} \phi_{\mathbf{v}})_K \\ &\quad - \langle \mathbf{n} \times \boldsymbol{\gamma}_t + \rho \mathbf{n}, \mathbf{v} - \operatorname{grad} \phi_{\mathbf{v}} \rangle_{\partial K}, \end{aligned}$$

(3.21c)

$$-(\mathbf{u}, \operatorname{grad} q)_K + \langle \widehat{\mathbf{u}} \cdot \mathbf{n}, q \rangle_{\partial K} = 0,$$

(3.21d)

$$(\mathbf{u}, \operatorname{grad} \xi)_K = (\operatorname{grad} \phi, \operatorname{grad} \xi)_K,$$

where

(3.21e)

$$(\widehat{\mathbf{u}})_t = (\mathbf{u})_t + \frac{1}{\tau_t} \mathbf{n} \times (\mathbf{W} - \boldsymbol{\gamma}_t),$$

(3.21f)

$$(\widehat{\mathbf{u}})_n = (\mathbf{u})_n + \frac{1}{\tau_n} (\mathcal{P} - \rho) \mathbf{n}.$$

A minor modification of the arguments in Proposition 2.1 shows unique solvability of (3.21); hence we can define a fourth local solver $\mathcal{L}^{\text{IV}} : \mathbf{L}^2(\partial\Omega_h) \times L^2(\partial\Omega_h) \times H^1(\Omega_h) \times \mathbf{L}^2(\Omega) \mapsto \mathbf{W}(K) \times \mathbf{V}(K) \times P(K)$ that takes $(\boldsymbol{\gamma}_t, \rho, \phi, \mathbf{f})$ to $(\mathbf{W}, \mathbf{u}, \mathcal{P})$.

Note that (3.21) is a discretization of the exact Stokes problem

$$\begin{aligned} \boldsymbol{\omega}_K - \mathbf{curl} \mathbf{u}_K &= 0 && \text{in } K, \\ \mathbf{curl} \boldsymbol{\omega}_K + \operatorname{grad} p_K &= \mathbf{f} && \text{in } K, \\ \operatorname{div} \mathbf{u}_K &= 0 && \text{in } K, \\ (\boldsymbol{\omega}_K)_t &= \boldsymbol{\gamma}_t && \text{on } \partial K, \\ p_K &= \rho && \text{on } \partial K, \end{aligned}$$

with the additional condition that the velocity field \mathbf{u}_K is L^2 -orthogonal to all gradients of harmonic functions, which is necessary for well-posedness.

Although we could have considered a global “Type IV boundary conditions” case in Proposition 2.1 through the addition of an equation like (3.21d), it does not appear to be very useful, because we do not know the data needed for the right-hand side. However, we can use Type IV boundary conditions locally to hybridize a global problem with Type II boundary conditions because we already have global solvability for the Type II boundary conditions case. We need only ensure that the local problems are solvable, and the reformulation of the local solvers with (3.21d) guarantees it.

Now, we proceed as in the previous cases to identify conditions on γ_t, ρ , and ϕ in such a way that $(\mathbf{W}, \mathbf{U}, \mathcal{P})$ is identical to $(\omega_h, \mathbf{u}_h, p_h)$. We begin by restricting the function (γ_t, ρ, ϕ) to the space $(\mathbf{G}_h)_t \times \Psi_h \times \Phi_h$, where

$$(3.22a) \quad (\mathbf{G}_h)_t := \{\delta_t \in L^2(\mathcal{E}_h) : \delta_t|_e \in \mathbf{G}(e) \quad \forall e \in \mathcal{E}_h^o\},$$

$$(3.22b) \quad \Psi_h := \{\psi \in L^2(\mathcal{E}_h) : \psi|_e \in \Psi(e) \quad \forall e \in \mathcal{E}_h\},$$

$$(3.22c) \quad \Phi_h := \{\xi \in H^1(\Omega_h) : \xi|_K \in \Phi(K) \quad \forall K \in \Omega_h\},$$

where, on each face $e \in \mathcal{E}_h$, we have finite-dimensional spaces $\mathbf{G}(e)$ and $\Psi(e)$ satisfying

$$(3.22d) \quad \mathbf{G}(e) \supseteq \{(\mathbf{v}_t + \mathbf{n} \times \boldsymbol{\tau})|_e : (\boldsymbol{\tau}, \mathbf{v}) \in \mathbf{W}(K) \times \mathbf{U}(K) \quad \forall K : e \subset \partial K\},$$

$$(3.22e) \quad \Psi(e) \supseteq \{(q + \mathbf{v} \cdot \mathbf{n})|_e : (\mathbf{v}, q) \in \mathbf{U}(K) \times P(K) \quad \forall K : e \subset \partial K\}.$$

The next theorem identifies the discrete analogues of the transmission conditions

$$\llbracket \mathbf{u} \times \mathbf{n} \rrbracket \Big|_{\mathcal{E}_h^o} = 0, \quad \llbracket \mathbf{u} \cdot \mathbf{n} \rrbracket \Big|_{\mathcal{E}_h^o} = 0$$

that recover the original solution. An additional condition also appears because of our reformulation of the local solvers.

THEOREM 3.7 (conditions for Type IV hybridization). *Suppose $(\omega_h, \mathbf{u}_h, p_h)$ is the solution of the HDG method defined by (2.2), (2.5), and (2.6). Assume that $(\gamma_t, \rho, \phi) \in \mathbf{M}_h \times \Psi_h \times \Phi_h$ is such that*

$$(3.23a) \quad \langle \llbracket \widehat{\mathbf{u}} \times \mathbf{n} \rrbracket, \delta_t \rangle_{\mathcal{E}_h} = \langle \mathbf{g} \times \mathbf{n}, \delta_t \rangle_{\partial\Omega} \quad \forall \delta_t \in \mathbf{G}_h,$$

$$(3.23b) \quad \langle \llbracket \widehat{\mathbf{u}} \cdot \mathbf{n} \rrbracket, \psi \rangle_{\mathcal{E}_h} = \langle \mathbf{g} \cdot \mathbf{n}, \psi \rangle_{\partial\Omega} \quad \forall \psi \in \Psi_h,$$

$$(3.23c) \quad \langle \mathbf{n} \times \gamma_t + \rho \mathbf{n}, \text{grad } \xi \rangle_{\partial\Omega_h} = \langle \mathbf{f}, \text{grad } \xi \rangle_{\Omega_h} \quad \forall \xi \in \Phi_h,$$

$$(3.23d) \quad (\mathcal{P}, 1)_\Omega = 0.$$

Then $(\mathbf{W}, \mathbf{U}, \mathcal{P}) = (\omega_h, \mathbf{u}_h, p_h)$, $\gamma_t = (\widehat{\omega}_h)_t$, and $\rho = \widehat{p}_h$.

Proof. The proof is similar to the analogous proofs in the previous three cases and begins with the observation that $(\mathbf{W}, \mathbf{U}, \mathcal{P})$ satisfies the weak formulation (2.2) by the definition of the local solver (3.21) and condition (3.23c). Next, the jump conditions (3.23a) and (3.23b) imply that

$$\llbracket \widehat{\mathbf{u}} \times \mathbf{n} \rrbracket = 0 \quad \text{and} \quad \llbracket \widehat{\mathbf{u}} \cdot \mathbf{n} \rrbracket = 0 \quad \text{on } \mathcal{E}_h^o.$$

Inserting the definition of the numerical traces (3.21e) and (3.21f), we readily obtain that, on \mathcal{E}_h° ,

$$\begin{aligned} \llbracket \mathbf{u} \times \mathbf{n} \rrbracket + \frac{1}{\tau_t^+} (\mathbf{W}^+)_t + \frac{1}{\tau_t^-} (\mathbf{W}^-)_t - \left(\frac{1}{\tau_t^+} + \frac{1}{\tau_t^-} \right) \gamma_t &= 0, \\ \llbracket \mathbf{u} \cdot \mathbf{n} \rrbracket + \frac{1}{\tau_n^+} \mathcal{P}_h^+ + \frac{1}{\tau_n^-} \mathcal{P}_h^- - \left(\frac{1}{\tau_n^+} + \frac{1}{\tau_n^-} \right) \rho &= 0, \end{aligned}$$

or, equivalently,

$$\begin{aligned} \gamma_t &= \left(\frac{\tau_t^- (\mathbf{W}^+)_t + \tau_t^+ (\mathbf{W}^-)_t}{\tau_t^- + \tau_t^+} \right) + \left(\frac{\tau_t^- \tau_t^+}{\tau_t^- + \tau_t^+} \right) \llbracket \mathbf{u} \times \mathbf{n} \rrbracket, \\ \rho &= \left(\frac{\tau_n^- \mathcal{P}_h^+ + \tau_n^+ \mathcal{P}_h^-}{\tau_n^- + \tau_n^+} \right) + \left(\frac{\tau_n^- \tau_n^+}{\tau_n^- + \tau_n^+} \right) \llbracket \mathbf{u} \cdot \mathbf{n} \rrbracket. \end{aligned}$$

Hence,

$$\begin{aligned} (\widehat{\mathbf{u}})_t &= \left(\frac{\tau_t^+ (\mathbf{u}^+)_t + \tau_t^- (\mathbf{u}^-)_t}{\tau_t^- + \tau_t^+} \right) + \left(\frac{\tau_t^+ \tau_t^-}{\tau_t^- + \tau_t^+} \right) \llbracket \mathbf{n} \times \mathbf{W} \rrbracket, \\ (\widehat{\mathbf{u}})_n &= \left(\frac{\tau_n^+ (\mathbf{u}^+)_n + \tau_n^- (\mathbf{u}^-)_n}{\tau_n^- + \tau_n^+} \right) + \left(\frac{1}{\tau_n^- + \tau_n^+} \right) \llbracket \mathcal{P} \mathbf{n} \rrbracket. \end{aligned}$$

In other words, $(\mathbf{W}, \mathbf{u}, \mathcal{P})$ satisfies (2.2), (2.5), and (2.6). By the uniqueness result of Proposition 2.1(2), we can now conclude that the approximation $(\mathbf{W}, \mathbf{u}, \mathcal{P})$ coincides with $(\boldsymbol{\omega}_h, \mathbf{u}_h, p_h)$ and consequently $\gamma_t = (\widehat{\boldsymbol{\omega}}_h)_t$ and $\rho = \widehat{p}_h$. \square

Next, we give a characterization of the approximate solution in terms of the local solutions

$$\begin{aligned} (\mathbf{W}_{\gamma_t}, \mathbf{u}_{\gamma_t}, \mathcal{P}_{\gamma_t}) &:= \mathcal{L}^{\text{IV}}(\gamma_t, 0, 0, \mathbf{0}), & (\mathbf{W}_\rho, \mathbf{u}_\rho, \mathcal{P}_\rho) &:= \mathcal{L}^{\text{IV}}(\mathbf{0}, \rho, 0, \mathbf{0}), \\ (\mathbf{W}_\phi, \mathbf{u}_\phi, \mathcal{P}_\phi) &:= \mathcal{L}^{\text{IV}}(\mathbf{0}, 0, \phi, \mathbf{0}), & (\mathbf{W}_f, \mathbf{u}_f, \mathcal{P}_f) &:= \mathcal{L}^{\text{IV}}(\mathbf{0}, \mathbf{0}, 0, \mathbf{f}). \end{aligned}$$

Note that

$$(3.24) \quad (\mathbf{W}_\phi, \mathbf{u}_\phi, \mathcal{P}_\phi) = (\mathbf{0}, \text{grad } \phi, 0)$$

by direct verification in (3.21). The next theorem gives a mixed problem for the numerical traces γ_t, ρ together with the volumetric unknown ϕ . The presence of the variable ϕ defined within the elements (and not element boundaries, as in the previous cases) may appear to annul the potential advantages of dimensional reduction brought about by hybridization. However, this is not the case because ϕ is completely determined by its values on element boundaries.

THEOREM 3.8 (characterization of the approximate solution). *We have that*

$$\begin{aligned} \boldsymbol{\omega}_h &= \mathbf{W}_{\gamma_t} + \mathbf{W}_\rho + \mathbf{W}_f, \\ \mathbf{u}_h &= \mathbf{u}_{\gamma_t} + \mathbf{u}_\rho + \mathbf{u}_f + \text{grad } \phi, \\ p_h &= \mathcal{P}_{\lambda_t^\circ} + \mathcal{P}_\rho + \mathcal{P}_f, \end{aligned}$$

where (γ_t, ρ, ϕ) is the only element of $(\mathbf{G}_h)_t \times \Psi_h \times \Phi_h$ such that

$$\begin{aligned} a_h(\gamma_t, \boldsymbol{\delta}_t) + b_h(\rho, \boldsymbol{\delta}_t) + c_h(\phi, \boldsymbol{\delta}_t) &= \ell_1(\boldsymbol{\mu}_t), \\ b_h(\psi, \gamma_t) + d_h(\rho, \psi) + e_h(\phi, \psi) &= \ell_2(\psi), \\ -c_h(\xi, \gamma_t) - e_h(\xi, \rho) &= \ell_3(\xi) \end{aligned}$$

for all $(\boldsymbol{\delta}_t, \psi, \xi) \in (\mathbf{G}_h)_t \times \Psi_h \times \Phi_h$, and

$$(\mathcal{P}_{\boldsymbol{\gamma}_t} + \mathcal{P}_\rho + \mathcal{P}_{\mathbf{f}}, 1)_\Omega = 0.$$

Here

$$\begin{aligned} a_h(\boldsymbol{\gamma}_t, \boldsymbol{\delta}_t) &:= (\mathcal{W}_{\boldsymbol{\gamma}_t}, \mathcal{W}_{\boldsymbol{\delta}_t})_{\Omega_h} + \left\langle \frac{1}{\tau_n} \mathcal{P}_{\boldsymbol{\gamma}_t}, \mathcal{P}_{\boldsymbol{\delta}_t} \right\rangle_{\partial\Omega_h} \\ &\quad + \left\langle \frac{1}{\tau_t} \mathbf{n} \times (\boldsymbol{\gamma}_t - \mathcal{W}_{\boldsymbol{\gamma}_t}), \mathbf{n} \times (\boldsymbol{\delta}_t - \mathcal{W}_{\boldsymbol{\delta}_t}) \right\rangle_{\partial\Omega_h}, \\ b_h(\rho, \boldsymbol{\delta}_t) &:= - \left\langle \mathbf{u}_{\boldsymbol{\delta}_t} + \frac{1}{\tau_n} \mathcal{P}_{\boldsymbol{\delta}_t}, \rho \right\rangle_{\partial\Omega_h}, \\ c_h(\phi, \boldsymbol{\delta}) &:= \langle \mathbf{n} \times \text{grad } \phi, \boldsymbol{\delta}_t \rangle_{\partial\Omega_h}, \\ d_h(\rho, \psi) &:= (\mathcal{W}_\rho, \mathcal{W}_\psi)_{\Omega_h} + \left\langle \frac{1}{\tau_n} (\rho - \mathcal{P}_\rho), (\psi - \mathcal{P}_\psi) \right\rangle_{\partial\Omega_h} \\ &\quad + \left\langle \frac{1}{\tau_t} \mathbf{n} \times \mathcal{W}_\rho, \mathbf{n} \times \mathcal{W}_\psi \right\rangle_{\partial\Omega_h}, \\ e_h(\phi, \psi) &:= - \langle \text{grad } \phi \cdot \mathbf{n}, \psi \rangle_{\partial\Omega_h}, \end{aligned}$$

and

$$\begin{aligned} \ell_1(\boldsymbol{\delta}_t) &:= -(\mathbf{f}, \mathbf{u}_{\boldsymbol{\delta}_t})_{\Omega_h} - \langle \mathbf{g} \times \mathbf{n}, \boldsymbol{\delta}_t \rangle_{\partial\Omega}, \\ \ell_2(\psi) &:= -(\mathbf{f}, \mathbf{u}_\psi)_{\Omega_h} - \langle \mathbf{g} \cdot \mathbf{n}, \psi \rangle_{\partial\Omega}, \\ \ell_3(\psi) &:= +(\mathbf{f}, \text{grad } \xi)_{\Omega_h}. \end{aligned}$$

3.5. Summary. We have shown how to hybridize the HDG methods in four different ways according to the choice of globally coupled variables. These variables are described in Table 3.1 for each of the hybridizations we considered. They are referred to as *unknowns* therein since all the other variables can be eliminated from the original equations. The corresponding discrete transmission conditions appear alongside under the heading *jump conditions*. The primary motivation for all these hybridizations is the reduction in the number of global degrees of freedom achieved by the elimination of volumetric unknowns $\boldsymbol{\omega}_h$, \mathbf{u}_h , and p_h . The variational equations on the mesh faces that we derived in each type result in significantly smaller systems, especially in the high order case.

TABLE 3.1
The unknowns and jump conditions for the hybridizations.

Type	Unknowns			Jump conditions	
I	$(\hat{\mathbf{u}}_h)_t$	\hat{p}_h		$[\![\mathbf{n} \times (\hat{\boldsymbol{\omega}}_h)_t]\!] = \mathbf{0}$	$[\![\hat{\mathbf{u}}_h]_n \cdot \mathbf{n}]\!] = 0$
II	$(\hat{\mathbf{u}}_h)_t$	$(\hat{\mathbf{u}}_h)_n$	\bar{p}_h	$[\![\mathbf{n} \times (\hat{\boldsymbol{\omega}}_h)_t]\!] = \mathbf{0}$	$[\![\hat{p}_h \mathbf{n}]\!] = \mathbf{0}$
III	$(\hat{\boldsymbol{\omega}}_h)_t$	$(\hat{\mathbf{u}}_h)_n$	\bar{p}_h	$[\![\hat{\mathbf{u}}_h]_t \times \mathbf{n}]\!] = \mathbf{0}$	$[\![\hat{p}_h \mathbf{n}]\!] = \mathbf{0}$
IV	$(\hat{\boldsymbol{\omega}}_h)_t$	\hat{p}_h	ϕ_h	$[\![\hat{\mathbf{u}}_h]_t \times \mathbf{n}]\!] = \mathbf{0}$	$[\![\hat{\mathbf{u}}_h]_n \cdot \mathbf{n}]\!] = 0$

For DG methods, the possibility of deriving a hybridized formulation is strongly dependent on the structure of the numerical traces. Although we gave expressions for

the numerical traces in the traditional DG format as in (2.5), we should note that the numerical traces on which the jump conditions are imposed can be expressed element by element. Indeed, on the boundary of *each* mesh element K , the numerical traces on which the jump conditions are imposed have the following expressions using the values of variables from just that element:

$$(3.25) \quad \text{Type I:} \quad \begin{cases} (\widehat{\boldsymbol{\omega}}_h)_t = (\boldsymbol{\omega}_h)_t + \tau_t (\mathbf{u}_h - (\widehat{\mathbf{u}}_h)_t) \times \mathbf{n} & \text{on } \partial K \\ (\widehat{\mathbf{u}}_h)_n = (\mathbf{u}_h)_n + \frac{1}{\tau_n} (p_h - \widehat{p}_h) \mathbf{n} & \text{on } \partial K, \end{cases}$$

$$(3.26) \quad \text{Type II:} \quad \begin{cases} (\widehat{\boldsymbol{\omega}}_h)_t = (\boldsymbol{\omega}_h)_t + \tau_t (\mathbf{u}_h - (\widehat{\mathbf{u}}_h)_t) \times \mathbf{n} & \text{on } \partial K, \\ \widehat{p}_h = p_h + \tau_n (\mathbf{u}_h - (\widehat{\mathbf{u}}_h)_n) \cdot \mathbf{n} & \text{on } \partial K, \end{cases}$$

$$(3.27) \quad \text{Type III:} \quad \begin{cases} (\widehat{\mathbf{u}}_h)_t = (\mathbf{u}_h)_t + \frac{1}{\tau_t} \mathbf{n} \times (\boldsymbol{\omega}_h - (\widehat{\boldsymbol{\omega}}_h)_t) & \text{on } \partial K, \\ \widehat{p}_h = p_h + \tau_n (\mathbf{u}_h - (\widehat{\mathbf{u}}_h)_n) \cdot \mathbf{n}, & \text{on } \partial K, \end{cases}$$

$$(3.28) \quad \text{Type IV:} \quad \begin{cases} (\widehat{\mathbf{u}}_h)_t = (\mathbf{u}_h)_t + \frac{1}{\tau_t} \mathbf{n} \times (\boldsymbol{\omega}_h - (\widehat{\boldsymbol{\omega}}_h)_t) & \text{on } \partial K, \\ (\widehat{\mathbf{u}}_h)_n = (\mathbf{u}_h)_n + \frac{1}{\tau_n} (p_h - (\widehat{p}_h)_t) \mathbf{n} & \text{on } \partial K. \end{cases}$$

Finally, let us note that in the rewritten expressions of the numerical traces above, it is easy to formally set the parameters τ_t, τ_n to either zero or infinity, which gives rise to numerical methods we can think of as being limiting cases of the HDG methods. In Table 3.2, for each of these limiting cases, we give the associated continuity properties of some of the components of the approximate solution as well as the corresponding natural hybridizations.

TABLE 3.2
The continuity properties induced by the formal limits.

Formal limit	Continuity property	Hybridization type
$\tau_t = 0$	$\boldsymbol{\omega}_h \in H(\mathbf{curl}, \Omega)$	I, II
$\frac{1}{\tau_t} = 0$	$\mathbf{u}_h \in H(\mathbf{curl}, \Omega)$	III, IV
$\tau_n = 0$	$p_h \in C^0(\Omega)$	II, III
$\frac{1}{\tau_n} = 0$	$\mathbf{u}_h \in H(\text{div}, \Omega)$	I, IV

In particular, if we use the hybridizations of Type I or IV and formally set $\tau_n = \infty$ in (3.25) or (3.28), we immediately obtain that $\mathbf{u}_h \in H(\text{div}, \Omega)$ by the jump condition (3.6c) (respectively, jump condition (3.23b)) for the Type I (respectively, Type IV) boundary conditions. We also immediately see that the discrete incompressibility condition (2.2c) becomes

$$(\text{div } \mathbf{u}_h, q)_{\Omega_h} = 0 \quad \forall q \in P_h,$$

and if we assume, as in Proposition 2.1, that

$$\text{div } \mathbf{V}(K) \subset P(K) \quad \forall K \in \Omega_h,$$

we can conclude that our approximate velocity \mathbf{u}_h is strongly incompressible. That is, the distributional divergence of the numerical velocity approximation satisfies

$\operatorname{div} \mathbf{u}_h = 0$ in all Ω . It is interesting that even though the space \mathbf{V}_h is a space of completely discontinuous functions, we are able to recover such a velocity approximation. The first DG methods producing strongly incompressible velocities were introduced, in the framework of the Navier–Stokes equations, in [12] and were later more explicitly developed in [13]; see also [21], where this idea is applied to square and cube elements. Another DG method able to provide strongly incompressible velocities is the method introduced in [3]. It uses a velocity space \mathbf{V}_h of exactly divergence-free velocities and uses a hybridization technique to avoid the almost impossible task of constructing its bases.

Unfortunately, the above-mentioned methods do not fit into our setting. The methods in [12, 13] do not use the vorticity as an unknown; instead, they use the gradient of the velocity. The method in [3] almost fits into our setting except for the fact that the numerical traces for the tangential vorticity and the tangential velocity do not coincide for any finite values of τ_t^\pm . If, on the other hand, we formally set $\tau_t^- = \infty$ and then take $\tau_t^+ = 0$, we do recover the general form of the numerical traces considered in [3]. However, in that case, the numerical trace for the tangential vorticity becomes independent of the tangential velocity. This is certainly not the case for the scheme treated in [3].

In Table 3.3, we describe four special limiting cases. Most finite element methods for the Stokes problem use approximate velocities \mathbf{u}_h in $\mathbf{H}^1(\Omega)$ (see [2]); they thus correspond to the case $\frac{1}{\tau_t} = \frac{1}{\tau_n} = 0$. The method introduced by Nédélec in [17] corresponds to the case $\tau_t = \frac{1}{\tau_n} = 0$; its hybridization was carried out in [7, 8].

TABLE 3.3
Four special formal limits of HDG methods.

	$\tau_t = 0$	$\frac{1}{\tau_t} = 0$
$\tau_n = 0$	$\omega_h \in H(\mathbf{curl} \Omega)$ $p_h \in C^0(\Omega)$ Type II hybridization	$\mathbf{u}_h \in H(\mathbf{curl} \Omega)$ $p_h \in C^0(\Omega)$ Type III hybridization
$\frac{1}{\tau_n} = 0$	$\omega_h \in H(\mathbf{curl} \Omega)$ $\mathbf{u}_h \in H(\operatorname{div}, \Omega)$ Type I hybridization	$\mathbf{u}_h \in H(\mathbf{curl} \Omega)$ $\mathbf{u}_h \in H(\operatorname{div}, \Omega)$ Type IV hybridization

4. Proofs of the characterization theorems.

4.1. Preliminaries. We begin by proving an auxiliary identity that we will use in all our proofs. It is stated in terms of functions $(\mathbf{w}_h, \mathbf{u}_h, p_h)$ in $\mathbf{W}_h \times \mathbf{V}_h \times P_h$ that are assumed to satisfy the equations

$$(4.1a) \quad (\mathbf{w}_h, \boldsymbol{\tau})_{\Omega_h} - (\mathbf{u}_h, \mathbf{curl} \boldsymbol{\tau})_{\Omega_h} = -\langle \widehat{\mathbf{u}}_h, \mathbf{n} \times \boldsymbol{\tau} \rangle_{\partial\Omega_h},$$

$$(4.1b) \quad (\mathbf{w}_h, \mathbf{curl} \mathbf{v})_{\Omega_h} - (p_h, \operatorname{div} \mathbf{v})_{\Omega_h} = (\mathbf{f}, \mathbf{v} - \mathbf{P}\mathbf{v})_{\Omega_h} - \langle \widehat{\mathbf{p}}_h, (\mathbf{v} - \mathbf{P}\mathbf{v}) \cdot \mathbf{n} \rangle_{\partial\Omega_h} \\ - \langle \widehat{\mathbf{w}}_h, (\mathbf{v} - \mathbf{P}\mathbf{v}) \times \mathbf{n} \rangle_{\partial\Omega_h},$$

$$(4.1c) \quad -(\mathbf{u}_h, \operatorname{grad} q)_{\Omega_h} = -\langle \widehat{\mathbf{u}}_h \cdot \mathbf{n}, q - \mathbf{P}q \rangle_{\partial\Omega_h}$$

for all $(\boldsymbol{\tau}, \mathbf{v}, q) \in \mathbf{W}_h \times \mathbf{V}_h \times P_h$. Here \mathbf{P} is a projection from P_h , and \mathbf{P} is a projection from \mathbf{V}_h . Their ranges are denoted by $\overline{\psi}_h$ and \mathbf{H}_h , respectively. The symbols $\widehat{\mathbf{w}}_h, \widehat{\mathbf{u}}_h$, and $\widehat{\mathbf{p}}_h$, while evocative of numerical traces, are *not* assumed to be related to the

variables $(\mathbf{w}_h, \mathbf{u}_h, \mathbf{p}_h)$ as in (2.5), nor are they assumed to be single valued on \mathcal{E}_h . They simply denote some given functions on $\partial\Omega_h$.

LEMMA 4.1. *Let $(\mathbf{w}_h, \mathbf{u}_h, \mathbf{p}_h)$ be a function satisfying (4.1a) and (4.1c), and let $(\mathbf{w}'_h, \mathbf{u}'_h, \mathbf{p}'_h)$ be a function satisfying (4.1b) with \mathbf{f} , $\widehat{\mathbf{w}}_h$, and $\widehat{\mathbf{p}}_h$ replaced by \mathbf{f}' , $\widehat{\mathbf{w}}'_h$, and $\widehat{\mathbf{p}}'_h$, respectively. Then*

$$\begin{aligned} -\langle \widehat{\mathbf{u}}_h, \mathbf{n} \times \widehat{\mathbf{w}}'_h + \mathbf{n} \widehat{\mathbf{p}}'_h \rangle_{\partial\Omega_h} &= (\mathbf{w}_h, \mathbf{w}'_h)_{\Omega_h} \\ &\quad - \langle \widehat{\mathbf{u}}_h - \mathbf{u}_h, \mathbf{n} \times (\widehat{\mathbf{w}}'_h - \mathbf{w}'_h) + \mathbf{n} (\widehat{\mathbf{p}}'_h - \mathbf{p}'_h) \rangle_{\partial\Omega_h} \\ &\quad - (\mathbf{u}_h, \mathbf{f}')_{\Omega_h} \end{aligned}$$

whenever $(\mathbf{P}\mathbf{u}_h, \mathbf{P}\mathbf{p}'_h) = (\mathbf{0}, 0)$.

Proof. By (4.1a) with $\boldsymbol{\tau} := \mathbf{w}'_h$, we have that

$$(\mathbf{w}_h, \mathbf{w}'_h)_{\Omega_h} = (\mathbf{u}_h, \mathbf{curl} \mathbf{w}'_h)_{\Omega_h} - \langle \widehat{\mathbf{u}}_h, \mathbf{n} \times \mathbf{w}'_h \rangle_{\partial\Omega_h}$$

and so, after integration by parts,

$$(\mathbf{w}_h, \mathbf{w}'_h)_{\Omega_h} = (\mathbf{curl} \mathbf{u}_h, \mathbf{w}'_h)_{\Omega_h} + \langle \mathbf{u}_h - \widehat{\mathbf{u}}_h, \mathbf{n} \times \mathbf{w}'_h \rangle_{\partial\Omega_h}.$$

By (4.1b) with $\mathbf{v} := \mathbf{u}_h$, and with $\mathbf{w}_h, \mathbf{u}_h, \mathbf{p}_h$, and \mathbf{f} replaced by $\mathbf{w}'_h, \mathbf{u}'_h, \mathbf{p}'_h$, and \mathbf{f}' , respectively, we get

$$\begin{aligned} (\mathbf{w}_h, \mathbf{w}'_h)_{\Omega_h} &= -\langle \widehat{\mathbf{w}}'_h, (\mathbf{u}_h - \mathbf{P}\mathbf{u}_h) \times \mathbf{n} \rangle_{\partial\Omega_h} + \langle \mathbf{u}_h - \widehat{\mathbf{u}}_h, \mathbf{n} \times \mathbf{w}'_h \rangle_{\partial\Omega_h} \\ &\quad + (\mathbf{p}'_h, \mathbf{div} \mathbf{u}_h)_{\Omega_h} - \langle \widehat{\mathbf{p}}'_h, (\mathbf{u}_h - \mathbf{P}\mathbf{u}_h) \cdot \mathbf{n} \rangle_{\partial\Omega_h} + (\mathbf{f}', \mathbf{u}_h - \mathbf{P}\mathbf{u}_h)_{\Omega_h} \\ &= -\langle \mathbf{u}_h, \mathbf{n} \times \widehat{\mathbf{w}}'_h \rangle_{\partial\Omega_h} + \langle \mathbf{u}_h - \widehat{\mathbf{u}}_h, \mathbf{n} \times \mathbf{w}'_h \rangle_{\partial\Omega_h} \\ &\quad + (\mathbf{div} \mathbf{u}_h, \mathbf{p}'_h)_{\Omega_h} - \langle \mathbf{u}_h \cdot \mathbf{n}, \widehat{\mathbf{p}}'_h \rangle_{\partial\Omega_h} + (\mathbf{u}_h, \mathbf{f}')_{\Omega_h} \end{aligned}$$

since $\mathbf{P}\mathbf{u}_h = \mathbf{0}$. If we now integrate by parts, we get

$$\begin{aligned} (\mathbf{w}_h, \mathbf{w}'_h)_{\Omega_h} &= -\langle \mathbf{u}_h, \mathbf{n} \times \widehat{\mathbf{w}}'_h \rangle_{\partial\Omega_h} + \langle \mathbf{u}_h - \widehat{\mathbf{u}}_h, \mathbf{n} \times \mathbf{w}'_h \rangle_{\partial\Omega_h} \\ &\quad - (\mathbf{u}_h, \mathbf{grad} \mathbf{p}'_h)_{\Omega_h} - \langle \mathbf{u}_h \cdot \mathbf{n}, \widehat{\mathbf{p}}'_h - \mathbf{p}'_h \rangle_{\partial\Omega_h} + (\mathbf{u}_h, \mathbf{f}')_{\Omega_h}, \end{aligned}$$

and by (4.1c) with $q := \mathbf{p}'_h$,

$$\begin{aligned} (\mathbf{w}_h, \mathbf{w}'_h)_{\Omega_h} &= -\langle \mathbf{u}_h, \mathbf{n} \times \widehat{\mathbf{w}}'_h \rangle_{\partial\Omega_h} + \langle \mathbf{u}_h - \widehat{\mathbf{u}}_h, \mathbf{n} \times \mathbf{w}'_h \rangle_{\partial\Omega_h} \\ &\quad - \langle \widehat{\mathbf{u}}_h \cdot \mathbf{n}, \mathbf{p}'_h - \mathbf{P}\mathbf{p}'_h \rangle_{\partial\Omega_h} - \langle \mathbf{u}_h \cdot \mathbf{n}, \widehat{\mathbf{p}}'_h - \mathbf{p}'_h \rangle_{\partial\Omega_h} + (\mathbf{u}_h, \mathbf{f}')_{\Omega_h} \\ &= -\langle \widehat{\mathbf{u}}_h, \mathbf{n} \times \widehat{\mathbf{w}}'_h \rangle_{\partial\Omega_h} + \langle \mathbf{u}_h - \widehat{\mathbf{u}}_h, \mathbf{n} \times (\mathbf{w}'_h - \widehat{\mathbf{w}}'_h) \rangle_{\partial\Omega_h} \\ &\quad - \langle \widehat{\mathbf{u}}_h \cdot \mathbf{n}, \widehat{\mathbf{p}}'_h \rangle_{\partial\Omega_h} + \langle (\widehat{\mathbf{u}}_h - \mathbf{u}_h) \cdot \mathbf{n}, \widehat{\mathbf{p}}'_h - \mathbf{p}'_h \rangle_{\partial\Omega_h} + (\mathbf{u}_h, \mathbf{f}')_{\Omega_h} \end{aligned}$$

since $\mathbf{P}\mathbf{p}'_h = 0$. The result now follows after a simple rearrangement of terms. This completes the proof. \square

The following immediate consequence of this result will also be useful.

COROLLARY 4.2. *Let $(\mathbf{w}_h, \mathbf{u}_h, \mathbf{p}_h)$ be a function satisfying (4.1), and let $(\mathbf{w}'_h, \mathbf{u}'_h, \mathbf{p}'_h)$ be a function satisfying (4.1) with \mathbf{f} , $\widehat{\mathbf{w}}_h$, $\widehat{\mathbf{u}}_h$, and $\widehat{\mathbf{p}}_h$ replaced by \mathbf{f}' , $\widehat{\mathbf{w}}'_h$, $\widehat{\mathbf{u}}'_h$, and $\widehat{\mathbf{p}}'_h$, respectively. Then we have*

$$-\langle \widehat{\mathbf{u}}_h, \mathbf{n} \times \widehat{\mathbf{w}}'_h + \mathbf{n} \widehat{\mathbf{p}}'_h \rangle_{\partial\Omega_h} + (\mathbf{u}_h, \mathbf{f}')_{\Omega_h} = -\langle \widehat{\mathbf{u}}'_h, \mathbf{n} \times \widehat{\mathbf{w}}_h + \mathbf{n} \widehat{\mathbf{p}}_h \rangle_{\partial\Omega_h} + (\mathbf{u}'_h, \mathbf{f})_{\Omega_h},$$

provided $(\mathbf{P}\mathbf{u}_h, \mathbf{P}\mathbf{p}_h) = (\mathbf{P}\mathbf{u}'_h, \mathbf{P}\mathbf{p}'_h) = (\mathbf{0}, 0)$ and

$$-\langle \widehat{\mathbf{u}}_h - \mathbf{u}_h, \mathbf{n} \times (\widehat{\mathbf{w}}'_h - \mathbf{w}'_h) + \mathbf{n} (\widehat{\mathbf{p}}'_h - \mathbf{p}'_h) \rangle_{\partial\Omega_h} = -\langle \widehat{\mathbf{u}}'_h - \mathbf{u}'_h, \mathbf{n} \times (\widehat{\mathbf{w}}_h - \mathbf{w}_h) + \mathbf{n} (\widehat{\mathbf{p}}_h - \mathbf{p}_h) \rangle_{\partial\Omega_h}.$$

4.2. Proof of the characterization of Theorem 3.2. To prove the characterization of Theorem 3.2, we are going to use several key identities gathered in the following result. Recall the definitions of specific local solutions in (3.7) (such as $\mathcal{W}_{\lambda_t}, \mathbf{u}_{\lambda_t}$, etc.). We denote by $\widehat{\mathcal{W}}_{\odot}$ and $\widehat{\mathbf{u}}_{\odot}$ the corresponding numerical traces, for all choices of the subscript “ \odot ” that make sense in the discussion of this hybridization case:

$$(4.2a) \quad \widehat{\mathcal{W}}_{\lambda_t} = \mathcal{W}_{\lambda_t} + \tau_t (\mathbf{u}_{\lambda_t} - \lambda_t) \times \mathbf{n}, \quad \widehat{\mathbf{u}}_{\lambda_t} = \mathbf{u}_{\lambda_t} + \frac{1}{\tau_n} \mathcal{P}_{\lambda_t} \mathbf{n},$$

$$(4.2b) \quad \widehat{\mathcal{W}}_{\rho} = \mathcal{W}_{\rho} + \tau_t (\mathbf{u}_{\rho} \times \mathbf{n}), \quad \widehat{\mathbf{u}}_{\rho} = \mathbf{u}_{\rho} + \frac{1}{\tau_n} (\mathcal{P}_{\rho} - \rho) \mathbf{n},$$

$$(4.2c) \quad \widehat{\mathcal{W}}_{\mathbf{f}} = \mathcal{W}_{\mathbf{f}} + \tau_t (\mathbf{u}_{\mathbf{f}} \times \mathbf{n}), \quad \widehat{\mathbf{u}}_{\mathbf{f}} = \mathbf{u}_{\mathbf{f}} + \frac{1}{\tau_n} \mathcal{P}_{\mathbf{f}} \mathbf{n}.$$

Clearly these equations are inherited from the definitions (3.4d) and (3.4e).

LEMMA 4.3 (elementary identities). *For any $\lambda_t, \mu_t \in \mathbf{L}^2(\mathcal{E}_h)$, any $\rho, \psi \in L^2(\mathcal{E}_h)$, and any $\mathbf{f} \in \mathbf{L}^2(\Omega)$, we have*

$$\begin{aligned} -\langle \llbracket \mathbf{n} \times \widehat{\mathcal{W}}_{\lambda_t} \rrbracket, \mu_t \rangle_{\mathcal{E}_h} &= (\mathcal{W}_{\lambda_t}, \mathcal{W}_{\mu_t})_{\Omega_h} + \langle \tau_t (\lambda_t - \mathbf{u}_{\lambda_t})_t, (\mu_t - \mathbf{u}_{\mu_t})_t \rangle_{\partial\Omega_h} \\ &\quad + \left\langle \frac{1}{\tau_n} \mathcal{P}_{\lambda_t}, \mathcal{P}_{\mu_t} \right\rangle_{\partial\Omega_h} \\ -\langle \llbracket \mathbf{n} \times \widehat{\mathcal{W}}_{\rho} \rrbracket, \mu_t \rangle_{\mathcal{E}_h} &= \langle \llbracket \widehat{\mathbf{u}}_{\mu_t} \cdot \mathbf{n} \rrbracket, \rho \rangle_{\mathcal{E}_h}, \\ -\langle \llbracket \mathbf{n} \times \widehat{\mathcal{W}}_{\mathbf{f}} \rrbracket, \mu_t \rangle_{\mathcal{E}_h} &= -(\mathbf{f}, \mathbf{u}_{\mu_t})_{\Omega_h} \end{aligned}$$

and

$$\begin{aligned} -\langle \llbracket \widehat{\mathbf{u}}_{\lambda_t} \cdot \mathbf{n} \rrbracket, \psi \rangle_{\mathcal{E}_h} &= \langle \llbracket \mathbf{n} \times \widehat{\mathcal{W}}_{\psi} \rrbracket, \lambda_t \rangle_{\mathcal{E}_h}, \\ -\langle \llbracket \widehat{\mathbf{u}}_{\rho} \cdot \mathbf{n} \rrbracket, \psi \rangle_{\mathcal{E}_h} &= (\mathcal{W}_{\rho}, \mathcal{W}_{\psi})_{\Omega_h} + \langle \tau_t (\mathbf{u}_{\rho})_t, (\mathbf{u}_{\psi})_t \rangle_{\partial\Omega_h} \\ &\quad + \left\langle \frac{1}{\tau_n} (\mathcal{P}_{\rho} - \rho), (\mathcal{P}_{\psi} - \psi) \right\rangle_{\partial\Omega_h}, \\ -\langle \llbracket \widehat{\mathbf{u}}_{\mathbf{f}} \cdot \mathbf{n} \rrbracket, \psi \rangle_{\mathcal{E}_h} &= +(\mathbf{f}, \mathbf{u}_{\psi})_{\Omega_h}. \end{aligned}$$

Proof. In all the applications of Lemma 4.1 and Corollary 4.2 in this proof, we take $(\mathbf{P}, \mathbf{P}) = (\mathbf{0}, \mathbf{0})$. Observe that (4.1) is satisfied by $(\mathbf{w}_h, \mathbf{u}_h, \mathbf{p}_h) = (\mathcal{W}_{\mu_t}, \mathbf{u}_{\mu_t}, \mathcal{P}_{\mu_t})$ if we set

$$(\widehat{\mathbf{w}}_h, (\widehat{\mathbf{u}}_h)_t, (\widehat{\mathbf{u}}_h)_n, \widehat{\mathbf{p}}_h, \mathbf{f}) = (\widehat{\mathcal{W}}_{\mu_t}, \mu_t, (\widehat{\mathbf{u}}_{\mu_t})_n, 0, \mathbf{0}).$$

The system (4.1) is also satisfied by $(\mathbf{w}'_h, \mathbf{u}'_h, \mathbf{p}'_h) = (\mathcal{W}_{\lambda_t}, \mathbf{u}_{\lambda_t}, \mathcal{P}_{\lambda_t})$ if we set

$$(\widehat{\mathbf{w}}'_h, (\widehat{\mathbf{u}}'_h)_t, (\widehat{\mathbf{u}}'_h)_n, \widehat{\mathbf{p}}'_h, \mathbf{f}') = (\widehat{\mathcal{W}}_{\lambda_t}, \lambda_t, (\widehat{\mathbf{u}}_{\lambda_t})_n, 0, \mathbf{0}).$$

Hence, by Lemma 4.1,

$$\begin{aligned} -\langle \llbracket \mathbf{n} \times \widehat{\mathcal{W}}_{\lambda_t} \rrbracket, \mu_t \rangle_{\mathcal{E}_h} &= (\mathcal{W}_{\lambda_t}, \mathcal{W}_{\mu_t})_{\Omega_h} - \langle \mu_t - \mathbf{u}_{\mu_t}, \mathbf{n} \times (\widehat{\mathcal{W}}_{\lambda_t} - \mathcal{W}_{\lambda_t}) \rangle_{\partial\Omega_h} \\ &\quad - \langle \widehat{\mathbf{u}}_{\mu_t} - \mathbf{u}_{\mu_t}, \mathbf{n} (0 - \mathcal{P}_{\lambda_t}) \rangle_{\partial\Omega_h}. \end{aligned}$$

The first identity of the lemma follows from this and the identities defining the numerical traces such as (4.2).

The second identity of the lemma follows just as the fourth; see below. The third identity follows from Corollary 4.2. It is easy to check that the conditions of the corollary are satisfied with

$$\begin{aligned} (\mathbf{w}_h, \mathbf{u}_h, \mathbf{p}_h) &= (\mathcal{W}_{\mu_t}, \mathbf{U}_{\mu_t}, \mathcal{P}_{\mu_t}), & (\widehat{\mathbf{w}}_h, (\widehat{\mathbf{u}}_h)_t, (\widehat{\mathbf{u}}_h)_n, \widehat{\mathbf{p}}_h, \mathbf{f}) &= (\widehat{\mathcal{W}}_{\mu_t}, \mu_t, (\widehat{\mathbf{U}}_{\mu_t})_n, \mathbf{0}, \mathbf{0}), \\ (\mathbf{w}'_h, \mathbf{u}'_h, \mathbf{p}'_h) &= (\mathcal{W}_f, \mathbf{U}_f, \mathcal{P}_f), & (\widehat{\mathbf{w}}'_h, (\widehat{\mathbf{u}}'_h)_t, (\widehat{\mathbf{u}}'_h)_n, \widehat{\mathbf{p}}'_h, \mathbf{f}') &= (\widehat{\mathcal{W}}_f, \mathbf{0}, (\widehat{\mathbf{U}}_f)_n, \mathbf{0}, \mathbf{f}). \end{aligned}$$

Hence the corollary implies that

$$\begin{aligned} -\langle \mathbf{u}_t, \mathbf{n} \times \widehat{\mathcal{W}}_f \rangle_{\partial\Omega_h} + (\mathbf{U}_{\mu_t}, \mathbf{f})_{\Omega_h} &= -\langle (\widehat{\mathbf{u}}_h)_t, \mathbf{n} \times \widehat{\mathbf{w}}'_h \rangle_{\partial\Omega_h} + (\mathbf{u}_h, \mathbf{f}')_{\Omega_h} \\ &= -\langle (\widehat{\mathbf{u}}'_h)_t, \mathbf{n} \times \widehat{\mathbf{w}}_h \rangle_{\partial\Omega_h} + (\mathbf{u}'_h, \mathbf{f})_{\Omega_h} \\ &= -\langle \mathbf{0}, \mathbf{n} \times \widehat{\mathcal{W}}_{\mu_t} \rangle_{\partial\Omega_h} \\ &= 0, \end{aligned}$$

and the required identity follows.

The fourth identity also follows from Corollary 4.2 after verifying its conditions with

$$\begin{aligned} (\mathbf{w}_h, \mathbf{u}_h, \mathbf{p}_h) &= (\mathcal{W}_{\lambda_t}, \mathbf{U}_{\lambda_t}, \mathcal{P}_{\lambda_t}), & (\widehat{\mathbf{w}}_h, (\widehat{\mathbf{u}}_h)_t, (\widehat{\mathbf{u}}_h)_n, \widehat{\mathbf{p}}_h, \mathbf{f}) &= (\widehat{\mathcal{W}}_{\lambda_t}, \lambda_t, (\widehat{\mathbf{U}}_{\lambda_t})_n, \mathbf{0}, \mathbf{0}), \\ (\mathbf{w}'_h, \mathbf{u}'_h, \mathbf{p}'_h) &= (\mathcal{W}_{\psi}, \mathbf{U}_{\psi}, \mathcal{P}_{\psi}), & (\widehat{\mathbf{w}}'_h, (\widehat{\mathbf{u}}'_h)_t, (\widehat{\mathbf{u}}'_h)_n, \widehat{\mathbf{p}}'_h, \mathbf{f}') &= (\widehat{\mathcal{W}}_{\psi}, \mathbf{0}, (\widehat{\mathbf{U}}_{\psi})_n, \psi, \mathbf{0}). \end{aligned}$$

The fifth identity follows from Lemma 4.1 with

$$\begin{aligned} (\mathbf{w}_h, \mathbf{u}_h, \mathbf{p}_h) &= (\mathcal{W}_{\rho}, \mathbf{U}_{\rho}, \mathcal{P}_{\rho}), & (\widehat{\mathbf{w}}_h, (\widehat{\mathbf{u}}_h)_t, (\widehat{\mathbf{u}}_h)_n, \widehat{\mathbf{p}}_h, \mathbf{f}) &= (\widehat{\mathcal{W}}_{\rho}, \mathbf{0}, (\mathbf{U}_{\rho})_n, \rho, \mathbf{0}), \\ (\mathbf{w}'_h, \mathbf{u}'_h, \mathbf{p}'_h) &= (\mathcal{W}_{\psi}, \mathbf{U}_{\psi}, \mathcal{P}_{\psi}), & (\widehat{\mathbf{w}}'_h, (\widehat{\mathbf{u}}'_h)_t, (\widehat{\mathbf{u}}'_h)_n, \widehat{\mathbf{p}}'_h, \mathbf{f}') &= (\widehat{\mathcal{W}}_{\psi}, \mathbf{0}, (\widehat{\mathbf{U}}_{\psi})_n, \psi, \mathbf{0}). \end{aligned}$$

The sixth identity follows from Corollary 4.2 after verifying its conditions with

$$\begin{aligned} (\mathbf{w}_h, \mathbf{u}_h, \mathbf{p}_h) &= (\mathcal{W}_f, \mathbf{U}_f, \mathcal{P}_f), & (\widehat{\mathbf{w}}_h, (\widehat{\mathbf{u}}_h)_t, (\widehat{\mathbf{u}}_h)_n, \widehat{\mathbf{p}}_h, \mathbf{f}) &= (\widehat{\mathcal{W}}_{\mu_t}, \mathbf{0}, (\widehat{\mathbf{U}}_f)_n, \mathbf{0}, \mathbf{f}), \\ (\mathbf{w}'_h, \mathbf{u}'_h, \mathbf{p}'_h) &= (\mathcal{W}_{\psi}, \mathbf{U}_{\psi}, \mathcal{P}_{\psi}), & (\widehat{\mathbf{w}}'_h, (\widehat{\mathbf{u}}'_h)_t, (\widehat{\mathbf{u}}'_h)_n, \widehat{\mathbf{p}}'_h, \mathbf{f}') &= (\widehat{\mathcal{W}}_{\psi}, \mathbf{0}, (\widehat{\mathbf{U}}_{\psi})_n, \psi, \mathbf{0}). \end{aligned}$$

This completes the proof of the lemma. \square

Proof of Theorem 3.2. By the jump conditions (3.6b) and (3.6c),

$$\begin{aligned} -\langle [\mathbf{n} \times (\widehat{\mathcal{W}}_{\lambda_t^o} + \widehat{\mathcal{W}}_{\rho})], \mu_t \rangle_{\mathcal{E}_h} &= \langle [\mathbf{n} \times (\widehat{\mathcal{W}}_f + \widehat{\mathcal{W}}_g)], \mu_t \rangle_{\mathcal{E}_h}, \\ -\langle [(\widehat{\mathbf{U}}_{\lambda_t^o} + \widehat{\mathbf{U}}_{\rho}) \cdot \mathbf{n}], \psi \rangle_{\mathcal{E}_h} &= \langle [(\widehat{\mathbf{U}}_f + \widehat{\mathbf{U}}_g) \cdot \mathbf{n}], \psi \rangle_{\mathcal{E}_h} - \langle \mathbf{g} \cdot \mathbf{n}, \psi \rangle_{\partial\Omega}. \end{aligned}$$

By Lemma 4.3, we have that

$$\begin{aligned} -\langle [\mathbf{n} \times \widehat{\mathcal{W}}_{\lambda_t^o}], \mu_t \rangle_{\mathcal{E}_h} &= a_h(\lambda_t^o, \mu_t), & -\langle [(\widehat{\mathbf{U}}_{\lambda_t^o}) \cdot \mathbf{n}], \psi \rangle_{\mathcal{E}_h} &= -b_h(\psi, \lambda_t^o), \\ -\langle [\mathbf{n} \times \widehat{\mathcal{W}}_{\rho}], \mu_t \rangle_{\mathcal{E}_h} &= b_h(\rho, \mu_t), & -\langle [(\widehat{\mathbf{U}}_{\rho}) \cdot \mathbf{n}], \psi \rangle_{\mathcal{E}_h} &= c_h(\rho, \psi). \end{aligned}$$

In order to prove (3.8a) and (3.8b), we now have only to show that $\ell_1 = \widetilde{\ell}_1$ and $\ell_2 = \widetilde{\ell}_2$, where

$$\begin{aligned} \widetilde{\ell}_1(\mu_t) &:= \langle [\mathbf{n} \times \widehat{\mathcal{W}}_f], \mu_t \rangle_{\mathcal{E}_h} + \langle [\mathbf{n} \times \widehat{\mathcal{W}}_g], \mu_t \rangle_{\mathcal{E}_h}, \\ \widetilde{\ell}_2(\psi) &:= \langle [(\widehat{\mathbf{U}}_f \cdot \mathbf{n})], \psi \rangle_{\mathcal{E}_h} + \langle [(\widehat{\mathbf{U}}_g \cdot \mathbf{n})], \psi \rangle_{\mathcal{E}_h} - \langle \mathbf{g} \cdot \mathbf{n}, \psi \rangle_{\partial\Omega}. \end{aligned}$$

But, again by Lemma 4.3, we have

$$\begin{aligned}\tilde{\ell}_1(\boldsymbol{\mu}_t) &= (\mathbf{f}, \mathbf{u}_{\boldsymbol{\mu}_t})_{\Omega_h} - a_h(\mathbf{g}, \boldsymbol{\mu}_t) \\ &= \ell_1(\boldsymbol{\mu}_t).\end{aligned}$$

Similarly, applying Lemma 4.3 one more time,

$$\begin{aligned}\tilde{\ell}_2(\psi) &= -(\mathbf{f}, \mathbf{u}_\psi)_{\Omega_h} - \langle \mathbf{g}, [\mathbf{n} \times \widehat{\mathbf{W}}_\psi] \rangle_{\mathcal{E}_h} - \langle \mathbf{g} \cdot \mathbf{n}, \psi \rangle_{\partial\Omega} \\ &= -(\mathbf{f}, \mathbf{u}_\psi)_{\Omega_h} + b_h(\psi, \mathbf{g}_t) - \langle \mathbf{g} \cdot \mathbf{n}, \psi \rangle_{\partial\Omega} \\ &= \ell_2(\psi).\end{aligned}$$

It now only remains to prove that $(\boldsymbol{\lambda}_t^\circ, \rho)$ is the only solution of (3.8a)–(3.8c). First observe that the above arguments in fact show that the jump conditions (3.6b) and (3.6c) hold if and only if (3.8a) and (3.8b) hold, respectively. Hence if $(\tilde{\boldsymbol{\lambda}}_t^\circ, \tilde{\rho})$ is another solution of (3.8a)–(3.8c), then the numerical traces generated by $\mathcal{L}^I(\tilde{\boldsymbol{\lambda}}_t^\circ, \tilde{\rho}, \mathbf{f})$ will also satisfy (3.6b) and (3.6c). But then, since (3.8c) implies (3.6d), we find that all the conditions of Theorem 3.1 are verified, so we conclude that $\tilde{\boldsymbol{\lambda}}_t^\circ + \mathbf{g}_t = (\widehat{\mathbf{u}}_h)_t$ and $\tilde{\rho} = \widehat{p}_h$. Since we also have $(\boldsymbol{\lambda}_t^\circ + \mathbf{g}_t, \rho) = ((\widehat{\mathbf{u}}_h)_t, \widehat{p}_h)$, we conclude that $(\tilde{\boldsymbol{\lambda}}_t^\circ, \tilde{\rho}) = (\boldsymbol{\lambda}_t^\circ, \rho)$. This completes the proof of Theorem 3.2. \square

4.3. Proof of the characterization of Theorem 3.4. To prove Theorem 3.4, we proceed as in the previous case and gather several key identities in the following result. Recall the definitions of specific local solutions in (3.15) (such as $\mathbf{W}_\lambda, \mathbf{u}_\lambda$, etc.). The numerical traces $\widehat{\mathbf{W}}_\odot$ and $\widehat{\mathcal{P}}_\odot$ are given by (3.11) for the choices of subscript \odot that make sense here, such as when \odot is $\lambda, \bar{\rho}$, or \mathbf{f} , e.g.,

$$\widehat{\mathcal{P}}_\lambda = \mathcal{P}_\lambda + \tau_n (\mathbf{u}_\lambda - \boldsymbol{\lambda}) \cdot \mathbf{n}, \quad \widehat{\mathbf{W}}_{\bar{\rho}} = \mathbf{W}_{\bar{\rho}} + \tau_t \mathbf{u}_{\bar{\rho}} \times \mathbf{n},$$

just as in the previous case.

LEMMA 4.4 (elementary identities). *For any $\boldsymbol{\lambda}, \boldsymbol{\mu} \in \mathbf{L}^2(\mathcal{E}_h)$, any $\bar{\rho} \in \ell^2(\partial\Omega_h)$, and any $\mathbf{f} \in \mathbf{L}^2(\Omega)$, we have*

$$\begin{aligned}-\langle [\mathbf{n} \times \widehat{\mathbf{W}}_\lambda + \widehat{\mathcal{P}}_\lambda \mathbf{n}], \boldsymbol{\mu} \rangle_{\mathcal{E}_h} &= (\mathbf{W}_\lambda, \mathbf{W}_\mu)_{\Omega_h} + \langle \tau_t (\boldsymbol{\lambda} - \mathbf{u}_\lambda)_t, (\boldsymbol{\mu} - \mathbf{u}_\mu)_t \rangle_{\partial\Omega_h} \\ &\quad + \langle \tau_n (\boldsymbol{\lambda} - \mathbf{u}_\lambda)_n, (\boldsymbol{\mu} - \mathbf{u}_\mu)_n \rangle_{\partial\Omega_h}, \\ -\langle [\mathbf{n} \times \widehat{\mathbf{W}}_{\bar{\rho}} + \widehat{\mathcal{P}}_{\bar{\rho}} \mathbf{n}], \boldsymbol{\mu} \rangle_{\mathcal{E}_h} &= -\langle \bar{\rho}, \boldsymbol{\mu} \cdot \mathbf{n} \rangle_{\partial\Omega_h}, \\ -\langle [\mathbf{n} \times \widehat{\mathbf{W}}_{\mathbf{f}} + \widehat{\mathcal{P}}_{\mathbf{f}} \mathbf{n}], \boldsymbol{\mu} \rangle_{\mathcal{E}_h} &= -(\mathbf{f}, \mathbf{u}_\mu)_{\Omega_h}.\end{aligned}$$

Proof. The second identity immediately follows because by (3.16),

$$\mathbf{n} \times \widehat{\mathbf{W}}_{\bar{\rho}} + \mathbf{n} \widehat{\mathcal{P}}_{\bar{\rho}} = +\bar{\rho} \mathbf{n}.$$

To prove the remaining identities, we set $\mathbf{P} = \mathbf{0}$ and $\mathbf{P}\psi = \bar{\psi}$ (where $\bar{\psi}$ is as defined in (3.12)) and apply Lemma 4.1 and Corollary 4.2 appropriately. Indeed, to prove the first identity, first observe that (4.1) is satisfied by

$$\begin{aligned}(\mathbf{w}'_h, \mathbf{u}'_h, \mathbf{p}'_h) &= (\mathbf{W}_\lambda, \mathbf{u}_\lambda, \mathcal{P}_\lambda) \quad \text{with} \quad (\widehat{\mathbf{w}}'_h, \widehat{\mathbf{u}}'_h, \widehat{\mathbf{p}}'_h, \mathbf{f}') = (\widehat{\mathbf{W}}_\lambda, \boldsymbol{\lambda}, \widehat{\mathcal{P}}_\lambda, \mathbf{0}), \quad \text{and} \\ (\mathbf{w}_h, \mathbf{u}_h, \mathbf{p}_h) &= (\mathbf{W}_\mu, \mathbf{u}_\mu, \mathcal{P}_\mu) \quad \text{with} \quad (\widehat{\mathbf{w}}_h, \widehat{\mathbf{u}}_h, \widehat{\mathbf{p}}_h, \mathbf{f}) = (\widehat{\mathbf{W}}_\mu, \boldsymbol{\mu}, \widehat{\mathcal{P}}_\mu, \mathbf{0}).\end{aligned}$$

Furthermore, $\mathbf{P}\mathcal{P}_\lambda = 0$. Hence the first identity follows by applying Lemma 4.1.

Similarly, the last identity follows from Corollary 4.2, setting

$$\begin{aligned} (\mathbf{w}'_h, \mathbf{u}'_h, \mathbf{p}'_h) &= (\mathcal{W}_f, \mathbf{U}_f, \mathcal{P}_f), & (\widehat{\mathbf{w}}'_h, \widehat{\mathbf{u}}'_h, \widehat{\mathbf{p}}'_h, \mathbf{f}') &= (\widehat{\mathcal{W}}_f, \mathbf{0}, \widehat{\mathcal{P}}_f, \mathbf{f}) \quad \text{and} \\ (\mathbf{w}_h, \mathbf{u}_h, \mathbf{p}_h) &= (\mathcal{W}_\mu, \mathbf{U}_\mu, \mathcal{P}_\mu), & (\widehat{\mathbf{w}}_h, \widehat{\mathbf{u}}_h, \widehat{\mathbf{p}}_h, \mathbf{f}) &= (\widehat{\mathcal{W}}_\mu, \boldsymbol{\mu}, \widehat{\mathcal{P}}_\mu, \mathbf{0}). \end{aligned}$$

This completes the proof of the identities. \square

Proof of Theorem 3.4. By the jump conditions (3.14b) and (3.14c),

$$\begin{aligned} -\langle [\mathbf{n} \times \widehat{\mathcal{W}}_{\lambda^\circ} + \widehat{\mathcal{P}}_{\lambda^\circ} \mathbf{n} + \mathbf{n} \times \widehat{\mathcal{W}}_{\bar{\rho}} + \widehat{\mathcal{P}}_{\bar{\rho}} \mathbf{n}], \boldsymbol{\mu} \rangle_{\mathcal{E}_h} \\ = \langle [\mathbf{n} \times \widehat{\mathcal{W}}_f + \widehat{\mathcal{P}}_f \mathbf{n} + \mathbf{n} \times \widehat{\mathcal{W}}_g + \widehat{\mathcal{P}}_g \mathbf{n}], \boldsymbol{\mu} \rangle_{\mathcal{E}_h}. \end{aligned}$$

By Lemma 4.4, we have that

$$\begin{aligned} -\langle [\mathbf{n} \times \widehat{\mathcal{W}}_{\lambda^\circ} + \widehat{\mathcal{P}}_{\lambda^\circ} \mathbf{n}], \boldsymbol{\mu} \rangle_{\mathcal{E}_h} &= a_h(\boldsymbol{\lambda}^\circ, \boldsymbol{\mu}), \\ -\langle [\mathbf{n} \times \widehat{\mathcal{W}}_{\bar{\rho}} + \widehat{\mathcal{P}}_{\bar{\rho}} \mathbf{n}], \boldsymbol{\mu} \rangle_{\mathcal{E}_h} &= b_h(\bar{\rho}, \boldsymbol{\mu}). \end{aligned}$$

It remains to show that the form $\ell(\cdot)$ of the theorem coincides with $\tilde{\ell}$ defined by

$$\tilde{\ell}(\boldsymbol{\mu}) := \langle [\mathbf{n} \times \widehat{\mathcal{W}}_f + \widehat{\mathcal{P}}_f \mathbf{n}], \boldsymbol{\mu} \rangle_{\mathcal{E}_h} + \langle [\mathbf{n} \times \widehat{\mathcal{W}}_g + \widehat{\mathcal{P}}_g \mathbf{n}], \boldsymbol{\mu} \rangle_{\mathcal{E}_h}.$$

But, again by Lemma 4.4, we have

$$\begin{aligned} \tilde{\ell}(\boldsymbol{\mu}) &= (\mathbf{f}, \mathbf{U}_\mu)_{\Omega_h} - a_h(\mathbf{g}, \boldsymbol{\mu}) \\ &= \ell(\boldsymbol{\mu}). \end{aligned}$$

The proof of uniqueness of the trace solution $(\boldsymbol{\lambda}^\circ, \bar{\rho})$ proceeds as in the Type I case, so we omit it. \square

4.4. Proof of the characterization of Theorem 3.6. We now prove Theorem 3.6, using the identities gathered in the next lemma. The notation for the numerical traces of the form $\widehat{\mathbf{U}}_\circ$ and $\widehat{\mathcal{P}}_\circ$ have meanings inherited from (3.18e) and (3.18f) as in the previous cases.

LEMMA 4.5 (elementary identities). *For any $\gamma_t, \delta_t \in \mathbf{L}^2(\mathcal{E}_h)$, any $\boldsymbol{\lambda}_n, \boldsymbol{\mu}_n \in \mathbf{L}^2(\mathcal{E}_h)$, any $\rho, \psi \in L^2(\mathcal{E}_h)$, and any $\mathbf{f} \in \mathbf{L}^2(\Omega)$, we have*

$$\begin{aligned} -\langle [\widehat{\mathbf{U}}_{\gamma_t} \times \mathbf{n}], \boldsymbol{\delta}_t \rangle_{\mathcal{E}_h} &= (\mathcal{W}_{\gamma_t}, \mathcal{W}_{\delta_t})_{\Omega_h} + \langle \tau_n(\mathbf{U}_{\gamma_t})_n, (\mathbf{U}_{\delta_t})_n \rangle_{\partial\Omega_h} \\ &\quad + \left\langle \frac{1}{\tau_t} \mathbf{n} \times (\gamma_t - \mathcal{W}_{\gamma_t}), \mathbf{n} \times (\delta_t - \mathcal{W}_{\delta_t}) \right\rangle_{\partial\Omega_h}, \\ -\langle [\widehat{\mathbf{U}}_{\boldsymbol{\lambda}_n} \times \mathbf{n}], \boldsymbol{\delta}_t \rangle_{\mathcal{E}_h} &= \langle [\widehat{\mathcal{P}}_{\delta_t} \mathbf{n}], \boldsymbol{\lambda}_n \rangle_{\mathcal{E}_h}, \\ -\langle [\widehat{\mathbf{U}}_{\bar{\rho}} \times \mathbf{n}], \boldsymbol{\delta}_t \rangle_{\mathcal{E}_h} &= 0, \\ -\langle [\widehat{\mathbf{U}}_{\mathbf{f}} \times \mathbf{n}], \boldsymbol{\delta}_t \rangle_{\mathcal{E}_h} &= (\mathbf{f}, \mathbf{U}_{\delta_t})_{\Omega_h} \end{aligned}$$

and

$$\begin{aligned} -\langle [\widehat{\mathcal{P}}_{\gamma_t} \mathbf{n}], \boldsymbol{\mu}_n \rangle_{\mathcal{E}_h} &= \langle [\widehat{\mathbf{U}}_{\boldsymbol{\mu}_n} \times \mathbf{n}], \gamma_t \rangle_{\mathcal{E}_h}, \\ -\langle [\widehat{\mathcal{P}}_{\boldsymbol{\lambda}_n} \mathbf{n}], \boldsymbol{\mu}_n \rangle_{\mathcal{E}_h} &= (\mathcal{W}_{\boldsymbol{\lambda}_n}, \mathcal{W}_{\boldsymbol{\mu}_n})_{\Omega_h} + \left\langle \frac{1}{\tau_t} \mathbf{n} \times \mathcal{W}_{\boldsymbol{\lambda}_n}, \mathbf{n} \times \mathcal{W}_{\boldsymbol{\mu}_n} \right\rangle_{\partial\Omega_h} \\ &\quad + \langle \tau_n(\boldsymbol{\lambda}_n - \mathbf{U}_{\boldsymbol{\lambda}_n})_n, (\boldsymbol{\mu}_n - \mathbf{U}_{\boldsymbol{\mu}_n})_n \rangle_{\partial\Omega_h}, \\ -\langle [\widehat{\mathcal{P}}_{\bar{\rho}} \mathbf{n}], \boldsymbol{\mu}_n \rangle_{\mathcal{E}_h} &= -\langle \bar{\rho}, \boldsymbol{\mu}_n \cdot \mathbf{n} \rangle_{\partial\Omega_h} \\ -\langle [\widehat{\mathcal{P}}_{\mathbf{f}} \mathbf{n}], \boldsymbol{\mu}_n \rangle_{\mathcal{E}_h} &= -(\mathbf{f}, \mathbf{U}_{\boldsymbol{\mu}_n})_{\Omega_h}. \end{aligned}$$

Proof. The third and seventh identities immediately follow because $\widehat{\mathbf{u}}_{\bar{p}} = \mathbf{0}$ and $\widehat{\mathcal{P}}_{\bar{p}} = \bar{p}$.

For proving the remaining identities, we apply Lemma 4.1 and Corollary 4.2 with $\mathbf{P} = \mathbf{0}$ and $\mathcal{P}\psi = \bar{\psi}$. To prove the first identity, observe that (4.1) is satisfied by

$$(\mathbf{w}'_h, \mathbf{u}'_h, \mathbf{p}'_h) = (\mathcal{W}_{\delta_t}, \mathbf{u}_{\delta_t}, \mathcal{P}_{\delta_t}) \text{ with } (\widehat{\mathbf{w}}'_h, (\widehat{\mathbf{u}}'_h)_t, (\widehat{\mathbf{u}}'_h)_n, \widehat{\mathbf{p}}'_h, \mathbf{f}') = (\delta_t, (\widehat{\mathbf{u}}_{\delta_t})_t, 0, \widehat{\mathcal{P}}_{\delta_t}, \mathbf{0}).$$

Equation (4.1) is also satisfied by

$$(\mathbf{w}'_h, \mathbf{u}'_h, \mathbf{p}'_h) = (\mathcal{W}_{\gamma_t}, \mathbf{u}_{\gamma_t}, \mathcal{P}_{\gamma_t}) \text{ with } (\widehat{\mathbf{w}}'_h, (\widehat{\mathbf{u}}'_h)_t, (\widehat{\mathbf{u}}'_h)_n, \widehat{\mathbf{p}}'_h, \mathbf{f}') = (\gamma_t, (\widehat{\mathbf{u}}_{\gamma_t})_t, 0, \widehat{\mathcal{P}}_{\gamma_t}, \mathbf{0}).$$

Since we also have $\mathcal{P}\mathcal{P}_{\delta_t} = 0$ because of (3.18d), all the conditions for applying Lemma 4.1 are satisfied. Thus the first identity follows from Lemma 4.1.

The second identity follows like the fifth; see below. The fourth identity follows from Corollary 4.2 with

$$\begin{aligned} (\mathbf{w}'_h, \mathbf{u}'_h, \mathbf{p}'_h) &= (\mathcal{W}_{\delta_t}, \mathbf{u}_{\delta_t}, \mathcal{P}_{\delta_t}), & (\widehat{\mathbf{w}}'_h, (\widehat{\mathbf{u}}'_h)_t, (\widehat{\mathbf{u}}'_h)_n, \widehat{\mathbf{p}}'_h, \mathbf{f}') &= (\delta_t, (\widehat{\mathbf{u}}_{\delta_t})_t, 0, \widehat{\mathcal{P}}_{\delta_t}, \mathbf{0}), \\ (\mathbf{w}_h, \mathbf{u}_h, \mathbf{p}_h) &= (\mathcal{W}_{\mathbf{f}}, \mathbf{u}_{\mathbf{f}}, \mathcal{P}_{\mathbf{f}}), & (\widehat{\mathbf{w}}_h, (\widehat{\mathbf{u}}_h)_t, (\widehat{\mathbf{u}}_h)_n, \widehat{\mathbf{p}}_h, \mathbf{f}) &= (\mathbf{0}, (\widehat{\mathbf{u}}_{\mathbf{f}})_t, \mathbf{0}, \widehat{\mathcal{P}}_{\mathbf{f}}, \mathbf{f}). \end{aligned}$$

The fifth identity follows from Corollary 4.2 with

$$\begin{aligned} (\mathbf{w}'_h, \mathbf{u}'_h, \mathbf{p}'_h) &= (\mathcal{W}_{\gamma_t}, \mathbf{u}_{\gamma_t}, \mathcal{P}_{\gamma_t}), & (\widehat{\mathbf{w}}'_h, (\widehat{\mathbf{u}}'_h)_t, (\widehat{\mathbf{u}}'_h)_n, \widehat{\mathbf{p}}'_h, \mathbf{f}') &= (\gamma_t, (\widehat{\mathbf{u}}_{\gamma_t})_t, 0, \widehat{\mathcal{P}}_{\gamma_t}, \mathbf{0}), \\ (\mathbf{w}_h, \mathbf{u}_h, \mathbf{p}_h) &= (\mathcal{W}_{\mu_n}, \mathbf{u}_{\mu_n}, \mathcal{P}_{\mu_n}), & (\widehat{\mathbf{w}}_h, (\widehat{\mathbf{u}}_h)_t, (\widehat{\mathbf{u}}_h)_n, \widehat{\mathbf{p}}_h, \mathbf{f}) &= (\mathbf{0}, (\widehat{\mathbf{u}}_{\mu_n})_t, \mu_n, \widehat{\mathcal{P}}_{\mu_n}, \mathbf{0}). \end{aligned}$$

The sixth identity follows from Lemma 4.1 with

$$\begin{aligned} (\mathbf{w}_h, \mathbf{u}_h, \mathbf{p}_h) &= (\mathcal{W}_{\mu_n}, \mathbf{u}_{\mu_n}, \mathcal{P}_{\mu_n}), & (\widehat{\mathbf{w}}_h, (\widehat{\mathbf{u}}_h)_t, (\widehat{\mathbf{u}}_h)_n, \widehat{\mathbf{p}}_h, \mathbf{f}) &= (\mathbf{0}, (\widehat{\mathbf{u}}_{\mu_n})_t, \mu_n, \widehat{\mathcal{P}}_{\mu_n}, \mathbf{0}), \\ (\mathbf{w}'_h, \mathbf{u}'_h, \mathbf{p}'_h) &= (\mathcal{W}_{\lambda_n}, \mathbf{u}_{\lambda_n}, \mathcal{P}_{\lambda_n}), & (\widehat{\mathbf{w}}'_h, (\widehat{\mathbf{u}}'_h)_t, (\widehat{\mathbf{u}}'_h)_n, \widehat{\mathbf{p}}'_h, \mathbf{f}') &= (\mathbf{0}, (\widehat{\mathbf{u}}_{\lambda_n})_t, \lambda_n, \widehat{\mathcal{P}}_{\lambda_n}, \mathbf{0}). \end{aligned}$$

The eighth identity follows from Corollary 4.2 with

$$\begin{aligned} (\mathbf{w}_h, \mathbf{u}_h, \mathbf{p}_h) &= (\mathcal{W}_{\mu_n}, \mathbf{u}_{\mu_n}, \mathcal{P}_{\mu_n}), & (\widehat{\mathbf{w}}_h, (\widehat{\mathbf{u}}_h)_t, (\widehat{\mathbf{u}}_h)_n, \widehat{\mathbf{p}}_h, \mathbf{f}) &= (\mathbf{0}, (\widehat{\mathbf{u}}_{\mu_n})_t, \mu_n, \widehat{\mathcal{P}}_{\mu_n}, \mathbf{0}), \\ (\mathbf{w}'_h, \mathbf{u}'_h, \mathbf{p}'_h) &= (\mathcal{W}_{\mathbf{f}}, \mathbf{u}_{\mathbf{f}}, \mathcal{P}_{\mathbf{f}}), & (\widehat{\mathbf{w}}'_h, (\widehat{\mathbf{u}}'_h)_t, (\widehat{\mathbf{u}}'_h)_n, \widehat{\mathbf{p}}'_h, \mathbf{f}') &= (\mathbf{0}, (\widehat{\mathbf{u}}_{\mathbf{f}})_t, \mathbf{0}, \widehat{\mathcal{P}}_{\mathbf{f}}, \mathbf{f}). \end{aligned}$$

This completes the proof. \square

Proof of Theorem 3.6. By the jump conditions (3.20b) and (3.20c),

$$\begin{aligned} -\langle [(\widehat{\mathbf{u}}_{\gamma_t} + \widehat{\mathbf{u}}_{\lambda_n^o} + \widehat{\mathbf{u}}_{\bar{p}}) \times \mathbf{n}], \delta_t \rangle_{\mathcal{E}_h} &= \langle [(\widehat{\mathbf{u}}_{\mathbf{f}_t} + \widehat{\mathbf{u}}_{\mathbf{g}_n}) \times \mathbf{n}], \delta_t \rangle_{\mathcal{E}_h} - \langle \mathbf{g}_t \times \mathbf{n}, \delta_t \rangle_{\partial\Omega}, \\ -\langle [(\widehat{\mathcal{P}}_{\gamma_t} + \widehat{\mathcal{P}}_{\lambda_n^o} + \widehat{\mathcal{P}}_{\bar{p}}) \mathbf{n}], \mu_n \rangle_{\mathcal{E}_h} &= \langle [(\widehat{\mathcal{P}}_{\mathbf{f}_t} + \widehat{\mathcal{P}}_{\mathbf{g}_n}) \mathbf{n}], \mu_n \rangle_{\mathcal{E}_h}. \end{aligned}$$

By Lemma 4.5, we have that

$$\begin{aligned} -\langle [(\widehat{\mathbf{u}}_{\gamma_t} \times \mathbf{n}), \delta_t]_{\mathcal{E}_h} &= a_h(\lambda_t^o, \delta_t), & -\langle [(\widehat{\mathcal{P}}_{\gamma_t} \mathbf{n}), \mu_n]_{\mathcal{E}_h} &= -b_h(\mu_n, \gamma_t), \\ -\langle [(\widehat{\mathbf{u}}_{\lambda_n^o} \times \mathbf{n}), \delta_t]_{\mathcal{E}_h} &= b_h(\lambda_n^o, \delta_t), & -\langle [(\widehat{\mathcal{P}}_{\lambda_n^o} \mathbf{n}), \mu_n]_{\mathcal{E}_h} &= c_h(\lambda_n^o, \mu_n), \\ -\langle [(\widehat{\mathbf{u}}_{\bar{p}} \times \mathbf{n}), \delta_t]_{\mathcal{E}_h} &= 0, & -\langle [(\widehat{\mathcal{P}}_{\bar{p}} \mathbf{n}), \mu_n]_{\mathcal{E}_h} &= d(\bar{p}, \mu_n). \end{aligned}$$

It remains to show that $\ell_1 = \widetilde{\ell}_1$ and $\ell_2 = \widetilde{\ell}_2$, where

$$\begin{aligned} \widetilde{\ell}_1(\delta_t) &:= -\langle [(\widehat{\mathbf{u}}_{\mathbf{f}} + \widehat{\mathbf{u}}_{\mathbf{g}_n}) \times \mathbf{n}], \delta_t \rangle_{\mathcal{E}_h} - \langle \mathbf{g}_t \times \mathbf{n}, \delta_t \rangle_{\partial\Omega}, \\ \widetilde{\ell}_2(\psi) &:= \langle [(\widehat{\mathcal{P}}_{\mathbf{f}_t} + \widehat{\mathcal{P}}_{\mathbf{g}_n}) \mathbf{n}], \mu_n \rangle_{\mathcal{E}_h}. \end{aligned}$$

But, again by Lemma 4.5, we have

$$\begin{aligned} \tilde{\ell}_1(\boldsymbol{\delta}_t) &= -(\mathbf{f}, \mathbf{u}_{\boldsymbol{\delta}_t})_{\Omega_h} - b_h(\mathbf{g}_n, \boldsymbol{\delta}_t) - \langle \mathbf{g}_t \times \mathbf{n}, \boldsymbol{\delta}_t \rangle_{\partial\Omega} \\ &= \ell_1(\boldsymbol{\delta}_t), \end{aligned}$$

and, similarly, by Lemma 4.5,

$$\begin{aligned} \tilde{\ell}_2(\boldsymbol{\mu}_n) &= (\mathbf{f}, \mathbf{u}_{\boldsymbol{\mu}_n})_{\Omega_h} - c_h(\mathbf{g}_n, \boldsymbol{\mu}_n) \\ &= \ell_2(\boldsymbol{\mu}_n). \end{aligned}$$

The proof of Theorem 3.6 is completed by also establishing the uniqueness as in the previous cases. \square

4.5. Proof of the characterization of Theorem 3.8. To prove Theorem 3.8, we use the identities below. The numerical traces of the form $\widehat{\mathbf{u}}_{\odot}$ appearing in these identities are defined using (3.21e) and (3.21f) as in the previous cases for all possible choices of the subscripts \odot that make sense for this case.

LEMMA 4.6 (elementary identities). *For any $\boldsymbol{\gamma}_t, \boldsymbol{\delta}_t \in \mathbf{L}^2(\mathcal{E}_h)$, any $\rho, \psi \in L^2(\mathcal{E}_h)$, any $\phi \in H^1(\Omega_h)$, and any $\mathbf{f} \in \mathbf{L}^2(\Omega)$, we have*

$$\begin{aligned} -\langle [\widehat{\mathbf{u}}_{\boldsymbol{\gamma}_t} \times \mathbf{n}], \boldsymbol{\delta}_t \rangle_{\mathcal{E}_h} &= (\mathbf{W}_{\boldsymbol{\gamma}_t}, \mathbf{W}_{\boldsymbol{\delta}_t})_{\Omega_h} + \left\langle \frac{1}{\tau_n} \mathcal{P}_{\boldsymbol{\gamma}_t}, \mathcal{P}_{\boldsymbol{\delta}_t} \right\rangle_{\partial\Omega_h} \\ &\quad + \left\langle \frac{1}{\tau_t} \mathbf{n} \times (\boldsymbol{\gamma}_t - \mathbf{W}_{\boldsymbol{\gamma}_t}), \mathbf{n} \times (\boldsymbol{\delta}_t - \mathbf{W}_{\boldsymbol{\delta}_t}) \right\rangle_{\partial\Omega_h}, \\ -\langle [\widehat{\mathbf{u}}_{\rho} \times \mathbf{n}], \boldsymbol{\delta}_t \rangle_{\mathcal{E}_h} &= -\langle [\widehat{\mathbf{u}}_{\boldsymbol{\delta}_t} \cdot \mathbf{n}], \rho \rangle_{\mathcal{E}_h}, \\ -\langle [\widehat{\mathbf{u}}_{\phi} \times \mathbf{n}], \boldsymbol{\delta}_t \rangle_{\mathcal{E}_h} &= \langle \mathbf{n} \times \text{grad } \phi, \boldsymbol{\delta}_t \rangle_{\partial\Omega_h}, \\ -\langle [\widehat{\mathbf{u}}_{\mathbf{f}} \times \mathbf{n}], \boldsymbol{\delta}_t \rangle_{\mathcal{E}_h} &= +(\mathbf{f}, \mathbf{u}_{\boldsymbol{\delta}_t})_{\Omega_h} \end{aligned}$$

and

$$\begin{aligned} -\langle [\widehat{\mathbf{u}}_{\boldsymbol{\gamma}_t} \cdot \mathbf{n}], \psi \rangle_{\mathcal{E}_h} &= -\langle [\widehat{\mathbf{u}}_{\boldsymbol{\psi}} \times \mathbf{n}], \boldsymbol{\gamma}_t \rangle_{\mathcal{E}_h} \\ -\langle [\widehat{\mathbf{u}}_{\rho} \cdot \mathbf{n}], \psi \rangle_{\mathcal{E}_h} &= (\mathbf{W}_{\rho}, \mathbf{W}_{\boldsymbol{\psi}})_{\Omega_h} + \left\langle \frac{1}{\tau_n} (\rho - \mathcal{P}_{\rho}), (\boldsymbol{\psi} - \mathcal{P}_{\boldsymbol{\psi}}) \right\rangle_{\partial\Omega_h} \\ &\quad + \left\langle \frac{1}{\tau_t} \mathbf{n} \times \mathbf{W}_{\rho}, \mathbf{n} \times \mathbf{W}_{\boldsymbol{\psi}} \right\rangle_{\partial\Omega_h}, \\ -\langle [\widehat{\mathbf{u}}_{\phi} \cdot \mathbf{n}], \psi \rangle_{\mathcal{E}_h} &= -\langle \text{grad } \phi \cdot \mathbf{n}, \boldsymbol{\psi} \rangle_{\partial\Omega_h}, \\ -\langle [\widehat{\mathbf{u}}_{\mathbf{f}} \cdot \mathbf{n}], \psi \rangle_{\mathcal{E}_h} &= +(\mathbf{f}, \mathbf{u}_{\boldsymbol{\psi}})_{\Omega_h}. \end{aligned}$$

Proof. The third and seventh identities are immediate because (3.24) implies that

$$\widehat{\mathbf{u}}_{\phi} = \text{grad } \phi.$$

In the remainder of the proof, whenever we apply Lemma 4.1 or Corollary 4.2 we take $\mathbf{P}\mathbf{v} = \text{grad } \phi_{\mathbf{v}}$ and $\mathbf{P} = 0$. To prove the first identity, we proceed as in the previous cases and apply Lemma 4.1 (now additionally noting that $\mathbf{P}\mathbf{u}_{\boldsymbol{\gamma}_t} = \mathbf{0}$) with

$$\begin{aligned} (\mathbf{w}_h, \mathbf{u}_h, \mathbf{p}_h) &= (\mathbf{W}_{\boldsymbol{\gamma}_t}, \mathbf{u}_{\boldsymbol{\gamma}_t}, \mathcal{P}_{\boldsymbol{\gamma}_t}), & (\widehat{\mathbf{w}}_h, \widehat{\mathbf{u}}_h, \widehat{\mathbf{p}}_h, \mathbf{f}) &= (\boldsymbol{\gamma}_t, \widehat{\mathbf{u}}_{\boldsymbol{\gamma}_t}, 0, \mathbf{0}), \\ (\mathbf{w}'_h, \mathbf{u}'_h, \mathbf{p}'_h) &= (\mathbf{W}_{\boldsymbol{\delta}_t}, \mathbf{u}_{\boldsymbol{\delta}_t}, \mathcal{P}_{\boldsymbol{\delta}_t}), & (\widehat{\mathbf{w}}'_h, \widehat{\mathbf{u}}'_h, \widehat{\mathbf{p}}'_h, \mathbf{f}') &= (\boldsymbol{\delta}_t, \widehat{\mathbf{u}}_{\boldsymbol{\delta}_t}, 0, \mathbf{0}). \end{aligned}$$

The second identity is proved just like the fifth; see below. The fourth identity follows from Corollary 4.2 with

$$\begin{aligned} (\mathbf{w}_h, \mathbf{u}_h, \mathbf{p}_h) &= (\mathcal{W}_f, \mathcal{U}_f, \mathcal{P}_f), & (\widehat{\mathbf{w}}_h, \widehat{\mathbf{u}}_h, \widehat{\mathbf{p}}_h, \mathbf{f}) &= (0, \widehat{\mathbf{U}}_f, 0, \mathbf{f}), \\ (\mathbf{w}'_h, \mathbf{u}'_h, \mathbf{p}'_h) &= (\mathcal{W}_{\delta_t}, \mathcal{U}_{\delta_t}, \mathcal{P}_{\delta_t}), & (\widehat{\mathbf{w}}'_h, \widehat{\mathbf{u}}'_h, \widehat{\mathbf{p}}'_h, \mathbf{f}') &= (\delta_t, \widehat{\mathbf{U}}_{\delta_t}, 0, \mathbf{0}), \end{aligned}$$

The fifth identity follows from Corollary 4.2 with

$$\begin{aligned} (\mathbf{w}_h, \mathbf{u}_h, \mathbf{p}_h) &= (\mathcal{W}_{\gamma_t}, \mathcal{U}_{\gamma_t}, \mathcal{P}_{\gamma_t}), & (\widehat{\mathbf{w}}_h, \widehat{\mathbf{u}}_h, \widehat{\mathbf{p}}_h, \mathbf{f}) &= (\gamma_t, \widehat{\mathbf{U}}_{\gamma_t}, 0, \mathbf{0}), \\ (\mathbf{w}'_h, \mathbf{u}'_h, \mathbf{p}'_h) &= (\mathcal{W}_\psi, \mathcal{U}_\psi, \mathcal{P}_\psi), & (\widehat{\mathbf{w}}'_h, \widehat{\mathbf{u}}'_h, \widehat{\mathbf{p}}'_h, \mathbf{f}') &= (0, \widehat{\mathbf{U}}_\psi, \psi, \mathbf{0}). \end{aligned}$$

The sixth identity follows from Lemma 4.1 with

$$\begin{aligned} (\mathbf{w}_h, \mathbf{u}_h, \mathbf{p}_h) &= (\mathcal{W}_\rho, \mathcal{U}_\rho, \mathcal{P}_\rho), & (\widehat{\mathbf{w}}_h, \widehat{\mathbf{u}}_h, \widehat{\mathbf{p}}_h, \mathbf{f}) &= (0, \widehat{\mathbf{U}}_\rho, \rho, \mathbf{0}), \\ (\mathbf{w}'_h, \mathbf{u}'_h, \mathbf{p}'_h) &= (\mathcal{W}_\psi, \mathcal{U}_\psi, \mathcal{P}_\psi), & (\widehat{\mathbf{w}}'_h, \widehat{\mathbf{u}}'_h, \widehat{\mathbf{p}}'_h, \mathbf{f}') &= (0, \widehat{\mathbf{U}}_\psi, \psi, \mathbf{0}). \end{aligned}$$

The eighth identity follows from Corollary 4.2 with

$$\begin{aligned} (\mathbf{w}_h, \mathbf{u}_h, \mathbf{p}_h) &= (\mathcal{W}_f, \mathcal{U}_f, \mathcal{P}_f), & (\widehat{\mathbf{w}}_h, \widehat{\mathbf{u}}_h, \widehat{\mathbf{p}}_h, \mathbf{f}) &= (0, \widehat{\mathbf{U}}_f, 0, \mathbf{f}), \\ (\mathbf{w}'_h, \mathbf{u}'_h, \mathbf{p}'_h) &= (\mathcal{W}_\psi, \mathcal{U}_\psi, \mathcal{P}_\psi), & (\widehat{\mathbf{w}}'_h, \widehat{\mathbf{u}}'_h, \widehat{\mathbf{p}}'_h, \mathbf{f}') &= (0, \widehat{\mathbf{U}}_\psi, \psi, \mathbf{0}). \quad \square \end{aligned}$$

Proof of Theorem 3.8. By the jump conditions (3.23b) and (3.23c),

$$\begin{aligned} -\langle [(\widehat{\mathbf{U}}_{\gamma_t} + \widehat{\mathbf{U}}_\rho + \widehat{\mathbf{U}}_\phi) \times \mathbf{n}], \delta_t \rangle_{\mathcal{E}_h} &= \langle [(\widehat{\mathbf{U}}_f \times \mathbf{n})], \delta_t \rangle_{\mathcal{E}_h} - \langle \mathbf{g} \times \mathbf{n}, \delta_t \rangle_{\partial\Omega}, \\ -\langle [(\widehat{\mathbf{U}}_{\gamma_t} + \widehat{\mathbf{U}}_\rho + \widehat{\mathbf{U}}_\phi) \cdot \mathbf{n}], \psi \rangle_{\mathcal{E}_h} &= \langle [(\widehat{\mathbf{U}}_f \cdot \mathbf{n})], \psi \rangle_{\mathcal{E}_h} - \langle \mathbf{g} \cdot \mathbf{n}, \psi \rangle_{\partial\Omega}. \end{aligned}$$

By Lemma 4.6, we have that

$$\begin{aligned} -\langle [(\widehat{\mathbf{U}}_{\gamma_t} \times \mathbf{n})], \delta_t \rangle_{\mathcal{E}_h} &= a_h(\gamma_t, \delta_t), & -\langle [(\widehat{\mathbf{U}}_{\gamma_t} \cdot \mathbf{n})], \psi \rangle_{\mathcal{E}_h} &= b_h(\psi, \gamma_t), \\ -\langle [(\widehat{\mathbf{U}}_\rho \times \mathbf{n})], \delta_t \rangle_{\mathcal{E}_h} &= b_h(\rho, \delta_t), & -\langle [(\widehat{\mathbf{U}}_\rho \cdot \mathbf{n})], \psi \rangle_{\mathcal{E}_h} &= d_h(\rho, \psi), \\ -\langle [(\widehat{\mathbf{U}}_\phi \times \mathbf{u})], \delta_t \rangle_{\mathcal{E}_h} &= c_h(\phi, \delta_t), & -\langle [(\widehat{\mathbf{U}}_\phi \cdot \mathbf{n})], \psi \rangle_{\mathcal{E}_h} &= e_h(\phi, \psi), \end{aligned}$$

and that

$$\begin{aligned} \langle [(\widehat{\mathbf{U}}_f \times \mathbf{n})], \delta_t \rangle_{\mathcal{E}_h} - \langle \mathbf{g} \times \mathbf{n}, \delta_t \rangle_{\partial\Omega} &= \ell_1(\delta_t), \\ \langle [(\widehat{\mathbf{U}}_f \cdot \mathbf{n})], \psi \rangle_{\mathcal{E}_h} - \langle \mathbf{g} \cdot \mathbf{n}, \psi \rangle_{\partial\Omega} &= \ell_2(\psi). \end{aligned}$$

The proof of Theorem 3.8 is now completed by a uniqueness argument as in the previous cases. \square

5. Concluding remarks. In this paper, we introduced a new HDG method for the Stokes system and showed four different ways of hybridizing it. In order for these methods to be competitive with previously known ones [14, 20, 18, 19, 12, 15, 3, 13], they need to be not only efficiently implemented, but also efficiently solved. We would like to emphasize that our characterization theorems are a first step towards such a goal since they shed light on the structure of the corresponding equations. However, we feel that a meaningful study of those equations deserves a separate paper. The design of efficient solvers for these methods constitutes work in progress.

Another subject that constitutes the subject of ongoing work is the analysis of the accuracy of the methods. A careful a priori error analysis of the HDG methods

should reveal the effect of the choice of the stabilization parameters τ_n and τ_t on their accuracy. Let us recall that, in the context of second-order elliptic problems, the HDG methods [10] were shown to be *more* accurate than all previously known DG methods when their stabilization parameters are suitably chosen. In particular, when using polynomial approximations of the same degree for both the solution and its gradient, *both* approximations were shown to converge with optimal order; see [4, 11]. It is thus reasonable to expect that by a proper choice of the parameters τ_n and τ_t , the HDG method using polynomial approximations of the same degree for the vorticity, velocity, and pressure will also converge optimally in *all* three variables. This is work in progress.

REFERENCES

- [1] D. N. ARNOLD, F. BREZZI, B. COCKBURN, AND L. D. MARINI, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.
- [2] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
- [3] J. CARRERO, B. COCKBURN, AND D. SCHÖTZAU, *Hybridized, globally divergence-free LDG methods. Part I: The Stokes problem*, Math. Comp., 75 (2006), pp. 533–563.
- [4] B. COCKBURN, B. DONG, AND J. GUZMÁN, *A superconvergent LDG-hybridizable Galerkin method for second-order elliptic problems*, Math. Comp., 77 (2008), pp. 1887–1916.
- [5] B. COCKBURN AND J. GOPALAKRISHNAN, *A characterization of hybridized mixed methods for second order elliptic problems*, SIAM J. Numer. Anal., 42 (2004), pp. 283–301.
- [6] B. COCKBURN AND J. GOPALAKRISHNAN, *Error analysis of variable degree mixed methods for elliptic problems via hybridization*, Math. Comp., 74 (2005), pp. 1653–1677.
- [7] B. COCKBURN AND J. GOPALAKRISHNAN, *Incompressible finite elements via hybridization. Part I: The Stokes system in two space dimensions*, SIAM J. Numer. Anal., 43 (2005), pp. 1627–1650.
- [8] B. COCKBURN AND J. GOPALAKRISHNAN, *Incompressible finite elements via hybridization. Part II: The Stokes system in three space dimensions*, SIAM J. Numer. Anal., 43 (2005), pp. 1651–1672.
- [9] B. COCKBURN AND J. GOPALAKRISHNAN, *New hybridization techniques*, GAMM-Mitt., 2 (2005), pp. 154–183.
- [10] B. COCKBURN, J. GOPALAKRISHNAN, AND R. LAZAROV, *Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second-order elliptic problems*, SIAM J. Numer. Anal., to appear.
- [11] B. COCKBURN, J. GUZMÁN, AND H. WANG, *Superconvergent discontinuous Galerkin methods for second-order elliptic problems*, Math. Comp., 78 (2009), pp. 1–24.
- [12] B. COCKBURN, G. KANSCHAT, AND D. SCHÖTZAU, *A locally conservative LDG method for the incompressible Navier-Stokes equations*, Math. Comp., 74 (2005), pp. 1067–1095.
- [13] B. COCKBURN, G. KANSCHAT, AND D. SCHÖTZAU, *A note on discontinuous Galerkin divergence-free solutions of the Navier-Stokes equations*, J. Sci. Comput., 31 (2007), pp. 61–73.
- [14] B. COCKBURN, G. KANSCHAT, D. SCHÖTZAU, AND C. SCHWAB, *Local discontinuous Galerkin methods for the Stokes system*, SIAM J. Numer. Anal., 40 (2002), pp. 319–343.
- [15] V. GIRAULT, B. RIVIÈRE, AND M. F. WHEELER, *A discontinuous Galerkin method with non-overlapping domain decomposition for the Stokes and Navier-Stokes problems*, Math. Comp., 74 (2005), pp. 53–84.
- [16] M. D. GUNZBURGER, *Finite Element Methods for Viscous Incompressible Flows: A Guide to Theory, Practice and Algorithms*, Academic Press, New York, 1989.
- [17] J.-C. NÉDÉLEC, *Éléments finis mixtes incompressibles pour l'équation de Stokes dans \mathbf{R}^3* , Numer. Math., 39 (1982), pp. 97–112.
- [18] D. SCHÖTZAU, C. SCHWAB, AND A. TOSELLI, *Mixed hp-DGFEM for incompressible flows*, SIAM J. Numer. Anal., 40 (2003), pp. 2171–2194.
- [19] D. SCHÖTZAU, C. SCHWAB, AND A. TOSELLI, *Stabilized hp-DGFEM for incompressible flow*, Math. Models Methods Appl. Sci., 13 (2003), pp. 1413–1436.
- [20] A. TOSELLI, *hp-discontinuous Galerkin approximations for the Stokes problem*, Math. Models Methods Appl. Sci., 12 (2002), pp. 1565–1616.
- [21] J. WANG AND X. YE, *New finite element methods in computational fluid dynamics by $H(\text{div})$ elements*, SIAM J. Numer. Anal., 45 (2007), pp. 1269–1286.

NUMERICAL ANALYSIS OF A FINITE ELEMENT/VOLUME PENALTY METHOD*

BERTRAND MAURY†

Abstract. We present here some contributions to the numerical analysis of the penalty method in the finite element context. We are especially interested in the ability provided by this approach to use Cartesian, non boundary-fitted meshes to solve elliptic problems in complicated domain. In the spirit of fictitious domains, the initial problem is replaced by a penalized one, posed over a simply shaped domain which covers the original one. This method relies on two parameters, namely h (space-discretization parameter) and ε (penalty parameter). We propose here a general strategy to estimate the error in both parameters, and we present how it can be applied to various situations. We pay special attention to a scalar version of the rigid motion constraint for fluid-particle flows.

Key words. finite element method, penalty, Poisson's problem, error estimate

AMS subject classifications. 65N30, 65N12, 49M30

DOI. 10.1137/080712799

1. Introduction. Because of its conceptual simplicity and the fact that it is straightforward to implement, the penalty method has been widely used to incorporate constraints in numerical optimization. The general principle can be seen as a relaxed version of the following fact: given a proper functional J over a set X , and K a subset of X , minimizing J over K is equivalent to minimizing $J_K = J + I_K$ over X , where I_K is the indicatrix of K :

$$I_K(x) = \begin{cases} 0 & \text{if } x \in K, \\ +\infty & \text{if } x \notin K. \end{cases}$$

Assume now that K is defined as $K = \{x \in X, \Psi(x) = 0\}$, where Ψ is a nonnegative function; the penalty method consists in considering relaxed functionals J_ε defined as

$$J_\varepsilon = J + \frac{1}{\varepsilon}\Psi, \quad \varepsilon > 0.$$

By definition of K , the function Ψ/ε approaches I_K pointwisely:

$$\frac{1}{\varepsilon}\Psi(x) \longrightarrow I_K(x) \text{ as } \varepsilon \text{ goes to } 0 \quad \forall x \in X.$$

If J_ε admits a minimum u^ε , for any ε , one can expect u^ε to approach a (or *the*) minimizer of J over K , if it exists.

In the finite element context, some u_h^ε is computed as the solution to a finite dimensional problem, where h is a space-discretization parameter. The work we present here is motivated by the fact that, even if the penalty method for the continuous problem is convergent and the discretization procedure is sound, the rate of convergence of u_h^ε toward the exact solution is not straightforward to obtain. A huge literature is

*Received by the editors January 9, 2008; accepted for publication (in revised form) November 6, 2008; published electronically February 19, 2009.

<http://www.siam.org/journals/sinum/47-2/71279.html>

†Laboratoire de Mathématiques, Université Paris-Sud, 91405 Orsay Cedex, France (Bertrand.Maury@math.u-psud.fr).

dedicated to the situation where the constraint is distributed over the domain, like the divergence-free constraint for incompressible Stokes flows (see [BF91, GR79]). In this context, the penalty approach makes it possible to use mixed finite element methods which do not fulfill the so-called Babuska–Brezzi–Ladyzhenskaya (or inf-sup) condition. The penalty approach is also commonly used to prescribe (possibly nonhomogeneous) Dirichlet boundary conditions on a boundary. The pioneering papers [Nit71] and [Bab73] already addressed in the early 70’s the problem of error estimation with respect to both parameters h and ε . Those works have been widely used since then, and this area has recently experienced a regain of interest, triggered by problems arising in domain decomposition (see, e.g., [BHS03]), discontinuous Galerkin methods [BE07], or handling of discontinuities for elliptic problems with discontinuous coefficients [HH02].

We will focus here on another type of constraints, namely geometrical ones: we are interested in solving an elliptic problem on a domain $\Omega \setminus \overline{\mathcal{O}}$, where Ω is a simply shaped domain (e.g., a rectangle) and \mathcal{O} a set of holes, and we aim at replacing it by a new problem posed over the global domain Ω . The simplest situation one may consider consists in solving a Poisson problem in a perforated, rectangular domain Ω , with homogeneous Dirichlet boundary conditions on the holes and over the external boundary. In the purpose of using a Cartesian mesh which covers the whole domain (which can be of great interest if the holes are intended to move), it is natural to consider the penalized version of the problem, which consists in minimizing (\mathcal{O} designs the subdomain covered by the holes)

$$\frac{1}{2} \int_{\Omega} |\nabla v|^2 - \int_{\Omega} f v + \frac{1}{2\varepsilon} \int_{\mathcal{O}} (v^2 + |\nabla v|^2)$$

over $H_0^1(\Omega)$. Another situation where the penalty approach has already proved to be quite efficient is the modeling of fluid-particle flows (see [RPVC05] or [JLM05]). The scalar version of this problem, which we shall address in detail in the following pages, consists in minimizing the standard functional

$$J(v) = \frac{1}{2} \int_{\Omega} |\nabla v|^2 - \int_{\Omega} f v$$

over all those functions which are constant on each connected component of the set of holes \mathcal{O} . Again, the constraint is easily relaxed by adding to J a term which penalizes the H^1 seminorm of v over \mathcal{O} .

Two points advocate for the use of this approach:

1. The use of a Cartesian mesh makes this approach quite easy to implement: both cases reduce to a few lines of instructions within user-friendly finite element solvers like Freefem++ [FFp] for two-dimensional problems, or Freefem3D [FFp] for three-dimensional ones. Note that the penalty terms do not preserve the spectrum of the discrete Laplacian matrix, which prevents us from using standard fast solvers like fast Fourier transform (to the contrary of Lagrange multiplier based fictitious domain methods [PG02, GG95], which do preserve the structure of the matrix, at the price of an iterative algorithm on the Lagrange multipliers). A harmful effect upon the condition number of the solution matrix is furthermore to be expected. Yet, as the penalty parameter does not need to be taken too small, the method remains quite competitive for reasonably sized problems.

2. This method provides, with no extra computational cost, an approximation of the Lagrange multiplier associated with the constraint, which is of great significance from the modeling standpoint in many situations. For example, in the first situation we considered, which can be seen as the stationary heat equation, it is quite straightforward that, if we denote by u^ε the solution to the discretized problem, $\xi^\varepsilon \in H^{-1}$ defined as

$$\langle \xi^\varepsilon, v \rangle = \frac{1}{\varepsilon} \int_{\mathcal{O}} (u^\varepsilon v + \nabla u^\varepsilon \cdot \nabla v)$$

approximates the heat source which is necessary to fulfill the constraint. We shall establish that this natural outcome of the method is still provided by the discretized/penalized version. Note that this property has already been used to handle numerically the motion of a three-dimensional turbine in a Navier–Stokes fluid (see [DPM07]).

As for the theoretical analysis of the method, the error due to the fact that the mesh is not boundary fitted is analyzed in [AR08, RAB07]. See also [SMSTT05] for similar estimates used to establish the convergence of a method to handle the motion of a rigid motion in the limit $\varepsilon = 0$. Yet, to the best of our knowledge, a full error estimate (simultaneous convergence of h and ε toward 0) has not yet been provided for the type of volume penalty approach we propose here. We aim here at showing that the global error can be controlled, as expected, by the sum of the penalty error and the space-discretization error, under quite general assumptions.

This paper is organized as follows: in section 2, we recall some standard properties of the penalty method in the framework of constrained quadratic minimization, including some general facts about the space discretization of those problems. Section 3 is devoted to the main result: an abstract estimate for the primal and the dual parts of the discretized/penalized problem. The next section is concerned with a model problem, in the spirit of fluid-particle flows, for which we present in detail how the abstract estimate can be applied. Finally, we present in section 5 some other typical situations where the abstract estimate can be used.

2. Preliminaries, abstract framework.

2.1. Continuous problem. We recall here some standard properties concerning the penalty method applied to infinite dimensional problems. Most of those properties are established in [BF91], with a slightly different formalism. We consider the following set of assumptions:

$$(2.1) \quad \left. \begin{array}{l} V \text{ is a Hilbert space, } \varphi \in V', \\ a(\cdot, \cdot) \text{ bilinear, symmetric, continuous, elliptic } (a(v, v) \geq \alpha |v|^2), \\ b(\cdot, \cdot) \text{ bilinear, symmetric, continuous, nonnegative,} \\ K = \{u \in V, b(u, u) = 0\} = \ker b, \\ J(v) = \frac{1}{2}a(v, v) - \langle \varphi, v \rangle, \quad u = \arg \min_K J, \\ J_\varepsilon(v) = \frac{1}{2}a(v, v) + \frac{1}{2\varepsilon}b(v, v) - \langle \varphi, v \rangle, \quad u^\varepsilon = \arg \min_V J_\varepsilon. \end{array} \right\}$$

PROPOSITION 2.1. *Under assumptions (2.1), the solution u^ε to the penalized problem converges to u .*

Proof. As the family (J_ε) is uniformly elliptic, $|u^\varepsilon|$ is bounded. We extract a subsequence, still denoted by (u^ε) , which converges weakly to some $z \in V$. As $J_\varepsilon \geq J$ and $b(u, u) = 0$, we have

$$(2.2) \quad J(u^\varepsilon) \leq J_\varepsilon(u^\varepsilon) \leq J_\varepsilon(u) = J(u) \quad \forall \varepsilon > 0,$$

so that (J is convex and continuous) $J(z) \leq \liminf J(u^\varepsilon) \leq J(u)$. As

$$J(u^\varepsilon) + \frac{1}{2\varepsilon}b(u^\varepsilon, u^\varepsilon) \leq J(u),$$

$b(u^\varepsilon, u^\varepsilon)/\varepsilon$ is bounded, so that $b(u^\varepsilon, u^\varepsilon)$ goes to 0 with ε . Consequently, it holds that $0 \leq b(z, z) \leq \liminf b(u^\varepsilon, u^\varepsilon) = 0$, which implies $z \in K$, so that $z = u$.

To establish the strong character of the convergence, we show that u^ε converges toward u for the norm associated with $a(\cdot, \cdot)$, which is equivalent to the original norm. As u^ε converges weakly to u for this scalar product ($a(u^\varepsilon, v) \rightarrow a(u, v)$ for any $v \in V$), it is sufficient to establish the convergence of $|u^\varepsilon|_a = a(u^\varepsilon, u^\varepsilon)^{1/2}$ toward $|u|_a$. First, $|u|_a \leq \liminf |u^\varepsilon|_a$, and the other inequality comes from (2.2):

$$\frac{1}{2}a(u^\varepsilon, u^\varepsilon) - \langle \varphi, u^\varepsilon \rangle \leq \frac{1}{2}a(u, u) - \langle \varphi, u \rangle,$$

so that $\limsup |u^\varepsilon|_a \leq |u|_a$. □

The proposition does not say anything about the rate of convergence, and it can be very poor, as the following example illustrates.

Example 2.1. Consider $I =]0, 1[$, $V = H^1(I)$, and the problem which consists in minimizing the functional

$$J(v) = \frac{1}{2} \int_I |v'|^2$$

over $K = \{v \in V, v(x) = 0 \text{ a.e. in } \mathcal{O} =]0, 1/2[\}$. The solution to that problem is obviously $u = \max(0, 2(x - 1/2))$. Now let us denote by u^ε the minimum of the penalized functional

$$J_\varepsilon = \frac{1}{2} \int_I |u'|^2 + \frac{1}{2\varepsilon} \int_{\mathcal{O}} |u|^2.$$

The solution to the penalized problem can be computed exactly:

$$u^\varepsilon = k_\varepsilon(x) \operatorname{sh} \left(\frac{x}{\sqrt{\varepsilon}} \right) \text{ in }]0, 1/2[\text{ with } k_\varepsilon(x) = \left(\operatorname{sh} \left(\frac{x}{\sqrt{\varepsilon}} \right) + \frac{1}{2\sqrt{\varepsilon}} \operatorname{ch} \left(\frac{x}{\sqrt{\varepsilon}} \right) \right)^{-1},$$

and u^ε affine in $]1/2, 1[$, continuous at $1/2$. This makes it possible to estimate $|u^\varepsilon - u|$, which turns out to behave like $\varepsilon^{1/4}$.

Yet, in many situations, convergence can be shown to be of order 1, given some assumptions are fulfilled. Let us introduce $\xi \in V'$ as the unique linear functional such that

$$(2.3) \quad a(u, v) + \langle \xi, v \rangle = \langle \varphi, v \rangle \quad \forall v \in V.$$

Before stating the first order convergence result, we show here that the penalty method provides an approximation of ξ .

PROPOSITION 2.2. Let $\xi^\varepsilon \in V'$ be defined by

$$v \in V \longmapsto \langle \xi^\varepsilon, v \rangle = \frac{1}{\varepsilon} b(u^\varepsilon, v).$$

Then ξ^ε converges (strongly) to ξ in V' , at least as fast as u^ε converges to u .

Proof. The variational formulation of the penalized problem reads

$$(2.4) \quad a(u^\varepsilon, v) + \frac{1}{\varepsilon} b(u^\varepsilon, v) = \langle \varphi, v \rangle \quad \forall v \in V.$$

The result is then a direct consequence of the identity which we obtain by subtracting (2.3) and (2.4):

$$\langle \xi, v \rangle - \frac{1}{\varepsilon} b(u^\varepsilon, v) = a(u - u^\varepsilon, v) \quad \forall v \in V,$$

which yields $\|\xi - \xi^\varepsilon\|_{V'} \leq C|u - u^\varepsilon|$. \square

Let us now establish the first order convergence, provided an extra compatibility condition between $b(\cdot, \cdot)$ and ξ is met.

PROPOSITION 2.3. Under assumptions (2.1), we assume in addition that there exists $\tilde{\xi} \in V$ such that $b(\tilde{\xi}, v) = \langle \xi, v \rangle$ for all $v \in V$. Then $|u^\varepsilon - u| = \mathcal{O}(\varepsilon)$.

Proof. First of all, notice that it is possible to pick $\tilde{\xi}$ in K^\perp (if not, we project it onto K^\perp). Now following the idea which is proposed in [Bab73] in a slightly different context (see the proof of Thm. 3.2 therein), we introduce

$$R_\varepsilon(v) = \frac{1}{2} a(u - v, u - v) + \frac{1}{2\varepsilon} b(\varepsilon\tilde{\xi} - v, \varepsilon\tilde{\xi} - v),$$

which can be written

$$R_\varepsilon(v) = \frac{1}{2} a(u, u) + \frac{\varepsilon}{2} b(\tilde{\xi}, \tilde{\xi}) + \frac{1}{2} a(v, v) + \frac{1}{2\varepsilon} b(v, v) - a(u, v) - b(\tilde{\xi}, v).$$

As $b(\tilde{\xi}, v) = \langle \xi, v \rangle$ and $-a(u, v) - \langle \xi, v \rangle = -\langle \varphi, v \rangle$, the functional R_ε is equal to J_ε up to a constant. Therefore minimizing R_ε amounts to minimizing J_ε . Let us now introduce $w = \varepsilon\tilde{\xi} + u$. We have

$$R_\varepsilon(w) = \frac{\varepsilon^2}{2} a(\tilde{\xi}, \tilde{\xi}) + 0 \quad \text{because } u \in K = \ker b,$$

so that $|R_\varepsilon(w)| \leq C\varepsilon^2$. As u^ε minimizes R_ε ,

$$0 \leq R_\varepsilon(u^\varepsilon) = \frac{1}{2} a(u - u^\varepsilon, u - u^\varepsilon) + \frac{1}{2\varepsilon} b(\varepsilon\tilde{\xi} - u^\varepsilon, \varepsilon\tilde{\xi} - u^\varepsilon) \leq C\varepsilon^2,$$

from which we deduce, as $a(\cdot, \cdot)$ is elliptic, $|u - u^\varepsilon| = \mathcal{O}(\varepsilon)$. \square

COROLLARY 2.4. Under assumptions (2.1), we assume in addition that $b(\cdot, \cdot)$ can be written $b(u, v) = (Bu, Bv)$, where B is a linear continuous operator onto a Hilbert space Λ , with closed range. Then $|u^\varepsilon - u| = \mathcal{O}(\varepsilon)$.

Proof. Let us show that the assumption of Proposition 2.3 is met. It is sufficient to prove that any $\xi \in V'$ which vanishes over K identifies through $b(\cdot, \cdot)$ with some $\tilde{\xi} \in V$; i.e., there exists $\tilde{\xi} \in V$ such that

$$\langle \xi, v \rangle = b(\tilde{\xi}, v) \quad \forall v \in V.$$

Note that, as ξ vanishes over K , it can be seen as a linear functional defined on K^\perp , so that it is equivalent to establish that $T : V \rightarrow (K^\perp)'$ defined by

$$\tilde{\xi} \mapsto \xi : \langle \xi, v \rangle = b(\tilde{\xi}, v) \quad \forall v \in K^\perp$$

is surjective. We denote by $T^* \in \mathcal{L}(K^\perp, V)$ the adjoint of T . For all $w \in K^\perp$,

$$|T^*w| = \sup_{v \neq 0} \frac{(T^*w, v)}{|v|} = \sup_{v \neq 0} \frac{b(w, v)}{|v|} = \sup_{v \neq 0} \frac{(Bw, Bv)}{|v|} \geq \frac{|Bw|^2}{|w|}.$$

As B has closed range, $|Bw| \geq C|w|$ for all w in $(\ker B)^\perp = K^\perp$, so that

$$|T^*w| \geq C^2|w| \quad \forall w \in K^\perp,$$

from which we conclude that T is surjective. \square

Remark 2.1. Note that Proposition 2.3 is strictly stronger than its corollary. Indeed, consider the handling of homogeneous Dirichlet boundary conditions by penalty: $V = H^1(\Omega)$, where Ω is a smooth, bounded domain, $a(u, v) = \int \nabla u \cdot \nabla v$, and $\langle \varphi, v \rangle = \int f v$, where f is in $L^2(\Omega)$, and $b(v, v) = \int_{\partial\Omega} v^2$. In this situation the corollary cannot be used, because the trace operator from $H^1(\Omega)$ onto $L^2(\partial\Omega)$ does not have a close range. On the other hand one can establish that

$$\langle \xi, v \rangle = \int_{\partial\Omega} \frac{\partial u}{\partial n} v,$$

and, as the solution u is regular ($u \in H^2(\Omega)$), its normal derivative (in $H^{1/2}(\partial\Omega)$) can be built as the trace of a function $\tilde{\xi}$ in $H^1(\Omega)$, so that Proposition 2.3 holds true.

We conclude this section by some considerations concerning the saddle-point formulation of the constrained problem, which will be useful in the following. We consider again the closed situation.

PROPOSITION 2.5. *Under the assumptions of Corollary 2.4, there exists $\lambda \in \Lambda$ such that*

$$(2.5) \quad a(u, v) + (\lambda, Bv) = \langle \varphi, v \rangle \quad \forall v \in V.$$

The solution is unique in $B(V)$ (which identifies with $\Lambda/\ker B^$).*

Proof. The proof of this standard property can be found in [BF91]. In fact, it has just been established in the proof of Corollary 2.4: λ is simply $B\tilde{\xi}$. Uniqueness is straightforward. \square

PROPOSITION 2.6. *Under the assumptions of Proposition 2.5 (assumptions (2.1) and $B(V)$ is closed), we introduce*

$$\lambda^\varepsilon = \frac{1}{\varepsilon} B u^\varepsilon.$$

Then $|\lambda^\varepsilon - \lambda| = \mathcal{O}(\varepsilon)$, where λ is the unique solution of (2.5) in $B(V)$.

Proof. Subtracting the variational formulations for u and u^ε , we get

$$(\lambda^\varepsilon - \lambda, Bv) = a(u^\varepsilon - u, v) \quad \forall v \in V.$$

Now, as the range of B is closed, and $\lambda^\varepsilon - \lambda \in B(V) = (\ker B^*)^\perp$, we have the inf-sup condition (see, e.g., [BF91])

$$\sup_{v \in V} \frac{(\lambda^\varepsilon - \lambda, Bv)}{|v|} \geq \beta |\lambda^\varepsilon - \lambda|,$$

so that

$$\beta |\lambda^\varepsilon - \lambda| \leq \sup \frac{(\lambda^\varepsilon - \lambda, Bv)}{|v|} = \sup \frac{a(u^\varepsilon - u, v)}{|v|} \leq \|a\| |u^\varepsilon - u|,$$

which ensures the first order convergence thanks to Corollary 2.4. \square

COROLLARY 2.7. *For any $z \in V$ such that $Bz = \lambda$, there exists a sequence (v^ε) in $\ker B$ such that*

$$\left| \frac{u^\varepsilon}{\varepsilon} - v^\varepsilon - z \right| = \mathcal{O}(\varepsilon).$$

Proof. This is a direct consequence of the fact that, $B(V)$ being closed, the restriction of B to $\ker B^\perp$ is a bicontinuous bijection between $\ker B^\perp$ and $B(V)$. The convergence is therefore obtained by taking $v^\varepsilon = P_{\ker B}(u^\varepsilon/\varepsilon - z)$. \square

2.2. Discretized problem. We consider now a family $(V_h)_h$ of inner approximation spaces $(V_h \subset V)$ and the associated penalized/discretized problems

$$(2.6) \quad \begin{cases} \text{Find } u_h^\varepsilon \in V_h \text{ such that } J^\varepsilon(u_h^\varepsilon) = \inf_{v_h \in V_h} J^\varepsilon(v_h), \\ J^\varepsilon(v_h) = \frac{1}{2}a(v_h, v_h) + \frac{1}{2\varepsilon}b(v_h, v_h) - \langle \varphi, v_h \rangle. \end{cases}$$

As far as we know, there does not exist any general theory which would give an upper bound for the error $|u - u_h^\varepsilon|$ as the sum of a discretization error (typically h or $h^{1/2}$ for volume penalty, depending on whether the mesh is boundary-fitted or not), and a penalty error (typically ε for closed-range penalty terms, possibly poorer in general situations, as in Example 2.1). We propose here two general properties which are direct consequences of standard arguments. They are suboptimal in the sense that neither of them is optimal from both standpoints (discretization and penalty), but, at least, they make it possible to recover the behavior in extreme situations (when ε goes to 0 much quicker than h , and the opposite).

The first proposition uses the following lemma.

LEMMA 2.8. *Under assumptions (2.1), there exists $C > 0$ such that*

$$b(u^\varepsilon, u^\varepsilon) \leq C\varepsilon |u - u^\varepsilon|.$$

Proof. By definition of u^ε ,

$$J_\varepsilon(u^\varepsilon) = \frac{1}{2}a(u^\varepsilon, u^\varepsilon) - \langle \varphi, u^\varepsilon \rangle + \frac{1}{2\varepsilon}b(u^\varepsilon, u^\varepsilon) \leq J_\varepsilon(u) = \frac{1}{2}a(u, u) - \langle \varphi, u \rangle,$$

so that

$$\begin{aligned} 0 \leq \frac{1}{2\varepsilon}b(u^\varepsilon, u^\varepsilon) &\leq \frac{1}{2}a(u, u) - \frac{1}{2}a(u^\varepsilon, u^\varepsilon) + \langle \varphi, u^\varepsilon - u \rangle \\ &\leq \frac{1}{2}a(u + u^\varepsilon, u - u^\varepsilon) + \langle \varphi, u^\varepsilon - u \rangle, \end{aligned}$$

which yields the estimate by continuity of $a(\cdot, \cdot)$ and φ . \square

PROPOSITION 2.9. *Under assumptions (2.1), we denote by u_h^ε the solution to problem (2.6). Then*

$$|u_h^\varepsilon - u| \leq C \left(\min_{v_h \in V_h \cap K} |v_h - u| + \sqrt{|u^\varepsilon - u|} \right).$$

Proof. As u_h^ε minimizes $a(v - u^\varepsilon, v - u^\varepsilon) + b(v - u^\varepsilon, v - u^\varepsilon)/\varepsilon$ over V_h ,

$$\begin{aligned} \alpha |u_h^\varepsilon - u^\varepsilon|^2 &\leq a(u_h^\varepsilon - u^\varepsilon, u_h^\varepsilon - u^\varepsilon) \\ &\leq a(u_h^\varepsilon - u^\varepsilon, u_h^\varepsilon - u^\varepsilon) + \frac{1}{\varepsilon} b(u_h^\varepsilon - u^\varepsilon, u_h^\varepsilon - u^\varepsilon) \\ &\leq \min_{v_h \in V_h} \left(a(v_h - u^\varepsilon, v_h - u^\varepsilon) + \frac{1}{\varepsilon} b(v_h - u^\varepsilon, v_h - u^\varepsilon) \right) \\ &\leq \min_{v_h \in V_h \cap K} \left(a(v_h - u^\varepsilon, v_h - u^\varepsilon) + \frac{1}{\varepsilon} b(v_h - u^\varepsilon, v_h - u^\varepsilon) \right). \end{aligned}$$

As v_h is in K , the second term is $b(u^\varepsilon, u^\varepsilon)/\varepsilon$, which is bounded by $C|u^\varepsilon - u|$ (by Lemma 2.8). Finally, we get

$$|u_h^\varepsilon - u^\varepsilon| \leq C \left(\min_{v_h \in V_h \cap K} |v_h - u^\varepsilon| + \sqrt{|u^\varepsilon - u|} \right),$$

from which we conclude. \square

PROPOSITION 2.10. *Under assumptions (2.1), $V_h \subset V$, and u_h^ε being the solution to (2.6), it holds that*

$$|u_h^\varepsilon - u| \leq \frac{C}{\sqrt{\varepsilon}} \inf_{v_h \in V_h} |u^\varepsilon - v_h| + |u^\varepsilon - u|.$$

Proof. One has

$$|u_h^\varepsilon - u| \leq |u_h^\varepsilon - u^\varepsilon| + |u^\varepsilon - u|,$$

and we control the first term by Céa’s lemma applied to the bilinear form $a + b/\varepsilon$, whose ellipticity constant behaves like $1/\varepsilon$. \square

The following example illustrates how those estimates can be used in practice.

Example 2.2. The simplest example of penalty formulation one may think about is the following: the constraint to vanish on the boundary of a subdomain $\mathcal{O} \subset \subset \Omega$ is handled by minimizing the functional

$$(2.7) \quad J_\varepsilon(v) = \frac{1}{2} \int_\Omega |\nabla v|^2 - \int_\Omega f v + \frac{1}{2\varepsilon} \int_\mathcal{O} u^2.$$

Now considering the L^2 penalty method in \mathcal{O} , if we admit the $\varepsilon^{1/4}$ convergence of $|u^\varepsilon - u|$, Proposition 2.9 provides an estimate in $h^{1/2} + \varepsilon^{1/8}$. This estimate is optimal in h : the natural space discretization order is obtained if ε is small enough ($\varepsilon = h^4$ in the present case).

Symmetrically, the natural order in ε can be recovered if h is small enough: Indeed, if we admit that u^ε can be approximated at the same order as u over Ω , which is $1/2$, then the choice $\varepsilon = h^{4/3}$ in Proposition 2.10 gives

$$|u_h^\varepsilon - u| \leq \frac{C}{\varepsilon^{1/2}} \varepsilon^{3/4} + \varepsilon^{1/4} = \mathcal{O}(\varepsilon^{1/4}).$$

Note that if we replace u^2 by $u^2 + |\nabla u|^2$ in the integral over \mathcal{O} in (2.7), assumptions of Corollary 2.4 are fulfilled, so that convergence holds at the first order in ε . As a consequence, $|u - u_h^\varepsilon|$ is bounded by $C(h^{1/2} + \varepsilon^{1/2})$ (by Proposition 2.9), which suggests the choice $\varepsilon = h$.

3. Full error estimate. As shall be made clear below, a full and optimal error estimate calls for a uniform discrete inf-sup condition. In the case of a nonconforming mesh, it appears immediately that the penalty term has to be modified. To anticipate this difficulty, we introduce a modified version of B , namely B_h , in this abstract approach. No assumption is made a priori on B_h in terms of approximation properties, but the estimate we establish below will not express any convergence property unless B_h approaches B in some sense.

Besides (2.1), we consider the following set of additional assumptions and notation:

$$(3.1) \quad \left. \begin{aligned} &b(v, v) = (Bv, Bv), \text{ where } B \in \mathcal{L}(V, \Lambda) \text{ has a closed range,} \\ &(V_h)_h \text{ family of approximation spaces, } V_h \subset V, \\ &B_h \in \mathcal{L}(V, \Lambda), \ker B \subset \ker B_h, \|B_h\| \text{ bounded, } \Lambda_h = B_h(V_h), \\ &J_h^\varepsilon(v_h) = J(v_h) + \frac{1}{\varepsilon}(B_h v_h, B_h v_h), \\ &u_h^\varepsilon = \arg \min_{V_h} J_h^\varepsilon, \lambda_h^\varepsilon = \frac{1}{\varepsilon} B_h u_h^\varepsilon \in \Lambda_h, \\ &\sup_{v_h \in V_h} \frac{(B_h v_h, \lambda_h)}{|v_h|} \geq \beta |\lambda_h|_{\Lambda_h} \quad \forall \lambda_h \in \Lambda_h. \end{aligned} \right\}$$

THEOREM 3.1 (primal/dual error estimate). *Under assumptions (2.1) and (3.1), we have the following error estimate:*

$$(3.2) \quad |u - u_h^\varepsilon| + |\lambda - \lambda_h^\varepsilon| \leq C \left(\varepsilon + \inf_{\tilde{u}_h \in V_h} |\tilde{u}_h - u| + \inf_{\tilde{\lambda}_h \in \Lambda_h} |\tilde{\lambda}_h - \lambda| + |(B_h^* - B^*)\lambda| + |(B_h - B)z| \right),$$

where z is such that $\lambda = Bz$.

Proof. The proof relies on some general properties of the continuous penalty method which we established in the beginning of this section, and an abstract stability estimate for saddle-point-like problems with stabilization (see Proposition 3.2 below).

First of all, note that, as the range of B is closed, the convergence of u^ε toward u holds at the first order (by Corollary 2.4). As another consequence, $\lambda^\varepsilon = Bu^\varepsilon/\varepsilon$ is such that $|\lambda - \lambda^\varepsilon| = \mathcal{O}(\varepsilon)$ (by Proposition 2.6).

We write the continuous penalized problem

$$\begin{cases} a(u^\varepsilon, v) + (\lambda^\varepsilon, Bv) = \langle \varphi, v \rangle & \forall v \in V, \\ (Bu^\varepsilon, \mu) - \varepsilon(\lambda^\varepsilon, \mu) = 0 & \forall \mu \in \Lambda \end{cases}$$

and the discrete penalized problem in a saddle-point form

$$\begin{cases} a(u_h^\varepsilon, v_h) + (\lambda_h^\varepsilon, B_h v_h) = \langle \varphi, v_h \rangle & \forall v_h \in V_h, \\ (B_h u_h^\varepsilon, \mu_h) - \varepsilon(\lambda_h^\varepsilon, \mu_h) = 0 & \forall \mu_h \in \Lambda_h. \end{cases}$$

As Λ_h is exactly $B_h(V_h)$, this problem admits a unique solution $(u_h^\varepsilon, \lambda_h^\varepsilon)$ (see Proposition 2.5). For any $(\tilde{u}_h, \tilde{\lambda}_h) \in V_h \times \Lambda_h$, $v_h \in V_h$, $\mu_h \in \Lambda_h$,

$$\begin{cases} a(\tilde{u}_h - u_h^\varepsilon, v_h) + (\tilde{\lambda}_h - \lambda_h^\varepsilon, B_h v_h) = a(\tilde{u}_h - u^\varepsilon, v_h) + (\tilde{\lambda}_h - \lambda^\varepsilon, B_h v_h) \\ \quad + \langle (B_h^* - B^*)\lambda^\varepsilon, v_h \rangle, \\ (B_h(\tilde{u}_h - u_h^\varepsilon), \mu_h) - \varepsilon(\tilde{\lambda}_h - \lambda_h^\varepsilon, \mu_h) = (B_h(\tilde{u}_h - u^\varepsilon), \mu_h) - \varepsilon(\tilde{\lambda}_h - \lambda^\varepsilon, \mu_h) \\ \quad + \langle (B_h - B)u^\varepsilon, \mu_h \rangle. \end{cases}$$

Our purpose is to use Proposition 3.2 (V_h and Λ_h play the role of V and Λ in the proposition, respectively) with

$$(3.3) \quad \langle \varphi, v_h \rangle = a(\tilde{u}_h - u^\varepsilon, v_h) + (\tilde{\lambda}_h - \lambda^\varepsilon, B_h v_h) + \langle (B_h^* - B^*)\lambda^\varepsilon, v_h \rangle,$$

$$(3.4) \quad \langle \Psi, \mu_h \rangle = (B_h(\tilde{u}_h - u^\varepsilon), \mu_h) - \varepsilon(\tilde{\lambda}_h - \lambda^\varepsilon, \mu_h) + ((B_h - B)u^\varepsilon, \mu_h).$$

The last term of (3.3) is transformed as follows:

$$(B_h^* - B^*)\lambda^\varepsilon = (B_h^* - B^*)\lambda + (B_h^* - B^*)(\lambda^\varepsilon - \lambda),$$

where $\lambda \in B(V)$ is the exact Lagrange multiplier defined by Proposition 2.5. So, defining

$$c(\mu, \mu') = \varepsilon(\mu, \mu'), \quad w = \tilde{u}_h - u^\varepsilon, \quad \gamma = -(\tilde{\lambda}_h - \lambda^\varepsilon) + (B_h - B)\frac{u^\varepsilon}{\varepsilon}$$

(see (3.7) for the meaning of w and γ), Proposition 3.2 ensures existence of a constant $C > 0$ (which does not depend on h) such that $|\tilde{u}_h - u_h^\varepsilon| + |\tilde{\lambda}_h - \lambda_h^\varepsilon|$ is less than

$$C \left(|\tilde{u}_h - u^\varepsilon| + |\tilde{\lambda}_h - \lambda^\varepsilon| + \|(B_h^* - B^*)\lambda\| + |\gamma| \right).$$

The second contribution to γ can be written, thanks to Corollary 2.7 and the fact that $\ker B \subset \ker B_h$,

$$(B_h - B)\frac{u^\varepsilon}{\varepsilon} = (B_h - B)\left(\frac{u^\varepsilon}{\varepsilon} - v^\varepsilon - z\right) + (B_h - B)z,$$

where $v^\varepsilon \in \ker B$, and z is such that $Bz = \lambda$, which yields

$$|\gamma| \leq |\tilde{\lambda}_h - \lambda^\varepsilon| + \mathcal{O}(\varepsilon) + |(B_h - B)z|.$$

We finally obtain that $|u^\varepsilon - u_h^\varepsilon| + |\lambda^\varepsilon - \lambda_h^\varepsilon|$ is less than

$$C \left(\inf_{\tilde{u}_h \in V_h} |\tilde{u}_h - u^\varepsilon| + \inf_{\tilde{\lambda}_h \in \Lambda_h} |\tilde{\lambda}_h - \lambda^\varepsilon| + \|(B_h^* - B^*)\lambda\| + \varepsilon + |(B_h - B)z| \right),$$

so that, by eliminating u^ε in the left-hand side, and again using $|u^\varepsilon - u| = \mathcal{O}(\varepsilon)$ and $|\lambda^\varepsilon - \lambda| = \mathcal{O}(\varepsilon)$ (see Corollary 2.4 and Proposition 2.6), we obtain the error estimate. \square

PROPOSITION 3.2 (abstract stability estimate). *Let V and Λ be two Hilbert spaces, $B \in \mathcal{L}(V, \Lambda)$, $a(\cdot, \cdot)$ and $c(\cdot, \cdot)$ bilinear continuous functionals, which we suppose elliptic. Then the problem*

$$(3.5) \quad \begin{cases} a(u, v) + (\lambda, Bv) = \langle \varphi, v \rangle & \forall v \in V, \\ (Bu, \mu) - c(\lambda, \mu) = \langle \Psi, \mu \rangle & \forall \mu \in \Lambda \end{cases}$$

admits a unique solution $(u, \lambda) \in V \times \Lambda$. We assume furthermore that there exists a constant $\beta > 0$ such that¹

$$(3.6) \quad \beta |P_{(\ker B)^\perp} v| \leq |Bv|, \quad \sup_{v \in V} \frac{(\mu, Bv)}{|v|} \geq \beta \|\mu\|_{\Lambda / \ker B^*},$$

¹As the second inequality of (3.6) is a direct consequence of the first one, it could be suppressed. We keep both assumptions for clarity reasons.

that Ψ can be written

$$(3.7) \quad \langle \Psi, \mu \rangle = (Bw, \mu) + c(\gamma, \mu),$$

and finally that $c(\cdot, \cdot)$ verifies

$$(3.8) \quad \mu_1 \perp \mu_2 \implies c(\mu_1, \mu_2) = 0.$$

Then we have the following estimate:

$$(3.9) \quad |u| + |\lambda| \leq C(\|\varphi\| + |w| + |\gamma|),$$

where C is a locally bounded expression of $\|a\|, 1/\alpha, 1/\beta, \|B\|, \|c\|$ (α is the coercivity constant of $a(\cdot, \cdot)$). Note that C does not depend upon the coercivity constant of $c(\cdot, \cdot)$.

Proof. The first part of the proposition is trivial. With obvious notation, problem (3.5) can be written

$$(3.10) \quad \begin{cases} Au + B^*\lambda = \varphi, \\ Bu - M\lambda = \Psi, \end{cases}$$

so that (u, λ) is uniquely determined as

$$u = (A + B^*M^{-1}B)^{-1}(\varphi + B^*M^{-1}\Psi), \quad \lambda = M^{-1}(Bu - \Psi).$$

In order to get an upper bound of $|u|$ which does not degenerate with $c(\cdot, \cdot)$, we introduce, following [BF91],

$$(3.11) \quad u = \underbrace{u_0}_{\in \ker B} + \underbrace{u^\perp}_{\in (\ker B)^\perp}, \quad \lambda = \underbrace{\lambda_0}_{\in \ker B^*} + \underbrace{\lambda^\perp}_{\in (\ker B^*)^\perp}.$$

From (3.6) and the first line of (3.5), we have

$$(3.12) \quad \beta | \lambda^\perp | = \beta \| \lambda \|_{\Lambda / \ker B^*} \leq \sup \frac{(\lambda, Bv)}{|v|} \leq \|a\| |u| + \|\varphi\|.$$

From (3.6) again and the second line of (3.5), we get

$$(3.13) \quad \beta |u^\perp| = \beta |P_{(\ker B)^\perp} u| \leq |Bu| = \sup \frac{(Bu, \mu)}{|\mu|} \leq \|\Psi\| + \|c\|^{1/2} c(\lambda, \lambda)^{1/2}.$$

From the ellipticity of $a(\cdot, \cdot)$ and the first line of (3.5),

$$(3.14) \quad \begin{aligned} \alpha |u_0| &\leq a\left(u_0, \frac{u_0}{|u_0|}\right) \leq \sup_{v_0 \in \ker B} \frac{a(u_0, v_0)}{|v_0|} = \sup_{v_0 \in \ker B} \frac{a(u, v_0) - a(u^\perp, v_0)}{|v_0|} \\ &\leq \|\varphi\| + \|a\| |u^\perp|. \end{aligned}$$

From (3.13) and (3.14), we have

$$(3.15) \quad \begin{aligned} |u| &\leq |u^\perp| + |u_0| \leq \frac{1}{\beta} \left(\|\Psi\| + \|c\|^{1/2} c(\lambda, \lambda)^{1/2} \right) + \frac{1}{\alpha} (\|\varphi\| + \|a\| |u^\perp|) \\ &\leq \frac{1}{\beta} \left(\|\Psi\| + \|c\|^{1/2} c(\lambda, \lambda)^{1/2} \right) \left(1 + \frac{\|a\|}{\alpha} \right) + \frac{\|\varphi\|}{\alpha}. \end{aligned}$$

Now subtracting the two lines of (3.5) with $v = u$ and $\mu = \lambda$, we obtain

$$\begin{aligned} a(u, u) + c(\lambda, \lambda) &= \langle \varphi, u \rangle - \langle \Psi, \lambda \rangle = \langle \varphi, u \rangle - (Bw, \lambda) - c(\gamma, \lambda) \\ &\leq \|\varphi\| |u| + \|B\| |w| |\lambda^\perp| + c(\gamma, \gamma)^{1/2} c(\lambda, \lambda)^{1/2}, \end{aligned}$$

so that, from (3.15) and (3.12),

$$\begin{aligned} (3.16) \quad a(u, u) + c(\lambda, \lambda) &\leq \left(\|\varphi\| + \frac{\|B\|}{\beta} |w| \|a\| \right) \left(\frac{\|\Psi\|}{\beta} \left(1 + \frac{\|a\|}{\alpha} \right) + \frac{\|\varphi\|}{\alpha} \right) \\ &+ c(\lambda, \lambda)^{1/2} \left(c(\gamma, \gamma)^{1/2} + \frac{1}{\beta} \|c\|^{1/2} \left(1 + \frac{\|a\|}{\alpha} \right) \left(\|\varphi\| + \frac{\|B\|}{\beta} |w| \frac{\|a\|}{\alpha} \right) \right), \end{aligned}$$

which can be written

$$a(u, u) + c(\lambda, \lambda) \leq P_0(\|\varphi\|, \|\Psi\|, |w|, |\gamma|_c) + c(\lambda, \lambda)^{1/2} P_1(\|\varphi\|, \|\Psi\|, |w|, |\gamma|_c),$$

where P_0 (resp., P_1) is an homogeneous polynomial of degree 2 (resp., 1) in its four variables. The coefficients of those polynomials are polynomial in $\|B\|$, $\|a\|$, $1/\beta$, $1/\alpha$, $\|c\|^{1/2}$ with positive coefficients. We write $X = c(\lambda, \lambda)^{1/2}$, so that $X^2 \leq P_1 X + P_0$, which implies $|X| \leq P_1 + \sqrt{P_0}$, and finally

$$c(\lambda, \lambda) = X^2 \leq 2P_1^2 + 2P_0 = P_2(\|\varphi\|, \|\Psi\|, |w|, |\gamma|_c),$$

where P_2 is an homogeneous polynomial of degree 2. It is dominated by the square of the sum of the modulus of its variables, so that

$$c(\lambda, \lambda)^{1/2} \leq C(\|\varphi\| + \|\Psi\| + |w| + |\gamma|_c).$$

Again using (3.16) (we keep C to design a generic constant, or more precisely a polynomial in $\|B\|$, $\|a\|$, $1/\beta$, $1/\alpha$, $\|c\|^{1/2}$), we obtain immediately

$$|u| \leq C(\|\varphi\| + \|\Psi\| + |w| + |\gamma|_c).$$

Finally, we write the second line of (3.5) with $\mu \in \ker B^*$. As $c(\cdot, \cdot)$ verifies (3.8), it yields $\lambda_0 = P_{\ker B^*} \gamma$, so that $|\lambda_0| \leq |\gamma|$. As $|\gamma|_c \leq \|c\|^{1/2} |\gamma|$, and $\|\Psi\| \leq |w| + |\gamma|$, estimate (3.9) is obtained. \square

4. Application. This section is dedicated to the application of Theorem 3.1 to a particular problem, namely a scalar version of the rigidity constraint for fluid-particle flows.

4.1. Model problem. In order to present explicit constructions when needed, we consider a particular situation. We introduce $\Omega =]-2, 2[^2$, and $\mathcal{O} = B(0, 1) \subset\subset \Omega$ (see Figure 4.1). The case of more general situations is addressed in Remark 4.2, at the end of this paper. We consider the following problem:

$$(4.1) \quad \begin{cases} -\Delta u = f & \text{in } \Omega \setminus \overline{\mathcal{O}}, \\ u = 0 & \text{on } \partial\Omega, \\ u = U & \text{on } \partial\mathcal{O}, \\ \int_{\partial\mathcal{O}} \frac{\partial u}{\partial n} = 0, \end{cases}$$

where U is an unknown constant, and $f \in L^2(\Omega \setminus \overline{\mathcal{O}})$. The scalar field u can be seen as a temperature and \mathcal{O} as a zone with infinite conductivity.

DEFINITION 4.1. *We say that u is a weak solution to (4.1) if $u \in V = H_0^1(\Omega)$, there exists $U \in \mathbb{R}$ such that $u = U$ a.e. in \mathcal{O} , and*

$$\int_{\Omega} \nabla u \cdot \nabla v = \int_{\Omega} f v \quad \forall v \in \mathcal{D}_{\mathcal{O}}(\Omega),$$

where $\mathcal{D}_{\mathcal{O}}(\Omega)$ is the set of all those functions which are compactly supported, C^∞ on Ω , and which are constant over \mathcal{O} .

PROPOSITION 4.2. *Problem (4.1) admits a unique weak solution $u \in V = H_0^1(\Omega)$, which is characterized as the solution to the minimization problem*

$$(4.2) \quad \begin{cases} \text{Find } u \in K \text{ such that} \\ J(u) = \inf_{v \in K} J(v), \quad \text{with } J(v) = \frac{1}{2} \int_{\Omega} |\nabla u|^2 - \int_{\Omega} f v, \\ K = \{v \in H_0^1(\Omega), \nabla v = 0 \text{ a.e. in } \mathcal{O}\}, \end{cases}$$

where f has been extended by 0 inside \mathcal{O} . Furthermore the restriction of u to the domain $\Omega \setminus \overline{\mathcal{O}}$ is in $H^2(\Omega \setminus \overline{\mathcal{O}})$.

Proof. Existence and uniqueness are direct consequences of the Lax–Milgram theorem applied in $K = \{v \in V, \nabla v = 0 \text{ a.e. in } \mathcal{O}\}$, which gives in addition the characterization of u as the solution to (4.2). Now $u|_{\Omega \setminus \overline{\mathcal{O}}}$ satisfies $-\Delta u = f$, with regular Dirichlet boundary conditions on the boundary of $\Omega \setminus \overline{\mathcal{O}}$ which decomposes as $\partial\mathcal{O} \cup \partial\Omega$. As Ω is a convex polygon and $\partial\mathcal{O}$ is smooth, standard theory ensures that $u|_{\Omega \setminus \overline{\mathcal{O}}} \in H^2(\Omega \setminus \overline{\mathcal{O}})$. \square

PROPOSITION 4.3 (saddle-point formulation). *Let u be the weak solution to (4.1). There exists a unique $\lambda \in \Lambda = L^2(\mathcal{O})^2$ such that λ is a gradient, and*

$$\int_{\Omega} \nabla u \cdot \nabla v + \int_{\mathcal{O}} \lambda \cdot \nabla v = \int_{\Omega} f v \quad \forall v \in V.$$

In addition λ is in $H^1(\mathcal{O})^2$.

Proof. The first part is a consequence of Proposition 2.5, where B is defined by

$$B : v \in H_0^1(\Omega) \mapsto \nabla v \in L^2(\mathcal{O})^2.$$

Let us prove that B has a closed range. Considering $\mu \in \Lambda$ with $\mu = \nabla v$, we define $w \in H_0^1(\mathcal{O})$ as $w = v - m(v)$, where $m(v)$ is the mean value of v over \mathcal{O} . By the Poincaré–Wirtinger inequality, one has

$$\|w\|_{H^1(\mathcal{O})} \leq C \|\mu\|_{L^2(\mathcal{O})^2}.$$

Now, as $\mathcal{O} \subset\subset \Omega$, there exists a continuous extension operator from $H^1(\mathcal{O})$ to $H_0^1(\Omega)$, so that we can extend w to obtain $\tilde{w} \in H_0^1(\Omega)$ with a norm controlled by $\|\mu\|_{L^2(\mathcal{O})^2}$, which proves the closed character of $B(V)$, and consequently the existence of $\lambda \in \Lambda$, and its uniqueness in $B(V)$.

Let us now describe λ . We have

$$\int_{\Omega} \nabla u \cdot \nabla v + \int_{\mathcal{O}} \lambda \cdot \nabla v = \int_{\Omega} f v,$$

so that, by taking test functions in $\mathcal{D}(\mathcal{O})$, we get $\lambda \in H_{\text{div}}(\mathcal{O})$ with $\nabla \cdot \lambda = 0$. Taking now test functions which do not vanish on the boundary of \mathcal{O} , we identify the normal trace of λ with $\partial u / \partial n \in H^{1/2}(\partial \mathcal{O})$. Therefore λ is defined as the unique divergence-free vector field in \mathcal{O} , with normal derivative equal to $\partial u / \partial n$ on $\partial \mathcal{O}$, which, in addition, is a gradient. In other words $\lambda = \nabla \Phi$, with

$$\begin{cases} \Delta \Phi = 0 & \text{in } \mathcal{O}, \\ \frac{\partial \Phi}{\partial n} = \frac{\partial u}{\partial n} & \text{on } \partial \mathcal{O}. \end{cases}$$

As \mathcal{O} is smooth, $\Phi \in H^2(\mathcal{O})$, so that $\lambda = \nabla \Phi \in H^1(\mathcal{O})^2$. \square

We introduce the penalized version of problem (4.2)

$$(4.3) \quad \begin{cases} \text{Find } u^\varepsilon \in V \text{ such that } J^\varepsilon(u^\varepsilon) = \inf_{v \in V} J^\varepsilon(v), \\ J^\varepsilon(v) = \frac{1}{2} \int_{\Omega} |\nabla v|^2 + \frac{1}{2\varepsilon} \int_{\mathcal{O}} |\nabla v|^2 - \int_{\Omega} f v. \end{cases}$$

Now we consider the family of Cartesian triangulations (T_h) of the square Ω (see Figure 4.1), and we denote by V_h the standard finite element space of continuous, piecewise affine function with respect to T_h :

$$V_h = \{v_h \in V, v_{|T} \text{ is affine } \forall T \in T_h\}.$$

It is tempting to define the fully discretized problem as the problem which consists in minimizing J^ε over V_h . But this straightforward approach (which does not correspond to what is done in actual computations; see Remark 4.1) raises some problems in relation to the discrete inf-sup condition which we need to establish the error estimate (see Proposition 4.7). It is related to the fact that we cannot control the size of intersections of triangles with \mathcal{O} (relative to the size of the whole triangle, which is $h^2/2$). To overcome this problem, many strategies can be adopted, all of them leading to change B onto a new discrete operator B_h . We propose here a radical method, which simply consists in removing in the penalty integral all squares (two-triangle sets) which intersect the boundary of \mathcal{O} . It will be made clear that the convergence

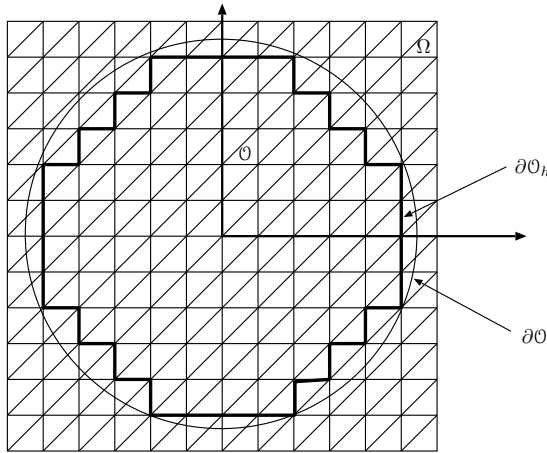


FIG. 4.1. Domains Ω , \mathcal{O} , \mathcal{O}_h , and the mesh T_h .

result is not sensitive to what is actually done in the neighborhood of $\partial\Omega$. The proof simply requires that the reduced obstacle is included in the exact one, and that the difference set $\mathcal{O} \setminus \mathcal{O}_h$ lies in a narrow band whose width goes to 0 like h .

DEFINITION 4.4. *The reduced obstacle $\mathcal{O}_h \subset \mathcal{O}$ is defined as the union of the triangles which belong to an elementary square which is contained in the disk \mathcal{O} (see Figure 4.1).*

DEFINITION 4.5. *We recall that $V = H_0^1(\Omega)$, Λ is $L^2(\mathcal{O})^2$, and $B \in \mathcal{L}(V, \Lambda)$ is the gradient operator (see Proposition 4.3). We define $B_h \in \mathcal{L}(V, \Lambda)$ as*

$$v \in V \mapsto \mu = B_h v = \mathbb{1}_{\mathcal{O}_h} \nabla v,$$

where $\mathbb{1}_{\mathcal{O}_h}$ is the characteristic function of \mathcal{O}_h (see Definition 4.4). Finally, the discretization space $\Lambda_h \subset \Lambda = L^2(\mathcal{O})^2$ is the set of all those vector fields μ_h such that their restriction to \mathcal{O}_h is the gradient of a scalar field $v_h \in V_h$, and which vanish a.e. in $\mathcal{O} \setminus \mathcal{O}_h$, which we can express as

$$\Lambda_h = \{\mu_h \in \Lambda, \exists v_h \in V_h, \mu_h = B_h v_h\} = B_h(V_h).$$

The fully discretized problem reads

$$(4.4) \quad \begin{cases} \text{Find } u_h^\varepsilon \in V_h \text{ such that } J_h^\varepsilon(u^\varepsilon) = \inf_{v_h \in V_h} J_h^\varepsilon(v_h), \\ J_h^\varepsilon(v_h) = \frac{1}{2} \int_{\Omega} |\nabla v_h|^2 + \frac{1}{2\varepsilon} \int_{\mathcal{O}_h} |\nabla v_h|^2 - \int_{\Omega} f v_h. \end{cases}$$

4.2. Error estimate for the model problem.

PROPOSITION 4.6 (primal/dual error estimate for (4.1), nonconforming case). *Let u be the weak solution to (4.1), u_h^ε the solution to (4.4), and λ the Lagrange multiplier (see Proposition 4.3), and let $\lambda_h^\varepsilon = B_h u_h^\varepsilon / \varepsilon$ (see Definition 4.5). We have the following error estimate:*

$$(4.5) \quad |u - u_h^\varepsilon| + |\lambda - \lambda_h^\varepsilon| \leq C(h^{1/2} + \varepsilon).$$

Proof. The proof is based on the abstract estimate in Theorem 3.1. All technical ingredients are put off until the end of the section. We shall simply refer here to the corresponding properties. The crucial requirement is the discrete inf-sup condition, which can be established for this choice of B_h (see Proposition 4.7). The terms

$$\inf_{\tilde{u}_h \in V_h} |\tilde{u}_h - u| \quad \text{and} \quad \inf_{\tilde{\lambda}_h \in \Lambda_h} |\tilde{\lambda}_h - \lambda|$$

can be shown to behave like $h^{1/2}$ (see Propositions 4.8 and 4.9, respectively). The last two terms can be handled the same way as $|\tilde{\lambda}_h - \lambda|$. Indeed,

$$|(B_h^* - B^*)\lambda| \leq |\lambda|_{0, \mathcal{O} \setminus \overline{\mathcal{O}_h}},$$

which is a $\mathcal{O}(h^{1/2})$ (it is the L^2 norm of a function with H^1 regularity, on a neighborhood of $\partial\mathcal{O}$). The very same argument holds for $|(B_h - B)z|$ (in our case, both quantities are the same). \square

PROPOSITION 4.7 (discrete inf-sup condition). *Let Ω and \mathcal{O} be defined as in the beginning of section 4. We introduce $h = 1/N$, $N \in \mathbb{N}$, and T_h is the regular triangulation with step h , so that the center of \mathcal{O} is a vertex of T_h . According to*

Definitions 4.4 and 4.5, \mathcal{O}_h is the reduced obstacle, and $\Lambda_h \subset L^2(\mathcal{O})^2 = \Lambda$ is the set of all those vector fields which are the gradient of a piecewise affine function in \mathcal{O}_h , and which vanish in $\mathcal{O} \setminus \mathcal{O}_h$.

There exists $\beta > 0$ such that, for all $h (= 1/N)$,

$$(4.6) \quad \beta |P_{(\ker B_h)^\perp} v_h| \leq |B_h v_h| \quad \forall v_h \in V_h, \quad \sup_{v_h \in V_h} \frac{(B_h v_h, \lambda_h)}{|v_h|} \geq \beta \|\lambda_h\|_{\Lambda_h}.$$

Proof. Let $v_h \in V_h$ be given. If we are able to build $w_h \in V_h$ such that $B_h w_h = B_h v_h$, with $\|w_h\| \leq C \|B_h v_h\|$, we obtain

$$|P_{(\ker B_h)^\perp} v_h| = \inf_{\tilde{v}_h \in \ker B_h} |v_h - \tilde{v}_h| \leq |v_h - (w_h - v_h)| = |w_h| \leq C |B_h v_h|,$$

and the first inequality is proven. Let us describe how this $w_h \in V_h$ can be built in five steps. First, we introduce $w_h^1 = v_h - \bar{v}_h$, where \bar{v}_h is the mean value of w_h over \mathcal{O}_h . Note that w_h^1 is not in V_h (it does not vanish on $\partial\Omega$), but we consider only its restriction to \mathcal{O}_h . We have $B_h w_h^1 = B_h v_h$, and the norm of w_h^1 is controlled: $\|w_h^1\|_{H^1(\mathcal{O}_h)} \leq C_1 \|B_h v_h\|_{L^2(\mathcal{O}_h)^2}$ by the Poincaré–Wirtinger inequality (with a constant which does not depend on h , as can be checked easily).

We shall now describe how we plan to extend w_h^1 in the first quadrant, the three others being done the same way. This construction is illustrated by Figure 4.2. The first step consists in extending w_h^1 in the polygonal domain $CA_3A'_2A_1$ on each horizontal segment by symmetry (see Figure 4.2). A similar construction extends w_h^1 in

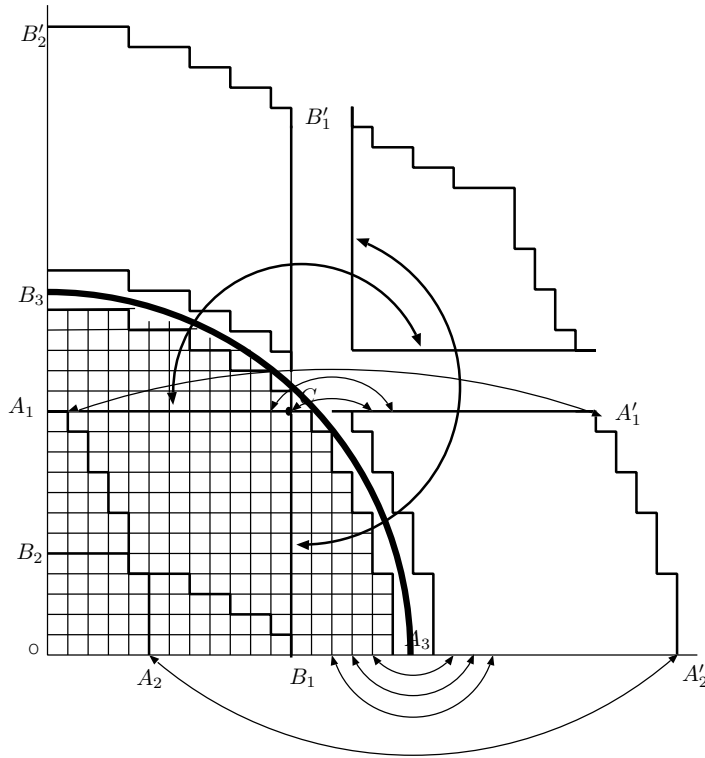


FIG. 4.2. Construction of w_h^2 .



FIG. 4.3. *Stretching of w_h^2 (detail).*

$CB_1^1B_2^1B_3$. Now the function is simply extended in the upper right zone by symmetry around C . To show that the H^1 seminorm of the newly defined function w_h^2 is under control, we first remark that the shift between two consecutive lines does not exceed one cell. Now consider the detail in Figure 4.3. On the left we represented a detail of the triangulated domain in \mathcal{O} where w_h^2 is already defined; the u_i 's and v_i 's represent the values of w_h^2 at some vertices. Now by applying the “symmetry” described previously, we obtain the stretched function which we represent on a single element. To control the effect of this stretching, we use Lemma 4.10 in the following way: The square of the H^1 seminorm of the new function is a quadratic nonnegative form q_1 in the six variables, and the square of the H^1 seminorm corresponding to the left-hand situation itself is a scale invariant quadratic, nonnegative form q_2 in the same variables, so that Lemma 4.10 ensures the existence of a universal constant C such that $q_1 \leq Cq_2$. As a consequence, the H^1 seminorm of the stretched function (in $CA_3A_2^1A_1^1$) is controlled by the H^1 seminorm of the initial function (in $CA_1A_2A_3$). As the new function in $CA_1^1B_1^1$ is obtained by standard symmetry, the H^1 seminorm identifies with the one of the initial function in CA_1B_1 .

This leads to a new function w_h^2 defined on \mathcal{O}_h^2 , subtriangulation of T_h , with $|w_h^2|_{1,\mathcal{O}_h^4} \leq C_2 \|B_h v_h\|_{L^2(\Omega)^2}$. As w_h^2 has zero mean value in $B(0, 1/2)$, one has

$$\|w_h^2\|_{H^1(\mathcal{O}_h^2)} \leq C'_2 \|B_h v_h\|_{L^2(\Omega)^2}.$$

Finally, \mathcal{O}_h^2 contains a ball strictly larger than \mathcal{O} , say $B(0, 1 + \sqrt{2}/4)$. Considering now a smooth function ρ which is equal to 1 in $B(0, (1 + r)/2)$, and 0 outside $B(0, r)$, we define w_h^3 as $I_h(\rho w_h^2)$ on \mathcal{O}_h^2 , and 0 in $\Omega \setminus \mathcal{O}_h^2$, where I_h is the standard interpolation operator. This function is in $V_h \cap H_0^1(\Omega)$, and it verifies

$$B_h w_h^3 = \lambda_h, \quad \|w_h^3\|_{H^1(\Omega)} \leq C_3 \|B_h v_h\|_{L^2(\Omega)^2},$$

so that the first inequality of (4.6) holds, with $\beta = 1/C_3$.

The second one is a direct consequence of the first one: given $\lambda_h = B_h u_h$, one considers $w_h = P_{(\ker B_h)^\perp} v_h$, so that

$$\sup_{v_h \in V_h} \frac{(B_h v_h, \lambda_h)}{|v_h|} \geq \frac{(B_h w_h, \lambda_h)}{|w_h|} = \frac{|B_h w_h|^2}{|w_h|} \geq \beta |B_h w_h| = \beta \|\lambda_h\|_{\Lambda_h},$$

which ends the proof. \square

PROPOSITION 4.8 (approximation of u). *We make the same assumptions as in Proposition 4.7, and we consider $u \in H_0^1(\Omega)$ such that $u = U \in \mathbb{R}$ a.e. in \mathcal{O} , $u_{\Omega \setminus \overline{\mathcal{O}}} \in H^2(\Omega \setminus \overline{\mathcal{O}})$. There exists $C > 0$ such that*

$$\inf_{\tilde{u}_h \in V_h} \|u - \tilde{u}_h\|_{H^1(\Omega)} \leq Ch^{1/2}.$$

Proof. We recall that I_h is the standard interpolation operator from $C(\Omega)$ onto V_h . Let us assume here that the constant value U on \mathcal{O} is O (which can be achieved by subtracting a smooth extension of this constant outside \mathcal{O}). Now we define $\tilde{\mathcal{O}}_h$ as the union of all those triangles of T_h which have a nonempty intersection with \mathcal{O} . We define \tilde{u}_h as the function in V_h which is 0 in $\tilde{\mathcal{O}}_h$ and which identifies with $I_h u$ at all other vertices. We introduce a narrow band around \mathcal{O} :

$$(4.7) \quad \omega_h = \left\{ x \in \Omega, x \notin \bar{\mathcal{O}}, d(x, \bar{\mathcal{O}}) < 2\sqrt{2}h \right\}.$$

As $u|_{\Omega \setminus \bar{\mathcal{O}}} \in H^2(\Omega \setminus \bar{\mathcal{O}})$, standard finite element estimates give

$$(4.8) \quad |u - \tilde{u}_h|_{0, L^2(\Omega \setminus (\mathcal{O} \cup \bar{\omega}_h))} \leq Ch^2 |u|_{H^2(\Omega \setminus \bar{\mathcal{O}})},$$

$$(4.9) \quad |u - \tilde{u}_h|_{1, L^2(\Omega \setminus (\mathcal{O} \cup \bar{\omega}_h))} \leq Ch |u|_{H^2(\Omega \setminus \bar{\mathcal{O}})}.$$

By construction, both L^2 and H^1 errors in \mathcal{O} are zero. There remains to estimate the error in the band ω_h . The principle is the following: \tilde{u}_h is a poor approximation of u in ω_h , but it is not very harmful because ω_h is small. Note that similar estimates are proposed in [SMSTT05] or [AR08]. For the sake of completeness, and because it is essential to understand why a better order than $1/2$ cannot be expected, we shall detail here the proof. First of all, we write

$$(4.10) \quad \|u - \tilde{u}_h\| \leq |u|_{0, \omega_h} + |u|_{1, \omega_h} + |u_h|_{0, \omega_h} + |u_h|_{1, \omega_h} = A + B + C + D.$$

Lemma 4.13 ensures $B \leq Ch^{1/2}$, and $A \leq Ch^{3/2}$. As for \tilde{u}_h (terms C and D in (4.10)), the proof is less trivial. It relies on the technical lemmas (Lemmas 4.11, 4.12, and 4.14 (see section 4.3)) which can be used as follows. The problematic triangles are those on which \tilde{u}_h identifies neither with 0, nor with $I_h u$. On such triangles, \tilde{u}_h sticks to $I_h u$ at 1 or 2 vertices, and vanishes at 2 or 1 vertices. As a consequence, the L^∞ norm of \tilde{u}_h is less than the L^∞ norm of $I_h u$. Let T be such a triangle. We write (using Lemma 4.11, the latter remark, the fact that I_h is a contraction from L^∞ onto L^∞ , Lemma 4.11 again, and Lemma 4.14)

$$\begin{aligned} \|\tilde{u}_h\|_{L^2(T)}^2 &\leq C' |T| \|\tilde{u}_h\|_{L^\infty(T)}^2 \leq C' |T| \|I_h u\|_{L^\infty(T)}^2 \\ &\leq \frac{C'}{C} \|I_h u\|_{L^2(T)}^2 \leq C'' \left(\|u\|_{L^2(T)}^2 + h^4 |u|_{2,T}^2 \right). \end{aligned}$$

By summing up all these contributions over all triangles which intersect ω_h , and using the fact that the L^2 norm of u on ω_h behaves like $h^{3/2} |u|_{2,T}$, we obtain

$$\|\tilde{u}_h\|_{L^2(\omega_h)}^2 \leq \sum_{T \cap \omega_h \neq \emptyset} \|\tilde{u}_h\|_{L^2(T)}^2 \leq h^3 |u|_{2,T}^2,$$

which gives the expected $h^{3/2}$ estimate for C . The last term of (4.10) is directly obtained by the previous estimate combined with the inverse inequality expressed by Lemma 4.12. \square

PROPOSITION 4.9 (approximation of λ). *Let $\lambda \in H^1(\mathcal{O})^2$ be given, with $\lambda = \nabla w$, $w \in H^2(\mathcal{O})$. There exists a constant $C > 0$ such that*

$$\inf_{\tilde{\lambda}_h \in \Lambda_h} \left\| \lambda - \tilde{\lambda}_h \right\|_{L^2(\mathcal{O})} \leq Ch^{1/2} |\lambda|_{1, \mathcal{O}},$$

where Λ_h is defined in section 3 (see Definition 4.5).

Proof. First of all, we extend w on $\Omega \setminus \bar{\mathcal{O}}$, to obtain a function (still denoted by w) in $H_0^1(\Omega) \cap H^2(\Omega)$. Let us define w_h as the standard interpolate of w over T_h . One has $|w - w_h|_{1,\mathcal{O}} \leq Ch$. We define $\tilde{\lambda}_h \in \Lambda_h$ as the piecewise constant function which identifies with ∇w_h on \mathcal{O}_h (see Definition 4.4), and which vanishes in $\mathcal{O} \setminus \mathcal{O}_h$. One has

$$\left\| \nabla w_h - \tilde{\lambda}_h \right\|_{L^2(\mathcal{O})} = \left\| \nabla w_h - \tilde{\lambda}_h \right\|_{L^2(\mathcal{O} \setminus \mathcal{O}_h)} = \|\nabla w_h\|_{L^2(\mathcal{O} \setminus \mathcal{O}_h)} \leq C \|\nabla w\|_{L^2(\mathcal{O} \setminus \mathcal{O}_h)},$$

which is the H^1 seminorm of a function in H^2 , in a narrow domain. Therefore it behaves like $h^{1/2}$ times the H^2 seminorm of u (see Lemma 4.13 and Remark 4.3), which is the H^1 seminorm of λ . Finally, one gets

$$\left\| \lambda - \tilde{\lambda}_h \right\|_{L^2(\mathcal{O})} \leq |w - w_h|_{1,\mathcal{O}} + \left\| \nabla w_h - \tilde{\lambda}_h \right\|_{L^2(\mathcal{O})} \leq C(h + h^{1/2}) |\lambda|_{1,\mathcal{O}},$$

which ends the proof. \square

Remark 4.1 (boundary fitted meshes). Although it is somewhat in contradiction with its original purpose, the penalty method can be used together with a discretization based on a boundary fitted mesh. In that case, the approximation error behaves no longer like $h^{1/2}$ but like h .

Remark 4.2 (technical assumptions). Some assumptions we made are only technical and can surely be relaxed without changing the convergence results. For example the inclusion, which we supposed circular, could be a collection of smooth domains. Note that a convex polygon is not acceptable, as it is seen from the outside, so that u may no longer be in H^2 , which rules out some of the approximation properties we made. Concerning the mesh, we have good confidence in the fact that the result generalizes to any kind of unstructured mesh, but the proof of Proposition 4.7 in the general case can no longer be based on an explicit construction.

4.3. Technical lemmas. We gather here some elementary properties which are used in the proofs of Propositions 4.6, 4.7, 4.8, and 4.9.

LEMMA 4.10. *Let E be a finite dimensional real vector space, with q_1 and q_2 two nonnegative quadratic forms with $\ker q_2 \subset \ker q_1$. There exists $C > 0$ such that $q_1 \leq Cq_2$.*

Proof. As q_2 is nonnegative, $\tilde{v} \mapsto |\tilde{v}|_{q_2(v)} = \sqrt{q_2(v)}$ is a norm for $E/\ker q_2$. Now we define

$$\tilde{q}_1 : \tilde{v} \in E/\ker q_2 \longmapsto \tilde{q}_1(\tilde{v}) = q_1(v) \in \mathbb{R}.$$

As $\ker q_1$ contains $\ker q_2$, this functional is well defined. As it is quadratic over a finite dimensional space, it is continuous for the norm $\sqrt{q_2}$, so that

$$q_1(v) = \tilde{q}_1(\tilde{v}) \leq C |v|_{q_2}^2 = q_2(v),$$

which ends the proof. \square

LEMMA 4.11. *There exist constants C and C' such that, for any nondegenerated triangle T , for any function w_h affine in T ,*

$$(4.11) \quad C |T| \|w_h\|_{L^\infty(T)}^2 \leq \|w_h\|_{L^2(T)}^2 \leq C' |T| \|w_h\|_{L^\infty(T)}^2.$$

Proof. It is a consequence of the fact that, when deforming the supporting triangle T , the L^∞ norm is unchanged whereas the L^2 norm scales like $|T|^{1/2}$. \square

LEMMA 4.12. *There exists a constant C such that, for any nondegenerated triangle T , for any function w_h affine in T ,*

$$|w_h|_{1,K}^2 \leq C \frac{|T|}{\rho_K^2} \|w_h\|_{L^\infty(T)}^2,$$

where ρ_K is the diameter of the inscribed circle.

Proof. Again, it is a straightforward consequence of the fact that, when deforming the supporting triangle T , the L^∞ norm is unchanged whereas the gradient (which is constant over the triangle) scales like $1/\rho_k$, so that the H^1 seminorm scales like $|T|^{1/2}/\rho_K$. \square

The next lemma establishes some Poincaré-like inequalities in narrow domains.

LEMMA 4.13. *Let $\Theta \subset \mathbb{R}^2$ be the unit disk, strongly included in a domain Ω , and let ω_η be the narrow band (note that this definition differs slightly from (4.7), which is of no consequence):*

$$\omega_\eta = \{x \in \Omega, x \notin \bar{\Theta}, d(x, \bar{\Theta}) < \eta\}, \text{ with } \eta > 0.$$

Denoting by $|\cdot|_{p,\omega}$ the H^p seminorm over ω , we have the following estimates:

$$\begin{aligned} |\varphi|_{0,\omega_\eta} &\leq C\eta^{1/2} |\varphi|_{1,\Omega \setminus \bar{\Theta}} \quad \forall \varphi \in H^1(\Omega \setminus \bar{\Theta}), \quad \varphi|_{\partial\Omega} = 0, \\ |\varphi|_{1,\omega_\eta} &\leq C\eta^{1/2} |\varphi|_{2,\Omega \setminus \bar{\Theta}} \quad \forall \varphi \in H^2(\Omega \setminus \bar{\Theta}), \quad \varphi|_{\partial\Omega} = 0, \\ |\varphi|_{0,\omega_\eta} &\leq C\eta^{3/2} |\varphi|_{2,\Omega \setminus \bar{\Theta}} \quad \forall \varphi \in H^2(\Omega \setminus \bar{\Theta}), \quad \varphi|_{\partial\Omega} = 0, \quad \varphi|_{\partial\Theta} = 0. \end{aligned}$$

Proof. We assume here that φ is C^1 in $\Omega \setminus \bar{\Theta}$ (the general case is obtained immediately by density). Using polar coordinates, we write $u(r, \theta) = u(1, \theta) + \int_1^r \partial_r u dr$, so that

$$\begin{aligned} |u|_{0,\omega_h}^2 &\leq 2 \int_0^{2\pi} \int_1^{1+\eta} |u(1, \theta)|^2 r dr d\theta + 2 \int_0^{2\pi} \int_1^{1+\eta} \left| \int_1^r \partial_r \varphi ds \right|^2 r dr d\theta \\ &\leq C \left(\eta |\varphi|_{0,\partial\Theta}^2 + \eta^2 |\varphi|_{1,\omega_\eta}^2 \right) \leq C\eta |\varphi|_{1,\Omega \setminus \bar{\Theta}}^2, \end{aligned}$$

from which we deduce the first estimate.

This same approach can be applied to $\partial_i \varphi$ for $\varphi \in H^2$. As φ is supposed to vanish over $\partial\Omega$, one has

$$|\partial_i \varphi| \leq C \|\nabla \varphi\|_{H^1(\Omega \setminus \bar{\Theta})} \leq C' |\varphi|_{2,\Omega \setminus \bar{\Theta}},$$

which leads to the second estimate. As for the third one, simply notice that the boundary term (L^2 norm over $\partial\Theta$) vanishes in the equation above:

$$|\varphi|_{0,\omega_\eta} \leq \eta |\varphi|_{1,\omega_\eta} \leq \eta^{3/2} |\varphi|_{2,\omega_\eta},$$

which ends the proof. \square

Remark 4.3. The previous lemma extends straightforwardly to the case of any smooth inclusion (C^2 regularity of the boundary is sufficient) strongly included in a

domain Ω (for a detailed proof of a similar property, see [GLM06]) or to the case where the function is defined within the subdomain (in that case, ω_η is defined as an inner narrow band).

The last lemma quantifies how one can control the L^2 norm of the interpolate of a regular function on a triangle, by means of the L^2 norm and the H^2 seminorm of the function.

LEMMA 4.14. *There exists a constant C such that, for any regular triangle T (see below), for any $u \in H^2(T)$,*

$$\|I_h u\|_{L^2(T)}^2 \leq C \left(\|u\|_{L^2(T)}^2 + h^4 |u|_{2,T}^2 \right).$$

By regular we mean that T runs over a set of triangles such that the flatness $\text{diam}(T)/\rho_K$ is bounded.

Proof. The interpolation operator $I_h : H^2(T) \rightarrow L^2(T)$ is continuous, and $|u|_{2,T}$ scales like $h/\rho_K^2 \approx 1/h$ whereas the L^2 norms scale like h . \square

5. Additional examples, concluding remarks. The approach can be checked to be applicable to some standard situations, like the constraint to vanish in an inclusion $\mathcal{O} \subset\subset \Omega$ (see Example 2.2), as soon as H^1 -penalty is used. The functional to minimize is then

$$J_\varepsilon(v) = \frac{1}{2} \int_\Omega |\nabla v|^2 - \int_\Omega f v + \frac{1}{2\varepsilon} \int_{\mathcal{O}} \left(u^2 + |\nabla u|^2 \right),$$

so that B identifies with the restriction operator from $H_0^1(\Omega)$ to $H^1(\mathcal{O})$. The discrete inf-sup condition, as well as the approximation properties, are essentially the same as in the case of an inclusion with infinite conductivity.

Another straightforward application of the abstract framework presented in section 3 is the numerical modeling of a rigid inclusion in a material which obeys Lamé's equations of linear elasticity. The penalized functional is then

$$J_\varepsilon(\mathbf{v}) = \frac{1}{2} \int_\Omega \mu |e(\mathbf{v})|^2 + \frac{1}{2} \int_\Omega \lambda |\nabla \cdot \mathbf{v}|^2 - \int_\Omega \mathbf{f} \cdot \mathbf{v} + \frac{1}{2\varepsilon} \int_{\mathcal{O}} |e(\mathbf{v})|^2,$$

where $e(\mathbf{v}) = (\nabla \mathbf{v} + (\nabla \mathbf{v})^T)/2$ is the strain tensor.

We conclude this section by some remarks on the proof itself and on possible extensions of this approach.

Remark 5.1 (conditioning issues). The fact that there is no need to choose ε too small (both errors balance for ε of the order of \sqrt{h}) is of particular importance in terms of conditioning. Indeed, considering the matrix A_h^ε resulting from the two-dimensional discrete minimization problem (4.4), it can be checked easily that its smallest eigenvalue scales like h^2 , whereas its largest eigenvalue behaves like $1/\varepsilon$, leading to a condition number of the order of $1/\varepsilon h^2$. Following the ε - h balance suggested by the error estimates, the condition number finally scales like $1/h^{5/2}$, which compares reasonably to the standard $1/h^2$. Note also that some special fixed point algorithms, recently proposed in [BFM08], can be used to circumvent the problem of ill-conditioning.

Remark 5.2 (convergence in space). The poor rate of convergence in h is optimal for a uniform mesh, at least if we consider the H^1 error over all Ω . Indeed, as the solution is constant inside \mathcal{O} , nonconstant outside with a jump in the normal derivative, the error within each element intersecting $\partial\mathcal{O}$ is a $\mathcal{O}(1)$ in this L^∞ norm. By summing

up over all those triangles, which cover a zone whose measure scales like h , we end up with this $h^{1/2}$ error. Note that a better convergence could be expected, in theory, if one considers only the error in the domain of interest $\Omega \setminus \overline{\mathcal{O}}$, the question now being whether the bad convergence in the neighborhood of $\partial\mathcal{O}$ pollutes the overall approximation. Our feeling is that this pollution actually occurs, because nothing is done in the present approach to distinguish both sides of $\partial\mathcal{O}$, so that the method tends to balance the errors on both sides. An interesting way to give priority to the side of interest is proposed in [DP02] for a boundary penalty method; it consists in having the diffusion coefficient vanish within Ω . Note that other methods have been proposed to reach the optimal convergence rate on nonboundary fitted mesh (see [Mau01]), but they are less straightforward to implement.

The simplest way to improve the actual order of convergence is to carry out a local refinement strategy in the neighborhood of $\partial\mathcal{O}$, as proposed in [RAB07].

Remark 5.3 (nonregular domains). The method can be implemented straightforwardly to nonregular domains (e.g., with corners or cusps), but the numerical analysis presented here is no longer valid. In particular, the inf-sup condition established in Proposition 4.7 and approximation properties for u (see Proposition 4.8) may no longer hold. Notice that Propositions 2.9 and 2.10 do not require any regularity assumption, so that convergence can be established for some sequences (h, ε) tending to $(0, 0)$, but the optimal order of convergence is lost. Practical tests suggest a reasonably good behavior of the method in such situations, like in the case where \mathcal{O} consists of two tangent discs (this situation is of special interest for practical applications in the context of fluid particle flows, when two particles are in contact; see, for example, [Lef07]).

Remark 5.4. Note that having ε go to 0 for any $h > 0$ leads to an estimate for a fictitious domain method (à la Glowinski, i.e., based on the use of Lagrange multipliers). In [GG95], an error estimate is obtained for such a method; it relies on two independent meshes for the primal and dual components of the solution (conditionally to some compatibility conditions between the sizes of the two meshes). We recover this estimate in the situation where the local mesh is simply the restriction of the covering mesh to the obstacle (to the reduced obstacle \mathcal{O}_h , to be more precise).

REFERENCES

- [AR08] P. ANGOT AND I. RAMIÈRE, *Convergence analysis of the Q1-finite element method for elliptic problems with non boundary-fitted meshes*, Internat. J. Numer. Methods Engrg., 75 (2008), pp. 1007–1052.
- [Bab73] I. BABUŠKA, *The finite element method with penalty*, Math. Comp., 27 (1973), pp. 221–228.
- [BE07] E. BURMAN AND A. ERN, *A continuous finite element method with face penalty to approximate Friedrichs' systems*, M2AN Math. Model. Numer. Anal., 41 (2007), pp. 55–76.
- [BF91] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer Ser. Comput. Math. 15, Springer-Verlag, New York, 1991.
- [BFM08] T. T. C. BUI, P. FREY, AND B. MAURY, *Méthode du second membre modifié pour la gestion de rapports de viscosité importants dans le problème de Stokes bifluide*, C. R. Mécanique, 336 (2008), pp. 524–529.
- [BHS03] R. BECKER, P. HANSBO, AND R. STENBERG, *A finite element method for domain decomposition with non-matching grids*, M2AN Math. Model. Numer. Anal., 37 (2003), pp. 209–225.
- [DP02] S. DEL PINO, *Une méthode d'éléments finis pour la résolution d'EDP dans des domaines décrits par géométrie constructive*, Ph.D. thesis, Université Pierre et Marie Curie, Paris, France, 2002.

- [DPM07] S. DEL PINO AND B. MAURY, *2d/3d turbine simulations with freefem*, in Numerical Analysis and Scientific Computing for PDEs and Their Challenging Applications, J. Haataja, R. Stenberg, J. Periaux, P. Raback, and P. Neittaanmaki, eds., CIMNE, Barcelona, Spain, 2008.
- [FFp] FREEFEM++; <http://www.freefem.org/>.
- [GG95] V. GIRAULT AND R. GLOWINSKI, *Error analysis of a fictitious domain method applied to a Dirichlet problem*, Japan J. Indust. Appl. Math., 12 (1995), pp. 487–514.
- [GLM06] V. GIRAULT, H. LÓPEZ, AND B. MAURY, *One time-step finite element discretization of the equation of motion of two-fluid flows*, Numer. Methods Partial Differential Equations, 22 (2006), pp. 680–707.
- [GR79] V. GIRAULT AND P.-A. RAVIART, *Finite Element Approximation of the Navier-Stokes Equations*, Lecture Notes in Math. 749, Springer-Verlag, Berlin, 1979.
- [HH02] A. HANSBO AND P. HANSBO, *An unfitted finite element method, based on Nitsche’s method, for elliptic interface problems*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 5537–5552.
- [JLM05] J. JANELA, A. LEFEBVRE, AND B. MAURY, *A penalty method for the simulation of fluid-rigid body interaction*, in CEMRACS 2004—Mathematics and Applications to Biology and Medicine, ESAIM Proc. 14, EDP Sciences, Les Ulis, France, 2005, pp. 115–123.
- [Lef07] A. LEFEBVRE, *Fluid-particle simulations with FreeFem++*, in Paris-Sud Working Group on Modelling and Scientific Computing 2006–2007, ESAIM Proc. 18, EDP Sciences, Les Ulis, France, 2007, pp. 120–132.
- [Mau01] B. MAURY, *A fat boundary method for the Poisson problem in a domain with holes*, J. Sci. Comput., 16 (2001), pp. 319–339.
- [Nit71] J. NITSCHKE, *Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind*, Abh. Math. Sem. Univ. Hamburg, 36 (1971), pp. 9–15.
- [PG02] T.-W. PAN AND R. GLOWINSKI, *Direct simulation of the motion of neutrally buoyant circular cylinders in plane Poiseuille flow*, J. Comput. Phys., 181 (2002), pp. 260–279.
- [RAB07] I. RAMIÈRE, P. ANGOT, AND M. BELLIARD, *A fictitious domain approach with spread interface for elliptic problems with general boundary conditions*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 766–781.
- [RPVC05] T. N. RANDRIANARIVELO, G. PIANET, S. VINCENT, AND J. P. CALTAGIRONE, *Numerical modelling of solid particle motion using a new penalty method*, Internat. J. Numer. Methods Fluids, 47 (2005), pp. 1245–1251.
- [SMSTT05] J. SAN MARTÍN, J.-F. SCHEID, T. TAKAHASHI, AND M. TUCSNAK, *Convergence of the Lagrange–Galerkin method for the equations modelling the motion of a fluid-rigid system*, SIAM J. Numer. Anal., 43 (2005), pp. 1536–1571.

MODIFIED COMBINED FIELD INTEGRAL EQUATIONS FOR ELECTROMAGNETIC SCATTERING*

O. STEINBACH[†] AND M. WINDISCH[†]

Abstract. The boundary integral formulation of exterior boundary value problems for the Maxwell system may not be equivalent to the original uniquely solvable problem if the wave number corresponds to an eigenvalue of an associated interior eigenvalue problem. To avoid these spurious modes one may use a combined boundary integral approach. To analyze the resulting boundary integral equations in the energy function spaces suitable regularizations have to be introduced. Here we formulate and analyze a modified boundary integral equation which is based on the use of standard boundary integral operators only. A first numerical example shows the applicability of the proposed approach.

Key words. combined field integral equations, electromagnetic scattering, Maxwell system

AMS subject classifications. 65N38, 78A45

DOI. 10.1137/070698063

1. Introduction. The modeling of electromagnetic scattering at a perfect conductor in the exterior of a bounded domain $\Omega \subset \mathbb{R}^3$ leads to the Dirichlet boundary value problem [12, 18, 22, 23]

$$(1.1) \quad \mathbf{curl} \mathbf{curl} \mathbf{U}(x) - \kappa^2 \mathbf{U}(x) = 0 \quad \text{for } x \in \Omega^c = \mathbb{R}^3 \setminus \overline{\Omega},$$

$$(1.2) \quad \mathbf{n}_x \times (\mathbf{U}(x) \times \mathbf{n}_x) = \mathbf{g}(x) \quad \text{for } x \in \Gamma = \partial\Omega,$$

where $\kappa \in \mathbb{R}_+$ is the wave number, and \mathbf{n}_x is the exterior unit normal vector for almost all $x \in \Gamma$. In addition to the exterior boundary value problem (1.1) we need to formulate the radiation condition of electromagnetic scattering, i.e., the Silver–Müller radiation condition

$$(1.3) \quad \lim_{r=|x| \rightarrow \infty} \int_{\partial B_r} |\mathbf{curl} \mathbf{U}(x) \times \mathbf{n}_x - i\kappa(\mathbf{n}_x \times \mathbf{U}(x)) \times \mathbf{n}_x|^2 ds_x = 0,$$

where B_r is a ball around zero with radius r . Note that the exterior Dirichlet boundary value problem (1.1)–(1.3) admits a unique solution. According to the partial differential operator in (1.1) we can formulate Green’s first formula which is valid for sufficiently smooth functions as

$$(1.4) \quad \int_{\Omega} \mathbf{curl} \mathbf{curl} \mathbf{U}(x) \cdot \overline{\mathbf{V}}(x) dx = \int_{\Omega} \mathbf{curl} \mathbf{U}(x) \cdot \mathbf{curl} \overline{\mathbf{V}}(x) dx \\ - \int_{\Gamma} (\mathbf{curl} \mathbf{U}(x)|_{\Gamma} \times \mathbf{n}_x) \cdot (\mathbf{n}_x \times (\overline{\mathbf{V}}(x)|_{\Gamma} \times \mathbf{n}_x)) ds_x.$$

*Received by the editors July 24, 2007; accepted for publication (in revised form) November 12, 2008; published electronically February 19, 2009. This work was supported by the Austrian Science Fund (FWF) within the project “Data sparse boundary and finite element domain decomposition methods in electromagnetics” under grant P19255.

<http://www.siam.org/journals/sinum/47-2/69806.html>

[†]Institute of Computational Mathematics, Graz University of Technology, Steyrergasse 30, 8010 Graz, Austria (o.steinbach@tugraz.at, markus.windisch@tugraz.at).

Based on (1.4) related Sobolev spaces and corresponding trace operators can be introduced [4, 5, 6, 7, 8]; these results will be summarized in section 2. Then, the well-known Stratton–Chu representation formula will be discussed which implies the definition of appropriate potential and boundary integral operators [6, 8, 11, 13, 16, 17, 20, 21, 23]. The corresponding boundary integral equations can be used for a numerical treatment of the problem by means of boundary element methods [3, 6, 8, 11, 12, 13, 19, 23]. But although the exterior boundary value problem (1.1)–(1.3) is uniquely solvable, the standard boundary integral equations are not uniquely solvable if the wave number κ corresponds to an eigenvalue of an associated interior eigenvalue problem. To avoid these spurious modes Brakhage and Werner [1] introduced a combined boundary integral approach for the acoustic problem in 1965. In the same year Panich discussed this approach for the electromagnetic case [24]. But the analysis of the approach of Brakhage and Werner is applicable for smooth boundaries only. Hence modified boundary integral equations were discussed in [10] for the acoustic case and in [9] for the electromagnetic case. In [14] an alternative approach was introduced for the acoustic case. Here we want to generalize this idea to obtain modified combined boundary integral equations for the electromagnetic case.

The paper is structured as follows: In section 2 we first summarize the definitions of Sobolev spaces to handle the variational formulation of the Maxwell system, and introduce potential operators and related boundary integral operators as needed later. We also discuss standard boundary integral approaches to solve the exterior Dirichlet boundary value problem, and comment on combined and already existing stabilized boundary integral formulations. An alternative modified boundary integral equation is formulated and analyzed in section 3. In particular, we present a new boundary integral formulation which is based on the use of standard, and therefore already available, boundary integral operators, and which is stable for all wave numbers. In section 4 we describe a first numerical example to show the applicability of the proposed approach. We finally end up with some conclusions and an outlook on ongoing work.

2. Function spaces and boundary integral equations. The formulation of boundary integral equations for the Maxwell system requires the use of the correct function spaces. Here we will recall only the definitions and the properties of Sobolev spaces for the Maxwell system; for a more detailed description see, e.g., [4, 5].

Let $\Omega \subset \mathbb{R}^3$ be a Lipschitz polyhedron [4] with a Lipschitz boundary $\Gamma = \partial\Omega$ which is the union of plane faces Γ_i , i.e., $\Gamma = \bigcup_i \Gamma_i$, where \mathbf{n}_i is the exterior normal vector on Γ_i .

The partial differential equation in (1.1) and Green’s first formula (1.4) motivate the definition of the energy space

$$\mathbf{H}(\mathbf{curl}, \Omega) := \{\mathbf{V} \in \mathbf{L}_2(\Omega) : \mathbf{curl} \mathbf{V} \in \mathbf{L}_2(\Omega)\}$$

as well as the space of the natural solutions

$$\mathbf{H}(\mathbf{curl}^2, \Omega) := \{\mathbf{V} \in \mathbf{H}(\mathbf{curl}, \Omega) : \mathbf{curl} \mathbf{curl} \mathbf{V} \in \mathbf{L}_2(\Omega)\}.$$

In addition we need to introduce appropriate Sobolev spaces on the boundary. For $|s| \leq 1$ and for scalar functions on the boundary the usual Sobolev spaces are denoted by $H^s(\Gamma)$. Let us define the Dirichlet traces

$$\gamma_D \mathbf{U} := \mathbf{n} \times (\mathbf{U}|_\Gamma \times \mathbf{n}) = \mathbf{n} \times \gamma_\times \mathbf{U}, \quad \gamma_\times \mathbf{U} := \mathbf{U}|_\Gamma \times \mathbf{n}$$

and the Neumann trace

$$\gamma_N \mathbf{U} := \mathbf{curl} \mathbf{U}|_\Gamma \times \mathbf{n}$$

which all are mappings into tangential spaces. Hence we introduce the space

$$\mathbf{L}_{2,t}(\Gamma) := \{\mathbf{u} \in \mathbf{L}_2(\Gamma) : \mathbf{u} \cdot \mathbf{n} = 0\}$$

of tangential $\mathbf{L}_2(\Gamma)$ integrable functions. For higher order Sobolev spaces we use the piecewise definition

$$\mathbf{H}_{pw,t}^s(\Gamma) := \{\mathbf{u} \in \mathbf{L}_{2,t}(\Gamma) : \mathbf{u} \in \mathbf{H}^s(\Gamma_k), k = 1, \dots, N_\Gamma\}.$$

The trace spaces $\gamma_D \mathbf{H}^1(\Omega)$ and $\gamma_\times \mathbf{H}^1(\Omega)$ are denoted by $\mathbf{H}_\parallel^{1/2}(\Gamma)$ and $\mathbf{H}_\perp^{1/2}(\Gamma)$, respectively; for an alternative definition see [4]. The dual spaces with respect to $\mathbf{L}_{2,t}(\Gamma)$ are denoted by $\mathbf{H}_\parallel^{-1/2}(\Gamma)$ and $\mathbf{H}_\perp^{-1/2}(\Gamma)$.

Before introducing the trace spaces of $\mathbf{H}(\mathbf{curl} \Omega)$ we need to define some boundary differential operators. Here we just give definitions for smooth boundaries; for Lipschitz polyhedrons see [4, 5]. For a scalar function u defined on Γ we denote by \tilde{u} an arbitrary bounded extension into a three-dimensional neighborhood of Γ . Then we can define the boundary differential operators

$$\nabla_\Gamma u := [\mathbf{n} \times (\nabla \tilde{u} \times \mathbf{n})]_{|\Gamma}, \quad \mathbf{curl}_\Gamma u := [\mathbf{curl}(\tilde{u} \mathbf{n})]_{|\Gamma},$$

where

$$\nabla_\Gamma : H^1(\Omega) \rightarrow \mathbf{L}_{2,t}(\Gamma), \quad \mathbf{curl}_\Gamma : H^1(\Omega) \rightarrow \mathbf{L}_{2,t}(\Gamma).$$

In addition, we introduce the adjoint operators of $-\nabla_\Gamma$ and of \mathbf{curl}_Γ , i.e.,

$$\operatorname{div}_\Gamma : \mathbf{L}_{2,t}(\Gamma) \rightarrow H_*^{-1}(\Omega), \quad \operatorname{curl}_\Gamma : \mathbf{L}_{2,t}(\Gamma) \rightarrow H_*^{-1}(\Omega),$$

where

$$H_*^{-1}(\Omega) = \{v \in H^{-1}(\Omega) : \langle v, 1 \rangle_\Omega = 0\}.$$

With the help of these operators we can finally define the Hilbert spaces

$$\mathbf{H}_\perp^{-1/2}(\mathbf{curl}_\Gamma, \Gamma) := \left\{ \mathbf{u} \in \mathbf{H}_\perp^{-1/2}(\Gamma) : \mathbf{curl}_\Gamma \mathbf{u} \in \mathbf{H}^{-1/2}(\Gamma) \right\},$$

$$\mathbf{H}_\parallel^{-1/2}(\operatorname{div}_\Gamma, \Gamma) := \left\{ \mathbf{u} \in \mathbf{H}_\parallel^{-1/2}(\Gamma) : \operatorname{div}_\Gamma \mathbf{u} \in H^{-1/2}(\Gamma) \right\}.$$

These spaces are dual to each other with respect to $\mathbf{L}_{2,t}(\Gamma)$ and represent the trace spaces $\gamma_D \mathbf{H}(\mathbf{curl}, \Omega)$ and $\gamma_\times \mathbf{H}(\mathbf{curl}, \Omega)$, respectively. Furthermore, there holds the following theorem [4, Theorems 2.7 and 2.8] and [5, Theorem 4.5].

THEOREM 2.1. *The operators*

$$\gamma_D : \mathbf{H}(\mathbf{curl}, \Omega) \rightarrow \mathbf{H}_\perp^{-1/2}(\mathbf{curl}_\Gamma, \Gamma),$$

$$\gamma_N : \mathbf{H}(\mathbf{curl} \operatorname{curl}, \Omega) \rightarrow \mathbf{H}_\parallel^{-1/2}(\operatorname{div}_\Gamma, \Gamma)$$

are linear, continuous, and surjective.

Now we are able to introduce some potential and boundary integral operators which are relevant for electromagnetic scattering [11]. The solution of the exterior Dirichlet boundary value problem (1.1)–(1.3) can be described by using the Stratton–Chu representation formula [13, 17, 23]

$$(2.1) \quad \mathbf{U}(x) = -\Psi_M^\kappa(\gamma_D^c \mathbf{U})(x) - \Psi_S^\kappa(\gamma_N^c \mathbf{U})(x) \quad \text{for } x \in \Omega^c,$$

where the Maxwell single layer potential is given by

$$\Psi_S^\kappa(\boldsymbol{\mu}) := \Psi_A^\kappa(\boldsymbol{\mu}) + \frac{1}{\kappa^2} \mathbf{grad} \Psi_V^\kappa(\operatorname{div}_\Gamma(\boldsymbol{\mu})),$$

and the Maxwell double layer potential is defined by

$$\Psi_M^\kappa(\boldsymbol{\lambda})(x) := \mathbf{curl} \Psi_A^\kappa(\boldsymbol{\lambda} \times \mathbf{n})(x).$$

The operators Ψ_A^κ and Ψ_V^κ are the vectorial and the scalar single layer potentials which are given by

$$\Psi_A^\kappa(\boldsymbol{\lambda})(x) := \int_\Gamma g_\kappa(x, y) \boldsymbol{\lambda}(y) ds_y, \quad \Psi_V^\kappa(\lambda)(x) := \int_\Gamma g_\kappa(x, y) \lambda(y) ds_y,$$

whereas $g_\kappa(x, y)$ is the fundamental solution of the Helmholtz equation,

$$g_\kappa(x, y) = \frac{1}{4\pi} \frac{e^{i\kappa|x-y|}}{|x-y|}.$$

To use an indirect approach to represent the solution of (1.1)–(1.3) the following result is essential; see, e.g., [11, Theorem 3.8] or [13, section 6].

THEOREM 2.2. *The Maxwell single and double layer potentials are solutions of the partial differential equation in (1.1) and fulfill the Silver–Müller radiation condition (1.3). Moreover, the following mapping properties are valid:*

$$\Psi_S^\kappa : \mathbf{H}_\parallel^{-1/2}(\operatorname{div}_\Gamma, \Gamma) \rightarrow \mathbf{H}_{\text{loc}}(\mathbf{curl}^2, \Omega \cup \Omega^c),$$

$$\Psi_M^\kappa : \mathbf{H}_\perp^{-1/2}(\operatorname{curl}_\Gamma, \Gamma) \rightarrow \mathbf{H}_{\text{loc}}(\mathbf{curl}^2, \Omega \cup \Omega^c).$$

Hence we can represent the solution of the exterior Dirichlet boundary value problem (1.1)–(1.3) either by the single layer potential

$$(2.2) \quad \mathbf{U}(x) = \Psi_S^\kappa(\boldsymbol{\mu})(x) \quad \text{for } x \in \Omega^c$$

or by using the double layer potential

$$(2.3) \quad \mathbf{U}(x) = \Psi_M^\kappa(\boldsymbol{\lambda})(x) \quad \text{for } x \in \Omega^c.$$

To find the unknown density functions $\boldsymbol{\mu} \in \mathbf{H}_\parallel^{-1/2}(\operatorname{div}_\Gamma, \Gamma)$ and $\boldsymbol{\lambda} \in \mathbf{H}_\perp^{-1/2}(\operatorname{curl}_\Gamma, \Gamma)$ we have to formulate appropriate boundary integral equations which can be derived from the Dirichlet boundary condition (1.2). For this we first use the trace operators γ_D and γ_N as given in Theorem 2.1 to define related boundary integral operators; in

particular for the interior trace we obtain

$$\begin{aligned} \gamma_D \Psi_S^\kappa \boldsymbol{\mu}(x) &=: \mathbf{S}_\kappa \boldsymbol{\mu}(x), \\ \gamma_D \Psi_M^\kappa \boldsymbol{\lambda}(x) &=: \left(\frac{1}{2} I + \mathbf{C}_\kappa \right) \boldsymbol{\lambda}(x), \\ \gamma_N \Psi_S^\kappa \boldsymbol{\mu}(x) &=: \left(\frac{1}{2} I + \mathbf{B}_\kappa \right) \boldsymbol{\mu}(x), \\ \gamma_N \Psi_M^\kappa \boldsymbol{\lambda}(x) &=: \mathbf{N}_\kappa \boldsymbol{\lambda}(x), \end{aligned}$$

while for the exterior trace we get

$$\begin{aligned} \gamma_D^c \Psi_S^\kappa \boldsymbol{\mu}(x) &=: \mathbf{S}_\kappa \boldsymbol{\mu}(x), \\ \gamma_D^c \Psi_M^\kappa \boldsymbol{\lambda}(x) &=: \left(-\frac{1}{2} I + \mathbf{C}_\kappa \right) \boldsymbol{\lambda}(x), \\ \gamma_N^c \Psi_S^\kappa \boldsymbol{\mu}(x) &=: \left(-\frac{1}{2} I + \mathbf{B}_\kappa \right) \boldsymbol{\mu}(x), \\ \gamma_N^c \Psi_M^\kappa \boldsymbol{\lambda}(x) &=: \mathbf{N}_\kappa \boldsymbol{\lambda}(x). \end{aligned}$$

Note that

$$\mathbf{S}_\kappa : \mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma) \rightarrow \mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma)$$

and

$$\mathbf{N}_\kappa : \mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma) \rightarrow \mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma).$$

Moreover, with respect to the complex duality pairing

$$\langle \boldsymbol{\lambda}, \boldsymbol{\mu} \rangle = \int_\Gamma \boldsymbol{\lambda}(x) \cdot \overline{\boldsymbol{\mu}(x)} \, ds_x,$$

we have for $\kappa \in \mathbb{R} \setminus \{0\}$

$$\langle \mathbf{S}_\kappa \boldsymbol{\mu}, \mathbf{w} \rangle = \langle \boldsymbol{\mu}, \mathbf{S}_{-\kappa} \mathbf{w} \rangle \quad \text{for all } \boldsymbol{\mu}, \mathbf{w} \in \mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma),$$

$$\langle \mathbf{N}_\kappa \boldsymbol{\lambda}, \mathbf{v} \rangle = \langle \boldsymbol{\lambda}, \mathbf{N}_{-\kappa} \mathbf{v} \rangle \quad \text{for all } \boldsymbol{\lambda}, \mathbf{v} \in \mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma),$$

while the double layer potentials \mathbf{C}_κ and \mathbf{B}_κ are related to each other as follows.

LEMMA 2.3. *For all $\boldsymbol{\mu} \in \mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma)$ and $\boldsymbol{\lambda} \in \mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma)$ there holds*

$$\langle \mathbf{B}_\kappa \boldsymbol{\mu}, \boldsymbol{\lambda} \rangle = -\langle \boldsymbol{\mu}, \mathbf{C}_{-\kappa} \boldsymbol{\lambda} \rangle \quad \text{for all } \kappa \in \mathbb{R} \setminus \{0\}.$$

Proof. Since $\mathbf{U} = \Psi_S^\kappa \boldsymbol{\mu}$ and $\mathbf{V} = \Psi_M^{-\kappa} \boldsymbol{\lambda}$ are solutions of the homogeneous Maxwell equations, we can write Green's first formula (1.4) for the bounded domain Ω as

$$\begin{aligned} \int_\Omega \text{curl } \mathbf{U} \cdot \text{curl } \overline{\mathbf{V}} \, dx &= \int_\Omega \text{curl } \text{curl } \mathbf{U} \cdot \overline{\mathbf{V}} \, dx + \langle \gamma_N \mathbf{U}, \gamma_D \mathbf{V} \rangle \\ &= \int_\Omega \kappa^2 \mathbf{U} \cdot \overline{\mathbf{V}} \, dx + \langle \gamma_N \mathbf{U}, \gamma_D \mathbf{V} \rangle \end{aligned}$$

and

$$\int_{\Omega} \mathbf{curl} \bar{\mathbf{V}} \cdot \mathbf{curl} \mathbf{U} dx = \int_{\Omega} \kappa^2 \bar{\mathbf{V}} \cdot \mathbf{U} dx + \langle \gamma_N \bar{\mathbf{V}}, \gamma_D \bar{\mathbf{U}} \rangle.$$

Hence we first conclude

$$\langle \gamma_N \mathbf{U}, \gamma_D \mathbf{V} \rangle = \langle \gamma_N \bar{\mathbf{V}}, \gamma_D \bar{\mathbf{U}} \rangle.$$

On the other hand, for a bounded domain $B_r \setminus \bar{\Omega}$ we have

$$\int_{B_r \setminus \bar{\Omega}} \mathbf{curl} \mathbf{U} \cdot \mathbf{curl} \bar{\mathbf{V}} dx = \int_{B_r \setminus \bar{\Omega}} \kappa^2 \mathbf{U} \cdot \bar{\mathbf{V}} dx + \int_{\partial B_r} \gamma_N \mathbf{U} \cdot \gamma_D \bar{\mathbf{V}} ds_x - \langle \gamma_N^c \mathbf{U}, \gamma_D^c \mathbf{V} \rangle$$

and

$$\int_{B_r \setminus \bar{\Omega}} \mathbf{curl} \bar{\mathbf{V}} \cdot \mathbf{curl} \mathbf{U} dx = \int_{B_r \setminus \bar{\Omega}} \kappa^2 \bar{\mathbf{V}} \cdot \mathbf{U} dx + \int_{\partial B_r} \gamma_N \bar{\mathbf{V}} \cdot \gamma_D \mathbf{U} ds_x - \langle \gamma_N^c \bar{\mathbf{V}}, \gamma_D^c \mathbf{U} \rangle.$$

Hence we also conclude

$$\langle \gamma_N^c \mathbf{U}, \gamma_D^c \mathbf{V} \rangle = \int_{\partial B_r} \gamma_N \mathbf{U} \cdot \gamma_D \bar{\mathbf{V}} ds_x - \int_{\partial B_r} \gamma_N \bar{\mathbf{V}} \cdot \gamma_D \mathbf{U} ds_x + \langle \gamma_N^c \bar{\mathbf{V}}, \gamma_D^c \mathbf{U} \rangle$$

and therefore, for $r \rightarrow \infty$,

$$\langle \gamma_N^c \mathbf{U}, \gamma_D^c \mathbf{V} \rangle = \langle \gamma_N^c \bar{\mathbf{V}}, \gamma_D^c \bar{\mathbf{U}} \rangle = \langle \gamma_N \bar{\mathbf{V}}, \gamma_D \bar{\mathbf{U}} \rangle = \langle \gamma_N \mathbf{U}, \gamma_D \mathbf{V} \rangle.$$

Note that $\mathbf{U} = \Psi_S^\kappa \boldsymbol{\mu}$ and $\bar{\mathbf{V}} = \overline{\Psi_M^{-\kappa} \boldsymbol{\lambda}} = \Psi_M^\kappa \bar{\boldsymbol{\lambda}}$ are both solutions of the homogeneous Maxwell equations (1.1) satisfying the radiation condition (1.3); see also [11, Lemma 3.10].

With the interior and exterior Neumann traces,

$$\gamma_N \mathbf{U} = \gamma_N \Psi_S^\kappa \boldsymbol{\mu} = \left(\frac{1}{2} I + B_\kappa \right) \boldsymbol{\mu}, \quad \gamma_N^c \mathbf{U} = \gamma_N \Psi_S^\kappa \boldsymbol{\mu} = \left(-\frac{1}{2} I + B_\kappa \right) \boldsymbol{\mu},$$

we further obtain

$$\gamma_N \mathbf{U} + \gamma_N^c \mathbf{U} = 2B_\kappa \boldsymbol{\mu}, \quad \gamma_N \mathbf{U} - \gamma_N^c \mathbf{U} = \boldsymbol{\mu}.$$

On the other hand, when considering the interior and exterior Dirichlet traces this gives

$$\gamma_D \mathbf{V} = \gamma_D \Psi_M^{-\kappa} \boldsymbol{\lambda} = \left(\frac{1}{2} I + C_{-\kappa} \right) \boldsymbol{\lambda}, \quad \gamma_D^c \mathbf{V} = \left(-\frac{1}{2} I + C_{-\kappa} \right) \boldsymbol{\lambda},$$

and therefore

$$\gamma_D \mathbf{V} + \gamma_D^c \mathbf{V} = 2C_{-\kappa} \boldsymbol{\lambda}, \quad \gamma_D \mathbf{V} - \gamma_D^c \mathbf{V} = \boldsymbol{\lambda}.$$

Hence we finally obtain

$$\begin{aligned} 2\langle B_\kappa \boldsymbol{\mu}, \boldsymbol{\lambda} \rangle &= \langle \gamma_N \mathbf{U} + \gamma_N^c \mathbf{U}, \gamma_D \mathbf{V} - \gamma_D^c \mathbf{V} \rangle \\ &= \langle \gamma_N \mathbf{U}, \gamma_D \mathbf{V} \rangle + \langle \gamma_N^c \mathbf{U}, \gamma_D \mathbf{V} \rangle - \langle \gamma_N \mathbf{U}, \gamma_D^c \mathbf{V} \rangle - \langle \gamma_N^c \mathbf{U}, \gamma_D^c \mathbf{V} \rangle \\ &= \langle \gamma_N^c \mathbf{U}, \gamma_D^c \mathbf{V} \rangle + \langle \gamma_N^c \mathbf{U}, \gamma_D \mathbf{V} \rangle - \langle \gamma_N \mathbf{U}, \gamma_D^c \mathbf{V} \rangle - \langle \gamma_N \mathbf{U}, \gamma_D \mathbf{V} \rangle \\ &= \langle \gamma_N^c \mathbf{U} - \gamma_N \mathbf{U}, \gamma_D^c \mathbf{V} + \gamma_D \mathbf{V} \rangle \\ &= -2\langle \boldsymbol{\mu}, C_{-\kappa} \boldsymbol{\lambda} \rangle. \quad \square \end{aligned}$$

When using the single layer potential (2.2) we have to find $\boldsymbol{\mu} \in \mathbf{H}_{\parallel}^{-1/2}(\text{div}_{\Gamma}, \Gamma)$ by solving the boundary integral equation

$$(2.4) \quad \mathbf{S}_{\kappa} \boldsymbol{\mu}(x) = \mathbf{g}(x) \quad \text{for } x \in \Gamma,$$

while for the double layer potential (2.3) $\boldsymbol{\lambda} \in \mathbf{H}_{\perp}^{-1/2}(\text{curl}_{\Gamma}, \Gamma)$ is the solution of the boundary integral equation

$$(2.5) \quad -\frac{1}{2} \boldsymbol{\lambda}(x) + \mathbf{C}_{\kappa} \boldsymbol{\lambda}(x) = \mathbf{g}(x) \quad \text{for } x \in \Gamma.$$

When applying the exterior Dirichlet and the exterior Neumann traces to the Stratton–Chu representation formula (2.1) we obtain a system of boundary integral equations,

$$(2.6) \quad \begin{aligned} \gamma_D^c \mathbf{U} &= -\mathbf{S}_{\kappa} \gamma_N^c \mathbf{U} &+ (\frac{1}{2}I - \mathbf{C}_{\kappa}) \gamma_D^c \mathbf{U}, \\ \gamma_N^c \mathbf{U} &= (\frac{1}{2}I - \mathbf{B}_{\kappa}) \gamma_N^c \mathbf{U} &+ -\mathbf{N}_{\kappa} \gamma_D^c \mathbf{U}. \end{aligned}$$

In particular, to describe the solution of the exterior Dirichlet boundary value problem (1.1)–(1.3) we may use the first boundary integral equation in (2.6) to find $\gamma_N^c \mathbf{U} \in \mathbf{H}_{\parallel}^{-1/2}(\text{div}_{\Gamma}, \Gamma)$ such that

$$(2.7) \quad \mathbf{S}_{\kappa} \gamma_N^c \mathbf{U}(x) = -\frac{1}{2} \mathbf{g}(x) - \mathbf{C}_{\kappa} \mathbf{g}(x) \quad \text{for } x \in \Gamma.$$

PROPOSITION 2.4 (see [12]). *Let $\lambda = \kappa^2$ be an eigenvalue of the interior Maxwell eigenvalue problem*

$$\mathbf{curl} \mathbf{curl} \mathbf{U}_{\lambda}(x) = \lambda \mathbf{U}_{\lambda}(x) \quad \text{for } x \in \Omega.$$

Then, in the case of the interior Dirichlet eigenvalue problem

$$(2.8) \quad \mathbf{curl} \mathbf{curl} \mathbf{U}_{\lambda}(x) = \lambda \mathbf{U}_{\lambda}(x) \quad \text{for } x \in \Omega, \quad \gamma_D \mathbf{U}_{\lambda}(x) = 0 \quad \text{for } x \in \Gamma,$$

$\gamma_N \mathbf{U}_{\lambda}(x)$ is in the kernel of \mathbf{S}_{κ} and $(-\frac{1}{2}I + \mathbf{B}_{\kappa})$, i.e.,

$$\mathbf{S}_{\kappa} \gamma_N \mathbf{U}_{\lambda} = 0, \quad \left(\frac{1}{2}I - \mathbf{B}_{\kappa} \right) \gamma_N \mathbf{U}_{\lambda} = 0.$$

On the other hand, if κ^2 is not an eigenvalue of the interior Dirichlet eigenvalue problem (2.8), then $\mathbf{S}_{\kappa} \mathbf{w} = 0$ implies $\mathbf{w} = 0$.

Moreover, in the case of the interior Neumann eigenvalue problem

$$(2.9) \quad \mathbf{curl} \mathbf{curl} \mathbf{V}_{\lambda}(x) = \lambda \mathbf{V}_{\lambda}(x) \quad \text{for } x \in \Omega, \quad \gamma_N \mathbf{V}_{\lambda}(x) = 0 \quad \text{for } x \in \Gamma,$$

$\gamma_D \mathbf{V}_{\lambda}(x)$ is in the kernel of \mathbf{N}_{κ} and $(\frac{1}{2}I - \mathbf{C}_{\kappa})$, i.e.,

$$\mathbf{N}_{\kappa} \gamma_D \mathbf{V}_{\lambda} = 0, \quad \left(\frac{1}{2}I - \mathbf{C}_{\kappa} \right) \gamma_D \mathbf{V}_{\lambda} = 0.$$

Hence, if $\lambda = \kappa^2$ is an eigenvalue of the interior Maxwell eigenvalue problem, we conclude that the single layer potential operator \mathbf{S}_{κ} is not invertible, and therefore

the boundary integral equations (2.4) and (2.7) are in general not solvable. However, due to

$$\left\langle -\frac{1}{2}\mathbf{g} - C_\kappa\mathbf{g}, \gamma_N\mathbf{U}_\lambda \right\rangle = \left\langle \mathbf{g}, \left(-\frac{1}{2}I + B_{-\kappa}\right)\gamma_N\mathbf{U}_\lambda \right\rangle = 0$$

we conclude that the right-hand side of the boundary integral equation (2.7) is in the image of the single layer potential S_κ ; i.e., the boundary integral equation (2.7) of the direct approach is solvable, but the solution is not unique. Moreover, the boundary integral operator $\frac{1}{2}I - C_\kappa$ is also not invertible, and therefore the boundary integral equation (2.5) of the indirect approach is in general not solvable.

To overcome the problem of nonsolvability of boundary integral equations due to interior eigenfrequencies one may use a combined approach such as the formulation of Brakhage and Werner, who introduced a combined field integral equation for the acoustic scattering problem [1]. The same idea was used by Panich in [24] for the electromagnetic case. In general, the idea is to consider complex linear combinations of the single and double layer potential, i.e.,

$$\mathbf{U}(x) = -i\eta\Psi_S^\kappa\mathbf{w}(x) - \Psi_M^\kappa\mathbf{w}(x) \quad \text{for } x \in \Omega^c,$$

where $\eta \in \mathbb{R}_+$ is some parameter to be chosen. The unknown density $\mathbf{w} \in L_2(\Gamma)$ can then be determined from the resulting boundary integral equation

$$(2.10) \quad \gamma_D^c\mathbf{U}(x) = -i\eta S_\kappa\mathbf{w}(x) + \left(\frac{1}{2}I - C_\kappa\right)\mathbf{w}(x) = \mathbf{g}(x) \quad \text{for } x \in \Gamma$$

which can be proved to be uniquely solvable if the boundary $\Gamma = \partial\Omega$ is sufficiently smooth. But this proof is essentially based on the compactness of the double layer potential operator C_κ which is not satisfied if Ω is a Lipschitz polyhedron. Hence one may introduce a regularization operator $B : \mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma) \rightarrow \mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma)$ such that the stabilized boundary integral equation

$$(2.11) \quad \gamma_D^c\mathbf{U}(x) = -i\eta S_\kappa\mathbf{w}(x) + \left(\frac{1}{2}I - C_\kappa\right)B\mathbf{w}(x) = \mathbf{g}(x) \quad \text{for } x \in \Gamma$$

admits a unique solution $w \in \mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma)$. A suitable compact operator B was introduced by Buffa and Hiptmair in [9]. The unique solvability of the stabilized boundary integral equation (2.11) is then based on a generalized Gårding inequality for the single layer potential S_κ and on the injectivity of the composed boundary integral operator in (2.11).

In the next section we will describe an alternative approach which generalizes modified boundary integral equations for the Helmholtz case [14]. To analyze the proposed modified boundary integral formulation we will need some auxiliary results as given in the following.

Due to the boundary integral equations (2.6) we define, for general $\sigma \in \mathbb{C}$, the Calderon projector

$$\mathcal{C} = \begin{pmatrix} \frac{1}{2}I - C_\sigma & -S_\sigma \\ -N_\sigma & \frac{1}{2}I - B_\sigma \end{pmatrix}$$

which satisfies the projection property

$$(2.12) \quad \mathcal{C}^2 \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = \mathcal{C} \begin{pmatrix} \lambda \\ \mu \end{pmatrix}$$

for all $\boldsymbol{\lambda} \in \mathbf{H}_{\perp}^{-1/2}(\text{curl}_{\Gamma}, \Gamma)$ and $\boldsymbol{\mu} \in \mathbf{H}_{\parallel}^{-1/2}(\text{div}_{\Gamma}, \Gamma)$. As a corollary of the projection property (2.12) we then conclude the relations

$$(2.13) \quad S_{\sigma} N_{\sigma} = \frac{1}{4} I - C_{\sigma}^2,$$

$$(2.14) \quad N_{\sigma} S_{\sigma} = \frac{1}{4} I - B_{\sigma}^2,$$

$$(2.15) \quad -N_{\sigma} C_{\sigma} = B_{\sigma} N_{\sigma},$$

$$(2.16) \quad -C_{\sigma} S_{\sigma} = S_{\sigma} B_{\sigma}.$$

Note that the case $\sigma = \kappa \in \mathbb{R}$ corresponds to the Maxwell equation (1.1), while the purely imaginary case $\sigma = i\kappa, \kappa \in \mathbb{R}$, corresponds to the Yukawa-type equation

$$\mathbf{curl} \mathbf{curl} \mathbf{U}(x) + \kappa^2 \mathbf{U}(x) = 0 \quad \text{for } x \in \Omega^c,$$

and the associated fundamental solution is given by

$$g_{i\kappa}(x, y) = \frac{1}{4\pi} \frac{e^{-\kappa|x-y|}}{|x-y|}.$$

In this case, i.e., for $\sigma = i\kappa, \kappa \in \mathbb{R}$, the single layer boundary integral operator S_{σ} and the hypersingular integral operator N_{σ} are self-adjoint with respect to the complex duality pairing, while for the related double layer potentials we have

$$\langle B_{\sigma} \boldsymbol{\mu}, \boldsymbol{\lambda} \rangle = -\langle \boldsymbol{\mu}, C_{\sigma} \boldsymbol{\lambda} \rangle.$$

If the single layer potential operator S_{σ} is invertible, we can define the Steklov–Poincaré operator

$$(2.17) \quad T_{\sigma} := S_{\sigma}^{-1} \left(\frac{1}{2} I - C_{\sigma} \right) \quad : \quad \mathbf{H}_{\perp}^{-1/2}(\text{curl}_{\Gamma}, \Gamma) \rightarrow \mathbf{H}_{\parallel}^{-1/2}(\text{div}_{\Gamma}, \Gamma)$$

which allows an alternative symmetric representation

$$(2.18) \quad T_{\sigma} := N_{\sigma} + \left(\frac{1}{2} I + B_{\sigma} \right) S_{\sigma}^{-1} \left(\frac{1}{2} I - C_{\sigma} \right).$$

THEOREM 2.5. *The operators*

$$A_0 = \gamma_D^c \boldsymbol{\Psi}_A^0 : \mathbf{H}_{\parallel}^{-1/2}(\Gamma) \rightarrow \mathbf{H}_{\parallel}^{1/2}(\Gamma)$$

and

$$V_0 = \gamma_D^c \Psi_V^0 : H^{-1/2}(\Gamma) \rightarrow H^{1/2}(\Gamma)$$

are self-adjoint as well as $\mathbf{H}_{\parallel}^{-1/2}(\Gamma)$ - and $H^{-1/2}(\Gamma)$ -elliptic, respectively. Moreover, for $\sigma = i\kappa, \kappa \in \mathbb{R}_+$, the single layer potential

$$S_{\sigma} : \mathbf{H}_{\parallel}^{-1/2}(\text{div}_{\Gamma}, \Gamma) \rightarrow \mathbf{H}_{\perp}^{-1/2}(\text{curl}_{\Gamma}, \Gamma)$$

is $\mathbf{H}_{\parallel}^{-1/2}(\text{div}_{\Gamma}, \Gamma)$ -elliptic and self-adjoint.

Proof. For the mapping properties of the boundary integral operators A_0 and V_0 see [6, Theorem 4]. The ellipticity of S_{σ} follows as in the case of the Laplace operator; see, e.g., [27]. \square

3. Modified boundary integral equations. In this section we propose an alternative approach of a modified boundary integral equation to solve the exterior Dirichlet boundary value problem (1.1)–(1.3). Because of symmetry reasons we choose

$$B = S_0^{*-1} \left(\frac{1}{2}I + B_{-\kappa} \right) : \mathbf{H}_{\parallel}^{-1/2}(\text{div}_{\Gamma}, \Gamma) \rightarrow \mathbf{H}_{\perp}^{-1/2}(\text{curl}_{\Gamma}, \Gamma),$$

whereas $S_0^* : \mathbf{H}_{\perp}^{-1/2}(\text{curl}_{\Gamma}, \Gamma) \rightarrow \mathbf{H}_{\parallel}^{-1/2}(\text{div}_{\Gamma}, \Gamma)$ is given by

$$S_0^* \mathbf{u} := \mathbf{n} \times A_0(\mathbf{u} \times \mathbf{n}) + \text{curl}_{\Gamma} V_0 \text{curl}_{\Gamma} \mathbf{u}.$$

By using Theorem 2.5 one can prove that S_0^* is $\mathbf{H}_{\perp}^{-1/2}(\text{curl}_{\Gamma}, \Gamma)$ -elliptic and self-adjoint.

Now we can describe the solution of the exterior Dirichlet boundary value problem (1.1)–(1.3) by

$$\mathbf{U}(x) = \Psi_S^{\kappa} \mathbf{w}(x) - i\eta \Psi_M^{\kappa} B \mathbf{w}(x) \quad \text{for } x \in \Omega^c.$$

When applying the exterior Dirichlet trace we can find the unknown density $\mathbf{w} \in \mathbf{H}_{\parallel}^{-1/2}(\text{div}_{\Gamma}, \Gamma)$ from the modified boundary integral equation

$$(3.1) \quad Z_{\kappa} \mathbf{w}(x) = S_{\kappa} \mathbf{w}(x) + i\eta \left(\frac{1}{2}I - C_{\kappa} \right) S_0^{*-1} \left(\frac{1}{2}I + B_{-\kappa} \right) \mathbf{w}(x) = \mathbf{g}(x) \quad \text{for } x \in \Gamma.$$

To establish the unique solvability of the modified boundary integral equation (3.1) we first prove that Z_{κ} is coercive. In contrast to the approach in [14] we show the coercivity in the second part, because the single layer potential S_{κ} does not fulfill a Gårding inequality.

To prove the coercivity of the operator Z_{κ} we first define an appropriate equivalent norm in $\mathbf{H}_{\perp}^{-1/2}(\text{curl}_{\Gamma}, \Gamma)$, see Theorem 2.5; i.e., for $\sigma = i\kappa, \kappa \in \mathbb{R}_+$,

$$\|\mathbf{u}\|_{S_{\sigma}^{-1}} := \sqrt{\langle S_{\sigma}^{-1} \mathbf{u}, \mathbf{u} \rangle}, \quad \mathbf{u} \in \mathbf{H}_{\perp}^{-1/2}(\text{curl}_{\Gamma}, \Gamma).$$

As in the case of a formally elliptic partial differential operator [28] we can prove a contraction property of the associated double layer potential $\frac{1}{2}I - C_{\sigma}, \sigma = i\kappa, \kappa \in \mathbb{R}_+$.

THEOREM 3.1. *For all $\mathbf{u} \in \mathbf{H}_{\perp}^{-1/2}(\text{curl}_{\Gamma}, \Gamma)$ and for $\sigma = i\kappa, \kappa \in \mathbb{R}_+$, there holds*

$$(1 - c_K) \|\mathbf{u}\|_{S_{\sigma}^{-1}} \leq \left\| \left(\frac{1}{2}I - C_{\sigma} \right) \mathbf{u} \right\|_{S_{\sigma}^{-1}} \leq c_K \|\mathbf{u}\|_{S_{\sigma}^{-1}},$$

where

$$c_K = \frac{1}{2} + \sqrt{\frac{1}{4} - c_1^S c_1^N} < 1,$$

and c_1^S, c_1^N are the ellipticity constants of the single layer potential S_{σ} and of the hypersingular operator N_{σ} .

Proof. The proof follows as in the case of a formally elliptic partial differential operator; see [28, Theorem 3.1].

For $\mathbf{u} \in \mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma)$ with $\|\mathbf{u}\|_{\mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma)} > 0$ we first have

$$\left\| \left(\frac{1}{2}I - C_\sigma \right) \mathbf{u} \right\|_{S_\sigma^{-1}}^2 = \left\langle S_\sigma^{-1} \left(\frac{1}{2}I - C_\sigma \right) \mathbf{u}, \left(\frac{1}{2}I - C_\sigma \right) \mathbf{u} \right\rangle = \langle T_\sigma \mathbf{u}, \mathbf{u} \rangle - \langle N_\sigma \mathbf{u}, \mathbf{u} \rangle,$$

where the Steklov–Poincaré operator T_σ is defined as in (2.18). Let

$$J : \mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma) \rightarrow \mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma)$$

be the Riesz operator; then

$$A := JS_\sigma^{-1} : \mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma) \rightarrow \mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma)$$

is self-adjoint and $\mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma)$ -elliptic.

Hence we can consider the splitting $A = A^{1/2}A^{1/2}$ to obtain

$$\begin{aligned} \langle T_\sigma \mathbf{u}, \mathbf{u} \rangle &= \left\langle S_\sigma^{-1} \left(\frac{1}{2}I - C_\sigma \right) \mathbf{u}, \mathbf{u} \right\rangle \\ &= \left\langle JS_\sigma^{-1} \left(\frac{1}{2}I - C_\sigma \right) \mathbf{u}, \mathbf{u} \right\rangle_{\mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma)} \\ &= \left\langle A^{1/2} \left(\frac{1}{2}I - C_\sigma \right) \mathbf{u}, A^{1/2} \mathbf{u} \right\rangle_{\mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma)} \\ &\leq \left\| A^{1/2} \left(\frac{1}{2}I - C_\sigma \right) \mathbf{u} \right\|_{\mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma)} \left\| A^{1/2} \mathbf{u} \right\|_{\mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma)}. \end{aligned}$$

With

$$\begin{aligned} \|A^{1/2} \mathbf{v}\|_{\mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma)}^2 &= \langle A^{1/2} \mathbf{v}, A^{1/2} \mathbf{v} \rangle_{\mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma)} \\ &= \langle JS_\sigma^{-1} \mathbf{v}, \mathbf{v} \rangle_{\mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma)} = \langle S_\sigma^{-1} \mathbf{v}, \mathbf{v} \rangle = \|\mathbf{v}\|_{S_\sigma^{-1}}^2 \end{aligned}$$

we then obtain

$$\langle T_\sigma \mathbf{u}, \mathbf{u} \rangle \leq \left\| \left(\frac{1}{2}I - C_\sigma \right) \mathbf{u} \right\|_{S_\sigma^{-1}} \|\mathbf{u}\|_{S_\sigma^{-1}}.$$

On the other hand, for the hypersingular boundary integral operator we have

$$\langle N_\sigma \mathbf{u}, \mathbf{u} \rangle \geq c_1^N \|\mathbf{u}\|_{\mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma)}^2 \geq c_1^N c_1^S \langle S_\sigma^{-1} \mathbf{u}, \mathbf{u} \rangle = c_1^N c_1^S \|\mathbf{u}\|_{S_\sigma^{-1}}^2.$$

Altogether, this gives

$$\begin{aligned} \left\| \left(\frac{1}{2}I - C_\sigma \right) \mathbf{u} \right\|_{S_\sigma^{-1}}^2 &= \langle T_\sigma \mathbf{u}, \mathbf{u} \rangle - \langle N_\sigma \mathbf{u}, \mathbf{u} \rangle \\ &\leq \left\| \left(\frac{1}{2}I - C_\sigma \right) \mathbf{u} \right\|_{S_\sigma^{-1}} \|\mathbf{u}\|_{S_\sigma^{-1}} - c_1^N c_1^S \|\mathbf{u}\|_{S_\sigma^{-1}}^2, \end{aligned}$$

which is equivalent to

$$\left(\frac{a}{b}\right)^2 - \frac{a}{b} + c_1^N c_1^S \leq 0,$$

where

$$a := \left\| \left(\frac{1}{2}I - C_\sigma \right) \mathbf{u} \right\|_{S_\sigma^{-1}} \geq 0, \quad b := \|\mathbf{u}\|_{S_\sigma^{-1}} > 0.$$

Hence we finally conclude

$$\frac{1}{2} - \sqrt{\frac{1}{4} - c_1^N c_1^S} \leq \frac{a}{b} \leq \frac{1}{2} + \sqrt{\frac{1}{4} - c_1^N c_1^S},$$

which gives the assertion. \square

A similar estimate can also be shown for the operator $\frac{1}{2}I + C_\sigma$.

THEOREM 3.2. *For $\mathbf{v} \in \mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma)$, $\sigma = i\kappa$, $\kappa \in \mathbb{R}_+$, there holds*

$$(1 - c_K) \|\mathbf{v}\|_{S_\sigma^{-1}} \leq \left\| \left(\frac{1}{2}I + C_\sigma \right) \mathbf{v} \right\|_{S_\sigma^{-1}} \leq c_K \|\mathbf{v}\|_{S_\sigma^{-1}}.$$

Proof. The proof follows as in the case of a formally elliptic partial differential operator; see [28, Theorem 3.2].

With the contraction property of $\frac{1}{2}I - C_\sigma$ we obtain

$$\begin{aligned} \|\mathbf{v}\|_{S_\sigma^{-1}} &= \left\| \left(\frac{1}{2}I + C_\sigma \right) \mathbf{v} + \left(\frac{1}{2}I - C_\sigma \right) \mathbf{v} \right\|_{S_\sigma^{-1}} \\ &\leq \left\| \left(\frac{1}{2}I + C_\sigma \right) \mathbf{v} \right\|_{S_\sigma^{-1}} + \left\| \left(\frac{1}{2}I - C_\sigma \right) \mathbf{v} \right\|_{S_\sigma^{-1}} \\ &\leq \left\| \left(\frac{1}{2}I + C_\sigma \right) \mathbf{v} \right\|_{S_\sigma^{-1}} + c_K \|\mathbf{v}\|_{S_\sigma^{-1}} \end{aligned}$$

and therefore the first inequality. On the other hand, by using the representations (2.17) and (2.18) we get

$$\begin{aligned} \left\| \left(\frac{1}{2}I + C_\sigma \right) \mathbf{v} \right\|_{S_\sigma^{-1}}^2 &= \left\| \left(I - \left(\frac{1}{2}I - C_\sigma \right) \right) \mathbf{v} \right\|_{S_\sigma^{-1}}^2 \\ &= \|\mathbf{v}\|_{S_\sigma^{-1}}^2 + \left\| \left(\frac{1}{2}I - C_\sigma \right) \mathbf{v} \right\|_{S_\sigma^{-1}}^2 - 2 \left\langle S_\sigma^{-1} \left(\frac{1}{2}I - C_\sigma \right) \mathbf{v}, \mathbf{v} \right\rangle \\ &= \|\mathbf{v}\|_{S_\sigma^{-1}}^2 + \left\| \left(\frac{1}{2}I - C_\sigma \right) \mathbf{v} \right\|_{S_\sigma^{-1}}^2 - 2 \langle T_\sigma \mathbf{v}, \mathbf{v} \rangle \\ &= \|\mathbf{v}\|_{S_\sigma^{-1}}^2 - \left\| \left(\frac{1}{2}I - C_\sigma \right) \mathbf{v} \right\|_{S_\sigma^{-1}}^2 - 2 \langle N_\sigma \mathbf{v}, \mathbf{v} \rangle \\ &\leq [1 - (1 - c_K)^2 - 2c_1^S c_1^N] \|\mathbf{v}\|_{S_\sigma^{-1}}^2 = c_K^2 \|\mathbf{v}\|_{S_\sigma^{-1}}^2 \end{aligned}$$

and therefore the upper estimate. \square

As for the operators $\frac{1}{2}I \pm C_\sigma$ we can prove related estimates for the operators $\frac{1}{2}I \pm B_\sigma$ when considering an equivalent norm in $\mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma)$ which is induced by the single layer potential S_σ ; i.e., for $\sigma = i\kappa, \kappa \in \mathbb{R}_+$ there holds

$$(3.2) \quad (1 - c_K) \|\mathbf{w}\|_{S_\sigma} \leq \left\| \left(\frac{1}{2}I \pm B_\sigma \right) \mathbf{w} \right\|_{S_\sigma} \leq c_K \|\mathbf{w}\|_{S_\sigma}$$

for all $\mathbf{w} \in \mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma)$.

For $\mathbf{u} \in \mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma)$ and $\kappa \in \mathbb{R}_+$ we finally define the operator

$$S_{\kappa,0}\mathbf{u} := A_0\mathbf{u} - \frac{1}{\kappa^2} \nabla_\Gamma V_0 \text{div}_\Gamma \mathbf{u}.$$

Now we are able to prove the coercivity of the operator Z_κ .

THEOREM 3.3. *Let $\kappa \in \mathbb{R}_+$. The operator*

$$Z_\kappa = S_\kappa + i\eta \left(\frac{1}{2}I - C_\kappa \right) S_0^{*-1} \left(\frac{1}{2}I + B_{-\kappa} \right) : \mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma) \rightarrow \mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma)$$

satisfies a Gårding inequality; i.e., there holds

$$\Im[\langle Z_\kappa \boldsymbol{\mu}, \boldsymbol{\mu} \rangle + c_1(\boldsymbol{\mu}, \boldsymbol{\mu})] \geq c_Z \|\boldsymbol{\mu}\|_{\mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma)}^2$$

for all $\boldsymbol{\mu} \in \mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma)$ with a positive constant c_Z where $c_1(\boldsymbol{\mu}, \boldsymbol{\mu})$ is a compact bilinear form.

Proof. Since $\langle S_{\kappa,0}\mathbf{w}, \mathbf{w} \rangle$ is real, the same holds true for the duality product

$$\left\langle S_0^{*-1} \left(\frac{1}{2}I + B_{-\kappa} \right) \mathbf{w}, \left(\frac{1}{2}I + B_{-\kappa} \right) \mathbf{w} \right\rangle \in \mathbb{R}.$$

Because of the contraction property (3.2) we get, for $\sigma = i\kappa$,

$$\left\| \left(\frac{1}{2}I + B_\sigma \right) \mathbf{w} \right\|_{\mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma)} \geq c \|\mathbf{w}\|_{\mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma)}$$

for all $\mathbf{w} \in \mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma)$. Since the operator S_0^{*-1} is $\mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma)$ -elliptic, we have

$$\left\langle S_0^{*-1} \left(\frac{1}{2}I + B_\sigma \right) \mathbf{w}, \left(\frac{1}{2}I + B_\sigma \right) \mathbf{w} \right\rangle \geq c \|\mathbf{w}\|_{\mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma)}^2$$

for all $\mathbf{w} \in \mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma)$. The operator Z_κ can now be written in the following form:

$$Z_\kappa = S_{\kappa,0} + \underbrace{(S_\kappa - S_{\kappa,0})}_{\text{compact}} + i\eta \left(\left(\frac{1}{2}I - C_\sigma \right) S_0^{*-1} \left(\frac{1}{2}I + B_\sigma \right) + \underbrace{(C_\sigma - C_\kappa) S_0^{*-1} \left(\frac{1}{2}I + B_{-\kappa} \right) + \left(\frac{1}{2}I - C_\sigma \right) S_0^{*-1} (B_{-\kappa} - B_\sigma)}_{\text{compact}} \right),$$

which implies

$$\begin{aligned} & \Im [\langle Z_\kappa \mathbf{w}, \mathbf{w} \rangle + c_1(\mathbf{w}, \mathbf{w})] \\ &= \Im \left[\langle S_{\kappa,0} \mathbf{w}, \mathbf{w} \rangle + i\eta \left\langle S_0^{*-1} \left(\frac{1}{2}I + B_\sigma \right) \mathbf{w}, \left(\frac{1}{2}I + B_\sigma \right) \mathbf{w} \right\rangle \right] \\ &= \eta \left\langle S_0^{*-1} \left(\frac{1}{2}I + B_\sigma \right) \mathbf{w}, \left(\frac{1}{2}I + B_\sigma \right) \mathbf{w} \right\rangle \\ &\geq c \|\mathbf{w}\|_{\mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma)}^2. \end{aligned}$$

Note that the compactness of $S_\kappa - S_{\kappa,0}$, $C_\sigma - C_\kappa$, and $B_{-\kappa} - B_\sigma$ follows as for the Helmholtz case; see, e.g., [26, 27, 29]. \square

Hence, to use Fredholm’s alternative to establish the unique solvability of the modified boundary integral equation (3.1) it remains to prove the injectivity of the operator Z_κ . This can be done as for the Helmholtz equation; see [14].

THEOREM 3.4. *For a positive wave number $\kappa \in \mathbb{R}_+$ there holds*

$$\Im[\langle S_\kappa \mathbf{w}, \mathbf{w} \rangle] \geq 0$$

for all $\mathbf{w} \in \mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma)$.

Proof. Let $\mathbf{U}(x) = \Psi_S^\kappa \mathbf{w}(x)$, $x \in \Omega$, be a solution of the partial differential equation (1.1). From Green’s first formula (1.4) we then have

$$\int_\Omega [\mathbf{curl} \mathbf{U}(x) \cdot \mathbf{curl} \mathbf{V}(x) - \kappa^2 \mathbf{U}(x) \cdot \mathbf{V}(x)] dx = \int_\Gamma \gamma_N \mathbf{U}(x) \cdot \gamma_D \mathbf{V}(x) ds_x.$$

For $\mathbf{V} = \bar{\mathbf{U}}$ it follows that

$$\int_\Omega (|\mathbf{curl} \mathbf{U}(x)|^2 - \kappa^2 |\mathbf{U}(x)|^2) dx = \int_\Gamma \gamma_N \mathbf{U}(x) \cdot \gamma_D \bar{\mathbf{U}}(x) ds_x.$$

With

$$\gamma_N \Psi_S^\kappa \mathbf{w}(x) = \frac{1}{2} \mathbf{w}(x) + B_\kappa \mathbf{w}(x),$$

$$\gamma_D \Psi_S^\kappa \mathbf{w}(x) = S_\kappa \mathbf{w}(x),$$

we then obtain

$$\int_\Omega (|\mathbf{curl} \mathbf{U}(x)|^2 - \kappa^2 |\mathbf{U}(x)|^2) dx = \langle \gamma_N \mathbf{U}, \gamma_D \bar{\mathbf{U}} \rangle = \left\langle \frac{1}{2} \mathbf{w} + B_\kappa \mathbf{w}, S_\kappa \mathbf{w} \right\rangle.$$

To handle the exterior domain Ω^c we first consider the bounded domain $B_r \setminus \bar{\Omega}$,

$$\begin{aligned} & \int_{B_r \setminus \bar{\Omega}} (|\mathbf{curl} \mathbf{U}(x)|^2 - \kappa^2 |\mathbf{U}(x)|^2) dx \\ &= \int_{\partial B_r} \gamma_N \mathbf{U}(x) \cdot \gamma_D \bar{\mathbf{U}}(x) ds_x - \int_\Gamma \gamma_N^c \mathbf{U}(x) \cdot \gamma_D^c \bar{\mathbf{U}}(x) ds_x. \end{aligned}$$

For the exterior traces of $\mathbf{U}(x) = \Psi_S^\kappa \mathbf{w}(x)$, $x \in \Omega^c$, we have for $x \in \Gamma$

$$\begin{aligned} \gamma_N^c \Psi_S^\kappa \mathbf{w}(x) &= -\frac{1}{2} \mathbf{w}(x) + \mathbf{B}_\kappa \mathbf{w}(x), \\ \gamma_D^c \Psi_S^\kappa \mathbf{w}(x) &= \mathbf{S}_\kappa \mathbf{w}(x), \end{aligned}$$

and therefore

$$\begin{aligned} &\int_{B_r \setminus \bar{\Omega}} (|\mathbf{curl} \mathbf{U}(x)|^2 - \kappa^2 |\mathbf{U}(x)|^2) dx \\ &= \int_{\partial B_r} \gamma_N \mathbf{U}(x) \cdot \gamma_D \bar{\mathbf{U}}(x) ds_x + \left\langle \frac{1}{2} \mathbf{w} - \mathbf{B}_\kappa \mathbf{w}, \mathbf{S}_\kappa \mathbf{w} \right\rangle. \end{aligned}$$

Hence we find by summing up the above expressions

$$\int_{B_r} (|\mathbf{curl} \mathbf{U}(x)|^2 - \kappa^2 |\mathbf{U}(x)|^2) dx = \langle \mathbf{w}, \mathbf{S}_\kappa \mathbf{w} \rangle + \int_{\partial B_r} \gamma_N \mathbf{U}(x) \cdot \gamma_D \bar{\mathbf{U}}(x) ds_x,$$

and therefore

$$\Im[\langle \mathbf{w}, \mathbf{S}_\kappa \mathbf{w} \rangle] = -\Im \left[\int_{\partial B_r} \gamma_N \mathbf{U}(x) \cdot \gamma_D \bar{\mathbf{U}}(x) ds_x \right].$$

From the Silver–Müller radiation condition, i.e.,

$$\lim_{r=|x| \rightarrow 0} \int_{\partial B_r} |\mathbf{curl} \mathbf{U}(x) \times \mathbf{n} - i\kappa(\mathbf{n} \times \mathbf{U}(x)) \times \mathbf{n}|^2 ds_x = 0,$$

we further conclude

$$\begin{aligned} &\int_{\partial B_r} |\gamma_N \mathbf{U}(x) - i\kappa \gamma_D \mathbf{U}(x)|^2 ds_x \\ &= \int_{\partial B_r} \left(|\gamma_N \mathbf{U}(x)|^2 + |\kappa \gamma_D \mathbf{U}(x)|^2 - 2\Re[\gamma_N \mathbf{U}(x) \cdot \overline{i\kappa \gamma_D \mathbf{U}(x)}] \right) ds_x \\ &= \int_{\partial B_r} \left(|\gamma_N \mathbf{U}(x)|^2 + |\kappa \gamma_D \mathbf{U}(x)|^2 - 2\kappa \Im[\gamma_N \mathbf{U}(x) \cdot \gamma_D \bar{\mathbf{U}}(x)] \right) ds_x \\ &= \int_{\partial B_r} \left(|\gamma_N \mathbf{U}(x)|^2 + |\kappa \gamma_D \mathbf{U}(x)|^2 \right) ds_x + 2\kappa \Im[\langle \mathbf{w}, \mathbf{S}_\kappa \mathbf{w} \rangle] \rightarrow 0 \end{aligned}$$

as $r \rightarrow \infty$, which implies

$$2\kappa \Im[\langle \mathbf{w}, \mathbf{S}_\kappa \mathbf{w} \rangle] \leq 0$$

and thus

$$2\kappa \Im[\langle \mathbf{S}_\kappa \mathbf{w}, \mathbf{w} \rangle] \geq 0. \quad \square$$

Now we are in a position to prove the injectivity of Z_κ .

THEOREM 3.5. *For $\kappa \in \mathbb{R}_+$ and $\eta \in \mathbb{R}_+$ the modified boundary integral operator*

$$Z_\kappa = \mathbf{S}_\kappa + i\eta \left(\frac{1}{2} I - \mathbf{C}_\kappa \right) \mathbf{S}_0^{*-1} \left(\frac{1}{2} I + \mathbf{B}_{-\kappa} \right) : \mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma) \rightarrow \mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma)$$

is injective.

Proof. Let $w \in \mathbf{H}_{\parallel}^{-1/2}(\operatorname{div}_{\Gamma}, \Gamma)$ be a solution of the homogeneous equation

$$\mathbf{Z}_{\kappa} \mathbf{w}(x) = 0 \quad \text{for } x \in \Gamma.$$

Then it follows that

$$0 = \langle \mathbf{Z}_{\kappa} \mathbf{w}, \mathbf{w} \rangle = \langle \mathbf{S}_{\kappa} \mathbf{w}, \mathbf{w} \rangle + i\eta \left\langle \mathbf{S}_0^{*-1} \left(\frac{1}{2} I + \mathbf{B}_{-\kappa} \right) \mathbf{w}, \left(\frac{1}{2} I + \mathbf{B}_{-\kappa} \right) \mathbf{w} \right\rangle$$

and therefore

$$\Im \left[\langle \mathbf{S}_{\kappa} \mathbf{w}, \mathbf{w} \rangle + i\eta \left\langle \mathbf{S}_0^{*-1} \left(\frac{1}{2} I + \mathbf{B}_{-\kappa} \right) \mathbf{w}, \left(\frac{1}{2} I + \mathbf{B}_{-\kappa} \right) \mathbf{w} \right\rangle \right] = 0.$$

By using Theorem 3.4 we then get

$$\eta \left\langle \mathbf{S}_0^{*-1} \left(\frac{1}{2} I + \mathbf{B}_{-\kappa} \right) \mathbf{w}, \left(\frac{1}{2} I + \mathbf{B}_{-\kappa} \right) \mathbf{w} \right\rangle = -\Im[\langle \mathbf{S}_{\kappa} \mathbf{w}, \mathbf{w} \rangle] \leq 0,$$

and hence we conclude

$$\left(\frac{1}{2} I + \mathbf{B}_{-\kappa} \right) \mathbf{w} = 0.$$

But then we also have

$$\mathbf{S}_{\kappa} \mathbf{w}(x) = 0 \quad \text{for } x \in \Gamma,$$

which admits only a nontrivial solution $\mathbf{w} = \gamma_N \mathbf{U}_{\lambda}$ if $\kappa^2 = \lambda$ is an eigenvalue of the interior Dirichlet eigenvalue problem (2.8) implying

$$\left(\frac{1}{2} I - \mathbf{B}_{\pm\kappa} \right) \mathbf{w} = 0,$$

i.e.,

$$\left(\frac{1}{2} I + \mathbf{B}_{-\kappa} \right) \mathbf{w} = 0, \quad \left(\frac{1}{2} I - \mathbf{B}_{-\kappa} \right) \mathbf{w} = 0.$$

Hence we conclude $\mathbf{w} = 0$ for all frequencies $\kappa > 0$. □

When combining the coercivity (Theorem 3.3) and the injectivity (Theorem 3.4) of the operator \mathbf{Z}_{κ} we therefore conclude the unique solvability of the modified boundary integral equation (3.1). The related variational formulation is to find $\mathbf{w} \in \mathbf{H}_{\parallel}^{-1/2}(\operatorname{div}_{\Gamma}, \Gamma)$ such that

$$(3.3) \quad \langle \mathbf{S}_{\kappa} \mathbf{w}, \boldsymbol{\tau} \rangle + i\eta \left\langle \left(\frac{1}{2} I - \mathbf{C}_{\kappa} \right) \mathbf{S}_0^{*-1} \left(\frac{1}{2} I + \mathbf{B}_{-\kappa} \right) \mathbf{w}, \boldsymbol{\tau} \right\rangle = \langle \mathbf{g}, \boldsymbol{\tau} \rangle$$

is satisfied for all test functions $\boldsymbol{\tau} \in \mathbf{H}_{\parallel}^{-1/2}(\operatorname{div}_{\Gamma}, \Gamma)$. Note that the variational problem (3.3) has a similar structure as the symmetric boundary integral representation of the Steklov–Poincaré operator. Due to the composite structure a direct Galerkin discretization of (3.3) will not be possible. Hence we introduce

$$\mathbf{z} = \mathbf{S}_0^{*-1} \left(\frac{1}{2} I + \mathbf{B}_{-\kappa} \right) \mathbf{w} \in \mathbf{H}_{\perp}^{-1/2}(\operatorname{curl}_{\Gamma}, \Gamma),$$

which is the unique solution of the variational problem such that

$$\langle S_0^* \mathbf{z}, \mathbf{v} \rangle = \left\langle \left(\frac{1}{2} I + B_{-\kappa} \right) \mathbf{w}, \mathbf{v} \right\rangle$$

is satisfied for all $\mathbf{v} \in \mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma)$. Finally we obtain a saddle point formulation to find $(\mathbf{w}, \mathbf{z}) \in \mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma) \times \mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma)$ such that

$$(3.4) \quad \begin{aligned} \langle S_\kappa \mathbf{w}, \boldsymbol{\tau} \rangle &+ i\eta \langle (\frac{1}{2} I - C_\kappa) \mathbf{z}, \boldsymbol{\tau} \rangle &= \langle \mathbf{g}, \boldsymbol{\tau} \rangle \\ - \langle (\frac{1}{2} I + B_{-\kappa}) \mathbf{w}, \mathbf{v} \rangle &+ \langle S_0^* \mathbf{z}, \mathbf{v} \rangle &= 0 \end{aligned}$$

is satisfied for all $(\boldsymbol{\tau}, \mathbf{v}) \in \mathbf{H}_\parallel^{-1/2}(\text{div}_\Gamma, \Gamma) \times \mathbf{H}_\perp^{-1/2}(\text{curl}_\Gamma, \Gamma)$. Since the modified boundary integral equation (3.1) is the Schur complement system of the mixed formulation (3.4) the unique solvability of (3.4) follows immediately.

Remark 3.6. In this paper we just presented a modified boundary integral formulation for the exterior Dirichlet boundary value problem (1.1)–(1.3). For an exterior Neumann boundary value problem a similar modified formulation can be derived and analyzed as well [29].

4. Numerical example. As a numerical example to show the applicability of the proposed approach we consider the exterior Dirichlet boundary value problem (1.1)–(1.3) where $\Omega = (0, 1)^3$ is the unit cube whose boundary $\Gamma = \partial\Omega$ is decomposed into N triangular plane elements. For this domain we can easily deduce the eigenvalues and eigenfrequencies of the interior Dirichlet eigenvalue problem. In particular we will consider the smallest eigenvalue which corresponds to the wave number $k = \sqrt{2}\pi \approx 4.44288$. As exact solution of the exterior Dirichlet boundary value problem (1.1)–(1.3) we consider [2]

$$\mathbf{U}(x) = \left[\frac{\kappa^2 r^2 + \kappa r + 1}{r^3} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} - \frac{\kappa^2 r^2 + 3\kappa r + 3}{r^5} (x_1 - \hat{x}_1) \begin{pmatrix} x_1 - \hat{x}_1 \\ x_2 - \hat{x}_2 \\ x_3 - \hat{x}_3 \end{pmatrix} \right] e^{\kappa r}$$

for $x \in \Omega^c$, where the source point is $\hat{x} = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2})^\top \in \Omega$, and $r = |x - \hat{x}|$. For a comparison of different approaches we consider the indirect single layer potential ansatz leading to the boundary integral equation (2.4), the proposed modified formulation ($\eta = 1$) where we have to solve (3.1), and a direct approach which results in the boundary integral equation (2.7). In all cases the Galerkin discretization is done by using linear Raviart–Thomas elements; see, e.g., [2, 25] for details. The resulting linear systems are solved by a GMRES method with a relative error reduction of $\varepsilon = 10^{-8}$. Then we compute approximate solutions \mathbf{U}_h and the related pointwise error in the evaluation point $\bar{x} = (1.4, 1.8, 2.0)^\top \in \Omega^c$. All results are documented in Table 1.

It is obvious that the indirect single layer potential approach fails since the wave number k corresponds to an eigenvalue of the interior Dirichlet eigenvalue problem. The results of the modified formulation (3.1) and of the direct approach (2.7) are comparable in this example. However, for the latter one has to ensure a solvability condition also in the discrete case which requires in general the knowledge of the related eigenfrequency. Here we considered only a direct Galerkin discretization of (2.7) which may fail in more general situations.

TABLE 1
Number of GMRES iterations and pointwise error.

N	Indirect, (2.4)		Modified, (3.1)		Direct, (2.7)	
	Iter	$ \mathbf{U}(\bar{x}) - \mathbf{U}_h(\bar{x}) $	Iter	$ \mathbf{U}(\bar{x}) - \mathbf{U}_h(\bar{x}) $	Iter	$ \mathbf{U}(\bar{x}) - \mathbf{U}_h(\bar{x}) $
72	53	7.64	110	1.27632	53	0.64908
288	107	10.85	197	0.19541	107	0.19153
1152	238	15.52	280	0.04874	209	0.04677
4608	554	43.20	403	0.01308	469	0.01222
18432			665	0.00730	834	0.00529

Related to the numerical results there are several points to be discussed, first of all the numerical analysis to establish the quadratic order of pointwise convergence. Moreover, we have to investigate a suitable choice of the scaling parameter $\eta \in \mathbb{R}_+$ and the construction of efficient preconditioned iterative solution methods. It is obvious that these questions are strongly related to the case of exterior boundary value problems for the Helmholtz equation [15]. Note that the formulation corresponds to the symmetric formulation of boundary integral equations as used in domain decomposition methods, or to solve boundary value problems with boundary conditions of mixed Dirichlet and Neumann type [27].

5. Conclusions. In this paper we have described and analyzed a modified boundary integral equation to solve an exterior Dirichlet boundary value problem for the Maxwell system which is stable for all wave numbers. Note that a similar formulation can be given in the case of an exterior Neumann boundary value problem as well. The proposed regularization operator relies on boundary integral operators which are already available when considering standard boundary integral equations for the Maxwell system. The modified boundary integral equation is finally reformulated as a saddle point formulation which allows a direct Galerkin discretization. A first numerical example shows the applicability of the proposed approach.

In a forthcoming paper we will present the numerical analysis of the related boundary element method to solve the saddle point formulation (3.4). This may also include the use of fast boundary element methods, and the design of preconditioned iterative solution strategies to solve the resulting linear systems of algebraic equations.

Acknowledgment. The authors would like to express their thanks to the anonymous referees for many helpful hints and advice.

REFERENCES

- [1] H. BRAKHAGE AND P. WERNER, *Über das Dirichletsche Aussenraumproblem für die Helmholtzsche Schwingungsgleichung*, Arch. Math., 16 (1965), pp. 325–329.
- [2] J. BREUER, *Schnelle Randelementmethoden zur Simulation von elektrischen Wirbelstromfeldern sowie ihrer Wärmeproduktion und Kühlung*, Dissertation, Universität Stuttgart, Stuttgart, Germany, 2005.
- [3] A. BUFFA, *Remarks on the discretization of some noncoercive operator with applications to heterogeneous Maxwell equations*, SIAM J. Numer. Anal., 43 (2005), pp. 1–18.
- [4] A. BUFFA AND P. CIARLET, *On traces for functional spaces related to Maxwell's equations. I. An integration by parts formula in Lipschitz polyhedra*, Math. Methods Appl. Sci., 24 (2001), pp. 9–30.
- [5] A. BUFFA AND P. CIARLET, *On traces for functional spaces related to Maxwell's equations. II. Hodge decompositions on the boundary of Lipschitz polyhedra and applications*, Math. Methods Appl. Sci., 24 (2001), pp. 31–48.
- [6] A. BUFFA, M. COSTABEL, AND C. SCHWAB, *Boundary element methods for Maxwell's equations on non-smooth domains*, Numer. Math., 92 (2002), pp. 679–710.

- [7] A. BUFFA, M. COSTABEL, AND D. SHEEN, *On traces for $\mathbf{H}(\mathbf{curl}, \Omega)$ in Lipschitz domains*, J. Math. Anal. Appl., 276 (2002), pp. 845–867.
- [8] A. BUFFA AND R. HIPTMAIR, *Galerkin boundary element methods for electromagnetic scattering*, in Topics in Computational Wave Propagation, Lect. Notes Comput. Sci. Eng. 31, Springer, Berlin, 2003, pp. 83–124.
- [9] A. BUFFA AND R. HIPTMAIR, *A coercive combined field integral equation for electromagnetic scattering*, SIAM J. Numer. Anal., 42 (2004), pp. 621–640.
- [10] A. BUFFA AND R. HIPTMAIR, *Regularized combined field integral equations*, Numer. Math., 100 (2005), pp. 1–19.
- [11] A. BUFFA, R. HIPTMAIR, T. VON PETERSDORFF, AND C. SCHWAB, *Boundary element methods for Maxwell transmission problems in Lipschitz domains*, Numer. Math., 95 (2003), pp. 459–485.
- [12] D. COLTON AND R. KRESS, *Integral Equation Methods in Scattering Theory*, John Wiley and Sons, New York, 1983.
- [13] D. COLTON AND R. KRESS, *Inverse Acoustic and Electromagnetic Scattering Theory*, Appl. Math. Sci. 93, Springer, Berlin, 1998.
- [14] S. ENGLER AND O. STEINBACH, *Modified boundary integral formulations for the Helmholtz equation*, J. Math. Anal. Appl., 331 (2007), pp. 396–407.
- [15] S. ENGLER AND O. STEINBACH, *Stabilized boundary element methods for exterior Helmholtz problems*, Numer. Math., 110 (2008), pp. 145–160.
- [16] R. HIPTMAIR, *Symmetric coupling for eddy current problems*, SIAM J. Numer. Anal., 40 (2002), pp. 41–65.
- [17] R. HIPTMAIR, *Boundary element methods for eddy current computation*, in Computational Electromagnetics (Kiel, 2001), Lect. Notes Comput. Sci. Eng. 28, Springer, Berlin, 2003, pp. 103–126.
- [18] R. HIPTMAIR, *Coupling of finite elements and boundary elements in electromagnetic scattering*, SIAM J. Numer. Anal., 41 (2003), pp. 919–944.
- [19] R. HIPTMAIR AND C. SCHWAB, *Natural boundary element methods for the electric field integral equation on polyhedra*, SIAM J. Numer. Anal., 40 (2002), pp. 66–86.
- [20] G. C. HSIAO, *Mathematical foundations for the boundary field equation methods in acoustic and electromagnetic scattering*, in Analysis and Computational Methods in Scattering and Applied Mathematics. A Volume in the Memory of Ralph Ellis Kleinman, Chapman & Hall/CRC Res. Notes Math. 417, F. Santosa and I. Stakgold, eds., Chapman & Hall/CRC, Boca Raton, FL, 2000, pp. 149–163.
- [21] G. C. HSIAO AND R. E. KLEINMAN, *Mathematical foundations for error estimation in numerical solutions of integral equations in electromagnetics*, IEEE Trans. Antennas and Propagation, 45 (1997), pp. 316–328.
- [22] P. MONK, *Finite Element Methods for Maxwell's Equations*, Numer. Math. Sci. Comput., Oxford University Press, New York, 2003.
- [23] J.-C. NÉDÉLEC, *Acoustic and Electromagnetic Equations. Integral Representations for Harmonic Problems*, Appl. Math. Sci. 144, Springer, New York, 2001.
- [24] O. I. PANICH, *On the question of the solvability of the exterior boundary value problems for the wave equation and Maxwell's equations*, Uspekhi Mat. Nauk., 20 (1965), pp. 221–226 (in Russian).
- [25] P.-A. RAVIART AND J. M. THOMAS, *A mixed finite element method for 2nd order elliptic problems*, in Mathematical Aspects of Finite Element Methods (Rome, 1975), Lecture Notes in Math. 606, Springer, Berlin, 1977, pp. 292–315.
- [26] S. A. SAUTER AND C. SCHWAB, *Randelementmethoden. Analyse, Numerik und Implementierung schneller Algorithmen*, B. G. Teubner, Stuttgart, Leipzig, Wiesbaden, 2004.
- [27] O. STEINBACH, *Numerical Approximation Methods for Elliptic Boundary Value Problems. Finite and Boundary Elements*, Springer, New York, 2008.
- [28] O. STEINBACH AND W. L. WENDLAND, *On C. Neumann's method for second order elliptic systems in domains with non-smooth boundaries*, J. Math. Anal. Appl., 262 (2001), pp. 733–748.
- [29] M. WINDISCH, *Modifizierte Randintegralgleichungen für elektromagnetische Streuprobleme*, Diplomarbeit, Institut für Numerische Mathematik, TU Graz, Graz, Austria, 2007.

A FAST METHOD FOR LINEAR WAVES BASED ON GEOMETRICAL OPTICS*

CHRISTIAAN C. STOLK[†]

Abstract. We develop a fast method for solving the one-dimensional wave equation based on geometrical optics. From geometrical optics (e.g., Fourier integral operator theory or WKB approximation) it is known that high-frequency waves split into forward and backward propagating parts, each propagating with the wave speed, with amplitude that is slowly changing depending on the medium coefficients, under the assumption that the medium coefficients vary slowly compared to the wavelength. Based on this we construct a method of optimal, $O(N)$ complexity, with basically the following steps: 1. decouple the wavefield into an approximately forward and an approximately backward propagating part; 2. propagate each component explicitly along the characteristics over a time step that is small compared to the medium scale but can be large compared to the wavelength; 3. apply a correction to account for the errors in the explicit propagation; repeat steps 2 and 3 over the necessary amount of time steps; and 4. reconstruct the full field by adding forward and backward propagating components again. Due to step 3 the method accurately computes the full wavefield. A variant of the method was implemented and outperformed a standard order (4,4) finite difference method by a substantial factor. The general principle is applicable also in higher dimensions, but requires efficient implementations of Fourier integral operators which are still the subject of current research.

Key words. wave equation, numerical method, multiscale method, geometrical optics, integrating factor

AMS subject classifications. 65M25, 76Q05

DOI. 10.1137/070698919

1. Introduction. Consider waves propagating in an inhomogeneous medium which varies slowly on the scale of the wavelength. In this case, waves behave much like in the constant coefficient case. For example, in one dimension an initial pulse approximately splits into a forward propagating pulse and a backward propagating pulse, each propagating with the wave speed, and with slowly varying amplitude. Indeed for small times, the wave “sees” only a small, approximately constant part of the medium. This can be made precise using WKB, or geometrical optics theory, or the more general and advanced theory of Fourier integral operators. One finds that the above picture is true in the limit for high-frequency waves; these have the just described relatively simple interaction with the medium. For the low-frequency part the interaction with the medium is of course more complicated; e.g., reflections occur.

Simulating high-frequency waves using finite differences or finite elements is notoriously expensive, especially in three dimensions. One reason for this is the large number of time steps that is generally needed, since in conventional methods the time step is bounded by the space discretization length. In one dimension this leads to cost at least $O(N^2)$ if N is the number of space discretization points. This on the one hand is quite understandable: The wavefield is computed over a finite part of the (x, t) -plane with resolution $1/N$ in both the x and the t direction. On the other hand,

*Received by the editors August 1, 2007; accepted for publication (in revised form) November 17, 2008; published electronically February 19, 2009. This research was supported by the Netherlands Organisation for Scientific Research through VIDI grant 639.032.509. This work was done while the author was employed at the University of Twente.

<http://www.siam.org/journals/sinum/47-2/69891.html>

[†]Korteweg-de Vries Institute for Mathematics, University of Amsterdam, 1018 TV Amsterdam, The Netherlands (C.C.Stolk@uva.nl).

if we are interested only in the map from initial to final values, one can argue that there is room for improvement: The high frequencies are well described by translation and scaling over quantities that follow from the smoothly varying medium. The low frequencies still need to be computed by some discretization, but with a coarse grid. In this paper we will show that in fact we can devise a scheme that follows this pattern and is of complexity $O(N)$, i.e., optimal.

The observation about the high cost of simulating high-frequency waves is not new, and various authors have sought to deal with this, e.g., [12] in one dimension, [2, 9] in higher dimensions. The paper [12] uses the observation that the matrix that describes the propagator $P(t)$ (the operator exponent e^{tM} in the notation below, that maps initial values at time 0 to values at a later time t , assuming time-independent coefficients) can be compressed by wavelet compression. High-frequency signals in the propagator are concentrated around the characteristics. Low-frequency signals are not. Due to the separation in space and scale that is obtained using wavelets, this leads to many small entries that, if omitted, give only a small error to the matrix. The matrix is compressed in this way, and it becomes possible to store it. The operator exponent is then first computed for a small time τ , and subsequently for longer times by repeated squaring $P(2\tau) = P(\tau)^2$, $P(4\tau) = P(2\tau)^2$, etc. Unlike our method this idea is restricted to time-independent coefficients. Curvelet frames [15, 4] have been proposed to extend this idea to multiple dimensions.

In this paper we introduce a new, different concept to reduce computational cost. We explicitly separate forward and backward propagating parts of the waves, as made possible by high-frequency asymptotic theory, and propagate these explicitly. No matrix compression is used. Roughly speaking the method involves the following steps, that are repeated over a number of time steps to obtain the final result:

1. Decouple the wavefield into a forward and a backward propagating part, like for the constant coefficient medium where we can find two functions F and B such that the solution is given by $U_1(x, t) = B(x + ct) + F(x - ct)$.
2. Propagate each component explicitly over a time step that is small compared to the medium scale but large compared to the wavelength.
3. Apply a correction to account for the errors in the explicit propagation.
4. Reconstruct the full field by adding forward and backward propagating components again.

For higher dimensions one could perhaps devise a similar scheme; however, at this point in time it is not clear how to efficiently compute the Fourier integral operators needed in step 2.

Two methods according to this outline will be described. First we derive a relatively straightforward method, that is implemented numerically and tested. The goal of this is to get a first impression of what kind of numerical results can be obtained. Compared with an order (4,4) finite difference method we find improvements in speed of factors up to 20, depending on the smoothness of the medium.

A second method is derived using several more innovations, in particular a new multiscale time-stepping method; see section 6 and thereafter. For this method we study error estimates and the complexity, and we show that it has *optimal* $O(N)$ complexity. The $O(N)$ complexity is better than that in [12], but we also have another improvement compared to the repeated squaring method, namely that our method is also applicable in media with time-dependent coefficients.

Let us discuss in more mathematical terms the ideas behind the method. We consider the one-dimensional acoustic wave equation

$$(1.1) \quad (\partial_t \circ a(x, t) \partial_t - \partial_x \circ b(x, t) \partial_x) U_1(x, t) = 0,$$

with domain given by a circle Ω of integer length L . It will be convenient to write this as a first-order system; let

$$(1.2) \quad U_2 = a \partial_t U_1, \quad U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}, \quad M = \begin{pmatrix} 0 & a^{-1} \\ \partial_x \circ b \partial_x & 0 \end{pmatrix}.$$

Then (1.1) becomes

$$(1.3) \quad \frac{d}{dt} U = MU.$$

We view this as an ODE with values in a function space, which explains the notation $\frac{d}{dt}$ in this equation.

We are interested in the initial value problem where $U(t_0) = U_0$ is given and $U(t_1)$ is to be determined. The natural space to consider the equation is $U(t) \in H^1 \times L^2$, where $H^s = H^s(\Omega)$ denotes the Sobolev space of order s . With coefficients that are $C^{k,1}$ in space, and with time derivative that is also $C^{k,1}$ in space, there is existence, uniqueness, and stable dependence on initial values for $U_0 \in H^{s+1} \times H^s$, with

$$U(t) \in C([t_0, t_1], H^{s+1} \times H^s),$$

for $-k-1 \leq s \leq k$ [14, 16].

Let us consider now where there is room for improvement in standard finite difference or finite element methods. Suppose U_1, U_2 are discretized on Ω by finite differences, using a regular grid with grid distance h and $N = L/h$ grid points. Then the operator M is discretized, and the time evolution is computed with some time-stepping procedure. The operator M behaves like a first-order operator, mapping $H^{s+1} \times H^s$ to $H^s \times H^{s-1}$. Its norm is proportional to h^{-1} . Accuracy and stability of a discrete approximation now require that the time step is of order h , $\Delta t \lesssim h/c(x, t)$, with $c = \sqrt{b/a}$ the velocity (the Courant–Friedrichs–Lewy condition). The cost for given N is therefore at least $O((\# \text{ of time steps}) \cdot N) = O(N^2)$.

To have lower cost, we will attack the number of time steps, by using larger time steps. An idea that has been used for this purpose is operator splitting with an *integrating factor method*. Suppose M is of the form

$$(1.4) \quad M = A + B.$$

Operator splitting is the idea that the matrix exponential $e^{\Delta t(A+B)}$ is approximated by products of factors $e^{\Delta t_j A}$ and $e^{\Delta t_k B}$. One way to derive an operator splitting method is the integrating factor method. Let $E(t, t_0)$ be a solution operator for $U' = AU$, i.e., an operator that maps $U(t_0)$ to the solution $U(t)$ of $U' = AU$. For the time-independent case $E(t, t_0) = e^{(t-t_0)A}$. Then we can define

$$(1.5) \quad V = E(t, t_0)^{-1} U.$$

The term $E(t, t_0)^{-1}$ is then an integrating factor. Differentiating the equivalent equation $E(t, t_0)V = U$ gives that

$$(A + B)U = \frac{dU}{dt} = AE(t, t_0)V + E(t, t_0) \frac{dV}{dt}.$$

Therefore, solving for $\frac{dV}{dt}$,

$$(1.6) \quad \frac{dV}{dt} = E(t, t_0)^{-1} B E(t, t_0) V.$$

To apply this usefully, the operator on the right-hand side must have smaller norm than the original operator M , so that time-stepping can be performed with larger time steps. This is applied in some nonlinear equations with a diffusive part, for which the time evolution can be computed efficiently in the Fourier domain [20]. Because of this use of an integrating factor, we call our method a geometrical optics integrating factor method.

A similar idea is used in the Egorov theorem of microlocal analysis. In this theory, a Fourier integral operator (FIO) $E(t, t_0)$ is constructed [11, 10, 19, 21], such that the field $V(t) = E(t, t_0)^{-1} U(t)$ satisfies

$$(1.7) \quad \frac{\partial}{\partial t} V(t) = R(t, t_0) V(t),$$

where the operator R is smoothing, in the sense that it maps $H^{s+1} \times H^s \rightarrow H^{s+1+K} \times H^{s+K}$ for any K desired (the order K depends on the amount of terms in the asymptotic series for the amplitude in the FIO $E(t, t_0)$). The fact that R is bounded means that a properly discretized version can be bounded independent of h . By the above reasoning the stepsize requirement would become independent of h (of course an estimate of the time discretization error is needed to establish this). For small h , as the number of time steps would become large due to the CFL condition, one might expect to have a gain in computation speed for the transformed differential equation (1.7).

Continuing this line of reasoning, the time step could become independent of the number of space discretization points N , assuming the desired accuracy stays fixed. For example, having initial conditions double in frequency, with the same medium and accuracy, one can conjecture that the time step could stay the same.

While Fourier integral operator theory has been developed for any space dimension, for dimension 2 or higher it is not clear how to efficiently obtain numerical approximations of Fourier integral operators (see for work in this direction, e.g., the recent paper [3]). Here we therefore treat the one-dimensional case.

In this case, it is convenient not to work with the field U , or with V in (1.7) directly, but instead work with forward and backward propagating components. These will be denoted by u_1 and u_2 . An operator Q and its inverse will be constructed such that $u = (u_1, u_2)^T = Q^{-1} U$ (this gives step 1 and 3). We will show that in terms of these variables the differential equation (1.3) becomes

$$(1.8) \quad \frac{d}{dt} u = (T + R) u$$

with

$$T = \begin{pmatrix} \sqrt{b/a} \partial_x + f_1 & 0 \\ 0 & -\sqrt{b/a} \partial_x + f_2 \end{pmatrix},$$

f_1, f_2 functions given below, and R a remainder operator, that is explicitly derived and is continuous $H^{s+1} \times H^{s+1} \rightarrow H^{s+2} \times H^{s+2}$ (for time-independent coefficients $f_1 = f_2 = 0$). Versions of R with off-diagonal terms that are even more smoothing can also be constructed; see further on in the paper.

Equation (1.8) will be used for operator splitting. The equation $u' = T u$ corresponds to two transport equations (step 2 in the outline above). These are solved

using the method of characteristics. This yields a geometrical optics approximation of the propagator. The term R then yields the correction mentioned in step 3 of the four points above.

Computing with the characteristics is cheaper than computing directly on the wavefield, e.g., using a discretization of the transport equation. The explanation for this is that the time steps in an ODE solver needed for solving for the characteristics depend on the medium smoothness, and not on the smoothness of the wavefield, and can therefore be longer than the time steps in a discretization of the transport equation. Similarly, it is not necessary to compute a characteristic for each grid point because interpolation can be used. After computing the characteristics, applying the flow along the characteristics becomes a standard interpolation problem.

The computation of flow along characteristics is related to the use of moving grids in scalar conservation laws. Originally the reason to have the grid moving with the singularities of solutions was that an adapted (locally refined) grid would stay adapted to the singularities. But it was also observed that this could lead to larger time steps [13].

As mentioned we have both numerical and theoretical results. First we derive a relatively simple method following the above ideas. This method has been implemented and compared with a standard order (4,4) finite-difference method described in [6]. Factors of order 10 to 20 of improvement in the computation speed were obtained in examples.

In the second part of the paper we study error estimates and complexity. It turns out that the method described in sections 2 to 4 does not yet have the best possible complexity. With several enhancements we construct a method (or a class of methods) with optimal complexity $O(N)$ to solve the initial value problem. These additional features are the use of higher-order decoupling, and of a multiscale decomposition where each scale has its own time step (multiscale time-stepping). They will be further introduced in section 6.

The remainder of the paper will be organized as follows. In section 2 we describe the separation of the forward and backward propagating parts of the wavefield (decoupling). The differential equation is then transformed into one to which operator splitting and the integrating factor method can be applied. This is discussed in section 3. We then describe a simple space discretization and the resulting algorithm in section 4. Section 5 contains some numerical results. Section 6 introduces the main additional ideas behind the method for which we establish $O(N)$ complexity. These are further worked out and proved in sections 7, 8, and 9. We end with a short discussion of the results.

2. Decoupling the equation. The splitting in (1.4)–(1.6) is not directly applied to M ; first the equation is transformed to new variables as announced in (1.8). We define new variables by

$$U(t) = Q(t)u(t),$$

with Q an invertible matrix operator. The operator Q is independent of t if M is independent of t , and may otherwise depend on t . The equation for u is then (' denoting time differentiation)

$$(2.1) \quad u' = (Q^{-1}MQ - Q^{-1}Q')u.$$

The purpose of this section is to find a suitable operator Q , such that the resulting differential equation is of the form

$$(2.2) \quad \frac{d}{dt} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} \sqrt{b/a}\partial_x + f_1 & 0 \\ 0 & -\sqrt{b/a}\partial_x + f_2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} + R \begin{pmatrix} u_1 \\ u_2 \end{pmatrix},$$

with Q , f_1 , and f_2 and the remainder operator R to be determined. In fact, we will derive an explicit expression for

$$(2.3) \quad QR = \underbrace{MQ}_A - \underbrace{Q \begin{pmatrix} \sqrt{b/a}\partial_x + f_1 & 0 \\ 0 & -\sqrt{b/a}\partial_x + f_2 \end{pmatrix}}_B - \underbrace{Q'}_C.$$

The notations A, B, C will be used below in evaluating the product. Note that R is not given directly, but has to be computed as the product of Q^{-1} and QR which are given; the reason for this is that we want to minimize the use of inverse differential operators, and here the only place where those occur is in Q^{-1} . We will find that the operator R belongs to a class of pseudodifferential operators of order -1 .

In the remainder of the section the actual computation is done. We treat separately the cases where a, b are time-independent, resp., the general case with time-dependent a, b . For convenience we collect the results in the following lemma.

LEMMA 2.1. *For the time-independent case, with Q given by (2.5), and $f_1 = f_2 = 0$, QR is given by (2.6) and (2.7). For the time-dependent case, with Q, f_1, f_2 given by (2.8), (2.9), and (2.12), QR is given by (2.10), (2.11), (2.13), and (2.14).*

Computation for the time-independent case. In this case we will take Q independent of t , so that $C = 0$, and such that f_1 and f_2 vanish. Consider first the following choice for Q :

$$Q^{(0)} = \begin{pmatrix} 1 & 1 \\ \sqrt{ab}\partial_x & -\sqrt{ab}\partial_x \end{pmatrix}.$$

A quick computation shows that

$$(2.4) \quad Q^{(0)}R^{(0)} = MQ^{(0)} - Q^{(0)} \begin{pmatrix} \sqrt{b/a}\partial_x & 0 \\ 0 & -\sqrt{b/a}\partial_x \end{pmatrix} = \begin{pmatrix} \text{order}(0) & \text{order}(0) \\ \text{order}(1) & \text{order}(1) \end{pmatrix},$$

so to highest order this is a good choice.

Next we modify Q so that (1) it is invertible, and (2) the components of QR vanish to one order lower. The operator Q becomes invertible when the derivative is replaced by a regularized derivative, which will be denoted by $\tilde{\partial}_x$, defined in the Fourier domain by multiplication with $ik + \frac{\alpha}{\beta k^2 + 1}$, with α, β suitable positive, real constants that remain to be chosen. To eliminate the order 0 and order 1 terms in (2.4), the columns of Q will be normalized by a weight function; we will try

$$(2.5) \quad Q = \begin{pmatrix} f(x) & f(x) \\ f(x)\sqrt{ab}\tilde{\partial}_x & -f(x)\sqrt{ab}\tilde{\partial}_x \end{pmatrix}, \quad Q^{-1} = \frac{1}{2} \begin{pmatrix} f^{-1} & \tilde{\partial}_x^{-1}f^{-1}\frac{1}{\sqrt{ab}} \\ f^{-1} & -\tilde{\partial}_x^{-1}f^{-1}\frac{1}{\sqrt{ab}} \end{pmatrix},$$

with f given by $f = a^{-1/4}b^{-1/4}$.

For contribution A we then find

$$A_{11} = -A_{12} = f(x)\sqrt{\frac{b}{a}}\partial_x + f(x)\sqrt{\frac{b}{a}}(\tilde{\partial}_x - \partial_x) = a^{-3/4}b^{1/4}\partial_x + a^{-3/4}b^{1/4}(\tilde{\partial}_x - \partial_x)$$

and

$$\begin{aligned} A_{21} &= A_{22} = \partial_x b \partial_x f \\ &= a^{1/4}b^{1/4}\partial_x a^{-1/2}b^{1/2}\partial_x + R_1 \end{aligned}$$

with

$$\begin{aligned} R_1 &= a^{-1/4}b^{3/4} \left[\left(\frac{1}{4}\partial_x \log a - \frac{3}{4}\partial_x \log b \right) \left(\frac{1}{4}\partial_x \log a + \frac{1}{4}\partial_x \log b \right) \right. \\ &\quad \left. - \left(\frac{1}{4}\partial_x^2 \log a + \frac{1}{4}\partial_x^2 \log b \right) \right]. \end{aligned}$$

Contribution B is given by

$$B_{11} = -B_{12} = a^{-3/4}b^{1/4}\partial_x$$

and

$$B_{21} = B_{22} = a^{1/4}b^{1/4}\partial_x a^{-1/2}b^{1/2}\partial_x + a^{1/4}b^{1/4}(\tilde{\partial}_x - \partial_x)a^{-1/2}b^{1/2}\partial_x.$$

We thus find the following for QR :

$$(2.6) \quad (QR)_{11} = -(QR)_{12} = a^{-3/4}b^{1/4}(\tilde{\partial}_x - \partial_x)$$

and

$$(2.7) \quad (QR)_{21} = (QR)_{22} = R_1 - a^{1/4}b^{1/4}(\tilde{\partial}_x - \partial_x)a^{-1/2}b^{1/2}\partial_x.$$

The time-dependent case. In this case we try

$$(2.8) \quad Q = \begin{pmatrix} f(x) & f(x) \\ f(x)\sqrt{ab}\tilde{\partial}_x + c_1 & -f(x)\sqrt{ab}\tilde{\partial}_x + c_2 \end{pmatrix},$$

with f as above, and f_1, f_2, c_1, c_2 to be determined. The inverse of Q will be discussed below. We find

$$\begin{aligned} A_{11} &= a^{-3/4}b^{1/4}\partial_x + a^{-3/4}b^{1/4}(\tilde{\partial}_x - \partial_x) + a^{-1}c_1, \\ A_{12} &= -a^{-3/4}b^{1/4}\partial_x - a^{-3/4}b^{1/4}(\tilde{\partial}_x - \partial_x) + a^{-1}c_2; \end{aligned}$$

A_{21} and A_{22} remain unchanged. For the coefficients of the matrix operator B we find

$$\begin{aligned} B_{11} &= a^{-3/4}b^{1/4}\partial_x + (ab)^{-1/4}f_1, \\ B_{12} &= -a^{-3/4}b^{1/4}\partial_x + (ab)^{-1/4}f_2, \\ B_{21} &= a^{1/4}b^{1/4}\partial_x a^{-1/2}b^{1/2}\partial_x + a^{1/4}b^{1/4}(\tilde{\partial}_x - \partial_x)a^{-1/2}b^{1/2}\partial_x \\ &\quad + c_1\sqrt{b/a}\partial_x + (ab)^{1/4}\tilde{\partial}_x \circ f_1 + c_1f_1, \\ B_{22} &= a^{1/4}b^{1/4}\partial_x a^{-1/2}b^{1/2}\partial_x + a^{1/4}b^{1/4}(\tilde{\partial}_x - \partial_x)a^{-1/2}b^{1/2}\partial_x \\ &\quad - c_2\sqrt{b/a}\partial_x - (ab)^{1/4}\tilde{\partial}_x \circ f_2 + c_2f_2. \end{aligned}$$

For C we have

$$\begin{aligned} C_{11} &= \partial_t(ab)^{-1/4}, \\ C_{12} &= \partial_t(ab)^{-1/4}, \\ C_{21} &= \partial_t(ab)^{1/4}\tilde{\partial}_x + \partial_t c_1, \\ C_{22} &= -\partial_t(ab)^{1/4}\tilde{\partial}_x + \partial_t c_2. \end{aligned}$$

Adding all the contributions we find that

$$(QR)_{11} = +a^{-3/4}b^{1/4}(\tilde{\partial}_x - \partial_x) + a^{-1}c_1 - (ab)^{-1/4}f_1 - \partial_t(ab)^{-1/4}$$

and

$$\begin{aligned} (QR)_{21} &= R_1 - a^{1/4}b^{1/4}(\tilde{\partial}_x - \partial_x)a^{-1/2}b^{1/2}\partial_x - c_1\sqrt{b/a}\partial_x - (ab)^{1/4}\tilde{\partial}_x f_1 \\ &\quad - c_1 f_1 - \partial_t(ab)^{1/4}\tilde{\partial}_x - \partial_t c_1. \end{aligned}$$

The lower-order terms vanish if

$$\begin{aligned} (2.9) \quad c_1 &= -a^{3/4}b^{-1/4}((ab)^{-1/4}\partial_t(ab)^{1/4}), \\ f_1 &= 0. \end{aligned}$$

What results is

$$(2.10) \quad (QR)_{11} = a^{-3/4}b^{1/4}(\tilde{\partial}_x - \partial_x)$$

and

$$\begin{aligned} (2.11) \quad (QR)_{21} &= R_1 - a^{1/4}b^{1/4}(\tilde{\partial}_x - \partial_x)a^{-1/2}b^{1/2}\partial_x - \partial_t(ab)^{1/4}(\tilde{\partial}_x - \partial_x) \\ &\quad + \partial_t(\sqrt{a/b}\partial_t(ab)^{1/4}). \end{aligned}$$

Similarly we have for the 12 and 22 components

$$\begin{aligned} (QR)_{12} &= -a^{-3/4}b^{1/4}(\tilde{\partial}_x - \partial_x) + a^{-1}c_2 - (ab)^{-1/4}f_2 - \partial_t(ab)^{-1/4}, \\ (QR)_{22} &= R_1 - a^{1/4}b^{1/4}(\tilde{\partial}_x - \partial_x)a^{-1/2}b^{1/2}\partial_x + c_2\sqrt{b/a}\partial_x + (ab)^{1/4}\tilde{\partial}_x f_2 \\ &\quad - c_2 f_2 + \partial_t(ab)^{1/4}\tilde{\partial}_x - \partial_t c_2, \end{aligned}$$

with lower-order terms vanishing if

$$\begin{aligned} (2.12) \quad c_2 &= -a^{3/4}b^{-1/4}((ab)^{-1/4}\partial_t(ab)^{1/4}), \\ f_2 &= 0. \end{aligned}$$

The result for $(QR)_{12}$ and $(QR)_{22}$ are

$$\begin{aligned} (2.13) \quad (QR)_{12} &= -a^{-3/4}b^{1/4}(\tilde{\partial}_x - \partial_x), \\ (2.14) \quad (QR)_{22} &= R_1 - a^{1/4}b^{1/4}(\tilde{\partial}_x - \partial_x)a^{-1/2}b^{1/2}\partial_x + \partial_t(ab)^{1/4}(\tilde{\partial}_x - \partial_x) \\ &\quad + \partial_t(\sqrt{a/b}\partial_t(ab)^{1/4}). \end{aligned}$$

This completes the time-dependent case, except for the inverse of Q .

For the inversion, rewrite Q as

$$Q = \begin{pmatrix} f(x) & f(x) \\ f(x)\sqrt{ab}(\tilde{\partial}_x + \bar{c}_1) & -f(x)\sqrt{ab}(\tilde{\partial}_x - \bar{c}_2) \end{pmatrix},$$

with $\bar{c}_j = \frac{c_j}{f\sqrt{ab}}$. It turns out that Q can be inverted, according to the following explicit formula:

$$(2.15) \quad Q^{-1} = \frac{1}{2} \begin{pmatrix} \tilde{\partial}^{-1}(\tilde{\partial}_x - \bar{c}_2)f^{-1} & \tilde{\partial}^{-1}\frac{1}{\sqrt{ab}}f^{-1} \\ \tilde{\partial}^{-1}(\tilde{\partial}_x + \bar{c}_1)f^{-1} & -\tilde{\partial}^{-1}\frac{1}{\sqrt{ab}}f^{-1} \end{pmatrix}.$$

This is basically due to the fact that $\bar{c}_1 = \bar{c}_2$.

3. Operator splitting and time-stepping. The equation for the decoupled wavefields u is now

$$(3.1) \quad \frac{d}{dt}u = (T + R)u$$

with R as derived in the previous section and T given by

$$T = \begin{pmatrix} \sqrt{b/a}\partial_x & 0 \\ 0 & -\sqrt{b/a}\partial_x \end{pmatrix}.$$

The integrating factor will be $E(t, t_0)^{-1}$, where $E(t, t_0)$ solves $\frac{d}{dt}E(t, t_0) = TE(t, t_0)$, $E(t_0, t_0) = \text{Id}$, and we will define a field v by

$$v(t, t_0) = E(t, t_0)^{-1}u(t),$$

which satisfies the differential equation

$$(3.2) \quad \frac{dv}{dt} = E(t, t_0)^{-1}RE(t, t_0)v.$$

Applying Euler forward time-stepping for this equation gives

$$v(t + \Delta t, t) \approx (1 + \Delta t E(t + \Delta t, t)^{-1}RE(t + \Delta t, t))v(t),$$

using that $v(t, t) = u(t)$. Hence

$$u(t + \Delta t) \approx (1 + \Delta t R)E(t + \Delta t, t)u(t).$$

A symmetric form of splitting (cf. Strang splitting [17]) leads to the following time-stepping, expressed in time-stepping for u :

$$(3.3) \quad u(t + \Delta t) \approx (1 + \frac{1}{2}\Delta t R)E(t + \Delta t, t)(1 + \frac{1}{2}\Delta t R)u(t).$$

Let us now explain in more detail the computation of $E(t, t_0)$. This is a diagonal 2×2 matrix operator. We take the forward propagating component (the $E_{2,2}$ component, which acts on the u_2 field); the backward propagating component is done similarly. The characteristic equation is

$$(3.4) \quad \frac{dx}{dt} = c(x, t).$$

For the time-independent case, we can solve this ODE for $x(t)$ with initial value $x(t_0) = x_0$ by separating the variables, which yields the equation $\int_{x_0}^x c(\xi)^{-1} d\xi = t - t_0$, so the computation can be done from a primitive $\int c(x) dx$. For the time-dependent case (3.4) is solved directly. Let $X(x_0, t, t_0)$ denote the solution $x(t)$ with initial values $x(t_0) = x_0$. Then we have

$$(3.5) \quad E_{2,2}(t, t_0)u_2 = u_2(t_0, X(x, t_0, t))$$

(the characteristic is computed backward). If $\Phi_2(t, t_0)$ denotes the characteristic flow mapping x_0 to $X(x_0, t, t_0)$, this equals the pull back $E_{2,2}(t, t_0)u_2(t_0) = \Phi_2(t_0, t)^*u_2(t_0)$.

4. Numerical implementation. For a numerical implementation, it remains to perform the space discretization. We chose to work with finite differences, which are easy to implement. The following operators were discretized:

1. ∂_x . This operator was discretized using central differences.
2. $\tilde{\partial}_x, \tilde{\partial}_x^{-1}, \tilde{\partial}_x - \partial_x$. These are applied in the Fourier domain, with a regularized version of central differences. There computation involves an FFT and an inverse FFT, which, due to the $O(N \log N)$ cost of this operation, will form the bulk of the computations.
3. Multiplications with coefficients and derivatives of coefficients. Derivatives of coefficients are computed again using central differences.
4. The *translation operator* $E(t, t_0)$ is computed for the time-independent case using the primitive $\int c(x)^{-1} dx$, mentioned above, and using a Runge–Kutta ODE solver otherwise. Then third-order Lagrange interpolation is applied. For the time-independent case a sparse matrix is precomputed, that performs the translation over a given time step Δt .

In this way a simple numerical implementation of the method given by (3.3) was made.

5. Numerical results. In the numerical results we concentrate on the method for the time-independent case. For this case comparisons of computation time were made. For the time-dependent case it was observed that solutions are well approximated. But we feel the results for the time-independent case give sufficient indication of the effectiveness of the method.

For this method, with the assumption of medium smoothness it is of course an important question *just how smooth the medium coefficients need to be* in order that the method demonstrates an improvement compared to more conventional methods. Therefore numerical results were computed for media with increasing smoothness. The media were chosen parameterized by B-splines of order 3; the coefficients a of the media were randomly chosen, uniformly distributed between 0.4 and 1.6. The increasing smoothness was obtained by increasing the node distance, for which we took the values 1, 2, 4, and 8. The b coefficient was chosen equal to 1. The initial value for U_1 was a pulse of approximately unit width; the initial value for U_2 was chosen equal to zero. In Figure 5.1 one such medium is displayed. In Figure 5.2 the initial value for U_1 is plotted. The propagation was over approximately 100 wavelengths.

The results were compared with the result of an order (4,4) finite difference method; see [6]. Both methods were implemented in MATLAB. For our method the main cost was in the Fourier transform used for computing $\tilde{\partial}_x$ and its inverse. In the standard finite difference methods, for each time step a sparse matrix was applied, and this constituted almost 100% of the cost.

The first check was that the method actually approximates the solutions well. This was indeed the case. In Table 5.1 some numerical results are given, where

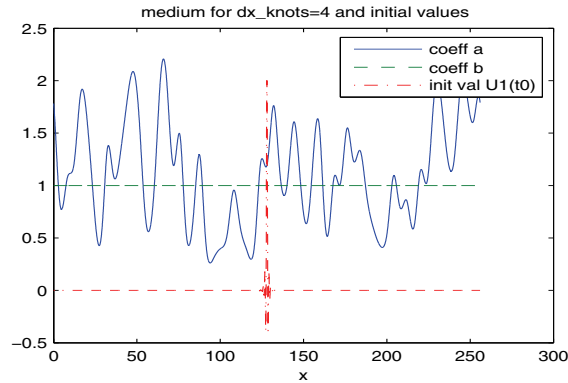


FIG. 5.1. Medium coefficients with random B-splines with knot distance 4.

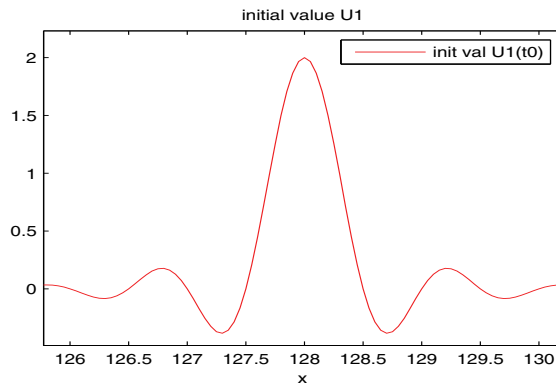


FIG. 5.2. Initial value for U_1 used in the numerical tests.

TABLE 5.1

Comparison of the cost of the method of section 4 with an order (4,4) finite difference method. h_{FD} is the space stepsize taken in the finite difference scheme for which the comparison is made.

Medium scale	h_{FD}	$\frac{\text{Cost FD}}{\text{Cost GOIF}}$
1	0.05	0.37
	0.025	0.30
2	0.05	3.3
	0.025	5.4
4	0.05	9.0
	0.025	15.7
8	0.05	17.4
	0.025	25.3

computation time is compared. For the new method we required the error to be smaller in both the supremum and the L^2 sense, or at most 10% larger in one of the two, but better when both are taken into account. As can be seen, knot distance 1 is not sufficient to obtain any gain, but from knot distance 2 considerable gain is obtained, up to a factor of about 20 for very smooth media.

As this is only a first implementation we feel this is strong encouragement to further analyze geometrical optics based methods.

6. An optimal complexity method: Overview. For the method introduced above there were no rigorous error estimates given. The complexity is, however, at least $O(N \log N)$ since the regularized derivative $\tilde{\partial}_x$ and its inverse were computed in the Fourier domain and needed to be computed for each time step. In this section we present a more elaborate algorithm, for which we establish that the complexity is $O(N)$, where N denotes the number of grid points in the space discretization.

So the task in the remaining sections is on the one hand to control the error in a numerical method and on the other hand control the cost. The discretization will be done for the differential equation

$$(6.1) \quad \frac{dv}{dt} = E(t, t_0)^{-1} R E(t, t_0) v,$$

that resulted from (3.1) after applying the integrating factor. It follows from the results in section 8 below that the transformation from the original equation (1.3) to this form and back can be done at cost $O(N)$ and with error satisfying bounds that are sufficient.

We will provide precise error estimates of classical type; i.e., we assume the input has a certain amount of additional regularity, we consider the discretization error in the result given that the input has to be approximated in an N -dimensional space of (spline-) functions, and we then show that the total error in the output is of the same order in N as the discretization error. Evolution according to (6.1) maps initial values $v(t_0) = v_0$ in $H^1 \times H^1$ to final values $v(t_1)$ that are also in $H^1 \times H^1$. We will assume that v_0 is in $H^{1+\alpha} \times H^{1+\alpha}$, i.e., has α additional orders of regularity. The discretization error that results from putting v_0 in an N -dimensional spline space can then be estimated by $CN^{-\alpha}$. We will show that, for a method with cost that can be bounded by CN , the final result satisfies an estimate of the type

$$\|v_{\text{approx}}(t_1) - v(t_1)\|_{H^1 \times H^1} \leq CN^{-\alpha}$$

(the letter C may mean a different constant in different equations).

A naive approach would be to simply take the differential equation (6.1), first apply a discretization in space, and then subsequently apply discretization in time. The time discretization should preferably be of higher order. There are two main problems with this approach, which will lead to additional special features of our method. These new features are the following:

1. *Higher order decoupling.* Control of the time discretization error in higher-order time-stepping, say of order K , requires bounds on the time derivatives of the operator $E(t, t_0)^{-1} R E(t, t_0)$ occurring on the right-hand side of (6.1). The first time derivative contains a commutator $[R, T]$ (which is of order 0 and hence bounded), but higher time derivatives contain higher-order commutators, that are of positive order, and hence do not satisfy the required bounds. To address this issue we will introduce higher-order decoupling. In section 7 we will construct a new operator R , with off-diagonal terms that are smoothing operators of order K , and show that its time derivatives of order $0, \dots, K$ are bounded on a sufficiently large range of Sobolev spaces. The higher-order decoupling is obtained by adapting an argument of Taylor [19, Chapter 9] or [18].
2. *Multiscale time-stepping.* The second problem that needs to be addressed is that in our complexity estimates, with increasing N , the error must decrease. This in turn means that the time step must decrease, which would lead to

superlinear complexity. To address this issue we introduce *multiscale time-stepping*. The idea is that the coarse scales are propagated with a small time step. The coarse scales are parameterized with relatively few coefficients but contain most of the energy. It is therefore affordable to use a smaller time step, and at the same time this leads to a big improvement in the error. For the fine scales, that contain relatively little energy, larger time steps are used. Incidentally this is very much in agreement with the philosophy of asymptotic methods, where the high frequencies are well approximated. Each time step amounts to a correction to the purely asymptotic approximation, so few are needed for the high frequencies. The idea of multiscale time-stepping is new to our knowledge.

Because of the multiscale time-stepping, we assume the use of a wavelet based multiscale discretization in space. We will use [5] as our main reference for wavelet discretization; see also [7].

In the next three sections we will work out the above issues in detail and prove the $O(N)$ complexity result. Section 7 concerns the higher-order decoupling. Discretization and operator approximation will be discussed in section 8. Section 9 will contain the ideas on multiscale time-stepping and the final parts of the proof that combine all the intermediate results.

7. Higher-order decoupling. By the transformation $u = Q^{-1}U$ in section 2, the original system (1.3) was transformed to

$$u' = (T + R)u,$$

where $T + R = Q^{-1}MQ - Q^{-1}Q'$. We had

$$T = \begin{pmatrix} \sqrt{b/a}\partial_x & 0 \\ 0 & -\sqrt{b/a}\partial_x \end{pmatrix}.$$

The operator R is a matrix pseudodifferential operator, with components that are of order

$$(7.1) \quad R = \begin{pmatrix} \text{order}(-1) & \text{order}(-1) \\ \text{order}(-1) & \text{order}(-1) \end{pmatrix}.$$

Here by $\text{order}(-1)$ we mean that it is bounded $H^s \rightarrow H^{s+1}$ for a suitable range of s .

In this section we explain how to construct Q such that R has the property that

$$(7.2) \quad \frac{d^j}{dt^j}(E(t, t_0)^{-1}RE(t, t_0)) \text{ is bounded on } H^1 \times H^1 \text{ for } j = 0, 1, \dots, K,$$

with K a positive integer indicating, as mentioned, the order of the time-stepping that is going to be used.

We first argue that property (7.1) is not sufficient if $K > 1$. Take for example the first time derivative of $E(t, t_0)^{-1}RE(t, t_0)$:

$$(7.3) \quad \frac{d}{dt}(E(t, t_0)RE(t, t_0)) = E(t, t_0)^{-1} \left([R, T] + \frac{dR}{dt} \right) E(t, t_0).$$

Consider the commutator $[R, T]$ occurring inside the brackets:

$$(7.4) \quad [R, T] = \begin{pmatrix} [R_{1,1}, T_{1,1}] & R_{1,2}T_{2,2} - T_{1,1}R_{1,2} \\ R_{2,1}T_{1,1} - T_{2,2}R_{2,1} & [R_{2,2}, T_{2,2}] \end{pmatrix}.$$

To get the idea assume that the coefficients a and b are C^∞ , so that R and T have smooth symbols. What we see from this expression is the following:

- The diagonal terms $[R, T]_{1,1}$ and $[R, T]_{2,2}$ are commutators of scalar pseudodifferential operators, and their order equals the order of $R_{1,1}$, resp., $R_{2,2}$.
- For the off-diagonal terms $[R, T]_{1,2}$ and $[R, T]_{2,1}$ this is not true; their order is increased by 1 compared to $R_{1,2}$, resp., $R_{2,1}$.

This has nothing to do with the specific form of R ; if R is replaced by a different matrix pseudodifferential operator, these two statements remain true. So consider the second-order time derivative of $E(t, t_0)RE(t, t_0)$. This contains the higher-order commutator $[[R, T], T]$. Assuming (7.1) and using (7.4) twice, it follows that the off-diagonal terms $[[R, T], T]_{1,2}$ and $[[R, T], T]_{2,1}$ are (a priori) of order 1, implying that (7.2) is violated.

To address this problem we will construct a modified operator Q , such that

$$(7.5) \quad R = \begin{pmatrix} \text{order}(-1) & \text{order}(-K) \\ \text{order}(-K) & \text{order}(-1) \end{pmatrix}.$$

The old operators Q and R will be referred to as $Q^{(-1)}$ and $R^{(-1)}$, because of (7.1). The new operators will be referred to as $Q^{(-K)}$ and $R^{(-K)}$. This way, we can handle K time derivatives, each of which can increase the order of the off-diagonal term by 1.

We write $\tilde{\partial}_x = \partial_x + \Psi$, where from now on we assume that Ψ is smoothing in the sense that it is continuous $H^s \rightarrow H^{s+K}$, $1 - K \leq s \leq 1$. The reason is that then any term that is a product of Ψ and other operators, none of which is of positive order, automatically is of order $(-K)$ and is hence “safe” (see (7.5)). For Ψ , we could use for example

$$\Psi = \frac{\alpha}{\beta(-\partial_x^2)^{\lceil K/2 \rceil} + 1}$$

with symbol $\frac{\alpha}{\beta k^{2\lceil K/2 \rceil} + 1}$. This is a modification with respect to the original definition of $\tilde{\partial}$ in section 2. However, it does not affect equations like (2.6), (2.7), (2.10), (2.11), (2.13), and (2.14), because the specific form of $\tilde{\partial}_x - \partial_x$ is not used in their derivation.

The main result of this section is captured in the following theorem, a short explanation of which is given after its formulation.

THEOREM 7.1. *Assume a, b are at least $C^{2K+1,1}$. There exists an operator $Q^{(-K)}$ of the form*

$$Q^{(-K)} = Q^{(-1)} \begin{pmatrix} 1 & E \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ F & 1 \end{pmatrix}$$

such that the operator $R^{(-K)}$ satisfies (7.2). The operators E, F can be chosen of the form

$$\sum_{j=2}^K c_{E^{(-j)}}(x, t) \tilde{\partial}_x^{-j}, \quad \sum_{j=2}^K c_{F^{(-j)}}(x, t) \tilde{\partial}_x^{-j},$$

where the $c_{E^{(-j)}}(x, t)$, $c_{F^{(-j)}}(x, t)$ are (x, t) dependent coefficients that depend on $a(x, t)$, $b(x, t)$ and derivatives of order up to j of a, b . The operators that form the matrix elements of $R^{(-K)}$ are sums of products of the following basic operators: operator Ψ , operators $\tilde{\partial}^{-k}$ for $k \geq 0$, and multiplication by coefficients that are functions

of a, b and derivatives of order at most $K + 1$ of a and b . This can be done such that all the terms for the off-diagonal elements of $R^{(-K)}$ are explicitly of order $-K$ in the sense that they contain a factor of Ψ or at least K powers of $\tilde{\partial}_x^{-1}$.

The description as a sum of products of basic operators is such that the operators involved can be numerically approximated with the techniques described in section 8. We note in particular that there are no cancellations between terms of $R^{(-K)}$ of order $> -K$. This is important, to avoid the situation where $R^{(-K)}$ consists of several contributions whose highest-order parts cancel analytically but not numerically due to the errors made in the numerical approximation. In the proof we will also describe a calculational scheme to compute the $c_{E^{(-j)}}(x, t), c_{F^{(-j)}}(x, t)$. (We have not calculated any case $K > 1$ explicitly.)

Proof. We write temporarily

$$T + R^{(-1)} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

We will first assume that a, b are C^∞ , so that all pseudodifferential operators involved have smooth symbols; later we will investigate how much smoothness for the coefficients is needed. Using a transformation with a matrix pseudodifferential operator of the form $\begin{pmatrix} 1 & E \\ 0 & 1 \end{pmatrix}$ the operator B will be removed to the highest $K - 1$ orders.

Replacing Q by $Q\begin{pmatrix} 1 & E \\ 0 & 1 \end{pmatrix}$ yields the following for the new operator R ; see (2.1):

$$(7.6) \quad \begin{pmatrix} 1 & E \\ 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} 1 & E \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 & E \\ 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 & E' \\ 0 & 0 \end{pmatrix} \\ = \begin{pmatrix} A - EC & B + AE - ED - E' \\ C & D + CE \end{pmatrix},$$

where we used the explicit inverse $\begin{pmatrix} 1 & E \\ 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & -E \\ 0 & 1 \end{pmatrix}$. The first problem is to find E such that $B + AE - ED - E'$ is of the desired lower order. Next we do a transformation with a matrix $\begin{pmatrix} 1 & 0 \\ F & 1 \end{pmatrix}$ of the matrix in (7.6). After this second transformation, the new operator R becomes

$$\begin{pmatrix} 1 & 0 \\ F & 1 \end{pmatrix}^{-1} \begin{pmatrix} A - EC & B + AE - ED - E' \\ C & D + CE \end{pmatrix} \begin{pmatrix} 1 & 0 \\ F & 1 \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ F & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 0 \\ F' & 0 \end{pmatrix} \\ = \begin{pmatrix} A - EC + (B + AE - ED - E')F & B + AE - ED - E' \\ C + (D + CE)F - F(A - EC) - F' & D + CE - F(B + AE - ED - E') \end{pmatrix}.$$

Just like E we must then choose F , such that $C + (D + CE)F - F(A - EC) - F'$ is of the desired lower order. The new Q is then $Q\begin{pmatrix} 1 & E \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ F & 1 \end{pmatrix}$ (using the factor $\begin{pmatrix} 1 & E \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ F & 1 \end{pmatrix}$ is convenient compared to $\begin{pmatrix} 1 & E \\ F & 1 \end{pmatrix}$ because it has an explicit inverse, easy numerically).

Let us consider the construction of E . This follows a standard pattern in pseudodifferential operator theory, choosing E order by order. We let

$$E = E^{(-2)} + E^{(-3)} + \dots + E^{(-K)},$$

and set

$$B^{(-2)} = B^{(-1)} + AE^{(-2)} - E^{(-2)}D - E^{(-2)'},$$

$$B^{(-3)} = B^{(-2)} + AE^{(-3)} - E^{(-3)}D - E^{(-3)'},$$

etc., until $B^{(-K)} = B + AE - ED - E'$.

The principal symbol of $B^{(-k)}$ is of the form $c_{B^{(-k)}(x)}(i\xi)^{-k}$, while those of A and $-D$ are both equal to $\sqrt{b/a}(i\xi)$. Hence if we choose the principal symbol of $E^{(-k-1)}$ equal to $-\frac{c_{B^{(-k)}(x)}}{2\sqrt{b/a}}(i\xi)^{-k-1}$, then the principal symbol of $B^{(-k-1)}$ vanishes, with as a result that $B^{(-k-1)}$ becomes an operator of order $-k - 1$ as desired. So we set

$$c_{E^{(-k-1)}} = -\frac{c_{B^{(-k)}}}{2\sqrt{b/a}} \quad \text{and} \quad E^{(-k-1)} = c_{E^{(-k-1)}}\tilde{\partial}_x^{-k-1}.$$

The operators $E^{(-k)}$ follow from this scheme. The coefficients $c_{B^{(-k)}}$ and $c_{E^{(-k)}}$ are determined inductively. This can be done on the symbol level using pseudodifferential operator calculus, or directly, as we will demonstrate now.

We further investigate this construction of the $c_{B^{(-k)}}$ and $c_{E^{(-k)}}$ and of the remainders $R^{(-k)}$. It is convenient to just take the matrix $R^{(-1)}$, which is the starting point of the induction, and apply a few steps of the recipe. Doing this, the key properties that allow the successful construction will become clear, without becoming overly formal.

The matrix $R^{(-1)}$ follows in the time-independent case from (2.6), (2.7), and (2.5). Omitting anything involving $\tilde{\partial}_x - \partial_x$ (which is smoothing by definition), we have the following terms relevant for the higher-order decoupling:

$$R^{(-1)} = \begin{pmatrix} \tilde{\partial}_x^{-1}a^{-1/4}b^{-1/4}R_1 & \tilde{\partial}_x^{-1}a^{-1/4}b^{-1/4}R_1 \\ -\tilde{\partial}_x^{-1}a^{-1/4}b^{-1/4}R_1 & -\tilde{\partial}_x^{-1}a^{-1/4}b^{-1/4}R_1 \end{pmatrix} + \text{order}(-K).$$

So we set, following the above scheme,

$$E^{(-2)} = -\frac{a^{-1/4}b^{-1/4}R_1}{2\sqrt{b/a}}\tilde{\partial}_x^{-2}.$$

We then find

$$\begin{aligned} B^{(-2)} &= B^{(-1)} + \left(\sqrt{b/a}\partial_x + \tilde{\partial}_x a^{-1/4}b^{-1/4}R_1\right) E - E \left(-\sqrt{b/a}\partial_x - \tilde{\partial}_x a^{-1/4}b^{-1/4}R_1\right) \\ &\quad - E' \\ &= \tilde{\partial}_x^{-1}a^{-1/4}b^{-1/4}R_1 - \sqrt{b/a}\partial_x \frac{a^{-1/4}b^{-1/4}R_1}{2\sqrt{b/a}}\tilde{\partial}_x^{-2} - \frac{a^{-1/4}b^{-1/4}R_1}{2\sqrt{b/a}}\tilde{\partial}_x^{-2}\sqrt{b/a}\partial_x \\ (7.7) \quad &+ \text{order}(-3). \end{aligned}$$

In the first term we need to commute $\tilde{\partial}_x^{-1}$ to the right, in the second term we need to commute ∂_x to the right, and in the third term we need to commute $\tilde{\partial}_x^{-2}$ to the right.

To continue an understanding of the commutator of $\tilde{\partial}_x^{-1}$ with (multiplication by) some function $g(x)$ is needed. Such a commutator yields the following:

$$\begin{aligned} [\tilde{\partial}_x^{-1}, g] &= -\tilde{\partial}_x^{-1}[\partial_x + S, g]\tilde{\partial}_x^{-1} \\ &= -\tilde{\partial}_x^{-1}(\partial_x g)\tilde{\partial}_x^{-1} - \tilde{\partial}_x^{-1}(Sg - gS)\tilde{\partial}_x^{-1}. \end{aligned}$$

The first term in this expression for the commutator is of order -2 and contains a coefficient with one more derivative. The second term is of order less than $-K$ and is hence to be disregarded.

After the commutations the highest-order terms in $B^{(-2)}$ cancel, and what remains are commutator terms and other lower-order terms.

Several more remarks are in order. First the general form, involving as basic operations the $\tilde{\partial}_x^j$, the operator $\Psi = \tilde{\partial}_x - \partial_x$, and multiplications with coefficients and derivatives and powers of coefficients, remains conserved in each step.

Concerning the order of derivatives of the coefficients that occur, in $B^{(-1)}$ and $E^{(-2)}$ we have at most second-order derivatives, in $B^{(-2)}$ and $E^{(-3)}$ at most third order, and inductively we find that in $B^{(-j)}$ and $E^{(-j-1)}$ we have at most $j + 1$ order of derivatives. One of the assumptions is that the coefficients are $C^{2K+1,1}$, which implies that in $R^{(-K)}$ the coefficients are still $C^{K,1}$.

Does this also hold for the time derivatives; i.e., do we have (7.2)? We must then carefully study (7.3) and (7.4). It turns out that each time derivative leads to a loss of at most one derivative in the regularity of the coefficients of a coefficient multiplication operator. With K time derivatives, we need $C^{0,1}$ smoothness to have a bounded map on $H^1 \times H^1$ (L^∞ would be enough if the operator was considered on $L^2 \times L^2$). Therefore $C^{K,1}$ in the coefficients occurring in $R^{(-K)}$ is sufficient and (7.2) follows.

The operator F can be determined in a similar fashion. This completes the proof of Theorem 7.1. \square

8. Discretization and operator approximation. The multiscale discretization will be done using wavelets. We follow the book of Cohen [5], which gives an excellent description of one-dimensional wavelet discretization theory; see also [7]. In a wavelet discretization functions in $L^2(\Omega)$ and $H^s(\Omega)$ are approximated by elements of increasingly large finite-dimensional subspaces of $L^2(\Omega)$ given by a multiresolution analysis V_j , $j = 0, 1, 2, \dots$. The spaces V_j are spanned by translates and scalings of the scaling function ϕ :

$$\phi_{j,k} = 2^{j/2} \phi(2^j \cdot -k), \quad k \in \mathbb{Z}/(2^j L\mathbb{Z}).$$

The V_j are assumed to form an increasing sequence $V_j \subset V_{j+1}$, $\bigcup_{j=0}^\infty V_j = L^2(\Omega)$.

In our case, where the domain is a circle of integer length L , the space V_j has $L2^j$ elements. We denote by J the final level of discretization, so that $N = L2^J$. Typically we will denote by f_j an approximation of a function f in V_j , and by A_j the approximation of an operator A on V_j .

The multiscale decomposition is obtained from the wavelet spaces. The wavelet space W_j is such that $V_{j+1} = V_j \oplus W_j$. It is spanned by the translates and scalings $\psi_{j,k}$ of a mother wavelet ψ . This leads to the multiscale decomposition

$$V_j = V_0 \oplus W_0 \oplus \dots \oplus W_{j-1}.$$

The scaling function can be chosen with compact support, and with any order C^k smoothness. Together with the V_j , a dual multiresolution analysis \tilde{V}_j can be

constructed, spanned by translates and scalings of a dual scaling function $\tilde{\phi}$, such that the basis functions satisfy the biorthogonality property $\langle \tilde{\phi}_{j,k}, \phi_{j,k'} \rangle = \delta_{k,k'}$. One of $\phi, \tilde{\phi}$ can be chosen as a compactly supported spline, we assume ϕ is a spline, and V is a spline space of a certain order. The space V_j can be made to satisfy $V_j \subset H^s$ for any s by choosing wavelets of sufficiently high order of smoothness. Throughout the analysis we will assume sufficient smoothness of the wavelets, without specifying this precisely.

The error estimates and assumptions on the smoothness of initial values are formulated in terms of regularity in L^2 based Sobolev spaces. That is natural and convenient for wave equations (where physical energy conservation holds). It is also easy to handle in wavelet discretizations, because of norm equivalences. The Sobolev norms $\|\cdot\|_{H^s}$ are equivalent to weighted norms of the wavelet coefficients. If

$$f = \sum_{k=0}^{K-1} c_{-1,k} \phi_{0,k} + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j L} c_{j,k} \psi_{j,k},$$

and the wavelets are sufficiently smooth, then there is the norm equivalence

$$\|f\|_{H^\alpha(\Omega)}^2 \sim \sum_{j=-1}^{\infty} \sum_k |2^{\alpha j} c_{j,k}|^2.$$

From these norm equivalences one can easily derive an important approximation result. Assume that f is in H^α ; then the projections $\Pi_{V_j} f$ of f to the V_j satisfy

$$\|f - \Pi_{V_j} f\|_{L^2(\Omega)} \leq C 2^{-\alpha j} \|f\|_{H^\alpha(\Omega)}.$$

In our application we typically deal with products of operators that are applied after each other, in discrete form, to a discretized function. We first derive a criterion, that we call *order k approximation operator*, for each of the operators to satisfy, such that such products converge. After this we will argue that the operators in our application can be approximated such that the approximation indeed satisfies the approximation property.

Suppose A is some operator $H^{s_1} \rightarrow H^{s_2}$, and A_j is a discrete approximation to A . As pointed out, convergence estimates are done using additional regularity, say k additional orders of regularity. For our operator A from $H^{s_1} \rightarrow H^{s_2}$ we therefore assume its argument, say f is in H^{s_1+k} . The result Af may be the argument of another operator, so we will require $Af \in H^{s_2+k}$; in other words we will assume

$$A \text{ is continuous } H^{s_1+s} \rightarrow H^{s_2+s} \text{ for } 0 \leq s \leq k.$$

Next we discuss a property that ensures that $A_j f_j$ approximates Af if f_j approximates f .

Definition. Let A be as just described; then we say A and A_j satisfy the *order k approximation property* if

$$\|A - A_j\|_{H^{s_1+k} \rightarrow H^{s_2}} \leq C 2^{-jk}.$$

This also implies that A_j is continuous $H^{s_1+s} \rightarrow H^{s_2+s}$ for $0 \leq s \leq k$. This implies that if a function $f \in H^{s_1+k}$ is approximated in H^{s_1} by functions f_j , with

the convergence as expected from the additional regularity, i.e., $\|f - f_j\|_{H^{s_1}} \leq C2^{-kj} \|f\|_{H^{s_1+k}}$, then $A_j f_j$ approximates Af in the same way, since

$$\begin{aligned} \|Af - A_j f_j\|_{H^{s_2}} &\leq \|A_j(f - f_j)\|_{H^{s_2}} + \|(A - A_j)f\|_{H^{s_2}} \\ &\leq C2^{-jk} \|f\|_{H^{s_1+k}}. \end{aligned}$$

We will assume that k is an integer, although this does not seem essential, and that $k \geq 1$.

The basic operators needed here are partial differential operators, the operator $(-\partial_x^2 + 1)^{-1}$ or inverses of higher-order elliptic operators for the approximation of the operator S of section 7, and the pull back along the characteristic flow (which is a smooth coordinate transformation). Here we discuss partial differential operators and constant coefficient inverse partial differential operators; the pull back will be discussed in the last part of this section. We state the result on the approximation of $R^{(-K)}$, $Q^{(-K)}$ as a lemma.

LEMMA 8.1. *Assume the coefficients a, b are $C^{k+K+1,1}$. Then numerical approximations to the operators $R^{(-K)}$ on $H^1 \times H^1$, $Q^{(-K)}$ from $H^1 \times H^1 \rightarrow H^1 \times L^2$ and $(Q^{(-K)})^{-1}$ from $H^1 \times L^2 \rightarrow H^1 \times H^1$ can be constructed that satisfy the order k approximation property.*

Proof. Multiplication by polynomials and differentiation operators can be discretized using results of [8]; see that reference or section 2.5 of [5]. They can be discretized at cost $O(N)$, in such a way that the above order k approximation property is satisfied. For multiplication operators with functions other than polynomials, the coefficient is locally approximated by polynomials. As for the regularity requirement on the coefficients, for an approximate multiplication operator on H^{s_1} to have the order k approximation property, it is sufficient to have $C^{k+s_1-1,1}$ coefficients, since a $C^{k-1,1}$ function can be approximated to error 2^{-jk} by polynomials on regions of size order 2^{-j} .

In the case of the approximation of $R^{(-K)}$ on $H^1 \times H^1$, the coefficients in the remainder term need to be $C^{k,1}$. It follows that the coefficients a and b must be in $C^{k+K+1,1}$.

The operator $(-\partial_x^2 + 1)^{-1}$ can be computed in $O(N)$ cost using a multigrid algorithm [1]; a wavelet variant of this algorithm was given in [5]. To show that the approximation property holds, a slight change in the argument about multilevel preconditioning in example 4 in section 3.11 of [5] is needed; namely, n_j is chosen such that $\rho^{n_j} \leq 2^{-t'j}$, with $t' > t$. Similar arguments work for the higher-order inverse elliptic operator Ψ . This concludes the proof. \square

Next we will show a similar result for $E(t, t_0)$. This operator was diagonal with $E_{2,2}$ given by (see (3.5))

$$(8.1) \quad E_{2,2}(t, t_0)u_2(x) = u_2(t_0, X(x, t_0, t)).$$

The 1,1 component of $E(t, t_0)$ is given by a similar formula.

We will first discuss the approximation of $X(x, t, t_0)$; then the next lemma will contain the result on $E(t, t_0)$.

Let $X_j(x, t, t_0)$ denote a numerical approximation used at level j . This must be computed for a set of points x . We require increasing accuracy as j increases, with error bounded by $C2^{-j(k+1)}$. It is allowed that, as j increases, the computational cost increases as 2^j . We find that for the time-independent case C^{k+1} smoothness of

the coefficients is sufficient, while for the time-dependent case C^{2k+2} smoothness is sufficient for this computation, as we will now show.

For the time-independent case, the evaluation of (8.1) can be done by solving $X = X(x, t, t_0)$ from

$$(8.2) \quad \int_x^X c(\xi)^{-1} d\xi = t - t_0.$$

First the primitive $\int_0^x c(\xi)^{-1} d\xi$ is computed for all x in the periodic grid with grid distance 2^{-j} . Assuming that c is C^{k+1} , this can be done at cost $O(2^j)$, with error $\leq C2^{-j(k+1)}$. Next the solution of (8.2) can be done for a set of 2^j points x using interpolation, which conserves the order of error, i.e., with error still bounded by $C2^{-j(k+1)}$.

For the time-dependent case we solve for the characteristics using a Runge–Kutta method of order $2k + 2$. We require C^{2k+2} smoothness of c ; then we can take order $2^{j/2}$ points with distance between them of $2^{-j/2}$ and solve with time steps of order $2^{-j/2}$. The total error is then bounded by $C2^{-j(k+1)}$.

Next we discuss how (8.1) can be computed numerically such that the order k approximation property is satisfied.

LEMMA 8.2. *Assume the coefficients a, b are C^{k+1} for the time-independent case or C^{2k+2} in the time-dependent case, and the wavelets are order $k+1$ splines. Then a numerical approximation to the operator $E(t, t_0)$ on $H^1 \times H^1$ can be constructed that satisfies the order k approximation property.*

Proof. We consider the approximation at level J of $E_{2,2}(t, t_0)f$, with f an element of V_J . We have that $E_{2,2}(t, t_0)f(x) = f(X(x, t_0, t))$. For brevity we will write $X(x)$ instead of $X(x, t_0, t)$. We will write $h(x) = f(X(x))$. We want to compute $c_{J,\tilde{k}} = \langle \tilde{\phi}_{J,\tilde{k}}, h \rangle$. The computation of matrix elements of polynomials, i.e., $\langle \tilde{\phi}_{J,\tilde{k}}, p \rangle$, when p is a polynomial, is basically exact; see the method of section 2.5 of [5]. To compute matrix elements of other smooth functions, it is common to approximate these locally by polynomials, and we will also use this in this argument. So to compute the approximate coefficient of the scaling function $\phi_{J,\tilde{k}}$, the function h is approximated around the support of $\phi_{J,\tilde{k}}$ by a polynomial p . The approximate value of the coefficient is then $\tilde{c}_{J,\tilde{k}} = \langle \tilde{\phi}_{J,\tilde{k}}, p \rangle$ and is obtained according to the mentioned section of [5].

Thus we must define how to approximate h locally by a polynomial. This can simply be done by polynomial interpolation with an order k polynomial. A function h in $C^{k,1}$ can be approximated by interpolation on a grid of size 2^{-J} up to an error bounded by

$$\sup_{x \in S_{J,\tilde{k}}} |h(x) - p(x)| \leq C2^{-(k+1)J} \|h\|_{C^{k,1}(S_{J,\tilde{k}})}.$$

We will apply this to a wavelet, $f = \psi_{j,\hat{k}}$. We assume that the wavelet ψ is $C^{k,1}$ and use that X is also $C^{k,1}$. The function $h(x) = \psi_{j,\hat{k}}(X(x))$ satisfies $\|\psi_{j,\hat{k}}(X(\cdot))\|_{C^{k,1}(S_{J,\tilde{k}})} \leq C2^{j(k+3/2)}$. Thus the error with p an exact interpolating polynomial is given by

$$|c_{J,\tilde{k}} - \tilde{c}_{J,\tilde{k}}| \leq \|\tilde{\phi}_{J,\tilde{k}}\|_{L^1} \sup_{x \in S_{J,\tilde{k}}} |h(x) - p(x)| \leq C2^{J(-k-3/2)+j(k+3/2)} \leq C2^{(k+1)(j-J)}.$$

Here we used that $\|\tilde{\phi}_{J,\tilde{k}}\|_{L^1}$ can be bounded by $C2^{-J/2}$ (which has to do with the normalization; the L^2 norm of $\tilde{\phi}_{J,\tilde{k}}$ is normalized to unity). Thus we find that the

map from f to the error $\sum_{\hat{k}}(c_{J,\hat{k}} - \tilde{c}_{J,\hat{k}})\phi_{J,\hat{k}}$ is bounded by $C2^{-(k+1)J}$ from H^{k+1} to L^2 , and hence by $C2^{-kJ}$ from H^{k+1} to H^1 .

A second source of error is that $X_J(x)$ is used instead of the exact value $X(x)$. For these errors we have

$$\psi_{j,\hat{k}}(X_J(x)) - \psi_{j,\hat{k}}(X(x)) = \int_{X(x)}^{X_J(x)} \frac{d\psi_{j,\hat{k}}}{dx}(s)ds.$$

Since $\frac{d\psi_{j,\hat{k}}}{dx}$ is bounded by $C2^{3j/2}$, and $|X_J(x) - X(x)| < C2^{-J(k+1)}$, these errors satisfy

$$|\psi_{j,\hat{k}}(X_J(x)) - \psi_{j,\hat{k}}(X(x))| \leq C2^{3j/2 - J(k+1)}.$$

From this a bound $C2^{-J(k+1)}$ follows for the map from input to this error, considered in spaces $H^{3/2} \rightarrow L^2$, and a bound $C2^{-JK}$ from $H^{3/2} \rightarrow H^1$, which is better than or equal to the bound for the interpolation error, since $k > 1/2$. \square

9. Multiscale time-stepping and proof of the theorem. In this section multiscale time-stepping is introduced to finally obtain an $O(N)$ algorithm. The results of section 7 enable the use of higher-order time-stepping methods and lead to estimates for the time discretization errors. The results of section 8 allow us to estimate the errors due to space discretization. Here we will combine space and time discretization, choose parameters, like the order of space and time discretization, and establish the complexity of the algorithm by estimating error and cost of the algorithm.

We solve the equivalent of differential equation (3.2) with higher-order decoupling, after the application of the integrating factor; i.e., we solve

$$(9.1) \quad \frac{dv}{dt}(t) = S(t, t_0)v(t),$$

with

$$S(t, t_0) = E(t, t_0)^{-1}R^{(-K)}E(t, t_0),$$

where $R^{(-K)}$ is as constructed in section 7. We will approximate the solution $v(t_1)$ starting from t_0 . The approximation is done in $H^1 \times H^1$. The initial values $v_0 = u_0$ also must be in $H^1 \times H^1$. We assume they have α additional orders of regularity; i.e., they are in fact in $H^{1+\alpha} \times H^{1+\alpha}$. It follows from the results of sections 7 and 8 that we can transform the values $U(t)$ of the original system (1.3) to those of the transformed system (9.1) and back with complexity $O(N)$.

Operators will be approximated with the order k approximation property, with $k > \alpha$. A minimum value for k is derived below. Regularity assumptions follow from these assumptions according to the previous sections. Note that this is different from the previous section, where the order k corresponded to the order of additional regularity of functions that operators acted on, while here $k > \alpha$. By S_j we denote an approximation of S in $V_j \times V_j$ with the order k approximation property, according to the methods of section 8. (Note that $S_j \neq \Pi_{V_j} S \Pi_{V_j}$.)

In general in an integrating factor method it is common to frequently reset t_0 , so that $E(t, t_0)$ propagates only over small time intervals. We will refrain from doing so, as this is not needed in this context, and the frequent application of $E(t, t_0)$ to the full

signal (i.e., not only the addition made during a small time interval by a Runge–Kutta time step) may cause additional errors.

As motivated in section 6, we will make a multiscale decomposition of the signal and do time-stepping separately for each scale. The initial values are decomposed as follows:

$$u_0 = \sum_{j=0}^J w_{0,j},$$

with $w_{0,0} = \Pi_{V_0} u_0$, and $w_{0,j} = \Pi_{W_{j-1}} u_0$, for $j = 1, \dots, J$. Here Π_{V_j}, Π_{W_j} denote the projection on $V_j \times V_j$ and $W_j \times W_j$, respectively. The field $v(t)$ will also be decomposed. The j th component, corresponding to initial values in $W_{j-1} \times W_{j-1}$, will not be approximated in $V_j \times V_j$, however (nor in $W_{j-1} \times W_{j-1}$), but in a space $V_{l(j)} \times V_{l(j)}$, $j \leq l(j) \leq J$. To indicate this we write the components of the sum as $v_{j,l(j)}$. We will show that $v(t)$ can be approximated like

$$v(t) \approx \sum_{j=0}^J v_{j,l(j)}.$$

The motivation for doing this is simple: Large errors would result in the time propagation in $V_j \times V_j$ of the $w_{0,j}$, while large cost would result if we would work in the full space $V_J \times V_J$. By working in an intermediate space both cost and errors can be controlled.

The final numerical approximation will be a sum of components $w_{j,l(j),\Delta t_j}$. The terms describe the discrete time propagation with time step Δt_j , using the space discretized operators $S_{l(j)}(t)$, applied to the initial values $w_{0,j}$.

For purposes of error estimation we consider two sets of fields in addition to $w_{j,l(j),\Delta t_j}$. We assume the fields $v_{j,l(j)}$ introduced above describe the continuous time propagation of the operator $\Pi_{V_{l(j)}} S \Pi_{V_{l(j)}}$, and the field $v_{j,l(j),\Delta t_j}$ will describe the discrete time propagation of $\Pi_{V_{l(j)}} S \Pi_{V_{l(j)}}$.

We first establish that $v_J(t_1)$ can be approximated like

$$v_J(t_1) \approx \sum_{k=0}^J v_{j,l(j)}(t_1).$$

LEMMA 9.1. *Suppose $l(j)$ is such that*

$$(9.2) \quad k(l(j) - j) = \alpha(J - j).$$

Then

$$(9.3) \quad \left\| \sum_{j=0}^J v_{j,l(j)}(t_1) - v(t_1) \right\|_{H^1 \times H^1} \leq C 2^{-\alpha J} \|u_0\|_{H^{1+\alpha} \times H^{1+\alpha}}.$$

Proof. Let $v_{j,\infty}$ denote the solution of the exact differential equation with initial value $w_{0,j}$. It satisfies $\frac{dv_{j,\infty}}{dt} = S v_{j,\infty}$. As S is bounded on $H^{1+s} \times H^{1+s}$, $0 \leq s \leq k$, it follows that $v_{j,\infty}(t)$ satisfies the bound

$$\|v_{j,\infty}(t)\|_{H^{1+s} \times H^{1+s}} \leq C \|w_{0,j}\|_{H^{1+s} \times H^{1+s}}$$

for $0 \leq s \leq k$, $t_0 \leq t \leq t_1$.

We have $\frac{dv_{j,l(j)}}{dt} = \Pi_{V_{l(j)}} S \Pi_{V_{l(j)}} v_{j,l(j)}$, so the difference $v_{j,l(j)} - v_{j,\infty}$ satisfies

$$(9.4) \quad \frac{dv_{j,l(j)} - v_{j,\infty}}{dt} = \Pi_{V_{l(j)}} S \Pi_{V_{l(j)}} (v_{j,l(j)} - v_{j,\infty}) + (\Pi_{V_{l(j)}} S \Pi_{V_{l(j)}} - S)v_{j,\infty}.$$

By standard estimates for ODEs we have that

$$\begin{aligned} \|v_{j,l(j)}(t) - v_{j,\infty}(t)\|_{H^{1+s} \times H^{1+s}} &\leq C_1 \|v_{j,l(j)}(t_0) - v_{j,\infty}(t_0)\|_{H^{1+s} \times H^{1+s}} \\ &\quad + C_2 \int_{t_0}^t \|(\Pi_{V_{l(j)}} S \Pi_{V_{l(j)}} - S)v_{j,\infty}(s)\|_{H^{1+s} \times H^{1+s}} ds. \end{aligned}$$

The first term on the right-hand side is zero. For the second term we use that by the regularity assumptions we have

$$(9.5) \quad \|\Pi_{V_{l(j)}} S \Pi_{V_{l(j)}} - S\|_{H^{1+k} \times H^{1+k} \rightarrow H^1 \times H^1} \leq C2^{-kl(j)}.$$

The components of the initial values $w_{0,j}$ are bounded according to

$$(9.6) \quad \|w_{0,j}\|_{H^{1+k} \times H^{1+k}} \leq C2^{j(k-\alpha)} \|w_{0,j}\|_{H^{1+\alpha}},$$

and the same is true for $v_{j,\infty}(t)$ for $t_0 < t < t_1$. The inhomogeneous term in (9.4) can therefore be bounded by

$$\begin{aligned} \|(\Pi_{V_{l(j)}} S \Pi_{V_{l(j)}} - S)v_{j,\infty}(t)\|_{H^1 \times H^1} &\leq C2^{-kl(j)+j(k-\alpha)} \|w_{0,j}\|_{H^{1+\alpha} \times H^{1+\alpha}} \\ &= C2^{-\alpha J} \|w_{0,j}\|_{H^{1+\alpha} \times H^{1+\alpha}}. \end{aligned}$$

The error $v_{j,l(j)}(t_1) - v_{j,\infty}(t_1)$ therefore satisfies the bound

$$(9.7) \quad \|v_{j,\infty}(t_1) - v_{j,l(j)}(t_1)\|_{H^1 \times H^1} \leq C2^{-\alpha J} \|w_{0,j}\|_{H^{1+\alpha} \times H^{1+\alpha}}.$$

Adding the estimates for each j results in (9.3). □

The second step in the estimation of the error is to estimate the time discretization error for the field $v_{j,l(j)}$. We will argue that the fields $v_{j,l(j)}$ can be sufficiently accurately approximated using Runge–Kutta time discretization. By $v_{j,l(j),\Delta t_j}$ we denote the time-discretized fields. We assume the use of an order K Runge–Kutta method for the time-stepping.

LEMMA 9.2. *Suppose that the time step Δt_j satisfies the inequality*

$$(9.8) \quad \Delta t_j \leq C2^{-\alpha(J-j)/K},$$

and that the coefficients a, b are at least $C^{2K+1,1}$; then we have

$$(9.9) \quad \left\| \sum_{j=0}^J v_{j,l(j),\Delta t_j}(t_1) - \sum_{j=0}^J v_{j,l(j)}(t_1) \right\|_{H^1 \times H^1} \leq C2^{-\alpha J} \|u_0\|_{H^{1+\alpha} \times H^{1+\alpha}}.$$

Proof. The error per time step in

$$\|v_{j,l(j)} - v_{j,l(j),\Delta t_j}\|_{H^1 \times H^1}$$

is bounded by

$$(\Delta t_j)^{K+1} \sup_{\tau \in [t, t+\Delta t_j]} \left\| \frac{d^{K+1} v_{j,l(j)}(\tau)}{dt^{K+1}} \right\|_{H^1 \times H^1}.$$

Using the differential equation, the higher-order time derivative $\frac{d^{K+1} v_j(\tau)}{dt^{K+1}}$ can be expanded as a sum of terms that are each given by a product of factors $\frac{d^\gamma}{dt^\gamma} \Pi_{V_{l(j)}} S \Pi_{V_{l(j)}}$ (total sum of the γ 's is $\leq K$) acting on $v_{j,l(j)}(\tau)$. In section 7 it was shown that with the given smoothness assumption on a, b , the time derivatives $\frac{d^j S}{dt^j}$ were bounded operators on $H^1 \times H^1$ for $j = 0, \dots, K$. The same is true for $\frac{d^j}{dt^j} \Pi_{V_{l(j)}} S \Pi_{V_{l(j)}}$. It follows that the error per time step is bounded by

$$C(\Delta t_j)^{K+1} \sup_{\tau \in [t, t+\Delta t_j]} \|v_{j,l(j)}(\tau)\|_{H^1 \times H^1}.$$

Using standard arguments to go from local to global error, we find that the error at time t_1 can be estimated by

$$\|v_{j,l(j)}(t_1) - v_{j,l(j),\Delta t_j}(t_1)\|_{H^1 \times H^1} \leq C(\Delta t_j)^K \|w_{0,j}\|_{H^1 \times H^1}.$$

We have that

$$\sum_{j=0}^J (2^{\alpha j} \|w_{0,j}\|_{H^1 \times H^1})^2$$

is bounded. We therefore require that

$$(9.10) \quad (\Delta t_j)^K \leq C 2^{\alpha j} 2^{-\alpha J};$$

then (9.9) follows. The conditions (9.8) and (9.10) are of course equivalent. \square

For the estimate of the time discretization error it turned out to be convenient to work with $\Pi_{V_{l(j)}} S \Pi_{V_{l(j)}}$, an exact discretization that is not practical to compute, instead of $S_{l(j)}$, the approximate discretization discussed in section 8. The reason is that the errors made in $S_{l(j)}$ are not differentiable. So the next step is to take into account the difference between $S_{l(j)}$ and $\Pi_{V_{l(j)}} S \Pi_{V_{l(j)}}$.

LEMMA 9.3. *Assume still (9.2). We have the estimate*

$$(9.11) \quad \left\| \sum_{j=0}^J w_{j,l(j),\Delta t_j}(t_1) - \sum_{j=0}^J v_{j,l(j),\Delta t_j}(t_1) \right\|_{H^1 \times H^1} \leq C 2^{-\alpha J} \|u_0\|_{H^{1+\alpha} \times H^{1+\alpha}}.$$

Proof. The difference $S_{l(j)} - \Pi_{V_{l(j)}} S \Pi_{V_{l(j)}}$ satisfies a similar estimate as the difference $\Pi_{V_{l(j)}} S \Pi_{V_{l(j)}} - S$, which was considered in the proof of Lemma 9.1. The proof of (9.11) therefore proceeds similarly as the proof of Lemma 9.1, except that difference equations are used instead of differential equations. The difference $w_{j,l(j),\Delta t_j} - v_{j,l(j),\Delta t_j}$ satisfies the linear inhomogeneous difference equation

$$\begin{aligned} & w_{j,l(j),\Delta t_j}(t + \Delta t) - v_{j,l(j),\Delta t_j}(t + \Delta t) \\ &= \Delta t \text{RKStep}(t, \Delta t, S_{l(j)})(w_{j,l(j),\Delta t_j}(t) - v_{j,l(j),\Delta t_j}(t)) \\ &+ \Delta t (\text{RKStep}(t, \Delta t, S_{l(j)}) - \text{RKStep}(t, \Delta t, \Pi_{V_{l(j)}} S \Pi_{V_{l(j)}})) v_{j,l(j),\Delta t_j}(t), \end{aligned}$$

where $\Delta t \text{RKStep}(t, \Delta t, A)y$ denotes the Runge–Kutta step for the equation $y' = Ay$, which is a linear map on y . It follows that

$$\begin{aligned} & \|w_{j,l(j),\Delta t_j}(\hat{t}) - v_{j,l(j),\Delta t_j}(\hat{t})\|_{H^1 \times H^1} \leq C \Delta t_j \\ \times \sum_{t\text{-values} < \hat{t}} & \|(\text{RKStep}(t, \Delta t, S_{l(j)}) - \text{RKStep}(t, \Delta t, \Pi_{V_{l(j)}} S \Pi_{V_{l(j)}}))v_{j,l(j),\Delta t_j}(t)\|_{H^1 \times H^1}. \end{aligned}$$

The difference $\text{RKStep}(t, \Delta t, S_{l(j)}) - \text{RKStep}(t, \Delta t, \Pi_{V_{l(j)}} S \Pi_{V_{l(j)}})$ can be worked out. It is a product of $S_{l(j)} - \Pi_{V_{l(j)}} S \Pi_{V_{l(j)}}$ and of operators that are bounded on H^{1+s} , $0 \leq s \leq k$. It follows that we have the estimate

$$\|\text{RKStep}(t, \Delta t, S_{l(j)}) - \text{RKStep}(t, \Delta t, \Pi_{V_{l(j)}} S \Pi_{V_{l(j)}})\|_{H^{1+k} \times H^{1+k} \rightarrow H^1 \times H^1} \leq C 2^{-kl(j)}.$$

Furthermore

$$\|v_{j,l(j),\Delta t_j}(t)\|_{H^{1+k} \times H^{1+k}} \leq 2^{j(k-\alpha)} \|w_{0,j}\|_{H^{1+\alpha}}.$$

It follows that we can estimate

$$\begin{aligned} \|w_{j,l(j),\Delta t_j}(t_1) - v_{j,l(j),\Delta t_j}(t_1)\| & \leq C 2^{-kl(j)+j(k-\alpha)} \|w_{0,j}\|_{H^{1+\alpha} \times H^{1+\alpha}} \\ & = C 2^{-\alpha J} \|w_{0,j}\|_{H^{1+\alpha} \times H^{1+\alpha}}. \end{aligned}$$

The estimate (9.11) trivially follows from this. \square

This ends our estimation of the error. The *cost* of this time-stepping is

$$\begin{aligned} C \sum_{j=0}^J (\Delta t_j)^{-1} 2^{l(j)} & = C \sum_{j=0}^J 2^{\alpha(J-j)/K + \alpha(J-j)/k + j} \\ & = C 2^J \sum_{j=0}^J 2^{(-1+\alpha/K + \alpha/k)(J-j)}. \end{aligned}$$

The requirement is that the cost is bounded by CN , and hence that $-1 + \alpha/K + \alpha/k < 0$. If we allow logarithmic cost $O(N \log N)$, equality is also allowed. We hence have our final result.

THEOREM 9.4. *If a K th-order Runge–Kutta scheme is used, if the operators S_j are approximated using the order k approximation property, with, in particular, order $k+1$ spline wavelets, if the initial data u_0 are in $H^{1+\alpha} \times H^{1+\alpha}$, if coefficient functions are at least $C^{K+1+\max(k,K),1}$, and if*

$$(9.12) \quad 1/K + 1/k < 1/\alpha,$$

then the algorithm above with $N = L2^J$ degrees of freedom computes an approximation with error bound

$$(9.13) \quad \left\| \sum_{j=0}^J w_{j,l(j),\Delta t_j}(t_1) - v(t_1) \right\|_{H^1 \times H^1} \leq CN^{-\alpha} \|u_0\|_{H^{1+\alpha} \times H^{1+\alpha}}$$

at a cost $O(N)$. If

$$(9.14) \quad 1/K + 1/k = 1/\alpha,$$

it satisfies the same error bound at cost $O(N \log N)$.

The requirement that u_0 is in $H^{1+\alpha} \times H^{1+\alpha}$ means that the initial values U_0 for the original system (1.3) must be in $H^{1+\alpha} \times H^\alpha$.

In (9.13) it may look like we are summing J functions of N sample points, with cost $O(JN) = O(N \log N)$. However, this is not the case. The terms $w_{j,l(j),\Delta t_j}(t_1)$ have $C2^{l(j)}$ sample points (being in $V_{l(j)}$). Using the wavelet spaces, and the fast wavelet transform (which is $O(N)$ for N sample points), the summation can be done at cost $C \sum_{j=0}^J 2^{l(j)} \leq C2^J = O(N)$.

10. Discussion. A numerical method for wave propagation in smooth media was developed. The numerical results in section 5 show that the method certainly has potential in applications with relatively smooth media. Further improvements might be possible to further improve computation speed or weaken the requirements of medium smoothness. One step that could possibly give an improvement is a coordinate change that makes the wave speed equal to unity. We refrained from doing this since it has no equivalent in higher dimensions, but it could reduce the error in the application of the operator T .

The material of sections 6 to 9 not only leads to the $O(N)$ complexity result but also suggests ways to possibly improve the method.

The main question for future research is in our view about the generalization to higher-dimensional cases. For the multidimensional case, curvelets form a redundant basis (frame) with respect to which the solution operator can be made sparse [4]. Potentially it could be used for computations. However, one needs to be able to implement operators that give the approximate effect of wave propagation, such as translation, rotation, and deformation, efficiently in a curvelet basis. Perhaps other fast implementations of Fourier integral operators could be used (cf. [3]) to compute the approximate wave propagation. In dimension 2 and higher the remainder operator R becomes, at least in the continuous setting, a pseudodifferential operator, which is more challenging to implement. But a priori there is no reason why the principle of combining an approximate solution operator with lower-order, exact “corrections” could not be extended to higher dimensions.

REFERENCES

- [1] R. E. BANK AND T. DUPONT, *An optimal order process for solving finite element equations*, Math. Comp., 36 (1981), pp. 35–51.
- [2] G. BEYLKIN AND K. SANDBERG, *Wave propagation using bases for bandlimited functions*, Wave Motion, 41 (2005), pp. 263–291.
- [3] E. CANDÈS, L. DEMANET, AND L. YING, *Fast computation of Fourier integral operators*, SIAM J. Sci. Comput., 29 (2007), pp. 2464–2493.
- [4] E. J. CANDÈS AND L. DEMANET, *The curvelet representation of wave propagators is optimally sparse*, Comm. Pure Appl. Math., 58 (2005), pp. 1472–1528.
- [5] A. COHEN, *Numerical Analysis of Wavelet Methods*, Stud. Math. Appl. 32, North-Holland, Amsterdam, 2003.
- [6] G. C. COHEN, *Higher-Order Numerical Methods for Transient Wave Equations*, Sci. Comput., Springer-Verlag, Berlin, 2002.
- [7] W. DAHMEN, *Wavelet and multiscale methods for operator equations*, in Acta Numerica, 1997, Cambridge University Press, Cambridge, UK, 1997, pp. 55–228.
- [8] W. DAHMEN AND C. A. MICCHELLI, *Using the refinement equation for evaluating integrals of wavelets*, SIAM J. Numer. Anal., 30 (1993), pp. 507–537.
- [9] L. DEMANET AND L. YING, *Wave atoms and time upscaling of wave equations*, Numer. Math., to appear.
- [10] J. J. DUISTERMAAT, *Fourier Integral Operators*, Birkhäuser, Boston, 1996.

- [11] J. J. DUISTERMAAT AND L. HÖRMANDER, *Fourier integral operators II*, Acta. Math., 128 (1972), pp. 183–269.
- [12] B. ENGQUIST, S. OSHER, AND S. ZHONG, *Fast wavelet based algorithms for linear evolution equations*, SIAM J. Sci. Comput., 15 (1994), pp. 755–775.
- [13] R. J. LEVEQUE, *Convergence of a large time step generalization of Godunov’s method for conservation laws*, Comm. Pure Appl. Math., 37 (1984), pp. 463–477.
- [14] J. L. LIONS AND E. MAGENES, *Non-homogeneous Boundary Value Problems and Applications*, Vol. 1, Springer-Verlag, Berlin, 1972.
- [15] H. F. SMITH, *A Hardy space for Fourier integral operators*, J. Geom. Anal., 8 (1998), pp. 629–653.
- [16] C. C. STOLK, *On the Modeling and Inversion of Seismic Data*, Ph.D. thesis, Utrecht University, Utrecht, The Netherlands, 2000.
- [17] G. STRANG, *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal., 5 (1968), pp. 506–517.
- [18] M. E. TAYLOR, *Reflection of singularities of solutions to systems of differential equations*, Comm. Pure Appl. Math., 28 (1975), pp. 457–478.
- [19] M. E. TAYLOR, *Pseudodifferential Operators*, Princeton University Press, Princeton, NJ, 1981.
- [20] L. N. TREFETHEN, *Spectral Methods in MATLAB*, Software Environ. Tools 10, SIAM, Philadelphia, 2000.
- [21] F. TREVES, *Introduction to Pseudodifferential and Fourier Integral Operators*, Vol. 2, Plenum Press, New York, 1980.

STABLE AND COMPATIBLE POLYNOMIAL EXTENSIONS IN THREE DIMENSIONS AND APPLICATIONS TO THE p AND h - p FINITE ELEMENT METHOD*

BENQI GUO[†] AND JIANMING ZHANG[‡]

Abstract. Polynomial extensions play a vital role in the analysis of the p and h - p finite element method (FEM) and the spectral element method. We construct explicitly polynomial extensions on standard elements: cubes and triangular prisms, which together with the extension on tetrahedrons are used by the p and h - p FEM in three dimensions. These extensions are proved to be stable and compatible with FEM subspaces on tetrahedrons, cubes, and prisms and realize a continuous mapping: $H_{00}^{1/2}(T)$ (or $H_{00}^{1/2}(S)$) $\rightarrow H^1(\Omega_{st})$, where Ω_{st} denotes one of these standard elements and T and S are their triangular and square faces. Applications of these polynomial extensions to the p and h - p FEM are illustrated.

Key words. the p and h - p version, finite element method, polynomial extension, tetrahedron, hexahedron, prism, pyramid, cube, Sobolev spaces, Jacobi polynomials

AMS subject classifications. 65N30, 65N25, 35D10

DOI. 10.1137/070688006

1. Introduction. In the analysis of the high-order finite element method (FEM), such as the p and h - p versions of FEM and the spectral element method, we need to construct a globally continuous and piecewise polynomial which has the optimal estimation for its approximation error and satisfies homogeneous or nonhomogeneous Dirichlet boundary conditions. The construction of such a polynomial is started with local polynomial projections on each element for the best rate of convergence. Unfortunately, a union of local polynomial projections is not globally continuous and does not satisfy the homogeneous Dirichlet boundary conditions or the nonhomogeneous Dirichlet boundary conditions. In the context of the continuous Galerkin method in two and three dimensions, we have to adjust these local polynomial projections by a special technique called polynomial extension or lifting. Hence, it is essential for us to build a polynomial extension compatible with FEM subspaces, by which the union of local polynomial projections can be modified to a globally continuous polynomial without degrading the best order of approximation error. Compatible polynomial extensions together with local projections led to the best estimation in the approximation error for the p and h - p FEM [1, 2, 5, 6, 16, 21].

Babuška and Suri [5] proposed an extension F on a triangle T with $I = (0, 1)$ as one of its sides, which realizes a continuous mapping $H^{1/2}(I) \rightarrow H^1(T)$ such that $Ff \in \mathcal{P}_p^1(T)$ for $f \in \mathcal{P}_p(I)$. The extension is the convolution of f and a characteristic function. Using this extension they proved the existence of the continuous extension

*Received by the editors April 12, 2007, accepted for publication (in revised form) September 22, 2008; published electronically February 25, 2009.

<http://www.siam.org/journals/sinum/47-2/68800.html>

[†]Department of Mathematics, Shanghai Normal University, Shanghai, China and Department of Mathematics, University of Manitoba, Winnipeg, MB R3T 2N2, Canada (guo@cc.umanitoba.ca). The work of this author was partially supported by NSERC of Canada under grant OGP0046726 and partially supported by the Computational Science E-Institute of Shanghai Universities under project E03004.

[‡]Department of Mathematics, University of Manitoba, Winnipeg, MB R3T 2N2, Canada (umzhan37@cc.umanitoba.ca). The work of this author was partially supported by the University of Manitoba and by NSERC of Canada under grant OGP0046726.

R: $H_{00}^{1/2}(\Gamma) \rightarrow H^1(T)$ [3, 5] such that $Rf \in \mathcal{P}_p^1(T)$ for $f \in \mathcal{P}_p^0(I)$. They generalize the extension on a square $S = (-1, 1)^2$, which realizes a continuous mapping $H_{00}^{1/2}(\Gamma) \rightarrow H^1(S)$ and $Rf \in \mathcal{P}_p^2(S)$ for $f \in \mathcal{P}_p^0(I)$. Hereafter, $\mathcal{P}_p(I)$ denotes a set of polynomial of degree $\leq p$ and $\mathcal{P}_p^0(I)$ is its subset of polynomial vanishing at the endpoints of I , $\mathcal{P}_p^1(\Omega)$ and $\mathcal{P}_p^2(\Omega)$ denote sets of polynomials of total and separate degree $\leq p$ on a domain Ω in \mathbb{R}^n , $n = 2, 3$, respectively, and $\mathcal{P}_p^{m,0}(\Omega)$ is its subset of polynomials vanishing on the boundary of Ω . These polynomial extensions are compatible with FEM subspaces and have been successfully applied to the p and h - p versions of FEM in two dimensions, which lead to the optimal estimate for approximation error in the finite element solution of the p and h - p versions on quasi-uniform meshes with triangular and quadrilateral elements [1, 2, 5, 6, 16]. It was shown [20] that the extension on a triangle or a square defined in [5] is stable in Sobolev spaces. The polynomial extensions in weighted Sobolev spaces on a square were studied in [9] to improve the error estimation of the spectral collection method for an approximation of the Stokes equation. The polynomial extensions in high-order Sobolev spaces were studied in [8].

The extension of convolution-type has been generalized to tetrahedrons [21] and cubes [7] in three dimensions. Muñoz-Sola creatively developed the polynomial extension of convolution-type on tetrahedron K from a triangular face T by introducing the extension operator R (see (2.2)) and gave an explicit proof of continuity of the mapping $H_{00}^{1/2}(T) \rightarrow H^1(K)$ such that $R_K f \in \mathcal{P}_p^1(K)$ if $f \in \mathcal{P}_p^{1,0}(T)$, which is compatible with the FEM subspaces on tetrahedral elements. The polynomial extension R_K together with local projections leads to an error estimation for the h - p FEM on tetrahedral meshes [21]. Unfortunately, the polynomial extension of convolution-type on a cube D is not compatible with FEM subspaces on hexahedral element. Namely, if $f \in \mathcal{P}_p^{2,0}(S)$ where S is a square face of D , the extended polynomial by the convolution will not be in $\mathcal{P}_p^2(D)$, instead, in $\mathcal{P}_p^2(S) \times \mathcal{P}_{2p}(I)$. Also, if $f \in \mathcal{P}_p^{1,0}(S)$, the extended polynomial is in $\mathcal{P}_p^2(D)$. Obviously, $\mathcal{P}_p^1(S)$ is not a trace space of $\mathcal{P}_p^2(D)$ and $\mathcal{P}_p^2(S) \times \mathcal{P}_{2p}(I) \not\subseteq \mathcal{P}_p^2(D)$. It seems that the extension of convolution-type works only for polynomial spaces of total degree $\leq p$ on elements in three dimensions, e.g., $\mathcal{P}_p^1(K)$, but does not work for polynomial spaces of separate degree $\leq p$, e.g., $\mathcal{P}_p^2(D)$. Therefore, we need to develop a new type of extension operator R_D without using convolution.

In this paper we design polynomial extension on cubes by using spectral solutions of the eigenvalue problem of Poisson equation on a square face S and two-point value problem on an interval I . A polynomial extension using eigen-polynomials which forms an L^2 and H^1 orthogonal basis of $\mathcal{P}_p^{2,0}(S)$ and spectral solutions of two-point value problems associated with the eigenvalues realize a continuous mapping $R_D : H_{00}^{1/2}(S) \rightarrow H^1(D)$ and $R_D f \in \mathcal{P}_p^2(D)$ for $f \in \mathcal{P}_p^{2,0}(S)$. Besides tetrahedrons(simplices) and hexahedrons(cubes), triangular prisms are commonly used for FEM in three dimensions. There are two types of different faces of triangular prism: triangle and square. Therefore, we need to construct a polynomial extension from a triangular face and a polynomial extension from a square face. The former one is based on the convolution-type extension on a tetrahedron, and the later one is based on a new extension on a triangle from a side. Both are compatible with FEM subspaces and realize continuous mapping $H_{00}^{1/2}(T) \rightarrow H^1(G)$ and $H^2(S) \cap H_0^1(S) \rightarrow H^1(G)$, respectively.

The rest of the paper is organized as follows. In section 2, after quoting the results on polynomial extension R on tetrahedrons K from [21], a polynomial extension R_K^T from a triangular face T to a triangular prism G is introduced, which is based on the extension on a truncated tetrahedron K_H incorporated with a trilinear mapping of G onto K_H . The continuity of the mapping is proved, and the compatibility with

FEM subspace is verified. Another polynomial extension R_G^S from a square face S to a triangular prism G is constructed, which is as important as R_K^T in the error analysis of FEM on prism elements. In section 3, we construct an extension on a cube D without using convolution, instead using spectral solutions of an eigenvalue problem on a square and a two-point value problem on an interval. It is shown that this polynomial extension realizes a continuous mapping: $H_{00}^{1/2}(S) \rightarrow H^1(D)$ and compatible with FEM subspaces on cubic elements. Applications of the polynomial extensions to error estimation for the p -version of FEM in three dimensions are illustrated in the last section.

2. Polynomial extension on a triangular prism.

2.1. Polynomial extension on a tetrahedron. For the construction of polynomial extensions on a triangular prism, we need to quote results on the extension on a tetrahedron from [21]. We denote, by K , a standard tetrahedron $\{(x_1, x_2, x_3) | x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_1 + x_2 + x_3 \leq 1\}$ in \mathbb{R}^3 shown in Figure 2.1, and ∂K denotes the boundary of K . Let $T = \{(x_1, x_2) | x_1 \geq 0, x_2 \geq 0, x_1 + x_2 \leq 1\}$ be a standard triangle in \mathbb{R}^2 , and let $\Gamma_i, 1 \leq i \leq 3$ be faces of K contained in the plane $x_i = 0$ and Γ_4 be the oblique face.

Muñoz-Sola introduced the following operators [21]:

$$(2.1) \quad F_K f(x_1, x_2, x_3) = \frac{2}{x_3^2} \int_{x_1}^{x_1+x_3} d\xi_1 \int_{x_2}^{x_1+x_2+x_3-\xi_1} f(\xi_1, \xi_2) d\xi_2$$

and

$$(2.2) \quad R_K f(x_1, x_2, x_3) = (1 - x_1 - x_2 - x_3)x_1x_2F_K \tilde{f}(x_1, x_2, x_3),$$

with

$$\tilde{f}(x_1, x_2) = \frac{f(x_1, x_2)}{x_1x_2(1 - x_1 - x_2)}.$$

The operator R_K has the following decomposition:

$$(2.3) \quad R_K f(x_1, x_2, x_3) = (1 - x_1 - x_2 - x_3)R_{12}f(x_1, x_2, x_3) + x_2R_{13}f(x_1, x_2, x_3) + x_1R_{23}f(x_1, x_2, x_3),$$

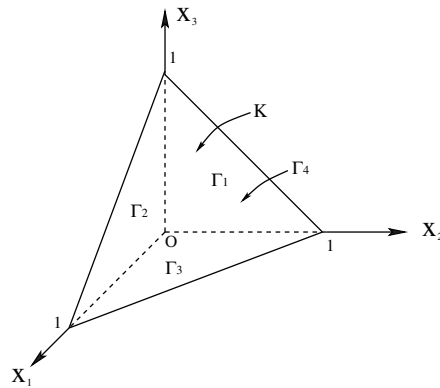


FIG. 2.1. The tetrahedron K .

where

$$(2.4) \quad R_{12}f(x_1, x_2, x_3) = x_1x_2F_K\tilde{f}_{12}(x_1, x_2, x_3), \quad \tilde{f}_{12}(x_1, x_2) = \frac{f(x_1, x_2)}{x_1x_2},$$

$$(2.5) \quad R_{i3}f(x_1, x_2, x_3) = (1 - x_1 - x_2 - x_3)x_iF_K\tilde{f}_{i3}(x_1, x_2, x_3),$$

with

$$\tilde{f}_{i3}(x_1, x_2) = \frac{f(x_1, x_2)}{x_i(1 - x_1 - x_2)}, \quad i = 1, 2.$$

The following theorems were proved in [21].

THEOREM 2.1. *Let R_K be the operator defined by (2.2). Then $R_Kf(x) \in \mathcal{P}_p^1(K)$ for $f \in \mathcal{P}_p^{1,0}(\Gamma_3)$, and*

$$(2.6) \quad \|R_Kf\|_{H^1(K)} \leq C\|f\|_{H_{00}^{\frac{1}{2}}(\hat{\Gamma}_3)},$$

$$(2.7) \quad R_Kf|_{\Gamma_3} = f, \quad R_Kf|_{\Gamma_i} = 0, \quad i = 1, 2, 4,$$

where C is a constant independent of f and p .

THEOREM 2.2. *For $f \in \mathcal{P}_p^1(\partial K) = \{f \in C^0(\partial K) \mid f|_{\Gamma_i} \in \mathcal{P}_p^1(\Gamma_i), 1 \leq i \leq 4\}$, there exists a polynomial $E_Kf \in \mathcal{P}_p^1(K)$ such that $E_Kf|_{\partial K} = f$ and*

$$(2.8) \quad \|E_Kf\|_{H^1(K)} \leq C\|f\|_{H^{1/2}(\partial K)},$$

where C is a constant independent of f and p .

2.2. Polynomial extension on prisms from a triangular face. Let $G = T \times I$ be a triangular prism with faces $\Gamma_i, 1 \leq i \leq 5$ shown in Figure 2.2, where $T = \{(\tilde{x}_1, \tilde{x}_2) \mid \tilde{x}_1 \geq 0, \tilde{x}_2 \geq 0, \tilde{x}_1 + \tilde{x}_2 \leq 1\}$ and $I = [0, 1]$. $\Gamma_i, 1 \leq i \leq 3$ are on the planes $\tilde{x}_i = 0$, Γ_5 is the face of G contained in the plane $\tilde{x}_3 = 1$, and Γ_4 is the face of G contained in the plane $\tilde{x}_1 + \tilde{x}_2 = 1$. Then $\Gamma_3 = T$ and $\Gamma_2 = S = I \times I$. By $\mathcal{P}_p^1(T) \times \mathcal{P}_p(I)$, we denote a set of polynomials with the subtotal degree in \tilde{x}_1 and $\tilde{x}_2 \leq p$ and with the degree $\leq p$ in \tilde{x}_3 . Obviously $\mathcal{P}_p^1(G) \subset \mathcal{P}_p^1(T) \times \mathcal{P}_p(I) \subset \mathcal{P}_p^2(G)$, it is denoted by $\mathcal{P}_p^{1,5}(G)$.

We shall establish polynomial extensions from the triangle T to the prism G .

Since the mapping M :

$$(2.9) \quad x_1 = \tilde{x}_1(1 - H\tilde{x}_3), \quad x_2 = \tilde{x}_2(1 - H\tilde{x}_3), \quad x_3 = H\tilde{x}_3$$

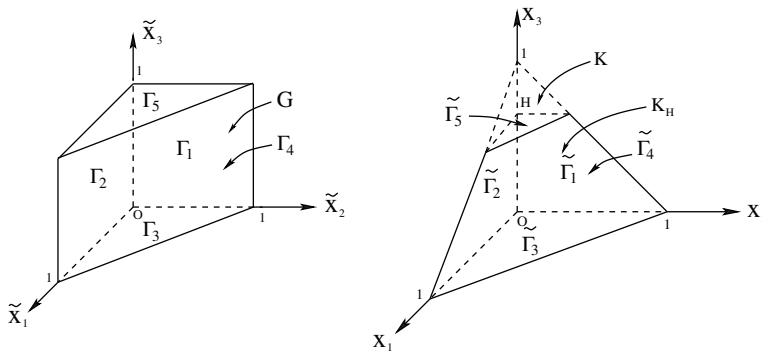


FIG. 2.2. The prism G and truncated tetrahedron K_H .

maps the prism G onto a truncated tetrahedron $K_H = \{(x_1, x_2, x_3) | x_1 \geq 0, x_2 \geq 0, H \geq x_3 \geq 0, x_1 + x_2 + x_3 \leq 1\}$, with $H \in (0, 1)$ shown in Figure 2.2. $\tilde{\Gamma}_i, i = 1, 2, 3, 4, 5$ are the faces of K_H , $\tilde{\Gamma}_3$ and $\tilde{\Gamma}_5$ are contained in the planes $x_3 = 0$ and $x_3 = H$, respectively, and $\tilde{\Gamma}_i, i = 1, 2, 4$ are portions of the faces of the tetrahedron K . Hence, we need to construct a polynomial extension operator $R_H : \mathcal{P}_p^{1,0}(T) \rightarrow \mathcal{P}_p^1(K_H) \oplus \mathcal{P}_p^1(T) \times \mathcal{P}_1(I_H)$ with desired properties, where $I_H = (0, H)$, which can lead to a polynomial extension from a triangular face to a whole prism.

We now introduce polynomial lifting operator R_H on K_H defined by

$$(2.10) \quad R_H f(x_1, x_2, x_3) = R_K f(x_1, x_2, x_3) - \frac{x_3}{H} R_K f(x_1, x_2, H),$$

where R_K is the lifting operator on K given in (2.2).

THEOREM 2.3. *Let R_H be the operator given in (2.10). Then, $R_H f(x) \in \mathcal{P}_p^1(K_H) \oplus \mathcal{P}_p^1(T) \times \mathcal{P}_1(I_H)$ for $f \in \mathcal{P}_p^{1,0}(T)$ such that $R_H f(x) |_{\tilde{\Gamma}_3} = f, R_H f |_{\tilde{\Gamma}_i} = 0, i = 1, 2, 4, 5$, and*

$$(2.11) \quad \|R_H f\|_{H^1(K_H)} \leq C \|f\|_{H_{00}^{\frac{1}{2}}(\tilde{\Gamma}_3)},$$

where $I_H = (0, H)$ and $T_H = \{(x_1, x_2) | x_1 \geq 0, x_2 \geq 0, x_1 + x_2 \leq 1 - H\}$ and C is a constant independent of f and p .

Combining the operator R_H and the mapping M , we construct an extension R_G^T by

$$(2.12) \quad R_G^T f(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) = R_H f \circ M = U(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) - \tilde{x}_3 U(\tilde{x}_1, \tilde{x}_2, 1),$$

where $U(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) = R_K f \circ M$. Suppose that $R_K f(x_1, x_2, x_3) = \sum_{i+j+k \leq p} a_{ijk} x_1^i x_2^j x_3^k$, then

$$\begin{aligned} R_K f \circ M(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) &= U(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) \\ &= \sum_{i+j+k \leq p} a_{ijk} H^k \tilde{x}_1^i \tilde{x}_2^j \tilde{x}_3^k (1 - H \tilde{x}_3)^{i+j} \in \mathcal{P}_p^1(T) \times \mathcal{P}_p(I) \end{aligned}$$

and

$$\frac{x_3}{H} R_K f(x_1, x_2, H) \circ M = \tilde{x}_3 U(\tilde{x}_1, \tilde{x}_2, 1) \in \mathcal{P}_p^1(T) \times \mathcal{P}_1(I).$$

Therefore, $R_G^T f(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) = R_H f \circ M \in \mathcal{P}_p^{1,0}(T) \times \mathcal{P}_p(I)$ if $f \in \mathcal{P}_p^1(T)$. We are able to establish the polynomial extension from a triangular face to a prism.

THEOREM 2.4. *Let R_G^T be the extension defined in (2.12). Then, $R_G^T f \in \mathcal{P}_p^1(T) \times \mathcal{P}_p(I)$ for $f \in \mathcal{P}_p^{1,0}(T)$, $R_G^T f |_{\Gamma_3} = f$ and vanishes on $\partial G \setminus \Gamma_3$, and*

$$(2.13) \quad \|R_G^T f\|_{H^1(G)} \leq C \|f\|_{H_{00}^{\frac{1}{2}}(\Gamma_3)},$$

where C is a constant independent of f and p .

Proof. Obviously, $R_G^T : \mathcal{P}_p^{1,0}(T) \rightarrow \mathcal{P}_p^{1,0}(T) \times \mathcal{P}_p(I)$, and $R_G^T f_{\Gamma_3} = f$ for $f \in \mathcal{P}_p^{1,0}(T)$, $R_G^T f |_{\Gamma_i} = 0, i = 1, 2, 4, 5$. Since the mapping M is trilinear,

$$\|R_G^T f\|_{H^1(G)} \leq C \|R_H f\|_{H^1(K_H)}.$$

Then, (2.13) follows from (2.11) easily. \square

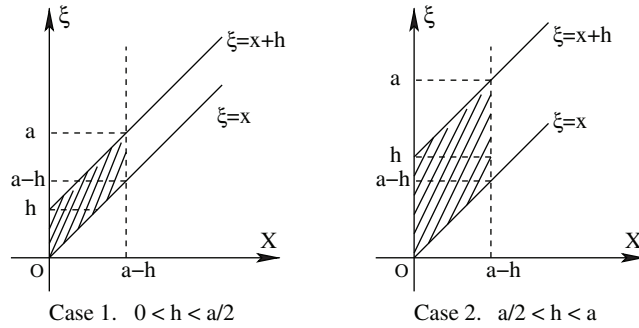


FIG. 2.3. Case 1 and Case 2.

It remained to prove Theorem 2.3. To this end, we need the following lemmas.

LEMMA 2.5. For $0 < h < a$ and any function $g \in L^2(0, a)$, it holds that

$$(2.14) \quad \int_0^{a-h} \left| \frac{1}{h} \int_x^{x+h} g(\xi) d\xi \right|^2 dx \leq \int_0^a |g(x)|^2 dx.$$

Also, there hold

$$(2.15) \quad \int_0^{a-h} \left| \frac{1}{h} \int_x^{x+h} g(\xi) d\xi \right|^2 dx \leq \frac{1}{h} \int_0^a x |g(x)|^2 dx$$

and

$$(2.16) \quad \int_0^{a-h} \left| \frac{1}{h} \int_x^{x+h} g(\xi) d\xi \right|^2 dx \leq \frac{1}{h} \int_0^a (a-x) |g(x)|^2 dx.$$

Proof. By Schwarz inequality, we have

$$\int_0^{a-h} \left| \frac{1}{h} \int_x^{x+h} g(\xi) d\xi \right|^2 dx \leq \int_0^{a-h} \left| \frac{1}{h} \int_x^{x+h} |g(\xi)| d\xi \right|^2 dx \leq \int_0^{a-h} dx \int_x^{x+h} \frac{|g(\xi)|^2}{h} d\xi.$$

Case 1 : $0 < h \leq a/2$ (see Figure 2.3). There holds

$$\begin{aligned} & \int_0^{a-h} \left| \frac{1}{h} \int_x^{x+h} g(\xi) d\xi \right|^2 dx \leq \int_0^{a-h} dx \int_x^{x+h} \frac{|g(\xi)|^2}{h} d\xi \\ &= \int_0^h d\xi \int_0^\xi \frac{|g(\xi)|^2}{h} dx + \int_h^{a-h} d\xi \int_{\xi-h}^\xi \frac{|g(\xi)|^2}{h} dx + \int_{a-h}^a d\xi \int_{\xi-h}^{a-h} \frac{|g(\xi)|^2}{h} dx \\ &= \int_0^h \frac{\xi |g(\xi)|^2}{h} d\xi + \int_h^{a-h} \frac{h |g(\xi)|^2}{h} d\xi + \int_{a-h}^a \frac{(a-\xi) |g(\xi)|^2}{h} d\xi. \end{aligned}$$

Hence, we have

$$\int_0^{a-h} \left| \frac{1}{h} \int_x^{x+h} g(\xi) d\xi \right|^2 dx \leq \int_0^a |g(\xi)|^2 d\xi$$

and

$$\int_0^{a-h} \left| \frac{1}{h} \int_x^{x+h} g(\xi) d\xi \right|^2 dx \leq \frac{1}{h} \int_0^a \xi |g(\xi)|^2 d\xi.$$

Case 2 : $a/2 < h < a$ (see Figure 2.3). Similarly, there holds

$$\begin{aligned} & \int_0^{a-h} \left| \frac{1}{h} \int_x^{x+h} g(\xi) d\xi \right|^2 dx \leq \int_0^{a-h} dx \int_x^{x+h} \frac{|g(\xi)|^2}{h} d\xi \\ &= \int_0^{a-h} d\xi \int_0^\xi \frac{|g(\xi)|^2}{h} dx + \int_{a-h}^h d\xi \int_0^{a-h} \frac{|g(\xi)|^2}{h} dx + \int_h^a d\xi \int_{\xi-h}^{a-h} \frac{|g(\xi)|^2}{h} dx \\ &= \int_0^{a-h} \frac{\xi |g(\xi)|^2}{h} d\xi + \int_{a-h}^h \frac{(a-h) |g(\xi)|^2}{h} d\xi + \int_h^a \frac{(a-\xi) |g(\xi)|^2}{h} d\xi, \end{aligned}$$

which implies

$$\int_0^{a-h} \left| \frac{1}{h} \int_x^{x+h} g(\xi) d\xi \right|^2 dx \leq \int_0^a |g(\xi)|^2 d\xi$$

and

$$\int_0^{a-h} \left| \frac{1}{h} \int_x^{x+h} g(\xi) d\xi \right|^2 dx \leq \frac{1}{h} \int_0^a \xi |g(\xi)|^2 d\xi.$$

Therefore, we always have (2.14) and (2.15) for $0 < h \leq a/2$ or $a/2 < h < a$.

Letting $\eta = a - \xi$ and $\hat{x} = a - h - x$ and using (2.15), we obtain

$$\begin{aligned} & \int_0^{a-h} \left| \frac{1}{h} \int_x^{x+h} g(\xi) d\xi \right|^2 dx = \int_0^{a-h} \left| \frac{1}{h} \int_{\hat{x}}^{\hat{x}+h} g(a-\eta) d\eta \right|^2 d\hat{x} \\ & \leq \frac{1}{h} \int_0^a \hat{x} |g(a-\hat{x})|^2 d\hat{x} = \frac{1}{h} \int_0^a (a-z) |g(z)|^2 dz, \end{aligned}$$

which yields (2.16). \square

LEMMA 2.6. Let $R_{12}(x_1, x_2, H)$ and $R_{i3}(x_1, x_2, H)$ be the operators given in (2.4) and (2.5), with $x_3 = H$. Then

$$(2.17) \quad \|R_{12}f(x_1, x_2, H)\|_{L^2(K_H)} \leq C \left\| (x_1 x_2)^{\frac{1}{2}} f(x_1, x_2) \right\|_{L^2(T)}$$

and for $i = 1, 2$,

$$(2.18) \quad \|R_{i3}f(x_1, x_2, H)\|_{L^2(K_H)} \leq C \left\| x_i^{\frac{1}{2}} (1-x_1-x_2)^{\frac{1}{2}} f(x_1, x_2) \right\|_{L^2(T)},$$

where C is a constant independent of f .

Proof. Note that

$$\|R_{12}f(x_1, x_2, H)\|_{L^2(K_H)}^2 \leq \frac{4}{H^2} \int_0^H dx_3 \int_0^{1-x_3} dx_2 \int_0^{1-x_2-x_3} \left| \frac{1}{H} \int_{x_1}^{x_1+H} g_1(\xi_1) d\xi_1 \right|^2 dx_1,$$

with $g_1(\xi_1) = \int_{x_2}^{x_2+H} |\tilde{f}(\xi_1, \xi_2)| d\xi_2$. Hereafter, \tilde{f} denotes the extension of f by zero outside T . We apply here Lemma 2.5 to $g_1(\xi_1)$ with $a = 1 - x_2 - x_3, h = H, x = x_1, \xi = \xi_1$. Then we get

$$\int_0^{1-x_2-x_3} \left(\frac{1}{H} \int_{x_1}^{x_1+H} g_1(\xi_1) d\xi_1 \right)^2 dx_1 \leq \frac{1}{H} \int_0^{1-x_2-x_3+H} x_1 |g_1(x_1)|^2 dx_1,$$

which implies

$$\begin{aligned} (2.19) \quad & \int_0^{1-x_3} dx_2 \int_0^{1-x_2-x_3} \left| \frac{1}{H} \int_{x_1}^{x_1+H} g_1(\xi_1) d\xi_1 \right|^2 dx_1 \\ & \leq \frac{1}{H} \int_0^{1-x_3} dx_2 \int_0^{1-x_2-x_3+H} x_1 \left| \int_{x_2}^{x_2+H} |\tilde{f}(x_1, \xi_2)| d\xi_2 \right|^2 dx_1 \\ & = H \left\{ \int_0^H x_1 dx_1 \int_0^{1-x_3} \left| \frac{1}{H} \int_{x_2}^{x_2+H} |\tilde{f}(x_1, \xi_2)| d\xi_2 \right|^2 dx_2 \right. \\ & \quad \left. + \int_H^{1-x_3+H} x_1 dx_1 \int_0^{1-x_1-x_3+H} \left| \frac{1}{H} \int_{x_2}^{x_2+H} |\tilde{f}(x_1, \xi_2)| d\xi_2 \right|^2 dx_2 \right\}. \end{aligned}$$

Applying Lemma 2.5 again, we have

$$\int_0^{1-x_3} \left| \frac{1}{H} \int_{x_2}^{x_2+H} |\tilde{f}(x_1, \xi_2)| d\xi_2 \right|^2 dx_2 \leq \frac{1}{H} \int_0^{1-x_3+H} x_2 |\tilde{f}(x_1, x_2)|^2 dx_2$$

and

$$\int_0^{1-x_1-x_3+H} \left| \frac{1}{H} \int_{x_2}^{x_2+H} |\tilde{f}(x_1, \xi_2)| d\xi_2 \right|^2 dx_2 \leq \frac{1}{H} \int_0^{1-x_1-x_3+2H} x_2 |\tilde{f}(x_1, x_2)|^2 dx_2,$$

which together with (2.19) yields

$$\begin{aligned} & \int_0^{1-x_3} dx_2 \int_0^{1-x_2-x_3} \left(\frac{1}{H} \int_{x_1}^{x_1+H} d\xi_1 \int_{x_2}^{x_2+H} |\tilde{f}(\xi_1, \xi_2)| d\xi_2 \right)^2 dx_1 \\ & \leq \left(\int_0^H dx_1 \int_0^{1+H} + \int_H^{1-x_3+H} dx_1 \int_0^{1-x_1+2H} \right) x_2 x_1 |\tilde{f}(x_1, x_2)|^2 dx_2 \\ & \leq \left(\int_0^H dx_1 \int_0^{1+H} + \int_H^{1+H} dx_1 \int_0^{1-x_1+2H} \right) x_2 x_1 |\tilde{f}(x_1, x_2)|^2 dx_2 \leq 2 \left\| (x_1 x_2)^{\frac{1}{2}} f \right\|_{L^2(T)}^2. \end{aligned}$$

Therefore, (2.17) follows immediately.

Let Q_1 be the mapping

$$(2.20) \quad x_1 = \hat{x}_2, \quad x_2 = 1 - \hat{x}_1 - \hat{x}_2 - \hat{x}_3, \quad x_3 = \hat{x}_3,$$

which maps K_H onto itself, and let W_1 be the mapping

$$(2.21) \quad \xi_1 = \hat{\xi}_2, \quad \xi_2 = 1 - \hat{\xi}_1 - \hat{\xi}_2,$$

which maps T onto itself. Then $\hat{f}(\hat{\xi}_1, \hat{\xi}_2) = f(\xi_1, \xi_2) \circ W_1 = f(\hat{\xi}_2, 1 - \hat{\xi}_1 - \hat{\xi}_2)$ and $R_{12}f(\hat{x}_1, \hat{x}_2, H) = R_{13}f(x_1, x_2, x_3) \circ Q_1|_{x_3=H}$. Therefore,

$$\begin{aligned} \|R_{13}f(x_1, x_2, H)\|_{L^2(K_H)} &\leq \|R_{12}\hat{f}(\hat{x}_1, \hat{x}_2, H)\|_{L^2(K_H)} \leq C \left\| \left(\hat{\xi}_1\hat{\xi}_2\right)^{\frac{1}{2}} \hat{f} \right\|_{L^2(T)} \\ &\leq C \left\| \xi_1^{\frac{1}{2}}(1 - \xi_1 - \xi_2)^{\frac{1}{2}} f \right\|_{L^2(T)}. \end{aligned}$$

For $R_{23}f$, we introduce mapping Q_2 and W_2 :

$$(2.22) \quad Q_2 : \quad x_1 = 1 - \hat{x}_1 - \hat{x}_2 - \hat{x}_3, \quad x_2 = \hat{x}_1, \quad x_3 = \hat{x}_3,$$

which maps K_H onto itself, and

$$(2.23) \quad W_2 : \quad \xi_1 = 1 - \hat{\xi}_1 - \hat{\xi}_2, \quad \xi_2 = \hat{\xi}_1,$$

which maps T onto itself. Similarly, there holds

$$\begin{aligned} \|R_{23}f(x_1, x_2, H)\|_{L^2(K_H)} &\leq \|R_{12}\hat{f}(\hat{x}_1, \hat{x}_2, H)\|_{L^2(K_H)} \leq C \left\| \left(\hat{\xi}_1\hat{\xi}_2\right)^{\frac{1}{2}} \hat{f} \right\|_{L^2(T)} \\ &\leq C \left\| \xi_2^{\frac{1}{2}}(1 - \xi_1 - \xi_2)^{\frac{1}{2}} f \right\|_{L^2(T)}. \quad \square \end{aligned}$$

LEMMA 2.7. *Let $R_{12}(x_1, x_2, H)$ and $R_{i3}(x_1, x_2, H)$ be the operators given in (2.4) and (2.5), with $x_3 = H$. Then for $i = 1, 2$,*

$$(2.24) \quad \left\| \frac{\partial R_{12}f(x_1, x_2, H)}{\partial x_i} \right\|_{L^2(K_H)} \leq C \left\| x_i^{-\frac{1}{2}} f \right\|_{L^2(T)},$$

and $t = 1, 2$

$$(2.25) \quad \left\| \frac{\partial R_{i3}f(x_1, x_2, H)}{\partial x_t} \right\|_{L^2(K_H)} \leq C \left(\left\| x_t^{-\frac{1}{2}} f \right\|_{L^2(T)} + \left\| (1 - x_1 - x_2)^{-\frac{1}{2}} f \right\|_{L^2(T)} \right),$$

where C is a constant independent of f .

Proof. Note that

$$\begin{aligned} \frac{\partial R_{12}f(x_1, x_2, H)}{\partial x_1} &= \frac{2x_2}{H^2} \int_{x_1}^{x_1+H} d\xi_1 \int_{x_2}^{x_1+x_2+H-\xi_1} \frac{f(\xi_1, \xi_2)}{\xi_1\xi_2} d\xi_2 \\ &- \frac{2x_2}{H^2} \int_{x_2}^{x_2+H} \frac{f(x_1, \xi_2)}{\xi_2} d\xi_2 + \frac{2x_1x_2}{H^2} \int_{x_1}^{x_1+H} \frac{f(\xi_1, x_1+x_2+H-\xi_1)}{\xi_1(x_1+x_2+H-\xi_1)} d\xi_1 \end{aligned}$$

and

$$(2.26) \quad \left| \frac{\partial R_{12}f(x_1, x_2, H)}{\partial x_1} \right| \leq I_1 + I_2 + I_3,$$

where

$$\begin{aligned} I_1 &= \frac{2}{H^2} \int_{x_1}^{x_1+H} d\xi_1 \int_{x_2}^{x_2+H} \frac{|f(\xi_1, \xi_2)|}{\xi_1} d\xi_2, \quad I_2 = \frac{2}{H^2} \int_{x_2}^{x_2+H} |f(x_1, \xi_2)| d\xi_2, \\ I_3 &= \frac{2}{H^2} \int_{x_1}^{x_1+H} |f(\xi_1, x_1+x_2+H-\xi_1)| d\xi_1. \end{aligned}$$

Note that

$$\|I_1\|_{L^2(K_H)}^2 = \frac{4}{H^2} \int_0^H dx_3 \int_0^{1-x_3} dx_2 \int_0^{1-x_2-x_3} \left(\frac{1}{H} \int_{x_1}^{x_1+H} g_1(\xi_1) d\xi_1 \right)^2 dx_1,$$

with $g_1(\xi_1) = \int_{x_2}^{x_2+H} \frac{|\tilde{f}(\xi_1, \xi_2)|}{\xi_1} d\xi_2$. Applying Lemma 2.5 to $g_1(\xi_1)$ with $a = 1 - x_2 - x_3, h = H, x = x_1, \xi = \xi_1$, we have

$$\int_0^{1-x_2-x_3} \left| \frac{1}{H} \int_{x_1}^{x_1+H} g_1(\xi_1) d\xi_1 \right|^2 dx_1 \leq \frac{1}{H} \int_0^{1-x_2-x_3+H} x_1 \left| \int_{x_2}^{x_2+H} \frac{\tilde{f}(x_1, \xi_2)}{x_1} d\xi_2 \right|^2 dx_1,$$

which implies

$$\begin{aligned} & \int_0^{1-x_3} dx_2 \int_0^{1-x_2-x_3} \left| \frac{1}{H} \int_{x_1}^{x_1+H} g_1(\xi_1) d\xi_1 \right|^2 dx_1 \\ & \leq \frac{1}{H} \int_0^{1-x_3} dx_2 \int_0^{1-x_2-x_3+H} \frac{1}{x_1} \left| \int_{x_2}^{x_2+H} \tilde{f}(x_1, \xi_2) d\xi_2 \right|^2 dx_1 \\ & \leq H \left\{ \int_0^H \frac{1}{x_1} dx_1 \int_0^{1-x_3} \left| \frac{1}{H} \int_{x_2}^{x_2+H} \tilde{f}(x_1, \xi_2) d\xi_2 \right|^2 dx_2 \right. \\ & \quad \left. + \int_H^{1-x_3+H} \frac{1}{x_1} dx_1 \int_0^{1-x_1-x_3+H} \left| \frac{1}{H} \int_{x_2}^{x_2+H} \tilde{f}(x_1, \xi_2) d\xi_2 \right|^2 dx_2 \right\}. \end{aligned}$$

Applying Lemma 2.5 again to the function $g_2(\xi_2) = \tilde{f}(x_1, \xi_2)$, we have

$$\begin{aligned} (2.27) \quad \|I_1\|_{L^2(K_H)}^2 & \leq \frac{4}{H^2} \int_0^H dx_3 \int_0^H \frac{1}{x_1} dx_1 \int_0^{1-x_3+H} |\tilde{f}(x_1, x_2)|^2 dx_2 \\ & \quad + \frac{4}{H^2} \int_0^H dx_3 \int_H^{1-x_3+H} \frac{1}{x_1} dx_1 \int_0^{1-x_1-x_3+2H} |\tilde{f}(x_1, x_2)|^2 dx_2 \\ & \leq \frac{4}{H} \left(\int_0^H dx_1 \int_0^{1+H} + \int_H^{1+H} dx_1 \int_0^{1-x_1+2H} \right) \frac{|\tilde{f}(x_1, x_2)|^2}{x_1} dx_2 \\ & \leq \frac{8}{H} \|x_1^{-\frac{1}{2}} f\|_{L^2(T)}^2. \end{aligned}$$

Similarly, we have by Lemma 2.5,

$$\begin{aligned} (2.28) \quad \|I_2\|_{L^2(K_H)}^2 & = \frac{4}{H^2} \int_0^H dx_3 \int_0^{1-x_3} dx_1 \int_0^{1-x_1-x_3} \left| \frac{1}{H} \int_{x_2}^{x_2+H} |f(x_1, \xi_2)| d\xi_2 \right|^2 dx_2 \\ & \leq \frac{4}{H^3} \int_0^H dx_3 \int_0^{1-x_3} dx_1 \int_0^{1-x_1-x_3+H} x_2 |f(x_1, x_2)|^2 dx_2 \\ & \leq \frac{4}{H^3} \int_0^H dx_3 \int_0^1 dx_1 \int_0^{1-x_1+H} x_2 |\tilde{f}(x_1, x_2)|^2 dx_2 \\ & = \frac{4}{H^2} \|x_2^{\frac{1}{2}} f\|_{L^2(T)}^2 \end{aligned}$$

and

$$\begin{aligned} & \|I_3\|_{L^2(K_H)}^2 \\ &= \frac{4}{H^2} \int_0^H dx_3 \int_0^{1-x_3} dx_2 \int_0^{1-x_2-x_3} \left| \frac{1}{H} \int_{x_1}^{x_1+H} \tilde{f}(\xi_1, x_1 + x_2 + H - \xi_1) d\xi_1 \right|^2 dx_1 \\ &\leq \frac{4}{H^3} \int_0^H dx_3 \int_0^{1-x_3} dx_2 \int_0^{1-x_2-x_3+H} x_1 \left| \tilde{f}(x_1, x_2 + H) \right|^2 dx_1 \\ &\leq \frac{4}{H^3} \int_0^H dx_3 \int_0^1 dx_2 \int_0^{1-x_2+H} x_1 \left| \tilde{f}(x_1, x_2 + H) \right|^2 dx_1 \\ &= \frac{4}{H^2} \int_0^1 dx_2 \int_0^{1-x_2+H} x_1 \left| \tilde{f}(x_1, x_2 + H) \right|^2 dx_1. \end{aligned}$$

Letting $z = x_2 + H$, we have

$$\begin{aligned} & \frac{4}{H^2} \int_0^1 dx_2 \int_0^{1-x_2+H} x_1 \left| \tilde{f}(x_1, x_2 + H) \right|^2 dx_1 \\ &= \frac{4}{H^2} \int_H^{1+H} dz \int_0^{1-z+2H} x_1 \left| \tilde{f}(x_1, z) \right|^2 dx_1 \\ &= \frac{4}{H^2} \int_H^1 dz \int_0^{1-z} x_1 \left| \tilde{f}(x_1, z) \right|^2 dx_1 \leq \frac{4}{H^2} \left\| x_1^{\frac{1}{2}} f \right\|_{L^2(T)}^2, \end{aligned}$$

which implies

$$(2.29) \quad \|I_3\|_{L^2(K_H)}^2 \leq \frac{4}{H^2} \left\| x_1^{\frac{1}{2}} f \right\|_{L^2(T)}^2.$$

Combining (2.26)–(2.29), we have

$$\left\| \frac{\partial R_{12}f(x_1, x_2, H)}{\partial x_1} \right\|_{L^2(K_H)} \leq C \left\| x_1^{-\frac{1}{2}} f \right\|_{L^2(T)}.$$

Similarly, we can prove

$$\left\| \frac{\partial R_{12}f(x_1, x_2, H)}{\partial x_2} \right\|_{L^2(K_H)} \leq C \left\| x_2^{-\frac{1}{2}} f \right\|_{L^2(T)}.$$

Let Q_i and W_i ($i=1,2$) be the mapping as defined in (2.20)–(2.23). Then, for $t = 1, 2$,

$$\begin{aligned} \left\| \frac{\partial R_{13}f(x_1, x_2, H)}{\partial x_t} \right\|_{L^2(K_H)} &\leq \sum_{i=1,2} \left\| \frac{\partial R_{12}\hat{f}(\hat{x}_1, \hat{x}_2, H)}{\partial \hat{x}_i} \right\|_{L^2(K_H)} \leq C \sum_{i=1,2} \left\| \hat{\xi}_i^{-\frac{1}{2}} \hat{f} \right\|_{L^2(T)} \\ &\leq C \left(\left\| \xi_1^{-\frac{1}{2}} f \right\|_{L^2(T)} + \left\| (1 - \xi_1 - \xi_2)^{-\frac{1}{2}} f \right\|_{L^2(T)} \right). \end{aligned}$$

Similarly, we have for $t = 1, 2$,

$$\begin{aligned} \left\| \frac{\partial R_{23}f(x_1, x_2, H)}{\partial x_t} \right\|_{L^2(K_H)} &\leq \sum_{i=1,2} \left\| \frac{\partial R_{12}\hat{f}(\hat{x}_1, \hat{x}_2, H)}{\partial \hat{x}_i} \right\|_{L^2(K_H)} \leq C \sum_{i=1,2} \left\| \hat{\xi}_i^{-\frac{1}{2}} \hat{f} \right\|_{L^2(T)} \\ &\leq C \left(\left\| \xi_2^{-\frac{1}{2}} f \right\|_{L^2(T)} + \left\| (1 - \xi_1 - \xi_2)^{-\frac{1}{2}} f \right\|_{L^2(T)} \right). \quad \square \end{aligned}$$

Proof of Theorem 2.3. Obviously, $R_H f(x) \in \mathcal{P}_p^1(K_H) \oplus \mathcal{P}_p^{1,0}(T) \times \mathcal{P}_1(I_H)$ for $f \in \mathcal{P}_p^{1,0}(T)$. Due to (2.10), we have

$$(2.30) \quad \begin{aligned} \|R_H f(x_1, x_2, x_3)\|_{H^1(K_H)} &\leq \|R_K f(x_1, x_2, x_3)\|_{H^1(K_H)} \\ &\quad + \left\| \frac{x_3}{H} R_K f(x_1, x_2, H) \right\|_{H^1(K_H)}. \end{aligned}$$

By Theorem 2.1, there holds

$$(2.31) \quad \|R_K f(x_1, x_2, x_3)\|_{H^1(K_H)} \leq \|R_K f(x_1, x_2, x_3)\|_{H^1(K)} \leq C \|f(x_1, x_2)\|_{H_{00}^{\frac{1}{2}}(T)}$$

and by (2.3) and Lemma 2.6–Lemma 2.7, it holds that

$$\begin{aligned} &\left\| \frac{x_3}{H} R_K f(x_1, x_2, H) \right\|_{H^1(K_H)} \\ &\leq C \left(\|R_{12} f(x_1, x_2, H)\|_{H^1(K_H)} + \sum_{i=1,2} \|R_{i3} f(x_1, x_2, H)\|_{H^1(K_H)} \right) \\ &\leq C \left(\|f\|_{H^{\frac{1}{2}}(T)} + \sum_{i=1,2} \|x_i^{-\frac{1}{2}} f\|_{L^2(T)} + \|(1-x_1-x_2)^{-\frac{1}{2}} f\|_{L^2(T)} \right) \leq C \|f\|_{H_{00}^{\frac{1}{2}}(T)}, \end{aligned}$$

which together with (2.30)–(2.31) leads to (2.11) immediately. \square

2.3. Polynomial extension on prisms from a square face. We shall construct a polynomial extension on prisms from a square face $S = \{x = (x_1, x_2, x_3) \mid 0 \leq x_1, x_3 \leq 1\}$, which is as important as the extension from a triangular face for error analysis and preconditioning of high-order FEM in three dimensions [15, 18].

LEMMA 2.8. *Let $T = \{(x_1, x_2) \mid 0 < x_2 < 1 - x_1, 0 \leq x_1 < 1\}$ be the standard triangle and $I = (0, 1)$. Then there is a polynomial extension operator $R_T^* : H_0^1(I) \rightarrow H^1(T)$ such that $R_T^* f \in \mathcal{P}_p^1(T)$ if $f(x_1) \in \mathcal{P}_p^0(I)$, and*

$$(2.32) \quad R_T^* f|_I = f(x_1), \quad R_T^* f|_{\partial T \setminus I} = 0,$$

$$(2.33) \quad \|R_T^* f\|_{H^t(T)} \leq C \left(p^{t-\frac{3}{2}} \|f\|_{H^1(I)} + p^{t-\frac{1}{2}} \|f\|_{L^2(I)} \right), \quad t = 0, 1,$$

with C independent of f and p .

Proof. Let $\psi(x_2) = (1 - x_2)^p$. Then for $t \geq 0$,

$$(2.34) \quad \|\psi\|_{H^t(I)} \leq C p^{t-\frac{1}{2}}.$$

We introduce a function $\Psi \in \mathcal{P}_{2p+1}^1(T)$ by

$$\Psi(x_1, x_2) = \psi(x_2)((1 - x_1 - x_2)f(x_1) + x_1 f(x_1 + x_2)).$$

Then $\Psi(x_1, 0) = f(x_1)$, $\Psi(1, x_2) = \Psi(x_1, 1 - x_1) = 0$, and

$$(2.35) \quad \|\Psi\|_{L^2(T)} \leq C p^{-\frac{1}{2}} \|f\|_{L^2(I)},$$

$$(2.36) \quad \|\Psi\|_{H^1(T)} \leq C \left(p^{-\frac{1}{2}} \|f\|_{H^1(I)} + p^{\frac{1}{2}} \|f\|_{L^2(I)} \right).$$

By the lifting theorem on the triangle T [17], there exists a lifting operator $R_T : H_{00}^{\frac{1}{2}}(I) \rightarrow H^1(T)$

$$R_T f = \frac{x_1(1 - x_1 - x_2)}{x_2^2} \int_{x_1}^{x_1+x_2} \frac{f(\xi)}{\xi(1-\xi)} d\xi$$

such that $R_T f \in \mathcal{P}_p^1(T)$, $R_T f|_I = f$, $R_T f|_{\partial T \setminus I} = 0$, and

$$\|R_T f\|_{H^1(T)} \leq C \|f\|_{H_{00}^{\frac{1}{2}}(I)},$$

which implies that R_T satisfies (2.33) with $t = 1$. Unfortunately, the extension does not give precise information on $\|R_T f\|_{L^2(T)}$, and the desired estimation (2.33) with $t = 0$ may not be true for R_T . Therefore, we have to construct a new extension operator R_T^* .

Note that $\Psi - R_T f = 0$ on ∂T . By Π_T , we denote the orthogonal projection operator $H_0^1(T) \rightarrow \mathcal{P}_p^{1,0}(T)$, and let

$$w_p = R_T f + \Pi_T(\Psi - R_T f).$$

Then $w_p(x_1, 0) = f(x_1)$, $w_p(1, x_2) = w_p(x_1, 1 - x_1) = 0$, and

$$(2.37) \quad \Psi - w_p = (I - \Pi_T)(\Psi - R_T f).$$

Due to the continuity of operator R_T and a trace theorem, we obtain

$$(2.38) \quad \begin{aligned} \|w_p\|_{H^1(T)} &\leq \|\Psi\|_{H^1(T)} + \|\Psi - w_p\|_{H^1(T)} \leq \|\Psi\|_{H^1(T)} + \|\Psi - R_T f\|_{H^1(T)} \\ &\leq 2\|\Psi\|_{H^1(T)} + \|R_T f\|_{H^1(T)} \leq C \left(\|\Psi\|_{H^1(T)} + \|f\|_{H_{00}^{\frac{1}{2}}(I)} \right) \\ &\leq C \left(\|\Psi\|_{H^1(T)} + \|\Psi\|_{H^{\frac{1}{2}}(\partial T)} \right) \leq C \|\Psi\|_{H^1(T)}. \end{aligned}$$

Let $R_T^* f = w_p$. Then (2.36) and (2.38) lead to (2.32) and (2.33) with $t = 1$. Note that $\Pi_T(\Psi - R_T f)$ is the finite element solution in $\mathcal{P}_p^{1,0}(T)$ for the the boundary value problem

$$\begin{aligned} -\Delta u + u &= \tilde{f} && \text{in } T \\ u|_{\partial T} &= 0, \end{aligned}$$

with $\tilde{f} = -\Delta(\Psi - R_T f) + \Psi - R_T f$. By the Nitsche's trick, we have

$$\|(I - \Pi_T)(\Psi - R_T f)\|_{L^2(T)} \leq Cp^{-1} \|(I - \Pi_T)(\Psi - R_T f)\|_{H^1(T)} \leq Cp^{-1} \|\Psi\|_{H^1(T)},$$

which implies

$$(2.39) \quad \|\Psi - w_p\|_{L^2(T)} = \|(I - \Pi_T)(\Psi - R_T f)\|_{L^2(T)} \leq Cp^{-1} \|\Psi\|_{H^1(T)}.$$

Combining (2.39) and (2.36) we have (2.33) for $t = 0$. \square

We construct a polynomial extension from a square face to the prism G with help of the extension R_T^* in triangle T :

$$(2.40) \quad R_G^S f(x_1, x_2, x_3) = R_T^* f(\cdot, x_3).$$

THEOREM 2.9. *Let $\Gamma_2 = S$ be a square face of the prism G as shown in Figure 2.2, and let R_G^S be the extension operator defined as in (2.40). Then, $R_G^S f \in \mathcal{P}_p^1(T) \times \mathcal{P}_p(I)$ for $f \in \mathcal{P}_p^{2,0}(\Gamma_2)$, and*

$$(2.41) \quad R_G^S f = f \text{ on } \Gamma_2, \quad R_G^S f = 0 \text{ on } \partial G \setminus \Gamma_2,$$

$$(2.42) \quad \|R_G^S f\|_{H^1(G)} \leq C \left(p^{-\frac{3}{2}} \|f_{x_3}\|_{H^1(\Gamma_2)} + p^{-\frac{1}{2}} \|f\|_{H^1(\Gamma_2)} + p^{\frac{1}{2}} \|f\|_{L^2(\Gamma_2)} \right),$$

$$(2.43) \quad \|R_G^S f\|_{L^2(G)} \leq C \left(p^{-\frac{3}{2}} \|f\|_{H^1(\Gamma_2)} + p^{-\frac{1}{2}} \|f\|_{L^2(\Gamma_2)} \right).$$

Proof. Obviously, $R_G^S f \in \mathcal{P}_p^1(T) \times \mathcal{P}_p(I)$ and (2.41) holds. Due to (2.40),

$$\begin{aligned} \|R_G^S f\|_{L^2(G)}^2 &= \int_0^1 \left(\int_T |R_G^S f|^2 dx_1 dx_2 \right) dx_3 \leq \int_0^1 \|R_T^* f\|_{L^2(T)}^2 dx_3 \\ &\leq C \int_0^1 \left(p^{-3} \|f(\cdot, x_3)\|_{H^1(I)}^2 + p^{-1} \|f(\cdot, x_3)\|_{L^2(I)}^2 \right) dx_3 \\ &\leq C \left(p^{-3} \|f\|_{H^1(S)}^2 + p^{-1} \|f\|_{L^2(S)}^2 \right), \end{aligned}$$

which leads to (2.43).

Applying (2.40) to $f(x_1, x_3)$ and $f_{x_3}(x_1, x_3)$, respectively, we have

$$\begin{aligned} |R_G^S f|_{H^1(G)}^2 &\leq \int_0^1 \left(|R_T^* f|_{H^1(T)}^2 + |R_T^* f_{x_3}|_{L^2(T)}^2 \right) dx_3 \\ &\leq C \int_0^1 \left(p^{-1} \|f\|_{H^1(I)}^2 + p \|f\|_{L^2(I)}^2 + p^{-3} \|f_{x_3}\|_{L^2(I)}^2 \right) dx_3, \end{aligned}$$

which implies (2.42). \square

Remark 2.1. It is an open problem whether there exists a polynomial extension operator R_G^S such that

$$(2.44) \quad \|R_G^S f\|_{H^1(G)} \leq C \|f\|_{H_0^{1/2}(S)}.$$

Although (2.42) is not strong as the desired stability of (2.44), it gives the dependence of $\|R_G^S f\|_{H^1(G)}$ on $\|f\|_{H^t(S)}$, $t = 1, 0$ and $\|f_{x_3}\|_{H^1(S)}$ furnished precisely with weights $p^{-1/2}$, $p^{1/2}$, and $p^{-3/2}$, respectively. This estimation is sufficient while we apply the extension to a pair of elements sharing a common square face for constructing a continuous piecewise polynomial in $\mathcal{P}_p^{1.5}(G)$ without degrading the best order of approximation error. Hence, the extension R_G^S defined as in (2.40) is weakly stable, and Theorem 2.9 plays an important role in error analysis for the p and h - p versions of the FEM in three dimensions on meshes containing triangular prism elements. For the detail of the application of this extension for the construction of a continuous piecewise polynomial, we refer to [15, 18].

3. Polynomial extension on a cube. Let D be a cube and $\Gamma_i, i = 1, 2, \dots, 6$ be faces of D shown in Figure 3.1, and let $\gamma_{ij} = \Gamma_i \cap \Gamma_j, i = 1, 2, \dots, 6$. As usual, $I = [-1, 1]$ and $S = [-1, 1]^2$.

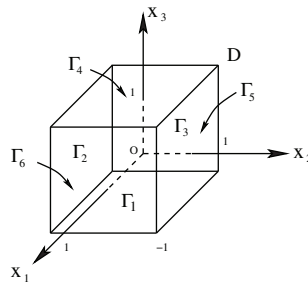


FIG. 3.1. A cube D .

3.1. Polynomial extension from a face. Let $J_j^{\alpha,\beta}(x)$ be the Jacobi polynomial of degree j :

$$(3.1) \quad J_j^{\alpha,\beta}(x) = \frac{(-1)^j(1-x)^{-\alpha}(1+x)^{-\beta}}{2^j j!} \frac{d^j(1-x)^{j+\alpha}(1+x)^{j+\beta}}{dx^j}, \quad j \geq 0,$$

with weights $\alpha, \beta > -1$, and let

$$(3.2) \quad \varphi_i(x) = \frac{1-x^2}{\sqrt{\gamma_{i-1}^{2,2}}} J_{i-1}^{2,2}(x), \quad i = 1, 2, 3, \dots,$$

where $\gamma_{i-1}^{2,2} = \frac{2^5 i(i+1)}{(2i+3)(i+2)(i+3)}$.

PROPOSITION 3.1. $\varphi_i(x), i = 1, 2, \dots, p-1$ form an orthogonal basis of $\mathcal{P}_p^0(I)$,

$$(3.3) \quad \langle \varphi_i(x), \varphi_j(x) \rangle_{L^2(I)} = \delta_{ij}, \quad 1 \leq i, j \leq p-1.$$

Proof. Due to the orthonormality of Jacobi polynomials,

$$\langle \varphi_i(x), \varphi_j(x) \rangle_{L^2(I)} = \frac{1}{\sqrt{\gamma_{i-1}^{2,2}} \sqrt{\gamma_{j-1}^{2,2}}} \int_I (1-x^2)^2 J_{i-1}^{2,2}(x) J_{j-1}^{2,2}(x) dx = \delta_{ij}. \quad \square$$

We introduce

$$(3.4) \quad \varphi_n(x_1, x_2) = \varphi_i(x_1) \varphi_j(x_2) = \frac{(1-x_1^2)(1-x_2^2)}{\sqrt{\gamma_{i-1}^{2,2}} \sqrt{\gamma_{j-1}^{2,2}}} J_{i-1}^{2,2}(x_1) J_{j-1}^{2,2}(x_2), \quad 1 \leq i, j \leq p-1,$$

with $n = (p-1)(i-1) + j$.

PROPOSITION 3.2. $\{\varphi_n(x_1, x_2), n = 1, 2, \dots, (p-1)^2\}$ forms an orthonormal basis of $\mathcal{P}_p^{2,0}(S)$ in $L^2(S)$, i.e.,

$$(3.5) \quad \langle \varphi_n, \varphi_m \rangle_{L^2(S)} = \delta_{nm}, \quad 1 \leq n, m \leq N_p = (p-1)^2.$$

Proof. Let $n = (p-1)(i-1) + j$ and $m = (p-1)(i'-1) + j'$. Then

$$\begin{aligned} & \langle \varphi_n, \varphi_m \rangle_{L^2(S)} \\ &= \int_I \frac{(1-x_1^2)^2}{\sqrt{\gamma_{i-1}^{2,2}} \sqrt{\gamma_{i'-1}^{2,2}}} J_{i-1}^{2,2}(x_1) J_{i'-1}^{2,2}(x_1) dx_1 \int_I \frac{(1-x_2^2)^2}{\sqrt{\gamma_{j-1}^{2,2}} \sqrt{\gamma_{j'-1}^{2,2}}} J_{j-1}^{2,2}(x_2) J_{j'-1}^{2,2}(x_2) dx_2 \\ &= \delta_{i,i'} \delta_{j,j'} = \delta_{nm}. \quad \square \end{aligned}$$

We consider an eigenvalue problem

$$(3.6) \quad \begin{cases} -\Delta u = \lambda u & \text{in } S = (-1, 1)^2, \\ u|_{\Gamma} = 0, \end{cases}$$

and its spectral solution (λ_p, ψ_p) , with $\psi_p \in \mathcal{P}_p^{2,0}(S)$, which satisfies

$$(3.7) \quad \int_S \nabla \psi_p \nabla q dx_1 dx_2 = \lambda_p \int_S \psi_p q dx_1 dx_2 \quad \forall q \in \mathcal{P}_p^{2,0}(S).$$

Selecting the basis $\{\varphi_n(x_1, x_2), n = 1, 2, \dots, N_p\}$ as in (3.4), with $N_p = (p - 1)^2$ and letting $\psi_p(x_1, x_2) = \sum_{i=1}^{N_p} c_i \varphi_i(x_1, x_2)$, we have the corresponding system of linear algebraic equations:

$$K \vec{C} = \lambda M \vec{C} = \lambda \vec{C},$$

where $\vec{C} = (c_1, c_2, \dots, c_{N_p})^T, K = (k_{ij})_{i,j=1}^{N_p}$, with $k_{ij} = \int_S \nabla \varphi_i \nabla \varphi_j dx_1 dx_2$. Here, we used the orthonormality of $\varphi_n(x_1, x_2)$ in $L^2(S)$, which implies the matrix $M = I$. Therefore, the spectral solution of eigenvalue problem (3.7) is equivalent to the eigenvalue problem of matrix K . Since K is symmetric and positive definite, the eigenvalues $\lambda_{p,k} > 0, k = 1, 2, \dots, N_p$ and the corresponding eigenvectors $\vec{C}^{(k)}$ are orthonormal, i.e.,

$$\langle \vec{C}^{(k)}, \vec{C}^{(l)} \rangle = \sum_{i=1}^{N_p} c_i^{(k)} c_i^{(l)} = \delta_{k,l}, 1 \leq k, l \leq N_p.$$

The corresponding eigen-polynomial $\psi_{p,k} = \sum_{n=1}^{N_p} c_n^{(k)} \varphi_n(x_1, x_2)$. Then, due to the properties of eigenvalues and vectors of K , we have the following theorem.

THEOREM 3.3. *The problem (3.7) has N_p real eigenvalues, and the corresponding eigen-polynomials $\{\psi_{p,k}(x_1, x_2), 1 \leq k \leq N_p\}$ are orthogonal in $L^2(S)$ and $H^1(S)$, which form an L^2 -orthonormal basis of $\mathcal{P}_p^{2,0}(S)$.*

Proof. The problem (3.7) has N_p real eigenvalues because the corresponding stiffness matrix K is positive definite and there hold for $1 \leq k, k' \leq N_p$

$$\langle \psi_{p,k}, \psi_{p,k'} \rangle_{L^2(S)} = \sum_{j=1}^{N_p} \sum_{i=1}^{N_p} c_i^{(k)} c_j^{(k')} \langle \varphi_i, \varphi_j \rangle_{L^2(S)} = \langle \vec{C}^{(k)}, \vec{C}^{(k')} \rangle = \delta_{k,k'}$$

and

$$\int_S \nabla \psi_{p,k} \nabla \psi_{p,k'} dx_1 dx_2 = \lambda_k \int_S \psi_{p,k} \psi_{p,k'} dx_1 dx_2 = \lambda_k \delta_{k,k'}.$$

Therefore, $\{\psi_{p,k}, k = 1, 2, \dots, N_p\}$ is orthogonal in $L^2(S)$ and $H^1(S)$ and forms an orthonormal basis in $L^2(S)$. \square

We next consider a two-point boundary value problem

$$(3.8) \quad \begin{cases} -v_{p,k}''(x_3) + \lambda_{p,k} v_{p,k}(x_3) = 0, & x_3 \in I = (-1, 1), \\ v_{p,k}(-1) = 1, \quad v_{p,k}(1) = 0, \end{cases}$$

and its spectral solution $\phi_{p,k} \in \mathcal{P}_p(I)$ such that $\phi_{p,k}(-1) = 1, \phi_{p,k}(1) = 0$ and

$$(3.9) \quad \int_I (\phi_{p,k}' q' + \lambda_{p,k} \phi_{p,k} q) dx_3 = 0,$$

which is equivalent to finding $\phi_{p,k} = \tilde{\phi}_{p,k} + \frac{1-x_3}{2}$, with $\tilde{\phi}_{p,k} \in P_p^0(I)$ satisfying

$$(3.10) \quad \begin{aligned} & \int_I (\tilde{\phi}_{p,k}'(x_3) q'(x_3) + \lambda_{p,k} \tilde{\phi}_{p,k}(x_3) q(x_3)) dx_3 \\ &= \frac{1}{2} \int_I (q'(x_3) - \lambda_{p,k} (1 - x_3) q(x_3)) dx_3. \end{aligned}$$

Since the corresponding bilinear form is coercive and continuous on $H_0^1(I) \times H_0^1(I)$, the solution $\tilde{\phi}_{p,k}(x_3)$ uniquely exists in $P_p^0(I)$ for each $\lambda_{p,k}$.

LEMMA 3.4 (Inverse inequality).

$$(3.11) \quad \int_S |\nabla \psi_{p,k}|^2 dx_1 dx_2 \leq Cp^4 \int_S |\psi_{p,k}|^2 dx_1 dx_2,$$

where C is a constant independent of p and k .

Proof. It is a typical inverse inequality in two dimensions; for the proof, we refer to, e.g., [11]. \square

LEMMA 3.5. Let $\lambda_{p,k}$ be an eigenvalue of the problem (3.7), and let $\phi_{p,k}(x_3)$ be the corresponding solution of two-point value problem (3.8). Then

$$(3.12) \quad \int_{-1}^1 \left(|\phi'_{p,k}|^2 + \lambda_{p,k} |\phi_{p,k}|^2 \right) dx_3 \leq C\sqrt{\lambda_{p,k}}, \quad k = 1, 2, \dots, N_p.$$

Proof. Since $\lambda_{p,k}$ is an eigenvalue of the problem (3.7), then

$$\lambda_{p,k} = \int_S (\nabla \psi_{p,k})^2 dx_1 dx_2.$$

By Lemma 3.4, there exists a constant $\eta > 0$ independent of p and k such that $0 < \lambda_{p,k} \leq \eta p^4$. Then for each k , we always can find a unique integer $1 \leq M_k \leq p$ satisfying

$$(3.13) \quad \eta(M_k - 1)^4 \leq \lambda_{p,k} \leq \eta M_k^4.$$

For each k , correspondingly we introduce the knots and the weights $\xi_i, \omega_i (i = 0, 1, \dots, M_k)$ of the Gauss–Legendre–Lobatto quadrature formula of order M_k on the interval $[-1, 1]$. We assume that the knots are ordered in such a way that $\xi_0 = -1$. Let χ_k be the Lagrange interpolation polynomial of degree M_k such that

$$\chi_k(\xi_i) = \begin{cases} 1, & \text{if } i = 0, \\ 0, & \text{otherwise.} \end{cases}$$

By the equivalence of discrete and continuous L^2 norms over $\mathcal{P}_{M_k}(-1, 1)$ (see [11]), there exists a constant $c_1 > 0$ independent of M_k such that

$$\int_{-1}^1 |\chi_k(x_1)|^2 dx_1 \leq c_1 \sum_{i=0}^{M_k} \chi_k^2(\xi_i) \omega_i = c_1 \omega_0.$$

Since $\omega_0 = \frac{2}{M_k(M_k+1)}$ (see [13]), we obtain

$$\int_{-1}^1 |\chi_k(x_1)|^2 dx_1 \leq \frac{c_2}{M_k^2},$$

and, by the inverse inequality, we have

$$\int_{-1}^1 |\chi'_k(x_1)|^2 dx_1 \leq c_2 \eta M_k^2.$$

Setting $q = \phi_{p,k} - \chi_k$ in (3.10) and by using the Cauchy–Schwarz inequality, we obtain

$$\int_{-1}^1 \left((\phi'_{p,k})^2 + \lambda_{p,k}(\phi_{p,k})^2 \right) dx_3 \leq CM_k^2.$$

Lemma 3.5 follows immediately by this inequality and (3.13). \square

Since $f(x_1, x_2) \in \mathcal{P}_p^{2,0}(S)$ and $\{\psi_{p,k}(x_1, x_2), 1 \leq k \leq N_p\}$ is an orthonormal basis of $\mathcal{P}_p^{2,0}(S)$,

$$f(x_1, x_2) = \sum_{k=1}^{N_p} \beta_k \psi_{p,k}(x_1, x_2),$$

with $\beta_k = \int_S f(x_1, x_2) \psi_{p,k}(x_1, x_2) dx_1 dx_2$. Let

$$(3.14) \quad R_D f = \sum_{k=1}^{N_p} \beta_k \psi_{p,k}(x_1, x_2) \phi_{p,k}(x_3).$$

Obviously,

$$R_D f|_{\Gamma_1} = \sum_{k=1}^{N_p} \beta_k \psi_{p,k}(x_1, x_2) = f(x_1, x_2),$$

where $\Gamma_1 = \{(x_1, x_2, -1) \mid -1 < x_1, x_2 < 1\}$.

THEOREM 3.6. *Let $D = (-1, 1)^3$ and $\Gamma_1 = \{(x_1, x_2, -1) \mid -1 < x_1, x_2 < 1\}$, then for $f \in \mathcal{P}_p^{2,0}(\Gamma_1)$, there exists $R_D f \in \mathcal{P}_p^2(D)$ such that $R_D f|_{\Gamma_1} = f$, $R_D f|_{\partial D \setminus \Gamma_1} = 0$, and*

$$(3.15) \quad \|R_D f\|_{H^1(D)} \leq C \|f\|_{H_{00}^{\frac{1}{2}}(\Gamma_1)},$$

where C is a constant, which is independent of p and f .

Proof. Let $\psi_{p,k}$ and $\phi_{p,k}$ be defined as in (3.7) and (3.10), and let $R_D f$ be given in (3.14), then

$$R_D f|_{\Gamma_1} = f, \quad R_D f|_{\partial D \setminus \Gamma_1} = 0.$$

Due to the orthogonality of the $\psi_{p,k} \in L^2(S)$ and $H^1(S)$ and by using (3.7) and Lemma 3.5, we have

$$\|R_D f\|_{L^2(D)}^2 = \sum_{k=1}^{N_p} \beta_k^2 \frac{1}{\sqrt{\lambda_{p,k}}}$$

and

$$\begin{aligned} |R_D f|_{H^1(D)}^2 &= \int_D \left(\left| \frac{\partial R_D f}{\partial x_1} \right|^2 + \left| \frac{\partial R_D f}{\partial x_2} \right|^2 + \left| \frac{\partial R_D f}{\partial x_3} \right|^2 \right) dx_1 dx_2 dx_3 \\ &= \sum_{k=1}^{N_p} \beta_k^2 \left(\int_S |\psi_{p,k}|^2 dx_1 dx_2 \int_I |\phi'_{p,k}|^2 dx_3 + \int_S |\nabla \psi_{p,k}|^2 dx_1 dx_2 \int_I |\phi_{p,k}|^2 dx_3 \right) \\ &= \sum_{k=1}^{N_p} \beta_k^2 \int_I \left(|\phi'_{p,k}|^2 + \lambda_{p,k} |\phi_{p,k}|^2 \right) dx_3 \leq C \sum_{k=1}^{N_p} \beta_k^2 \sqrt{\lambda_{p,k}}. \end{aligned}$$

Therefore,

$$(3.16) \quad \|R_D f\|_{H_0^1(D)}^2 \leq C \sum_{k=1}^{N_p} \beta_k^2 \left(1 + \sqrt{\lambda_{p,k}}\right).$$

Note that

$$\|f\|_{L^2(\Gamma_1)}^2 = \sum_{k=1}^{N_p} \beta_k^2, \quad \|f\|_{H_0^1(\Gamma_1)}^2 = \sum_{k=1}^{N_p} \beta_k^2 (1 + \lambda_{p,k}).$$

By interpolation space theory [8, 10, 19],

$$\|f\|_{H_{00}^{\frac{1}{2}}(\Gamma_1)}^2 \approx \sum_{k=1}^{N_p} \beta_k^2 (1 + \lambda_{p,k})^{\frac{1}{2}} \approx \sum_{k=1}^{N_p} \beta_k^2 \left(1 + \sqrt{\lambda_{p,k}}\right),$$

which together with (3.16) implies (3.15). \square

Analogously, we consider spectral solutions in either $\mathcal{P}_p^2(\Gamma_1)$ or ${}^0\mathcal{P}_p^2(\Gamma_1) = \{\varphi \in \mathcal{P}_p^2(\Gamma_1) \mid \varphi(\pm 1, x_2) = 0\}$ for the corresponding eigenvalue problems. Obviously, $\{\sqrt{(2i+1)(2j+1)}/2L_i(x_1)L_j(x_2), 0 \leq i, j \leq p\}$ and $\{\sqrt{2j+1}(1-x_1^2)/\sqrt{2\gamma_{i-1}^{2,2}}J_{i-1}^{2,2}(x_1)L_j(x_2), 1 \leq i \leq p-1, 0 \leq j \leq p\}$ are the orthonormal bases of $\mathcal{P}_p^2(\Gamma_1)$ and ${}^0\mathcal{P}_p^2(\Gamma_1)$, respectively, where $L_i(x_1)$ and $J_{i-1}^{2,2}(x_1)$ denote the Legendre and the Jacobi polynomials. The arguments for Theorem 3.6 can be carried out except replacing $\mathcal{P}_p^{2,0}(\Gamma_1)$ by $\mathcal{P}_p^2(\Gamma_1)$ or ${}^0\mathcal{P}_p^2(\Gamma_1)$. Therefore, we have the following two theorems which are parallel to Theorem 3.6.

THEOREM 3.7. *Let $D = [0, 1]^3$ and $\Gamma_1 = \{(x_1, x_2, 0) \mid 0 < x_1, x_2 < 1\}$, then for $f \in \mathcal{P}_p^2(\Gamma_1)$, there exists $U \in \mathcal{P}_p^2(D)$ such that $U|_{\Gamma_1} = f, U|_{\Gamma_4} = 0$ and*

$$(3.17) \quad \|U\|_{H^1(D)} \leq C \|f\|_{H^{\frac{1}{2}}(\Gamma_1)},$$

where C is a constant independent of p and f .

THEOREM 3.8. *Let $D = [0, 1]^3$ and $\Gamma_1 = \{(x_1, x_2, 0) \mid 0 < x_1, x_2 < 1\}$, then for $f \in \mathcal{P}_p^2(\Gamma_1)$, $f|_{\gamma_{12}} = 0, f|_{\gamma_{15}} = 0$, there exists $U \in \mathcal{P}_p^2(D)$ such that $U|_{\Gamma_1} = f, U|_{\Gamma_4} = 0, U|_{\Gamma_2} = 0, U|_{\Gamma_5} = 0$ and*

$$(3.18) \quad \|U\|_{H^1(D)} \leq C \|f\|_{H_{00}^{\frac{1}{2}}(\Gamma_1, \gamma_{12} \cup \gamma_{15})},$$

where C is a constant independent of p and f , and

$$(3.19) \quad \|u\|_{H_{00}^{\frac{1}{2}}(\Gamma_i, \gamma_{il} \cup \gamma_{im})}^2 = \|u\|_{H^{\frac{1}{2}}(\Gamma_i)}^2 + \int_{\Gamma_i} \frac{|u|^2}{\text{dist}(x, \gamma_{il})} dS_x + \int_{\Gamma_i} \frac{|u|^2}{\text{dist}(x, \gamma_{im})} dS_x.$$

Remark 3.1. Theorem 3.6 can be proved on a cube $(0, 1)^3$ by a simple mapping. Hereafter, $D = (0, 1)^3$ shall be the standard cube for the convenience in following sections.

Remark 3.2. The polynomial extension without using convolution was first proposed by Canuto and Funaro for the extension in square [10]. Since the polynomial extension of convolution-type is sufficient on triangle and square elements, the generalization of this approach to a cube is much more significant because it is the only polynomial extension compatible to FEM subspace on a cube.

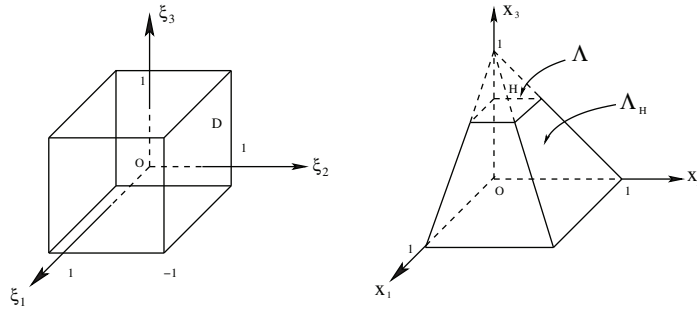


FIG. 3.2. A cube and a truncated pyramid Λ_H .

Remark 3.3. In [7] a similar extension was proposed by using spectral solutions of two eigenvalue problems in one dimension and one boundary value problem on an interval without rigorous proof. Recently, the same approach was developed with a proof in [12]. A genuine generalization of Canuto and Funaro’s approach from a square to a cube should be based on the spectral solution of an eigenvalue problem on a square, which is much better than the spectral solutions of two eigenvalue problems on an interval. More significantly, this approach can be used for a prism with nonsquare bases on which the eigenvalue problem cannot be decomposed into two one dimensional problems, e.g., a prism with a triangular base. The polynomial extension from a triangular base to a prism given in Theorem 2.4 can be proved by using this approach, but we will not elaborate the details here.

Remark 3.4. As an analogue to the extension on a square via a convolution-type extension on a triangle and a mapping of a square onto a truncated triangle [5, 18], we are able to construct an extension via a convolution-type extension on a tetrahedron and a mapping of a cube onto a truncated tetrahedron. It was shown that there is a convolution-type extension R_Λ from a square base S to a pyramid Λ such that R_Λ realizes a continuous mapping $H_{00}^{\frac{1}{2}}(S) \rightarrow H^1(\Lambda)$ and $R_\Lambda f|_S = f, R_\Lambda f|_{\partial\Lambda \setminus S} = 0$ [22]. Then a convolution-type extension \tilde{R}_D on a cube D is defined as

$$\begin{aligned} \tilde{R}_D f &= R_{\Lambda_H} f \circ M, \\ \tilde{R}_{\Lambda_H} f(x_1, x_2, x_3) &= \tilde{R}_\Lambda f(x_1, x_2, x_3) - \frac{x_3}{H} R_\Lambda f(x_1, x_2, H), \end{aligned}$$

where the mapping

$$M : x_i = \frac{\xi_i + 1}{2} \left(1 - \frac{H(\xi_3 + 1)}{2} \right), \quad i = 1, 2, \quad x_3 = \frac{H(\xi_3 + 1)}{2}$$

maps the cube D onto a truncated pyramid Λ_H as shown in Figure 3.2. It is easy to see that $\tilde{R}_D f \in \mathcal{P}_p^1(D), \tilde{R}_D f|_S = f, \tilde{R}_D f|_{\partial D \setminus S} = 0$ if $f \in \mathcal{P}_p^{1,0}(S)$. Note that $\tilde{R}_D f \notin \mathcal{P}_p^2(D)$, instead, $\tilde{R}_D f \in \mathcal{P}_p^{1,0}(S) \times \mathcal{P}_p^1(I)$ if $f \in \mathcal{P}_p^{2,0}(S)$. Hence, the convolution-type extension \tilde{R}_D is not compatible with the finite element space on the cube D and is not applicable to analysis of the p and h - p finite element solutions on meshes containing hexahedral elements.

3.2. Polynomial extension from whole boundary. We shall construct a polynomial extension E which lifts a polynomial on a whole boundary of a cube D in three steps, which is proved to be a continuous operator: $H^{\frac{1}{2}}(\partial D) \rightarrow H^1(T)$.

THEOREM 3.9. *Let $D = [0, 1]^3$ be the cube and $f \in \mathcal{P}_p^2(\partial D) = \{f \in C^0(\partial D), f|_{\Gamma_i} = f_i \in \mathcal{P}_p^2(\Gamma_i), i = 1, \dots, 6\}$, where Γ_i 's are the faces of cube D . Then there exists $E_D f \in \mathcal{P}_p^2(D)$ such that $E_D f|_{\partial D} = f$ and*

$$(3.20) \quad \|E_D f\|_{H^1(D)} \leq C \|f\|_{H^{\frac{1}{2}}(\partial D)},$$

where C is a constant independent of p and f , ∂D is the boundary of D .

Proof. By Theorem 3.7, there exist $U_1, U_4 \in \mathcal{P}_p^2(D)$ such that $U_1|_{\Gamma_1} = f_1, U_1|_{\Gamma_4} = 0; U_4|_{\Gamma_4} = f_4, U_4|_{\Gamma_1} = 0$, and

$$(3.21) \quad \|U_1\|_{H^1(D)} \leq C \|f_1\|_{H^{\frac{1}{2}}(\Gamma_1)}, \quad \|U_4\|_{H^1(D)} \leq C \|f_4\|_{H^{\frac{1}{2}}(\Gamma_4)}.$$

Let $g_2 = f_2 - U_1|_{\Gamma_2} - U_4|_{\Gamma_2}$ and $g_5 = f_5 - U_1|_{\Gamma_5} - U_4|_{\Gamma_5}$, then g_2 vanishes at the sides γ_{12} and γ_{24} of Γ_2 , and g_5 vanishes at the sides γ_{15} and γ_{45} of Γ_5 . By Theorem 3.8, there exist $U_2, U_5 \in \mathcal{P}_p^2(D)$ such that $U_2|_{\Gamma_2} = g_2, U_2|_{\Gamma_i} = 0, i = 1, 4, 5, U_5|_{\Gamma_5} = g_5, U_5|_{\Gamma_j} = 0, j = 1, 2, 4$, and

$$(3.22) \quad \|U_2\|_{H^1(D)} \leq C \|g_2\|_{H^{\frac{1}{2}}(\Gamma_2, \gamma_{12} \cup \gamma_{24})}, \quad \|U_5\|_{H^1(D)} \leq C \|g_5\|_{H^{\frac{1}{2}}(\Gamma_5, \gamma_{15} \cup \gamma_{45})}.$$

Let

$$g_3 = f_3 - \sum_{i=1,2,4,5} U_i|_{\Gamma_3}, \quad g_6 = f_6 - \sum_{i=1,2,4,5} U_i|_{\Gamma_6},$$

then

$$\begin{aligned} g_3|_{\gamma_{13}} &= -U_2|_{\gamma_{13}} - U_5|_{\gamma_{13}}, & g_3|_{\gamma_{23}} &= 0, & g_3|_{\gamma_{34}} &= -U_2|_{\gamma_{34}} - U_5|_{\gamma_{34}}, & g_3|_{\gamma_{35}} &= 0, \\ g_6|_{\gamma_{16}} &= -U_2|_{\gamma_{16}} - U_5|_{\gamma_{16}}, & g_6|_{\gamma_{26}} &= 0, & g_6|_{\gamma_{46}} &= -U_2|_{\gamma_{46}} - U_5|_{\gamma_{46}}, & g_6|_{\gamma_{56}} &= 0. \end{aligned}$$

By Theorem 3.8, there exist $U_3, U_6 \in \mathcal{P}_p^2(D)$ such that $U_3|_{\Gamma_3} = g_3, U_3|_{\Gamma_i} = 0, i = 2, 5, 6$, and $U_6|_{\Gamma_6} = g_6, U_6|_{\Gamma_j} = 0, j = 2, 3, 5$, and

$$(3.23) \quad \|U_3\|_{H^1(D)} \leq C \|g_3\|_{H^{\frac{1}{2}}(\Gamma_3, \gamma_{23} \cup \gamma_{35})}, \quad \|U_6\|_{H^1(D)} \leq C \|g_6\|_{H^{\frac{1}{2}}(\Gamma_6, \gamma_{26} \cup \gamma_{56})}.$$

Let $U = U_1 + U_2 + U_3 + U_4 + U_5 + U_6$. Then it is easy to see that $U|_{\Gamma_i} = f_i, i = 2, 3, 5, 6$. Let $g_1 = f_1 - U|_{\Gamma_1}, g_4 = f_4 - U|_{\Gamma_4}$. Since $\gamma_{12} = \bar{\Gamma}_1 \cap \bar{\Gamma}_2$ and $U_1|_{\Gamma_1} = f_1, U_2|_{\Gamma_1} = U_4|_{\Gamma_1} = U_5|_{\Gamma_1} = U_3|_{\Gamma_2} = U_6|_{\Gamma_2} = 0$, there holds

$$\begin{aligned} g_1|_{\gamma_{12}} &= (f_1 - U|_{\Gamma_1})|_{\gamma_{12}} = f_1|_{\gamma_{12}} - ((U_1 + U_2 + U_3 + U_4 + U_5 + U_6)|_{\Gamma_1})|_{\gamma_{12}} \\ &= f_1|_{\gamma_{12}} - (f_1 + U_2|_{\Gamma_1} + U_3|_{\Gamma_2} + U_4|_{\Gamma_1} + U_5|_{\Gamma_1} + U_6|_{\Gamma_2})|_{\gamma_{12}} = 0, \end{aligned}$$

and since $U_3|_{\gamma_{13}} = g_3|_{\gamma_{13}} = (f_3 - U_1 + U_2 + U_4 + U_5)|_{\gamma_{13}}$ and $U_6|_{\Gamma_3} = 0$, it holds that

$$g_1|_{\gamma_{13}} = (f_1 - U|_{\Gamma_1})|_{\gamma_{13}} = f_1|_{\gamma_{13}} - (U|_{\Gamma_3})|_{\gamma_{13}} = f_1|_{\gamma_{13}} - f_3|_{\gamma_{13}} = 0.$$

Similarly, it can be shown that $g_1|_{\gamma_{15}} = g_1|_{\gamma_{16}} = 0$. Hence, $g_1|_{\partial \Gamma_1} = 0$. Due to the symmetry, it holds that $g_4|_{\partial \Gamma_4} = 0$.

By Theorem 3.6, there exist $V_1 \in \mathcal{P}_p^{2,0}(\Gamma_1)$ and $V_4 \in \mathcal{P}_p^{2,0}(\Gamma_4)$ such that

$$\begin{aligned} V_1|_{\Gamma_1} &= g_1, & V_1|_{\Gamma_i} &= 0, & i &= 2, 3, 4, 5, 6, \\ V_4|_{\Gamma_4} &= g_4, & V_4|_{\Gamma_i} &= 0, & i &= 1, 2, 3, 5, 6, \end{aligned}$$

and

$$\|V_1\|_{H^1(D)} \leq C\|g_1\|_{H_{00}^{\frac{1}{2}}(\Gamma_1)}, \quad \|V_4\|_{H^1(D)} \leq C\|g_4\|_{H_{00}^{\frac{1}{2}}(\Gamma_4)}.$$

Let $E_D f = U + V_1 + V_4$, then we have $E_D f|_{\Gamma_i} = f_i$, $i = 1, 2, 3, 4, 5, 6$, and

$$\begin{aligned} (3.24) \quad \|E_D f\|_{H^1(S)} &\leq \|U\|_{H^1(S)} + \|V_1\|_{H^1(S)} + \|V_4\|_{H^1(S)} \\ &\leq C \left(\|f_1\|_{H^{\frac{1}{2}}(\Gamma_1)} + \|f_4\|_{H^{\frac{1}{2}}(\Gamma_4)} + \|g_2\|_{H_{00}^{\frac{1}{2}}(\Gamma_2, \gamma_{12} \cup \gamma_{24})} \right. \\ &\quad + \|g_5\|_{H_{00}^{\frac{1}{2}}(\Gamma_5, \gamma_{15} \cup \gamma_{45})} + \|g_3\|_{H_{00}^{\frac{1}{2}}(\Gamma_3, \gamma_{23} \cup \gamma_{35})} \\ &\quad \left. + \|g_6\|_{H_{00}^{\frac{1}{2}}(\Gamma_6, \gamma_{26} \cup \gamma_{56})} + \|g_1\|_{H_{00}^{\frac{1}{2}}(\Gamma_1)} + \|g_4\|_{H_{00}^{\frac{1}{2}}(\Gamma_4)} \right). \end{aligned}$$

First, we prove that

$$(3.25) \quad \|g_2\|_{H_{00}^{\frac{1}{2}}(\Gamma_2, \gamma_{12} \cup \gamma_{24})} \leq C\|f\|_{H^{\frac{1}{2}}(\Gamma_1 \cup \Gamma_2 \cup \Gamma_4)}.$$

Due to (3.21), there holds

$$\begin{aligned} (3.26) \quad \|g_2\|_{H^{\frac{1}{2}}(\Gamma_2)} &\leq \|f_2\|_{H^{\frac{1}{2}}(\Gamma_2)} + \|U_1\|_{H^{\frac{1}{2}}(\Gamma_2)} + \|U_4\|_{H^{\frac{1}{2}}(\Gamma_2)} \\ &\leq \|f_2\|_{H^{\frac{1}{2}}(\Gamma_2)} + C\|U_1\|_{H^1(D)} + C\|U_4\|_{H^1(D)} \\ &\leq C \left(\|f_2\|_{H^{\frac{1}{2}}(\Gamma_2)} + \|f_1\|_{H^{\frac{1}{2}}(\Gamma_1)} + \|f_4\|_{H^{\frac{1}{2}}(\Gamma_4)} \right). \end{aligned}$$

For (3.25), by the definition (3.19) of $H_{00}^{\frac{1}{2}}(\Gamma_2, \gamma_{12} \cup \gamma_{24})$, we need to show that

$$(3.27) \quad \int_S \frac{|g_2|^2}{x_3} dx_1 dx_3 \leq C\|f\|_{H^{\frac{1}{2}}(\Gamma_1 \cup \Gamma_2 \cup \Gamma_4)}, \quad \int_S \frac{|g_2|^2}{1-x_3} dx_1 dx_3 \leq C\|f\|_{H^{\frac{1}{2}}(\Gamma_1 \cup \Gamma_2 \cup \Gamma_4)}.$$

Since $U_1(x_1, x_3, 0) = f_1(x_1, x_3)$ and $U_4(x_1, x_3, 0) = 0$, there holds

$$\begin{aligned} g_2(x_1, x_3) &= f_2(x_1, x_3) - \sum_{i=1,4} U_i(x_1, x_2, x_3)|_{\Gamma_2} = f_2(x_1, x_3) - \sum_{i=1,4} U_i(x_1, 0, x_3) \\ &= (f_2(x_1, x_3) - f_1(x_1, x_3)) + (U_1(x_1, x_3, 0) - U_1(x_1, 0, x_3)) \\ &\quad + (U_4(x_1, x_3, 0) - U_4(x_1, 0, x_3)). \end{aligned}$$

Due to following equivalent norms for the space $H^{\frac{1}{2}}(\Gamma_2 \cup \Gamma_1)$ [3, 14],

$$(3.28) \quad \|f\|_{H^{\frac{1}{2}}(\Gamma_2 \cup \Gamma_1)} \approx \left(\|f_2\|_{H^{\frac{1}{2}}(\Gamma_2)}^2 + \|f_1\|_{H^{\frac{1}{2}}(\Gamma_1)}^2 + D(f_2, f_1) \right)^{\frac{1}{2}},$$

where

$$D(f_2, f_1) = \int_S \frac{|f_2(t_1, 0, t_2) - f_1(t_1, t_2, 0)|^2}{t_2} dt_1 dt_2,$$

we have

$$\begin{aligned} \int_S \frac{|f_2(x_1, x_3) - f_1(x_1, x_3)|^2}{x_3} dx_1 dx_3 &\leq \|f\|_{H^{\frac{1}{2}}(\Gamma_1 \cup \Gamma_2)}^2, \\ \int_S \frac{|U_1(x_1, x_3, 0) - U_1(x_1, 0, x_3)|^2}{x_3} dx_1 dx_3 &= D(U_1|_{\Gamma_1}, U_1|_{\Gamma_2}) \leq C \|U_1\|_{H^{\frac{1}{2}}(\Gamma_1 \cup \Gamma_2)}^2 \\ &\leq C \|U_1\|_{H^1(D)}^2 \leq C \|f_1\|_{H^{\frac{1}{2}}(\Gamma_1)}^2, \end{aligned}$$

and

$$\begin{aligned} \int_S \frac{|U_4(x_1, x_3, 0) - U_4(x_1, 0, x_3)|^2}{x_3} dx_1 dx_3 &= D(U_4|_{\Gamma_1}, U_4|_{\Gamma_2}) \leq C \|U_4\|_{H^{\frac{1}{2}}(\Gamma_1 \cup \Gamma_2)}^2 \\ &\leq C \|U_4\|_{H^1(D)}^2 \leq C \|f_4\|_{H^{\frac{1}{2}}(\Gamma_4)}^2. \end{aligned}$$

Therefore, we obtain the first inequality of (3.27).

For the second inequality of (3.27), we shall decompose $g_2(x_1, x_3)$ differently. Since $U_4(x_1, x_3, 1) = f_4(x_1, x_3)$ and $U_1(x_1, x_3, 1) = 0$, there holds

$$\begin{aligned} g_2(x_1, x_3) &= f_2(x_1, x_3) - \sum_{i=1,4} U_i(x_1, x_2, x_3)|_{\Gamma_2} = f_2(x_1, x_3) - \sum_{i=1,4} U_i(x_1, 0, x_3) \\ &= (f_2(x_1, x_3) - f_4(x_1, x_3)) + (U_4(x_1, x_3, 1) - U_4(x_1, 0, x_3)) \\ &\quad + (U_1(x_1, x_3, 1) - U_1(x_1, 0, x_3)). \end{aligned}$$

Arguing as previously, we have the second inequality of (3.27). Then (3.25) follows immediately from (3.26)–(3.27). Due to the symmetry, we have analogously

$$(3.29) \quad \|g_5\|_{H^{\frac{1}{2}}(\Gamma_5, \gamma_{15} \cup \gamma_{45})} \leq C \|f\|_{H^{\frac{1}{2}}(\Gamma_1 \cup \Gamma_4 \cup \Gamma_5)}.$$

We shall next prove that

$$(3.30) \quad \|g_3\|_{H^{\frac{1}{2}}(\Gamma_3, \gamma_{23} \cup \gamma_{35})} \leq C \|f\|_{H^{\frac{1}{2}}(\partial D \setminus \Gamma_6)}, \quad \|g_6\|_{H^{\frac{1}{2}}(\Gamma_6, \gamma_{26} \cup \gamma_{65})} \leq C \|f\|_{H^{\frac{1}{2}}(\partial D \setminus \Gamma_3)}.$$

By (3.22), (3.25), and (3.29) we have

$$\begin{aligned} (3.31) \quad \|g_3\|_{H^{\frac{1}{2}}(\Gamma_3)} &= \left\| f_3 - \sum_{i=1,2,4,5} U_i|_{\Gamma_3} \right\|_{H^{\frac{1}{2}}(\Gamma_3)} \\ &\leq \|f_3\|_{H^{\frac{1}{2}}(\Gamma_3)} + C \sum_{i=1,2,4,5} \|U_i\|_{H^1(D)} \\ &\leq C \left(\|f_3\|_{H^{\frac{1}{2}}(\Gamma_3)} + \sum_{i=1,4} \|f_i\|_{H^{\frac{1}{2}}(\Gamma_i)} + \sum_{i=2,5} \|g_i\|_{H^{\frac{1}{2}}(\Gamma_i, \gamma_{1i} \cup \gamma_{i4})} \right) \\ &\leq C \left(\|f_3\|_{H^{\frac{1}{2}}(\Gamma_3)} + \sum_{i=1,4} \|f_i\|_{H^{\frac{1}{2}}(\Gamma_i)} + \sum_{i=2,5} \|f\|_{H^{\frac{1}{2}}(\Gamma_1 \cup \Gamma_i \cup \Gamma_4)} \right) \\ &\leq C \|f\|_{H^{\frac{1}{2}}(\partial D \setminus \Gamma_6)}. \end{aligned}$$

For the first inequality of (3.30), due to the definition (3.19) of $H_{00}^{\frac{1}{2}}(\Gamma_3, \gamma_{23} \cup \gamma_{35})$, it remains to show that

$$(3.32) \quad \int_S \frac{|g_3|^2}{x_2} dx_1 dx_3 \leq C \|f\|_{H^{\frac{1}{2}}(\partial D \setminus \Gamma_6)}, \quad \int_S \frac{|g_3|^2}{1-x_2} dx_1 dx_3 \leq C \|f\|_{H^{\frac{1}{2}}(\partial D \setminus \Gamma_6)}.$$

Since $U_2(x_2, 0, x_3) = g_2(x_2, x_3)$ and $U_5(x_2, 0, x_3) = 0$, we have

$$\begin{aligned} g_3(x_2, x_3) &= f_3(x_2, x_3) - g_2(x_2, x_3) + U_2(x_2, 0, x_3) - \sum_{i=1,2,4,5} U_i(0, x_2, x_3) \\ &= f_3(x_2, x_3) - (f_2(x_2, x_3) - U_1(x_2, 0, x_3) - U_4(x_2, 0, x_3)) + U_2(x_2, 0, x_3) \\ &\quad - U_1(0, x_2, x_3) - U_4(0, x_2, x_3) - U_2(0, x_2, x_3) - U_5(0, x_2, x_3) \\ &= (f_3(x_2, x_3) - f_2(x_2, x_3)) + (U_1(x_2, 0, x_3) - U_1(0, x_2, x_3)) + (U_4(x_2, 0, x_3) \\ &\quad - U_4(0, x_2, x_3)) + (U_2(x_2, 0, x_3) - U_2(0, x_2, x_3)) + (U_5(x_2, 0, x_3) - U_5(0, x_2, x_3)). \end{aligned}$$

By the equivalent norm of $H^{\frac{1}{2}}(\Gamma_2 \cup \Gamma_3)$ described in (3.28), we have

$$\begin{aligned} \int_S \frac{|f_3(x_2, x_3) - f_2(x_2, x_3)|}{x_2} dx_2 dx_3 &\leq \|f\|_{H^{\frac{1}{2}}(\Gamma_3 \cup \Gamma_2)}^2, \\ \int_S \frac{|U_1(x_2, 0, x_3) - U_1(0, x_2, x_3)|}{x_2} dx_2 dx_3 &= D(U_1|_{\Gamma_2}, U_1|_{\Gamma_3}) \leq C \|U_1\|_{H^{\frac{1}{2}}(\Gamma_2 \cup \Gamma_3)}^2 \\ &\leq C \|U_1\|_{H^1(D)}^2 \leq C \|f_1\|_{H^{\frac{1}{2}}(\Gamma_1)}^2, \\ \int_S \frac{|U_4(x_2, 0, x_3) - U_4(0, x_2, x_3)|}{x_2} dx_2 dx_3 &= D(U_4|_{\Gamma_2}, U_4|_{\Gamma_3}) \leq C \|U_4\|_{H^{\frac{1}{2}}(\Gamma_2 \cup \Gamma_3)}^2 \\ &\leq C \|U_4\|_{H^1(D)}^2 \leq C \|f_4\|_{H^{\frac{1}{2}}(\Gamma_4)}^2, \\ \int_S \frac{|U_2(x_2, 0, x_3) - U_2(0, x_2, x_3)|}{x_2} dx_2 dx_3 &= D(U_2|_{\Gamma_2}, U_2|_{\Gamma_3}) \leq C \|U_2\|_{H^{\frac{1}{2}}(\Gamma_2 \cup \Gamma_3)}^2 \\ &\leq C \|U_2\|_{H^1(D)}^2 \leq C \|g_2\|_{H_{00}^{\frac{1}{2}}(\Gamma_2, \gamma_{12} \cup \gamma_{24})}^2 \\ &\leq C \|f\|_{H^{\frac{1}{2}}(\Gamma_1 \cup \Gamma_2 \cup \Gamma_4)}^2, \\ \int_S \frac{|U_5(x_2, 0, x_3) - U_5(0, x_2, x_3)|}{x_2} dx_2 dx_3 &= D(U_5|_{\Gamma_2}, U_5|_{\Gamma_3}) \leq C \|U_5\|_{H^{\frac{1}{2}}(\Gamma_2 \cup \Gamma_3)}^2 \\ &\leq C \|U_5\|_{H^1(D)}^2 \leq C \|g_5\|_{H_{00}^{\frac{1}{2}}(\Gamma_5, \gamma_{12} \cup \gamma_{24})}^2 \\ &\leq C \|f\|_{H^{\frac{1}{2}}(\Gamma_1 \cup \Gamma_4 \cup \Gamma_5)}^2. \end{aligned}$$

Then the first inequality of (3.32) follows easily.

For the second inequality of (3.32), we shall decompose $g_3(x_2, x_3)$ in different way, i.e.,

$$\begin{aligned} g_3(x_2, x_3) &= (f_3(x_2, x_3) - f_5(x_2, x_3)) + (U_1(x_2, 1, x_3) - U_1(0, x_2, x_3)) \\ &\quad + (U_4(x_2, 1, x_3) - U_4(0, x_2, x_3)) + (U_5(x_2, 1, x_3) - U_5(0, x_2, x_3)) \\ &\quad + (U_2(x_2, 1, x_3) - U_2(0, x_2, x_3)). \end{aligned}$$

Arguing as previously, we obtain the second estimation of (3.32). A combination of (3.31) and (3.32) leads to the first inequality of (3.30). By the symmetry, we have the second one of (3.30).

Finally, we prove that

$$(3.33) \quad \|g_i\|_{H^{\frac{1}{2}}_{00}(\Gamma_1)} \leq C\|f\|_{H^{\frac{1}{2}}(\partial D)}, \quad i = 1, 4.$$

By (3.21)–(3.23) and (3.25), (3.29)–(3.30), there holds

$$(3.34) \quad \begin{aligned} \|g_1\|_{H^{\frac{1}{2}}(\Gamma_1)} &= \|f_1 - U|_{\Gamma_1}\|_{H^{\frac{1}{2}}(\Gamma_1)} \\ &\leq \|f_1\|_{H^{\frac{1}{2}}(\Gamma_1)} + \sum_{i=1}^6 \|U_i\|_{H^{\frac{1}{2}}(\Gamma_1)} \leq C\|f\|_{H^{\frac{1}{2}}(\partial D)}. \end{aligned}$$

For (3.33) with $i = 1$, we need to show that, for $j = 1, 2$,

$$(3.35) \quad \int_S \frac{|g_1(x_1, x_2)|^2}{x_j} dx_1 dx_2 \leq C\|f\|_{H^{\frac{1}{2}}(\partial D)}, \quad \int_S \frac{|g_1(x_1, x_2)|^2}{1 - x_j} dx_1 dx_2 \leq C\|f\|_{H^{\frac{1}{2}}(\partial D)}.$$

Since $U_2|_{\Gamma_2} = g_2$, $U_5|_{\Gamma_1} = 0$, $U_3|_{\Gamma_2} = 0$ and $U_6|_{\Gamma_2} = 0$, we have

$$\begin{aligned} g_1(x_1, x_2) &= f_1(x_1, x_2) - g_2(x_1, x_2) + U_2(x_1, 0, x_2) - U(x_1, x_2, x_3)|_{\Gamma_1} \\ &= f_1(x_1, x_2) - (f_2(x_1, x_2) - U_1(x_1, 0, x_2) - U_4(x_1, 0, x_2)) \\ &\quad + U_2(x_1, 0, x_2) - \sum_{1 \leq i \leq 6} U_i(x_1, x_2, 0) \\ &= (f_1(x_1, x_2) - f_2(x_1, x_2)) + (U_1(x_1, 0, x_2) - U_1(x_1, x_2, 0)) \\ &\quad + (U_4(x_1, 0, x_2) - U_4(x_1, x_2, 0)) + (U_2(x_1, 0, x_2) - U_2(x_1, x_2, 0)) \\ &\quad + (U_3(x_1, 0, x_2) - U_3(x_1, x_2, 0)) + (U_6(x_1, 0, x_2) - U_6(x_1, x_2, 0)). \end{aligned}$$

By the equivalent norm of $H^{\frac{1}{2}}(\Gamma_2 \cup \Gamma_1)$ described in (3.28), there hold

$$\begin{aligned} \int_S \frac{|f_1(x_1, x_2) - f_2(x_1, x_2)|^2}{x_2} dx_1 dx_2 &\leq \|f\|_{H^{\frac{1}{2}}(\Gamma_1 \cup \Gamma_2)}, \\ \int_S \frac{|U_1(x_1, 0, x_2) - U_1(x_1, x_2, 0)|^2}{x_2} dx_1 dx_2 &= D(U_1|_{\Gamma_2}, U_1|_{\Gamma_1}) \leq C\|U_1\|_{H^{\frac{1}{2}}(\Gamma_2 \cup \Gamma_1)}^2 \\ &\leq C\|U_1\|_{H^1(D)}^2 \leq C\|f_1\|_{H^{\frac{1}{2}}(\Gamma_1)}^2, \\ \int_S \frac{|U_4(x_1, 0, x_2) - U_4(x_1, x_2, 0)|^2}{x_2} dx_1 dx_2 &= D(U_4|_{\Gamma_2}, U_4|_{\Gamma_1}) \leq C\|U_4\|_{H^{\frac{1}{2}}(\Gamma_2 \cup \Gamma_1)}^2 \\ &\leq C\|U_4\|_{H^1(D)}^2 \leq C\|f_4\|_{H^{\frac{1}{2}}(\Gamma_4)}^2, \\ \int_S \frac{|U_2(x_1, 0, x_2) - U_2(x_1, x_2, 0)|^2}{x_2} dx_1 dx_2 &= D(U_2|_{\Gamma_2}, U_2|_{\Gamma_1}) \leq C\|U_2\|_{H^{\frac{1}{2}}(\Gamma_2 \cup \Gamma_1)}^2 \\ &\leq C\|U_2\|_{H^1(D)}^2 \leq C\|g_2\|_{H^{\frac{1}{2}}_{00}(\Gamma_2, \gamma_{12} \cup \gamma_{24})}^2 \\ &\leq C\|f\|_{H^{\frac{1}{2}}(\Gamma_1 \cup \Gamma_2 \cup \Gamma_4)}^2, \end{aligned}$$

$$\begin{aligned} \int_S \frac{|U_3(x_1, 0, x_2) - U_3(x_1, x_2, 0)|^2}{x_2} dx_1 dx_2 &= D(U_3|_{\Gamma_2}, U_3|_{\Gamma_1}) \leq C \|U_3\|_{H^{\frac{1}{2}}(\Gamma_2 \cup \Gamma_1)}^2 \\ &\leq C \|U_3\|_{H^1(D)}^2 \leq C \|g_3\|_{H^{\frac{1}{2}}(\Gamma_3, \gamma_{23} \cup \gamma_{35})}^2 \\ &\leq C \|f\|_{H^{\frac{1}{2}}(\partial D \setminus \Gamma_6)}^2, \end{aligned}$$

and

$$\begin{aligned} \int_S \frac{|U_6(x_1, 0, x_2) - U_6(x_1, x_2, 0)|^2}{x_2} dx_1 dx_2 &= D(U_6|_{\Gamma_2}, U_6|_{\Gamma_1}) \leq C \|U_6\|_{H^{\frac{1}{2}}(\Gamma_2 \cup \Gamma_1)}^2 \\ &\leq C \|U_6\|_{H^1(D)}^2 \leq C \|g_6\|_{H^{\frac{1}{2}}(\Gamma_6, \gamma_{26} \cup \gamma_{56})}^2 \\ &\leq C \|f\|_{H^{\frac{1}{2}}(\partial D \setminus \Gamma_3)}^2. \end{aligned}$$

The above inequalities lead to the first estimation of (3.35) for $j = 2$.

For the second inequality of (3.35) with $j = 2$, we shall decompose g_1 differently. Since $U_5|_{\Gamma_5} = g_5$, $U_2|_{\Gamma_5} = 0$, $U_3|_{\Gamma_5} = 0$, and $U_6|_{\Gamma_5} = 0$, there holds

$$\begin{aligned} g_1(x_1, x_2) &= (f_1(x_1, x_2) - f_5(x_1, x_2)) + (U_1(x_1, 1, x_2) - U_1(x_1, x_2, 0)) \\ &\quad + (U_4(x_1, 1, x_2) - U_4(x_1, x_2, 0)) + (U_5(x_1, 1, x_2) - U_5(x_1, x_2, 0)) \\ &\quad + (U_2(x_1, 1, x_2) - U_2(x_1, x_2, 0)) + (U_3(x_1, 1, x_2) - U_3(x_1, x_2, 0)) \\ &\quad + (U_6(x_1, 1, x_2) - U_6(x_1, x_2, 0)). \end{aligned}$$

Arguing as above, we can get the second inequality of (3.35) for $j = 2$.

For the first and second inequalities of (3.35) for $j = 1$, we decompose g_1 in two other ways. Since $U_3|_{\Gamma_3} = g_3$, $U_6|_{\Gamma_3} = 0$, $U_6|_{\Gamma_6} = g_6$, and $U_3|_{\Gamma_6} = 0$, we have

$$\begin{aligned} g_1(x_1, x_2) &= (f_1(x_1, x_2) - f_3(x_1, x_2)ig) + (U_1(0, x_1, x_2) - U_1(x_1, x_2, 0)) \\ &\quad + (U_4(0, x_1, x_2) - U_4(x_1, x_2, 0)) + (U_2(0, x_1, x_2) - U_2(x_1, x_2, 0)) \\ &\quad + (U_5(0, x_1, x_2) - U_5(x_1, x_2, 0)) + (U_3(x_1, 1, x_2) - U_3(x_1, x_2, 0)) \\ &\quad + (U_6(0, x_1, x_2) - U_6(x_1, x_2, 0)) \end{aligned}$$

and

$$\begin{aligned} g_1(x_1, x_2) &= (f_1(x_1, x_2) - f_6(x_1, x_2)) + (U_1(1, x_1, x_2) - U_1(x_1, x_2, 0)) \\ &\quad + (U_4(1, x_1, x_2) - U_4(x_1, x_2, 0)) + (U_2(1, x_1, x_2) - U_2(x_1, x_2, 0)) \\ &\quad + (U_5(1, x_1, x_2) - U_5(x_1, x_2, 0)) + (U_3(1, x_1, x_2) - U_3(x_1, x_2, 0)) \\ &\quad + (U_6(1, x_1, x_2) - U_6(x_1, x_2, 0)), \end{aligned}$$

respectively, which implies (3.35) for $j = 1$.

Combining (3.34) and (3.35), we obtain (3.33) for $i = 1$. Analogously, we have (3.33) for $i = 4$ due to the symmetry, which together with (3.24)–(3.25) and (3.29)–(3.30) leads to (3.20). Thus, we complete the proof. \square

4. Applications to the error analysis of p -version of FEM. Tetrahedrons(simplices), triangular prisms(wedges), and hexahedrons(cubes) are three commonly used elements for the FEM in three dimensions. We have established polynomial extensions R_G, R_Λ , and R_D on a triangular prism, a pyramid, and a cube, which, with the polynomial extension R_K on a tetrahedron [21], are sufficient for the

construction of a globally continuous and piecewise polynomial on a mesh containing tetrahedral elements, triangular prism elements, and hexahedral elements. Therefore, approximation errors in solutions of the p and h - p version can be proved to be as good as in local projections without comprising the optimal rate of the convergence. We will illustrate how to incorporate the local projection with polynomial extensions in the error analysis for the p -version of the FEM; the details of the proof are given in a coming paper [15].

Let Ω be a Lipschitz domain in R^3 , and let $\Delta = \{\Omega_j, 1 \leq j \leq J\}$ be a partition of Ω . Ω_j 's are shape-regular and surfaced tetrahedral, hexahedral, and triangular-prism elements. By M_j , we denote a mapping of standard element Ω_{st} onto Ω_j , where Ω_{st} is the standard tetrahedral K , or the standard triangular-prism G , or the standard hexahedron D which we defined in previous sections. Let $\mathcal{P}_{p_j}(\Omega_j)$ denote a set of pull-back polynomials φ on Ω_j such that $\varphi \circ M_j \in \mathcal{P}_{p_j}^\kappa(\Omega_{st})$, with $\kappa = 1$ if Ω_{st} is the tetrahedron K , $\kappa = 2$ if Ω_{st} is the hexahedron D , and $\mathcal{P}_p^{1.5}(\Omega_{st}) = \mathcal{P}_p^1(T) \times \mathcal{P}_p(I)$ if Ω_{st} is the triangular-prism G . By P , we denote the distribution of the element degrees. As usual, the finite element subspaces of piecewise pull-back and continuous polynomials are defined as

$$(4.1) \quad S_D^{P,1}(\Omega; \Delta) = S_D^P(\Omega; \Delta) \cap H_D^1(\Omega), \quad S_D^{P,1}(\Omega; \Delta) = \{\varphi | \varphi|_{\Omega_j} \in \mathcal{P}_p(\Omega_j), 1 \leq j \leq J\},$$

where $H_D^1(\Omega)$ denotes the set of $u \in H^1(\Omega)$, with $u = 0$ on Γ_D .

Incorporating the polynomial extensions with the approximation in the framework of Jacobi-weighted Sobolev spaces, we have the following theorem, which leads to the error estimates for the p -version of the FEM with a quasi-uniform degree distribution in three dimensions.

THEOREM 4.1. *Let $u \in H^k(\Omega), k \geq 1$, and let $S_D^{P,1}(\Omega; \Delta)$ be the finite element subspace defined with a uniform degree p as in (4.1). Then there exists a polynomial $\varphi \in S_D^{P,1}(\Omega; \Delta)$ such that*

$$(4.2) \quad \|u - \varphi\|_{H^1(\Omega)} \leq C(p + 1)^{-(k-1)} \|u\|_{H^k(\Omega)},$$

with a constant C independent of p and u .

We shall outline the proof and emphasize the essential role which the polynomial extensions play, and we refer readers to [15] for the details. To this end, we introduce three important propositions.

PROPOSITION 4.2. *Let $u \in H^k(\Omega_j), k > \frac{3}{2}$, where Ω_j is a tetrahedron, or a prism, or a cube with planar surfaces or nonplanar surfaces. Then there exists a polynomial $\phi \in \mathcal{P}_p^\kappa(\Omega_j)$, with $p \geq 1$ and $\kappa = 1, 1.5, 2$, respectively, such that for $0 \leq \ell \leq k$,*

$$(4.3) \quad \|u - \phi\|_{H^\ell(\Omega_j)} \leq Cp^{-(k-\ell)} \|u\|_{H^k(\Omega_j)},$$

and $u = \phi$ at vertices V_ℓ of $\Omega_j, 1 \leq \ell \leq L, L = 4$ or 6 or 8 , respectively.

PROPOSITION 4.3. *Let $\gamma = (-\frac{1}{2}, \frac{1}{2})$ and $u \in H^s(\gamma), s > 1/2$. Then there exists an operator $\pi_\gamma = H^s(\gamma) \rightarrow \mathcal{P}_p(\gamma)$ such that $u(\pm\frac{1}{2}) = \pi_\gamma u(\pm\frac{1}{2})$ and for $0 \leq l \leq s$,*

$$(4.4) \quad \|u - \pi_\gamma u\|_{H^l(\gamma)} \leq C(p + 1)^{-(s-l)} \|u\|_{H^s(\gamma)},$$

with a constant C independent of p and u .

PROPOSITION 4.4. *Let Ω_{st} be a standard tetrahedron, or triangular prism, or hexahedron, and let $u \in H^s(\Omega_{st}), s \geq 2$. Then there exists a polynomial $\varphi_j \in \mathcal{P}_p(\Omega_{st})$ such that $u(V_l) = \varphi_p(V_l)$ at the vertices V_l of Ω_j , and $\varphi_p|_\gamma = \pi_\gamma u$ on each edge of Ω_{st} ,*

$$(4.5) \quad \|u - \varphi_j\|_{H^l(\Omega_{st})} \leq C(p + 1)^{-(s-l)} \|u\|_{H^s(\Omega_{st})}$$

and on each face of Ω_{st}

$$(4.6) \quad \|u - \psi\|_{H^t(F_i)} \leq Cp^{-(k-t-\frac{1}{2})} \|u\|_{H^k(\Omega_{st})}, t = 0, 1,$$

with a constant C independent of p and u . If Ω_{st} is a standard triangular prism and $u \in H^s(\Omega_{st}), s \geq 3$, it holds that

$$(4.7) \quad \left\| \frac{\partial(u - \psi)}{\partial x_3} \right\|_{H^1(F_i)} \leq Cp^{-(k-\frac{5}{2})} \|u\|_{H^k(\Omega_{st})}.$$

The construction of the operator π_γ and the polynomial φ_p are started with the Jacobi projection with $\beta = -1/2$ (Chebyshev projection) and followed by the modification at vertices and on edges.

Proof of Theorem 4.1. We first assume that $k \geq 2$. Due to Proposition 4.4, we have a polynomial $\varphi_j \in \mathcal{P}_p(\Omega_j)$ in each element Ω_j such that $u = \varphi_j$ at each vertex V of Ω_j and $\varphi_j = \pi_\gamma u$ on each edge γ of Ω_j , where π_γ is the projection-like operator defined as in Proposition 4.3, and, for $0 \leq l \leq k$,

$$(4.8) \quad \|u - \varphi_j\|_{H^l(\Omega_j)} \leq C(p + 1)^{-(k-l)} \|u\|_{H^k(\Omega_j)}.$$

Suppose that $F = \bar{\Omega}_j \cap \bar{\Omega}_i$ is a common face of two neighboring elements Ω_j and Ω_i . We may assume without loss of generality that Ω_i and Ω_j are standard-size elements.

If F is a standard triangle T , there are three possible cases:

- (T1) both are tetrahedrons;
- (T2) both are triangular prisms;
- (T3) Ω_j is a tetrahedron and Ω_i is a triangular prism.

If F is a standard square face S , similarly, there are three possible cases:

- (S1) both Ω_j and Ω_i are hexahedrons;
- (S2) both Ω_j and Ω_i are triangular prisms;
- (S3) Ω_j is a hexahedron and Ω_i is a triangular prism.

We shall modify φ_i and φ_j in the cases (T1) and (S2); the treatment for other cases are similar with what follows.

In the case (T1), Ω_i and Ω_j are tetrahedrons. $\psi = (\varphi_i - \varphi_j)|_F \in \mathcal{P}_p^{1,0}(F)$. By Theorem 2.1, there is a polynomial $\Psi \in \mathcal{P}_p^1(\Omega_j)$ such that $\Psi|_F = \psi$ and $\Psi|_{\partial\Omega_j \setminus F} = 0$, and

$$(4.9) \quad \|\Psi\|_{H^1(\Omega_j)} \leq C \|\psi\|_{H_{00}^{\frac{1}{2}}(F)} = C \|\varphi_i - \varphi_j\|_{H_{00}^{\frac{1}{2}}(F)}.$$

Note that $(\varphi_i - \varphi_j) \in H_{00}^{\frac{1}{2}}(F) = (H^0(F), H_0^1(F))_{\frac{1}{2}, 2}$ and that for $t = 0, 1$,

$$\begin{aligned} \|\varphi_i - \varphi_j\|_{H^t(F)} &\leq C (\|\varphi_i - u\|_{H^t(F)} + \|\varphi_j - u\|_{H^t(F)}) \\ &\leq C(p + 1)^{-(k+t-1/2)} (\|u\|_{H^k(\Omega_j)} + \|u\|_{H^k(\Omega_i)}), \end{aligned}$$

which implies

$$(4.10) \quad \|\Psi\|_{H^1(\Omega_j)} \leq C(p + 1)^{-(k-1)} (\|u\|_{H^k(\Omega_j)} + \|u\|_{H^k(\Omega_i)}).$$

In the case (S2), by Proposition 4.4, there are $\varphi_i \in \mathcal{P}_p^{1.5}(\Omega_i)$ and $\varphi_j \in \mathcal{P}_p^{1.5}(\Omega_j)$ satisfying (4.5)–(4.7). Suppose that $F = \{x = (x_1, 0, x_3) \mid 0 \leq x_1, x_3 \leq 1\}$. Then

$\psi(x_1, x_3) = (\varphi_i - \varphi_j)|_F \in \mathcal{P}_p^{2,0}(F)$, and there exists a polynomial extension Ψ on Ω_j [18] such that $\Psi \in \mathcal{P}_p^{1,5}(\Omega_j)$, $\Psi|_F = \psi$ and $\Psi|_{\partial\Omega_j \setminus F} = 0$, and

$$\|\Psi\|_{H^1(\Omega_j)} \leq C \left((p+1)^{-\frac{3}{2}} \|\psi_{x_3}\|_{H^1(F)} + (p+1)^{-\frac{1}{2}} \|\psi\|_{H^1(F)} + (p+1)^{\frac{1}{2}} \|\psi\|_{L^2(F)} \right).$$

Due to (4.5) and (4.7), there hold for $t = 0, 1$,

$$\|\psi\|_{H^t(F)} \leq \|u - \varphi_j\|_{H^t(F)} + \|u - \varphi_i\|_{H^t(F)} \leq C(p+1)^{-(k-t-\frac{1}{2})} (\|u\|_{H^k(\Omega_j)} + \|u\|_{H^k(\Omega_i)})$$

and

$$\begin{aligned} \|\psi_{x_3}\|_{H^1(F)} &\leq \left\| \frac{\partial(u - \varphi_j)}{\partial x_3} \right\|_{H^1(F)} + \left\| \frac{\partial(u - \varphi_i)}{\partial x_3} \right\|_{H^1(F)} \\ &\leq C(p+1)^{-(k-\frac{5}{2})} (\|u\|_{H^k(\Omega_j)} + \|u\|_{H^k(\Omega_i)}), \end{aligned}$$

which implies (4.10).

Let $\tilde{\varphi}_j = \varphi_j + \Psi$ and $\tilde{\varphi}_i = \varphi_i$. Then $\tilde{\varphi}_j = \tilde{\varphi}_i$ on F , and by (4.9) and (4.10),

$$(4.11) \quad \begin{aligned} \|u - \tilde{\varphi}_j\|_{H^1(\Omega_j)} &\leq \|u - \varphi_j\|_{H^1(\Omega_j)} + \|\Psi\|_{H^1(\Omega_j)} \\ &\leq C(p+1)^{-(k-1)} (\|u\|_{H^k(\Omega_j)} + \|u\|_{H^k(\Omega_i)}) \end{aligned}$$

and

$$(4.12) \quad \|u - \tilde{\varphi}_i\|_{H^1(\Omega_i)} = \|u - \varphi_i\|_{H^1(\Omega_i)} \leq C(p+1)^{-(k-1)} \|u\|_{H^k(\Omega_i)}.$$

Adjusting φ_j on each face of Ω_j by the polynomial extension Ψ , we achieve the continuity across interfaces of elements. For the homogeneous Dirichlet boundary condition, we can adjust φ_j in similar way such that $\tilde{\varphi}_j \in \mathcal{P}_p^k(\Omega_j)$ and vanishes on $\Gamma_D \cap \partial\Omega_j$. Let $\phi = \tilde{\varphi}_j$ in $\Omega_j, 1 \leq j \leq J$. Then $\varphi \in S_D^{P,1}(\Omega; \Delta)$ and satisfies (4.2).

We next prove (4.2) for $1 < k < 3$. It was shown in [4] that $H^k(\Omega) \cap H_D^1(\Omega) = (H_D^1(\Omega), H^3(\Omega) \cap H_D^1(\Omega))_{\theta,2} \subset (H^1(\Omega), H^3(\Omega))_{\theta,2} \cap H_D^1(\Omega)$, with $\theta = \frac{k-1}{2} \in (0, 1)$ for $1 < k < 3$. Since $(H^1(\Omega), H^3(\Omega))_{\theta,2} \subset (H^1(\Omega), H^3(\Omega))_{\theta,\infty} = B^k(\Omega)$, $H^k(\Omega) \cap H_D^1(\Omega) \subset B^k(\Omega) \cap H_D^1(\Omega)$. Suppose that $v \in H_D^1(\Omega)$ and $w \in H^3(\Omega) \cap H_D^1(\Omega)$ form a decomposition of $u \in B^k(\Omega) \cap H_D^1(\Omega)$. Applying (4.2) for $k = 3$, we have a polynomial $\varphi \in S_D^{P,1}(\Omega; \Delta)$, with $p \geq 1$ such that

$$\|w - \varphi_p\|_{H^1(\Omega)} \leq C \frac{1}{(p+1)^2} \|w\|_{H^3(\Omega)}.$$

Therefore, we have for any decomposition v and w of u ,

$$\begin{aligned} \|u - \varphi\|_{H^1(\Omega)} &\leq \|v\|_{H^1(\Omega)} + \|w - \varphi_p\|_{H^1(\Omega)} \\ &\leq C \left(\|v\|_{H^1(\Omega)} + \frac{1}{(p+1)^2} \|w\|_{H^3(\Omega)} \right) \\ &= C (\|v\|_{H^1(\Omega)} + t_1 \|w\|_{H^3(\Omega)}), \end{aligned}$$

with $t_1 = \frac{1}{(p+1)^2}$ and C independent of v and w . Due to the definition of the Besov space $B^k(\Omega)$, we have

$$\begin{aligned} \|u - \varphi\|_{H^1(\Omega)} &\leq CK(u, t_1) \leq Ct_1^\theta \sup_{t>0} t^{-\theta} K(u, t) \\ &\leq Ct_1^\theta \|u\|_{B^k(\Omega)} \leq C(p+1)^{k-1} \|u\|_{H^k(\Omega)}. \end{aligned}$$

For $p = 0$ or $k = 1$, (4.2) is trivial by selecting $\varphi = 0$. Thus, the proof of the theorem is completed. \square

Remark 4.1. For elliptic problems, there holds the finite element solution $u_p \in S_D^{P,1}(\Omega; \Delta)$ satisfies

$$\|u - u_p\|_{H^l(\Omega)} \leq C \inf_{w \in S_D^{P,1}(\Omega; \Delta)} \|u - w\|_{H^l(\Omega)} \leq C(p+1)^{-(k-1)} \|u\|_{H^k(\Omega)},$$

which together with (4.2) leads to the convergence of the p -version of FEM.

Remark 4.2. For the sake of simplicity, we prove the theorem only for the p -version with uniform degree for problems with homogeneous Dirichlet boundary conditions, but the result of the theorem and the techniques in the proof can be generalized to the p -version with quasi-uniform degree distributions for problems with homogeneous and nonhomogeneous Dirichlet boundary conditions [15] and the h - p version [18] with quasi-uniform meshes and quasi-uniform degree distribution.

REFERENCES

- [1] I. BABUŠKA AND B. GUO, *Direct and inverse approximation theorems of the p -version of the finite element method in the framework of weighted Besov spaces. Part 1: Approximability of functions in the weighted Besov spaces*, SIAM J. Numer. Anal., 39 (2002), pp. 1512–1538.
- [2] I. BABUŠKA AND B. GUO, *Direct and inverse approximation theorems of the p -version of the finite element method in the framework of weighted Besov spaces, part 2: Optimal convergence of the p -version of the finite element method*, Math. Models Methods Appl. Sci., 12 (2002), pp. 689–719.
- [3] I. BABUŠKA, A. CRAIG, J. MANDEL, AND J. PITKÄRANTA, *Efficient preconditioning for the p -version finite element method in two dimensions*, SIAM J. Numer. Anal., 28 (1991), pp. 624–661.
- [4] I. BABUŠKA, R. KELLOGG, AND J. PITKÄRANTA, *Direct and inverse error estimates for finite elements with mesh refinements*, Numer. Math., 33 (1979), pp. 447–471.
- [5] I. BABUŠKA AND M. SURI, *The h - p version of the finite element method with quasiuniform meshes*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 199–238.
- [6] I. BABUŠKA AND M. SURI, *The optimal convergence rate of the p -version of the finite element method*, SIAM J. Numer. Anal., 24 (1987), pp. 750–776.
- [7] F. B. BELGACEM, *Polynomial extensions of compatible polynomial traces in three dimensions*, Comput. Methods Appl. Mech. Engrg., 116 (1994), pp. 235–241.
- [8] C. BERNARDI, M. DAUGE, AND Y. MADAY, *Polynomials in the Sobolev World*, Version 2, 2007, Institut de Recherche Mathématique de Rennes (IRMAR), Université Rennes I and Laboratoire Jacques-Louis Lions (LJLL), Paris VI, France, preprint.
- [9] C. BERNARDI AND Y. MADAY, *Relèvement polynomial de traces et applications*, Math. Anal. Numér., 24 (1990), pp. 557–611.
- [10] C. CANUTO AND D. FUNARO, *The Schwarz algorithm for spectral methods*, SIAM J. Numer. Anal., 25 (1988), pp. 24–40.
- [11] C. CANUTO AND A. QUARTERONI, *Approximation results for orthogonal polynomial in Sobolev spaces*, Math. Comp., 38 (1982), pp. 67–86.
- [12] M. COSTABEL, M. DAUGE, AND L. DEMCKOWICZ, *Polynomial Extension Operators in $h^1, h(\text{curl})$ and $h(\text{div})$ -Spaces in a Cube*, Math. Comp., 77 (2008), pp. 1967–1999.
- [13] P. DAVIS AND P. RABINOWITZ, *Methods of Numerical Integration*, Academic Press, New York, 1975.
- [14] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Pitman Publishing, Boston, London, Melbourne, 1985.
- [15] B. GUO, *Approximation theory of the p -version of the finite element method in three dimensions, part 2: Convergence of the p -version*, SIAM J. Numer., to appear.
- [16] B. GUO AND W. SUN, *The optimal convergence of the h - p version of the finite element method with quasi-uniform meshes*, SIAM J. Numer. Anal., 45 (2007), pp. 698–730.
- [17] B. GUO AND J. ZHANG, *Constructive Proof of Polynomial Extensions in Two Dimensions*, preprint, 2006.
- [18] B. GUO AND J. ZHANG, *The h - p version of the finite element method in three dimensions with quasi uniform meshes*, in preparation.

- [19] J. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, Springer, New York, 1972.
- [20] Y. MADAY, *Relèvements de traces polynomiales et interpolations Hilbertiennes entre espaces de polynômes*, C. R. Acad. Sci. Paris Sér. I Math., 309 (1989), pp. 463–468.
- [21] R. MUÑOZ-SOLA, *Polynomial liftings on a tetrahedron and applications to the h - p version of the finite element method in three dimensions*, SIAM J. Numer. Anal., 34 (1997), pp. 282–314.
- [22] J. ZHANG, *The h - p Version of the Finite Element Method in Three Dimensions*, Ph.D. thesis, Department of Mathematics, University of Manitoba, Winnipeg, 2008.

**MIXED FINITE ELEMENT METHODS FOR THE FULLY
NONLINEAR MONGE–AMPÈRE EQUATION BASED ON THE
VANISHING MOMENT METHOD***

XIAOBING FENG[†] AND MICHAEL NEILAN[†]

Abstract. This paper studies mixed finite element approximations of the viscosity solution to the Dirichlet problem for the fully nonlinear Monge–Ampère equation $\det(D^2u^0) = f (> 0)$ based on the vanishing moment method which was proposed recently by the authors in [X. Feng and M. Neilan, *J. Scient. Comp.*, DOI 10.1007/s10915-008-9221-9, 2008]. In this approach, the second-order fully nonlinear Monge–Ampère equation is approximated by the fourth order quasilinear equation $-\varepsilon\Delta^2u^\varepsilon + \det D^2u^\varepsilon = f$. It was proved in [X. Feng, *Trans. AMS*, submitted] that the solution u^ε converges to the unique convex viscosity solution u^0 of the Dirichlet problem for the Monge–Ampère equation. This result then opens a door for constructing convergent finite element methods for the fully nonlinear second-order equations, a task which has been impracticable before. The goal of this paper is threefold. First, we develop a family of Hermann–Miyoshi-type mixed finite element methods for approximating the solution u^ε of the regularized fourth-order problem, which computes simultaneously u^ε and the moment tensor $\sigma^\varepsilon := D^2u^\varepsilon$. Second, we derive error estimates, which track explicitly the dependence of the error constants on the parameter ε , for the errors $u^\varepsilon - u_h^\varepsilon$ and $\sigma^0 - \sigma_h^\varepsilon$. Finally, we present a detailed numerical study on the rates of convergence in terms of powers of ε for the error $u^0 - u_h^\varepsilon$ and $\sigma^\varepsilon - \sigma_h^\varepsilon$, and numerically examine what is the “best” mesh size h in relation to ε in order to achieve these rates. Due to the strong nonlinearity of the underlying equation, the standard perturbation argument for error analysis of finite element approximations of nonlinear problems does not work for the problem. To overcome the difficulty, we employ a fixed point technique which strongly relies on the stability of the linearized problem and its mixed finite element approximations.

Key words. fully nonlinear PDEs, Monge–Ampère type equations, moment solutions, vanishing moment method, viscosity solutions, mixed finite element methods, Hermann–Miyoshi element

AMS subject classifications. 65N30, 65M60, 35J60, 53C45

DOI. 10.1137/070710378

1. Introduction. This paper is the second in a sequence (cf. [19]) which concerns finite element approximations of viscosity solutions of the following Dirichlet problem for the fully nonlinear Monge–Ampère equation (cf. [22]):

$$(1.1) \quad \det(D^2u^0) = f \quad \text{in } \Omega \subset \mathbf{R}^n,$$

$$(1.2) \quad u^0 = g \quad \text{on } \partial\Omega,$$

where Ω is a convex domain with smooth boundary $\partial\Omega$. $D^2u^0(x)$ and $\det(D^2u^0(x))$ denote the Hessian of u^0 at $x \in \Omega$ and the determinant of $D^2u^0(x)$.

The Monge–Ampère equation is a prototype of fully nonlinear second-order PDEs which have a general form

$$(1.3) \quad F(D^2u^0, Du^0, u^0, x) = 0$$

with $F(D^2u^0, Du^0, u^0, x) = \det(D^2u^0) - f$. The Monge–Ampère equation arises naturally from differential geometry and from applications such as mass transportation,

*Received by the editors December 10, 2007; accepted for publication (in revised form) October 7, 2008; published electronically February 25, 2009. This work was partially supported by NSF grants DMS-0410266 and DMS-0710831.

<http://www.siam.org/journals/sinum/47-2/71037.html>

[†]Department of Mathematics, The University of Tennessee, Knoxville, TN 37996 (xfeng@math.utk.edu, neilan@math.utk.edu).

meteorology, and geostrophic fluid dynamics [4, 8]. It is well known that, for non-strictly convex domain Ω , the above problem does not have classical solutions in general even when f, g , and $\partial\Omega$ are smooth (see [21]). Classical result of Aleksandrov states that the Dirichlet problem with $f > 0$ has a unique generalized solution in the class of convex functions (cf. [1, 9]). Major progress on the analysis of problems (1.1)–(1.2) has been made later after the introduction and establishment of the viscosity solution theory (cf. [7, 12, 22]). We recall that the notion of viscosity solutions was first introduced by Crandall and Lions [11] in 1983 for the first-order fully nonlinear Hamilton–Jacobi equations. It was quickly extended to second-order fully nonlinear PDEs, with dramatic consequences in the wake of a breakthrough of Jensen’s maximum principle [24] and the Ishii’s discovery [23] that the classical Perron’s method could be used to infer existence of viscosity solutions. To continue our discussion, we need to recall the definition of viscosity solutions for the Dirichlet Monge–Ampère problem (1.1)–(1.2) (cf. [22]).

DEFINITION 1.1. *A convex function $u^0 \in C^0(\overline{\Omega})$ satisfying $u^0 = g$ on $\partial\Omega$ is called a viscosity subsolution (resp., viscosity supersolution) of (1.1) if for any $\varphi \in C^2$ there holds $\det(D^2\varphi(x_0)) \geq f(x_0)$ (resp., $\det(D^2\varphi(x_0)) \leq f(x_0)$) provided that $u^0 - \varphi$ has a local maximum (resp., a local minimum) at $x_0 \in \Omega$. $u^0 \in C^0(\overline{\Omega})$ is called a viscosity solution if it is both a viscosity subsolution and a viscosity supersolution.*

It is clear that the notion of viscosity solutions is not variational. It is based on a “*differentiation by parts*” approach, instead of the more familiar integration by parts approach. As a result, it is not possible to directly approximate viscosity solutions using Galerkin type numerical methods such as finite element, spectral, and discontinuous Galerkin methods, which all are based on variational formulations of PDEs. The situation also presents a big challenge and paradox for the numerical PDE community, since, on one hand, the “*differentiation by parts*” approach has worked remarkably well for establishing the viscosity solution theory for fully nonlinear second-order PDEs in the past two decades; on the other hand, it is extremely difficult (if all possible) to mimic this approach at the discrete level. It should be noted that, unlike in the case of fully nonlinear first-order PDEs, the terminology “viscosity solution” loses its original meaning in the case of fully nonlinear second-order PDEs.

Motivated by this difficulty and by the goal of developing convergent Galerkin type numerical methods for fully nonlinear second-order PDEs, very recently we proposed in [18] a new notion of weak solutions, called *moment solutions*, which is defined using a constructive method, called the *vanishing moment method*. The main idea of the vanishing moment method is to approximate a fully nonlinear second-order PDE by a quasilinear higher order PDE. The notion of moment solutions and the vanishing moment method are natural generalizations of the original definition of viscosity solutions and the vanishing viscosity method introduced for the Hamilton–Jacobi equations in [11]. We now briefly recall the definitions of moment solutions and the vanishing moment method, and refer the reader to [16, 18] for a detailed exposition.

The first step of the vanishing moment method is to approximate the fully nonlinear (1.3) by the following quasilinear fourth-order PDE:

$$(1.4) \quad -\varepsilon\Delta^2 u^\varepsilon + F(D^2 u^\varepsilon, Du^\varepsilon, u^\varepsilon, x) = 0 \quad (\varepsilon > 0),$$

which holds in domain Ω . Suppose the Dirichlet boundary condition $u^0 = g$ is prescribed on the boundary $\partial\Omega$, then it is natural to impose the same boundary condition

on u^ε , that is,

$$(1.5) \quad u^\varepsilon = g \quad \text{on } \partial\Omega.$$

However, boundary condition (1.5) alone is not sufficient to ensure uniqueness for fourth-order PDEs. An additional boundary condition must be imposed. In [16] the authors proposed to use one of the following (extra) boundary conditions:

$$(1.6) \quad \Delta u^\varepsilon = \varepsilon, \quad \text{or} \quad D^2 u^\varepsilon \nu \cdot \nu = \varepsilon \quad \text{on } \partial\Omega,$$

where ν stands for the unit outward normal to $\partial\Omega$. Although both boundary conditions work well numerically, the first boundary condition $\Delta u^\varepsilon = \varepsilon$ is more convenient for standard finite element methods, spectral, and discontinuous Galerkin methods (cf. [19]), while the second boundary condition $D^2 u^\varepsilon \nu \cdot \nu = \varepsilon$ fits better for mixed finite element methods, and hence, it will be used in this paper.

In summary, the vanishing moment method involves approximating second-order boundary value problem (1.2)–(1.3) by fourth-order boundary value problems (1.4), (1.5), and (1.6). In the case of the Monge–Ampère equation, this means that we approximate boundary value problem (1.1)–(1.2) by the following problem:

$$(1.7) \quad -\varepsilon \Delta^2 u^\varepsilon + \det(D^2 u^\varepsilon) = f \quad \text{in } \Omega,$$

$$(1.8) \quad u^\varepsilon = g \quad \text{on } \partial\Omega,$$

$$(1.9) \quad D^2 u^\varepsilon \nu \cdot \nu = \varepsilon \quad \text{on } \partial\Omega.$$

It was proved in [16] that, if $f > 0$ in Ω , then problem (1.7)–(1.9) has a unique solution u^ε which is a strictly convex function over Ω . Moreover, u^ε uniformly converges as $\varepsilon \rightarrow 0$ to the unique viscosity solution of (1.1)–(1.2). As a result, this shows that (1.1)–(1.2) possesses a unique moment solution that coincides with the unique viscosity solution. Furthermore, it was proved that there hold the following a priori bounds which will be used frequently later in this paper:

$$(1.10) \quad \|u^\varepsilon\|_{H^j} = O\left(\varepsilon^{-\frac{j-1}{2}}\right), \quad \|u^\varepsilon\|_{W^{2,\infty}} = O(\varepsilon^{-1}),$$

$$(1.11) \quad \|D^2 u^\varepsilon\|_{L^2} = O\left(\varepsilon^{-\frac{1}{2}}\right), \quad \|\text{cof}(D^2 u^\varepsilon)\|_{L^\infty} = O(\varepsilon^{-1})$$

for $j = 2, 3$, where $\text{cof}(D^2 u^\varepsilon)$ denotes the cofactor matrix of the Hessian, $D^2 u^\varepsilon$.

With the help of the vanishing moment methodology, the original difficult task of computing the unique convex viscosity solution of the fully nonlinear Monge–Ampère problem (1.1)–(1.2), which has multiple solutions (i.e., there are nonconvex solutions), is now reduced to a feasible task of computing the unique regular solution of the quasilinear fourth-order problem (1.7)–(1.9). This then opens a door to let one use and/or adapt the wealthy amount of existing numerical methods, in particular, finite element Galerkin methods to solve the original problem (1.1)–(1.2) via the problem (1.7)–(1.9).

The goal of this paper is to construct and analyze a class of Hermann–Miyoshi-type mixed finite element methods for approximating the solution of (1.7)–(1.9). In particular, we are interested in deriving error bounds that exhibit explicit dependence on ε . We like to point out that one of our motivations for developing mixed finite element methods for (1.7)–(1.9) is that our experience in [19] tells us that Galerkin methods are numerically expensive for solving the singularly perturbed problem (1.7)–(1.9) (see [18] for a detailed numerical study). Finite element approximations of fourth-

order PDEs, in particular, the biharmonic equation, were carried out extensively in the 1970s in the two-dimensional case (see [10] and the references therein), and have attracted renewed interest lately for generalizing the well known 2-D finite elements to the 3-D case (cf. [33, 34, 32]) and for developing discontinuous Galerkin methods in all dimensions (cf. [17, 26]). Clearly, all these methods can be readily adapted to discretize problem (1.7)–(1.9) although their convergence analysis do not come easy due to the strong nonlinearity of the PDE (1.7). We refer the reader to [19, 27] for further discussions in this direction.

A few attempts and results on numerical approximations of the Monge–Ampère as well as related equations have recently been reported in the literature. Oliker and Prussner [29] constructed a finite difference scheme for computing Aleksandrov measure induced by D^2u in 2-D and obtained the solution u of problem (1.7)–(1.9) as a by-product. Baginski and Whitaker [2] proposed a finite difference scheme for Gauss curvature equation (cf. [18] and the references therein) in 2-D by mimicking the unique continuation method (used to prove existence of the PDE) at the discrete level. In a series of papers (cf. [13] and the references therein) Dean and Glowinski proposed an augmented Lagrange multiplier method and a least squares method for problem (1.7)–(1.9) and the Pucci’s equation (cf. [7, 21]) in 2-D by treating the Monge–Ampère equation and Pucci’s equation as a constraint and using a variational criterion to select a particular solution. Very recently, Oberman [28] constructed some wide stencil finite difference schemes which fulfill the convergence criterion established by Barles and Souganidis in [3] for finite difference approximations of fully nonlinear second order PDEs. Consequently, the convergence of the proposed wide stencil finite difference scheme immediately follows from the general convergence framework of [3]. Numerical experiments results were reported in [29, 28, 2, 13]; however, convergence analysis was not addressed except in [28].

The remainder of this paper is organized as follows. In section 2, we first derive the Hermann–Miyoshi mixed weak formulation for problem (1.7)–(1.9) and then present our mixed finite element methods based on this weak formulation. Section 3 is devoted to studying the linearization of problem (1.7)–(1.9) and its mixed finite element approximations. The results of this section, which are of independent interests in themselves, will play a crucial role in our error analysis for the mixed finite element introduced in section 2. In section 4, we establish error estimates in the $H^1 \times L^2$ -norm for the mixed finite element solution $(u_h^\varepsilon, \sigma_h^\varepsilon)$. Our main ideas are to use a fixed point technique and to make strong use of the stability property of the linearized problem and its finite element approximations, which all are established in section 3. In addition, we derive the optimal order error estimate in the H^1 -norm for $u^\varepsilon - u_h^\varepsilon$ using a duality argument. Finally, in section 5, we first run some numerical tests to validate our theoretical error estimate results, and we then present a detailed computational study for determining the “best” choice of mesh size h in terms of ε in order to achieve the optimal rates of convergence, and for estimating the rates of convergence for both $u^0 - u_h^\varepsilon$ and $u^0 - u^\varepsilon$ in terms of powers of ε .

We conclude this section by remarking that standard space notations are adopted in this paper; we refer to [5, 21, 10] for their exact definitions. In addition, Ω denotes a bounded domain in \mathbf{R}^n for $n = 2, 3$. (\cdot, \cdot) and $\langle \cdot, \cdot \rangle$ denote the L^2 -inner products on Ω and on $\partial\Omega$, respectively. For a Banach space B , its dual space is denoted by B^* . C is used to denote a generic ε -independent positive constant.

2. Formulation of mixed finite element methods. There are several popular mixed formulations for fourth-order problems (cf. [6, 10, 15]). However, since the

Hessian matrix, D^2u^ε , appears in (1.7) in a nonlinear fashion, we cannot use Δu^ε alone as our additional variable, but rather we are forced to use $\sigma^\varepsilon := D^2u^\varepsilon$ as a new variable. Because of this, we rule out the family of Ciarlet–Raviart mixed finite elements (cf. [10]). On the other hand, this observation suggests to try Hermann–Miyoshi or Hermann–Johnson mixed elements (cf. [6, 15, 30, 31]), which both seek σ^ε as an additional unknown. In this paper, we shall only focus on developing Hermann–Miyoshi-type mixed methods.

We begin with a few more space notations:

$$\begin{aligned} V &:= H^1(\Omega), & W &:= \left\{ \mu \in [H^1(\Omega)]^{n \times n}; \mu_{ij} = \mu_{ji} \right\}, \\ V_0 &:= H_0^1(\Omega), & V_g &:= \{v \in V; v|_{\partial\Omega} = g\}, \\ W_\varepsilon &:= \{\mu \in W; \mu\nu \cdot \nu|_{\partial\Omega} = \varepsilon\}, & W_0 &:= \{\mu \in W; \mu\nu \cdot \nu|_{\partial\Omega} = 0\}. \end{aligned}$$

To define the Hermann–Miyoshi mixed formulation for problem (1.7)–(1.9), we rewrite the PDE into the following system of second-order equations:

$$\begin{aligned} (2.1) \quad & \sigma^\varepsilon - D^2u^\varepsilon = 0, \\ (2.2) \quad & -\varepsilon\Delta\text{tr}(\sigma^\varepsilon) + \det(\sigma^\varepsilon) = f. \end{aligned}$$

Testing (2.2) with $v \in V_0$ yields

$$(2.3) \quad \varepsilon \int_\Omega \text{div}(\sigma^\varepsilon) \cdot Dv \, dx + \int_\Omega \det(\sigma^\varepsilon)v \, dx = \int_\Omega f v \, dx.$$

Multiplying (2.1) by $\mu \in W_0$ and integrating over Ω we get

$$(2.4) \quad \int_\Omega \sigma^\varepsilon : \mu \, dx + \int_\Omega Du^\varepsilon \cdot \text{div}(\mu) \, dx = \sum_{k=1}^{n-1} \int_{\partial\Omega} \mu\nu \cdot \tau_k \frac{\partial g}{\partial \tau_k} \, ds,$$

where $\sigma^\varepsilon : \mu$ denotes the matrix inner product and $\{\tau_1(x), \tau_2(x), \dots, \tau_{n-1}(x)\}$ denotes the standard basis for the tangent space to $\partial\Omega$ at x .

From (2.3) and (2.4), we define the variational formulation for (2.1)–(2.2) as follows: Find $(u^\varepsilon, \sigma^\varepsilon) \in V_g \times W_\varepsilon$ such that

$$(2.5) \quad (\sigma^\varepsilon, \mu) + (\text{div}(\mu), Du^\varepsilon) = \langle \tilde{g}, \mu \rangle \quad \forall \mu \in W_0,$$

$$(2.6) \quad (\text{div}(\sigma^\varepsilon), Dv) + \frac{1}{\varepsilon} (\det\sigma^\varepsilon, v) = (f^\varepsilon, v) \quad \forall v \in V_0,$$

where

$$\langle \tilde{g}, \mu \rangle = \sum_{i=1}^{n-1} \left\langle \frac{\partial g}{\partial \tau_i}, \mu\nu \cdot \tau_i \right\rangle \quad \text{and} \quad f^\varepsilon = \frac{1}{\varepsilon} f.$$

To discretize (2.5)–(2.6), let T_h be a quasiuniform triangular or rectangular partition of Ω if $n = 2$ and be a quasiuniform tetrahedral or 3-D rectangular mesh if $n = 3$. Let $V^h \subset H^1(\Omega)$ be the Lagrange finite element space consisting of continuous piecewise polynomials of degree $k(\geq 2)$ associated with the mesh T_h . Let

$$\begin{aligned} V_g^h &:= V^h \cap V_g, & V_0^h &:= V^h \cap V_0, \\ W_\varepsilon^h &:= [V^h]^{n \times n} \cap W_\varepsilon, & W_0^h &:= [V^h]^{n \times n} \cap W_0. \end{aligned}$$

In the 2-D case, the above choices of V_0^h and W_0^h are known as the Hermann–Miyoshi mixed finite element for the biharmonic equation (cf. [6, 15]). They form a stable pair which satisfies the inf-sup condition. We like to note that it is easy to check that the Hermann–Miyoshi mixed finite element also satisfies the inf-sup condition in 3-D. See section 3.2 for the details.

Based on the weak formulation (2.5)–(2.6) and using the above finite element spaces, we now define our Hermann–Miyoshi-type mixed finite element method for (1.7)–(1.9) as follows: Find $(u_h^\varepsilon, \sigma_h^\varepsilon) \in V_g^h \times W_\varepsilon^h$ such that

$$(2.7) \quad (\sigma_h^\varepsilon, \mu_h) + (\operatorname{div}(\mu_h), Du_h^\varepsilon) = \langle \tilde{g}, \mu_h \rangle \quad \forall \mu_h \in W_0^h,$$

$$(2.8) \quad (\operatorname{div}(\sigma_h^\varepsilon), Dv_h) + \frac{1}{\varepsilon} (\det(\sigma_h^\varepsilon), v_h) = (f^\varepsilon, v_h) \quad \forall v_h \in V_0^h.$$

Let $(\sigma^\varepsilon, u^\varepsilon)$ be the solution to (2.5)–(2.6) and $(\sigma_h^\varepsilon, u_h^\varepsilon)$ solves (2.7)–(2.8). As mentioned in section 1, the primary goal of this paper is to derive error estimates for $u^\varepsilon - u_h^\varepsilon$ and $\sigma^\varepsilon - \sigma_h^\varepsilon$. To this end, we first need to prove existence and uniqueness of $(\sigma_h^\varepsilon, u_h^\varepsilon)$. It turns out both tasks are not easy to accomplish due to the strong nonlinearity in (2.8). Unlike in the continuous PDE case, where u^ε is proved to be convex for all ε (cf. [16]), it is far from clear if u_h^ε preserves the convexity even for small ε and h . Without a guarantee of convexity for u_h^ε , we could not establish any stability result for u_h^ε . This, in turn, makes proving existence and uniqueness a difficult and delicate task. In addition, again due to the strong nonlinearity, the standard perturbation technique for deriving error estimate for numerical approximations of mildly nonlinear problems does not work here. To overcome the difficulty, our idea is to adopt a combined fixed point and linearization technique which was used by the authors in [20], where a nonlinear singular second-order problem known as the inverse mean curvature flow was studied. We note that this combined fixed point and linearization technique kills three birds by one stone, that is, it simultaneously proves existence and uniqueness for u_h^ε and also yields the desired error estimates. In the next two sections, we shall give a detailed account about the technique and realize it for problem (2.7)–(2.8).

3. Linearized problem and its finite element approximations. To build the necessary technical tools, in this section we shall derive and present a detailed study of the linearization of (2.5)–(2.6) and its mixed finite element approximations. First, we recall the following divergence-free row property for the cofactor matrices, which will be frequently used in later sections. We refer to [14, p. 440] for a short proof of the lemma.

LEMMA 3.1. *Given a vector-valued function $\mathbf{v} = (v_1, v_2, \dots, v_n) : \Omega \rightarrow \mathbf{R}^n$. Assume $\mathbf{v} \in [C^2(\Omega)]^n$. Then the cofactor matrix $\operatorname{cof}(D\mathbf{v})$ of the gradient matrix $D\mathbf{v}$ of \mathbf{v} satisfies the following row divergence-free property:*

$$(3.1) \quad \operatorname{div}(\operatorname{cof}(D\mathbf{v}))_i = \sum_{j=1}^n \partial_{x_j} (\operatorname{cof}(D\mathbf{v}))_{ij} = 0 \quad \text{for } i = 1, 2, \dots, n,$$

where $(\operatorname{cof}(D\mathbf{v}))_i$ and $(\operatorname{cof}(D\mathbf{v}))_{ij}$ denote, respectively, the i th row and the (i, j) -entry of $\operatorname{cof}(D\mathbf{v})$.

3.1. Derivation of linearized problem. We note that for a given function w there holds

$$\det(D^2(u^\varepsilon + tw)) = \det(D^2u^\varepsilon) + t \operatorname{tr}(\Phi^\varepsilon D^2w) + \dots + t^n \det(D^2w),$$

where $\Phi^\varepsilon := \text{cof}(D^2u^\varepsilon)$. Thus, setting $t = 0$ after differentiating with respect to t we find the linearization of $M^\varepsilon(u^\varepsilon) := -\varepsilon\Delta^2u^\varepsilon + \det(D^2u^\varepsilon)$ at the solution u^ε to be

$$L_{u^\varepsilon}(w) := -\varepsilon\Delta^2w + \text{tr}(\Phi^\varepsilon D^2w) = -\varepsilon\Delta^2w + \Phi^\varepsilon : D^2w = -\varepsilon\Delta^2w + \text{div}(\Phi^\varepsilon Dw),$$

where we have used (3.1) with $\mathbf{v} = Du^\varepsilon$.

We now consider the following linear problem:

$$(3.2) \quad L_{u^\varepsilon}(w) = q \quad \text{in } \Omega,$$

$$(3.3) \quad w = 0 \quad \text{on } \partial\Omega,$$

$$(3.4) \quad D^2w\nu \cdot \nu = 0 \quad \text{on } \partial\Omega.$$

To introduce a mixed formulation for (3.2)–(3.4), we rewrite the PDE as

$$(3.5) \quad \chi - D^2w = 0,$$

$$(3.6) \quad -\varepsilon\Delta\text{tr}(\chi) + \text{div}(\Phi^\varepsilon Dw) = q.$$

Its variational formulation is then defined as: Given $q \in V_0^*$, find $(\chi, w) \in W_0 \times V_0$ such that

$$(3.7) \quad (\chi, \mu) + (\text{div}(\mu), Dw) = 0 \quad \forall \mu \in W_0,$$

$$(3.8) \quad (\text{div}(\chi), Dv) - \frac{1}{\varepsilon}(\Phi^\varepsilon Dw, Dv) = \frac{1}{\varepsilon}\langle q, v \rangle \quad \forall v \in V_0.$$

It is not hard to show that if (χ, w) solves (3.7)–(3.8), then $w \in H^2(\Omega) \cap H_0^1(\Omega)$ should be a weak solution to problem (3.2)–(3.4). On the other hand, by the elliptic theory for linear PDEs (cf. [25]), we know that if $q \in V_0^*$, then the solution to problem (3.2)–(3.4) satisfies $w \in H^3(\Omega)$, so that $\chi = D^2w \in H^1(\Omega)$. It is easy to verify that (w, χ) is a solution to (3.7)–(3.8).

3.2. Mixed finite element approximations of the linearized problem.

Our finite element method for (3.7)–(3.8) is defined by seeking $(\chi_h, w_h) \in W_0^h \times V_0^h$ such that

$$(3.9) \quad (\chi_h, \mu_h) + (\text{div}(\mu_h), Dw_h) = 0 \quad \forall \mu_h \in W_0^h,$$

$$(3.10) \quad (\text{div}(\chi_h), Dv_h) - \frac{1}{\varepsilon}(\Phi^\varepsilon Dw_h, Dv_h) = \langle q, v_h \rangle \quad \forall v_h \in V_0^h.$$

The objectives of this subsection are to first prove existence and uniqueness for problem (3.9)–(3.10) and then derive error estimates in various norms. First, we prove the following inf-sup condition for the mixed finite element pair (W_0^h, V_0^h) .

LEMMA 3.2. *For every $v_h \in V_0^h$, there exists a constant $\beta_0 > 0$, independent of h , such that*

$$(3.11) \quad \sup_{\mu_h \in W_0^h} \frac{(\text{div}(\mu_h), Dv_h)}{\|\mu_h\|_{H^1}} \geq \beta_0 \|v_h\|_{H^1}.$$

Proof. Given $v_h \in V_0^h$, set $\mu_h = I_{n \times n} v_h$. Then $(\text{div}(\mu_h), Dv_h) = \|Dv_h\|_{L^2}^2 \geq \beta_0 \|v_h\|_{H^1}^2 = \beta_0 \|v_h\|_{H^1} \|\mu_h\|_{H^1}$. Here we have used Poincaré inequality. \square

Remark 3.1. By [15, Proposition 1], (3.11) implies that there exists a linear operator $\Pi_h : W \rightarrow W^h$ such that

$$(3.12) \quad (\text{div}(\mu - \Pi_h \mu), Dv_h) = 0 \quad \forall v_h \in V_0^h,$$

and for $\mu \in W \cap [H^r(\Omega)]^{n \times n}$, $r \geq 1$, there holds

$$(3.13) \quad \|\mu - \Pi_h \mu\|_{H^j} \leq Ch^{l-j} \|\mu\|_{H^l} \quad j = 0, 1, \quad 1 \leq l \leq \min\{k + 1, r\}.$$

We note that the above results were proved in the 2-D case in [15]; however, they also hold in the 3-D case as (3.11) holds in 3-D.

THEOREM 3.1. *For any $q \in V_0^*$, there exists a unique solution $(\chi_h, w_h) \in W_0^h \times V_0^h$ to problem (3.9)–(3.10).*

Proof. Since we are in the finite dimensional case and the problem is linear, it suffices to show uniqueness. Thus, suppose $(\chi_h, w_h) \in W_0^h \times V_0^h$ solves

$$\begin{aligned} (\chi_h, \mu_h) + (\operatorname{div}(\mu_h), Dw_h) &= 0 & \forall \mu_h \in W_0^h, \\ (\operatorname{div}(\chi_h), Dv_h) - \frac{1}{\varepsilon}(\Phi^\varepsilon Dw_h, Dv_h) &= 0 & \forall v_h \in V_0^h. \end{aligned}$$

Let $\mu_h = \chi_h$, $v_h = w_h$, and subtract two equations to obtain

$$(\chi_h, \chi_h) + \frac{1}{\varepsilon}(\Phi^\varepsilon Dw_h, Dw_h) = 0.$$

Since u^ε is strictly convex, then Φ^ε is positive definite. Thus, there exists $\theta > 0$ such that

$$\|\chi_h\|_{L^2}^2 + \frac{\theta}{\varepsilon} \|Dw_h\|_{L^2}^2 \leq 0.$$

Hence, $\chi_h = 0$, $w_h = 0$, and the desired result follows. \square

THEOREM 3.2. *Let $(\chi, w) \in [H^r(\Omega)]^{n \times n} \cap W_0 \times H^r(\Omega) \cap V_0$ ($r \geq 2$) be the solution to (3.7)–(3.8) and $(\chi_h, w_h) \in W_0^h \times V_0^h$ solves (3.9)–(3.10). Then there hold*

$$(3.14) \quad \|\chi - \chi_h\|_{L^2} \leq C\varepsilon^{-\frac{3}{2}} h^{l-2} [\|\chi\|_{H^l} + \|w\|_{H^l}],$$

$$(3.15) \quad \|\chi - \chi_h\|_{H^1} \leq C\varepsilon^{-\frac{3}{2}} h^{l-3} [\|\chi\|_{H^l} + \|w\|_{H^l}],$$

$$(3.16) \quad \|w - w_h\|_{H^1} \leq C\varepsilon^{-3} h^{l-1} [\|\chi\|_{H^l} + \|w\|_{H^l}],$$

where $l := \min\{k + 1, r\}$. Moreover, for $k \geq 3$ there also holds

$$(3.17) \quad \|w - w_h\|_{L^2} \leq C\varepsilon^{-5} h^l [\|\chi\|_{H^l} + \|w\|_{H^l}].$$

Proof. Let $I_h w$ denote the standard finite element interpolant of w in V_0^h . Then

$$(3.18) \quad \begin{aligned} (\Pi_h \chi - \chi_h, \mu_h) + (\operatorname{div}(\mu_h), D(I_h w - w_h)) \\ = (\Pi_h \chi - \chi, \mu_h) + (\operatorname{div}(\mu_h), D(I_h w - w)), \end{aligned}$$

$$(3.19) \quad \begin{aligned} (\operatorname{div}(\Pi_h \chi - \chi_h), Dv_h) - \frac{1}{\varepsilon}(\Phi^\varepsilon D(I_h w - w_h), Dv_h) \\ = -\frac{1}{\varepsilon}(\Phi^\varepsilon D(I_h w - w), Dv_h). \end{aligned}$$

Let $\mu_h = \Pi_h - \chi_h$ and $v_h = I_h w - w_h$ and subtract (3.19) from (3.18) to get

$$\begin{aligned} (\Pi_h \chi - \chi_h, \Pi_h \chi - \chi_h) + \frac{1}{\varepsilon}(\Phi^\varepsilon D(I_h w - w_h), D(I_h w - w_h)) \\ = (\Pi_h \chi - \chi, \Pi_h \chi - \chi_h) + (\operatorname{div}(\Pi_h \chi - \chi_h), D(I_h w - w)) \\ + \frac{1}{\varepsilon}(\Phi^\varepsilon D(I_h w - w), D(I_h w - w_h)). \end{aligned}$$

Thus,

$$\begin{aligned}
& \|\Pi_h \chi - \chi_h\|_{L^2}^2 + \frac{\theta}{\varepsilon} \|D(I_h w - w_h)\|_{L^2}^2 \\
& \leq \|\Pi_h \chi - \chi\|_{L^2} \|\Pi_h \chi - \chi_h\|_{L^2} + \|\Pi_h \chi - \chi_h\|_{H^1} \|D(I_h w - w)\|_{L^2} \\
& \quad + \frac{C}{\varepsilon^2} \|D(I_h w - w)\|_{L^2} \|D(I_h w - w_h)\|_{L^2} \\
& \leq \|\Pi_h \chi - \chi\|_{L^2} \|\Pi_h \chi - \chi_h\|_{L^2} + Ch^{-1} \|\Pi_h \chi - \chi_h\|_{L^2} \|D(I_h w - w)\|_{L^2} \\
& \quad + \frac{C}{\varepsilon^2} \|D(I_h w - w)\|_{L^2} \|D(I_h w - w_h)\|_{L^2},
\end{aligned}$$

where we have used the inverse inequality.

Using the Schwarz inequality and rearranging terms yield

$$\begin{aligned}
(3.20) \quad & \|\Pi_h \chi - \chi_h\|_{L^2}^2 + \frac{1}{\varepsilon} \|D(I_h w - w_h)\|_{L^2}^2 \\
& \leq C (\|\Pi_h \chi - \chi\|_{L^2}^2 + h^{-2} \|I_h w - w\|_{H^1}^2 + \varepsilon^{-3} \|I_h w - w\|_{H^1}^2).
\end{aligned}$$

Hence, by the standard interpolation results [5, 10] we have

$$\begin{aligned}
\|\Pi_h \chi - \chi_h\|_{L^2} & \leq C \left(\|\Pi_h \chi - \chi\|_{L^2} + h^{-1} \|I_h w - w\|_{H^1} + \varepsilon^{-\frac{3}{2}} \|I_h w - w\|_{H^1} \right) \\
& \leq C \varepsilon^{-\frac{3}{2}} h^{l-2} (\|\chi\|_{H^l} + \|w\|_{H^l}),
\end{aligned}$$

which, by the triangle inequality, yield

$$\|\chi - \chi_h\|_{L^2} \leq C \varepsilon^{-\frac{3}{2}} h^{l-2} (\|\chi\|_{H^l} + \|w\|_{H^l}).$$

The above estimate and the inverse inequality yield

$$\begin{aligned}
\|\chi - \chi_h\|_{H^1} & \leq \|\chi - \Pi_h \chi\|_{H^1} + \|\Pi_h \chi - \chi_h\|_{H^1} \\
& \leq \|\chi - \Pi_h \chi\|_{H^1} + h^{-1} \|\Pi_h \chi - \chi_h\|_{L^2} \\
& \leq C \varepsilon^{-\frac{3}{2}} h^{l-3} (\|\chi\|_{H^l} + \|w\|_{H^l}).
\end{aligned}$$

Next, from (3.20) we have

$$\begin{aligned}
\|D(I_h w - w_h)\|_{L^2} & \leq \sqrt{\varepsilon} C \left[\|\Pi_h \chi - \chi\|_{L^2} + h^{-1} \|D(I_h w - w)\|_{L^2} + \varepsilon^{-\frac{3}{2}} \|I_h w - w\|_{H^1} \right] \\
(3.21) \quad & \leq C \varepsilon^{-1} h^{l-2} (\|\chi\|_{H^l} + \|w\|_{H^l}).
\end{aligned}$$

To derive (3.16), we appeal to a version of the Aubin–Nitsche duality argument (cf. [5, 10]). We consider the following auxiliary problem: Find $z \in H^2(\Omega) \cap H_0^1(\Omega)$ such that

$$\begin{aligned}
-\varepsilon \Delta^2 z + \operatorname{div}(\Phi^\varepsilon D z) & = -\Delta(w - w_h) & \text{in } \Omega, \\
D^2 z \nu \cdot \nu & = 0 & \text{on } \partial\Omega.
\end{aligned}$$

By the elliptic theory for linear PDEs (cf. [25]), we know that the above problem has a unique solution $z \in H_0^1(\Omega) \cap H^3(\Omega)$ and

$$(3.22) \quad \|z\|_{H^3} \leq C_b(\varepsilon) \|D(w - w_h)\|_{L^2} \quad \text{where } C_b(\varepsilon) = O(\varepsilon^{-1}).$$

Setting $\kappa = D^2z$, it is easy to verify that $(\kappa, z) \in W_0 \times V_0$ and

$$\begin{aligned} (\kappa, \mu) + (\operatorname{div}(\mu), Dz) &= 0 & \forall \mu \in W_0, \\ (\operatorname{div}(\kappa), Dv) - \frac{1}{\varepsilon}(\Phi^\varepsilon Dz, Dv) &= \frac{1}{\varepsilon}(D(w - w_h), Dv) & \forall v \in V_0. \end{aligned}$$

It is easy to check that (3.9)–(3.10) produce the following error equations:

$$(3.23) \quad (\chi - \chi_h, \mu_h) + (\operatorname{div}(\mu_h), D(w - w_h)) = 0 \quad \forall \mu_h \in W_0^h,$$

$$(3.24) \quad (\operatorname{div}(\chi - \chi_h), Dv_h) - \frac{1}{\varepsilon}(\Phi^\varepsilon D(w - w_h), Dv_h) = 0 \quad \forall v_h \in V_0^h.$$

Thus,

$$\begin{aligned} \frac{1}{\varepsilon} \|D(w - w_h)\|_{L^2}^2 &= (\operatorname{div}(\kappa), D(w - w_h)) - \frac{1}{\varepsilon}(\Phi^\varepsilon Dz, D(w - w_h)) \\ &= (\operatorname{div}(\kappa - \Pi_h \kappa), D(w - w_h)) - \frac{1}{\varepsilon}(\Phi^\varepsilon Dz, D(w - w_h)) \\ &\quad + (\operatorname{div}(\Pi_h \kappa), D(w - w_h)) \\ &= (\operatorname{div}(\kappa - \Pi_h \kappa), D(w - I_h w)) - \frac{1}{\varepsilon}(\Phi^\varepsilon Dz, D(w - w_h)) \\ &\quad + (\chi_h - \chi, \Pi_h \kappa) \\ &= (\operatorname{div}(\kappa - \Pi_h \kappa), D(w - I_h w)) - \frac{1}{\varepsilon}(\Phi^\varepsilon Dz, D(w - w_h)) \\ &\quad + (\chi_h - \chi, \Pi_h \kappa - \kappa) + (\chi_h - \chi, \kappa) \\ &= (\operatorname{div}(\kappa - \Pi_h \kappa), D(w - I_h w)) - \frac{1}{\varepsilon}(\Phi^\varepsilon Dz, D(w - w_h)) \\ &\quad + (\chi_h - \chi, \Pi_h \kappa - \kappa) + (\operatorname{div}(\chi - \chi_h), Dz) \\ &= (\operatorname{div}(\kappa - \Pi_h \kappa), D(w - I_h w)) + (\chi_h - \chi, \Pi_h \kappa - \kappa) \\ &\quad + (\operatorname{div}(\chi - \chi_h), D(z - I_h z)) - \frac{1}{\varepsilon}(\Phi^\varepsilon D(w - w_h), D(z - I_h z)) \\ &\leq \| \operatorname{div}(\kappa - \Pi_h \kappa) \|_{L^2} \|D(w - I_h w)\|_{L^2} + \| \chi_h - \chi \|_{L^2} \| \Pi_h \kappa - \kappa \|_{L^2} \\ &\quad + \| \operatorname{div}(\chi - \chi_h) \|_{L^2} \|D(z - I_h z)\|_{L^2} \\ &\quad + \frac{C}{\varepsilon^2} \|D(z - I_h z)\|_{L^2} \|D(w - w_h)\|_{L^2} \\ &\leq C \left[\|D(w - I_h w)\|_{L^2} + h \| \chi_h - \chi \|_{L^2} + h^2 \| \operatorname{div}(\chi - \chi_h) \|_{L^2} \right. \\ &\quad \left. + \frac{h^2}{\varepsilon^2} \|D(w - w_h)\|_{L^2} \right] \|z\|_{H^3}. \end{aligned}$$

Then, by (3.14), (3.15), (3.21), and (3.22), we have

$$\|D(w - w_h)\|_{L^2} \leq C_b(\varepsilon) \varepsilon^{-2} h^{l-1} [\| \chi \|_{H^l} + \|w\|_{H^l}].$$

Substituting $C_b(\varepsilon) = O(\varepsilon^{-1})$ we get (3.16).

To derive the L^2 -norm estimate for $w - w_h$, we consider the following auxiliary problem: Find $(\kappa, z) \in W_0 \times V_0$ such that

$$\begin{aligned} (\kappa, \mu) + (\operatorname{div}(\mu), Dz) &= 0 & \forall \mu \in W_0, \\ (\operatorname{div}(\kappa), Dv) - \frac{1}{\varepsilon}(\Phi^\varepsilon Dz, Dv) &= \frac{1}{\varepsilon}(w - w_h, v) & \forall v \in V_0. \end{aligned}$$

Assume the above problem is H^4 regular, that is, $z \in H^4(\Omega)$ and

$$(3.25) \quad \|z\|_{H^4} \leq C_b(\varepsilon) \|w - w_h\|_{L^2} \quad \text{with} \quad C_b(\varepsilon) = O(\varepsilon^{-1}).$$

We then have

$$\begin{aligned} \frac{1}{\varepsilon} \|w - w_h\|_{L^2}^2 &= (\operatorname{div}(\kappa), D(w - w_h)) - \frac{1}{\varepsilon} (\Phi^\varepsilon D(w - w_h), Dz) \\ &= (\operatorname{div}(\Pi_h \kappa), D(w - w_h)) - \frac{1}{\varepsilon} (\Phi^\varepsilon D(w - w_h), Dz) \\ &\quad + (\operatorname{div}(\kappa - \Pi_h \kappa), D(w - w_h)) \\ &= (\chi_h - \chi, \Pi_h \kappa) - \frac{1}{\varepsilon} (\Phi^\varepsilon Dz, D(w - w_h)) \\ &\quad + (\operatorname{div}(\kappa - \Pi_h \kappa), D(w - I_h w)) \\ &= (\chi_h - \chi, \kappa) + (\chi_h - \chi, \Pi_h \kappa - \kappa) \\ &\quad - \frac{1}{\varepsilon} (\Phi^\varepsilon Dz, D(w - w_h)) + (\operatorname{div}(\kappa - \Pi_h \kappa), D(w - I_h w)) \\ &= (\operatorname{div}(\chi - \chi_h), Dz) - \frac{1}{\varepsilon} (\Phi^\varepsilon D(w - w_h), Dz) \\ &\quad + (\chi_h - \chi, \Pi_h \kappa - \kappa) + (\operatorname{div}(\kappa - \Pi_h \kappa), D(w - I_h w)) \\ &= (\operatorname{div}(\chi - \chi_h), D(z - I_h z)) - \frac{1}{\varepsilon} (\Phi^\varepsilon D(w - w_h), D(z - I_h z)) \\ &\quad + (\chi_h - \chi, \Pi_h \kappa - \kappa) + (\operatorname{div}(\kappa - \Pi_h \kappa), D(w - I_h w)) \\ &\leq \left[\|\operatorname{div}(\chi - \chi_h)\|_{L^2} + \frac{C}{\varepsilon^2} \|D(w - w_h)\|_{L^2} \right] \|D(z - I_h z)\|_{L^2} \\ &\quad + \|\chi_h - \chi\|_{L^2} \|\Pi_h \kappa - \kappa\|_{L^2} + \|\operatorname{div}(\kappa - \Pi_h \kappa)\|_{L^2} \|D(w - I_h w)\|_{L^2} \\ &\leq Ch^3 \left[\|\chi - \chi_h\|_{H^1} + \frac{1}{\varepsilon^2} \|w - w_h\|_{H^1} \right] \|z\|_{H^4} \\ &\quad + Ch^2 \|\chi_h - \chi\|_{L^2} \|\kappa\|_{H^2} + Ch \|w - I_h w\|_{H^1} \|\kappa\|_{H^2} \\ &\leq C\varepsilon^{-5} h^l (\|\chi\|_{H^1} + \|w\|_{H^1}) \|z\|_{H^4} \\ &\leq CC_b(\varepsilon) \varepsilon^{-5} h^l (\|\chi\|_{H^1} + \|w\|_{H^1}) \|w - w_h\|_{L^2}, \end{aligned}$$

where we have used (3.14), (3.15), (3.16), (3.25), and the assumption $k \geq 3$. Dividing the above inequality by $\|w - w_h\|_{L^2}$ and substituting $C_b(\varepsilon) = O(\varepsilon^{-1})$ we get (3.17). The proof is complete. \square

4. Error analysis for finite element method (2.7)–(2.8). The goal of this section is to derive error estimates for the finite element method (2.7)–(2.8). Our main idea is to use a combined fixed point and linearization technique (cf. [20]).

DEFINITION 4.1. Let $T : W_\varepsilon^h \times V_g^h \rightarrow W_\varepsilon^h \times V_g^h$ be a linear mapping such that for any $(\mu_h, v_h) \in W_\varepsilon^h \times V_g^h$, $T(\mu_h, v_h) = (T^{(1)}(\mu_h, v_h), T^{(2)}(\mu_h, v_h))$ satisfies

$$(4.1) \quad \left(\mu_h - T^{(1)}(\mu_h, v_h), \kappa_h \right) + \left(\operatorname{div}(\kappa_h), D \left(v_h - T^{(2)}(\mu_h, v_h) \right) \right) \\ = (\mu_h, \kappa_h) + (\operatorname{div}(\kappa_h), Dv_h) - \langle \tilde{g}, \kappa_h \rangle \quad \forall \kappa_h \in W_0^h,$$

$$(4.2) \quad \left(\operatorname{div} \left(\mu_h - T^{(1)}(\mu_h, v_h) \right), Dz_h \right) - \frac{1}{\varepsilon} \left(\Phi^\varepsilon D \left(v_h - T^{(2)}(\mu_h, v_h) \right), Dz_h \right) \\ = (\operatorname{div}(\mu_h), Dz_h) + \frac{1}{\varepsilon} (\det(\mu_h), z_h) - (f^\varepsilon, z_h) \quad \forall z_h \in V_0.$$

By Theorem 3.1, we conclude that $T(\mu_h, v_h)$ is well defined. Clearly, any fixed point (χ_h, w_h) of the mapping T (i.e., $T(\chi_h, w_h) = (\chi_h, w_h)$) is a solution to problem (2.7)–(2.8), and vice-versa. The rest of this section shows that, indeed, the mapping T has a unique fixed point in a small neighborhood of $(I_h\sigma^\varepsilon, I_hu^\varepsilon)$. To this end, we define

$$\begin{aligned} \tilde{B}_h(\rho) &:= \left\{ (\mu_h, v_h) \in W_\varepsilon^h \times V_g^h; \|\mu_h - I_h\sigma^\varepsilon\|_{L^2} + \frac{1}{\sqrt{\varepsilon}}\|v_h - I_hu^\varepsilon\|_{H^1} \leq \rho \right\}. \\ \tilde{Z}_h &:= \left\{ (\mu_h, v_h) \in W_\varepsilon^h \times V_g^h; (\mu_h, \kappa_h) + (\operatorname{div}(\kappa_h), Dv_h) = \langle \tilde{g}, \kappa_h \rangle \forall \kappa_h \in W_0^h \right\}. \\ B_h(\rho) &:= \tilde{B}_h(\rho) \cap \tilde{Z}_h. \end{aligned}$$

We also assume $\sigma^\varepsilon \in H^r(\Omega)$ and set $l = \min\{k + 1, r\}$.

The next lemma measures the distance between the center of $B_h(\rho)$ and its image under the mapping T .

LEMMA 4.1. *The mapping T satisfies the following estimates:*

$$(4.3) \quad \left\| I_h\sigma^\varepsilon - T^{(1)}(I_h\sigma^\varepsilon, I_hu^\varepsilon) \right\|_{H^1} \leq C_1(\varepsilon)h^{l-3} [\|\sigma^\varepsilon\|_{H^l} + \|u^\varepsilon\|_{H^l}],$$

$$(4.4) \quad \left\| I_h\sigma^\varepsilon - T^{(1)}(I_h\sigma^\varepsilon, I_hu^\varepsilon) \right\|_{L^2} \leq C_2(\varepsilon)h^{l-2} [\|\sigma^\varepsilon\|_{H^l} + \|u^\varepsilon\|_{H^l}],$$

$$(4.5) \quad \left\| I_hu^\varepsilon - T^{(2)}(I_h\sigma^\varepsilon, I_hu^\varepsilon) \right\|_{H^1} \leq C_3(\varepsilon)h^{l-1} [\|\sigma^\varepsilon\|_{H^l} + \|u^\varepsilon\|_{H^l}],$$

where $C_1(\varepsilon) = O(\varepsilon^{-1})$, $C_2(\varepsilon) = O(\varepsilon^{-1})$, $C_3(\varepsilon) = O(\varepsilon^{-4})$ when $n = 2$, and $C_1(\varepsilon) = O(\varepsilon^{-\frac{5}{2}})$, $C_2(\varepsilon) = O(\varepsilon^{-\frac{5}{2}})$, $C_3(\varepsilon) = O(\varepsilon^{-\frac{11}{2}})$ when $n = 3$.

Proof. We divide the proof into four steps.

Step 1: To ease notation we set $\omega_h = I_h\sigma^\varepsilon - T^{(1)}(I_h\sigma^\varepsilon, I_hu^\varepsilon)$, $s_h = I_hu^\varepsilon - T^{(2)}(I_h\sigma^\varepsilon, I_hu^\varepsilon)$. By the definition of T , we have for any $(\mu_h, v_h) \in W_0^h \times V_0^h$

$$\begin{aligned} (\omega_h, \mu_h) + (\operatorname{div}(\mu_h), Ds_h) &= (I_h\sigma^\varepsilon, \mu_h) + (\operatorname{div}(\mu_h), D(I_hu^\varepsilon)) - \langle \tilde{g}, \mu_h \rangle, \\ (\operatorname{div}(\omega_h), Dv_h) - \frac{1}{\varepsilon}(\Phi^\varepsilon Ds_h, Dv_h) &= (\operatorname{div}(I_h\sigma^\varepsilon), Dv_h) + \frac{1}{\varepsilon}(\det(I_h\sigma^\varepsilon), v_h) - (f^\varepsilon, v_h). \end{aligned}$$

It follows from (2.5)–(2.6) that, for any $(\mu_h, v_h) \in W_0^h \times V_0^h$

$$(4.6) \quad (\omega_h, \mu_h) + (\operatorname{div}(\mu_h), Ds_h) = (I_h\sigma^\varepsilon - \sigma^\varepsilon, \mu_h) + (\operatorname{div}(\mu_h), D(I_hu^\varepsilon - u^\varepsilon)),$$

$$(4.7) \quad (\operatorname{div}(\omega_h), Dv_h) - \frac{1}{\varepsilon}(\Phi^\varepsilon Ds_h, Dv_h) = (\operatorname{div}(I_h\sigma^\varepsilon - \sigma^\varepsilon), Dv_h) + \frac{1}{\varepsilon}(\det(I_h\sigma^\varepsilon) - \det(\sigma^\varepsilon), v_h).$$

Letting $v_h = s_h$, $\mu_h = \omega_h$ in (4.6)–(4.7), subtracting the two equations and using the mean value theorem we get

$$\begin{aligned} (\omega_h, \omega_h) + \frac{1}{\varepsilon}(\Phi^\varepsilon Ds_h, Ds_h) &= (I_h\sigma^\varepsilon - \sigma^\varepsilon, \omega_h) + (\operatorname{div}(\omega_h), D(I_hu^\varepsilon - u^\varepsilon)) \\ &\quad + (\operatorname{div}(\sigma - I_h\sigma^\varepsilon), Ds_h) + \frac{1}{\varepsilon}(\det(\sigma^\varepsilon) - \det(I_h\sigma^\varepsilon), s_h) \\ &= (I_h\sigma^\varepsilon - \sigma^\varepsilon, \omega_h) + (\operatorname{div}(\omega_h), D(I_hu^\varepsilon - u^\varepsilon)) \\ &\quad + (\operatorname{div}(\sigma - I_h\sigma^\varepsilon), Ds_h) + \frac{1}{\varepsilon}(\Psi^\varepsilon : (\sigma^\varepsilon - I_h\sigma^\varepsilon), s_h), \end{aligned}$$

where $\Psi^\varepsilon = \operatorname{cof}(\tau I_h\sigma^\varepsilon + [1 - \tau]\sigma^\varepsilon)$ for $\tau \in [0, 1]$.

Step 2: The case $n = 2$. Since Ψ^ε is a 2×2 matrix whose entries are the same as those of $\tau I_h \sigma^\varepsilon + [1 - \tau] \sigma^\varepsilon$, then by (1.11) we have

$$\begin{aligned} \|\Psi^\varepsilon\|_{L^2} &= \|\text{cof}(\tau I_h \sigma^\varepsilon + [1 - \tau] \sigma^\varepsilon)\|_{L^2} = \|\tau I_h \sigma^\varepsilon + [1 - \tau] \sigma^\varepsilon\|_{L^2} \\ &\leq \|I_h \sigma^\varepsilon\|_{L^2} + \|\sigma^\varepsilon\|_{L^2} \leq C \|\sigma^\varepsilon\|_{L^2} = O\left(\varepsilon^{-\frac{1}{2}}\right). \end{aligned}$$

Step 3: The case $n = 3$. Note that $(\Psi^\varepsilon)_{ij} = (\text{cof}(\tau I_h \sigma^\varepsilon + [1 - \tau] \sigma^\varepsilon))_{ij} = \det(\tau I_h \sigma^\varepsilon|_{ij} + [1 - \tau] \sigma^\varepsilon|_{ij})$, where $\sigma^\varepsilon|_{ij}$ denotes the 2×2 matrix after deleting the i th row and j th column of σ^ε . We can, thus, conclude that

$$\begin{aligned} |(\Psi^\varepsilon)_{ij}| &\leq 2 \max_{s \neq i, t \neq j} (|\tau(I_h \sigma^\varepsilon)_{st} + [1 - \tau](\sigma^\varepsilon)_{st}|)^2 \\ &\leq C \max_{s \neq i, t \neq j} |(\sigma^\varepsilon)_{st}|^2 \leq C \|\sigma^\varepsilon\|_{L^\infty}^2. \end{aligned}$$

Thus, (1.11) implies that

$$\|\Psi^\varepsilon\|_{L^2} \leq C \|\sigma^\varepsilon\|_{L^\infty}^2 = O(\varepsilon^{-2}).$$

Step 4: Using the estimates of $\|\Psi^\varepsilon\|_{L^2}$ we have

$$\begin{aligned} \|\omega_h\|_{L^2}^2 + \frac{\theta}{\varepsilon} \|Ds_h\|_{L^2}^2 &\leq \|I_h \sigma^\varepsilon - \sigma^\varepsilon\|_{L^2} \|\omega_h\|_{L^2} + \|\omega_h\|_{H^1} \|D(I_h u^\varepsilon - u^\varepsilon)\|_{L^2} \\ &\quad + \|I_h \sigma^\varepsilon - \sigma^\varepsilon\|_{H^1} \|Ds_h\|_{L^2} + C\varepsilon^{\frac{3}{2}(1-n)} \|\sigma^\varepsilon - I_h \sigma^\varepsilon\|_{H^1} \|s_h\|_{H^1}, \end{aligned}$$

where we have used Sobolev inequality. It follows from Poincaré inequality, Schwarz inequality, and the inverse inequality that

$$\begin{aligned} (4.8) \quad \|\omega_h\|_{L^2}^2 + \frac{\theta}{\varepsilon} \|s_h\|_{H^1}^2 &\leq C\varepsilon^{4-3n} \|I_h \sigma^\varepsilon - \sigma^\varepsilon\|_{H^1}^2 + C \|\omega_h\|_{H^1} \|I_h u^\varepsilon - u^\varepsilon\|_{H^1} \\ &\leq C\varepsilon^{4-3n} h^{2l-2} \|\sigma^\varepsilon\|_{H^l}^2 + Ch^{-1} \|\omega_h\|_{L^2} \|I_h u^\varepsilon - u^\varepsilon\|_{H^1}. \end{aligned}$$

Hence,

$$\|\omega_h\|_{L^2}^2 + \frac{1}{\varepsilon} \|s_h\|_{H^1}^2 \leq C\varepsilon^{4-3n} h^{2l-2} \|\sigma^\varepsilon\|_{H^l}^2 + Ch^{2l-4} \|u^\varepsilon\|_{H^l}^2.$$

Therefore,

$$\|\omega_h\|_{L^2} \leq C_2(\varepsilon) h^{l-2} [\|\sigma^\varepsilon\|_{H^l} + \|u^\varepsilon\|_{H^l}],$$

which and the inverse inequality yield

$$\|\omega_h\|_{H^1} \leq C_1(\varepsilon) h^{l-3} [\|\sigma^\varepsilon\|_{H^l} + \|u^\varepsilon\|_{H^l}].$$

Next, from (4.6) we have

$$\begin{aligned} (\text{div}(\mu_h), Ds_h) &\leq \|\omega_h\|_{L^2} \|\mu_h\|_{L^2} + \|I_h \sigma^\varepsilon - \sigma^\varepsilon\|_{L^2} \|\mu_h\|_{L^2} \\ &\quad + \|\text{div}(\mu_h)\|_{L^2} \|D(I_h u^\varepsilon - u^\varepsilon)\|_{L^2} \\ &\leq C_2(\varepsilon) h^{l-2} [\|\sigma^\varepsilon\|_{H^l} + \|u^\varepsilon\|_{H^l}] \|\mu_h\|_{H^1}. \end{aligned}$$

It follows from (3.11) that

$$(4.9) \quad \|Ds_h\|_{L^2} \leq C(\varepsilon) h^{l-2} [\|\sigma^\varepsilon\|_{H^l} + \|u^\varepsilon\|_{H^l}].$$

To prove (4.5), let (κ, z) be the solution to

$$\begin{aligned} (\kappa, \mu) + (\operatorname{div}(\mu), Dz) &= 0 & \forall \mu \in W_0, \\ (\operatorname{div}(\kappa), Dv) - \frac{1}{\varepsilon}(\Phi^\varepsilon Dz, Dv) &= \frac{1}{\varepsilon}(Ds_h, Dv) & \forall v \in V_0, \end{aligned}$$

and satisfy

$$\|z\|_{H^3} \leq C_b(\varepsilon)\|Ds_h\|_{L^2}.$$

Then,

$$\begin{aligned} \frac{1}{\varepsilon}\|Ds_h\|_{L^2}^2 &= (\operatorname{div}(\kappa), Ds_h) - \frac{1}{\varepsilon}(\Phi^\varepsilon Dz, Ds_h) \\ &= (\operatorname{div}(\Pi_h \kappa), Ds_h) - \frac{1}{\varepsilon}(\Phi^\varepsilon Dz, Ds_h) \\ &= -(\omega_h, \Pi_h \kappa) - \frac{1}{\varepsilon}(\Phi^\varepsilon Dz, Ds_h) + (I_h \sigma^\varepsilon - \sigma^\varepsilon, \Pi_h \kappa) \\ &\quad + (\operatorname{div}(\Pi_h \kappa), D(I_h u^\varepsilon - u^\varepsilon)) \\ &= -(\omega_h, \kappa) + (\omega_h, \kappa - \Pi_h \kappa) - \frac{1}{\varepsilon}(\Phi^\varepsilon Dz, Ds_h) \\ &\quad + (I_h \sigma^\varepsilon - \sigma^\varepsilon, \Pi_h \kappa) + (\operatorname{div}(\Pi_h \kappa), D(I_h u^\varepsilon - u^\varepsilon)) \\ &= (\operatorname{div}(\omega_h), Dz) - \frac{1}{\varepsilon}(\Phi^\varepsilon Ds_h, Dz) + (\omega_h, \kappa - \Pi_h \kappa) \\ &\quad + (I_h \sigma^\varepsilon - \sigma^\varepsilon, \Pi_h \kappa) + (\operatorname{div}(\Pi_h \kappa), D(I_h u^\varepsilon - u^\varepsilon)) \\ &= (\operatorname{div}(\omega_h), D(z - I_h z)) - \frac{1}{\varepsilon}(\Phi^\varepsilon Ds_h, D(z - I_h z)) + (\omega_h, \kappa - \Pi_h \kappa) \\ &\quad + (I_h \sigma^\varepsilon - \sigma^\varepsilon, \Pi_h \kappa) + (\operatorname{div}(\Pi_h \kappa), D(I_h u^\varepsilon - u^\varepsilon)) \\ &\quad + (\operatorname{div}(\sigma^\varepsilon - I_h \sigma^\varepsilon), I_h z) + \frac{1}{\varepsilon}(\det(\sigma^\varepsilon) - \det(I_h \sigma^\varepsilon), I_h z) \\ &\leq \|\operatorname{div}(\omega_h)\|_{L^2}\|D(z - I_h z)\|_{L^2} + \frac{1}{\varepsilon}\|\Phi^\varepsilon\|_{L^\infty}\|Ds_h\|_{L^2}\|D(z - I_h z)\|_{L^2} \\ &\quad + \|\omega_h\|_{L^2}\|\kappa - \Pi_h \kappa\|_{L^2} + \|I_h \sigma^\varepsilon - \sigma^\varepsilon\|_{L^2}\|\Pi_h \kappa\|_{L^2} \\ &\quad + \|\operatorname{div}(\Pi_h \kappa)\|_{L^2}\|D(I_h u^\varepsilon - u^\varepsilon)\|_{L^2} \\ &\quad + \|\operatorname{div}(\sigma^\varepsilon - I_h \sigma^\varepsilon)\|_{L^2}\|I_h z\|_{L^2} + \frac{C}{\varepsilon}\|\Psi^\varepsilon\|_{L^2}\|\sigma^\varepsilon - I_h \sigma^\varepsilon\|_{H^1}\|I_h z\|_{H^1} \\ &\leq Ch^2 \left(\|\omega\|_{H^1} + \frac{1}{\varepsilon^2}\|Ds_h\|_{L^2} \right) \|z\|_{H^3} \\ &\quad + C(\varepsilon)h^{l-1} (\|I_h z\|_{L^2} + \|I_h z\|_{H^1}) \|\sigma^\varepsilon\|_{H^1} \\ &\quad + Ch\|\omega_h\|_{L^2}\|\kappa\|_{H^1} + Ch^l\|\sigma^\varepsilon\|_{H^1}\|\Pi_h \kappa\|_{L^2} + Ch^{l-1}\|\Pi_h \kappa\|_{H^1}\|u^\varepsilon\|_{H^1} \\ &\leq C_2(\varepsilon)\varepsilon^{-2}h^{l-1} [\|u^\varepsilon\|_{H^1} + \|\sigma^\varepsilon\|_{H^1}] \|z\|_{H^3} \\ &\leq C_2(\varepsilon)\varepsilon^{-2}C_b(\varepsilon)h^{l-1} [\|u^\varepsilon\|_{H^1} + \|\sigma^\varepsilon\|_{H^1}] \|Ds_h\|_{L^2}. \end{aligned}$$

Dividing by $\|Ds_h\|_{L^2}$, we get (4.5). The proof is complete. \square

The next lemma shows the contractiveness of the mapping T .

LEMMA 4.2. *There exists an $h_0 = o(\varepsilon^{\frac{19}{12}})$ and $\rho_0 = o(\varepsilon^{\frac{19}{12}}|\log h|^{n-3}h^{\frac{n}{2}-1})$, such that for $h \leq h_0$, T is a contracting mapping in the ball $B_h(\rho_0)$ with a contraction*

factor $\frac{1}{2}$. That is, for any $(\mu_h, v_h), (\chi_h, w_h) \in B_h(\rho_0)$, there holds

$$(4.10) \quad \left\| T^{(1)}(\mu_h, v_h) - T^{(1)}(\chi_h, w_h) \right\|_{L^2} + \frac{1}{\sqrt{\varepsilon}} \left\| T^{(2)}(\mu_h, v_h) - T^{(2)}(\chi_h, w_h) \right\|_{H^1} \\ \leq \frac{1}{2} \left(\|\mu_h - \chi_h\|_{L^2} + \frac{1}{\sqrt{\varepsilon}} \|v_h - w_h\|_{H^1} \right).$$

Proof. We divide the proof into five steps.

Step 1: To ease notation, let

$$T^{(1)} = T^{(1)}(\mu_h, v_h) - T^{(1)}(\chi_h, w_h), \quad T^{(2)} = T^{(2)}(\mu_h, v_h) - T^{(2)}(\chi_h, w_h).$$

By the definition of $T^{(i)}$, we get

$$(4.11) \quad \left(T^{(1)}, \kappa_h \right) + \left(\operatorname{div}(\kappa_h), D \left(T^{(2)} \right) \right) = 0 \quad \forall \kappa_h \in W_0^h,$$

$$(4.12) \quad \left(\operatorname{div} \left(T^{(1)} \right), D z_h \right) - \frac{1}{\varepsilon} \left(\Phi^\varepsilon D \left(T^{(2)} \right), D z_h \right) \\ = \frac{1}{\varepsilon} \left[\left(\Phi^\varepsilon D(w_h - v_h), D z_h \right) + \left(\det(\chi_h) - \det(\mu_h), z_h \right) \right] \quad \forall z_h \in V_0^h.$$

Letting $z_h = T^{(2)}$ and $\kappa_h = T^{(1)}$, subtracting (4.12) from (4.11), and using the mean value theorem we have

$$\begin{aligned} & \left(T^{(1)}, T^{(1)} \right) + \frac{1}{\varepsilon} \left(\Phi^\varepsilon D T^{(2)}, D T^{(2)} \right) \\ &= \frac{1}{\varepsilon} \left[\left(\Phi^\varepsilon D(v_h - w_h), D T^{(2)} \right) + \left(\det(\mu_h) - \det(\chi_h), T^{(2)} \right) \right] \\ &= \frac{1}{\varepsilon} \left[\left(\Phi^\varepsilon D(v_h - w_h), D T^{(2)} \right) + \left(\Lambda_h : (\mu_h - \chi_h), T^{(2)} \right) \right] \\ &= \frac{1}{\varepsilon} \left[\left(\Phi^\varepsilon D(v_h - w_h), D T^{(2)} \right) + \left(\Phi^\varepsilon : (\mu_h - \chi_h), T^{(2)} \right) \right. \\ &\quad \left. + \left((\Lambda_h - \Phi^\varepsilon) : (\mu_h - \chi_h), T^{(2)} \right) \right] \\ &= \frac{1}{\varepsilon} \left[\left(\operatorname{div} \left(\Phi^\varepsilon T^{(2)} \right), D(v_h - w_h) \right) + \left(\mu_h - \chi_h, \Phi^\varepsilon T^{(2)} \right) \right. \\ &\quad \left. + \left((\Lambda_h - \Phi^\varepsilon) : (\mu_h - \chi_h), T^{(2)} \right) \right] \\ &= \frac{1}{\varepsilon} \left[\left(\operatorname{div} \left(\Pi_h \left(\Phi^\varepsilon T^{(2)} \right) \right), D(v_h - w_h) \right) + \left(\mu_h - \chi_h, \Phi^\varepsilon T^{(2)} \right) \right. \\ &\quad \left. + \left((\Lambda_h - \Phi^\varepsilon) : (\mu_h - \chi_h), T^{(2)} \right) \right] \\ &= \frac{1}{\varepsilon} \left[\left(\Phi^\varepsilon T^{(2)} - \Pi_h \left(\Phi^\varepsilon T^{(2)} \right), \mu_h - \chi_h \right) + \left((\Lambda_h - \Phi^\varepsilon) : (\mu_h - \chi_h), T^{(2)} \right) \right] \\ &\leq \frac{1}{\varepsilon} \left[\left\| \Phi^\varepsilon T^{(2)} - \Pi_h \left(\Phi^\varepsilon T^{(2)} \right) \right\|_{L^2} \|\mu_h - \chi_h\|_{L^2} \right. \\ &\quad \left. + C \|\Lambda_h - \Phi^\varepsilon\|_{L^2} \|\mu_h - \chi_h\|_{L^2} \left\| T^{(2)} \right\|_{L^\infty} \right] \\ &\leq \frac{1}{\varepsilon} \left[\left\| \Phi^\varepsilon T^{(2)} - \Pi_h \left(\Phi^\varepsilon T^{(2)} \right) \right\|_{L^2} \|\mu_h - \chi_h\|_{L^2} \right. \\ &\quad \left. + |\log h|^{\frac{3-n}{2}} h^{1-\frac{n}{2}} \|\Lambda_h - \Phi^\varepsilon\|_{L^2} \|\mu_h - \chi_h\|_{L^2} \left\| T^{(2)} \right\|_{H^1} \right], \end{aligned}$$

where $\Lambda_h = \text{cof}(\mu_h + \tau(\chi_h - \mu_h))$, $\tau \in [0, 1]$. $n = 2, 3$. We have used the inverse inequality to get the last inequality above.

Step 2: The case of $n = 2$. We bound $\|\Phi^\varepsilon - \Lambda_h\|_{L^2}$ as follows:

$$\begin{aligned} \|\Phi^\varepsilon - \Lambda_h\|_{L^2} &= \|\text{cof}(\sigma^\varepsilon) - \text{cof}(\mu_h + \tau(\chi_h - \mu_h))\|_{L^2} \\ &= \|\sigma^\varepsilon - \mu_h - \tau(\chi_h - \mu_h)\|_{L^2} \\ &\leq \|\sigma^\varepsilon - I_h\sigma^\varepsilon\|_{L^2} + \|I_h\sigma^\varepsilon - \mu_h\|_{L^2} + \|\chi_h - \mu_h\|_{L^2} \\ &\leq Ch^l \|\sigma^\varepsilon\|_{H^l} + 3\rho_0. \end{aligned}$$

Step 3: The case of $n = 3$. To bound $\|\Phi^\varepsilon - \Lambda_h\|_{L^2}$ in this case, we first write

$$\begin{aligned} \|(\Phi^\varepsilon - \Lambda_h)_{ij}\|_{L^2} &= \|(\text{cof}(\sigma^\varepsilon)_{ij} - \text{cof}(\mu_h + \tau(\chi_h - \mu_h))_{ij})\|_{L^2} \\ &= \|\det(\sigma^\varepsilon|_{ij}) - \det(\mu_h|_{ij} + \tau(\chi_h|_{ij} - \mu_h|_{ij}))\|_{L^2}, \end{aligned}$$

where $\sigma|_{ij}$ denotes the 2×2 matrix after deleting the i^{th} row and j^{th} column. Then, use the mean value theorem to get

$$\begin{aligned} \|(\Phi^\varepsilon - \Lambda_h)_{ij}\|_{L^2} &= \|\det(\sigma^\varepsilon|_{ij}) - \det(\mu_h|_{ij} + \tau(\chi_h|_{ij} - \mu_h|_{ij}))\|_{L^2} \\ &= \|\Lambda_{ij} : (\sigma^\varepsilon|_{ij} - \mu_h|_{ij} - \tau(\chi_h|_{ij} - \mu_h|_{ij}))\|_{L^2} \\ &\leq \|\Lambda_{ij}\|_{L^\infty} \|\sigma^\varepsilon|_{ij} - \mu_h|_{ij} - \tau(\chi_h|_{ij} - \mu_h|_{ij})\|_{L^2}, \end{aligned}$$

where $\Lambda_{ij} = \text{cof}(\sigma^\varepsilon|_{ij} + \lambda(\mu|_{ij} - \tau(\chi_h|_{ij} - \mu|_{ij}) - \sigma^\varepsilon|_{ij}))$, $\lambda \in [0, 1]$.

On noting that $\Lambda_{ij} \in \mathbf{R}^2$, we have

$$\begin{aligned} \|\Lambda_{ij}\|_{L^\infty} &= \|\text{cof}(\sigma^\varepsilon|_{ij} + \lambda(\mu|_{ij} - \tau(\chi_h|_{ij} - \mu|_{ij}) - \sigma^\varepsilon|_{ij}))\|_{L^\infty} \\ &= \|\sigma^\varepsilon|_{ij} + \lambda(\mu|_{ij} - \tau(\chi_h|_{ij} - \mu|_{ij}) - \sigma^\varepsilon|_{ij})\|_{L^\infty} \\ &\leq C\|\sigma^\varepsilon\|_{L^\infty} \leq \frac{C}{\varepsilon}. \end{aligned}$$

Combining the above estimates gives

$$\begin{aligned} \|(\Phi^\varepsilon - \Lambda_h)_{ij}\|_{L^2} &\leq \frac{C}{\varepsilon} \|\sigma^\varepsilon|_{ij} - \mu_h|_{ij} - \tau(\chi_h|_{ij} - \mu_h|_{ij})\|_{L^2} \\ &\leq \frac{C}{\varepsilon} (h^l \|\sigma^\varepsilon\|_{H^l} + \rho_0). \end{aligned}$$

Step 4: We now bound $\|\Phi^\varepsilon T^{(2)} - \Pi_h(\Phi^\varepsilon T^{(2)})\|_{L^2}$ as follows:

$$\begin{aligned} \left\| \Phi^\varepsilon T^{(2)} - \Pi_h \left(\Phi^\varepsilon T^{(2)} \right) \right\|_{L^2}^2 &\leq Ch^2 \left\| \Phi^\varepsilon T^{(2)} \right\|_{H^1}^2 \\ &= Ch^2 \left(\left\| \Phi^\varepsilon T^{(2)} \right\|_{L^2}^2 + \left\| D \left(\Phi^\varepsilon T^{(2)} \right) \right\|_{L^2}^2 \right) \\ &\leq Ch^2 \left(\left\| \Phi^\varepsilon T^{(2)} \right\|_{L^2}^2 + \left\| \Phi^\varepsilon DT^{(2)} \right\|_{L^2}^2 + \left\| D\Phi^\varepsilon T^{(2)} \right\|_{L^2}^2 \right) \\ &\leq Ch^2 \left(\left\| \Phi^\varepsilon \right\|_{L^4}^2 \left\| T^{(2)} \right\|_{L^4}^2 + \left\| \Phi^\varepsilon \right\|_{L^\infty} \left\| DT^{(2)} \right\|_{L^2}^2 + \left\| D\Phi^\varepsilon \right\|_{L^3}^2 \left\| T^{(2)} \right\|_{L^6}^2 \right) \\ &\leq Ch^2 \left(\left\| \Phi^\varepsilon \right\|_{L^4}^2 \left\| T^{(2)} \right\|_{H^1}^2 + \left\| \Phi^\varepsilon \right\|_{L^\infty}^2 \left\| DT^{(2)} \right\|_{L^2}^2 + \left\| D\Phi^\varepsilon \right\|_{L^3}^2 \left\| T^{(2)} \right\|_{H^1}^2 \right) \\ &\leq Ch^2 \left(\left\| \Phi^\varepsilon \right\|_{L^\infty}^2 + \left\| D\Phi^\varepsilon \right\|_{L^3}^2 \right) \left\| DT^{(2)} \right\|_{L^2}^2 \\ &\leq \frac{Ch^2}{\varepsilon^{\frac{13}{6}}} \left\| DT^{(2)} \right\|_{L^2}^2, \end{aligned}$$

where we have used Sobolev’s inequality followed by Poincaré’s inequality. Thus,

$$\left\| \Phi^\varepsilon T^{(2)} - \Pi_h \left(\Phi^\varepsilon T^{(2)} \right) \right\|_{L^2} \leq \frac{Ch}{\varepsilon^{\frac{13}{12}}} \left\| DT^{(2)} \right\|_{L^2}.$$

Step 5: Finishing up. Substituting all estimates from Steps 2–4 into Step 1, and using the fact that Φ^ε is positive definite we obtain for $n = 2, 3$

$$\left\| T^{(1)} \right\|_{L^2}^2 + \frac{\theta}{\varepsilon} \left\| DT^{(2)} \right\|_{L^2}^2 \leq C\varepsilon^{-\frac{25}{12}} \left(h + |\log h|^{\frac{3-n}{2}} h^{1-\frac{n}{2}} \rho_0 \right) \|\mu_h - \chi_h\|_{L^2} \left\| DT^{(2)} \right\|_{L^2}.$$

Using Schwarz’s inequality, we get

$$\left\| T^{(1)} \right\|_{L^2} + \frac{1}{\sqrt{\varepsilon}} \left\| T^{(2)} \right\|_{H^1} \leq C\varepsilon^{-\frac{19}{12}} \left(h + |\log h|^{\frac{3-n}{2}} h^{1-\frac{n}{2}} \rho_0 \right) \|\mu_h - \chi_h\|_{L^2}.$$

Choosing $h_0 = o(\varepsilon^{\frac{19}{12}})$, for $h \leq h_0$ and $\rho_0 = o(\varepsilon^{\frac{19}{12}} |\log h|^{\frac{n-3}{2}} h^{\frac{n}{2}-1})$, there holds

$$\begin{aligned} \left\| T^{(1)} \right\|_{L^2} + \frac{1}{\sqrt{\varepsilon}} \left\| T^{(2)} \right\|_{H^1} &\leq \frac{1}{2} \|\mu_h - \chi_h\|_{L^2} \\ &\leq \frac{1}{2} \left(\|\mu_h - \chi_h\|_{L^2} + \frac{1}{\sqrt{\varepsilon}} \|v_h - w_h\|_{H^1} \right). \end{aligned}$$

The proof is complete. \square

We are now ready to state and prove the main theorem of this paper.

THEOREM 4.1. *Let $\rho_1 = 2[C_2(\varepsilon)h^{l-2} + \frac{C_3(\varepsilon)}{\sqrt{\varepsilon}}h^{l-1}](\|\sigma^\varepsilon\|_{H^l} + \|u^\varepsilon\|_{H^l})$. Then there exists an $h_1 > 0$ such that for $h \leq \min\{h_0, h_1\}$, there exists a unique solution $(\sigma_h^\varepsilon, u_h^\varepsilon)$ to (2.7)–(2.8) in the ball $B_h(\rho_1)$. Moreover,*

$$(4.13) \quad \|\sigma^\varepsilon - \sigma_h^\varepsilon\|_{L^2} + \frac{1}{\sqrt{\varepsilon}} \|u^\varepsilon - u_h^\varepsilon\|_{H^1} \leq C_4(\varepsilon)h^{l-2} (\|\sigma^\varepsilon\|_{H^l} + \|u^\varepsilon\|_{H^l}),$$

$$(4.14) \quad \|\sigma^\varepsilon - \sigma_h^\varepsilon\|_{H^1} \leq C_5(\varepsilon)h^{l-3} (\|\sigma^\varepsilon\|_{H^l} + \|u^\varepsilon\|_{H^l}),$$

where $C_4(\varepsilon) = C_5(\varepsilon) = O(\varepsilon^{-\frac{9}{2}})$ when $n = 2$, $C_4(\varepsilon) = C_5(\varepsilon) = O(\varepsilon^{-6})$ when $n = 3$.

Proof. Let $(\mu_h, v_h) \in B_h(\rho_1)$ and choose $h_1 > 0$ such that

$$\begin{aligned} h_1 |\log h_1|^{\frac{3-n}{2l-n}} &\leq C \left(\frac{\varepsilon^{\frac{25}{12}}}{C_3(\varepsilon)(\|\sigma^\varepsilon\|_{H^l} + \|u^\varepsilon\|_{H^l})} \right)^{\frac{2}{2l-n}} \quad \text{and} \\ h_1 |\log h_1|^{\frac{3-n}{2l-n-2}} &\leq C \left(\frac{\varepsilon^{\frac{19}{12}}}{C_2(\varepsilon)(\|\sigma^\varepsilon\|_{H^l} + \|u^\varepsilon\|_{H^l})} \right)^{\frac{2}{2l-n-2}}. \end{aligned}$$

Then $h \leq \min\{h_0, h_1\}$ implies $\rho_1 \leq \rho_0$. Thus, using the triangle inequality and

Lemmas 4.1 and 4.2, we get

$$\begin{aligned}
 & \left\| I_h \sigma^\varepsilon - T^{(1)}(\mu_h, v_h) \right\|_{L^2} + \frac{1}{\sqrt{\varepsilon}} \left\| I_h u^\varepsilon - T^{(2)}(\mu_h, v_h) \right\|_{H^1} \\
 & \leq \left\| I_h \sigma^\varepsilon - T^{(1)}(I_h \sigma^\varepsilon, I_h u^\varepsilon) \right\|_{L^2} \\
 & \quad + \left\| T^{(1)}(I_h \sigma^\varepsilon, I_h u^\varepsilon) - T^{(1)}(\mu_h, v_h) \right\|_{L^2} + \frac{1}{\sqrt{\varepsilon}} \left\| I_h u^\varepsilon - T^{(2)}(I_h \sigma^\varepsilon, I_h u^\varepsilon) \right\|_{H^1} \\
 & \quad + \frac{1}{\sqrt{\varepsilon}} \left\| T^{(2)}(I_h \sigma^\varepsilon, I_h u^\varepsilon) - T^{(2)}(\mu_h, v_h) \right\|_{H^1} \\
 & \leq \left[C_2(\varepsilon) h^{l-2} + \frac{C_3(\varepsilon)}{\sqrt{\varepsilon}} h^{l-1} \right] (\|\sigma^\varepsilon\|_{H^l} + \|u^\varepsilon\|_{H^l}) \\
 & \quad + \frac{1}{2} \left(\|I_h \sigma^\varepsilon - \mu_h\|_{L^2} + \frac{1}{\sqrt{\varepsilon}} \|I_h u^\varepsilon - v_h\|_{H^1} \right) \\
 & \leq \frac{\rho_1}{2} + \frac{\rho_1}{2} = \rho_1 < 1.
 \end{aligned}$$

So, $T(\mu_h, v_h) \in B_h(\rho_1)$. Clearly, T is a continuous mapping. Thus, T has a unique fixed point $(\sigma_h^\varepsilon, u_h^\varepsilon) \in B_h(\rho_1)$, which is the unique solution to (2.7)–(2.8).

Next, we use the triangle inequality to get

$$\begin{aligned}
 \|\sigma^\varepsilon - \sigma_h^\varepsilon\|_{L^2} + \frac{1}{\sqrt{\varepsilon}} \|u^\varepsilon - u_h^\varepsilon\|_{H^1} & \leq \|\sigma^\varepsilon - I_h \sigma^\varepsilon\|_{L^2} + \|I_h \sigma^\varepsilon - \sigma_h^\varepsilon\|_{L^2} \\
 & \quad + \frac{1}{\sqrt{\varepsilon}} (\|u^\varepsilon - I_h u^\varepsilon\|_{H^1} + \|I_h u^\varepsilon - u_h^\varepsilon\|_{H^1}) \\
 & \leq \rho_1 + C h^{l-1} (\|\sigma^\varepsilon\|_{H^l} + \|u^\varepsilon\|_{H^l}) \\
 & \leq C_4(\varepsilon) h^{l-2} (\|\sigma^\varepsilon\|_{H^l} + \|u^\varepsilon\|_{H^l}).
 \end{aligned}$$

Finally, using the inverse inequality, we have

$$\begin{aligned}
 \|\sigma^\varepsilon - \sigma_h^\varepsilon\|_{H^1} & \leq \|\sigma^\varepsilon - I_h \sigma^\varepsilon\|_{H^1} + \|I_h \sigma^\varepsilon - \sigma_h^\varepsilon\|_{H^1} \\
 & \leq \|\sigma^\varepsilon - I_h \sigma^\varepsilon\|_{H^1} + C h^{-1} \|I_h \sigma^\varepsilon - \sigma_h^\varepsilon\|_{L^2} \\
 & \leq C h^{l-1} \|\sigma^\varepsilon\|_{H^l} + C h^{-1} \rho_1 \\
 & \leq C_5(\varepsilon) h^{l-3} [\|\sigma^\varepsilon\|_{H^l} + \|u^\varepsilon\|_{H^l}].
 \end{aligned}$$

The proof is complete. \square

Comparing with error estimates for the linearized problem in Theorem 3.2, we see that the above H^1 error for the scalar variable is not optimal. Next, we shall employ a similar duality argument as used in the proof of Theorem 3.2 to show that the estimate can be improved to optimal order.

THEOREM 4.2. *Under the same hypothesis of Theorem 4.1 there holds*

$$(4.15) \quad \|u^\varepsilon - u_h^\varepsilon\|_{H^1} \leq C_4(\varepsilon) \varepsilon^{-2} \left[h^{l-1} + C_5(\varepsilon) h^{2(l-2)} \right] (\|\sigma^\varepsilon\|_{H^l} + \|u^\varepsilon\|_{H^l}).$$

Proof. The regularity assumption implies that there exists $(\kappa, z) \in W_0 \times V_0 \cap H^3(\Omega)$ such that

$$(4.16) \quad (\kappa, \mu) + (\operatorname{div}(\mu), Dz) = 0 \quad \forall \mu \in W_0,$$

$$(4.17) \quad (\operatorname{div}(\kappa), Dv) - \frac{1}{\varepsilon} (\Phi^\varepsilon Dz, Dv) = \frac{1}{\varepsilon} (D(u^\varepsilon - u_h^\varepsilon), Dv) \quad \forall v \in V_0,$$

with

$$(4.18) \quad \|z\|_{H^3} \leq C_b(\varepsilon) \|D(u^\varepsilon - u_h^\varepsilon)\|_{L^2}.$$

It is easy to check that $\sigma^\varepsilon - \sigma_h^\varepsilon$ and $u^\varepsilon - u_h^\varepsilon$ satisfy the following error equations:

$$(4.19) \quad (\sigma^\varepsilon - \sigma_h^\varepsilon, \mu_h) + (\operatorname{div}(\mu_h), D(u^\varepsilon - u_h^\varepsilon)) = 0 \quad \forall \mu_h \in W_0^h,$$

$$(4.20) \quad (\operatorname{div}(\sigma^\varepsilon - \sigma_h^\varepsilon), Dv_h) + \frac{1}{\varepsilon}(\det(\sigma^\varepsilon) - \det(\sigma_h^\varepsilon), v_h) = 0 \quad \forall v_h \in V_0^h.$$

By (4.16)–(4.20) and the mean value theorem, we get

$$\begin{aligned} \frac{1}{\varepsilon} \|D(u^\varepsilon - u_h^\varepsilon)\|_{L^2}^2 &= (\operatorname{div}(\kappa), D(u^\varepsilon - u_h^\varepsilon)) - \frac{1}{\varepsilon} (\Phi^\varepsilon Dz, D(u^\varepsilon - u_h^\varepsilon)) \\ &= (\operatorname{div}(\Pi_h \kappa), D(u^\varepsilon - u_h^\varepsilon)) - \frac{1}{\varepsilon} (\Phi^\varepsilon D(u^\varepsilon - u_h^\varepsilon), Dz) + (\operatorname{div}(\kappa - \Pi_h \kappa), D(u^\varepsilon - u_h^\varepsilon)) \\ &= (\sigma_h^\varepsilon - \sigma^\varepsilon, \Pi_h \kappa) - \frac{1}{\varepsilon} (\Phi^\varepsilon D(u^\varepsilon - u_h^\varepsilon), Dz) + (\operatorname{div}(\kappa - \Pi_h \kappa), D(u^\varepsilon - u_h^\varepsilon)) \\ &= (\sigma_h^\varepsilon - \sigma^\varepsilon, \kappa) - \frac{1}{\varepsilon} (\Phi^\varepsilon D(u^\varepsilon - u_h^\varepsilon), Dz) \\ &\quad + (\operatorname{div}(\kappa - \Pi_h \kappa), D(u^\varepsilon - I_h u^\varepsilon)) + (\sigma_h^\varepsilon - \sigma^\varepsilon, \Pi_h \kappa - \kappa) \\ &= (\operatorname{div}(\sigma^\varepsilon - \sigma_h^\varepsilon), Dz) - \frac{1}{\varepsilon} (\Phi^\varepsilon D(u^\varepsilon - u_h^\varepsilon), Dz) \\ &\quad + (\operatorname{div}(\kappa - \Pi_h \kappa), D(u^\varepsilon - I_h u^\varepsilon)) + (\sigma_h^\varepsilon - \sigma^\varepsilon, \Pi_h \kappa - \kappa) \\ &= (\operatorname{div}(\sigma^\varepsilon - \sigma_h^\varepsilon), D(z - I_h z)) - \frac{1}{\varepsilon} (\Phi^\varepsilon D(u^\varepsilon - u_h^\varepsilon), D(z - I_h z)) \\ &\quad + (\operatorname{div}(\kappa - \Pi_h \kappa), D(u^\varepsilon - I_h u^\varepsilon)) + (\sigma_h^\varepsilon - \sigma^\varepsilon, \Pi_h \kappa - \kappa) \\ &\quad - \frac{1}{\varepsilon} (\det(\sigma^\varepsilon) - \det(\sigma_h^\varepsilon), I_h z) - \frac{1}{\varepsilon} (\Phi^\varepsilon D(u^\varepsilon - u_h^\varepsilon), D(I_h z)) \\ &= (\operatorname{div}(\sigma^\varepsilon - \sigma_h^\varepsilon), D(z - I_h z)) - \frac{1}{\varepsilon} (\Phi^\varepsilon D(u^\varepsilon - u_h^\varepsilon), D(z - I_h z)) \\ &\quad + (\operatorname{div}(\kappa - \Pi_h \kappa), D(u^\varepsilon - I_h u^\varepsilon)) + (\sigma_h^\varepsilon - \sigma^\varepsilon, \Pi_h \kappa - \kappa) \\ &\quad - \frac{1}{\varepsilon} (\Psi^\varepsilon : (\sigma^\varepsilon - \sigma_h^\varepsilon), I_h z) - \frac{1}{\varepsilon} (\Phi^\varepsilon D(u^\varepsilon - u_h^\varepsilon), D(I_h z)), \end{aligned}$$

where $\Psi^\varepsilon = \operatorname{cof}(\sigma^\varepsilon + \tau[\sigma_h^\varepsilon - \sigma^\varepsilon])$ for $\tau \in [0, 1]$.

Next, we note that

$$\begin{aligned} &(\Psi^\varepsilon : (\sigma^\varepsilon - \sigma_h^\varepsilon), I_h z) + (\Phi^\varepsilon D(u^\varepsilon - u_h^\varepsilon), D(I_h z)) \\ &= (\Phi^\varepsilon : (\sigma^\varepsilon - \sigma_h^\varepsilon), I_h z) + (\operatorname{div}(\Phi^\varepsilon I_h z), D(u^\varepsilon - u_h^\varepsilon)) + ((\Psi^\varepsilon - \Phi^\varepsilon) : (\sigma^\varepsilon - \sigma_h^\varepsilon), I_h z) \\ &= (\sigma^\varepsilon - \sigma_h^\varepsilon, \Phi^\varepsilon I_h z) + (\operatorname{div}(\Pi_h(\Phi^\varepsilon I_h z)), D(u^\varepsilon - u_h^\varepsilon)) + ((\Psi^\varepsilon - \Phi^\varepsilon) : (\sigma^\varepsilon - \sigma_h^\varepsilon), I_h z) \\ &\quad + (\operatorname{div}(\Phi^\varepsilon I_h z - \Pi_h(\Phi^\varepsilon I_h z)), D(u^\varepsilon - I_h u^\varepsilon)) \\ &= (\sigma^\varepsilon - \sigma_h^\varepsilon, \Phi^\varepsilon I_h z - \Pi_h(\Phi^\varepsilon I_h z)) + ((\Psi^\varepsilon - \Phi^\varepsilon) : (\sigma^\varepsilon - \sigma_h^\varepsilon), I_h z) \\ &\quad + (\operatorname{div}(\Phi^\varepsilon I_h z - \Pi_h(\Phi^\varepsilon I_h z)), D(u^\varepsilon - I_h u^\varepsilon)). \end{aligned}$$

Using this and the same technique used in Step 4 of Lemma 4.2, we have

$$\begin{aligned}
 \frac{1}{\varepsilon} \|D(u^\varepsilon - u_h^\varepsilon)\|_{L^2}^2 &= (\operatorname{div}(\sigma^\varepsilon - \sigma_h^\varepsilon), D(z - I_h z)) - \frac{1}{\varepsilon} (\Phi^\varepsilon D(u^\varepsilon - u_h^\varepsilon), D(z - I_h z)) \\
 &\quad + \frac{1}{\varepsilon} [((\Phi^\varepsilon - \Psi^\varepsilon) : (\sigma^\varepsilon - \sigma_h^\varepsilon), I_h z) + (\sigma^\varepsilon - \sigma_h^\varepsilon, \Pi_h(\Phi^\varepsilon I_h z) - \Phi^\varepsilon I_h z) \\
 &\quad + (\operatorname{div}(\Pi_h(\Phi^\varepsilon I_h z) - \Phi^\varepsilon I_h z), D(u^\varepsilon - I_h u^\varepsilon))] + (\sigma_h^\varepsilon - \sigma^\varepsilon, \Pi_h \kappa - \kappa) \\
 &\quad + (\operatorname{div}(\kappa - \Pi_h \kappa), D(u^\varepsilon - I_h u^\varepsilon)) \\
 &\leq \left[\|\operatorname{div}(\sigma^\varepsilon - \sigma_h^\varepsilon)\|_{L^2} + \frac{C}{\varepsilon^2} \|D(u^\varepsilon - u_h^\varepsilon)\|_{L^2} \right] \|D(z - I_h z)\|_{L^2} \\
 &\quad + \frac{C}{\varepsilon} \left[\|\Phi^\varepsilon - \Psi^\varepsilon\|_{L^2} \|\sigma^\varepsilon - \sigma_h^\varepsilon\|_{L^2} \|I_h z\|_{L^\infty} + \|\sigma^\varepsilon - \sigma_h^\varepsilon\|_{L^2} \|\Pi_h(\Phi^\varepsilon I_h z) - \Phi^\varepsilon I_h z\|_{L^2} \right. \\
 &\quad \left. + \|\operatorname{div}(\Pi_h(\Phi^\varepsilon I_h z) - \Phi^\varepsilon I_h z)\|_{L^2} \|D(u^\varepsilon - I_h u^\varepsilon)\|_{L^2} \right] + \|\kappa - \Pi_h \kappa\|_{L^2} \|\sigma^\varepsilon - \sigma_h^\varepsilon\|_{L^2} \\
 &\quad + \|\operatorname{div}(\kappa - \Pi_h \kappa)\|_{L^2} \|D(u^\varepsilon - I_h u^\varepsilon)\|_{L^2} \\
 &\leq Ch^2 \left(\|\sigma^\varepsilon - \sigma_h^\varepsilon\|_{H^1} + \frac{1}{\varepsilon^2} \|u^\varepsilon - u_h^\varepsilon\|_{H^1} \right) \|z\|_{H^3} \\
 &\quad + \frac{C}{\varepsilon^2} (\|\Phi^\varepsilon - \Psi^\varepsilon\|_{L^2} \|\sigma^\varepsilon - \sigma_h^\varepsilon\|_{L^2} + h \|\sigma^\varepsilon - \sigma_h^\varepsilon\|_{L^2} + \|u^\varepsilon - I_h u^\varepsilon\|_{H^1}) \|z\|_{H^3} \\
 &\quad + Ch \|\sigma^\varepsilon - \sigma_h^\varepsilon\|_{L^2} \|\kappa\|_{H^1} + C \|u^\varepsilon - I_h u^\varepsilon\|_{H^1} \|\kappa\|_{H^1} \\
 &\leq \left\{ \frac{(C_4(\varepsilon) + C_5(\varepsilon))h^{l-1}}{\varepsilon^{\frac{3}{2}}} [\|\sigma^\varepsilon\|_{H^1} + \|u^\varepsilon\|_{H^1}] + \frac{C_4(\varepsilon)h^{l-2}}{\varepsilon^2} \|\Phi^\varepsilon - \Psi^\varepsilon\|_{L^2} \right\} \|z\|_{H^3} \\
 &\leq C_b(\varepsilon) \left\{ \frac{(C_4(\varepsilon) + C_5(\varepsilon))h^{l-1}}{\varepsilon^{\frac{3}{2}}} [\|\sigma^\varepsilon\|_{H^1} + \|u^\varepsilon\|_{H^1}] \right. \\
 &\quad \left. + \frac{C_4(\varepsilon)h^{l-2}}{\varepsilon^2} \|\Phi^\varepsilon - \Psi^\varepsilon\|_{L^2} \right\} \|D(u^\varepsilon - u_h^\varepsilon)\|_{L^2}.
 \end{aligned}$$

We now bound $\|\Phi^\varepsilon - \Psi^\varepsilon\|_{L^2}$ separately for the cases $n = 2$ and $n = 3$. First, when $n = 2$ we have

$$\begin{aligned}
 \|\Phi^\varepsilon - \Psi^\varepsilon\|_{L^2} &= \|\operatorname{cof}(\sigma^\varepsilon) - \operatorname{cof}(\sigma_h^\varepsilon + \tau[\sigma^\varepsilon - \sigma_h^\varepsilon])\|_{L^2} = \|\sigma^\varepsilon - (\sigma_h^\varepsilon + \tau[\sigma^\varepsilon - \sigma_h^\varepsilon])\|_{L^2} \\
 &\leq C_4(\varepsilon)h^{l-2} [\|\sigma^\varepsilon\|_{H^1} + \|u^\varepsilon\|_{H^1}].
 \end{aligned}$$

Second, when $n = 3$, on noting that

$$\begin{aligned}
 |(\Phi^\varepsilon - \Psi^\varepsilon)_{ij}| &= |(\operatorname{cof}(\sigma^\varepsilon))_{ij} - (\operatorname{cof}(\sigma_h^\varepsilon + \tau[\sigma^\varepsilon - \sigma_h^\varepsilon]))_{ij}| \\
 &= |\det(\sigma^\varepsilon|_{ij}) - \det(\sigma^\varepsilon|_{ij} + \tau[\sigma^\varepsilon|_{ij} - \sigma_h^\varepsilon|_{ij}])|,
 \end{aligned}$$

and, using the mean value theorem and Sobolev inequality, we get

$$\begin{aligned}
 \|(\Psi^\varepsilon)_{ij} - (\Phi^\varepsilon)_{ij}\|_{L^2} &= (1 - \tau) \|(\Lambda^\varepsilon)^{ij} : (\sigma^\varepsilon|_{ij} - \sigma_h^\varepsilon|_{ij})\|_{L^2} \\
 &\leq \|(\Lambda^\varepsilon)^{ij}\|_{H^1} \|\sigma^\varepsilon|_{ij} - \sigma_h^\varepsilon|_{ij}\|_{H^1},
 \end{aligned}$$

where $(\Lambda^\varepsilon)^{ij} = \operatorname{cof}(\sigma^\varepsilon|_{ij} + \lambda[\sigma_h^\varepsilon|_{ij} - \sigma^\varepsilon|_{ij}])$ for $\lambda \in [0, 1]$. Since $(\Lambda^\varepsilon)^{ij} \in \mathbf{R}^{2 \times 2}$, then

$$\|(\Lambda^\varepsilon)^{ij}\|_{H^1} = \|\sigma^\varepsilon|_{ij} + \lambda(\sigma_h^\varepsilon|_{ij} - \sigma^\varepsilon|_{ij})\|_{H^1} \leq C \|\sigma^\varepsilon\|_{H^1} = O(\varepsilon^{-1}).$$

Thus,

$$\|\Phi^\varepsilon - \Psi^\varepsilon\|_{L^2} \leq C_4(\varepsilon)\varepsilon^{-1}h^{l-2} (\|\sigma^\varepsilon\|_{H^1} + \|u^\varepsilon\|_{H^1}).$$

Finally, combining the above estimates we obtain

$$\|D(u^\varepsilon - u_h^\varepsilon)\|_{L^2} \leq C_4(\varepsilon)\varepsilon^{-2} \left[h^{l-1} + C_4(\varepsilon)h^{2(l-2)} \right] (\|\sigma^\varepsilon\|_{H^1} + \|u^\varepsilon\|_{H^1}).$$

We note that $2(l-2) \geq l-1$ for $k \geq 2$. The proof is complete. \square

5. Numerical experiments and rates of convergence. In this section, we provide several 2-D numerical experiments to gauge the efficiency of the mixed finite element method developed in the previous sections. We numerically determine the “best” choice of the mesh size h in terms of ϵ , and rates of convergence for both $u^0 - u^\epsilon$ and $u^\epsilon - u_h^\epsilon$. All tests given below are done on domain $\Omega = [0, 1]^2$. We refer the reader to [18, 27] for more extensive 2-D and 3-D numerical simulations. Newton’s method is employed as the (nonlinear) solver in all our numerical tests. We like to remark that the mixed finite element methods we tested are often 10–20 times faster than the Argyris finite element Galerkin method studied in [19]. We refer the reader to [18] for more discussions and comparisons on the Galerkin and mixed methods.

Test 1. For this test, we calculate $\|u^0 - u_h^\epsilon\|$ for fixed $h = 0.015$, while varying ϵ in order to estimate $\|u^\epsilon - u^0\|$. We use quadratic Lagrange element for both variables and solve problem (2.5)–(2.6) with the following test functions:

$$\begin{aligned} \text{(a)} \quad u^0 &= e^{\frac{x^2+y^2}{2}}, & f &= (1+x^2+y^2)e^{\frac{x^2+y^2}{2}}, & g &= e^{\frac{x^2+y^2}{2}}, \\ \text{(b)} \quad u^0 &= x^4 + y^2, & f &= 24x^2, & g &= x^4 + y^2. \end{aligned}$$

After having computed the error, we divide it by various powers of ϵ to estimate the rate at which each norm converges. The left column of Figure 5.1, which is the log-log plots of the errors in various norms vs ϵ , clearly shows that $\|\sigma^0 - \sigma_h^\epsilon\|_{L^2} = O(\epsilon^{\frac{1}{4}})$. Since h is very small, we then have $\|u^0 - u^\epsilon\|_{H^2} \approx \|\sigma^0 - \sigma_h^\epsilon\|_{L^2} = O(\epsilon^{\frac{1}{4}})$. Based on this heuristic argument, we predict that $\|u^0 - u^\epsilon\|_{H^2} = O(\epsilon^{\frac{1}{4}})$. Similarly, from the left column of Figure 5.1, we see that $\|u^0 - u^\epsilon\|_{L^2} \approx O(\epsilon)$ and $\|u^0 - u^\epsilon\|_{H^1} \approx O(\epsilon^{\frac{3}{4}})$.

Test 2. The purpose of this test is to calculate the rate of convergence of $\|u^\epsilon - u_h^\epsilon\|$ for fixed ϵ in various norms. We use quadratic Lagrange element for both variables and solve problem (2.5)–(2.6) with boundary condition $D^2 u^\epsilon \nu \cdot \nu = \epsilon$ on $\partial\Omega$ being replaced by $D^2 u^\epsilon \nu \cdot \nu = h_\epsilon$ on $\partial\Omega$ and using the following test functions:

$$\begin{aligned} \text{(a)} \quad u^\epsilon &= 20x^6 + y^6, & f^\epsilon &= 18000x^4y^4 - \epsilon(7200x^2 + 360y^2), \\ g^\epsilon &= 20x^6 + y^6, & h^\epsilon &= 600x^4\nu_x^2 + 30y^4\nu_y^2. \\ \text{(b)} \quad u^\epsilon &= x\sin(x) + y\sin(y), & f^\epsilon &= (2\cos(x) - x\sin(x))(2\cos(y) - y\sin(y)) \\ & & & - \epsilon(x\sin(x) - 4\cos(x) + y\sin(y) - 4\cos(y)), \\ g^\epsilon &= x\sin(x) + y\sin(y), & h^\epsilon &= (2\cos(x) - x\sin(x))\nu_x^2 + (2\cos(y) - y\sin(y))\nu_y^2. \end{aligned}$$

After having computed the error in different norms, we divided each value by a power of h expected to be the convergence rate by the analysis in the previous section. As seen from the right column of Figure 5.1, which is the log-log plots of the errors in various norms vs h , the error converges exactly as expected in H^1 norm, but σ_h^ϵ appears to converge one order of h better than the analysis shows. In addition, the error seems to converge optimally in L^2 norm although a theoretical proof of such a result has not yet been proved.

Test 3. In this test, we fix a relation between ϵ and h , and then determine the “best” choice for h in terms of ϵ such that the global error $u^0 - u_h^\epsilon$ has the same convergence rate as that of $u^0 - u^\epsilon$. We solve problem (2.5)–(2.6) with the following test functions:

$$\text{(a)} \quad u^0 = x^4 + y^2, \quad f = 24x^2, \quad g = x^4 + y^2.$$

To see which relation gives the sought-after convergence rate, we compare the data with a function, $y = \beta x^\alpha$, where $\alpha = 1$ in the L^2 case, $\alpha = \frac{3}{4}$ in the H^1 case, and

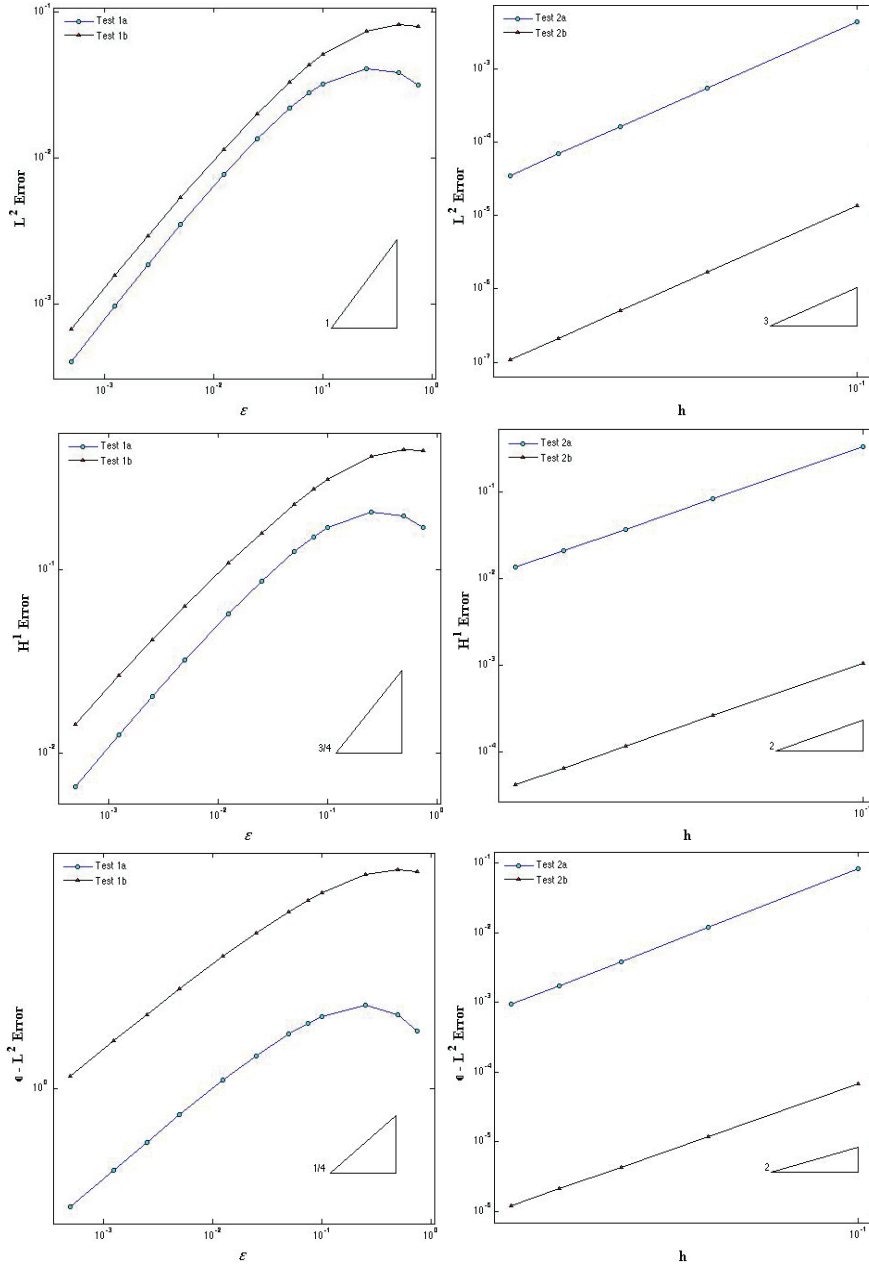


FIG. 5.1. Log-log plots of change of $\|u - u_h^\epsilon\|$ w.r.t. ϵ for Test 1 (left column) and log-log plots of change of $\|u - u_h^\epsilon\|$ w.r.t. h for Test 2 (right column).

$\alpha = \frac{1}{4}$ in the H^2 -case. The constant, β , is determined using a least squares fitting algorithm based on the data.

As seen in the figures below, the best $h - \epsilon$ relation depends on which norm one considers. Figures 5.2 and 5.3 indicate that when $h = \epsilon^{\frac{1}{2}}$, $\|u^0 - u_h^\epsilon\|_{L^2} \approx O(\epsilon)$, and $\|\sigma^0 - \sigma_h^\epsilon\|_{L^2} \approx O(\epsilon^{\frac{1}{4}})$. It can also be seen from Figure 5.4 that when $h = \epsilon$, $\|u^0 - u_h^\epsilon\|_{H^1} = O(\epsilon^{\frac{3}{4}})$.

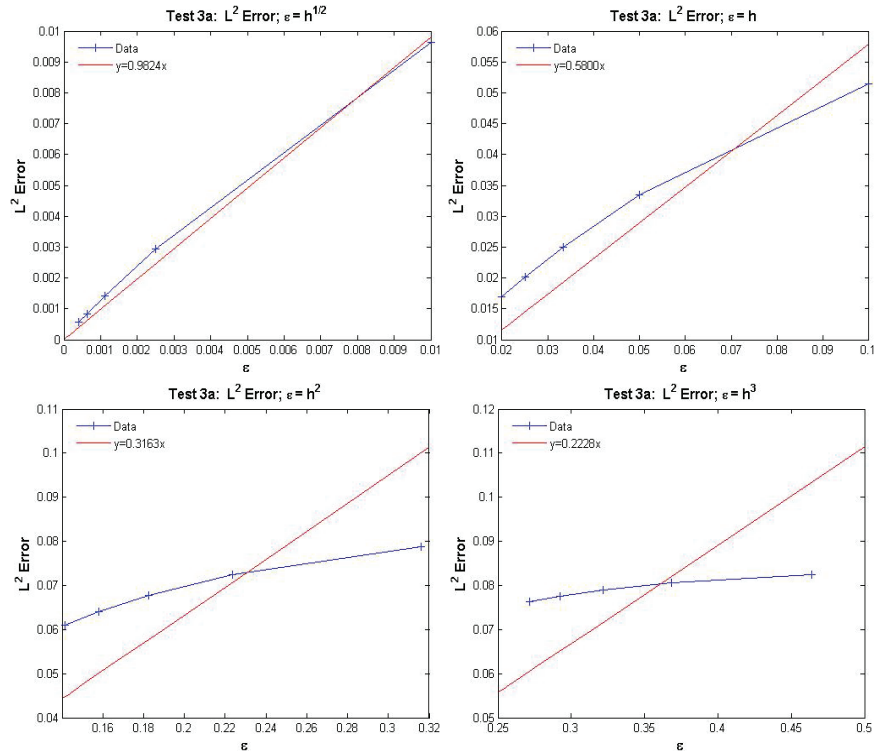


FIG. 5.2. Test 3a. L^2 -error of u_h^ϵ .

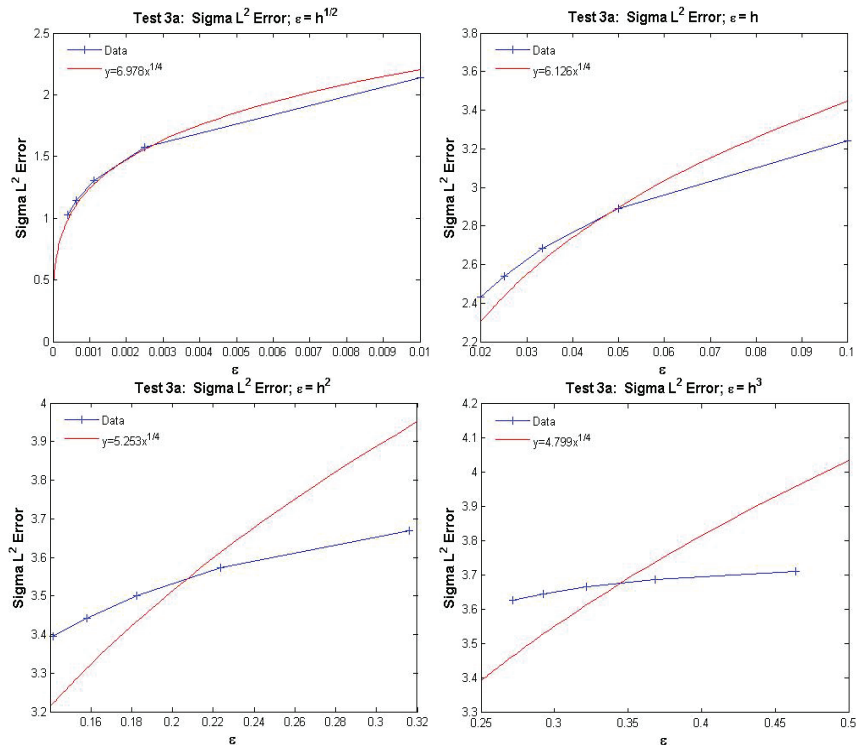
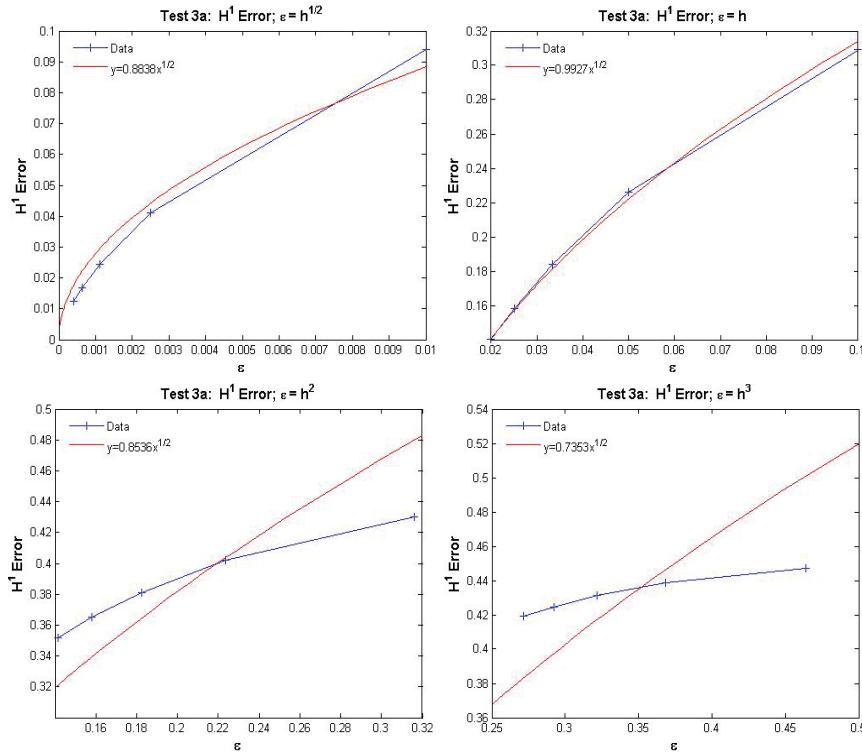


FIG. 5.3. Test 3a. L^2 -error of σ_h^ϵ .

FIG. 5.4. Test 3a. H^1 -error of u_h^ϵ .

REFERENCES

- [1] A. D. ALEKSANDROV, *Certain estimates for the Dirichlet problem*, Soviet Math. Dokl., 1 (1961), pp. 1151–1154.
- [2] F. E. BAGINSKI AND N. WHITAKER, *Numerical solutions of boundary value problems for K -surfaces in \mathbf{R}^3* , Numer. Methods for Partial Differential Equations, 12 (1996), pp. 525–546.
- [3] G. BARLES AND P. E. SOUGANIDIS, *Convergence of approximation schemes for fully nonlinear second order equations*, Asymptot. Anal., 4 (1991), pp. 271–283.
- [4] J.-D. BENAMOU AND Y. BRENIER, *A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem*, Numer. Math., 84 (2000), pp. 375–393.
- [5] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, 3rd edition, Springer, New York, 2008.
- [6] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, 1st edition, Springer-Verlag, Berlin, 1991.
- [7] L. A. CAFFARELLI AND X. CABRÉ, *Fully nonlinear elliptic equations*, American Mathematical Society Colloquium Publications 43, AMS, Providence, RI, 1995.
- [8] L. A. CAFFARELLI AND M. MILMAN, *Monge Ampère equation: Applications to geometry and optimization*, Contemporary Mathematics, AMS, Providence, RI, 1999.
- [9] S. Y. CHENG AND S. T. YAU, *On the regularity of the Monge-Ampère equation $\det(\partial^2 u / \partial x_i \partial x_j) = F(x, u)$* , Comm. Pure Appl. Math., 30 (1977), pp. 41–68.
- [10] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
- [11] M. G. CRANDALL AND P.-L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.
- [12] M. G. CRANDALL, H. ISHII, AND P.-L. LIONS, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.
- [13] E. J. DEAN AND R. GLOWINSKI, *Numerical methods for fully nonlinear elliptic equations of the Monge-Ampère type*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 1344–1386.

- [14] L. C. EVANS, *Partial differential equations*, Graduate Studies in Mathematics 19, AMS, Providence, RI, 1998.
- [15] R. S. FALK AND J. E. OSBORN, *Error estimates for mixed methods*, R.A.I.R.O. Anal. Numér., 14 (1980), pp. 249–277.
- [16] X. FENG, *Convergence of the vanishing moment method for the Monge-Ampère equation*, Trans. AMS, submitted.
- [17] X. FENG AND O. A. KARAKASHIAN, *Fully discrete dynamic mesh discontinuous Galerkin methods for the Cahn-Hilliard equation of phase transition*, Math. Comp. 76 (2007), pp. 1093–1117.
- [18] X. FENG AND M. NEILAN, *Vanishing moment method and moment solutions for second order fully nonlinear partial differential equations*, J. Scient. Comp., DOI 10.1007/s10915-008-9221-9, 2008.
- [19] X. FENG AND M. NEILAN, *Analysis of Galerkin methods for the fully nonlinear Monge-Ampère equation*, Math. Comp., to appear.
- [20] X. FENG, M. NEILAN, AND A. PROHL, *Error analysis of finite element approximations of the inverse mean curvature flow arising from the general relativity*, Numer. Math., 108 (2007), pp. 93–119.
- [21] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order, Classics in Mathematics*, Springer-Verlag, Berlin, 2001. Reprint of the 1998 edition.
- [22] C. E. GUTIERREZ, *The Monge-Ampère Equation*, volume 44 of Progress in Nonlinear Differential Equations and Their Applications, Birkhauser, Boston, MA, 2001.
- [23] H. ISHII, *On uniqueness and existence of viscosity solutions of fully nonlinear second order PDE's*, Comm. Pure Appl. Math., 42 (1989), pp. 14–45.
- [24] R. JENSEN, *The maximum principle for viscosity solutions of fully nonlinear second order partial differential equations*, Arch. Ration. Mech. Anal., 101 (1988), pp. 1–27.
- [25] O. A. LADYZHENSKAYA AND N. N. URAL'TSEVA, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.
- [26] I. MOZOLEVSKI AND E. SÜLI, *A priori error analysis for the hp-version of the discontinuous Galerkin finite element method for the biharmonic equation*, Comput. Methods Appl. Math., 3 (2003), pp. 596–607.
- [27] M. NEILAN, *Numerical methods for fully nonlinear second order partial differential equations*, Ph.D. Dissertation, The University of Tennessee, in preparation.
- [28] A. M. OBERMAN, *Wide stencil finite difference schemes for elliptic Monge-Ampère equation and functions of the eigenvalues of the Hessian*, Discrete Contin. Dyn. Syst. B, 10 (2008), pp. 221–238.
- [29] V. I. OLIKER AND L. D. PRUSSNER, *On the numerical solution of the equation $(\partial^2 z / \partial x^2)(\partial^2 z / \partial y^2) - ((\partial^2 z / \partial x \partial y))^2 = f$ and its discretizations. I.*, Numer. Math., 54 (1988), pp. 271–293.
- [30] A. OUKIT AND R. PIERRE, *Mixed finite element for the linear plate problem: The Hermann-Miyoshi model revisited*, Numer. Math., 74 (1996), pp. 453–477.
- [31] J. E. ROBERTS AND J. M. THOMAS, *Mixed and hybrid methods*, Handbook of Numerical Analysis, Vol. II, Finite Element Methods, North-Holland, Amsterdam, 1989.
- [32] T. NILSSEN, X.-C. TAI, AND R. WAGNER, *A robust nonconfirming H^2 element*, Math. Comp., 70 (2000), pp. 489–505.
- [33] M. WANG, Z. SHI, AND J. XU, *A new class of Zienkiewicz-type nonconforming elements in any dimensions*, Numer. Math., 106 (2007), pp. 335–347.
- [34] M. WANG AND J. XU, *Some tetrahedron nonconforming elements for fourth order elliptic equations*, Math. Comp., 76 (2007), pp. 1–18.

NONSMOOTH NEWTON METHODS FOR SET-VALUED SADDLE POINT PROBLEMS*

CARSTEN GRÄSER† AND RALF KORNHUBER†

Abstract. We present a new class of iterative schemes for large scale set-valued saddle point problems as arising, e.g., from optimization problems in the presence of linear and inequality constraints. Our algorithms can be regarded either as nonsmooth Newton-type methods for the nonlinear Schur complement or as Uzawa-type iterations with active set preconditioners. Numerical experiments with a control constrained optimal control problem and a discretized Cahn–Hilliard equation with obstacle potential illustrate the reliability and efficiency of the new approach.

Key words. set-valued saddle point problems, nonsmooth Newton methods, Uzawa algorithms, active set preconditioners

AMS subject classifications. 49M29, 65H20, 65N22, 90C46

DOI. 10.1137/060671012

1. Introduction. We consider the iterative solution of large scale saddle point problems of the form

$$(1.1) \quad u^* \in \mathbb{R}^n, w^* \in \mathbb{R}^m : \quad \begin{pmatrix} F & B^T \\ B & -C \end{pmatrix} \begin{pmatrix} u^* \\ w^* \end{pmatrix} \ni \begin{pmatrix} f \\ g \end{pmatrix},$$

where B and C are suitable matrices and the set-valued operator $F = \partial\varphi$ stands for the subdifferential of a strictly convex functional φ . Such kind of problems typically arise from the discretization of optimization or optimal control problems governed by partial differential equations with inequality constraints (cf., e.g., [32, 45]). In the case of a quadratic objective functional, we get

$$(1.2) \quad F = A + \partial I_K,$$

where I_K is denoting the indicator functional of the admissible set K , A is a self-adjoint positive definite, sometimes even diagonal matrix, and $C = 0$. Another rich and still growing class of problems of the form (1.1) consists of discretized phase field models, such as Cahn–Hilliard equations [5, 6, 8, 18, 19], Penrose–Fife equations [10], or Stefan-type problems [48]. For example, discretization of Cahn–Hilliard equations with logarithmic potential leads to the single-valued but singularly perturbed nonlinearity $F(u) = Au + T \log((1+u)/(1-u))$ where the logarithmic term is understood componentwise. Nonlinearities of the form (1.2) occur as singular limit for vanishing temperature T . The matrices A and C are essentially stiffness matrices of the Laplacian with A augmented by a nonlocal term reflecting mass conservation. Other possible applications include discretized plasticity problems [21, 43].

Saddle point problems of the form (1.1) with single-valued, Lipschitz continuous nonlinearities F have been considered in [12, 27]. Interior point methods (cf., e.g.,

*Received by the editors October 2, 2006; accepted for publication (in revised form) October 8, 2008; published electronically February 25, 2009. This work was funded in part by the Deutsche Forschungsgemeinschaft (DFG) under contract Ko 1806/3-1 and by the DFG Research Center Math-eon.

<http://www.siam.org/journals/sinum/47-2/67101.html>

†Institut für Mathematik II, Freie Universität Berlin, Arnimallee 6, D - 14195 Berlin, Germany (graeser@math.fu-berlin.de, kornhuber@math.fu-berlin.de).

[50, 51]) are based on suitable regularizations of set-valued nonlinearities (1.2). It is not immediately clear how this strategy should be generalized to single-valued but singularly perturbed nonlinearities. Existing primal-dual active set methods [26, 46] are based on the elimination of the state variables u_s and an active set approach to the resulting constrained minimization problem for the controls u_c . These methods are applicable to (1.1) with $u = (u_s, u_c)$, provided that the corresponding partitioning of $B = (B_s, B_c)$ generates an invertible matrix B_s , that the set-valued nonlinearity (1.2) constrains only u_c , and finally that $C = 0$. For example, discretized Cahn–Hilliard equations have none of these properties.

The novel approach presented in this paper relies on convexity rather than smoothness. It is motivated by the fact that a variety of practically relevant nonlinearities F can be either inverted in closed form or efficiently inverted by multigrid methods. This includes, e.g., the nonlinearities mentioned above [4, 3, 24, 30, 31, 29].

The basic idea is to reformulate (1.1) as an unconstrained convex minimization problem for the dual unknown w . The gradient of the objective functional h is just the nonlinear Schur complement H of (1.1) and, thus, involves F^{-1} . Minimization of h is carried out by well-known gradient-related descent methods (cf., e.g., [36, 37, 38]). Global convergence is enforced by standard Armijo damping [2] for simplicity. We particularly concentrate on nonsmooth Newton or Newton-like methods for nonlinearities of the form (1.2) taking into account that the nonlinear Schur complement H is Lipschitz but not differentiable in the classical sense. We prove global convergence and local exactness. Inexact versions are shown to be globally convergent.

In the special case of discretized optimal control problems with control constraints and diagonal matrix A , our algorithms reduce to well-known primal-dual active set methods [25]. Hence, the algorithms presented in this paper can be regarded as a new variational approach to primal-dual active set strategies, thus, providing a natural globalization and generalization of these methods. Extensions to single-valued but singularly perturbed nonlinearities F will be presented in a forthcoming paper [23]. Our approach also sheds new light on well-established algorithms in computational plasticity [49].

From a computational point of view, our algorithms can be reinterpreted as nonlinear Uzawa iterations with active set preconditioners [22]. For nonlinearities of the form (1.2), each iteration step requires the detection of the actual active set of $u^\nu = F^{-1}(f - B^T w^\nu)$ (not of u^ν itself!) and the sufficiently accurate evaluation of a corresponding linear saddle point problem (the actual preconditioner). We found in our numerical experiments with a discretized Cahn–Hilliard equation that, for bad initial iterates, the overall computational work was dominated by Armijo damping, because each Armijo test involves the exact evaluation of F^{-1} , i.e., the solution of a discrete elliptic obstacle problem. For reasonable initial iterates as obtained, e.g., from the preceding time step, almost no damping was necessary. In this case the (inexact) evaluation of the linear saddle point problem clearly dominated the overall computational cost.

The paper is organized as follows. After some notation and a precise formulation of the assumptions, we derive the equivalent unconstrained minimization problem which is fundamental for the rest of this paper. In section 3, we recall some general convergence results for gradient-related descent methods for unconstrained minimization, including damping strategies and inexact variants. Then we concentrate on the selection of suitable descent directions for the special case of nonlinearities of the form (1.2). More precisely, we investigate the B-subdifferential of F and later of H , giving rise to various nonsmooth Newton-type methods. The main convergence results are

collected in Theorems 4.1–4.3. Section 5 provides a more tangible reformulation of these abstract schemes in terms of quadratic obstacle problems and linear saddle point problems. Inexact evaluation of both of these subproblems and a heuristic damping strategy are also discussed. In our numerical computations, we consider a control constrained optimal control problem and a discretized Cahn–Hilliard equation. We found superlinear convergence and finite termination, supporting our theoretical findings.

2. Set-valued saddle point problems.

2.1. General assumptions and notation. Let $\langle \cdot, \cdot \rangle$ denote the euclidian inner product on \mathbb{R}^m . We equip \mathbb{R}^m with the norm $\|\cdot\|_M$,

$$\|x\|_M^2 = \langle Mx, x \rangle, \quad x \in \mathbb{R}^m,$$

induced by a fixed symmetric, positive definite (s.p.d.) matrix $M \in \mathbb{R}^{m,m}$. Linear mappings will be identified with their matrix representations with respect to the canonical basis vectors e_i with the coefficients $(e_i)_j = \delta_{i,j}$ (Kronecker- δ). Elements x' of the dual space $(\mathbb{R}^m)'$ will be represented as $x' = \langle x, \cdot \rangle$ with suitable $x \in \mathbb{R}^m$. Hence, using

$$|x'(y)| = |\langle x, y \rangle| \leq \|M^{-\frac{1}{2}}x\| \|M^{\frac{1}{2}}y\| = \|x\|_{M^{-1}} \|y\|_M,$$

the dual space $(\mathbb{R}^m, \|\cdot\|_M)'$ is identified with $(\mathbb{R}^m, \|\cdot\|_{M^{-1}})$.

We impose the following conditions on the saddle point problem (1.1).

- (A1) $F = \partial\varphi$ is the subdifferential of a proper, lower semicontinuous, strictly convex functional $\varphi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. The inverse $F^{-1} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is single-valued and Lipschitz continuous.
- (A2) $C \in \mathbb{R}^{m,m}$ is symmetric, positive semidefinite.
- (A3) $B \in \mathbb{R}^{m,n}$.
- (A4) The saddle point problem (1.1) has a unique solution.

Nonlinearities F satisfying condition (A1) occur, e.g., in discretized Cahn–Hilliard equations with logarithmic potential [5]. Later on, we will concentrate on the special case

$$F = A + \partial I_K,$$

where $A \in \mathbb{R}^{n,n}$ is s.p.d. and I_K denotes the indicator functional of a closed convex set K . In this case, (A1) holds with

$$\varphi(x) = \frac{1}{2} \langle Ax, x \rangle + I_K,$$

and $x = F^{-1}(y)$ is the unique solution of the variational inequality

$$(2.1) \quad x \in K : \quad \langle Ax - y, v - x \rangle \geq 0 \quad \forall v \in K.$$

It is well known that the corresponding mapping $F^{-1} : (\mathbb{R}^n, \|\cdot\|_{A^{-1}}) \rightarrow (\mathbb{R}^n, \|\cdot\|_A)$ is Lipschitz continuous with constant $L_{F^{-1}} \leq 1$ (cf., e.g., [28, p. 24]).

2.2. Nonlinear Schur complement and unconstrained minimization. Our aim is to reformulate the given saddle point problem as an *unconstrained* minimization problem. In the first step, the inclusion (1.1) is transformed into a single-valued equation.

PROPOSITION 2.1. *The saddle point problem (1.1) is equivalent to*

$$(2.2) \quad w^* \in \mathbb{R}^m : \quad H(w^*) = 0$$

with the Lipschitz continuous mapping

$$(2.3) \quad H(w) = -BF^{-1}(f - B^T w) + Cw + g, \quad w \in \mathbb{R}^m.$$

Proof. Using (A1), the equivalence is easily obtained by straightforward block elimination. Lipschitz continuity is clear since H consists of a sum and a composition of the Lipschitz continuous function F^{-1} with linear and constant functions. \square

The operator H can be regarded as a nonlinear version of the well-known Schur complement. In contrast to the linear case, the right-hand side f cannot be separated from the part depending on w . Note that H is single-valued, because $F^{-1} = (\partial\varphi)^{-1}$ is single-valued or, equivalently, the minimization of φ on \mathbb{R}^n admits a unique solution.

THEOREM 2.1. *There is a Fréchet-differentiable, convex functional $h : \mathbb{R}^m \rightarrow \mathbb{R}$ with the property $\nabla h = H$ and the representation*

$$(2.4) \quad h(w) = -\mathcal{L}(F^{-1}(f - B^T w), w), \quad w \in \mathbb{R}^m,$$

where

$$\mathcal{L}(u, w) = \varphi(u) - \langle f, u \rangle + \langle Bu - g, w \rangle - \frac{1}{2} \langle Cw, w \rangle$$

denotes the Lagrange functional associated with (1.1).

Proof. The polar (or conjugate) functional φ^* of φ is convex and, by Corollary 5.2 in [17, p. 22], has the property $\partial\varphi^* = (\partial\varphi)^{-1} = F^{-1}$. Since F^{-1} is single-valued, φ^* is Gâteaux-differentiable. The continuity of F^{-1} implies that φ^* is even Fréchet-differentiable with $\nabla\varphi^* = F^{-1}$. Setting

$$(2.5) \quad h(w) = \varphi^*(f - B^T w) + \frac{1}{2} \langle Cw, w \rangle + \langle g, w \rangle$$

we immediately get $\nabla h = H$ using the chain rule. By the definition of φ^* we have

$$\begin{aligned} \varphi^*(y) &= \sup_{x \in \mathbb{R}^n} (\langle y, x \rangle - \varphi(x)) = - \inf_{x \in \mathbb{R}^n} (\varphi(x) - \langle y, x \rangle) \\ &= - (\varphi(F^{-1}(y)) - \langle y, F^{-1}(y) \rangle), \quad y \in \mathbb{R}^n. \end{aligned}$$

Inserting this representation with $y = f - B^T w$ into (2.5), we get (2.4).

The convexity of φ implies the monotonicity of F^{-1} . In combination with the nonnegativity of C we get

$$(2.6) \quad \begin{aligned} \langle w_1 - w_2, H(w_1) - H(w_2) \rangle &= \langle (f - B^T w_1) - (f - B^T w_2), F^{-1}(f - B^T w_1) - F^{-1}(f - B^T w_2) \rangle \\ &\quad + \langle C(w_1 - w_2), w_1 - w_2 \rangle \geq 0 \end{aligned}$$

so that H is monotone. Therefore, h is convex. \square

Assuming, in addition to (A2), that C is positive definite, it is not difficult to show that h is strongly convex; i.e., there is a constant $\mu > 0$ such that

$$(2.7) \quad h(\lambda x + (1 - \lambda)y) \leq \lambda h(x) + (1 - \lambda)h(y) - \lambda(1 - \lambda) \frac{\mu}{2} \|x - y\|_M^2 \quad \forall \lambda \in [0, 1]$$

holds for all $x, y \in \mathbb{R}^m$. In general, however, h is not even strictly convex so that we had to require uniqueness separately.

Combining Proposition 2.1 with Theorem 2.1, we are ready to state the main result of this section.

COROLLARY 2.1. *The set-valued saddle point problem (1.1) is equivalent to the unconstrained convex minimization problem*

$$(2.8) \quad w^* \in \mathbb{R}^m : \quad h(w^*) \leq h(w) \quad \forall w \in \mathbb{R}^m.$$

Recall that the functional h is differentiable with Lipschitz continuous gradient $H = \nabla h$. However, the actual evaluation of $h(w)$ and $\nabla h(w)$ might be expensive, because it involves the solution of $F(u) = f - B^T w$.

3. Gradient-related methods. Exploiting Corollary 2.1, existing algorithms for the unconstrained minimization of convex, differentiable functionals now can be utilized to solve the constrained saddle point problem (1.1). In this section, we consider the fairly general class of gradient-related descent methods (see, for example, [37]). In agreement with section 2.2, we assume that $h : \mathbb{R}^m \rightarrow \mathbb{R}$ denotes a convex functional with Lipschitz continuous Fréchet derivative ∇h and the unique minimizer $w^* \in \mathbb{R}^m$.

3.1. Global convergence results. We consider the iteration

$$(3.1) \quad w^{\nu+1} = w^\nu + \rho^\nu d^\nu, \quad \nu = 0, 1, \dots,$$

with given initial guess $w^0 \in \mathbb{R}^m$. In each step, first a search direction d^ν is chosen according to the actual iterate w^ν and then a step size ρ^ν is fixed according to w^ν and d^ν , i.e.,

$$(3.2) \quad d^\nu = d(\nu, w^\nu), \quad \rho^\nu = \rho(\nu, w^\nu, d^\nu), \quad \nu = 0, 1, \dots,$$

with suitable mappings d, ρ .

The search directions d^ν should allow for a sufficient descent of h .

DEFINITION 3.1. *The search directions $d^\nu = d(\nu, w^\nu)$, $\nu \in \mathbb{N}$, are called gradient-related descent directions if for any sequence $(w^\nu) \subset \mathbb{R}^m$ the conditions*

$$(3.3) \quad \nabla h(w^\nu) = 0 \iff d^\nu = 0 \quad \forall \nu \in \mathbb{N}$$

and

$$(3.4) \quad -\langle \nabla h(w^\nu), d^\nu \rangle \geq c_D \|\nabla h(w^\nu)\|_{M^{-1}} \|d^\nu\|_M \quad \forall \nu \in \mathbb{N}$$

hold with a constant $c_D > 0$ independent of ν .

Note that the preconditioned gradients $d^\nu = -M^{-1}\nabla h(w^\nu)$ satisfy (3.4) with equality and $c_D = 1$. Obviously, (3.4) implies

$$(3.5) \quad -\langle \nabla h(w^\nu), d^\nu \rangle > 0$$

if $\nabla h(w^\nu) \neq 0$. Search directions $d^\nu = d(\nu, w^\nu)$, $\nu \in \mathbb{N}$, satisfying (3.3) and, instead of (3.4), the weaker condition (3.5) for arbitrary $(w^\nu) \in \mathbb{R}^m$ are called descent directions.

The step sizes ρ^ν should realize a sufficient portion of possible descent.

DEFINITION 3.2. *Let $d^\nu = d(\nu, w^\nu)$, $\nu \in \mathbb{N}$, be descent directions. Then the step sizes $\rho^\nu = \rho(\nu, w^\nu, d^\nu)$, $\nu \in \mathbb{N}$, are called efficient if for any sequence $(w^\nu) \subset \mathbb{R}^m$ the estimate*

$$(3.6) \quad h(w^\nu + \rho^\nu d^\nu) \leq h(w^\nu) - c_S \left(\frac{\langle \nabla h(w^\nu), d^\nu \rangle}{\|d^\nu\|_M} \right)^2$$

holds for all $\nu \in \mathbb{N}$ such that $\nabla h(w^\nu) \neq 0$ with a constant $c_S > 0$ independent of ν .

We are now ready to prove convergence.

THEOREM 3.1. *Assume that (3.2) provides gradient-related descent directions d^ν and efficient step sizes ρ^ν . Then, for arbitrary initial iterate $w^0 \in \mathbb{R}^m$, the iterates $w^\nu, \nu \in \mathbb{N}$, obtained from (3.1) converge to the minimizer w^* of h .*

Proof. Combining the properties of $d^\nu = d(\nu, w^\nu)$ and $\rho^\nu = \rho(\nu, w^\nu, d^\nu)$ we get

$$(3.7) \quad h(w^\nu) - h(w^{\nu+1}) \geq c_S c_D^2 \|\nabla h(w^\nu)\|_{M^{-1}}^2 \quad \forall \nu \in \mathbb{N}.$$

Since h has a global minimizer, the sequence $(h(w^\nu))$ is bounded from below and, by (3.7), is monotonically decreasing. Hence, $h(w^\nu)$ converges to some $h^* \in \mathbb{R}$. Using again (3.7), we get

$$(3.8) \quad 0 \leq c_S c_D^2 \|\nabla h(w^\nu)\|_{M^{-1}}^2 \leq h(w^\nu) - h(w^{\nu+1}) \rightarrow 0$$

so that $\nabla h(w^\nu)$ must tend to zero.

The section $S = \{w \in \mathbb{R}^m \mid h(w) \leq h(w^0)\}$ is bounded. Otherwise, there would be a sequence $(w_k) \subset S$ with the property $\lambda_k^{-1} := \|w_k - w^*\| \geq k$. Then, by compactness of the unit sphere with center w^* , the sequence $w'_k = w^* + (w_k - w^*)/\|w_k - w^*\|$ has a convergent subsequence $w'_{k_j} \rightarrow w^{**} \neq w^*$. By continuity and convexity of h this leads to

$$h(w^{**}) = \lim_{j \rightarrow \infty} h(w'_{k_j}) \leq \lim_{j \rightarrow \infty} \lambda_{k_j} h(w_{k_j}) + (1 - \lambda_{k_j})h(w^*) = h(w^*),$$

contradicting the uniqueness of w^* .

The section S is also closed and, therefore, compact. As a consequence, (w^ν) has a convergent subsequence $(w^{\nu_i}) \rightarrow w^{**}$. The continuity of ∇h provides $\nabla h(w^{**}) = 0$, and uniqueness implies $w^{**} = w^*$. Hence, each convergent subsequence must tend to w^* . This proves the assertion. \square

In the proof, we have made extensive use of Heine–Borel’s theorem which is restricted to finite dimensions. However, using weak compactness and the weak lower semicontinuity of h , weak convergence of the iterates w^ν can be shown by similar arguments in the infinite-dimensional case. Strong linear convergence can be shown in any dimension under the additional assumption that h is strongly convex. The proof is based on the following lemma summarizing well-known results (cf., e.g., [37]).

LEMMA 3.1. *Let h be strongly convex with constant $\mu > 0$. Then h satisfies the estimates*

$$(3.9) \quad \frac{\mu}{2} \|w - w^*\|_M^2 \leq h(w) - h(w^*) \leq \frac{1}{2\mu} \|\nabla h(w)\|_{M^{-1}}^2 \quad \forall w \in \mathbb{R}^m$$

with the minimizer w^* of h .

THEOREM 3.2. *Assume that the conditions of Theorem 3.1 are satisfied and, in addition, h is strongly convex with constant $\mu > 0$. Then the iterates $w^\nu, \nu \in \mathbb{N}$, produced by (3.1) satisfy the error estimate*

$$(3.10) \quad \|w^\nu - w^*\|_M^2 \leq q^\nu \frac{2}{\mu} (h(w^0) - h(w^*)),$$

where $0 \leq q = (1 - 2c_S c_D^2 \mu) < 1$ if $w^0 \neq w^*$.

The proof is straightforward using Lemma 3.1.

3.2. Damping strategies. A variety of algorithms for efficient step size control are available from surveys and textbooks like [16, 36, 37, 38]. For simplicity, we consider the standard Armijo strategy [2], [16, p. 121], and [37, p. 491] based on the actual decrease of the functional h . More precisely, for a fixed parameter $\delta \in (0, 1)$ and each $\nu \in \mathbb{N}$ a step size $\rho \geq 0$ is called *admissible* if

$$(3.11) \quad h(w^\nu + \rho d^\nu) \leq h(w^\nu) + \rho \delta \langle \nabla h(w^\nu), d^\nu \rangle$$

is satisfied.

PROPOSITION 3.1. *Let $(w^\nu) \subset \mathbb{R}^m$, and let $d^\nu = d(\nu, w^\nu)$, $\nu \in \mathbb{N}$, be descent directions. For suitably selected, fixed parameters $\alpha > 0$ and $\delta, \beta \in (0, 1)$ determine the step sizes $\rho^\nu = \rho(\nu, w^\nu, d^\nu) \geq 0$ by*

$$(3.12) \quad \rho^\nu = \max_{j \in \mathbb{N} \cup \{0\}} \left\{ \rho = \alpha_\nu \beta^j \mid \alpha_\nu \geq -\alpha \frac{\langle \nabla h(w^\nu), d^\nu \rangle}{\|d^\nu\|_M^2}, \rho \text{ admissible} \right\}$$

if $d^\nu \neq 0$ and set $\rho^\nu = 0$ otherwise. Then the efficiency condition (3.6) holds with

$$(3.13) \quad c_S = \delta \min \left\{ \alpha, \beta \left(\frac{1-\delta}{L} \right) \right\}.$$

Here L stands for the Lipschitz constant of ∇h , i.e.,

$$(3.14) \quad \|\nabla h(v) - \nabla h(w)\|_{M^{-1}} \leq L \|v - w\|_M \quad \forall v, w \in \mathbb{R}^m.$$

The proof of Proposition 3.1 adopts standard arguments, e.g., from [37]. Starting with $j = 0$, efficient step sizes can be computed from (3.12) by a finite number of tests. Observe that each of these tests might be expensive, because it requires the evaluation of h and, therefore, the evaluation of F^{-1} (cf. Theorem 2.1).

3.3. Inexact versions. We consider inexact search directions \tilde{d}^ν . This means that for given ν and w^ν the exact evaluation $d^\nu = d(\nu, w^\nu)$ is replaced by some approximation

$$(3.15) \quad \tilde{d}^\nu = \tilde{d}(\nu, w^\nu)$$

based on some approximation \tilde{d} of the exact mapping d .

PROPOSITION 3.2. *Let $d^\nu = d(\nu, w^\nu)$ be gradient-related descent directions with constant c_D . Assume that the approximations $\tilde{d}^\nu = \tilde{d}(\nu, w^\nu)$ satisfy (3.3) and the accuracy condition*

$$(3.16) \quad \|d^\nu - \tilde{d}^\nu\|_M \leq c \|\tilde{d}^\nu\|_M \quad \forall \nu \in \mathbb{N}, \quad c < \frac{c_D}{2},$$

for any sequence (w^ν) . Then the approximations $\tilde{d}^\nu = \tilde{d}(\nu, w^\nu)$ are also gradient-related descent directions.

Proof. Let $(w^\nu) \subset \mathbb{R}^m$. Then the vectors $d^\nu = d(\nu, w^\nu)$, $\nu \in \mathbb{N}$, satisfy (3.4) and we have to prove a similar estimate for the approximations \tilde{d}^ν . This is trivial for $\tilde{d}^\nu = 0$. Note that (3.16) implies $d^\nu = 0$ in this case. In light of (3.3) there is only the remaining case $d^\nu, \tilde{d}^\nu \neq 0$. Some elementary calculations involving the Cauchy-Schwarz inequality and the triangle inequality yield

$$\left| \left\langle \frac{\nabla h(w^\nu)}{\|\nabla h(w^\nu)\|_{M^{-1}}}, \frac{d^\nu}{\|d^\nu\|_M} - \frac{\tilde{d}^\nu}{\|\tilde{d}^\nu\|_M} \right\rangle \right| \leq 2 \frac{\|d^\nu - \tilde{d}^\nu\|_M}{\|\tilde{d}^\nu\|_M}.$$

As $\|d^\nu - \tilde{d}^\nu\|_M / \|\tilde{d}^\nu\|_M \leq c < c_D/2$, it is clear that

$$-\langle \nabla h(w^\nu), \tilde{d}^\nu \rangle \geq \tilde{c}_D \|\nabla h(w^\nu)\|_{M^{-1}} \|\tilde{d}^\nu\|_M$$

with $\tilde{c}_D = c_D - 2c > 0$. \square

Usually, the constant c_D occurring in the accuracy condition (3.16) is not known. Replacing (3.16) by the asymptotic criterion

$$(3.17) \quad \lim_{\nu \rightarrow \infty} \frac{\|d^\nu - \tilde{d}^\nu\|_M}{\|\tilde{d}^\nu\|_M} = 0$$

the approximate directions \tilde{d}^ν have the desired property (3.4) for sufficiently large ν .

4. Nonsmooth Newton methods and related algorithms. We now consider the question of how to choose the descent directions $d^\nu = d(w^\nu)$. We will concentrate on preconditioned gradients of h or, more precisely, on directions of the form

$$(4.1) \quad d^\nu = -S_\nu^{-1} H(w^\nu), \quad H = \nabla h,$$

with suitable s.p.d. matrices $S_\nu = S(\nu, w^\nu)$. If H would be sufficiently smooth, the derivative

$$S_\nu = H'(w^\nu) : \mathbb{R}^m \rightarrow \mathbb{R}^m$$

would provide the classical Newton iteration. From our assumptions (A1)–(A4) and the definition (2.3), we cannot expect H' to exist. Hence, related concepts from nonsmooth analysis will be applied. To this end, (A1) is from now on replaced by the stronger condition (A1’):

(A1’) $F = A + \partial I_K$, where $A \in \mathbb{R}^{n,n}$ is s.p.d. and I_K denotes the indicator functional of the closed convex set

$$(4.2) \quad K = \{x \in \mathbb{R}^n \mid a \leq x \leq b\}, \quad a, b \in (\mathbb{R} \cup \{-\infty, \infty\})^n, \quad a < 0 < b.$$

Recall that F is the subdifferential of $\varphi(x) = \frac{1}{2} \langle Ax, x \rangle + I_K$ and Lipschitz continuous with constant $L \leq 1$ in this case. Nonlinearities F satisfying (A1’) occur, e.g., in discretized optimal control problems with inequality constraints [32, 45] or discretized phase field models with obstacle potentials [6, 8]. The condition $a < 0 < b$ causes no loss of generality and will be notationally convenient in what follows.

4.1. The B-subdifferential of F^{-1} . Let $c \in K$ with $K \subset \mathbb{R}^n$ defined in (4.2). We introduce the subset of all active indices

$$N_c^\bullet := \{i \in N \mid a_i = c_i \text{ or } c_i = b_i\}$$

of the index set $N = \{1, \dots, n\}$. The mapping $T_c : \mathbb{R}^n \rightarrow \mathbb{R}^n$, defined by

$$T_c x := \sum_{i \in N \setminus N_c^\bullet} x_i e_i, \quad x \in \mathbb{R}^n,$$

truncates all coefficients with active indices. Note that T_c is an orthogonal projection with respect to the euclidian scalar product $\langle \cdot, \cdot \rangle$. The finite set

$$\mathcal{C} := \{c \in K \mid (I - T_c)c = c\}$$

represents all possible configurations of active coefficients, i.e., of coefficients with active indices. The active coefficients of $x \in K$ are given by

$$(4.3) \quad T_c x := (I - T_x)x \in \mathcal{C}.$$

As $F : K \rightarrow \mathbb{R}^n$ is invertible, K and \mathbb{R}^n can be decomposed according to

$$(4.4) \quad K = \bigcup_{c \in \mathcal{C}} \mathcal{I}_c, \quad \mathbb{R}^n = \bigcup_{c \in \mathcal{C}} F(\mathcal{I}_c), \quad \mathcal{I}_c := \{x \in K \mid T_c x = c\},$$

based on the subsets \mathcal{I}_c of vectors with the same active coefficients. Note that

$$(I - T_c)x = c \quad \forall x \in \mathcal{I}_c, \quad c \in \mathcal{C}.$$

We now investigate the restriction of F to \mathcal{I}_c . To this end, it is convenient to introduce the mapping

$$(4.5) \quad \widehat{A}_c := T_c A T_c + I - T_c : \mathbb{R}^n \rightarrow \mathbb{R}^n.$$

Observe that $\widehat{A}_c : \text{ran } T_c \rightarrow \text{ran } T_c$ and \widehat{A}_c reduces to the identity on the orthogonal complement $\text{ran}(I - T_c)$. Hence,

$$(4.6) \quad \widehat{A}_c T_c = T_c A T_c = T_c \widehat{A}_c, \quad \widehat{A}_c (I - T_c) = I - T_c.$$

Using

$$\langle \widehat{A}_c x, y \rangle = \langle A T_c x, T_c y \rangle + \langle (I - T_c)x, (I - T_c)y \rangle$$

it is easy to show that \widehat{A}_c is s.p.d. Multiplying (4.6) by \widehat{A}_c^{-1} we obtain

$$(4.7) \quad \widehat{A}_c^{-1} T_c = T_c \widehat{A}_c^{-1}, \quad \widehat{A}_c^{-1} (I - T_c) = I - T_c.$$

LEMMA 4.1. *Let $c \in \mathcal{C}$. Then the restriction of F to \mathcal{I}_c takes the form*

$$(4.8) \quad F(x) = Ax + \sum_{i \in N_c^\bullet} [0, \infty) s_i(c) e_i, \quad x \in \mathcal{I}_c,$$

denoting

$$s_i(c) = \begin{cases} +1 & \text{if } c_i = b_i, \\ -1 & \text{if } c_i = a_i, \end{cases} \quad i \in N_c^\bullet.$$

Conversely, the restriction of F^{-1} to $F(\mathcal{I}_c)$ takes the form

$$(4.9) \quad F^{-1}(y) = T_c \widehat{A}_c^{-1} T_c y + (I - T_c \widehat{A}_c^{-1} T_c A) c, \quad y \in F(\mathcal{I}_c).$$

Proof. Let $x \in \mathcal{I}_c$. Using the representation

$$I_K(x) = \sum_{i \in N} I_{[a_i, b_i]}(x_i), \quad x = \sum_{i \in N} x_i e_i,$$

of the characteristic functional I_K , we immediately get (cf. [17, p. 26])

$$\partial I_K(x) = \sum_{i \in N} \partial I_{[a_i, b_i]}(x_i) e_i = \sum_{i \in N_c^\bullet} [0, \infty) s_i(c) e_i.$$

This proves (4.8).

Let $x \in \mathcal{I}_c$ and $y \in F(x)$. We apply T_c to the representation (4.8), insert the splitting $x = T_c x + (I - T_c)x$, and use the identity $(I - T_c)x = c$ to obtain

$$T_c y = T_c A x = T_c A T_c x + T_c A c = \widehat{A}_c x - (I - T_c A)c.$$

Multiplication by \widehat{A}_c^{-1} and reordering terms, we get

$$(4.10) \quad x = \widehat{A}_c^{-1} T_c y + \widehat{A}_c^{-1} (I - T_c A)c.$$

The left identity in (4.7) yields

$$\widehat{A}_c^{-1} T_c = \widehat{A}_c^{-1} T_c T_c = T_c \widehat{A}_c^{-1} T_c.$$

Using $c = (I - T_c)c$ and the right identity in (4.7), we obtain

$$\widehat{A}_c^{-1} c = \widehat{A}_c^{-1} (I - T_c)c = (I - T_c)c = c.$$

Inserting these representations into (4.10) the assertion (4.9) follows. \square

As a consequence of (4.4) and (4.9), F^{-1} is piecewise affine linear on \mathbb{R}^n with the linear part $T_c \widehat{A}_c^{-1} T_c$ on each subset $F(\mathcal{I}_c)$, $c \in \mathcal{C}$. In the extreme case, $N_c^\bullet = N$, F^{-1} is even constant on $F(\mathcal{I}_c)$.

As F^{-1} is Lipschitz continuous, F^{-1} must be differentiable almost everywhere (cf. Rademacher's theorem [35]). Let $\mathcal{D}_{F^{-1}}$ denote the set where F^{-1} is differentiable. Then the B-subdifferential $\partial_B(F^{-1})$ (cf. [40, 46]) is defined by

$$\partial_B(F^{-1})(y) = \left\{ \lim_{\substack{y_n \rightarrow y \\ y_n \in \mathcal{D}_{F^{-1}}} } D(F^{-1})(y_n) \right\}.$$

Note that

$$\partial_B(F^{-1})(y) \subset \text{co} \partial_B(F^{-1})(y) = \partial(F^{-1})(y)$$

with $\partial(F^{-1})$ denoting Clarke's generalized derivative [13, Chapter 2].

PROPOSITION 4.1. *Let $y \in \mathbb{R}^n$ and $c = T_c(F^{-1}(y)) \in \mathcal{C}$. Then*

$$(4.11) \quad T_c \widehat{A}_c^{-1} T_c \in \partial_B(F^{-1})(y).$$

Proof. Note that $F^{-1}(y) \in \mathcal{I}_c$ by definition (4.4) of \mathcal{I}_c . Inserting the decomposition $x = T_c x + c$ of some arbitrary $x \in \mathcal{I}_c$ into (4.8), it turns out that $F(\mathcal{I}_c)$ is the parallelepiped translated from the origin by Ac and spanned by the nonzero column vectors of AT_c and of $I - T_c$ with coefficients $z_i \in (a_i, b_i)$, $i \in N \setminus N_c^\bullet$, and $z_i \in [0, \infty) s_i(c)$, $i \in N_c^\bullet$, respectively. Utilizing the identities $AT_c + I - T_c = \widehat{A}_c + (I - T_c)AT_c$, (4.7), and the orthogonality $T_c(I - T_c) = 0$, it is easily checked that

$$\left(\widehat{A}_c^{-1} - (I - T_c)AT_c \widehat{A}_c^{-1} \right) (AT_c + I - T_c) = I.$$

Hence, the interior of $F(\mathcal{I}_c)$ cannot be empty so that the convexity of $F(\mathcal{I}_c)$ yields

$$(4.12) \quad F(\mathcal{I}_c) \subset \overline{\text{int} F(\mathcal{I}_c)}.$$

If $y \in \text{int} F(\mathcal{I}_c)$, then the representation (4.9) implies

$$D(F^{-1})(y) = T_c \widehat{A}_c^{-1} T_c.$$

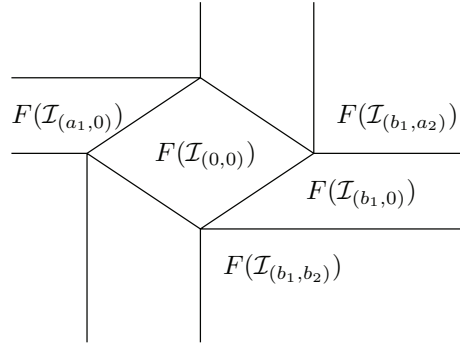


FIG. 4.1. Decomposition of \mathbb{R}^2 into parallelepipeds $F(\mathcal{I}_c)$, $c \in \mathcal{C}$.

If $y \in F(\mathcal{I}_c) \setminus \text{int } F(\mathcal{I}_c)$, then (4.12) implies that there is a sequence $(y_k) \subset \text{int } F(\mathcal{I}_c)$ with $y_k \rightarrow y$. Obviously,

$$\lim_{k \rightarrow \infty} DF^{-1}(y_k) = T_c \widehat{A}_c^{-1} T_c$$

which proves the assertion. \square

Figure 4.1 illustrates the decomposition of \mathbb{R}^n into the nondegenerating parallelepipeds $F(\mathcal{I}_c)$, $c \in \mathcal{C}$, for $n = 2$. The only bounded parallelepiped $F(\mathcal{I}_{(0,0)})$ is spanned by the column vectors of A .

4.2. Algorithms and convergence results. Proposition 4.1 suggests using B-subdifferentials $T_c \widehat{A}_c^{-1} T_c$, $c \in \mathcal{C}$, for the linearization of the Schur complement

$$H(w) = -BF^{-1}(f - B^T w) + Cw + g, \quad w \in \mathbb{R}^m,$$

as introduced in (2.3).

PROPOSITION 4.2. Assume that $\text{rank } B = n$. Then

$$(4.13) \quad S(c) = BT_c \widehat{A}_c^{-1} (BT_c)^T + C \in \partial_B H(w), \quad w \in \mathbb{R}^m,$$

where

$$(4.14) \quad c = c(w) = T_c F^{-1}(f - B^T w).$$

Proof. Let $G : \mathbb{R}^m \rightarrow \mathbb{R}^n$ be defined by $G(w) = F^{-1}(f - B^T w)$, $w \in \mathbb{R}^m$. We consider some fixed $w \in \mathbb{R}^m$ and $c = T_c G(w)$. As $\text{rank } B^T = n$, the mapping $B^T : \mathbb{R}^m \rightarrow \mathbb{R}^n$ is surjective. Hence, the preimage $G^{-1}(\mathcal{I}_c)$ of \mathcal{I}_c is still a nondegenerate parallelepiped. Therefore, we can use the same arguments as in the proof of Proposition 4.1 to show

$$-T_c \widehat{A}_c^{-1} T_c B^T \in \partial_B G(w).$$

As H is an affine transformation of G , the assertion follows. \square

Simple counterexamples show that (4.13) might not hold for $\text{rank } B^T < n$.

Let us check whether $S(c)$ is invertible. We immediately get

$$\langle S(c)x, y \rangle = \left\langle \widehat{A}_c^{-1} (BT_c)^T x, (BT_c)^T y \right\rangle + \langle Cx, y \rangle, \quad x, y \in \mathbb{R}^m.$$

Hence, $S(c)$ is symmetric and positive semidefinite. It is a sufficient (but not necessary) condition for the regularity of $S(c)$ that C is s.p.d.

LEMMA 4.2. *Assume that $S(c)$ is s.p.d. for all $c \in \mathcal{C}$. Then h is strongly convex.*

Proof. Consider $G(w) = F^{-1}(f - B^T w)$ as already introduced in the proof of Proposition 4.2. Let $c \in \mathcal{C}$. Then for all $w \in G^{-1}\mathcal{I}_c$ the representation $\nabla h(w) = H(w) = S(c)w + \tilde{g}(c)$ holds with suitable $\tilde{g}(c) \in \mathbb{R}^m$ independent of w (cf. Lemma 4.1). As $S(c)$ is s.p.d., we have

$$(4.15) \quad \langle S(c)w, w \rangle \geq \gamma_c \|w\|_M^2 \quad \forall w \in G^{-1}\mathcal{I}_c$$

with some constant $\gamma_c > 0$. This means that h is quadratic and strongly convex on each preimage $G^{-1}\mathcal{I}_c$. We now show strong convexity on the whole $\mathbb{R}^m = \bigcup_{c \in \mathcal{C}} G^{-1}\mathcal{I}_c$ with the constant $\mu = \min_{c \in \mathcal{C}} \gamma_c > 0$. To this end, we define the scalar functions

$$\begin{aligned} \psi_1(\lambda) &= \|x - y\|_M^{-2} h(\lambda x + (1 - \lambda)y), \\ \psi_2(\lambda) &= \|x - y\|_M^{-2} (\lambda h(x) + (1 - \lambda)h(y)) - \frac{\mu}{2} \lambda(1 - \lambda), \quad \lambda \in [0, 1], \end{aligned}$$

with some fixed $x \neq y \in \mathbb{R}^m$. It is sufficient to show $\psi_1 \leq \psi_2$. Obviously, ψ_1 is piecewise quadratic, ψ_2 is quadratic, and $\psi_1(\lambda) = \psi_2(\lambda)$ at the boundary $\lambda = 0, 1$. By definition,

$$\psi_1''(\lambda) \geq \min_{c \in \mathcal{C}} \gamma_c = \psi_2''(\lambda)$$

holds for almost all $\lambda \in [0, 1]$. Now $\psi_1 \leq \psi_2$ follows either from elementary arguments or from a weak maximum principle (cf. [20, Theorem 9.1]) as applied to $\psi_1 - \psi_2$. \square

We are ready to state the basic convergence result of this section.

THEOREM 4.1. *Assume that $S(c)$ is s.p.d. for all $c \in \mathcal{C}$. Then, for arbitrary initial iterate $w^0 \in \mathbb{R}^m$, the damped nonsmooth Newton-type method, as obtained by inserting the search directions*

$$(4.16) \quad d^\nu = -S_\nu^{-1} H(w^\nu), \quad H(w^\nu) = \nabla h(w^\nu),$$

with

$$S_\nu = S(c^\nu), \quad c^\nu = T_{\mathcal{C}} F^{-1}(f - B^T w^\nu),$$

and step sizes ρ^ν selected according to Proposition 3.1 into the basic algorithm (3.1), converges linearly to the solution w^* of (2.8). If (2.8) is nondegenerate in the sense that

$$(4.17) \quad F^{-1}(f - B^T w^*) \in \text{int } \mathcal{I}_{c^*}, \quad c^* = T_{\mathcal{C}} F^{-1}(f - B^T w^*),$$

then the algorithm terminates after a finite number of steps.

Proof. To prove convergence by Theorem 3.1, we have only to show that the directions d^ν as defined in (4.16) are gradient-related. Let $c \in \mathcal{C}$. Denoting the norm of the linear mapping $S(c) : (\mathbb{R}^m, \|\cdot\|_M)$ to $(\mathbb{R}^m, \|\cdot\|_{M^{-1}})$ by Γ_c and using the coercivity (4.15), we get

$$\langle \nabla h(w), S(c)^{-1} \nabla h(w) \rangle \geq \gamma_c \|S(c)^{-1} \nabla h(w)\|_M^2 \geq \frac{\gamma_c}{\Gamma_c} \|S(c)^{-1} \nabla h(w)\|_M \|\nabla h(w)\|_{M^{-1}}$$

for all $w \in \mathbb{R}^m$. Since \mathcal{C} is finite, (3.4) now holds with

$$c_D := \min_{c \in \mathcal{C}} \frac{\gamma_c}{\Gamma_c} > 0.$$

Utilizing Lemma 4.2, linear convergence immediately follows from Theorem 3.2. If (2.8) is nondegenerate, then $F^{-1}(f - B^T w^{\nu_0}) \in \mathcal{I}_{c^*}$ holds for sufficiently large ν_0 . This implies $w^{\nu_0+1} = w^*$, because H is affine on all w with $F^{-1}(f - B^T w) \in \mathcal{I}_{c^*}$. \square

Under the additional assumption $\text{rank } B = n$, we obtain (cf. Proposition 4.2)

$$S_\nu = S(c^\nu) \in \partial_B H(w) \quad \forall \nu \in \mathbb{N}$$

and, therefore, a nonsmooth Newton method. In order to allow for local superlinear or even quadratic convergence (cf. [39, 40]), it is essential that $\rho^\nu \rightarrow 1$ for $\nu \rightarrow \infty$ which, in general, does not hold for the standard Armijo strategy. Hence, nonsmooth analogues of well-known affine-invariant damping strategies [16, section 3.4] will be the subject of future research.

If h is not strongly convex, then $S(c)$ is not invertible for certain c . Therefore, we now modify $S(c)$ to ensure invertibility.

By symmetry we have $\ker S(c) = (\text{ran } S(c))^\perp$. We introduce the mapping $I(c) : \mathbb{R}^m \rightarrow \mathbb{R}^m$ by

$$(4.18) \quad I(c)|_{\ker S(c)} = I|_{\ker S(c)}, \quad I(c)|_{\text{ran } S(c)} = 0,$$

to define

$$(4.19) \quad \widehat{S}(c) = S(c) + I(c), \quad c \in \mathcal{C}.$$

Observe that the orthogonal subspaces $\ker S(c)$ and $\text{ran } S(c)$ are invariant with respect to $\widehat{S}(c)$. Decomposing x, y into their components from $\ker S(c)$ and $\text{ran } S(c)$, respectively, we get

$$\langle \widehat{S}(c)x, y \rangle = \langle S(c)x_{\text{ran}}, y_{\text{ran}} \rangle + \langle x_{\ker}, y_{\ker} \rangle$$

so that $\widehat{S}(c)$ is s.p.d. Note that $\widehat{S}(c)$ can be rewritten as

$$\widehat{S}(c) = S(c) + \sum_{i=1}^l \frac{k_i k_i^T}{\|k_i\|^2}$$

with k_1, \dots, k_l denoting an orthogonal basis of $\ker S(c)$. If $S(c)$ is replaced by $\widehat{S}(c)$, then nonsmooth Newton steps are carried out on $\text{ran } S_\nu$, i.e., if possible, while simple gradient steps are performed on $\ker S_\nu$.

THEOREM 4.2. *For arbitrary initial iterate $w^0 \in \mathbb{R}^m$, the nonsmooth Newton-like method, as obtained by inserting the search directions*

$$(4.20) \quad d^\nu = -\widehat{S}_\nu^{-1} H(w^\nu), \quad H(w^\nu) = \nabla h(w^\nu),$$

with

$$\widehat{S}_\nu = \widehat{S}(c^\nu), \quad c^\nu = T_{\mathcal{C}} F^{-1}(f - B^T w^\nu),$$

and step sizes ρ^ν selected according to Proposition 3.1 into the basic algorithm (3.1), converges to the solution w^* of (2.8). If the problem (2.8) is nondegenerate in the sense of (4.17) and $S(c^*)$, $c^* = T_{\mathcal{C}} F^{-1}(f - B^T w^*)$, is positive definite, then the algorithm terminates after a finite number of steps.

Proof. Using the same arguments as in the proof of Theorem 4.1 it can be shown that the modified search directions d^ν defined in (4.20) are gradient-related. Hence, convergence is a consequence of Theorem 3.1. Finite termination also follows by the reasoning as in the proof of Theorem 4.1. \square

Remark. In general, one would expect local superlinear convergence of a Newton-like method. However, straightforward application of this concept makes no sense in the present, piecewise affine case, because, in a sufficiently small neighborhood, the algorithms terminate with the exact solution after one step. Further insight could be obtained by showing that the domain of superlinear convergence is larger than the domain of one step termination and, in particular, does not depend on the dimension m .

In order to determine $d^\nu = -\widehat{S}_\nu^{-1}H(w^\nu)$, a linear saddle point problem associated with the Schur complement matrix $\widehat{S}_\nu = \widehat{S}(c^\nu)$ has to be solved (see section 5 below). A sufficiently accurate iterative solution preserves convergence.

THEOREM 4.3. *For arbitrary initial iterate $w^0 \in \mathbb{R}^m$, the inexact nonsmooth Newton-like method, as obtained by inserting search directions \tilde{d}^ν which satisfy (3.3) and the accuracy condition (3.16) with $d^\nu = -\widehat{S}_\nu^{-1}H(w^\nu)$ and step sizes ρ^ν selected according to Proposition 3.1 into the basic algorithm (3.1), converges to the solution w^* of (2.8). The iterates converge linearly if h is strongly convex, e.g., for positive definite C .*

Proof. As the directions d^ν are gradient-related (see the proof of Theorem 4.2 above) the convergence is an immediate consequence of Proposition 3.2. If C is positive definite, then h is strongly convex. In this case linear convergence follows from Theorem 3.2. \square

5. Computational aspects.

5.1. Preconditioned Uzawa methods. Denoting $u^\nu := F^{-1}(f - B^T w^\nu)$ the Newton-like method as introduced in Theorem 4.2 can be interpreted as the preconditioned Uzawa iteration

$$(5.1a) \quad u^\nu = F^{-1}(f - B^T w^\nu),$$

$$(5.1b) \quad w^{\nu+1} = w^\nu + \rho^\nu \widehat{S}_\nu^{-1}(Bu^\nu - Cw^\nu - g)$$

for the saddle point problem (1.1).

The first substep (5.1a) amounts to the solution of the quadratic obstacle problem

$$(5.2) \quad u^\nu = \arg \min_{v \in K} \left(\frac{1}{2} \langle Av, v \rangle - \langle f - B^T w^\nu, v \rangle \right),$$

which has been extensively treated in the literature (cf., e.g., [14, 21, 30, 34, 44, 3]).

Inserting the definitions (4.19) and (4.13) of \widehat{S}_ν and $S(c^\nu)$, the evaluation of the preconditioned residual

$$d^\nu = \widehat{S}_\nu^{-1}(Bu^\nu - Cw^\nu - g)$$

in the second substep (5.1b) can be rewritten as the solution of the linear saddle point problem

$$(5.3) \quad \begin{pmatrix} \widehat{A}_{c^\nu} & (BT_{c^\nu})^T \\ (BT_{c^\nu}) & -(C + I(c^\nu)) \end{pmatrix} \begin{pmatrix} \tilde{u}^\nu \\ d^\nu \end{pmatrix} = \begin{pmatrix} 0 \\ g + Cw^\nu - Bu^\nu \end{pmatrix},$$

where, according to (4.3), $c^\nu = T_C u^\nu$ identifies the active coefficients of u^ν . Recall that \widehat{A}_{c^ν} is obtained from A by replacing the i th row and the i th column by the unit

vector e_i if i is active, i.e., $c_i \in \{a_i, b_i\}$. BT_{c^ν} is obtained from B by annihilating the i th column if i is active. Finally, $I(c^\nu)$ has been defined in (4.18). Thus, the preconditioner \widehat{S}_ν is approximating the original set-valued operator by essentially eliminating the actual active coefficients [22]. A sufficiently accurate, iterative solution of (5.3) preserves convergence of the overall iteration (5.1) (cf. Theorem 4.3). In particular, multigrid methods have been investigated in [9, 42, 47, 52, 53].

5.2. Inexact evaluation of F^{-1} . The exact solution $u^\nu = F^{-1}(f - B^T w^\nu)$ appears on the right-hand side of the linear saddle point problem (5.3). However, it turns out that the preconditioned residual can be computed from w^ν and the *active coefficients* c^ν of u^ν alone.

PROPOSITION 5.1. *For given $w^\nu \in \mathbb{R}^m$ and $c^\nu = T_C u^\nu$ let $(\tilde{u}^\nu, \tilde{w}^\nu)$ be the solution of*

$$(5.4) \quad \begin{pmatrix} \widehat{A}_{c^\nu} & (BT_{c^\nu})^T \\ (BT_{c^\nu}) & -(C + I(c^\nu)) \end{pmatrix} \begin{pmatrix} \tilde{u}^\nu \\ \tilde{w}^\nu \end{pmatrix} = \begin{pmatrix} T_{c^\nu} f - T_{c^\nu} A c^\nu \\ g - B c^\nu - I(c^\nu) w^\nu \end{pmatrix}.$$

Then

$$\widehat{S}_\nu^{-1}(B u^\nu - C w^\nu - g) = \tilde{w}^\nu - w^\nu.$$

Proof. Let $d^\nu = \widehat{S}_\nu^{-1}(B u^\nu - C w^\nu - g) = -\widehat{S}_\nu^{-1} H(w^\nu)$. Utilizing the definitions (2.3) of H , the representation (4.9) of F^{-1} , and the definitions (4.19) and (4.13) of \widehat{S}_ν and $S(c^\nu)$, respectively, we get

$$\begin{aligned} \widehat{S}_\nu(w^\nu + d^\nu) &= \widehat{S}_\nu w^\nu - H(w^\nu) \\ &= \widehat{S}_\nu w^\nu + B T_{c^\nu} \widehat{A}_{c^\nu}^{-1} T_{c^\nu} (f - B^T w^\nu - A c^\nu) + B c^\nu - C w^\nu - g \\ &= (B T_{c^\nu}) \widehat{A}_{c^\nu}^{-1} (T_{c^\nu} f - T_{c^\nu} A c^\nu) - (g - B c^\nu - I(c^\nu) w^\nu). \end{aligned}$$

Hence, $\tilde{w}^\nu = w^\nu + d^\nu$ is the second component of the solution of (5.4). This completes the proof. \square

Usually, the active coefficients c^ν of u^ν can be computed much faster than u^ν itself: For nondegenerate problems monotone multigrid methods [30] or even simple projected Gauß–Seidel relaxations [21, Chapter V] provide c^ν in a finite number of steps. Using the a priori estimate (cf., e.g., [28, p. 24])

$$(5.5) \quad \|u^* - u^\nu\|_A \leq \|B(w^* - w^\nu)\|_{A^{-1}}$$

the accuracy of u^ν can be estimated without actual computation of u^ν .

In order to determine efficient step sizes ρ^ν by Armijo’s strategy (cf. Proposition 3.1), we have to evaluate F^{-1} for each test $j = 0, \dots$ in (3.12). Though it is possible to develop straightforward inexact variants of existing damping strategies, e.g., of the Curry–Altmann principle [37, p. 483], an even cheaper heuristic strategy will be applied in the numerical computations to be reported below: We set $\rho^\nu = 1$ if the condition

$$(5.6) \quad \|d^\nu\|_M \leq \sigma \|d^{\nu-1}\|_M$$

holds with some fixed parameter $\sigma \in (0, 1)$ and compute ρ^ν according to Armijo’s strategy otherwise. Note that it is not hard to show convergence if (5.6) holds for $d^\nu = \widehat{S}_\nu^{-1} H(w^\nu)$ and all $\nu \in \mathbb{N}$.

6. Numerical results. In the following examples $\Omega = (0, 1) \times (0, 1)$ denotes the unit square and the triangulation \mathcal{T}_J of Ω is resulting from J uniform refinement steps as applied to the initial partition \mathcal{T}_0 consisting of four congruent subtriangles. The uniform refinement \mathcal{T}_{j+1} of \mathcal{T}_j is obtained by connecting the midpoints of all triangles $T \in \mathcal{T}_j$. Hence, the mesh size of \mathcal{T}_J is $h_J = 2^{-J}$. The sequence $\mathcal{T}_0 \subset \mathcal{T}_1 \subset \dots \subset \mathcal{T}_J$ of triangulations gives rise to a nested sequence $\mathcal{S}_0 \subset \mathcal{S}_1 \subset \dots \subset \mathcal{S}_J$ of finite element spaces

$$\mathcal{S}_j = \{v \in C(\overline{\Omega}) \mid v|_T \text{ is linear } \forall T \in \mathcal{T}_j\} \subset H^1(\Omega), \quad j = 0, \dots, J.$$

The standard nodal basis of \mathcal{S}_J is denoted by $\lambda_p, p \in \mathcal{N}_J$, where \mathcal{N}_J stands for the set of vertices of \mathcal{T}_J . Homogeneous Dirichlet conditions give rise to the subspace

$$\mathcal{S}_{J,0} = \text{span}\{\lambda_p \mid p \in \mathcal{N}_{J,0}\} \subset H_0^1(\Omega), \quad \mathcal{N}_{J,0} = \mathcal{N}_J \cap \Omega.$$

The scalar product in $L^2(\Omega)$ and its lumped version in \mathcal{S}_J are denoted by (\cdot, \cdot) and $\langle \cdot, \cdot \rangle$, respectively. The linear space of piecewise constant functions

$$\mathcal{P}_J = \{v \in L^2(\Omega) \mid v|_T \text{ is constant } \forall T \in \mathcal{T}_J\} \subset L^2(\Omega)$$

is spanned by the canonical basis $\mu_T, T \in \mathcal{T}_J$, as defined by $\mu_T(x) = 1$ for $x \in \text{int } T$ and $\mu_T(x) = 0$ otherwise.

6.1. An optimal control problem with control constraints. For given $y_0 \in L^4(\Omega)$ and $\varepsilon > 0$, we consider the following optimal control problem [45].

Find $y \in H_0^1(\Omega)$ and $u \in L^\infty(\Omega)$ such that

$$(6.1) \quad \mathcal{J}(y, u) = \int_{\Omega} \frac{1}{2} \|y - y_0\|_{L^2(\Omega)}^2 + \frac{\varepsilon}{2} \|u\|_{L^2(\Omega)}^2 \, dx$$

is minimal over all functions in $H_0^1(\Omega)$ and $L^\infty(\Omega)$ subject to the state equation

$$(6.2) \quad (\nabla y, \nabla v) = (u, v) \quad \forall v \in H_0^1(\Omega)$$

and the control constraint

$$(6.3) \quad u \in \mathcal{K} = \{v \in L^\infty(\Omega) \mid |v(x)| \leq 1 \text{ a.e. in } \Omega\}.$$

Approximating $H_0^1(\Omega)$ by $\mathcal{S}_{J,0}$ and \mathcal{K} by

$$\mathcal{K}_J = \{v \in \mathcal{P}_J \mid |v|_T| \leq 1 \quad \forall T \in \mathcal{T}_J\} \subset \mathcal{K},$$

we obtain a discrete analogue of the continuous problem. For existence and error estimates, we refer to [1]. We restrict our considerations to this discretization only. However, the algorithm behaves similar for other discretizations, e.g., with linear finite elements for the control. After incorporating (6.2) by a Lagrange multiplier w , the Kuhn–Tucker conditions of the discretized problem can be rewritten in the form (1.1) with $n = |\mathcal{N}_{J,0}| + |\mathcal{T}_J|$, $m = |\mathcal{T}_J|$, $F = A + \partial\mathcal{K}_J$,

$$A = \begin{pmatrix} D_S & 0 \\ 0 & \varepsilon D_P \end{pmatrix}, \quad D_S = (\langle \lambda_p, \lambda_q \rangle)_{p,q \in \mathcal{N}_{J,0}}, \quad D_P = ((\mu_T, \mu_{T'}))_{T,T' \in \mathcal{T}_J},$$

$$B = (A_S \quad -D_{SP}), \quad A_S = ((\nabla \lambda_p, \nabla \lambda_q))_{p,q \in \mathcal{N}_{J,0}}, \quad D_{SP} = ((\lambda_p, \mu_T))_{p \in \mathcal{N}_{J,0}, T \in \mathcal{T}_J},$$

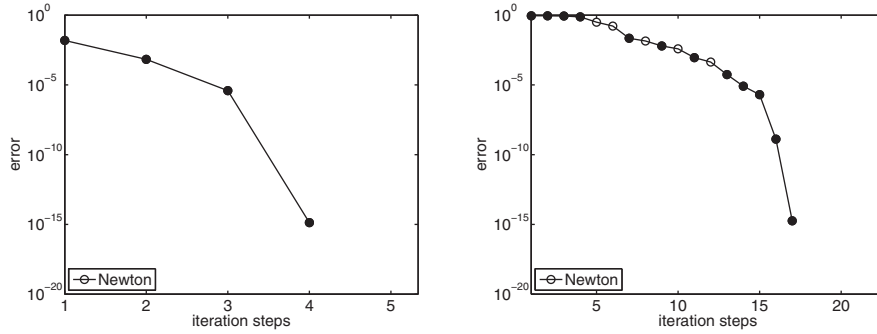


FIG. 6.1. Iteration history for $\varepsilon = 10^{-4}$ (left) and $\varepsilon = 10^{-8}$ (right). The filled dots indicate $\rho^\nu = 1$.

$C = 0$, and suitable right-hand sides f and g . It is easily checked that the assumptions (A1'), (A2), and (A3) are fulfilled. Moreover, it turns out that $S(c)$ is s.p.d. $\forall c \in \mathcal{C}$. As a consequence, h must be strongly convex (cf. Lemma 4.2) providing uniqueness (A4) and linear convergence of the Newton-type iteration to be called NEWTON as well as its inexact version (cf. Theorems 4.1 and 4.3). In general, we have $\text{rank } B = m < n$ so that it is not clear from our present analysis that $S_\nu = S(c^\nu) \in \partial_B(H(w^\nu))$ (cf. Proposition 4.2). As A is diagonal, the quadratic obstacle problems (5.2) arising in each iteration step can be easily solved by nodal projection. The linear saddle point problems (5.3) are evaluated by the direct solver UMFPACK [15].

Following [41, Chapter 5], we select the desired state

$$y_0(x) = 0.001 \begin{cases} 4 & \text{if } x \in [0, 0.75] \times [0, 0.5], \\ -10 & \text{if } x \in [0, 0.75] \times [0.5, 1], \\ -2 & \text{if } x \in [0.75, 1] \times [0, 0.5], \\ 50 & \text{if } x \in [0.75, 1] \times [0.5, 1] \end{cases}$$

in our numerical computations. The mesh size $h_J = 2^{-J}$ is resulting from $J = 7$ refinement steps. Finally, we choose the parameters

$$(6.4) \quad \alpha = 10^{-2}, \quad \alpha_\nu = \max \left\{ 1, -\alpha \frac{\langle \nabla h(w^\nu), d^\nu \rangle}{\|d^\nu\|_M^2} \right\}, \quad \beta = 0.5, \quad \delta = 0.5$$

in the associated Armijo strategy (cf. Proposition 3.1).

Figure 6.1 shows the algebraic error $\|w^* - w^\nu\|_M$ over the number of iteration steps for the two problem parameters $\varepsilon = 10^{-4}$ and $\varepsilon = 10^{-8}$, respectively. The algebraic error is measured in the energy norm induced by the Schur complement $M = BA^{-1}B^T$ providing

$$\|w^* - w^\nu\|_M = \|B^T(w^* - w^\nu)\|_{A^{-1}} \geq \|u^* - u^\nu\|_A$$

according to (5.5). The “exact” solution w^* is precomputed to round-off errors. In both cases, we observe superlinear convergence and finite termination, even exceeding the findings of Theorem 4.1. The condition number of (6.1) is increasing for decreasing regularization parameter ε . This is reflected by the large number of iteration steps for the small value $\varepsilon = 10^{-8}$. As the solution of the (diagonal!) obstacle problems (5.2) is almost for free and, in addition, no more than two tests are necessary in Armijo

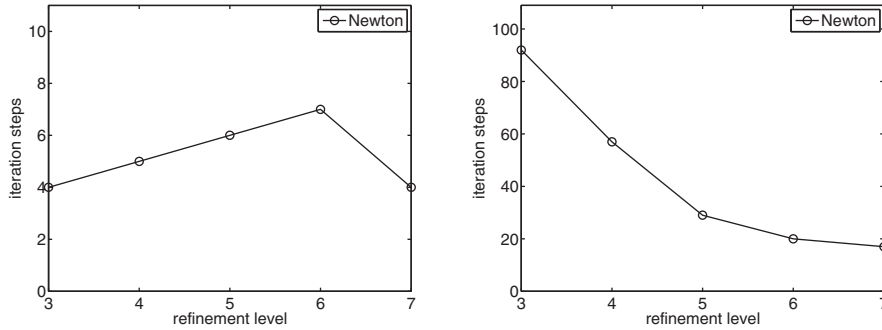


FIG. 6.2. Mesh dependence for $\epsilon = 10^{-4}$ (left) and $\epsilon = 10^{-8}$ (right).

damping, almost 100% of cpu time is consumed by the solution of the linear saddle point problems. For the given initial iterates the well-known (undamped) primal-dual algorithm converges only for $\epsilon = 10^{-4}$ but not for $\epsilon = 10^{-8}$ as indicated by Figure 6.1. On the other hand, in both cases the damping parameter $\rho^\nu = 1$ is accepted before the correct active set is detected in the last iteration step.

We now investigate the mesh dependence of NEWTON. The two pictures in Figure 6.2 show the number of iteration steps required for the solution to round-off errors over the refinement levels. For both values $\epsilon = 10^{-4}$ and $\epsilon = 10^{-8}$, the convergence speed seems to saturate with increasing refinement. It is interesting that coarser problems seem to become even harder for small ϵ . Note that the maximal number of Armijo tests is also increasing from two to ten on the coarsest mesh.

6.2. A Cahn–Hilliard problem. For given $\epsilon > 0$, final time $T > 0$, and initial condition $u_0 \in \mathcal{K} = \{v \in H^1(\Omega) \mid |v| \leq 1\}$, we consider the following initial value problem for the Cahn–Hilliard equation with an obstacle potential [7, 11, 18].

Find $u \in H^1(0, T; (H^1(\Omega))') \cap L^\infty(0, T; H^1(\Omega))$ and $w \in L^2(0, T; H^1(\Omega))$ with $u(0) = u_0$ such that $u(t) \in \mathcal{K}$ and

$$(6.5a) \quad \left\langle \frac{du}{dt}, v \right\rangle_{H^1(\Omega)} + (\nabla w, \nabla v) = 0 \quad \forall v \in H^1(\Omega),$$

$$(6.5b) \quad \epsilon(\nabla u, \nabla v - \nabla u) - (u, v - u) \geq (w, v - u) \quad \forall v \in \mathcal{K}$$

hold a.e. for $t \in (0, T)$.

Here $\langle \cdot, \cdot \rangle_{H^1(\Omega)}$ denotes the duality pairing of $H^1(\Omega)$ and $H^1(\Omega)'$. The unknown functions u and w are called order parameter and chemical potential, respectively. For existence and uniqueness results we refer to [7]. Semi-implicit Euler discretization in time and finite elements in space [6, 8] lead to the following discretized problem.

Find $u_J^k \in \mathcal{K}_J$ and $w_J^k \in \mathcal{S}_J$ such that

$$(6.6a) \quad \langle u_J^k, v \rangle + \tau(\nabla w_J^k, \nabla v) = \langle u_J^{k-1}, v \rangle \quad \forall v \in \mathcal{S}_J,$$

$$(6.6b) \quad \epsilon(\nabla u_J^k, \nabla(v - u_J^k)) - \langle w_J^k, v - u_J^k \rangle \geq \langle u_J^{k-1}, v - u_J^k \rangle \quad \forall v \in \mathcal{K}_J$$

hold for each $k = 1, \dots, N$.

We have chosen a uniform time step size $\tau = T/N$, and $\mathcal{K}_J = \mathcal{K} \cap \mathcal{S}_J$ is the nodal approximation of \mathcal{K} . The initial condition $u_J^0 \in \mathcal{K}_J$ is obtained by discrete L^2 projection $\langle u_J^0, v \rangle = (u_0, v) \forall v \in \mathcal{S}_J$. Existence, uniqueness, and error estimates have been established in [8]. More precisely, there exists a discrete solution (u_J^k, w_J^k) with

uniquely determined u_J^k , $k = 1, \dots, N$. Moreover, w_J^k is also unique, provided that the condition

$$(6.7) \quad \exists p \in \mathcal{N}_J : |u_J^k(p)| < 1$$

is fulfilled. Hence, (A4) is satisfied in this case. If (6.7) is violated, then either the triangulation \mathcal{T}_J is too coarse to resolve the diffuse interface or only one phase is present; i.e., u_J is constant. For the iterative solution of each spatial problem (6.6) a projected block Gauß–Seidel scheme [6] and an ADI-type iteration [33] are widely used. Both algorithms suffer from rapidly deteriorating convergence rates for increasing refinement.

Exploiting discrete mass conservation $\langle u_J^k, 1 \rangle = (u_0, 1)$, each spatial problem (6.6) takes the form (1.1) with $n = m = |\mathcal{N}_J|$, $F = A + \partial I_{\mathcal{K}_J}$,

$$A = \varepsilon (\langle \lambda_p, 1 \rangle \langle \lambda_q, 1 \rangle + (\nabla \lambda_p, \nabla \lambda_q))_{p,q \in \mathcal{N}_J},$$

$$B = -(\langle \lambda_p, \lambda_q \rangle)_{p,q \in \mathcal{N}_J}, \quad C = \tau ((\nabla \lambda_p, \nabla \lambda_q))_{p,q \in \mathcal{N}_J},$$

and suitable right-hand sides f and g . Assuming (6.7), it is easily checked that the assumptions (A1'), (A2), and (A3) are satisfied. Observe that A is the sum of a sparse stiffness matrix and a rank one matrix. We clearly have $\text{rank } B = n$ so that $S(c) \in \partial_B H(w)$ is a B-subdifferential of H (cf. Proposition 4.2). However, as C is only positive semidefinite, the kernel $\ker S(c)$ is trivial only if $N_c^\bullet \neq N$. In the singular case $N_c^\bullet = N$, $\ker S(c)$ is spanned by the constant vector $k_1 = (1, \dots, 1)^T$.

For our numerical computations, we select $\varepsilon = 10^{-4}$ and the time step $\tau = \varepsilon$, and the mesh size $h_J = 2^{-J}$ is resulting from $J = 9$ refinement steps. The initial condition u_0 takes the values $u_0(x) = \max\{\min\{2 \sin(4\pi x_1) \sin(4\pi x_2), 1\}, -1\}$.

We compare the nonsmooth Newton-like method (cf. Theorem 4.2) called NEWTON-LIKE, the inexact variant (cf. Theorem 4.3) called INEXACT, and the projected block Gauß–Seidel relaxation [6] called GAUSS–SEIDEL. The actual active coefficients are computed from the obstacle problem (5.2) by a monotone multigrid method [30]. The linear saddle point problems (5.4) are solved iteratively by a linear multigrid method with block Gauß–Seidel smoother and canonical restriction and prolongation. In the exact version NEWTON-LIKE the solution w^ν is computed to machine accuracy, and we use Armijo damping (cf. Proposition 3.1) with $\delta = 10^{-3}$ and the other parameters given in (6.4). In the ν th outer iteration of INEXACT we apply 3ν steps of the linear multigrid method with $V(3, 3)$ cycle to match the asymptotic accuracy condition (3.17), and we use heuristic damping (5.6) with $\sigma = 0.5$.

Figure 6.3 illustrates the algebraic error $\|w^* - w^\nu\|_M$ over the computational work for the first two spatial problems. We choose the discrete H^1 -norm induced by $M = D + C$ with $D = \tau (\langle \lambda_p, \lambda_q \rangle)_{p,q \in \mathcal{N}_J}$. Hence, $\|u^* - u^\nu\|_A \leq c \|w^* - w^\nu\|_M$ with a constant c independent of J (cf. (5.5) and Poincaré’s inequality). The “exact” solution w^* is precomputed to round-off errors. For a fair comparison, the computational work is now measured in work units (not in iteration steps). One work unit is the cpu time required by one linear multigrid $V(3, 3)$ cycle as applied to the linear saddle point problem (5.4). The left and the right picture in Figure 6.3 show the iteration histories for the spatial problems arising from the first and the second time step, respectively. Each marker refers to one iteration step of NEWTON-LIKE and INEXACT, respectively. As no initial data are available for the chemical potential w , we start with the bad initial iterate $w^0 = 0$ in the first problem, while the final approximation from the previous time step provides a reasonable initial iterate for the second

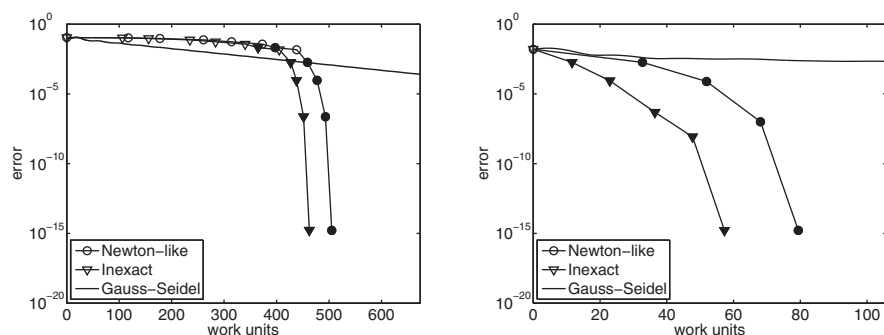


FIG. 6.3. Iteration histories for good initial iterates (left) and bad initial iterates (right). The filled dots indicate $\rho^\nu = 1$.

TABLE 6.1

Distribution of cpu time over the subtasks in each Uzawa step.

INEXACT	1	2	3	4	5	6	7	8	9	10	11
# tests	7	3	5	3	3	1	3	1	0	0	0
% Armijo	88.7	85.9	88.1	76.1	74.2	49.2	69.3	44.4	0.1	0.1	0.1
% obstacle	7.2	0.0	-0.0	-0.0	0.0	0.0	0.0	-0.0	0.0	27.2	24.0
% linear	4.1	14.0	11.9	23.8	25.7	50.7	30.7	55.5	99.7	72.6	75.7
work units	106.1	50.1	78.5	49.0	56.4	24.5	40.5	21.8	11.0	13.4	10.9

one. This makes quite a difference. For the bad initial iterate, it takes about 400 work units (about 6 iteration steps) until NEWTON-LIKE and INEXACT finally display superlinear convergence. GAUSS-SEIDEL is even more efficient in the beginning of the iteration, but not comparable later. For reasonable initial iterates, superlinear convergence starts immediately (observe the different scaling of the x -axis). In both cases, INEXACT turns out to be more efficient than NEWTON-LIKE.

Table 6.1 gives more detailed insight into the performance of the different building blocks of INEXACT as applied to the first problem. The number of tests involved in Armijo damping is given in the first line. Due to the bad initial iterate, a considerable number of tests are required in the beginning which later goes down to zero. The following three lines show the actual percentage of cpu time required by damping and the approximate solution of the obstacle problem and of the linear saddle point problem, respectively. These numbers do not sum to 100 because minor computations are neglected. Observe that the computational work is first dominated by Armijo damping and later by the increasing number of multigrid sweeps for the linear saddle point problem. Apart from the initial step, the detection of the active set takes not more than 5 monotone multigrid sweeps, each of which is cheaper than a multigrid sweep for the linear saddle point problem. As shown in the last line, the absolute amount of computational work strongly depends on the number of Armijo tests, which in turn strongly depends on the (problem dependent!) choice of the parameters. Hence, the performance of INEXACT could be probably improved by more careful tuning of the damping parameters. Observe that, for bad initial iterates, neither the exact nor the inexact method converges without damping. On the other hand, for both versions the damping parameter $\rho^\nu = 1$ is accepted before the correct active set is detected (cf. Figure 6.3). More efficient affine-invariant damping strategies for nonsmooth Newton-type algorithms will be the subject of future research.

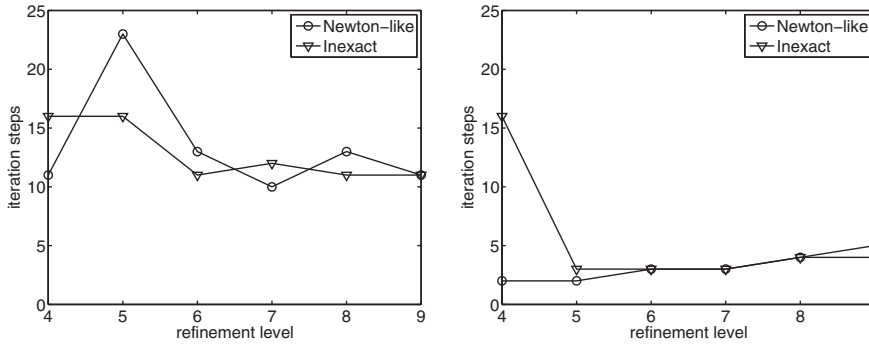


FIG. 6.4. Mesh dependence for good initial iterates (left) and bad initial iterates (right).

We now investigate the mesh dependence of NEWTON-LIKE and INEXACT. Figure 6.4 shows the number of iteration steps required for the solution to round-off errors over the refinement levels. For the first spatial problem (left), we always start with $w^\nu = 0$, while, for the second spatial problem (right), we always start from the previous time level. In both cases, the overall convergence speed seems to be scarcely affected by decreasing mesh size. It is astonishing that INEXACT sometimes even needs less iteration steps. Note that the averaged error reduction per work unit of INEXACT is about $\rho = 0.6$. We observed $\rho \approx 0.16$ for the linear multigrid solver as applied to the linear saddle point problems. Hence, for reasonable initial iterates, the solution of the discrete Cahn–Hilliard problem by straightforward inexact versions required about three to four times the cpu time for the solution of related linear saddle point problems by standard multigrid methods.

Acknowledgments. The authors would like to thank the unknown referees for their most valuable comments and suggestions.

REFERENCES

- [1] N. ARADA, E. CASAS, AND F. TRÖLTZSCH, *Error estimates for the numerical approximation of a semilinear elliptic control problem*, *Comput. Optim. Appl.*, 23 (2002), pp. 201–229.
- [2] L. ARMijo, *Minimization of functions having Lipschitz-continuous first partial derivatives*, *Pacific J. Math.*, 204 (1966), pp. 126–136.
- [3] L. BADEA, X.-C. TAI, AND J. WANG, *Convergence rate analysis of a multiplicative Schwarz method for variational inequalities*, *SIAM J. Numer. Anal.*, 41 (2003), pp. 1052–1073.
- [4] L. BADEA, *Convergence rate of a Schwarz multilevel method for the constrained minimization of nonquadratic functionals*, *SIAM J. Numer. Anal.*, 44 (2006), pp. 449–477.
- [5] J. W. BARRETT AND J. BLOWEY, *An error bound for the finite element approximation of the Cahn–Hilliard equation with logarithmic free energy*, *Numer. Math.*, 72 (1995), pp. 1–20.
- [6] J. W. BARRETT, R. NÜRNBERG, AND V. STYLES, *Finite element approximation of a phase field model for void electromigration*, *SIAM J. Numer. Anal.*, 42 (2004), pp. 738–772.
- [7] J. BLOWEY AND C. ELLIOTT, *The Cahn–Hilliard gradient theory for phase separation with non-smooth free energy, Part I: Mathematical analysis*, *European J. Appl. Math.*, 2 (1991), pp. 233–280.
- [8] J. BLOWEY AND C. ELLIOTT, *The Cahn–Hilliard gradient theory for phase separation with non-smooth free energy, Part II: Numerical analysis*, *European J. Appl. Math.*, 3 (1992), pp. 147–179.
- [9] D. BRAESS AND R. SARAZIN, *An efficient smoother for the Stokes problem*, *Appl. Numer. Math.*, 23 (1997), pp. 3–19.
- [10] M. BROKATE AND J. SPREKELS, *Hysteresis and Phase Transition*, *Appl. Math. Sci.* 121, Springer, Berlin, Heidelberg, New York, 1996.

- [11] J. CAHN AND J. HILLIARD, *Free energy of a nonuniform system I. Interfacial energy*, J. Chem. Phys., 28 (1958), pp. 258–267.
- [12] X. CHEN, *On preconditioned Uzawa methods and SOR methods for saddle-point problems*, J. Comput. Appl. Math., 100 (1998), pp. 207–224.
- [13] F. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.
- [14] R. COTTLE, J. PANG, AND R. STONE, *The Linear Complementarity Problem*, Academic Press, Boston, 1992.
- [15] T. A. DAVIS, *Algorithm 832: Umfpack v4.3 – an unsymmetric-pattern multifrontal method*, ACM Trans. Math. Software, 30 (2004), pp. 196–199.
- [16] P. DEUFLHARD, *Newton Methods for Nonlinear Problems*, Springer, Berlin, Heidelberg, 2004.
- [17] I. EKELAND AND R. TEMAM, *Convex Analysis*, North-Holland, Amsterdam, 1976.
- [18] C. ELLIOTT, *The Cahn-Hilliard model for the kinetics of phase separation*, in Mathematical Models for Phase Change Problems, J. Rodrigues, ed., Birkhäuser, Basel, Switzerland, 1989, pp. 35–73.
- [19] H. GARCKE AND B. STINNER, *Second order phase field asymptotics for multi-component systems*, Interfaces Free Bound., 8 (2006), pp. 131–157.
- [20] D. GILBARG AND N. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer, Berlin, 1988.
- [21] R. GLOWINSKI, *Numerical Methods for Nonlinear Variational Problems*, Springer, New York, 1984.
- [22] C. GRÄSER AND R. KORNUBER, *On preconditioned Uzawa-type iterations for a saddle point problem with inequality constraints*, in Domain Decomposition Methods in Science and Engineering XVI, Lect. Notes Comput. Sci. Eng., O. Widlund and D. Keyes, eds., Springer, Heidelberg, 2006, pp. 91–102.
- [23] C. GRÄSER AND R. KORNUBER, *Adaptive multigrid methods for the Cahn-Hilliard equation with logarithmic potential*, in preparation.
- [24] C. GRÄSER AND R. KORNUBER, *Multigrid methods for obstacle problems*, J. Comput. Math., to appear.
- [25] C. GRÄSER, *Globalization of nonsmooth Newton methods for optimal control problems*, in Numerical Mathematics and Advanced Applications, K. Kunisch, G. Of, and O. Steinbach, eds., Springer, Berlin, 2007, pp. 605–612.
- [26] M. HINTERMÜLLER, K. ITO, AND K. KUNISCH, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2002), pp. 865–888.
- [27] Q. HU AND J. ZOU, *Nonlinear inexact Uzawa algorithms for linear and nonlinear saddle-point problems*, SIAM J. Optim., 16 (2006), pp. 798–825.
- [28] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.
- [29] R. KORNUBER AND R. KRAUSE, *Robust multigrid methods for vector-valued Allen-Cahn equations with logarithmic free energy*, Comput. Vis. Sci., 9 (2006), pp. 103–116.
- [30] R. KORNUBER, *Monotone multigrid methods for elliptic variational inequalities I*, Numer. Math., 69 (1994), pp. 167–184.
- [31] R. KORNUBER, *On constrained Newton linearization and multigrid for variational inequalities*, Numer. Math., 91 (2002), pp. 699–721.
- [32] J. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer, Berlin, Heidelberg, New York, 1971.
- [33] P. LIONS AND B. MERCIER, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Anal., 16 (1979), pp. 964–979.
- [34] J. MANDEL, *A multilevel iterative method for symmetric, positive definite linear complementarity problems*, Appl. Math. Optim., 11 (1984), pp. 77–95.
- [35] A. NEKVINDA AND L. ZAJÍČEK, *A simple proof of the Rademacher theorem*, Časopis Pěst. Mat., 113 (1988), pp. 337–341.
- [36] J. NOCEDAL, *Theory of algorithms for unconstrained optimization*, Acta Numer., 1 (1992), pp. 199–242.
- [37] J. ORTEGA AND W. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
- [38] M. POWELL, *Direct search algorithms for optimization calculations*, Acta Numer., 7 (1998), pp. 287–336.
- [39] L. QI AND J. SUN, *A nonsmooth version of Newton's method*, Math. Program., 58 (1993), pp. 353–367.
- [40] L. QI, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–244.

- [41] A. SCHIELA AND M. WEISER, *Superlinear convergence of the control reduced interior point method for pde constrained optimization*, *Comput. Optim. Appl.*, 39 (2008), pp. 369–393.
- [42] J. SCHÖBERL AND W. ZULEHNER, *On Schwarz-type smoothers for saddle point problems*, *Numer. Math.*, 95 (2003), pp. 377–399.
- [43] J. SIMO AND T. HUGHES, *Computational Inelasticity*, Springer, Berlin, 1998.
- [44] X.-C. TAI, *Rate of convergence for some constraint decomposition methods for nonlinear variational inequalities*, *Numer. Math.*, 93 (2003), pp. 755–786.
- [45] F. TRÖLTZSCH, *Optimale Steuerung partieller Differentialgleichungen. Theorie, Verfahren und Anwendungen*, Vieweg, Wiesbaden, 2005.
- [46] M. ULBRICH, *Nonsmooth Newton-like Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, Habilitationsschrift, TU München, Munich, 2002.
- [47] S. VANKA, *Block-implicit multigrid solution of Navier-Stokes equations in primitive variables*, *J. Comput. Phys.*, 65 (1986), pp. 138–158.
- [48] A. VISINTIN, *Models of Phase Transitions*, Birkhäuser, Boston, 1996.
- [49] C. WIENERS, *Nonlinear solution methods for infinitesimal perfect plasticity*, *ZAMM Z. Angew. Math. Mech.*, 87 (2007), pp. 643–660.
- [50] S. J. WRIGHT, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, 1997.
- [51] Y. YE, *Interior Point Algorithms*, Wiley, Chichester, 1997.
- [52] W. ZULEHNER, *A class of smoothers for saddle point problems*, *Computing*, 65 (2000), pp. 227–246.
- [53] W. ZULEHNER, *Analysis of iterative methods for saddle point problems: A unified approach*, *Math. Comp.*, 71 (2002), pp. 479–505.

THE LOCAL L^2 PROJECTED C^0 FINITE ELEMENT METHOD FOR MAXWELL PROBLEM*

HUO-YUAN DUAN[†], FENG JIA[†], PING LIN[†], AND ROGER C. E. TAN[†]

Abstract. An element-local L^2 -projected C^0 finite element method is presented to approximate the nonsmooth solution being not in H^1 of the Maxwell problem on a nonconvex Lipschitz polyhedron with reentrant corners and edges. The key idea lies in that element-local L^2 projectors are applied to both curl and div operators. The C^0 linear finite element (enriched with certain higher degree bubble functions) is employed to approximate the nonsmooth solution. The coercivity in L^2 norm is established uniform in the mesh-size, and the condition number $\mathcal{O}(h^{-2})$ of the resulting linear system is proven. For the solution and its curl in H^r with $r < 1$ we obtain an error bound $\mathcal{O}(h^r)$ in an energy norm. Numerical experiments confirm the theoretical error bound.

Key words. Maxwell problem, nonsmooth solution, C^0 finite element method, L^2 projection

AMS subject classification. 65N30

DOI. 10.1137/070707749

1. Introduction. In this paper we shall study the C^0 finite element method for Maxwell equations with a nonsmooth solution (i.e., the solution not in H^1). Consider a simply connected nonconvex polyhedral domain $\Omega \subset \mathbb{R}^3$ with a connected Lipschitz continuous boundary Γ , and let \mathbf{u} denote an unknown field and \mathbf{f} a given function. The problem we shall consider is to find \mathbf{u} such that

$$(1.1) \quad \mathbf{curl\,curl\,u} = \mathbf{f} \quad \text{in } \Omega, \quad \mathbf{u} \times \mathbf{n} = \mathbf{0} \quad \text{on } \Gamma.$$

The $\mathbf{curl\,curl}$ operator in (1.1) represents the principal part of a large number of forms and models of Maxwell equations [15, 20], and problem (1.1) plays a central role in most mathematical issues associated with Maxwell equations, such as regularity-singularities (see [26, 23, 13, 27, 29]), solvability-uniqueness (see [12, 39, 24, 2, 34, 14, 35]), and numerical methods (see [13, 25, 41, 19, 7, 49, 43, 50, 51, 48, 5, 44, 37, 4, 18, 8, 42] and references therein). We are interested in using C^0 finite elements of piecewise polynomials for the numerical solution of (1.1) because of the availability of numerous software packages. Also, C^0 elements are highly preferred in practice for all unknown variables of those problems coupled with Maxwell equations, e.g., for Magnetohydrodynamics coupling with Navier–Stokes equations and Maxwell equations, since velocity and pressure in the Navier–Stokes equations part are approximated by C^0 elements; it is not desirable from the implementation point of view if using non C^0 elements to approximate the magnetic field in the Maxwell equations part. Although (1.1) looks quite simple, its discretization by the C^0 finite element method is not straightforward. This is associated with some main difficulties displayed in computational electromagnetics: (a) The infinite dimensional null-space (i.e., gradient field) of the curl operator badly pollutes the finite element solutions (cf. [41, 43]); (b) In the case where the solution is not in H^1 , the finite element solution would not converge

*Received by the editors November 9, 2007; accepted for publication (in revised form) October 29, 2008; published electronically February 25, 2009. This work was supported by NUS academic research grant R-146-000-064-112.

<http://www.siam.org/journals/sinum/47-2/70774.html>

[†]Department of Mathematics, National University of Singapore, 2 Science Drive 2, Singapore 117543, Singapore (scidhy@nus.edu.sg, fjia2005@gmail.com, matlinp@nus.edu.sg, scitance@nus.edu.sg).

to the true solution but to some other solution in H^1 ; (c) The indefiniteness of the resulting linear system would increase difficulty in implementation.

To avoid the problems of gradient field and indefiniteness, a plain regularization (PR) method is widely used in practice (see [39, 13, 26, 43]), with a divergence constraint imposed on \mathbf{u} for a given g :

$$(1.2) \quad \operatorname{div} \mathbf{u} = g \quad \text{in } \Omega.$$

Setting

$$(1.3) \quad U = \left\{ \mathbf{v} \in (L^2(\Omega))^3; \operatorname{curl} \mathbf{v} \in (L^2(\Omega))^3, \operatorname{div} \mathbf{v} \in L^2(\Omega), \mathbf{v} \times \mathbf{n}|_{\Gamma} = \mathbf{0} \right\}$$

and letting (\cdot, \cdot) denote the L^2 -inner product, the variational form of the PR method consists of finding $\mathbf{u} \in U$ such that

$$(1.4) \quad (\operatorname{curl} \mathbf{u}, \operatorname{curl} \mathbf{v}) + s(\operatorname{div} \mathbf{u}, \operatorname{div} \mathbf{v}) = (\mathbf{f}, \mathbf{v}) + s(g, \operatorname{div} \mathbf{v}) \quad \forall \mathbf{v} \in U,$$

where the real number $s > 0$ is referred to as *penalty or regularization parameter* and can be taken as any positive constant [26]. The PR formulation (1.4) is well suited for C^0 finite element discretizations depicted in [21], since (1.4) is a second-order elliptic problem with its bilinear form coercive on U (cf. [39, 13, 4, 24, 37, 34, 26]). Consequently, a globally C^0 finite element solution may be produced, and the resulting linear system can be solved by any of the numerous well-developed direct and iterative solvers (e.g., conjugate gradient method) [38, 47] for symmetric, positive definite linear systems.

Nevertheless, the C^0 finite element discretization of (1.4) does not give a correct approximation when the solution is not in H^1 . What is worse, even refining the meshes with more elements cannot improve this situation. Readers are referred to [27, 39, 13, 28, 41] for more details. The low regularity of the solution would occur near reentrant corners and edges of nonsmooth domains, even if the right-hand sides are smooth; see [26, 29]. Here we shall try to explain the incorrect convergence based on our intuitive observation. Such an observation, together with the well-known interpolation error estimate (1.6) below, essentially motivates the method developed in this paper. Take $s = 1$, and let \mathbf{u}_h denote the C^0 finite element solution of (1.4), with h being the mesh size of the finite element triangulation of Ω . As h tends to zero, the PR formulation (1.4) would force \mathbf{u}_h to converge to an element in H^1 , but not to the solution \mathbf{u} that does not belong to H^1 , due to the following fact (see [24, 27]) that

$$(1.5) \quad (\operatorname{curl} \mathbf{v}, \operatorname{curl} \mathbf{z}) + (\operatorname{div} \mathbf{v}, \operatorname{div} \mathbf{z}) = (\nabla \mathbf{v}, \nabla \mathbf{z}) \quad \text{for all } \mathbf{v}, \mathbf{z} \in U \cap (H^1(\Omega))^3.$$

On the other hand, any function u in L^2 (even in L^1) can be well approximated by C^0 finite elements:

$$(1.6) \quad \|u - \tilde{u}\|_0 \leq C h^r \|u\|_r \quad \text{if } u \in H^r, r \geq 0,$$

where \tilde{u} is a C^0 interpolation of u , and $\|\cdot\|_0, \|\cdot\|_r$ stand for L^2 - and H^r -norm, respectively, cf. [10, 11, 52, 22, 21, 53, 17]. So, when the solution is not in H^1 , there should be no problem in using C^0 elements to obtain a correct and good C^0 approximation, but we have to modify the PR formulation.

In this respect, there is an existing method: the weighted regularization (WR) method [25]. The WR method is theoretically and numerically proven to be good in

obtaining correct C^0 approximations. It adds a suitable weight function in front of the *div* operator in (1.4), i.e.,

$$(1.7) \quad (\mathbf{curl} \mathbf{u}, \mathbf{curl} \mathbf{v}) + s(\omega \operatorname{div} \mathbf{u}, \operatorname{div} \mathbf{v}) = (\mathbf{f}, \mathbf{v}) + s(\omega g, \operatorname{div} \mathbf{v}) \quad \forall \mathbf{v} \in U^\omega,$$

where $\omega(x)$ is a weight function and U^ω is a ω -weighted Hilbert space. The weight function ω is determined according to the geometric singularities of the domain boundary. To approximate the solution the WR method employs a C^0 finite element that is required to contain the gradient of a C^1 finite element space. Several C^1 elements [21] exist in two dimensions (2D), but, to our knowledge, in three-dimensional (3D) case, either few C^1 elements are known or C^1 elements involve too many degrees of freedoms and stringent conditions on the finite element triangulation of the domain [45, 1, 54, 33, 55]. Thus, either it is not easy to find a C^0 approximate space containing the gradient of a 3D C^1 element, or such a C^0 approximate space is of relatively little interest. It is also worth mentioning the singular function (SF) method [13, 39, 5, 6]. The SF method is successful for reduced 2D problems [40, 5]. Roughly speaking, the SF method uses the PR formulation (1.4) but augments the C^0 approximate space by the singular functions associated with reentrant corners and edges, which would span a space with an infinite dimension and should be precisely calculated in advance. Based on above reasons, it is rather inconvenient to apply these methods to 3D problems, especially when the geometric singularities of the domain boundary are not explicitly known. It is also worth mentioning the weighted least-squares method of a first-order system of (1.1) in [46] with additional independent variables, where linear elements are used with fewer degrees of freedom.

In this paper, we develop a new C^0 finite element method for solving problem (1.1)–(1.2), based on the spirit of the L^2 projection technique involved in the least-squares minimization of the L^2 projected residual of the Stokes first-order system [32]. In our case here, the PR formulation (1.4) is not a least-squares minimization of the residual of the curcurl-div second-order system (1.1)–(1.2), so we directly modify (1.4) by applying element-local L^2 projectors in front of both *curl* and *div* operators, with suitable mesh-dependent (element-local) bilinear and linear forms added. In the C^0 linear element (enriched by suitable face- and element-bubbles) an approximation behaving like (1.6) of the solution being not in H^1 to problem (1.1)–(1.2) is obtained.

Specifically, let \check{R}_h and R_h denote two local L^2 projectors, respectively, for *div* and *curl* operators, which are, respectively, defined element-by-element onto the discontinuous piecewise constant finite element space and the discontinuous piecewise linear finite element space, and let $\mathcal{S}_h(\cdot, \cdot)$ denote a mesh-dependent (element-local) bilinear form which is called the *stabilization term* and corresponds to a right-hand side mesh-dependent linear form $\mathcal{Z}_h(\cdot)$, and let $U_h \subset U \cap (H^1(\Omega))^3$ denote the approximate space. Then the L^2 projection method for solving problem (1.1)–(1.2) is to find $\mathbf{u}_h \in U_h$ such that

$$(1.8) \quad \begin{aligned} \mathcal{L}_h(\mathbf{u}_h, \mathbf{v}_h) &:= (R_h(\mathbf{curl} \mathbf{u}_h), R_h(\mathbf{curl} \mathbf{v}_h)) + s \left(\check{R}_h(\operatorname{div} \mathbf{u}_h), \check{R}_h(\operatorname{div} \mathbf{v}_h) \right) \\ &\quad + \alpha \mathcal{S}_h(\mathbf{u}_h, \mathbf{v}_h) \\ &= (\mathbf{f}, \mathbf{v}_h) + s \left(g, \check{R}_h(\operatorname{div} \mathbf{v}_h) \right) + \alpha \mathcal{Z}_h(\mathbf{v}_h) \quad \forall \mathbf{v}_h \in U_h, \end{aligned}$$

where the real number $\alpha > 0$ is referred to as a *stabilization parameter*. As the approximate space, U_h is chosen to be the C^0 linear element (enriched with certain higher degree face- and element-bubble functions; see (3.10)). We show that the

following coercivity holds:

$$(1.9) \quad \mathcal{L}_h(\mathbf{v}, \mathbf{v}) \geq C \|\mathbf{v}\|_0^2 \quad \forall \mathbf{v} \in U_h,$$

and obtain the condition number $\mathcal{O}(h^{-2})$ of the resulting linear system. With the help of the L^2 projectors and the face- and element-bubbles in U_h , we construct an appropriate C^0 interpolation $\tilde{\mathbf{u}} \in U_h$ such that the exact solution \mathbf{u} being not in H^1 and the finite element solution $\mathbf{u}_h \in U_h$ satisfies

$$(1.10) \quad \|\mathbf{u} - \mathbf{u}_h\|_0 \lesssim \|\mathbf{u} - \tilde{\mathbf{u}}\|_0.$$

Inequalities (1.10) and (1.6) indicate that even if \mathbf{u} is not in H^1 , a correct and good C^0 approximation of \mathbf{u} should be expected. In fact, when \mathbf{u} and $\mathbf{curl} \mathbf{u}$ are in H^r , with a smooth \mathbf{f} , we obtain the following desirable error estimate in an energy norm:

$$(1.11) \quad \|\mathbf{u} - \mathbf{u}_h\|_0 + \|R_h(\mathbf{curl}(\mathbf{u} - \mathbf{u}_h))\|_0 + \left\| \check{R}_h(\operatorname{div}(\mathbf{u} - \mathbf{u}_h)) \right\|_0 \leq C h^r.$$

Before closing this section, we make several remarks. Firstly, the implementation of the L^2 projection method is almost the same as that of the PR method (1.4), since in the former both additional L^2 projections and mesh-dependent terms are element-locally evaluated. Secondly, in comparison with the WR method (1.7), the L^2 projection method (1.8) does not involve the geometric singularities of the domain boundary, and the approximate space U_h is not required to contain the gradient of a C^1 element. As a matter of fact, U_h here does not contain the gradient of any known C^1 elements. Thirdly, if the approximate space is chosen to contain the gradient of some C^1 element, then we can drop the L^2 projector R_h before the curl operator and use the following bilinear form:

$$(1.12) \quad \mathcal{L}_h^*(\mathbf{u}, \mathbf{v}) := (\mathbf{curl} \mathbf{u}, \mathbf{curl} \mathbf{v}) + s \left(\check{R}_h(\operatorname{div} \mathbf{u}), \check{R}_h(\operatorname{div} \mathbf{v}) \right) + \alpha \mathcal{S}_h^*(\mathbf{u}, \mathbf{v}),$$

where \mathcal{S}_h^* is a part of the mesh-dependent bilinear form \mathcal{S}_h . We note that both (1.7) and (1.12) may employ the same approximate space containing the gradient of some C^1 element, but (1.12) involves only one element local L^2 projector \check{R}_h for the *div* operator and an element local stabilization term. No geometric singularities are explicitly involved in (1.12).

The outline of this paper is as follows. In section 2, we review the Maxwell equations. In section 3, we describe the local L^2 projected C^0 finite element method. Section 4 is devoted to the establishment of coercivity and the condition number. In section 5 we obtain error bounds in an energy norm. In section 6, numerical tests are performed to demonstrate the theoretical error bounds, and we make some conclusions in the last section.

2. Preliminaries. Let $\Omega \subset \mathbb{R}^3$ be a simply connected polyhedron with a connected Lipschitz continuous boundary Γ . Let \mathbf{n} denote the outward unit normal vector to Γ . In addition to the usual Hilbert spaces: $H^1(\Omega)$ with norm $\|\cdot\|_1$; $H_0^1(\Omega)$ and $H^1(\Omega)/\mathbb{R}$ with norm $|\cdot|_1$; $H^r(\Omega)$ with norm $\|\cdot\|_r$ for $r \in \mathbb{R}$, we introduce some

of the div and curl Hilbert spaces as follows:

$$\begin{aligned} H(\operatorname{div}; \Omega) &= \{ \mathbf{v} \in (L^2(\Omega))^3, \operatorname{div} \mathbf{v} \in L^2(\Omega) \}, \\ H_0(\operatorname{div}; \Omega) &= \{ \mathbf{v} \in H(\operatorname{div}; \Omega); \mathbf{v} \cdot \mathbf{n}|_{\Gamma} = 0 \}, \\ H(\operatorname{div}^0; \Omega) &= \{ \mathbf{v} \in H(\operatorname{div}; \Omega); \operatorname{div} \mathbf{v} = 0 \}, \\ H_0(\operatorname{div}^0; \Omega) &= H_0(\operatorname{div}; \Omega) \cap H(\operatorname{div}^0; \Omega), \\ H(\mathbf{curl}; \Omega) &= \left\{ \mathbf{v} \in (L^2(\Omega))^3, \mathbf{curl} \mathbf{v} \in (L^2(\Omega))^3 \right\}, \\ H_0(\mathbf{curl}; \Omega) &= \{ \mathbf{v} \in H(\mathbf{curl}; \Omega), \mathbf{v} \times \mathbf{n}|_{\Gamma} = \mathbf{0} \}, \\ H(\mathbf{curl}^0; \Omega) &= \{ \mathbf{v} \in H(\mathbf{curl}; \Omega); \mathbf{curl} \mathbf{v} = \mathbf{0} \}, \\ H_0(\mathbf{curl}^0; \Omega) &= H_0(\mathbf{curl}; \Omega) \cap H(\mathbf{curl}^0; \Omega), \end{aligned}$$

where these div and curl space are, respectively, equipped with norms: $\| \cdot \|_{0; \operatorname{div}}$ and $\| \cdot \|_{0; \mathbf{curl}}$:

$$\| \mathbf{v} \|_{0; \operatorname{div}}^2 = \| \mathbf{v} \|_0^2 + \| \operatorname{div} \mathbf{v} \|_0^2, \quad \| \mathbf{v} \|_{0; \mathbf{curl}}^2 = \| \mathbf{v} \|_0^2 + \| \mathbf{curl} \mathbf{v} \|_0^2,$$

where $\| \cdot \|_0$ stands for the L^2 -norm. We have for U defined as in (1.3)

$$U = H(\operatorname{div}; \Omega) \cap H_0(\mathbf{curl}; \Omega).$$

Assume that the right-hand sides

$$\mathbf{f} \in H(\operatorname{div}^0; \Omega) \quad \text{and} \quad g \in L^2(\Omega).$$

The 3D Maxwell problem we shall consider reads as follows:

Find $\mathbf{u} \in U$ such that

$$(2.1) \quad \mathbf{curl} \mathbf{curl} \mathbf{u} = \mathbf{f}, \quad \operatorname{div} \mathbf{u} = g \quad \text{in } \Omega,$$

$$(2.2) \quad \mathbf{u} \times \mathbf{n}|_{\Gamma} = \mathbf{0}.$$

Remark 2.1. Setting

$$(2.3) \quad \mathbf{z} := \mathbf{curl} \mathbf{u},$$

we see that \mathbf{z} satisfies

$$(2.4) \quad \mathbf{curl} \mathbf{curl} \mathbf{z} = \mathbf{curl} \mathbf{f}, \quad \operatorname{div} \mathbf{z} = 0 \quad \text{in } \Omega,$$

$$(2.5) \quad \mathbf{z} \cdot \mathbf{n}|_{\Gamma} = 0, \quad \mathbf{curl} \mathbf{z} \times \mathbf{n}|_{\Gamma} = \mathbf{f} \times \mathbf{n}|_{\Gamma},$$

if additionally $\mathbf{f} \in H(\mathbf{curl}; \Omega)$.

Remark 2.2. The time-harmonic Maxwell equations in 3D,

$$\begin{aligned} \mathbf{curl} \mathbf{E} - i\omega\mu\mathbf{H} &= \mathbf{0} \quad \text{and} \quad \mathbf{curl} \mathbf{H} + (i\varepsilon\omega - \sigma)\mathbf{E} = \mathbf{J} \quad \text{in } \Omega, \\ \mathbf{E} \times \mathbf{n}|_{\Gamma} &= \mathbf{0} \quad \text{and} \quad (\mu\mathbf{H}) \cdot \mathbf{n}|_{\Gamma} = 0, \end{aligned}$$

are often considered in practice, where \mathbf{E} is the electric field; \mathbf{H} is the magnetic field; $\omega > 0$ is the frequency of the vibrations; ε, μ, σ are, respectively, the permittivity, the permeability, and the conductivity of the materials occupying Ω ; and $\mathbf{J} \in H(\operatorname{div}; \Omega)$ is

the current density. Set $\mathbf{f}_\# := i\omega\mathbf{J}$, $\kappa_\#^2 := \omega^2(\varepsilon + i\sigma/\omega)$, and $g := \mathbf{J}/(i\omega)$. Eliminating \mathbf{H} we see that \mathbf{E} satisfies

$$\mathbf{curl} (\mu^{-1} \mathbf{curl} \mathbf{E}) - \kappa_\#^2 \mathbf{E} = \mathbf{f}_\#, \quad \operatorname{div} (\varepsilon + i\sigma/\omega) \mathbf{E} = g \quad \text{in } \Omega.$$

Similarly, setting $\mathbf{f}^\# := \mathbf{curl} (\varepsilon + i\sigma/\omega)^{-1} \mathbf{J}$ and $\kappa^\#{}^2 := \omega^2\mu$, and eliminating \mathbf{E} we see that \mathbf{H} satisfies

$$\mathbf{curl} ((\varepsilon + i\sigma/\omega)^{-1} \mathbf{curl} \mathbf{H}) - \kappa^\#{}^2 \mathbf{H} = \mathbf{f}^\#, \quad \operatorname{div} (\mu \mathbf{H}) = 0 \quad \text{in } \Omega.$$

In the case of $\mu = \varepsilon = 1$ and $\sigma = 0$, we have the following models of Maxwell equations:

$$(2.6) \quad \mathbf{curl} \mathbf{curl} \mathbf{u} - \omega^2 \mathbf{u} = \mathbf{f}, \quad \operatorname{div} \mathbf{u} = g \quad \text{in } \Omega,$$

$$(2.7) \quad \mathbf{u} \times \mathbf{n}|_\Gamma = \mathbf{0},$$

where \mathbf{u} stands for the electric field, with $\mathbf{f} = i\omega\mathbf{J}$; or

$$(2.8) \quad \mathbf{curl} \mathbf{curl} \mathbf{u} - \omega^2 \mathbf{u} = \mathbf{curl} \mathbf{f}, \quad \operatorname{div} \mathbf{u} = 0 \quad \text{in } \Omega,$$

$$(2.9) \quad \mathbf{u} \cdot \mathbf{n}|_\Gamma = 0, \quad \mathbf{curl} \mathbf{u} \times \mathbf{n}|_\Gamma = \mathbf{f} \times \mathbf{n}|_\Gamma,$$

where \mathbf{u} stands for the magnetic field, with $\mathbf{f} = \mathbf{J}$.

Since the corner and edge singularities of problem (2.1)–(2.2) (resp., (2.4)–(2.5)) have the same principal parts as those of problem (2.6)–(2.7) (resp., (2.8)–(2.9)), and since the main difficulty in the C^0 finite element discretization of (2.6)–(2.7) (resp., (2.8)–(2.9)) is due to the low regularity of the solution (not due to the presence of ω^2), it suffices for us to develop C^0 finite element methods for problem (2.1)–(2.2) (resp., (2.4)–(2.5)), which is in [26] called a *Maxwell problem*. In other words, the finite element method for problem (2.1)–(2.2) can be applied to problem (2.6)–(2.7) straightforwardly, as well as to the Maxwell eigenproblem (see Remark 2.3 below).

Remark 2.3. The 3D Maxwell eigenproblem relating to the source problem (2.6)–(2.7) is to find \mathbf{u} and ω^2 such that

$$(2.10) \quad \mathbf{curl} \mathbf{curl} \mathbf{u} = \omega^2 \mathbf{u}, \quad \operatorname{div} \mathbf{u} = 0 \quad \text{in } \Omega, \quad \mathbf{u} \times \mathbf{n}|_\Gamma = \mathbf{0}.$$

The PR variational formulation of (2.10) is to find $\mathbf{u} \in U$ and ω^2 such that (cf. [27])

$$(2.11) \quad (\mathbf{curl} \mathbf{u}, \mathbf{curl} \mathbf{v}) + s(\operatorname{div} \mathbf{u}, \operatorname{div} \mathbf{v}) = \omega^2 (\mathbf{u}, \mathbf{v}) \quad \forall \mathbf{v} \in U.$$

Note that, if the eigenfunction is not in H^1 , then (2.11) suffers the same difficulty as the source problem when discretized by the C^0 finite element method.

Now let us recall Green’s formula of integration by parts on Lipschitz domain D :

$$\begin{aligned} (\operatorname{div} \mathbf{v}, \phi)_{0,D} + (\mathbf{v}, \nabla \phi)_{0,D} &= \int_{\partial D} \mathbf{v} \cdot \mathbf{n} \phi \quad \forall \mathbf{v} \in H(\operatorname{div}; D), \forall \phi \in H^1(D), \\ (\mathbf{curl} \mathbf{v}, \phi)_{0,D} - (\mathbf{v}, \mathbf{curl} \phi)_{0,D} &= \int_{\partial D} \mathbf{v} \times \mathbf{n} \cdot \phi \quad \forall \mathbf{v} \in H(\mathbf{curl}; D), \forall \phi \in (H^1(D))^3, \end{aligned}$$

where $\mathbf{v} \cdot \mathbf{w} = \sum_{i=1}^3 v_i w_i$. Note that the last formula holds also for $\phi \in H(\mathbf{curl}; D)$ (in a suitable weak sense) on Lipschitz polyhedra, cf. [3], with $\int_{\partial D} \mathbf{v} \times \mathbf{n} \cdot \phi$ being

written as $\int_{\partial D} \mathbf{v} \times \mathbf{n} \cdot (\mathbf{n} \times \phi \times \mathbf{n})$. Here and in the sequel, $(\cdot, \cdot)_{0,D}$ denotes the L^2 inner product on D , and (\cdot, \cdot) stands solely for the L^2 -inner product on Ω .

Before closing this section, we define a notation in 3D. For any vector-valued function $\mathbf{v} = (v_1, v_2, v_3)$ and a scalar function q , we define a notation $(\mathbf{v}, q)_{0,D} \in \mathbb{R}^3$ by

$$(2.12) \quad (\mathbf{v}, q)_{0,D} := ((v_1, q)_{0,D}, (v_2, q)_{0,D}, (v_3, q)_{0,D}) \in \mathbb{R}^3.$$

For any $\mathbf{v} \in H(\mathbf{curl}; D)$ and $\phi \in H_0^1(D)$, we have from the above Green's formula

$$(2.13) \quad (\mathbf{curl} \mathbf{v}, \phi)_{0,D} = (((v_2, v_3), \mathbf{curl}_{23} \phi)_{0,D}, ((v_3, v_1), \mathbf{curl}_{31} \phi)_{0,D}, ((v_1, v_2), \mathbf{curl}_{12} \phi)_{0,D}) \in \mathbb{R}^3,$$

where $\mathbf{curl}_{ij} \phi = (\partial_j \phi, -\partial_i \phi)$ is the curl of the scalar function ϕ with respect to the coordinate components (x_i, x_j) , and we also have for $\mathbf{u}, \mathbf{v} \in H(\mathbf{curl}; D)$ and $\phi \in H_0^1(D)$

$$(2.14) \quad (\mathbf{curl} \mathbf{u}, \phi)_{0,D} \cdot (\mathbf{curl} \mathbf{v}, \phi)_{0,D} = ((u_2, u_3), \mathbf{curl}_{23} \phi)_{0,D} ((v_2, v_3), \mathbf{curl}_{23} \phi)_{0,D} + ((u_3, u_1), \mathbf{curl}_{31} \phi)_{0,D} ((v_3, v_1), \mathbf{curl}_{31} \phi)_{0,D} + ((u_1, u_2), \mathbf{curl}_{12} \phi)_{0,D} ((v_1, v_2), \mathbf{curl}_{12} \phi)_{0,D}.$$

3. The L^2 projected C^0 finite element method. Let \mathcal{C}_h denote the shape-regular triangulation (see [21, 16, 37]) of $\bar{\Omega}$ into tetrahedra, with diameters h_K for $K \in \mathcal{C}_h$ bounded by h . Let \mathcal{P}_k be the space of polynomials of degree not greater than $k \geq 0$, with k being a nonnegative integer. Set

$$(3.1) \quad P_h := \{q \in L^2(\Omega); q|_K \in \mathcal{P}_1(K), \forall K \in \mathcal{C}_h\},$$

$$(3.2) \quad Q_h := \{q \in L^2(\Omega); q|_K \in \mathcal{P}_0(K), \forall K \in \mathcal{C}_h\}.$$

Let $K \in \mathcal{C}_h$ be a tetrahedron with vertices $a_i, 1 \leq i \leq 4$, and let F_i be the face opposite a_i . Denote by λ_i the barycentric coordinate of a_i . In fact, $\mathcal{P}_1(K) = \text{span}\{\lambda_i, 1 \leq i \leq 4\}$ and λ_i is also called the *shape function* of $\mathcal{P}_1(K)$; cf. [21]. Introduce the element-bubble

$$(3.3) \quad b_K := \lambda_1 \lambda_2 \lambda_3 \lambda_4 \in H_0^1(K),$$

and the face bubbles

$$(3.4) \quad b_{F_1} = \lambda_2 \lambda_3 \lambda_4, \quad b_{F_2} = \lambda_1 \lambda_3 \lambda_4, \quad b_{F_3} = \lambda_1 \lambda_2 \lambda_4, \quad b_{F_4} = \lambda_1 \lambda_2 \lambda_3.$$

We see that these face bubbles satisfy

$$(3.5) \quad b_{F_i}|_{F_i} \in H_0^1(F_i), \quad b_{F_i}|_{F_j} = 0 \quad \text{for all } j \neq i.$$

Let $\phi_{F_i,j} = p_{F_i,j} b_{F_i} \in H^1(K), 1 \leq j \leq 3$, be the shape (basis) functions of $\mathcal{P}_4(K)$ on $F_i, 1 \leq i \leq 4$, where

$$(3.6) \quad \mathcal{P}_1(F_i) = \text{span}\{p_{F_i,j}|_{F_i}, 1 \leq j \leq 3\}.$$

Let

$$(3.7) \quad \mathbf{P}_{F_i} := \text{span}\{q_{F_i,l}, 1 \leq l \leq 9\} = (\text{span}\{p_{F_i,j}, 1 \leq j \leq 3\})^3.$$

Clearly, we have $\mathbf{P}_{F_i}|_{F_i} = (\mathcal{P}_1(F_i))^3$. Introduce

$$\begin{aligned}
 (3.8) \quad \Phi_h &:= \left\{ \mathbf{v} \in (H^1(\Omega))^3; \mathbf{v}|_K \in (\text{span}\{\phi_{F_i,j}, 1 \leq j \leq 3, 1 \leq i \leq 4\})^3, \forall K \in \mathcal{C}_h \right\} \\
 &= \left\{ \mathbf{v} \in (H^1(\Omega))^3; \mathbf{v}|_K \in \text{span}\{\mathbf{q}_{F_i,l} b_{F_i}, 1 \leq l \leq 9, 1 \leq i \leq 4\}, \forall K \in \mathcal{C}_h \right\}, \\
 (3.9) \quad \mathfrak{B}_h &:= \left\{ \mathbf{v} \in (H_0^1(\Omega))^3; \mathbf{v}|_K \in (\text{span}\{b_K\})^3, \forall K \in \mathcal{C}_h \right\} \\
 &= \left\{ \mathbf{v} \in (H_0^1(\Omega))^3; \mathbf{v}|_K \in (\mathcal{P}_0(K))^3 b_K, \forall K \in \mathcal{C}_h \right\}.
 \end{aligned}$$

Define the C^0 approximate space $U_h \subset (H^1(\Omega))^3 \cap H_0(\mathbf{curl}; \Omega) \subset U$ as follows:

$$(3.10) \quad U_h = (P_h \cap H^1(\Omega))^3 \cap H_0(\mathbf{curl}; \Omega) + \Phi_h \cap H_0(\mathbf{curl}; \Omega) + \mathfrak{B}_h.$$

Let $\theta_{K,l}$, $1 \leq l \leq m = 20$, denote the shape function of $\mathcal{P}_3(K)$. Introduce a local set of functions

$$(3.11) \quad \Upsilon_K = \{\theta_{K,l}, 1 \leq l \leq m = 20\},$$

and define mesh-dependent (elementwisely) bilinear and linear forms as follows:

$$(3.12) \quad S_{h,\text{div}}(\mathbf{u}, \mathbf{v}) := \sum_{K \in \mathcal{C}_h} \frac{\sum_{l=1}^m (\mathbf{u}, \nabla(\theta_{K,l} b_K))_{0,K} (\mathbf{v}, \nabla(\theta_{K,l} b_K))_{0,K}}{\sum_{l=1}^m \|\nabla(\theta_{K,l} b_K)\|_{0,K}^2},$$

$$(3.13) \quad Z_{h,\text{div}}(g; \mathbf{v}) := - \sum_{K \in \mathcal{C}_h} \frac{\sum_{l=1}^m (g, \theta_{K,l} b_K)_{0,K} (\mathbf{v}, \nabla(\theta_{K,l} b_K))_{0,K}}{\sum_{l=1}^m \|\nabla(\theta_{K,l} b_K)\|_{0,K}^2},$$

$$(3.14) \quad S_{h,\text{curl}}(\mathbf{u}, \mathbf{v}) := \sum_{K \in \mathcal{C}_h} \frac{\sum_{l=1}^m (\mathbf{curl} \mathbf{u}, \theta_{K,l} b_K)_{0,K} \cdot (\mathbf{curl} \mathbf{v}, \theta_{K,l} b_K)_{0,K}}{\sum_{l=1}^m \|\nabla(\theta_{K,l} b_K)\|_{0,K}^2},$$

where the notation in (2.12) was used in (3.14). We finally define $\check{R}_h(\text{div} \mathbf{v}) \in Q_h$ for a given $\mathbf{v} \in H(\text{div}; \Omega) \cap H(\mathbf{curl}; \Omega)$ by

$$(3.15) \quad \check{R}_h(\text{div} \mathbf{v})|_K := \frac{1}{|K|} \int_K \text{div} \mathbf{v} \quad \forall K \in \mathcal{C}_h,$$

where $|K|$ denotes the volume of K , and define $R_h(\mathbf{curl} \mathbf{v}) \in (P_h)^3$ by

$$(3.16) \quad (R_h(\mathbf{curl} \mathbf{v}), \mathbf{q})_{0,K} := (\mathbf{curl} \mathbf{v}, \mathbf{q})_{0,K} \quad \forall \mathbf{q} \in (\mathcal{P}_1(K))^3, \forall K \in \mathcal{C}_h.$$

Setting

$$(3.17) \quad \mathcal{S}_h(\mathbf{u}, \mathbf{v}) := S_{h,\text{div}}(\mathbf{u}, \mathbf{v}) + S_{h,\text{curl}}(\mathbf{u}, \mathbf{v}), \quad \mathcal{Z}_h(\mathbf{v}) := Z_{h,\text{div}}(g; \mathbf{v}),$$

and letting s, α be two positive constants, we define the bilinear form on $U_h \times U_h$ as follows:

$$(3.18) \quad \mathcal{L}_h(\mathbf{u}, \mathbf{v}) := (R_h(\mathbf{curl} \mathbf{u}), R_h(\mathbf{curl} \mathbf{v})) + s \left(\check{R}_h(\text{div} \mathbf{u}), \check{R}_h(\text{div} \mathbf{v}) \right) + \alpha \mathcal{S}_h(\mathbf{u}, \mathbf{v}),$$

and define the linear form on U_h as follows:

$$(3.19) \quad \mathcal{F}_h(\mathbf{v}) := (\mathbf{f}, \mathbf{v}) + s(g, \check{R}_h(\operatorname{div} \mathbf{v})) + \alpha \mathcal{Z}_h(\mathbf{v}).$$

The L^2 projected C^0 finite element method to numerically solve problem (2.1)–(2.2) reads as follows:

$$(3.20) \quad \begin{cases} \text{Find } \mathbf{u}_h \in U_h \text{ such that} \\ \mathcal{L}_h(\mathbf{u}_h, \mathbf{v}) = \mathcal{F}_h(\mathbf{v}) \quad \forall \mathbf{v} \in U_h. \end{cases}$$

Remark 3.1. The method (3.20) is not consistent in the usual sense [21], i.e., with \mathbf{u} the exact solution and \mathbf{u}_h the finite element solution (See Lemma 5.1 for more details),

$$(3.21) \quad \mathcal{L}_h(\mathbf{u} - \mathbf{u}_h, \mathbf{v}_h) \neq 0 \quad \forall \mathbf{v}_h \in U_h,$$

because the term $S_{h,\operatorname{curl}}(\mathbf{u}, \mathbf{v}_h)$ does not correspond to any right-hand side term and

$$(3.22) \quad (R_h(\operatorname{curl} \mathbf{u}), R_h(\operatorname{curl} \mathbf{v}_h)) = (\operatorname{curl} \mathbf{u}, R_h(\operatorname{curl} \mathbf{v}_h)) \neq (\mathbf{f}, \mathbf{v}_h) \quad \forall \mathbf{v}_h \in U_h,$$

where \mathbf{u} satisfies

$$(3.23) \quad (\operatorname{curl} \mathbf{u}, \operatorname{curl} \mathbf{v}) = (\mathbf{f}, \mathbf{v}) \quad \forall \mathbf{v} \in U.$$

As we shall see, the estimate of the inconsistency error in (3.22) will be involved with the profound result on the regular-singular decomposition stated in Proposition 5.2 in the regularity theory for the Maxwell equations.

Remark 3.2. The role of the face- and element-bubbles in U_h is to eliminate the effects of both *curl* and *div* partial derivatives on the solution \mathbf{u} with the help of the local L^2 projectors R_h and \check{R}_h (see (5.14) in Lemma 5.5). The local set Υ_K , defined as (3.11) and used in (3.12)–(3.14), ensures that the following element-local inclusion properties hold:

$$(3.24) \quad \operatorname{div}(\mathbf{v}|_K) \in \mathcal{P}_3(K), \quad \operatorname{curl}(\mathbf{v}|_K) \in (\mathcal{P}_3(K))^3 \quad \text{on } K \quad \forall \mathbf{v} \in U_h, \forall K \in \mathcal{C}_h,$$

where $\mathbf{v}|_K$ is the restriction of \mathbf{v} to $K \in \mathcal{C}_h$. From (3.24) we have certain coercivity properties for both $S_{h,\operatorname{div}}(\mathbf{u}, \mathbf{v})$ and $S_{h,\operatorname{curl}}(\mathbf{u}, \mathbf{v})$ (see Lemma 4.3). The stabilization term \mathcal{S}_h in (3.17) is to ‘remedy’ the loss in the coercivity, where the loss is caused by the introduction of the L^2 projectors in front of both curl and div operators (cf. the coercive PR form (1.4) without L^2 projectors); see (4.27) in proving the coercivity property stated in Theorem 4.1.

Remark 3.3. In 2D, we just take the approximate space as the \mathcal{P}_3 element:

$$(3.25) \quad U_h := \left\{ \mathbf{v} \in (H^1(\Omega))^2 \cap H_0(\operatorname{curl}; \Omega); \mathbf{v}|_K \in (\mathcal{P}_3(K))^2, \forall K \in \mathcal{C}_h \right\},$$

where $H_0(\operatorname{curl}; \Omega) = \{ \mathbf{v} \in (L^2(\Omega))^2; \operatorname{curl} \mathbf{v} \in L^2(\Omega), \mathbf{v} \cdot \boldsymbol{\tau}|_{\partial\Omega} = 0 \}$, with $\operatorname{curl} \mathbf{v} = \partial_1 v_2 - \partial_2 v_1$ and $\boldsymbol{\tau}$ being the tangential unit vector to $\partial\Omega$, and the local set of functions

$$(3.26) \quad \Upsilon_K := \{ \theta_{K,l}, 1 \leq l \leq m = 6 \},$$

where $\theta_{K,l}$, $1 \leq l \leq m = 6$, is chosen as the shape function of $\mathcal{P}_2(K)$, and other definitions can be easily adjusted.

Remark 3.4. If the approximate space could contain the gradient of some C^1 element (i.e., the continuity of the functions is also imposed on the first-order partial derivatives across adjacent finite elements), we can drop both the L^2 projector R_h of the *curl* operator and the mesh-dependent bilinear form $S_{h,\text{curl}}(\cdot, \cdot)$. Below for the 2D problem we propose two finite element methods for which the approximate space, respectively, contains the gradient of the *Argyris* C^1 triangle element and the *Hsieh–Clough–Tocher (HCT)* C^1 macro triangle element (see [21]). As for 3D, the approximate space containing the gradient of a C^1 element is of relatively little interest as pointed out in section 1.

The *Argyris* C^1 element consists of polynomials of degree not greater than 5. The *HCT* C^1 macro-element consists of piecewise \mathcal{P}_3 polynomials, i.e, let $T_i, 1 \leq i \leq 3$, denote the subtriangles which are obtained by connecting the barycentric point of the triangle $K \in \mathcal{C}_h$ to the three vertices of K , then the *HCT* functions are \mathcal{P}_3 on each T_i . Set

$$(3.27) \quad \mathcal{T}_{h/2} := \cup_{K \in \mathcal{C}_h} \cup_{i=1}^3 T_i.$$

Define two approximate spaces as follows:

$$(3.28) \quad U_h^* := \left\{ \mathbf{v} \in (H^1(\Omega))^2 \cap H_0(\text{curl}; \Omega); \mathbf{v}|_K \in (\mathcal{P}_4(K))^2, \forall K \in \mathcal{C}_h \right\},$$

$$(3.29) \quad U_h^{**} := \left\{ \mathbf{v} \in (H^1(\Omega))^2 \cap H_0(\text{curl}; \Omega); \mathbf{v}|_T \in (\mathcal{P}_2(T))^2, \forall T \in \mathcal{T}_{h/2} \right\},$$

where U_h^* contains the gradient of the *Argyris* C^1 element, and U_h^{**} contains the gradient of the *HCT* C^1 macro element. Corresponding to U_h^* , we introduce the local set of functions

$$(3.30) \quad \Upsilon_K^* := \{\theta_{K,l}, 1 \leq l \leq m = 10\},$$

where $\theta_{K,l}, 1 \leq l \leq m = 10$, is chosen as the shape function of $\mathcal{P}_3(K)$, and we define

$$(3.31) \quad \mathcal{L}_h^*(\mathbf{u}, \mathbf{v}) := (\mathbf{curl} \mathbf{u}, \mathbf{curl} \mathbf{v}) + s \left(\check{R}_h(\text{div} \mathbf{u}), \check{R}_h(\text{div} \mathbf{v}) \right) + \alpha S_{h,\text{div}}(\mathbf{u}, \mathbf{v}),$$

$$(3.32) \quad \mathcal{F}_h^*(\mathbf{v}) := (\mathbf{f}, \mathbf{v}) + s \left(g, \check{R}_h(\text{div} \mathbf{v}) \right) + \alpha Z_{h,\text{div}}(g; \mathbf{v}),$$

where $S_{h,\text{div}}(\mathbf{u}, \mathbf{v})$ and $Z_{h,\text{div}}(g; \mathbf{v})$ are, respectively, defined by (3.12) and (3.13) but those functions $\theta_{K,l}, 1 \leq l \leq m$, are in Υ_K^* given by (3.30), and \check{R}_h is defined by (3.15). The finite element method is, thus, stated as follows:

$$(3.33) \quad \begin{cases} \text{Find } \mathbf{u}_h^* \in U_h^* \text{ such that} \\ \mathcal{L}_h^*(\mathbf{u}_h^*, \mathbf{v}) = \mathcal{F}_h^*(\mathbf{v}) \quad \forall \mathbf{v} \in U_h^*. \end{cases}$$

While corresponding to U_h^{**} we introduce the local set of functions for $T \in \mathcal{T}_{h/2}$

$$(3.34) \quad \Upsilon_T^{**} := \{\theta_{T,l}; 1 \leq l \leq m = 3\},$$

where $\theta_{T,l}, 1 \leq l \leq m = 3$, is chosen as the shape function of $\mathcal{P}_1(T)$, and we define

$$(3.35) \quad \mathcal{L}_h^{**}(\mathbf{u}, \mathbf{v}) := (\mathbf{curl} \mathbf{u}, \mathbf{curl} \mathbf{v}) + s \left(\check{R}_h(\text{div} \mathbf{u}), \check{R}_h(\text{div} \mathbf{v}) \right) + \alpha S_{h/2,\text{div}}(\mathbf{u}, \mathbf{v}),$$

$$(3.36) \quad \mathcal{F}_h^{**}(\mathbf{v}) := (\mathbf{f}, \mathbf{v}) + s \left(g, \check{R}_h(\text{div} \mathbf{v}) \right) + \alpha Z_{h/2,\text{div}}(g; \mathbf{v}),$$

where, with respect to the subtriangulation $\mathcal{T}_{h/2}$ given by (3.27), $S_{h/2,\text{div}}(\mathbf{u}, \mathbf{v})$ and $Z_{h/2,\text{div}}(g; \mathbf{v})$ are defined similarly to those in (3.12) and (3.13) with the choice Υ_T^{**} given by (3.34), and \check{R}_h is still defined in (3.15) with respect to the triangulation \mathcal{C}_h . The finite element method reads as follows:

$$(3.37) \quad \begin{cases} \text{Find } \mathbf{u}_h^{**} \in U_h^{**} \text{ such that} \\ \mathcal{L}_h^{**}(\mathbf{u}_h^{**}, \mathbf{v}) = \mathcal{F}_h^{**}(\mathbf{v}) \quad \forall \mathbf{v} \in U_h^{**}. \end{cases}$$

It can be easily seen that both the methods (3.33) and (3.37) are consistent in the usual sense, i.e.,

$$(3.38) \quad \mathcal{L}_h^*(\mathbf{u} - \mathbf{u}_h^*, \mathbf{v}_h) = 0 \quad \forall \mathbf{v}_h \in U_h^*,$$

for example, where \mathbf{u} and \mathbf{u}_h^* are the exact solution and the finite element solution, respectively. As we shall see, the advantage of the consistency is allowing the right-hand side \mathbf{f} to be less regular; see (5.50) and Remark 5.3.

4. Coercivity and condition number. We first investigate properties of the mesh dependent bilinear forms.

LEMMA 4.1. *Under the shape-regular condition, there exist constants C_1, C_2 and C_3, C_4 , independent of h and K , such that*

$$(4.1) \quad C_1 h_K^3 \leq \sum_{l=1}^m \|\theta_{K,l} b_K\|_{0,K}^2 \leq C_2 h_K^3,$$

$$(4.2) \quad C_3 h_K \leq \sum_{l=1}^m \|\nabla(\theta_{K,l} b_K)\|_{0,K}^2 \leq C_4 h_K,$$

where $\theta_{K,l} \in \Upsilon_K$, $1 \leq l \leq m = 20$, with Υ_K given as in (3.11), and b_K is defined by (3.3).

Proof. Both (4.1) and (4.2) can be easily shown by the scaling argument [37, 21, 17], or by a direct approach as follows. Since $b_K = \lambda_1 \lambda_2 \lambda_3 \lambda_4$, and $\theta_{K,l}$ is either $\frac{1}{2} \lambda_i (3 \lambda_i - 1) (3 \lambda_i - 2)$ (at vertices), or $\frac{9}{2} \lambda_i \lambda_j (3 \lambda_i - 1)$, $\frac{9}{2} \lambda_i \lambda_j (3 \lambda_j - 1)$ (at two-edge Gaussian nodes), or $27 \lambda_i \lambda_j \lambda_k$ (at face barycentric nodes), using the following formula on tetrahedron K

$$\int_K \lambda_1^{n_1} \lambda_2^{n_2} \lambda_3^{n_3} \lambda_4^{n_4} = |K| \frac{(n_1)!(n_2)!(n_3)!(n_4)!}{(n_1 + n_2 + n_3 + n_4 + 3)!}, \quad (\text{for nonnegative integers } n_j)$$

under the shape-regular condition [37, 16], it is not difficult to show that (4.1) and (4.2) hold. \square

LEMMA 4.2. *We have*

$$(4.3) \quad |S_{h,\text{div}}(\mathbf{u}, \mathbf{v})| \leq \|\mathbf{u}\|_0 \|\mathbf{v}\|_0,$$

$$(4.4) \quad |S_{h,\text{curl}}(\mathbf{u}, \mathbf{v})| \leq 3 \|\mathbf{u}\|_0 \|\mathbf{v}\|_0,$$

$$(4.5) \quad 0 \leq S_{h,\text{div}}(\mathbf{v}, \mathbf{v}) \leq C \sum_{K \in \mathcal{C}_h} h_K^2 \|\text{div } \mathbf{v}\|_{0,K}^2,$$

$$(4.6) \quad 0 \leq S_{h,\text{curl}}(\mathbf{v}, \mathbf{v}) \leq C \sum_{K \in \mathcal{C}_h} h_K^2 \|\text{curl } \mathbf{v}\|_{0,K}^2.$$

Proof. The left-hand sides of (4.5)–(4.6) are obvious. We only prove (4.3) and the right-hand side of (4.5) as examples, while (4.4) and the right-hand side of (4.6) can

be estimated in the same way, only noting that (2.14) will be used in proving (4.4). We first prove (4.3). From the Cauchy–Schwarz inequality we have

$$\left| \sum_{l=1}^m (\mathbf{u}, \nabla(\theta_{K,l} b_K))_{0,K} (\mathbf{v}, \nabla(\theta_{K,l} b_K))_{0,K} \right| \leq \|\mathbf{u}\|_{0,K} \|\mathbf{v}\|_{0,K} \sum_{l=1}^m \|\nabla(\theta_{K,l} b_K)\|_{0,K}^2,$$

and

$$\begin{aligned} |S_{h,\text{div}}(\mathbf{u}, \mathbf{v})| &= \left| \sum_{K \in \mathcal{C}_h} \frac{\sum_{l=1}^m (\mathbf{u}, \nabla(\theta_{K,l} b_K))_{0,K} (\mathbf{v}, \nabla(\theta_{K,l} b_K))_{0,K}}{\sum_{l=1}^m \|\nabla(\theta_{K,l} b_K)\|_{0,K}^2} \right| \\ (4.7) \quad &\leq \sum_{K \in \mathcal{C}_h} \|\mathbf{u}\|_{0,K} \|\mathbf{v}\|_{0,K} \leq \left(\sum_{K \in \mathcal{C}_h} \|\mathbf{u}\|_{0,K}^2 \right)^{\frac{1}{2}} \left(\sum_{K \in \mathcal{C}_h} \|\mathbf{v}\|_{0,K}^2 \right)^{\frac{1}{2}} \\ &= \|\mathbf{u}\|_0 \|\mathbf{v}\|_0. \end{aligned}$$

We next prove the right-hand side of (4.5). Since $\theta_{K,l} b_K \in H_0^1(K)$, we have from Green’s formula of integration by parts

$$(4.8) \quad (\mathbf{v}, \nabla(\theta_{K,l} b_K))_{0,K} = -(\text{div } \mathbf{v}, \theta_{K,l} b_K)_{0,K},$$

and then

$$(4.9) \quad S_{h,\text{div}}(\mathbf{v}, \mathbf{v}) = \sum_{K \in \mathcal{C}_h} \frac{\sum_{l=1}^m ((\mathbf{v}, \nabla(\theta_{K,l} b_K))_{0,K})^2}{\sum_{l=1}^m \|\nabla(\theta_{K,l} b_K)\|_{0,K}^2} = \sum_{K \in \mathcal{C}_h} \frac{\sum_{l=1}^m ((\text{div } \mathbf{v}, \theta_{K,l} b_K)_{0,K})^2}{\sum_{l=1}^m \|\nabla(\theta_{K,l} b_K)\|_{0,K}^2},$$

where, from the Cauchy–Schwarz inequality and the right-hand side of (4.1),

$$(4.10) \quad \sum_{l=1}^m ((\text{div } \mathbf{v}, \theta_{K,l} b_K)_{0,K})^2 \leq \|\text{div } \mathbf{v}\|_{0,K}^2 \sum_{l=1}^m \|\theta_{K,l} b_K\|_{0,K}^2 \leq C h_K^3 \|\text{div } \mathbf{v}\|_{0,K}^2.$$

Combining (4.9)–(4.10) and the left-hand side of (4.2) obtains the right-hand side of (4.5). \square

Now we introduce a mesh-dependent norm on U_h :

$$(4.11) \quad \|\mathbf{v}\|_h^2 := \sum_{K \in \mathcal{C}_h} h_K^2 \|\text{div } \mathbf{v}\|_{0,K}^2 + \sum_{K \in \mathcal{C}_h} h_K^2 \|\mathbf{curl } \mathbf{v}\|_{0,K}^2.$$

LEMMA 4.3. *For all $\mathbf{v} \in U_h$ we have*

$$(4.12) \quad S_{h,\text{div}}(\mathbf{v}, \mathbf{v}) \geq C \sum_{K \in \mathcal{C}_h} h_K^2 \|\text{div } \mathbf{v}\|_{0,K}^2,$$

$$(4.13) \quad S_{h,\text{curl}}(\mathbf{v}, \mathbf{v}) \geq C \sum_{K \in \mathcal{C}_h} h_K^2 \|\mathbf{curl } \mathbf{v}\|_{0,K}^2.$$

As a consequence, for $S_h(\mathbf{u}, \mathbf{v})$, defined as in (3.17), there holds

$$(4.14) \quad S_h(\mathbf{v}, \mathbf{v}) \geq C \|\mathbf{v}\|_h^2.$$

Proof. We only prove (4.12), while (4.13) can be proven in the same way. Given any $\mathbf{v} \in U_h$. From the element-local inclusion properties in (3.24) we, thus, write on K

$$(4.15) \quad \operatorname{div} \mathbf{v} = \sum_{l=1}^m c_l \theta_{K,l},$$

where $c_l \in \mathbb{R}$ are coefficients, and $\theta_{K,l} \in \Upsilon_K$ defined by (3.11). We have

$$(4.16) \quad \sum_{l=1}^m ((\mathbf{v}, \nabla (\theta_{K,l} b_K))_{0,K})^2 = \sum_{l=1}^m ((\operatorname{div} \mathbf{v}, \theta_{K,l} b_K)_{0,K})^2 = \sum_{l=1}^m (\mathbf{c}' \mathbf{d}_l)^2 = \mathbf{c}' A_K^2 \mathbf{c},$$

where $\mathbf{c} = (c_1, \dots, c_m)' \in \mathbb{R}^m$, $\mathbf{d}_l = (d_{1,l}, \dots, d_{m,l})' \in \mathbb{R}^m$, $1 \leq l \leq m$, with $d_{i,l} = (\theta_{K,i}, \theta_{K,l} b_K)_{0,K}$, $1 \leq i, l \leq m$, and A_K is the ‘mass’ matrix with $A_K = [\mathbf{d}_1, \dots, \mathbf{d}_m] \in \mathbb{R}^{m \times m}$. Clearly, A_K is symmetric and positive definite. Let $T \in \mathbb{R}^{m \times m}$ be the orthogonal matrix such that $A_K = T' \operatorname{diag} (\lambda_1, \dots, \lambda_m) T$, where $0 < \lambda_1 \leq \dots \leq \lambda_m$ are the eigenvalues of A_K . Using the scaling argument, we can easily show

$$(4.17) \quad \lambda_1 \geq C h_K^3.$$

Let $\bar{\mathbf{c}} = T \mathbf{c} = (\bar{c}_1, \dots, \bar{c}_m)' \in \mathbb{R}^m$, we have from (4.16) that

$$(4.18) \quad \sum_{l=1}^m ((\mathbf{v}, \nabla (\theta_{K,l} b_K))_{0,K})^2 = \sum_{l=1}^m (\bar{c}_l \lambda_l)^2.$$

On the other hand, by a similar argument we have from (4.15) that

$$(4.19) \quad (\operatorname{div} \mathbf{v}, \operatorname{div} \mathbf{v} b_K)_{0,K} = \sum_{l=1}^m (\bar{c}_l)^2 \lambda_l.$$

We then obtain

$$(4.20) \quad \sum_{l=1}^m ((\mathbf{v}, \nabla (\theta_{K,l} b_K))_{0,K})^2 = \sum_{l=1}^m (\bar{c}_l \lambda_l)^2 \geq \lambda_1 \sum_{l=1}^m (\bar{c}_l)^2 \lambda_l = \lambda_1 (\operatorname{div} \mathbf{v}, \operatorname{div} \mathbf{v} b_K)_{0,K}.$$

But, using the scaling argument we can have

$$(4.21) \quad (\operatorname{div} \mathbf{v}, \operatorname{div} \mathbf{v} b_K)_{0,K} \geq C (\operatorname{div} \mathbf{v}, \operatorname{div} \mathbf{v})_{0,K}.$$

Hence, from (4.20), (4.21), and (4.17),

$$(4.22) \quad \sum_{l=1}^m ((\mathbf{v}, \nabla (\theta_{K,l} b_K))_{0,K})^2 \geq C h_K^3 \|\operatorname{div} \mathbf{v}\|_{0,K}^2.$$

Then we have from (4.22) and the right-hand side of (4.2)

$$(4.23) \quad S_{h,\operatorname{div}}(\mathbf{v}, \mathbf{v}) = \sum_{K \in \mathcal{C}_h} \frac{\sum_{l=1}^m ((\mathbf{v}, \nabla (\theta_{K,l} b_K))_{0,K})^2}{\sum_{l=1}^m \|\nabla (\theta_{K,l} b_K)\|_{0,K}^2} \geq C \sum_{K \in \mathcal{C}_h} h_K^2 \|\operatorname{div} \mathbf{v}\|_{0,K}^2.$$

This completes the proof. \square

Remark 4.1. With $\mathcal{S}_h(\mathbf{u}, \mathbf{v})$ defined in (3.17), Lemmas 4.2 and 4.3 lead to

$$C \|\mathbf{v}\|_h^2 \leq \mathcal{S}_h(\mathbf{v}, \mathbf{v}) \leq C' \|\mathbf{v}\|_h^2 \quad \forall \mathbf{v} \in U_h.$$

One might, thus, think that, instead of using $\mathcal{S}_h(\mathbf{u}, \mathbf{v})$, it would be more convenient to use the following stabilization term $\mathcal{S}_h^\sharp(\mathbf{u}, \mathbf{v})$:

$$(4.24) \quad \mathcal{S}_h^\sharp(\mathbf{u}, \mathbf{v}) := \sum_{K \in \mathcal{C}_h} h_K^2 (\operatorname{div} \mathbf{u}, \operatorname{div} \mathbf{v})_{0,K} + \sum_{K \in \mathcal{C}_h} h_K^2 (\operatorname{curl} \mathbf{u}, \operatorname{curl} \mathbf{v})_{0,K}.$$

In that case, however, a correct convergent finite element solution may not be obtained when the exact solution is not in H^1 . This was confirmed by our numerical experiments (which is not reported in this paper). Such an incorrect convergence may be explained as in section 1. In fact, taking $h_K = h$ for all K , we have

$$\mathcal{S}_h^\sharp(\mathbf{u}, \mathbf{v}) = h^2 (\operatorname{div} \mathbf{u}, \operatorname{div} \mathbf{v}) + h^2 (\operatorname{curl} \mathbf{u}, \operatorname{curl} \mathbf{v}) = h^2 (\nabla \mathbf{u}, \nabla \mathbf{v}) \quad \forall \mathbf{u}, \mathbf{v} \in U_h,$$

which may enforce a convergence of the finite element solution \mathbf{u}_h to an element in H^1 . On the other hand, the $\mathcal{S}_h(\mathbf{u}, \mathbf{v})$ defined as in (3.17) is suitable for the nonsmooth solution that does not belong to H^1 , since no partial differential derivatives are applied on both \mathbf{u} and \mathbf{v} (where, to see this point for $\mathcal{S}_{h, \operatorname{curl}}(\mathbf{u}, \mathbf{v})$, (2.14) was used).

For the analysis of coercivity, below we recall the L^2 -orthogonal decomposition and the regular-singular decomposition of vector fields on Lipschitz polyhedra. The following first two propositions are due to [34], see also [4, 14].

PROPOSITION 4.1. *We have the following L^2 -orthogonal decomposition of vector fields with respect to the L^2 inner product (\cdot, \cdot) :*

$$(L^2(\Omega))^3 = \nabla H_0^1(\Omega) \oplus \operatorname{curl} (H(\operatorname{curl}; \Omega) \cap H_0(\operatorname{div}^0; \Omega)).$$

PROPOSITION 4.2. *For any $\mathbf{v} \in H(\operatorname{curl}; \Omega) \cap H_0(\operatorname{div}^0; \Omega)$, or for any $\mathbf{v} \in H_0(\operatorname{curl}; \Omega) \cap H(\operatorname{div}^0; \Omega)$, we have*

$$\|\mathbf{v}\|_0 \leq C \|\operatorname{curl} \mathbf{v}\|_0.$$

PROPOSITION 4.3 ([12, 13]). *For any $\psi \in H(\operatorname{curl}; \Omega) \cap H_0(\operatorname{div}; \Omega)$, it can be written as the following regular-singular decomposition:*

$$\psi = \psi^0 + \nabla q,$$

where $\psi^0 \in H_0(\operatorname{div}; \Omega) \cap (H^1(\Omega))^3$ is called “regular part” and $q \in H^1(\Omega) \setminus \mathbb{R}$ “singular part,” satisfying

$$\|\psi^0\|_1 \leq C \{\|\psi\|_0 + \|\operatorname{curl} \psi\|_0 + \|\operatorname{div} \psi\|_0\}.$$

THEOREM 4.1. *Let the stabilization parameter $\alpha \geq \alpha_0 > 0$, with α_0 being determined according to (4.27) below, i.e., $\alpha \geq \alpha_0 = C_6$ as given in (4.28). We have*

$$(4.25) \quad \mathcal{L}_h(\mathbf{v}, \mathbf{v}) \geq C \|\mathbf{v}\|_0^2 \quad \forall \mathbf{v} \in U_h.$$

As a consequence of Lax–Milgram lemma, problem (3.20) has a unique solution.

Proof. Since

$$(4.26) \quad \mathcal{L}_h(\mathbf{v}, \mathbf{v}) = \|R_h(\operatorname{curl} \mathbf{v})\|_0^2 + s \|\check{R}_h(\operatorname{div} \mathbf{v})\|_0^2 + \alpha \mathcal{S}_h(\mathbf{v}, \mathbf{v}),$$

we need only prove that there exist positive constants C_5 and C_6 such that

$$(4.27) \quad \|R_h(\mathbf{curl} \mathbf{v})\|_0^2 + s \left\| \check{R}_h(\operatorname{div} \mathbf{v}) \right\|_0^2 \geq C_5 \|\mathbf{v}\|_0^2 - C_6 \mathcal{S}_h(\mathbf{v}, \mathbf{v}) \quad \forall \mathbf{v} \in U_h.$$

Then the theorem follows by choosing

$$(4.28) \quad \alpha \geq \alpha_0 := C_6.$$

Note that s may be chosen in advance as any given positive constant, say $s = 1$.

From Proposition 4.1 we write \mathbf{v} as the following L^2 -orthogonal decomposition with respect to the L^2 inner product:

$$(4.29) \quad \mathbf{v} = \nabla p + \mathbf{curl} \psi,$$

with $p \in H_0^1(\Omega)$ and $\psi \in H(\mathbf{curl}; \Omega) \cap H_0(\operatorname{div}^0; \Omega)$, satisfying

$$(4.30) \quad \|\mathbf{v}\|_0^2 = \|\nabla p\|_0^2 + \|\mathbf{curl} \psi\|_0^2.$$

We also have from Proposition 4.2

$$(4.31) \quad \|\psi\|_0 \leq C \|\mathbf{curl} \psi\|_0.$$

From Proposition 4.3 we further write ψ as

$$(4.32) \quad \psi = \psi^0 + \nabla q,$$

where $\psi^0 \in H_0(\operatorname{div}; \Omega) \cap (H^1(\Omega))^3$, $\nabla q \in H(\mathbf{curl}^0; \Omega)$ with $q \in H^1(\Omega)/\mathbb{R}$, and we have from Proposition 4.3 and (4.31)

$$(4.33) \quad \|\psi^0\|_1 \leq C \|\mathbf{curl} \psi\|_0.$$

According to two components (p, ψ) in (4.29), we divide the proof of (4.27) into two steps.

Step 1. We consider p . We take $\tilde{p} \in Q_h$ as the local L^2 projection of p such that [30, 36]

$$(4.34) \quad \tilde{p}|_K = \frac{1}{|K|} \int_K p \quad \forall K \in \mathcal{C}_h,$$

$$(4.35) \quad \left(\sum_{K \in \mathcal{C}_h} h_K^{-2} \|p - \tilde{p}\|_{0,K}^2 \right)^{\frac{1}{2}} + \|\tilde{p}\|_0 \leq C \|p\|_1.$$

Let $\delta > 0$ be a constant to be determined. We have

$$(4.36) \quad \left\| \check{R}_h(\operatorname{div} \mathbf{v}) \right\|_0^2 = \left\| \check{R}_h(\operatorname{div} \mathbf{v}) + \delta \tilde{p} \right\|_0^2 - \delta^2 \|\tilde{p}\|_0^2 - 2\delta \left(\check{R}_h(\operatorname{div} \mathbf{v}), \tilde{p} \right),$$

where

$$(4.37) \quad -\delta^2 \|\tilde{p}\|_0^2 \geq -\delta^2 \|p\|_0^2 \geq -\delta^2 C \|\nabla p\|_0^2,$$

$$(4.38) \quad \begin{aligned} -2\delta \left(\check{R}_h(\operatorname{div} \mathbf{v}), \tilde{p} \right) &= -2\delta \sum_{K \in \mathcal{C}_h} (\operatorname{div} \mathbf{v}, \tilde{p})_{0,K} \\ &= 2\delta \sum_{K \in \mathcal{C}_h} (\operatorname{div} \mathbf{v}, p - \tilde{p})_{0,K} - 2\delta \sum_{K \in \mathcal{C}_h} (\operatorname{div} \mathbf{v}, p)_{0,K}, \end{aligned}$$

$$(4.39) \quad -2\delta \sum_{K \in \mathcal{C}_h} (\operatorname{div} \mathbf{v}, p)_{0,K} = 2\delta (\mathbf{v}, \nabla p) = 2\delta \|\nabla p\|_0^2,$$

$$(4.40) \quad \begin{aligned} 2\delta \sum_{K \in \mathcal{C}_h} (\operatorname{div} \mathbf{v}, p - \tilde{p})_{0,K} &\geq -2\delta \left(\sum_{K \in \mathcal{C}_h} h_K^2 \|\operatorname{div} \mathbf{v}\|_{0,K}^2 \right)^{1/2} \left(\sum_{K \in \mathcal{C}_h} h_K^{-2} \|p - \tilde{p}\|_{0,K}^2 \right)^{1/2} \\ &\geq -2\delta C \left(\sum_{K \in \mathcal{C}_h} h_K^2 \|\operatorname{div} \mathbf{v}\|_{0,K}^2 \right)^{\frac{1}{2}} \|p\|_1 \\ &\geq - \sum_{K \in \mathcal{C}_h} h_K^2 \|\operatorname{div} \mathbf{v}\|_{0,K}^2 - C\delta^2 \|\nabla p\|_0^2. \end{aligned}$$

Summarizing (4.36)–(4.40) and choosing

$$(4.41) \quad 0 < \delta < 1/C,$$

we have

$$(4.42) \quad \left\| \check{R}_h(\operatorname{div} \mathbf{v}) \right\|_0^2 \geq \delta(2 - 2C\delta) \|\nabla p\|_0^2 - \sum_{K \in \mathcal{C}_h} h_K^2 \|\operatorname{div} \mathbf{v}\|_{0,K}^2.$$

Step 2. We consider ψ (The argument is similar to that in *Step 1*, but we still give the details). We take $\psi^0 \in (P_h)^3$ as the local L^2 projection of ψ^0 such that

$$(4.43) \quad \int_K \widetilde{\psi}^0 \cdot \mathbf{q} = \int_K \psi^0 \cdot \mathbf{q} \quad \forall \mathbf{q} \in (\mathcal{P}_1(K))^3, \forall K \in \mathcal{C}_h,$$

$$(4.44) \quad \left(\sum_{K \in \mathcal{C}_h} h_K^{-2} \|\psi^0 - \widetilde{\psi}^0\|_{0,K}^2 \right)^{\frac{1}{2}} + \|\widetilde{\psi}^0\|_0 \leq C \|\psi^0\|_1.$$

Let $\delta > 0$ be a constant to be determined. We have

$$(4.45) \quad \|R_h(\mathbf{curl} \mathbf{v})\|_0^2 = \left\| R_h(\mathbf{curl} \mathbf{v}) - \delta \widetilde{\psi}^0 \right\|_0^2 - \delta^2 \|\widetilde{\psi}^0\|_0^2 + 2\delta (R_h(\mathbf{curl} \mathbf{v}), \widetilde{\psi}^0),$$

where

$$(4.46) \quad -\delta^2 \|\widetilde{\psi}^0\|_0^2 \geq -\delta^2 C \|\psi^0\|_1^2 \geq -\delta^2 C \|\mathbf{curl} \psi\|_0^2, \quad (\text{by (4.44) and (4.33)})$$

$$(4.47) \quad \begin{aligned} 2\delta (R_h(\mathbf{curl} \mathbf{v}), \widetilde{\psi}^0) &= 2\delta \sum_{K \in \mathcal{C}_h} (\mathbf{curl} \mathbf{v}, \widetilde{\psi}^0)_{0,K} \\ &= 2\delta \sum_{K \in \mathcal{C}_h} (\mathbf{curl} \mathbf{v}, \widetilde{\psi}^0 - \psi^0)_{0,K} + 2\delta \sum_{K \in \mathcal{C}_h} (\mathbf{curl} \mathbf{v}, \psi^0)_{0,K}, \end{aligned}$$

$$(4.48) \quad 2\delta \sum_{K \in \mathcal{C}_h} (\mathbf{curl} \mathbf{v}, \psi^0)_{0,K} = 2\delta (\mathbf{v}, \mathbf{curl} \psi^0) = 2\delta (\mathbf{v}, \mathbf{curl} \psi) = 2\delta \|\mathbf{curl} \psi\|_0^2,$$

$$\begin{aligned}
 (4.49) \quad & 2\delta \sum_{K \in \mathcal{C}_h} \left(\mathbf{curl} \mathbf{v}, \widetilde{\psi}^0 - \psi^0 \right)_{0,K} \\
 & \geq -2\delta \left(\sum_{K \in \mathcal{C}_h} h_K^{-2} \|\widetilde{\psi}^0 - \psi^0\|_{0,K}^2 \right)^{\frac{1}{2}} \left(\sum_{K \in \mathcal{C}_h} h_K^2 \|\mathbf{curl} \mathbf{v}\|_{0,K}^2 \right)^{\frac{1}{2}} \\
 & \geq -2\delta C \|\psi^0\|_1 \left(\sum_{K \in \mathcal{C}_h} h_K^2 \|\mathbf{curl} \mathbf{v}\|_{0,K}^2 \right)^{\frac{1}{2}} \\
 & \geq -2\delta C \|\mathbf{curl} \psi\|_0 \left(\sum_{K \in \mathcal{C}_h} h_K^2 \|\mathbf{curl} \mathbf{v}\|_{0,K}^2 \right)^{1/2} \\
 & \geq - \sum_{K \in \mathcal{C}_h} h_K^2 \|\mathbf{curl} \mathbf{v}\|_{0,K}^2 - C \delta^2 \|\mathbf{curl} \psi\|_0^2.
 \end{aligned}$$

Summarizing (4.45)–(4.49) and choosing $0 < \delta < 1/C$, we have

$$(4.50) \quad \|R_h(\mathbf{curl} \mathbf{v})\|_0^2 \geq \delta(2 - 2C\delta) \|\mathbf{curl} \psi\|_0^2 - \sum_{K \in \mathcal{C}_h} h_K^2 \|\mathbf{curl} \mathbf{v}\|_{0,K}^2.$$

Finally, from (4.42), (4.50), (4.30), (4.14), and (4.11), we obtain

$$\begin{aligned}
 (4.51) \quad & \|R_h(\mathbf{curl} \mathbf{v})\|_0 + s \left\| \check{R}_h(\operatorname{div} \mathbf{v}) \right\|_0^2 \geq C_5 (\|\nabla p\|_0^2 + \|\mathbf{curl} \psi\|_0^2) - \|\mathbf{v}\|_h^2 \\
 & \geq C_5 \|\mathbf{v}\|_0^2 - C_6 \mathcal{S}_h(\mathbf{v}, \mathbf{v}),
 \end{aligned}$$

where C_5 and C_6 are two positive constants independent of h and K . The proof is finished. \square

Remark 4.2. In fact, the regularization parameter s and the stabilization parameter α can be both taken as any given positive constants, since $\mathcal{L}_h(\cdot, \cdot)$ is nonnegative no matter what $\alpha \geq 0$ and $s \geq 0$ are, i.e., for all $\alpha, s \in [0, +\infty)$,

$$\mathcal{L}_h(\mathbf{v}, \mathbf{v}) \geq 0 \quad \forall \mathbf{v} \in U_h.$$

For example, denoting by $\mathcal{L}_h^{1,1}$ the bilinear form in (3.20) for the choice $\alpha = s = 1$ and by $\mathcal{L}^{\alpha,s}$ for the choice (4.28) and any $s > 0$, we still have the coercivity as stated in (4.25) for $\mathcal{L}_h^{1,1}$, since we have from the above nonnegativeness property that

$$\mathcal{L}_h^{1,1}(\mathbf{v}, \mathbf{v}) = \|R_h(\mathbf{curl} \mathbf{v})\|_0^2 + \left\| \check{R}_h(\operatorname{div} \mathbf{v}) \right\|_0^2 + \mathcal{S}_h(\mathbf{v}, \mathbf{v}) \geq (\max(1, \alpha, s))^{-1} \mathcal{L}^{\alpha,s}(\mathbf{v}, \mathbf{v}).$$

On the other hand, a suitable large α will indeed yield smaller errors in their values, although whatever value of α does not affect the convergence rate, see the numerical experiments in section 6.

Remark 4.3. Regarding \mathcal{L}_h^* in (3.31), we can obtain the same coercivity as in (4.25) by a similar argument, but replacing *Step 2* by the following: since $\mathbf{v} \in H_0(\mathbf{curl}; \Omega)$, we have from (4.29)

$$(4.52) \quad \mathbf{curl} \psi \in H_0(\mathbf{curl}; \Omega) \cap H(\operatorname{div}^0; \Omega),$$

and applying Proposition 4.2 with $\mathbf{curl} \psi \in H_0(\mathbf{curl}; \Omega) \cap H(\operatorname{div}^0; \Omega)$ to obtain

$$(4.53) \quad \|\mathbf{curl} \mathbf{v}\|_0^2 = \|\mathbf{curl} \mathbf{curl} \psi\|_0^2 \geq C \|\mathbf{curl} \psi\|_0^2,$$

and both (4.53) and (4.42) yield an estimation similar to (4.27), i.e.,

$$(4.54) \quad \|\mathbf{curl} \mathbf{v}\|_0^2 + s \left\| \check{R}_h(\mathbf{div} \mathbf{v}) \right\|_0^2 \geq C_7 \|\mathbf{v}\|_0^2 - C_8 S_{h,\mathbf{div}}(\mathbf{v}, \mathbf{v}),$$

from which we have the following coercivity for \mathcal{L}_h^* with the stabilization parameter $\alpha > C_8$:

$$(4.55) \quad \mathcal{L}_h^*(\mathbf{v}, \mathbf{v}) \geq C \left(\|\mathbf{v}\|_{0,\mathbf{curl}}^2 + \sum_{K \in \mathcal{C}_h} h_K^2 \|\mathbf{div} \mathbf{v}\|_{0,K}^2 \right) \quad \forall \mathbf{v} \in U_h^*.$$

The above argument goes also to the \mathcal{L}_h^{**} in (3.35) in the same way, only noting that

$$(4.56) \quad S_{h/2,\mathbf{div}}(\mathbf{v}, \mathbf{v}) \geq C \sum_{T \in \mathcal{T}_{h/2}} h_T^2 \|\mathbf{div} \mathbf{v}\|_{0,T}^2 \geq C \sum_{K \in \mathcal{C}_h} h_K^2 \|\mathbf{div} \mathbf{v}\|_{0,K}^2$$

holds for all $\mathbf{v} \in U_h^{**}$, where (4.56) can be shown by a similar argument used for Lemma 4.3.

Before closing this section, we give the condition number of the resulting linear system.

THEOREM 4.2. *Assume that the meshes are uniform as usual. Then, the condition number of the resulting linear system of problem (3.20) is of $\mathcal{O}(h^{-2})$.*

Proof. Since both R_h and \check{R}_h are local L^2 projectors, we have from the inverse estimates [21] that for all $\mathbf{v} \in U_h$

$$(4.57) \quad \left\| \check{R}_h(\mathbf{div} \mathbf{v}) \right\|_0 + \|R_h(\mathbf{curl} \mathbf{v})\|_0 \leq \|\mathbf{div} \mathbf{v}\|_0 + \|\mathbf{curl} \mathbf{v}\|_0 \leq C h^{-1} \|\mathbf{v}\|_0.$$

On the other hand, from Lemma 4.2 we have for all $\mathbf{v} \in U_h$

$$S_h(\mathbf{v}, \mathbf{v}) = S_{h,\mathbf{div}}(\mathbf{v}, \mathbf{v}) + S_{h,\mathbf{curl}}(\mathbf{v}, \mathbf{v}) \leq C \|\mathbf{v}\|_0^2.$$

Hence, we have

$$(4.58) \quad \mathcal{L}_h(\mathbf{v}, \mathbf{v}) = \|R_h(\mathbf{curl} \mathbf{v})\|_0^2 + s \left\| \check{R}_h(\mathbf{div} \mathbf{v}) \right\|_0^2 + \alpha S_h(\mathbf{v}, \mathbf{v}) \leq C h^{-2} \|\mathbf{v}\|_0^2 \quad \forall \mathbf{v} \in U_h,$$

which, together with the L^2 coercivity property in Theorem 4.1 and the symmetry property of \mathcal{L}_h , leads to the result. \square

5. Error estimates. In this section, we establish in an energy norm the error bound between the exact solution and the finite element solution. This consists mainly of how to estimate the inconsistent errors caused by the L^2 projector R_h and how to construct an appropriate interpolant of the exact solution to eliminate the effects of the first order derivatives from both *div* and *curl* operators on the solution that is not in H^1 , i.e., “eliminating” the *div* and *curl* operators in the context of (5.14) later on. The former depends on a profound result on the regular-singular decomposition of the curl of the solution and the latter resorts to the two L^2 projectors.

We first give estimates of inconsistency errors from the curl operator.

LEMMA 5.1. *Let \mathbf{u} and \mathbf{u}_h be the exact solution to problem (2.1)–(2.2) and the finite element solution to problem (3.20), respectively. We have for all $\mathbf{v}_h \in U_h$*

$$(5.1) \quad \mathcal{L}_h(\mathbf{u} - \mathbf{u}_h, \mathbf{v}_h) = (\mathbf{curl} \mathbf{u}, R_h(\mathbf{curl} \mathbf{v}_h)) - (\mathbf{curl} \mathbf{u}, \mathbf{curl} \mathbf{v}_h) + \alpha S_{h,\mathbf{curl}}(\mathbf{u}, \mathbf{v}_h).$$

Proof. From (3.12), (3.13), and the second equation in (2.1) we clearly have on U_h

$$(5.2) \quad S_{h,\text{div}}(\mathbf{u}, \mathbf{v}_h) = Z_{h,\text{div}}(g; \mathbf{v}_h).$$

On the other hand, we have from (3.15), (3.16), (2.1), (2.2), and (3.23) on U_h

$$(5.3) \quad \left(\check{R}_h(\text{div } \mathbf{u}), \check{R}_h(\text{div } \mathbf{v}_h) \right) = \left(\text{div } \mathbf{u}, \check{R}_h(\text{div } \mathbf{v}_h) \right) = \left(g, \check{R}_h(\text{div } \mathbf{v}_h) \right),$$

$$(5.4) \quad \begin{aligned} (R_h(\mathbf{curl } \mathbf{u}), R_h(\mathbf{curl } \mathbf{v}_h)) &= (\mathbf{curl } \mathbf{u}, R_h(\mathbf{curl } \mathbf{v}_h)) \\ &= (\mathbf{curl } \mathbf{u}, R_h(\mathbf{curl } \mathbf{v}_h)) \\ &\quad - (\mathbf{curl } \mathbf{u}, \mathbf{curl } \mathbf{v}_h) + (\mathbf{f}, \mathbf{v}_h), \end{aligned}$$

and we obtain (5.1). \square

Remark 5.1. Regarding (3.33) or (3.37), as pointed out in Remark 3.4, there are no inconsistent errors, see (3.38).

LEMMA 5.2. *Let \mathbf{u} be the solution of problem (2.1)–(2.2). We have for all $\mathbf{v}_h \in U_h$*

$$(5.5) \quad |S_{h,\mathbf{curl}}(\mathbf{u}, \mathbf{v}_h)| \leq Ch \|\mathbf{curl } \mathbf{u}\|_0 \left(\sum_{K \in \mathcal{C}_h} h_K^2 \|\mathbf{curl } \mathbf{v}_h\|_{0,K}^2 \right)^{\frac{1}{2}}.$$

Proof. Equation (5.5) is derived from the same argument as in proving Lemma 4.2. \square

PROPOSITION 5.1. *For any $\mathbf{v} \in H_0(\mathbf{curl}; \Omega) \cap H(\text{div}; \Omega)$ or for any $\mathbf{v} \in H(\mathbf{curl}; \Omega) \cap H_0(\text{div}; \Omega)$ we have $\mathbf{v} \in (H^r(\Omega))^3$ for some real number $r > 1/2$, satisfying*

$$\|\mathbf{v}\|_r \leq C (\|\text{div } \mathbf{v}\|_0 + \|\mathbf{curl } \mathbf{v}\|_0).$$

LEMMA 5.3. *Let $\mathbf{u} \in U$ be the solution of problem (2.1)–(2.2). Then, we have $\mathbf{u}, \mathbf{curl } \mathbf{u} \in (H^r(\Omega))^3$ for some real number $r > 1/2$, satisfying*

$$\|\mathbf{u}\|_r \leq C (\|\mathbf{f}\|_0 + \|g\|_0), \quad \|\mathbf{curl } \mathbf{u}\|_r \leq C \|\mathbf{f}\|_0.$$

Proof. Since $\mathbf{u} \in U = H(\text{div}; \Omega) \cap H_0(\mathbf{curl}; \Omega)$ is the solution of problem (2.1)–(2.2), then for all $\mathbf{v} \in U$

$$(\mathbf{curl } \mathbf{u}, \mathbf{curl } \mathbf{v}) + (\text{div } \mathbf{u}, \text{div } \mathbf{v}) = (\mathbf{f}, \mathbf{v}) + (g, \text{div } \mathbf{v}),$$

which, together with Proposition 5.1, leads to the stated result. Moreover, since $\mathbf{z} = \mathbf{curl } \mathbf{u}$ satisfies

$$\mathbf{curl } \mathbf{z} = \mathbf{f}, \quad \text{div } \mathbf{z} = 0 \quad \text{in } \Omega, \quad \mathbf{z} \cdot \mathbf{n}|_{\Gamma} = 0,$$

we have from Proposition 5.1 again

$$\|\mathbf{curl } \mathbf{u}\|_r = \|\mathbf{z}\|_r \leq C \|\mathbf{curl } \mathbf{z}\|_0 = C \|\mathbf{f}\|_0. \quad \square$$

PROPOSITION 5.2 ([51, 29, 27, 26, 31]). *Additionally, assume that $\mathbf{f} \in H(\mathbf{curl}; \Omega) \cap (H^r(\Omega))^3$ for some real number $r > 1/2$. Let \mathbf{z} be given as in (2.3), satisfying (2.4)–(2.5). Then, \mathbf{z} can be written into the following regular-singular decomposition*

$$\mathbf{z} = \mathbf{z}_H + \nabla \varphi \quad \text{in } \Omega,$$

where $\mathbf{z}_H \in H(\mathbf{curl}; \Omega) \cap (H^{1+r}(\Omega))^3$ and $\varphi \in H^1(\Omega) \cap H^{1+r}(\Omega)$ satisfy

$$\|\mathbf{z}_H\|_{1+r} + \|\varphi\|_{1+r} \leq C (\|\mathbf{f}\|_r + \|\mathbf{curl} \mathbf{f}\|_0).$$

LEMMA 5.4. Let \mathbf{u} be the solution to problem (2.1)–(2.2), with the additional assumption that $\mathbf{f} \in H(\mathbf{curl}; \Omega) \cap (H^r(\Omega))^3$ for some real number $r > 1/2$. We have for all $\mathbf{v}_h \in U_h$

$$(5.6) \quad (\mathbf{curl} \mathbf{u}, R_h(\mathbf{curl} \mathbf{v}_h)) - (\mathbf{curl} \mathbf{u}, \mathbf{curl} \mathbf{v}_h) \leq C h^r (\|\mathbf{f}\|_r + \|\mathbf{curl} \mathbf{f}\|_0) \left(\|R_h(\mathbf{curl} \mathbf{v}_h)\|_0 + \left(\sum_{K \in \mathcal{C}_h} h_K^2 \|\mathbf{curl} \mathbf{v}_h\|_{0,K}^2 \right)^{\frac{1}{2}} \right).$$

Proof. According to the regular-singular decomposition of $\mathbf{z} = \mathbf{curl} \mathbf{u} = \mathbf{z}_H + \nabla \varphi$ in Proposition 5.2, we define $\widetilde{\mathbf{curl} \mathbf{u}} \in (P_h)^3$ as the interpolation to $\mathbf{curl} \mathbf{u}$ by

$$(5.7) \quad \widetilde{\mathbf{curl} \mathbf{u}} := \widetilde{\mathbf{z}}_H + \nabla \widetilde{\varphi},$$

where $\widetilde{\mathbf{z}}_H \in (P_h)^3$ is the local L^2 projection of \mathbf{z}_H , and $\widetilde{\varphi} \in P_h \cap H^1(\Omega)$ is the usual interpolant of φ . We have

$$(5.8) \quad \left(\sum_{K \in \mathcal{C}_h} h_K^{-2} \|\mathbf{z}_H - \widetilde{\mathbf{z}}_H\|_{0,K}^2 \right)^{1/2} \leq C h^r \|\mathbf{z}_H\|_{1+r}, \quad \|\varphi - \widetilde{\varphi}\|_1 \leq C h^r \|\varphi\|_{1+r},$$

$$(5.9) \quad \|\mathbf{z}_H - \widetilde{\mathbf{z}}_H\|_0 \leq C h^r \|\mathbf{z}_H\|_r.$$

We, thus, have

$$(5.10) \quad \left\| \mathbf{curl} \mathbf{u} - \widetilde{\mathbf{curl} \mathbf{u}} \right\|_0 \leq \|\mathbf{z}_H - \widetilde{\mathbf{z}}_H\|_0 + \|\nabla(\varphi - \widetilde{\varphi})\|_0 \leq C h^r (\|\mathbf{z}_H\|_r + \|\varphi\|_{1+r}),$$

$$(5.11) \quad (\nabla(\widetilde{\varphi} - \varphi), \mathbf{curl} \mathbf{v}_h) = 0 \quad \forall \mathbf{v}_h \in U_h.$$

Since we have from (3.16)

$$(5.12) \quad (\widetilde{\mathbf{curl} \mathbf{u}}, R_h(\mathbf{curl} \mathbf{v}_h)) = (\widetilde{\mathbf{curl} \mathbf{u}}, \mathbf{curl} \mathbf{v}_h),$$

we then have from (5.10)–(5.12)

$$(5.13) \quad \begin{aligned} & (\mathbf{curl} \mathbf{u}, R_h(\mathbf{curl} \mathbf{v}_h)) - (\mathbf{curl} \mathbf{u}, \mathbf{curl} \mathbf{v}_h) \\ &= (\mathbf{curl} \mathbf{u} - \widetilde{\mathbf{curl} \mathbf{u}}, R_h(\mathbf{curl} \mathbf{v}_h)) \\ & \quad + (\widetilde{\mathbf{curl} \mathbf{u}} - \mathbf{curl} \mathbf{u}, \mathbf{curl} \mathbf{v}_h) \\ &= (\mathbf{curl} \mathbf{u} - \widetilde{\mathbf{curl} \mathbf{u}}, R_h(\mathbf{curl} \mathbf{v}_h)) + (\widetilde{\mathbf{z}}_H - \mathbf{z}_H, \mathbf{curl} \mathbf{v}_h) \\ & \leq C h^r (\|\mathbf{z}_H\|_r + \|\varphi\|_{1+r}) \|R_h(\mathbf{curl} \mathbf{v}_h)\|_0 \\ & \quad + C h^r \|\mathbf{z}_H\|_{1+r} \left(\sum_{K \in \mathcal{C}_h} h_K^2 \|\mathbf{curl} \mathbf{v}_h\|_{0,K}^2 \right)^{1/2}, \end{aligned}$$

which, together with Proposition 5.2, leads to (5.6). \square

In what follows, we construct an interpolant $\widetilde{\mathbf{u}} \in U_h$ of the solution \mathbf{u} .

LEMMA 5.5. *Let $\mathbf{u} \in U = H(\operatorname{div}; \Omega) \cap H_0(\mathbf{curl}; \Omega)$ be the solution to problem (2.1)–(2.2). Then, there exists a $\tilde{\mathbf{u}} \in U_h$ defined as in (3.10) such that*

$$(5.14) \quad \left\| \check{R}_h(\operatorname{div}(\mathbf{u} - \tilde{\mathbf{u}})) \right\|_0^2 = \|R_h(\mathbf{curl}(\mathbf{u} - \tilde{\mathbf{u}}))\|_0^2 = 0,$$

$$(5.15) \quad \|\mathbf{u} - \tilde{\mathbf{u}}\|_0 \leq C h^r \|\mathbf{u}\|_r.$$

Proof. From Lemma 5.3 we know that $\mathbf{u} \in (H^r(\Omega))^3$ for some real number $r > 1/2$. We first let $\mathbf{u}^0 \in (P_h \cap H^1(\Omega))^3 \cap H_0(\mathbf{curl}; \Omega)$ be such that [10, 11, 22, 52, 53]

$$(5.16) \quad \|\mathbf{u} - \mathbf{u}^0\|_0 + \left(\sum_{K \in \mathcal{C}_h} \sum_{F \subset \partial K} h_F \|\mathbf{u} - \mathbf{u}^0\|_{0,F}^2 \right)^{1/2} \leq C h^r \|\mathbf{u}\|_r, \quad r > \frac{1}{2}.$$

We then define $\tilde{\mathbf{u}} \in U_h$ by the following (5.17)–(5.19):

$$(5.17) \quad \tilde{\mathbf{u}}(a) = \mathbf{u}^0(a) \quad \text{for all vertices } a,$$

$$(5.18) \quad \int_{F_i} (\tilde{\mathbf{u}} - \mathbf{u}) \cdot \mathbf{q}_{F_i,l} = 0 \quad \forall \mathbf{q}_{F_i,l} \in \mathbf{P}_{F_i}, \forall F_i \in \partial K, \forall K \in \mathcal{C}_h,$$

where \mathbf{P}_{F_i} is given by (3.7) and $\partial K = \{F_i, 1 \leq i \leq 4\}$,

$$(5.19) \quad \int_K (\tilde{\mathbf{u}} - \mathbf{u}) = \mathbf{0}.$$

According to (3.10), on K with boundary $\partial K = \{F_i, 1 \leq i \leq 4\}$, we write $\tilde{\mathbf{u}} \in U_h$ as the following form:

$$(5.20) \quad \tilde{\mathbf{u}} = \mathbf{u}^0 + \sum_{i=1}^4 \sum_{l=1}^9 c_{i,l} \mathbf{q}_{F_i,l} b_{F_i} + \mathbf{c}_K b_K =: \hat{\mathbf{u}} + \mathbf{c}_K b_K,$$

where $c_{i,l} \in \mathbb{R}$ and $\mathbf{c}_K \in \mathbb{R}^3$ are all coefficients to be determined. Since the face bubble and the element bubble take zero at all vertices, (5.17) determines the linear part of $\tilde{\mathbf{u}}$, and (5.18) is to determine the face bubble part because the element bubble takes zero along all faces, and (5.19) is for the element bubble part. From (5.18) the coefficients $c_{i,l}, 1 \leq l \leq 9$, are determined uniquely by

$$(5.21) \quad \sum_{l=1}^9 c_{i,l} \int_{F_i} \mathbf{q}_{F_i,l} \cdot \mathbf{q}_{F_i,k} b_{F_i} = \int_{F_i} (\mathbf{u} - \mathbf{u}^0) \cdot \mathbf{q}_{F_i,k} \quad 1 \leq k \leq 9,$$

and from (5.19) the coefficient \mathbf{c}_K is given by

$$(5.22) \quad \mathbf{c}_K = \frac{\int_K (\mathbf{u} - \hat{\mathbf{u}})}{\int_K b_K}.$$

Using the scaling argument, we can easily obtain

$$(5.23) \quad \|\mathbf{u} - \hat{\mathbf{u}}\|_{0,K} \leq C \|\mathbf{u} - \mathbf{u}^0\|_{0,K} + C \sum_{F \subset \partial K} h_F^{\frac{1}{2}} \|\mathbf{u} - \mathbf{u}^0\|_{0,F},$$

and

$$(5.24) \quad \|\mathbf{u} - \tilde{\mathbf{u}}\|_{0,K} \leq C \|\mathbf{u} - \hat{\mathbf{u}}\|_{0,K}.$$

From (5.24), (5.23), and (5.16) it follows that (5.15) holds.

Equation (5.14) holds from the construction of $\tilde{\mathbf{u}}$: we have from (3.15) and (5.18) that

$$\begin{aligned}
 (5.25) \quad \left\| \check{R}_h(\operatorname{div}(\mathbf{u} - \tilde{\mathbf{u}})) \right\|_0^2 &= \sum_{K \in \mathcal{C}_h} \left(\operatorname{div}(\mathbf{u} - \tilde{\mathbf{u}}), \check{R}_h(\operatorname{div}(\mathbf{u} - \tilde{\mathbf{u}})) \right)_{0,K} \\
 &= \sum_{K \in \mathcal{C}_h} \sum_{F \subset \partial K} \int_F (\mathbf{u} - \tilde{\mathbf{u}}) \cdot (\mathbf{n} \check{R}_h(\operatorname{div}(\mathbf{u} - \tilde{\mathbf{u}}))) = 0,
 \end{aligned}$$

since $\mathbf{n} \check{R}_h(\operatorname{div}(\mathbf{u} - \tilde{\mathbf{u}}))|_F \in \mathbf{P}_F|_F$. Similarly, we have from (3.16), (5.19), and (5.18) that

$$\begin{aligned}
 (5.26) \quad \left\| R_h(\operatorname{curl}(\mathbf{u} - \tilde{\mathbf{u}})) \right\|_0^2 &= \sum_{K \in \mathcal{C}_h} (\operatorname{curl}(\mathbf{u} - \tilde{\mathbf{u}}), R_h(\operatorname{curl}(\mathbf{u} - \tilde{\mathbf{u}})))_{0,K} \\
 &= \sum_{K \in \mathcal{C}_h} (\mathbf{u} - \tilde{\mathbf{u}}, \operatorname{curl} R_h(\operatorname{curl}(\mathbf{u} - \tilde{\mathbf{u}})))_{0,K} \\
 &\quad - \sum_{K \in \mathcal{C}_h} \sum_{F \subset \partial K} \int_F (\mathbf{u} - \tilde{\mathbf{u}}) \cdot (\mathbf{n} \times R_h(\operatorname{curl}(\mathbf{u} - \tilde{\mathbf{u}}))) = 0,
 \end{aligned}$$

since $\operatorname{curl} R_h(\operatorname{curl}(\mathbf{u} - \tilde{\mathbf{u}}))|_K \in (\mathcal{P}_0(K))^3$, and $\mathbf{n} \times R_h(\operatorname{curl}(\mathbf{u} - \tilde{\mathbf{u}}))|_F \in \mathbf{P}_F|_F$. \square

LEMMA 5.6. We have on $H(\operatorname{curl}; \Omega) \cap H(\operatorname{div}; \Omega)$

$$\mathcal{L}_h(\mathbf{u}, \mathbf{v}) \leq (\mathcal{L}_h(\mathbf{u}, \mathbf{u}))^{1/2} (\mathcal{L}_h(\mathbf{v}, \mathbf{v}))^{1/2}.$$

Proof. Both the symmetry and the coercivity properties of \mathcal{L}_h lead to the above generalized Cauchy–Schwarz inequality. \square

Setting

$$(5.27) \quad \|\mathbf{v}\|_{\mathcal{L}_h}^2 := \mathcal{L}_h(\mathbf{v}, \mathbf{v}),$$

we introduce an energy norm as follows:

$$\begin{aligned}
 (5.28) \quad \|\mathbf{v}\|_{0; \mathcal{L}_h}^2 &:= \|\mathbf{v}\|_0^2 + \|\mathbf{v}\|_{\mathcal{L}_h}^2 \\
 &= \|\mathbf{v}\|_0^2 + \|R_h(\operatorname{curl} \mathbf{v})\|_0^2 + s \left\| \check{R}_h(\operatorname{div} \mathbf{v}) \right\|_0^2 + \alpha \mathcal{S}_h(\mathbf{v}, \mathbf{v}).
 \end{aligned}$$

THEOREM 5.1. Let $\mathbf{u} \in U$ be the solution to problem (2.1)–(2.2) with the right-hand sides $\mathbf{f} \in H(\operatorname{div}^0; \Omega) \cap H(\operatorname{curl}; \Omega) \cap (H^r(\Omega))^3$ for some $r > 1/2$ and $g \in L^2(\Omega)$, and let $\mathbf{u}_h \in U_h$ be the solution to the finite element problem (3.20). Then

$$(5.29) \quad \|\mathbf{u} - \mathbf{u}_h\|_{0; \mathcal{L}_h} \leq C h^r (\|\mathbf{f}\|_{0; \operatorname{curl}} + \|\mathbf{f}\|_r + \|g\|_0).$$

Proof. Let $\tilde{\mathbf{u}} \in U_h$ be constructed as in Lemma 5.5. We have from Lemmas 5.1, 5.2, 5.4, and 5.6 that

$$\begin{aligned}
 \|\mathbf{u}_h - \tilde{\mathbf{u}}\|_{\mathcal{L}_h}^2 &= \mathcal{L}_h(\mathbf{u}_h - \tilde{\mathbf{u}}, \mathbf{u}_h - \tilde{\mathbf{u}}) \\
 &= \mathcal{L}_h(\mathbf{u} - \tilde{\mathbf{u}}, \mathbf{u}_h - \tilde{\mathbf{u}}) + \mathcal{L}_h(\mathbf{u}_h - \mathbf{u}, \mathbf{u}_h - \tilde{\mathbf{u}}) \\
 &\leq \|\mathbf{u} - \tilde{\mathbf{u}}\|_{\mathcal{L}_h} \|\mathbf{u}_h - \tilde{\mathbf{u}}\|_{\mathcal{L}_h} + C h^r (\|\mathbf{f}\|_r + \|\operatorname{curl} \mathbf{f}\|_0) \|\mathbf{u}_h - \tilde{\mathbf{u}}\|_{\mathcal{L}_h} \\
 &\leq C (\|\mathbf{u} - \tilde{\mathbf{u}}\|_{\mathcal{L}_h} + h^r (\|\mathbf{f}\|_r + \|\operatorname{curl} \mathbf{f}\|_0)) \|\mathbf{u}_h - \tilde{\mathbf{u}}\|_{\mathcal{L}_h},
 \end{aligned}$$

that is,

$$(5.30) \quad \|\mathbf{u}_h - \tilde{\mathbf{u}}\|_{\mathcal{L}_h} \leq C (\|\mathbf{u} - \tilde{\mathbf{u}}\|_{\mathcal{L}_h} + h^r (\|\mathbf{f}\|_r + \|\operatorname{curl} \mathbf{f}\|_0)),$$

where, from Lemma 5.5 and Lemma 4.2,

$$\begin{aligned} |||\mathbf{u} - \tilde{\mathbf{u}}|||_{\mathcal{L}_h}^2 &= \mathcal{L}_h(\mathbf{u} - \tilde{\mathbf{u}}, \mathbf{u} - \tilde{\mathbf{u}}) \\ &= \|R_h(\mathbf{curl}(\mathbf{u} - \tilde{\mathbf{u}}))\|_0^2 + s \left\| \check{R}_h(\operatorname{div}(\mathbf{u} - \tilde{\mathbf{u}})) \right\|_0^2 \\ &\quad + \alpha S_{h,\operatorname{div}}(\mathbf{u} - \tilde{\mathbf{u}}, \mathbf{u} - \tilde{\mathbf{u}}) + \alpha S_{h,\mathbf{curl}}(\mathbf{u} - \tilde{\mathbf{u}}, \mathbf{u} - \tilde{\mathbf{u}}) \\ &= \alpha S_{h,\operatorname{div}}(\mathbf{u} - \tilde{\mathbf{u}}, \mathbf{u} - \tilde{\mathbf{u}}) + \alpha S_{h,\mathbf{curl}}(\mathbf{u} - \tilde{\mathbf{u}}, \mathbf{u} - \tilde{\mathbf{u}}) \\ &\leq C \|\mathbf{u} - \tilde{\mathbf{u}}\|_0^2, \end{aligned}$$

that is,

$$(5.31) \quad |||\mathbf{u} - \tilde{\mathbf{u}}|||_{\mathcal{L}_h} \leq C \|\mathbf{u} - \tilde{\mathbf{u}}\|_0.$$

Therefore, we have from the L^2 coercivity in Theorem 4.1, (5.30), (5.31), and (5.15)

$$\begin{aligned} \|\mathbf{u} - \mathbf{u}_h\|_0 &\leq \|\mathbf{u} - \tilde{\mathbf{u}}\|_0 + \|\mathbf{u}_h - \tilde{\mathbf{u}}\|_0 \leq \|\mathbf{u} - \tilde{\mathbf{u}}\|_0 + C |||\mathbf{u}_h - \tilde{\mathbf{u}}|||_{\mathcal{L}_h} \\ (5.32) \quad &\leq \|\mathbf{u} - \tilde{\mathbf{u}}\|_0 + C (|||\mathbf{u} - \tilde{\mathbf{u}}|||_{\mathcal{L}_h} + h^r (\|\mathbf{f}\|_r + \|\mathbf{curl} \mathbf{f}\|_0)) \\ &\leq C \|\mathbf{u} - \tilde{\mathbf{u}}\|_0 + C h^r (\|\mathbf{f}\|_r + \|\mathbf{curl} \mathbf{f}\|_0) \\ &\leq C h^r (\|\mathbf{u}\|_r + \|\mathbf{f}\|_r + \|\mathbf{curl} \mathbf{f}\|_0), \end{aligned}$$

$$\begin{aligned} (5.33) \quad |||\mathbf{u} - \mathbf{u}_h|||_{\mathcal{L}_h} &\leq |||\mathbf{u} - \tilde{\mathbf{u}}|||_{\mathcal{L}_h} + |||\tilde{\mathbf{u}} - \mathbf{u}_h|||_{\mathcal{L}_h} \\ &\leq C \|\mathbf{u} - \tilde{\mathbf{u}}\|_0 + C h^r (\|\mathbf{f}\|_r + \|\mathbf{curl} \mathbf{f}\|_0) \\ &\leq C h^r (\|\mathbf{u}\|_r + \|\mathbf{f}\|_r + \|\mathbf{curl} \mathbf{f}\|_0), \end{aligned}$$

but from Lemma 5.3

$$(5.34) \quad \|\mathbf{u}\|_r \leq C (\|\mathbf{f}\|_0 + \|g\|_0),$$

we, therefore, add (5.32) and (5.33) to obtain (5.29). \square

Remark 5.2. For the finite element method (3.33), since we have no inconsistent errors, let $\tilde{\mathbf{u}}^* \in U_h^*$ be the interpolant to the solution \mathbf{u} of problem (2.1)–(2.2), we have

$$(5.35) \quad |||\mathbf{u}_h^* - \tilde{\mathbf{u}}^*|||_{\mathcal{L}_h^*} \leq C |||\mathbf{u} - \tilde{\mathbf{u}}^*|||_{\mathcal{L}_h^*},$$

following a similar argument as in proving Theorem 5.1, where

$$(5.36) \quad |||\mathbf{u} - \tilde{\mathbf{u}}^*|||_{\mathcal{L}_h^*}^2 = \|\mathbf{curl}(\mathbf{u} - \tilde{\mathbf{u}}^*)\|_0^2 + s \left\| \check{R}_h(\operatorname{div}(\mathbf{u} - \tilde{\mathbf{u}}^*)) \right\|_0^2 + \alpha S_{h,\operatorname{div}}(\mathbf{u} - \tilde{\mathbf{u}}^*, \mathbf{u} - \tilde{\mathbf{u}}^*).$$

We construct the interpolant $\tilde{\mathbf{u}}^* \in U_h^*$ to the solution \mathbf{u} in a bit different way from Lemma 5.5, but in a way similar to (5.7). So, we recall the regular-singular decomposition for the solution \mathbf{u} itself.

PROPOSITION 5.3 ([51, 29, 27, 26, 31]). *Let $\mathbf{u} \in U$ be the solution to problem (2.1)–(2.2), with the right-hand sides $\mathbf{f} \in \dot{H}(\operatorname{div}^0; \Omega)$ and $g \in L^2(\Omega)$. Then, \mathbf{u} can be written as the sum of a regular part and a singular part:*

$$(5.37) \quad \mathbf{u} = \mathbf{u}_H + \nabla \psi,$$

where

$$(5.38) \quad \mathbf{u}_H \in (H^{1+r}(\Omega))^3 \cap H_0(\mathbf{curl}; \Omega), \quad \psi \in H_0^1(\Omega) \cap H^{1+r}(\Omega)$$

for some $r > 1/2$, and

$$(5.39) \quad \|\mathbf{u}_H\|_{1+r} + \|\psi\|_{1+r} \leq C (\|\mathbf{f}\|_0 + \|g\|_0).$$

We define the interpolant $\tilde{\mathbf{u}}^* \in U_h^*$ to the solution \mathbf{u} as follows:

$$(5.40) \quad \tilde{\mathbf{u}}^* := \tilde{\mathbf{u}}_H + \nabla \tilde{\psi},$$

where $\tilde{\mathbf{u}}_H \in U_h^*$ is the interpolant to $\mathbf{u}_H \in (H^{1+r}(\Omega))^3$ with $r > 1/2$ and is constructed in a similar way as in Lemma 5.5 such that

$$(5.41) \quad \left\| \check{R}_h(\operatorname{div}(\mathbf{u}_H - \tilde{\mathbf{u}}_H)) \right\|_0 = 0,$$

$$(5.42) \quad \|\mathbf{u}_H - \tilde{\mathbf{u}}_H\|_0 + h \|\mathbf{u}_H - \tilde{\mathbf{u}}_H\|_1 \leq C h^{1+r} \|\mathbf{u}_H\|_{1+r},$$

while $\tilde{\psi}$ is the interpolant to $\psi \in H_0^1(\Omega) \cap H^{1+r}(\Omega)$ with $r > 1/2$ and is constructed in the *Argyris* C^1 triangle element [21] such that

$$(5.43) \quad \int_F \partial_n \tilde{\psi} = \int_F \partial_n \psi \quad \text{for all } F \in \partial K, \text{ for all } K \in \mathcal{C}_h,$$

$$(5.44) \quad \|\psi - \tilde{\psi}\|_1 \leq C h^r \|\psi\|_{1+r}.$$

From (5.43) we have

$$(5.45) \quad \int_K \operatorname{div} \nabla (\psi - \tilde{\psi}) = 0,$$

that is to say, we have

$$(5.46) \quad \left\| \check{R}_h(\operatorname{div} \nabla (\psi - \tilde{\psi})) \right\|_0 = 0.$$

The combination of (5.46) and (5.41) results in

$$(5.47) \quad \left\| \check{R}_h(\operatorname{div}(\mathbf{u} - \tilde{\mathbf{u}}^*)) \right\|_0 = 0.$$

We, therefore, have from the triangle-inequality, (5.35), (5.36), (5.42), (5.44), (5.47), and Lemma 4.2 that

$$(5.48) \quad \begin{aligned} \|\mathbf{u} - \mathbf{u}_h^*\|_{\mathcal{L}_h^*} &\leq \|\mathbf{u} - \tilde{\mathbf{u}}^*\|_{\mathcal{L}_h^*} + \|\mathbf{u}_h^* - \tilde{\mathbf{u}}^*\|_{\mathcal{L}_h^*} \\ &\leq C \|\mathbf{u} - \tilde{\mathbf{u}}^*\|_{\mathcal{L}_h^*} \\ &\leq C (\|\operatorname{curl}(\mathbf{u}_H - \tilde{\mathbf{u}}_H)\|_0 + \|\mathbf{u} - \tilde{\mathbf{u}}^*\|_0) \\ &\leq C h^r (\|\mathbf{u}_H\|_{1+r} + \|\psi\|_{1+r}), \end{aligned}$$

and from Remark 4.3, (5.35), (5.42), (5.44), and (5.48) that

$$(5.49) \quad \begin{aligned} \|\mathbf{u} - \mathbf{u}_h^*\|_0 &\leq \|\mathbf{u} - \tilde{\mathbf{u}}^*\|_0 + \|\tilde{\mathbf{u}}^* - \mathbf{u}_h^*\|_0 \\ &\leq C (\|\mathbf{u} - \tilde{\mathbf{u}}^*\|_0 + \|\tilde{\mathbf{u}}^* - \mathbf{u}_h^*\|_{\mathcal{L}_h^*}) \\ &\leq C (\|\mathbf{u} - \tilde{\mathbf{u}}^*\|_0 + \|\mathbf{u} - \tilde{\mathbf{u}}^*\|_{\mathcal{L}_h^*}) \\ &\leq C h^r (\|\mathbf{u}_H\|_{1+r} + \|\psi\|_{1+r}). \end{aligned}$$

Finally, from (5.48), (5.49), and (5.39) we have the following error estimate in the energy norm

$$(5.50) \quad \|\mathbf{u} - \mathbf{u}_h^*\|_{0;\mathcal{L}_h^*} \leq C h^r (\|\mathbf{f}\|_0 + \|g\|_0).$$

The above argument goes as well to the finite element method (3.37).

Remark 5.3. We see that (5.50) involves only the L^2 norm $\|\mathbf{f}\|_0$ of the right-hand side \mathbf{f} . So, when the approximate space contains the gradient of some C^1 element, the right-hand side \mathbf{f} can be less regular. In general, \mathbf{f} is required to be a little more regular (see (5.29)), since the regular-singular decomposition of the curl of the solution is used (see Proposition 5.2) in estimating the inconsistent error caused by the L^2 projected curl term.

6. Numerical experiments. In this section we shall report some numerical results which confirm the theoretical error bound, by considering a 3D source problem and a 2D eigenproblem.

A 3D source problem. Take the thick L-domain $\Omega = ([-1, 1]^2 \setminus ([0, 1] \times [-1, 0])) \times [0, 1] \subset \mathbb{R}^3$, and consider the Maxwell source problem: Find \mathbf{u} such that

$$\mathbf{curl} \mathbf{curl} \mathbf{u} = \mathbf{f}, \quad \operatorname{div} \mathbf{u} = g \quad \text{in } \Omega, \quad \mathbf{u} \times \mathbf{n} = 0 \quad \text{on } \Gamma = \partial\Omega,$$

where \mathbf{n} is the unit outer normal vector to Γ . We take the exact solution

$$\mathbf{u} = \eta(x, y, z) \nabla \left(\varrho^{\frac{2}{3}} \sin \left(\frac{2\theta}{3} \right) \right) = (u_1, u_2, u_3 = 0),$$

where $x = \varrho \cos(\theta)$, $y = \varrho \sin(\theta)$ and $z = z$, with ϱ being the distance to the reentrant edge along the z -axis starting from the origin $(0, 0, 0)$ of opening angle $3\pi/2$, and $\eta(x, y, z) = (1 - x^2)(1 - y^2)z(1 - z)$ is a cut-off function so that $\mathbf{u} \times \mathbf{n} = 0$ on Γ . The right-hand sides \mathbf{f} and g are obtained by evaluating the equations on the given exact solution.

We partition Ω into tetrahedra with uniform meshes. We employ the conjugate gradient method to solve the resulting symmetric and positive definite linear system, with the stopping tolerance 10^{-10} and with the null vector as an initial guess. In this numerical test we have two specific goals: (i) To verify the theoretical convergence rate, by computing the relative errors in L^2 norm using the exact solution $\mathbf{u} = (u_1, u_2, u_3)$ and the finite element solution $\mathbf{u}_h = (u_{1,h}, u_{2,h}, u_{3,h})$; (ii) To examine the effect of the stabilization parameter α , by considering several values of α as follows:

$$\alpha = 0.1, \quad 1, \quad 1000, \quad 10000.$$

In addition, we set the penalty/regularization parameter $s = 1$.

Since the regularity for the \mathbf{u} and its $\mathbf{curl} \mathbf{u}$ is $H^{\frac{2}{3}-\epsilon}$ for any $\epsilon \in (0, 1)$ (\mathbf{f} is also in $H^{\frac{2}{3}-\epsilon}$), from the theoretical convergence rate stated in Theorem 5.1 we expect that a mesh reduction of a factor of two (i.e., the mesh size decreases like $h = \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \dots$) should result in an error reduction of $2^{2/3} \approx 1.586$. This is clearly confirmed by the computed results listed in Tables 1–4. On the other hand, we observe that the stabilization parameter α *does not affect the error reduction ratio* (i.e., the ratios in Tables 1–4 are almost the same), although it affects the sizes of errors in the way that larger values of α yield smaller values of errors. This may be due to the fact that suitable larger α would enhance the stability (cf. (4.27)–(4.28)) and, thus, make the constant in front of the error bound (5.29) smaller. We also observe that the values of errors are the same for both u_1 and u_2 . This is because u_1 and u_2 are symmetric with respect to the $O - xyz$ coordinates system.

¹In Tables 1–4 the 3rd row is the L^2 -norm values of $u_{3,h}$ for different mesh sizes, since $u_3 = 0$.

TABLE 1
Relative errors in L^2 norm with $\alpha = 0.1$.

	$h = \frac{1}{4}$	$h = \frac{1}{8}$	$h = \frac{1}{16}$
$\frac{\ u_1 - u_{1,h}\ _0}{\ u_1\ _0} = \frac{\ u_2 - u_{2,h}\ _0}{\ u_2\ _0}$	5092.13	3177.736	1975.3958
$^1 \frac{\ u_3 - u_{3,h}\ _0}{\ u_3\ _0}$	471.240	303.274	190.983

TABLE 2
Relative errors in L^2 norm with $\alpha = 1.0$.

	$h = \frac{1}{4}$	$h = \frac{1}{8}$	$h = \frac{1}{16}$
$\frac{\ u_1 - u_{1,h}\ _0}{\ u_1\ _0} = \frac{\ u_2 - u_{2,h}\ _0}{\ u_2\ _0}$	509.238	317.792	197.552
$\frac{\ u_3 - u_{3,h}\ _0}{\ u_3\ _0}$	47.1187	30.3241	19.0963

TABLE 3
Relative errors in L^2 norm with $\alpha = 1000.0$.

	$h = \frac{1}{4}$	$h = \frac{1}{8}$	$h = \frac{1}{16}$
$\frac{\ u_1 - u_{1,h}\ _0}{\ u_1\ _0} = \frac{\ u_2 - u_{2,h}\ _0}{\ u_2\ _0}$	0.622315	0.400576	0.254468
$\frac{\ u_3 - u_{3,h}\ _0}{\ u_3\ _0}$	0.050283	0.033089	0.021106

TABLE 4
Relative errors in L^2 norm with $\alpha = 10000.0$.

	$h = \frac{1}{4}$	$h = \frac{1}{8}$	$h = \frac{1}{16}$
$\frac{\ u_1 - u_{1,h}\ _0}{\ u_1\ _0} = \frac{\ u_2 - u_{2,h}\ _0}{\ u_2\ _0}$	0.292675	0.202238	0.153258
$\frac{\ u_3 - u_{3,h}\ _0}{\ u_3\ _0}$	0.016049	0.0139586	0.010769

A 2D eigenproblem. As an illustration of the application of the L^2 projection method to Maxwell eigenproblem, we perform the numerical test for a 2D eigenproblem in the L-domain $\Omega = [-1, 1]^2 \setminus ([0, 1] \times [-1, 0]) \subset \mathbb{R}^2$: Find eigenvalues ω^2 and eigenfunctions \mathbf{u} such that

$$\mathbf{curl} \mathbf{curl} \mathbf{u} = \omega^2 \mathbf{u}, \quad \mathbf{div} \mathbf{u} = 0 \quad \text{in } \Omega, \quad \mathbf{u} \cdot \boldsymbol{\tau} = 0 \quad \text{on } \Gamma = \partial \Omega,$$

where $\boldsymbol{\tau}$ is the unit tangential vector along Γ .

We partition Ω into triangles with uniform meshes. As mentioned in Remark 3.3, the approximate space is of \mathcal{P}_3 element. We can set the penalty/regularization parameter s as any positive constant, say $s = 2$. Following the computational results in Table 4 for the source problem, we take the stabilization parameter α as $\alpha = 10000$.

We consider the benchmark example for the L-domain from the website at

<http://www.maths.univ-rennes1.fr/dauge/benchmax.html>,

and take the first two computed eigenvalues therein as true solutions, i.e.,

$$\omega_1^2 = 1.47562182408, \quad \omega_2^2 = 3.53403136678.$$

Note that the first eigenfunction has a strong singularity and is in $H^{\frac{2}{3}-\epsilon}$, and the second eigenfunction is smooth and belongs to $H^{\frac{4}{3}-\epsilon}$ for all $\epsilon > 0$ (see [28]). We would like to verify the error estimates in the case of eigenproblem: with the application of the result of [9], we can conclude from Theorem 5.1 that the following theoretical

TABLE 5
Relative errors and error reduction ratios of the first eigenvalue.

	$h = \frac{1}{4}$	$h = \frac{1}{8}$	$h = \frac{1}{16}$	$h = \frac{1}{32}$	$h = \frac{1}{64}$	$h = \frac{1}{128}$
$\frac{ \omega_1^2 - \omega_{1,h}^2 }{ \omega_1^2 }$	0.79882e0	0.48321e0	0.23809e0	0.10345e0	0.42512e - 1	0.17092e - 1
Ratio	—	1.65315	2.02953	2.30150	2.43343	2.48725

TABLE 6
Relative errors and error reduction ratios of the second eigenvalue.

	$h = \frac{1}{4}$	$h = \frac{1}{8}$	$h = \frac{1}{16}$	$h = \frac{1}{32}$	$h = \frac{1}{64}$	$h = \frac{1}{128}$
$\frac{ \omega_2^2 - \omega_{2,h}^2 }{ \omega_2^2 }$	0.39675e - 1	0.94427e - 2	0.21858e - 2	0.51238e - 3	0.12298e - 3	0.30034e - 4
Ratio	—	4.20166	4.32002	4.26597	4.16637	4.09469

convergence rate

$$|\omega_1^2 - \omega_{1,h}^2| \leq C h^{2r} \quad \text{with } r = \frac{2}{3} - \epsilon$$

holds for the first eigenvalue corresponding to eigenfunction in H^r . Thus, the error reduction ratio of the first eigenvalue should be about $2^{\frac{4}{3}} \approx 2.519$, with a mesh reduction of factor two. Regarding the second eigenvalue corresponding to a smooth eigenfunction in $H^{\frac{4}{3}-\epsilon}$, for the approximation of the \mathcal{P}_3 element, an error reduction ratio would be about $2^{\frac{8}{3}} \approx 6.349$ with a mesh reduction of factor two. But, due to the inconsistent errors caused by both the L^2 projected *curl* term and the mesh-dependent term $S_{h,\text{curl}}$, the error reduction ratio is 4 only; i.e., the theoretical convergence rate from Theorem 5.1 for the second eigenvalue is

$$|\omega_2^2 - \omega_{2,h}^2| \leq C h^2.$$

From the computed error reduction ratios of eigenvalues listed in Tables 5 and 6 we see that the computational ratios are very close to the ones as predicted above.

7. Conclusions. We have proposed the element-local L^2 projected C^0 finite element method for solving the Maxwell problem with the nonsmooth solution being not in H^1 . The key feature is that some element-local L^2 projectors are applied to both the curl and div operators in the well-known plain regularization variational formulation. The Maxwell problem under consideration is posed in a simply connected polyhedron with a connected Lipschitz continuous boundary and has a solution that may be in H^r with $r < 1$. We have established the coercivity and the condition number $\mathcal{O}(h^{-2})$ of the resulting linear system. We have also obtained the desired error bounds $\mathcal{O}(h^r)$ in an energy norm for the C^0 linear element (enriched by certain higher degree face- and element-bubble functions), when the solution and its curl are in H^r ($1/2 < r < 1$) with a smooth right-hand side. Performed for a 3D source problem and a 2D eigenproblem, both of which are posed on nonsmooth domains with reentrant corners and/or edges and have nonsmooth solutions being not in H^1 , the numerical experiments have produced good and correct C^0 approximations of nonsmooth solutions and confirmed the theoretical convergence rate obtained.

For this L^2 projection method, we do not require that the C^0 approximate space contain the gradient of some C^1 element and we do not impose the information of the geometric singularities of the domain boundary in the finite element variational

formulation. These make the L^2 projection method particularly attractive for Maxwell equations posed on more complex 3D domains.

In addition, for 2D Maxwell problem we proposed two more L^2 projection methods (only the divergence part involves the element-local L^2 projector), where the C^0 approximate space contains the gradient of the *Argyris* C^1 triangle element and the *Hsieh-Clough-Tocher* C^1 macro-triangle element, respectively. Coercivity is established and error estimates for nonsmooth solution being not in H^1 are obtained. These last two methods are consistent and allow less regular right-hand sides. For 3D Maxwell problem similar methods can be developed in the same routine.

A generalization of the L^2 projection method to Maxwell interface problems with discontinuous inhomogeneous anisotropic materials in a multiply connected nonsmooth domain (existing reentrant corners and edges) and with mixed boundary conditions is currently being studied and will be reported elsewhere.

Acknowledgments. The authors would like to thank the anonymous referees for their valuable comments and suggestions on the presentation of this paper.

REFERENCES

- [1] P. ALFELD, *A trivariate Clough-Tocher scheme for tetrahedral data*, Comput. Aided Geom. Design, 1 (1984), pp. 169–181.
- [2] A. ALONSO AND A. VALLI, *Some remarks on the characterization of the space of tangential traces of $H(\text{rot}; \Omega)$ and the construction of an extension operator*, Manuscripta Math., 89 (1996), pp. 159–178.
- [3] A. ALONSO RODRÍGUEZ, P. FERNANDERS, AND A. VALLI, *Weak and strong formulations for the time-harmonic eddy-current problem in general multi-connected domains*, European J. Appl. Math., 14 (2003), pp. 387–406.
- [4] C. AMROUCHE, C. BERNARDI, M. DAUGE, AND V. GIRAULT, *Vector potentials in three-dimensional non-smooth domains*, Math. Methods Appl. Sci., 21 (1998), pp. 823–864.
- [5] F. ASSOUS, P. CIARLET, JR., AND E. SONNENDRÜCKER, *Resolution of the Maxwell equations in a domain with reentrant corners*, M2AN Math. Model. Numer. Anal., 32 (1998), pp. 359–389.
- [6] F. ASSOUS, P. CIARLET, JR., P.-A. RAVAIPT, AND E. SONNENDRÜCKER, *Characterization of the singular part of the solution of Maxwell's equations in a polyhedral domain*, Math. Methods Appl. Sci., 22 (1999), pp. 485–499.
- [7] F. BEN BELGACEM AND C. BERNARDI, *Spectral element discretization of the Maxwell equations*, Math. Comp., 68 (1999), pp. 1497–1520.
- [8] A. BERMÚDEZ, R. RODRÍGUEZ, AND P. SALGADO, *A finite element method with Lagrangian multiplier for low-frequency harmonic Maxwell equations*, SIAM J. Numer. Anal., 40 (2002), pp. 1823–1849.
- [9] I. BABUŠKA, AND J. E. OSBORN, *Finite element-Galerkin approximation of the eigenvalues and eigenvectors of selfadjoint problems*, Math. Comp., 52 (1989), pp. 275–297.
- [10] C. BERNARDI, *Optimal finite element interpolation on curved domains*, SIAM J. Numer. Anal., 26 (1989), pp. 1212–1240.
- [11] C. BERNARDI AND V. GIRAULT, *A local regularization operator for triangular and quadrilateral finite elements*, SIAM J. Numer. Anal., 35 (1998), pp. 1893–1916.
- [12] M. BIRMAN AND M. SOLOMYAK, *L^2 -theory of the Maxwell operator in arbitrary domains*, Russian Math. Surveys, 42 (1987), pp. 75–96.
- [13] A.-S. BONNET-BEN DHIA, C. HAZARD, AND S. LOHRENGEL, *A singular field method for the solution of Maxwell's equations in polyhedral domains*, SIAM J. Appl. Math., 59 (1999), pp. 2028–2044.
- [14] A. BOSSAVIT, *Magnetostatic problems in multiply connected regions: Some properties of the curl operator*, IEEE Proc., 135 (1988), pp. 179–187.
- [15] A. BOSSAVIT, *Computational Electromagnetism: Variational Formulations, Complementarity, Edge Elements*, Academic Press, New York, 1998.
- [16] J. BRANDTS, S. KOROTOV, AND M. KRÍŽEK, *On the equivalence of regularity criteria for triangular and tetrahedral partitions*, Comput. Math. Appl., 55 (2008), pp. 2227–2233.

- [17] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, Berlin, 1996.
- [18] S. CAORSI, P. FERNANDES, AND M. RAFFETTO, *Spurious-free approximations of electromagnetic eigenproblems by means of Nédléc-type elements*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 331–354.
- [19] C. CARSTENSEN, S. FUNKEN, W. HACKBUSCH, R. H. W. HOPPE, AND P. MONK, *Computational Electromagnetics, Proceedings of the GAMM Workshop on Computational Electromagnetics*, Springer-Verlag, Berlin, 2003.
- [20] M. CESSENAT, *Mathematical Methods in Electromagnetism: Linear Theory and Applications*, World Scientific, 1996.
- [21] P. G. CIARLET, *Basic Error Estimates for Elliptic Problems*, in: Handbook of Numerical Analysis, Vol. II, Finite Element Methods (part 1), P. G. Ciarlet and J.-L. Lions, eds., North-Holland, Amsterdam, 1991.
- [22] P. CLÉMENT, *Approximation by finite element functions using local regularization*, RAIRO Numer. Anal., 9 (1975), pp. 77–84.
- [23] M. COSTABEL, *A remark on the regularity of solutions of Maxwell's equations on Lipschitz domains*, M3AS Math. Methods Appl. Sci., 12 (1990), pp. 365–368.
- [24] M. COSTABEL, *A coercive bilinear form for Maxwell's equations*, J. Math. Anal. Appl., 157 (1991), pp. 527–541.
- [25] M. COSTABEL AND M. DAUGE, *Weighted regularization of Maxwell equations in polyhedral domains*, Numer. Math., 93 (2002), pp. 239–277.
- [26] M. COSTABEL AND M. DAUGE, *Singularities of electromagnetic fields in polyhedral domains*, Arch. Rational Mech. Anal., 151 (2000), pp. 221–276.
- [27] M. COSTABEL AND M. DAUGE, *Maxwell and Lamé eigenvalues on polyhedra*, Math. Methods Appl. Sci., 22 (1999), pp. 243–258.
- [28] M. COSTABEL AND M. DAUGE, *Computation of resonance frequencies for Maxwell equations in non smooth domains*, in Lecture Notes in Comput. Sci. Eng. 31, M. Ainsworth, P. Davies, D. Duncan, P. Martin, and B. Rynne, eds., 2003, pp. 125–162.
- [29] M. COSTABEL, M. DAUGE, AND S. NICAISE, *Singularities of Maxwell interface problem*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 627–649.
- [30] M. CROUZEIX AND P.-A. RAVIART, *Conforming and nonconforming finite element methods for solving the stationary Stokes equations*, RAIRO Numer. Anal., 7 (1973), pp. 33–75.
- [31] M. DAUGE, *Private communication*, 2005.
- [32] H.-Y. DUAN, P. LIN, P. SAIKRISHNAN, AND R. C. E. TAN, *L^2 -projected least-squares finite element methods for the Stokes equations*, SIAM J. Numer. Anal., 44 (2006), pp. 732–752.
- [33] G. FARIN, *Triangular Bernstein-Bézier patches*, Comput. Aided Geom. Design, 3 (1986), pp. 83–127.
- [34] P. FERNANDES AND G. GILARDI, *Magnetostatic and Electrostatic problems in inhomogeneous anisotropic media with irregular boundary and mixed boundary conditions*, Math. Models Methods Appl. Sci., 7 (1997), pp. 957–991.
- [35] P. FERNANDES AND I. PERUGIA, *Vector potential formulation for magnetostatics and modelling of permanent magnets*, IMA J. Appl. Math., 66 (2001), pp. 293–318.
- [36] V. GIRAULT, *A local projection operator for quadrilateral finite elements*, Math. Comp., 64(1995), pp. 1421–1431.
- [37] V. GIRAULT AND P. A. RAVIART, *Finite Element Methods for Navier-Stokes Equations, Theory and Algorithms*, Springer-Verlag, Berlin, 1986.
- [38] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computation*, 3rd edition, Johns Hopkins University Press, Baltimore, MD, 1996.
- [39] C. HAZARD AND M. LENOIR, *On the solution of time-harmonic scattering problems for Maxwell's equations*, SIAM J. Math. Anal., 27 (1996), pp. 1597–1630.
- [40] C. HAZARD AND S. LOHRENGEL, *A singular field method for Maxwell's equations: Numerical aspects for 2D magnetostatics*, SIAM J. Numer. Anal., 40 (2003), pp. 1021–1040.
- [41] R. HIPTMAIR, *Finite elements in computational electromagnetism*, Acta Numer., 2002, pp. 237–339.
- [42] P. HOUSTON, I. PERUGIA, A. SCHNEEBELI, AND D. SCHÖTZAU, *Interior penalty method for indefinite time-harmonic Maxwell equations*, Numer. Math., 100 (2005), pp. 485–518.
- [43] J. M. JIN, *The Finite Element Method in Electromagnetics* (2nd Edition), John Wiley & Sons, New York, 2002.
- [44] F. KIKUCHI, *Mixed and penalty formulations for finite element analysis of an eigenvalue problem in electromagnetism*, Comput. Methods Appl. Mech. Engrg., 64 (1987), pp. 509–521.
- [45] M.-J. LAI AND A. LEMÉHAUTÉ, *A new kind of trivariate C^1 macro-element*, Adv. Comput. Math., 21 (2004), pp. 273–292.

- [46] E. J. LEE AND T. A. MANTEUFFEL, *FOSLL* method for the eddy current problem with three-dimensional edge singularities*, SIAM J. Numer. Anal., 45 (2007), pp. 787–809.
- [47] G. MEURANT, *Computer Solution of Large Linear Systems*, Elsevier, Singapore, 1999.
- [48] P. MONK, *A finite element method for approximating the time-harmonic Maxwell's equations*, Numer. Math., 63 (1992), pp. 243–261.
- [49] P. MONK, *Analysis of a finite element method for Maxwell's equations*, SIAM J. Numer. Anal., 29 (1992), pp. 714–729.
- [50] P. MONK, *Finite Element Methods for Maxwell Equations*, Clarendon Press, Oxford, 2003.
- [51] S. NICAISE, *Edge elements on anisotropic meshes and approximation of the Maxwell equations*, SIAM J. Numer. Anal., 39 (2001), pp. 784–816.
- [52] L. R. SCOTT AND S. ZHANG, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comput., 54 (1990), pp. 483–493.
- [53] O. STEINBACH, *On the stability of the L_2 projection in fractional Sobolev spaces*, Numer. Math., 88 (2000), pp. 367–379.
- [54] A. J. WORSEY AND G. FARIN, *An n -dimensional Clough-Tocher interpolant*, Constr. Approx., 3 (1987), pp. 99–110.
- [55] A. J. WORSEY AND B. PIPER, *A trivariate Powell-Sabin interpolant*, Comput. Aided Geom. Design, 5 (1988), pp. 177–186.

ON THE EXISTENCE OF EXPLICIT hp -FINITE ELEMENT METHODS USING GAUSS–LOBATTO INTEGRATION ON THE TRIANGLE*

B. T. HELENBROOK[†]

Abstract. Spectral-element simulations on quadrilaterals and hexahedra rely on the Gauss–Lobatto (GL) integration rule to enable explicit simulations with optimal spatial convergence rates. In this work, it is proved that a similar integration rule does not exist on triangles. The following properties of the rule are sought: a $(p+1)(p+2)/2$ point integration rule capable of exactly integrating the space given by $\mathcal{T}(2p-1) \equiv \{x^m y^n | 0 \leq m, n; m+n \leq 2p-1\}$, where p is an integer; integration points located at each of the triangle vertices; $p-1$ integration points located on each side; and $(p-1)(p-2)/2$ integration points located in the interior of the element. The proof hinges on the fact that the existence of such a rule implies the existence of a nodal basis with an approximate diagonal mass matrix that can be inverted to obtain exact Galerkin projections of functions in $\mathcal{T}(p-1)$. The proof shows that vertex functions of a basis having this property exist and are unique, but on a triangle these functions are not nodal, and therefore the GL rule does not exist. In spite of this, the existence of the vertex functions indicates that there may be a nonnodal basis that has the above property. This basis would enable explicit hp -finite element simulations on the triangle with optimal spatial accuracy. The methodology developed in the paper gives insight into a possible way to find such a basis.

Key words. triangles, quadrature, integration, Gauss, Lobatto, mass-lumping

AMS subject classifications. 65D32, 74S05

DOI. 10.1137/070685439

1. Introduction. Gauss–Lobatto (GL) integration [1, p. 888] provides the foundation for spectral element simulations [20]. Not only does it provide a numerical integration method, but the integration points also define a nodal basis that allows easy enforcement of continuity constraints at element boundaries and gives an approximately diagonal mass matrix. This last point enables unsteady simulations that do not require inversion of a globally coupled mass matrix and yet still obtain optimal spatial convergence rates [20]. These properties are the main reason that spectral element simulations can efficiently achieve a high order of accuracy.

Although GL integration rules can be defined for segments, quadrilaterals, and hexahedra [16, p. 143], an equivalent integration rule has not been found for triangles. This is not due to a lack of effort in searching. Much effort has been made to find optimal interpolation points on the triangle [18, 2, 3, 27, 15] and also to find a quadrature formula [28, 29, 14, 4, 5, 25, 17]. Cools and coworkers provide an excellent summary of the current status of quadrature rules on triangles as well as other geometries [9, 8, 10, 7, 6, 19, 11]. Because no completely satisfactory integration rule has been found, researchers are still experimenting with different techniques for performing high-order continuous finite element simulations on triangles [23, 24, 12, 30, 21].

In this work, it is proved that there is no GL integration rule for a triangle that has properties similar to those for segments, quadrilaterals, and hexahedra.

*Received by the editors March 16, 2007; accepted for publication (in revised form) October 31, 2008; published electronically February 25, 2009. This material is based upon work supported by the National Science Foundation under grant 0513380.

<http://www.siam.org/journals/sinum/47-2/68543.html>

[†]Mechanical & Aeronautical Engineering Department, Clarkson University, 8 Clarkson Avenue, Potsdam, NY 13699-5725 (helenbrk@clarkson.edu).

On quadrilaterals, the tensor-product GL integration rule for the space $\mathcal{Q}(p) \equiv \{x^m y^n \mid 0 \leq m, n; m \leq p; n \leq p\}$ has the following properties:

- a $\dim(\mathcal{Q}(p)) = (p + 1)^2$ point integration rule capable of exactly integrating the polynomial space $\mathcal{Q}(2p - 1)$;
- integration points located at each of the quadrilateral vertices;
- $p - 1$ integration points located on each quadrilateral side;
- $(p - 1)^2$ integration points located in the interior of the element.

On triangles, the function space typically used is $\mathcal{T}(p) \equiv \{x^m y^n \mid 0 \leq m, n; m + n \leq p\}$ [27]. For this space an integration rule is sought with the following properties:

- a $\dim(\mathcal{T}(p)) = (p + 1)(p + 2)/2$ point integration rule capable of exactly integrating the polynomial space $\mathcal{T}(2p - 1)$;
- integration points located at each of the triangle vertices;
- $p - 1$ integration points located on each side;
- $(p - 1)(p - 2)/2$ integration points located in the interior of the element.

Theoretical results for polynomial integration formulas on a triangle give a lower bound for the number of points required to exactly integrate the space $\mathcal{T}(2p - 1)$ of $p(p + 1)/2 + \lfloor p/2 \rfloor$, where the floor symbols $\lfloor \cdot \rfloor$ denote truncation [8]. The rule sought has more points than the lower bound for all p , but with special constraints on the positions. Note that in both the quadrilateral case and the triangle case the problem is overdetermined. On quadrilaterals, there are $4 + 2 \times 4(p - 1) + 3 \times (p - 1)^2 = 3p^2 + 2p - 1$ degrees of freedom for the positions and weights, and there are $4p^2$ accuracy constraints. However, this solution exists. On triangles, there are $3 + 2 \times 3(p - 1) + 3 \times (p - 1)(p - 2)/2 = 3(p^2 + p)/2$ degrees of freedom for the positions and weights and $2p^2 + p$ accuracy constraints.

The basic steps of the proof are given in one dimension as a demonstration and then subsequently applied to triangles. A positive result of the proof is that vertex modes are found that allow “diagonal projection.” This is defined to mean that a diagonal mass matrix can be inverted to obtain exact Galerkin projections of functions in $\mathcal{T}(p - 1)$. A full basis that allows diagonal projection will enable explicit-unsteady, continuous finite element simulations on the triangle with optimal spatial accuracy.

2. One-dimensional integration. The first part of the proof is to establish some basic features of the GL integration rule on the domain $x \in [-1, 1]$. It is of course well known that the GL integration rule exists on this domain, but nonetheless it is instructive to go through the process in one dimension before applying it to triangles. The GL integration rule in one dimension is defined by

$$(2.1) \quad \int_{-1}^1 f(x) dx \approx \sum_{i=1}^n w_i f(x_i),$$

where $f(x)$ is the function to be integrated, n is the number of points in the GL rule, w_i is the integration weight associated with each integration point, and x_i is the location of the integration point. The first and last integration points are constrained to be at the edge of the domain, $x_1 = -1$ and $x_n = 1$. The GL integration rule has the following properties:

- an n -point formula integrates polynomials of order $2n - 3$;
- the locations of the integration points are the roots of the derivative of the $(n - 1)$ st Legendre polynomials, $P'_{n-1}(x)$;

- the weights are given by

$$(2.2) \quad w_i = \frac{2}{n(n-1)[P_{n-1}(x_i)]^2}.$$

A $(p+1)$ -point GL integration rule can be used to generate an order p nodal polynomial basis. This basis is defined by

$$\phi_i(x) = \prod_{j=1, j \neq i}^{p+1} \frac{x - x_j}{x_i - x_j}, \quad i \in [1, p+1],$$

where ϕ_i is the i th function of the basis vector. $\phi_i(x)$ is zero at all of the GL integration points except the i th point where it has the value 1, i.e., $\phi_i(x_j) = \delta_{i,j}$, where $\delta_{i,j}$ is the Kronecker delta function. The basis ϕ is referred to as the Gauss–Lobatto–Lagrange (GLL) basis. It spans $\mathcal{P}(p)$, which is the space of polynomials of degree p .

The standard method of projecting a function onto this basis is defined as

$$(2.3) \quad \int_{\Omega} \phi \phi^T \vec{u} d\Omega = \int_{\Omega} \phi f(x) d\Omega,$$

where Ω is the domain $[-1, 1]$. This equation determines the coefficient vector, \vec{u} such that $\phi^T \vec{u}$ approximates $f(x)$. The matrix $\int_{\Omega} \phi \phi^T d\Omega$ is typically called the mass matrix, M . M is diagonal if the basis functions are orthogonal (Legendre polynomials). The above equation gives an exact representation of $f(x)$ if $f(x)$ is contained in the space spanned by ϕ .

The combination of the $(p+1)$ -point GL integration rule and the order p nodal basis leads to an approximate orthogonality property. If

$$\int_{\Omega} \phi_j \phi_k d\Omega$$

is approximated as

$$\sum_{i=1}^{p+1} w_i \phi_j(x_i) \phi_k(x_i),$$

this becomes

$$\sum_{i=1}^{p+1} w_i \delta_{j,i} \delta_{k,i} = \delta_{j,k} w_j.$$

This shows that the basis is orthogonal when integrated with the GL integration rule. Because the GL integration is accurate only for polynomials of order $2p-1$, and the integrand is of order $2p$, this is not equivalent to showing that the basis itself is orthogonal.

THEOREM 2.1. *The approximate orthogonality property of the GLL basis guarantees the existence of a diagonal projection operation that gives an exact representation of functions in $\mathcal{P}(p-1)$. The diagonal projection operation is defined as*

$$(2.4) \quad D\vec{u} = \int_{\Omega} \phi f(x) d\Omega,$$

where D is a diagonal matrix.

Proof. Let an entry of the matrix D be defined by

$$(2.5) \quad d_{j,k} = \sum_{i=1}^{p+1} w_i \phi_j(x_i) \phi_k(x_i) = \delta_{j,k} w_j.$$

Because of the approximate orthogonality property, D is diagonal. Furthermore, all of the weights of the GL integration rule are nonzero, so D is invertible. Thus there is a unique solution to (2.4). It remains to show that the exact solution satisfies (2.4) when $f(x)$ is a polynomial of order $p - 1$. If $f(x)$ is a polynomial of order $p - 1$ and the inversion is exact, then $\phi^T \vec{u}$ is also a polynomial of order $p - 1$. Furthermore, (2.4), with $d_{j,k}$ defined as in (2.5), is an approximation to (2.3). When $\phi^T \vec{u}$ is of order $p - 1$, the integrand on the left-hand side of (2.3) is of order $2p - 1$. Because the GL integration rule is exact for polynomials of order $2p - 1$, (2.4) and (2.5) are exact approximations to (2.3). Since the exact solution satisfies (2.3), it must also satisfy (2.4). \square

The above shows that a GL integration rule guarantees the existence of a nodal basis that allows exact “diagonal projection” for functions of degree $p - 1$. Next, it is shown that this basis can be derived based on accuracy considerations. First, the nodal basis is divided into interior modes and vertex modes. The interior modes are zero at element boundaries and can be constructed from the space

$$\mathcal{I}(p) \equiv \frac{1 - x^2}{4} \mathcal{P}(p - 2).$$

This is the space of all polynomials of degree $\leq p$ that are zero at both -1 and 1 .

Because the GL integration rule must have an integration point at -1 and 1 , the nodal basis will always have a left and right vertex mode. The left vertex mode can be defined as a polynomial that is 1 at $x = -1$ and 0 at $x = 1$. Polynomials of degree p that satisfy these constraints can be constructed as

$$\frac{1 - x}{2} + i(x) \text{ with } i(x) \in \mathcal{I}(p).$$

The function $\frac{1-x}{2}$ and all of the interior modes have a root at $x = 1$, and therefore the left vertex mode will always have a root at $x = 1$. Similar results hold for the right vertex mode.

THEOREM 2.2. *There is one and only one left vertex mode, ϕ_1 , that allows exact diagonal projection of polynomials of order $p - 1$.*

Proof. Let the function to be projected, $f(x)$, be described as

$$(2.6) \quad f(x) = a_1 \frac{1 - x}{2} + \frac{1 + x}{2} \sum_{i=2}^p a_i x^{i-2},$$

and let the left vertex function of the basis vector, ϕ_1 , be described as

$$(2.7) \quad \phi_1(x) = \frac{1 - x}{2} + \frac{1 - x^2}{4} \sum_{i=1}^{p-1} b_i x^{i-1}.$$

Let the projection be represented as $\phi^T(x) \vec{u}$. The first component of (2.4) is given by

$$d_{1,1} u_1 = \int_{-1}^1 \phi_1(x) f(x) dx,$$

which is equivalent to

$$d_{1,1}u_1 = \int_{-1}^1 \left[\frac{1-x}{2} + \frac{1-x^2}{4} \sum_{i=1}^{p-1} b_i x^{i-1} \right] \left[a_1 \frac{1-x}{2} + \frac{1+x}{2} \sum_{i=2}^p a_i x^{i-2} \right] dx.$$

This equation must be true for all \vec{a} . Equating $\phi^T(-1)\vec{u}$ to $f(-1)$ and using the fact that the left vertex function is the only nonzero basis function at $x = -1$ gives $u_1 = a_1$. a_1 then gives

$$d_{1,1} = \int_{-1}^1 \left[\frac{1-x}{2} + \frac{1-x^2}{4} \sum_{i=1}^{p-1} b_i x^{i-1} \right] \frac{1-x}{2} dx,$$

which determines $d_{1,1}$. Each of the remaining a 's give a row of the equations

$$(2.8) \int_{-1}^1 \begin{bmatrix} 1 \\ x \\ \vdots \\ x^{p-2} \end{bmatrix} \frac{1+x}{2} \frac{1-x^2}{4} [1, x, \dots, x^{p-2}] \vec{b} dx = - \int_{-1}^1 \frac{1-x^2}{4} \begin{bmatrix} 1 \\ x \\ \vdots \\ x^{p-2} \end{bmatrix} dx.$$

The above equations are a system of $p - 1$ equations in the $p - 1$ unknowns of \vec{b} . It has a unique solution if the matrix on the left-hand side has a nonzero determinant. This matrix is symmetric because any entry can be represented as

$$c_{i,j} = \int_{-1}^1 \left(\frac{1+x}{2} \right) \left(\frac{1-x^2}{4} \right) x^{i-1} x^{j-1} dx.$$

It is also positive definite because

$$\begin{aligned} \vec{b}^T \int_{-1}^1 \left(\frac{1+x}{2} \right) \left(\frac{1-x^2}{4} \right) \begin{bmatrix} 1 \\ x \\ \vdots \\ x^{p-2} \end{bmatrix} [1, x, \dots, x^{p-2}] dx \vec{b} \\ = \int_{-1}^1 \left(\frac{1+x}{2} \right) \left(\frac{1-x^2}{4} \right) b(x)^2 dx, \end{aligned}$$

where $b(x) = [1, x, \dots, x^{p-2}] \vec{b}$. The integrand is always positive over the domain $[-1, 1]$. Because it is symmetric and positive definite, it is invertible, which proves that the left vertex mode is unique. Similar results hold for the right vertex mode. \square

The next theorem is similar to a more general theorem given by Mysovskikh [22] for a multidimensional Gauss integration rule that states that “a necessary condition for the existence of a quadrature formula of degree $2k + 1$ with $N = \dim \mathbf{P}_k^d$ points is that the basic orthogonal polynomials of degree $k + 1$ have N common zeros” where \mathbf{P}_k^d is the space of polynomials in dimension d with total degree less than k . (See [8, Theorem 2].) The following theorem is more useful for analyzing the existence of GL integration rules.

THEOREM 2.3. *If the left and right vertex modes satisfying the diagonal projection property do not have $p-1$ roots at coincident locations in $(-1, 1)$, then a GL integration rule does not exist.*

Proof. Assume

1. a GL integration rule exists, and
2. a left and right vertex function exists satisfying the diagonal projection property but with roots at different locations in $(-1, 1)$.

By assumption 1 and Theorem 2.1, there exists a left and right vertex function satisfying the diagonal projection property. Furthermore, these functions are from a nodal basis and thus share the same roots in $(-1, 1)$. Because the left and right vertex functions are unique by Theorem 2.2, this contradicts item 2 above. Thus assumption 2 excludes the existence of the GL integration rule. \square

To verify whether a GL integration rule can exist or not, the location of the roots of the left (and right) vertex mode must be found using (2.8). By relaxing the form specified for the left vertex mode, one can obtain an explicit expression, which then makes it easy to determine the location of the roots. Instead of assuming the form given by (2.7), the following form is used:

$$(2.9) \quad \phi_1(x) = \frac{1-x}{2} \widehat{\phi}_1(x),$$

where $\widehat{\phi}_1(x) \in \mathcal{P}(p-1)$. This enforces the constraint that the left vertex mode have a root at $x = 1$, but does not constrain the value at -1 . Following the same procedure as used to prove Theorem 2.2, $\phi^T(-1)\vec{u}$ is equated to $f(-1)$, giving $u_1 \widehat{\phi}_1(-1) = a_1$. Plugging (2.6) and (2.9) into (2.4) gives equations that must be true for all \vec{a} . As before, the equation from a_1 determines $d_{1,1}$. The remaining equations can be written as

$$(2.10) \quad \int_{-1}^1 \frac{1-x^2}{4} \begin{bmatrix} 1 \\ x \\ \vdots \\ x^{p-2} \end{bmatrix} \widehat{\phi}_1 dx = 0.$$

This shows that the function $\widehat{\phi}_1$ must be orthogonal to $\mathcal{P}(p-2)$ with respect to the weighting $\frac{1-x^2}{4}$. The Jacobi polynomials satisfy

$$\int_{-1}^1 P_m^{(\alpha,\beta)} P_n^{(\alpha,\beta)} (1-x)^\alpha (1+x)^\beta = \delta_{m,n}.$$

Because the space $\mathcal{P}(p-2)$ can be represented using the Jacobi polynomials $P_n^{(1,1)}(x)$ for $n \in [0, p-2]$, the polynomial $P_{p-1}^{(1,1)}(x)$ will satisfy (2.10). The left vertex function can therefore be represented as $\frac{1-x}{2} P_{p-1}^{(1,1)}(x)$. Following the same procedure for the right vertex function shows that it can be represented as $\frac{1+x}{2} P_{p-1}^{(1,1)}(x)$. The roots of both polynomials in $(-1, 1)$ are determined by $P_{p-1}^{(1,1)}(x)$ and thus have the same locations. Not surprising, this shows that a GL integration rule may exist in one dimension. Based on the already known expression for the locations of the GL points, it also shows that $P'_p(x) = P_{p-1}^{(1,1)}(x)$.

3. Triangles. In this section, the same basic steps are used to show that a GL integration rule does not exist on triangles. First, it is shown that the existence of a GL integration rule with the properties defined in the introduction implies the existence of a nodal basis for the space $\mathcal{T}(p)$ that has a diagonal projection operation that is

exact for functions in the space $\mathcal{T}(p - 1)$. It is then shown that the basis satisfying this property is unique and not nodal, proving that the GL integration rule does not exist.

Before beginning, a standard triangle on which to perform the operations is defined by $\{r, s \mid -1 \leq r \leq 1, -1 \leq s \leq r\}$, as shown in Figure 1. Following Dubiner [13], we introduce coordinates $\xi = -1 + 2(1 + r)/(1 - s)$ and $\eta = s$, which are shown on the figure as well. In this coordinate system, the standard triangle is defined by $-1 \leq \xi \leq 1, -1 \leq \eta \leq 1$. Integration over the standard element is given by

$$\int_{-1}^1 \int_{-1}^r f(r, s) ds dr = \int_{-1}^1 \int_{-1}^1 f(\xi, \eta) \frac{1 - \eta}{2} d\eta d\xi$$

in these coordinate systems.

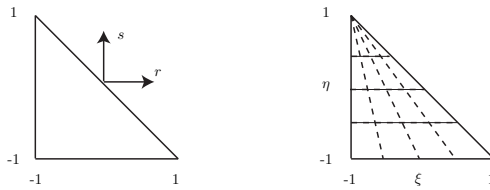


FIG. 1. Standard triangle and coordinate systems.

As in one dimension, it is assumed that the GL integration rule has the form

$$\int_{-1}^1 \int_{-1}^r f(r, s) ds dr \approx \sum_{i=1}^{N(p)} w_i f(r_i, s_i),$$

where $f(r, s)$ is the function to be integrated, $N(p) \equiv \dim(\mathcal{T}(p)) = (p + 1)(p + 2)/2$ is the number of points in the GL rule, and w_i is the weight associated with the point located at r_i, s_i . Three of the points are required to be at the triangle vertices, $r, s = (-1, -1), (-1, 1),$ and $(1, -1)$, and $p - 1$ points are required to be along each side of the element, $r = -1, s = -1,$ and $r = s$. The remaining $N(p - 3) = (p - 1)(p - 2)/2$ points are assumed to be in the interior of the element. A formula is sought that can integrate polynomials in the space $\mathcal{T}(2p - 1)$ exactly.

Some basic observations about the space $\mathcal{T}(p)$ are first given. This space can be decomposed into interior, side, and vertex modes. Interior modes are zero on all sides of the triangle and can be constructed from the space

$$\mathcal{I}(p) \equiv (r + 1)(s + 1)(r + s)\mathcal{T}(p - 3).$$

This is a general space for the interior modes, and it contains all polynomials in $\mathcal{T}(p)$ that have three component curves defined by $r = -1, s = -1,$ and $r = -s$. (See [26, section 1.8] for a definition of component curves.) In some cases, it will be convenient to have an explicit representation of the interior space. In this case, the interior modes of the modified Dubiner basis [13] will be used. These are described in ξ, η coordinates as

$$\phi_{\text{int}, m, n} = \left(\frac{1 + \xi}{2}\right) \left(\frac{1 - \xi}{2}\right) P_m^{2,2}(\xi) \left(\frac{1 - \eta}{2}\right)^{m+2} \left(\frac{1 + \eta}{2}\right) P_n^{2m+5,2}(\eta),$$

where $0 \leq m < p - 2$, $0 \leq n < p - 2 - m$. In some cases, a one-dimensional numbering of the interior modes will be needed, in which case $\phi_{\text{int},m,n}$ will be replaced by $\phi_{\text{int},j}$, where $j = N(m + n - 1) + n + 1$.

There are three distinct sets of side modes. The sides are numbered as shown in Figure 2, with side 1 being opposite to vertex 1. General spaces for constructing the side modes are

$$\begin{aligned} \mathcal{S}_1(p) &\equiv (r + 1)(r + s)\mathcal{T}(p - 2), \\ \mathcal{S}_2(p) &\equiv (r + 1)(s + 1)\mathcal{T}(p - 2), \\ \mathcal{S}_3(p) &\equiv (s + 1)(r + s)\mathcal{T}(p - 2). \end{aligned}$$

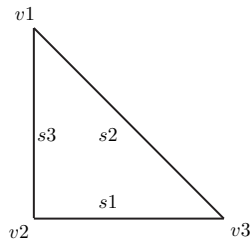


FIG. 2. Numbering of the vertices and sides of the triangle.

The side modes can be constructed from $p - 1$ modes that are nonzero along the side and any linear combination of interior modes. Thus, each of these spaces includes the interior space as a subset. For each side, the form of the $p - 1$ side modes in the modified Dubiner basis is given by

$$\begin{aligned} \phi_{s1,m} &= \left(\frac{1 + \xi}{2}\right) \left(\frac{1 - \xi}{2}\right) P_m^{2,2}(\xi) \left(\frac{1 - \eta}{2}\right)^{m+2}, \\ \phi_{s2,m} &= \left(\frac{1 + \xi}{2}\right) \left(\frac{1 - \eta}{2}\right) \left(\frac{1 + \eta}{2}\right) P_m^{2,2}(\eta), \\ \phi_{s3,m} &= (-1)^m \left(\frac{1 - \xi}{2}\right) \left(\frac{1 - \eta}{2}\right) \left(\frac{1 + \eta}{2}\right) P_m^{2,2}(\eta), \end{aligned}$$

where $(0 \leq m < p - 1)$.

Vertex modes are constrained to be one at one vertex and zero along the opposing side. General spaces for obtaining vertex modes are given by

$$\begin{aligned} \mathcal{V}_1 &= (1 + s)\mathcal{T}(p - 1), \\ \mathcal{V}_2 &= (r + s)\mathcal{T}(p - 1), \\ \mathcal{V}_3 &= (1 + r)\mathcal{T}(p - 1). \end{aligned}$$

Vertex modes can be constructed using a vertex function and any combination of modes from the two adjacent sides as well as interior modes. Thus, the vertex 1 space, for example, contains the \mathcal{S}_2 , \mathcal{S}_3 , and \mathcal{T} spaces as a subset. In the modified Dubiner basis, the three vertex modes are linear functions that are one at one vertex and zero

along the opposing side:

$$\begin{aligned}\phi_{v1} &= \left(\frac{1+\eta}{2}\right), \\ \phi_{v2} &= \left(\frac{1-\xi}{2}\right) \left(\frac{1-\eta}{2}\right), \\ \phi_{v3} &= \left(\frac{1+\xi}{2}\right) \left(\frac{1-\eta}{2}\right).\end{aligned}$$

The vertex, side, and interior modes of the modified Dubiner basis are assembled into a single basis vector, $\vec{\phi}$, by listing first the three vertex modes, then the side 1 modes, the side 2 modes, the side 3 modes, and lastly the interior modes. To distinguish different basis orders, the notation $\vec{\phi}_p$ is used.

As in one dimension, the first step is to show that the existence of a GL integration rule guarantees the existence of a nodal basis that allows exact diagonal projection for functions in $\mathcal{T}(p-1)$. The following theorem is slightly more difficult to prove in two dimensions.

THEOREM 3.1. *The existence of a GL integration rule on the triangle guarantees the existence of a nodal basis on the triangle.*

Proof. If a function in $\mathcal{T}(p)$, say $\vec{\phi}^T \vec{a}$, is to exactly reproduce the values of a function $u(r, s)$ at the GL points, the following must be true:

$$(3.1) \quad \sum_{j=1}^{N(p)} a_j \phi_j(r_k, s_k) = u(r_k, s_k) \quad \forall k \in [1, N(p)].$$

This can be written more compactly as

$$P\vec{a} = \vec{u},$$

where P is an $N(p) \times N(p)$ square matrix with entries given by

$$p_{j,k} = \phi_j(r_k, s_k),$$

and \vec{u} is a column vector containing the values of $u(r, s)$ at each GL point. To find the i th mode of the nodal basis, ψ_i , $u(r_k, s_k)$ is set to $\delta_{i,k}$. If P is invertible, then the nodal basis is uniquely determined. This is in agreement with Theorem 3.7-3 in [26] which proves a similar result and then goes on to investigate the properties of these functions.

Now assume P is singular. In this case, there are either an infinite number of solutions to (3.1) or no solutions. If \vec{u} is chosen to be evaluated using a function in $\mathcal{T}(p)$, then there is certainly a function in $\mathcal{T}(p)$ that can reproduce these values in this particular case. This shows that there is at least one solution. To prove that this solution is unique, assume that there are two distinct functions, u_1 and u_2 , in $\mathcal{T}(p)$ that produce the same values on the GL points. Let these functions be represented using the modified Dubiner basis. Because there is a GL point located at each vertex, the coefficients of the vertex modes for both functions must be identical. Furthermore, because there are $p-1$ GL points on each side, the coefficients of the side modes are also uniquely determined. u_1 and u_2 can therefore differ only in the coefficients of the interior modes. However, the GL integration rule integrates all polynomials in

$\mathcal{T}(2p - 1)$ exactly, and both functions are assumed to have the same values on the Gauss points. Therefore,

$$\int_{-1}^1 \int_{-1}^r \vec{\phi}_{p-3} u_1 ds dr = \int_{-1}^1 \int_{-1}^r \vec{\phi}_{p-3} u_2 ds dr.$$

This actually holds for $\vec{\phi}_{p-1}$, but the additional constraints are not necessary for the proof. u_1 and u_2 have the same side and vertex modes, so they can be eliminated from both sides of the equation. The interior space of functions can be represented as

$$\mathcal{I}(p) = \text{span} \left[(1 + s)(1 + r)(r + s)\vec{\phi}_{p-3} \right].$$

In the same way that showed that (2.8) is symmetric positive definite, it can be shown that the above equation results in a symmetric positive definite matrix. u_1 and u_2 must therefore be identical. Since there is a unique solution, then P is not singular and the nodal basis is uniquely determined. \square

Given the nodal basis and the GL integration rule, Theorem 2.1 can be extended to apply to triangles with no modification. This shows that if a GL integration rule exists, there is a nodal basis, and there is an exact diagonal projection operation for functions in $\mathcal{T}(p - 1)$. Following along with the one dimension logic, the next step is to prove the following theorem.

THEOREM 3.2. *The three triangle vertex modes that allow exact diagonal projection of functions from $\mathcal{T}(p - 1)$ are unique.*

Proof. Let the function to be projected, $f(r, s)$, be contained in $\mathcal{T}(p - 1)$ and described as

$$f(r, s) = \vec{\phi}_{p-1}^T \vec{a},$$

and let the projected function be represented by

$$u(r, s) = \vec{\psi}_p^T \vec{u},$$

where ψ is the basis allowing diagonal projection. Let the first vertex mode, ψ_1 , be described using the modified Dubiner basis as $(1 + s)\vec{\phi}_{p-1}^T \vec{b}$. Since ψ_1 is assumed to be a vertex mode, b_1 is not zero. The mode can be scaled by an arbitrary constant, so b_1 can be constrained to be 1. Because $u(r, s)$ must equal $f(r, s)$ at the vertex point, u_1 is then equal to a_1 . Diagonal projection requires that

$$d_{1,1} u_1 = d_{1,1} a_1 = \int_{\Omega} (1 + s) \vec{\phi}_{p-1}^T \vec{b} \vec{\phi}_{p-1}^T \vec{a} dr ds$$

hold for all \vec{a} . This again results in a set of symmetric positive definite matrices for the coefficients from b_2 to $b_{N(p-1)}$. To see this, the first component from the vectors \vec{b} and \vec{a} is explicitly extracted and then the remaining part of the vectors is represented as $\vec{b}_{/1} = b_2, \dots, b_{N(p-1)}$. In the following, all subscripts of /1 indicate the vector without the first component. The constraint corresponding to a_1 determines the diagonal projection constant $d_{1,1}$. The remaining constraints are given by

$$(3.2) \quad \int_{\Omega} (1 + s) \vec{\phi}_{p-1, /1}^T \vec{b}_{/1} \vec{\phi}_{p-1, /1}^T \vec{a}_{/1} dr ds = - \int_{\Omega} \frac{(1 + s)^2}{2} b_1 \vec{\phi}_{p-1, /1}^T \vec{a}_{/1} dr ds.$$

These are $N(p - 1) - 1$ equations in $N(p - 1) - 1$ unknowns (b_1 is set to one). That the matrix is positive definite can be seen by first letting $\vec{b}_{/1} = \vec{a}_{/1}$ and then defining $g(r, s)$ as $\vec{\phi}_{p-1, /1}^T \vec{a}_{/1}$. This results in

$$\int_{\Omega} (1 + s)(g(r, s))^2 dr ds,$$

which is positive over the triangle. Thus the matrix is positive definite, and the vertex mode that allows diagonal projection is unique. \square

THEOREM 3.3. *If the zero curves of the three vertex modes do not coincide at $p - 1$ locations along each side of the triangle, then a GL integration rule does not exist.*

Proof. Assume

1. a GL integration rule exists, and
2. vertex functions exist satisfying the diagonal projection property, but the zero curves of these functions do not intersect at $p - 1$ locations along any side of the triangle.

By assumption 1 and Theorem 3.1, there exist vertex functions satisfying the diagonal projection property. Furthermore, these functions are from a nodal basis, and there are $p - 1$ nodes along each side. This implies that all three functions are zero at $p - 1$ locations on each triangle side. Because the vertex functions are unique by Theorem 3.2, this contradicts item 2 above. Thus assumption 2 excludes the existence of the GL integration rule. \square

The final step is to determine analytic expressions for the vertex functions. The easiest way to find the vertex functions is to simply invert (3.2) numerically. This result was used as a guide to determine an analytic description of the vertex functions. The analytic expression can be found most easily using ξ, η coordinates on the triangle. Treating b_1 as an unknown and letting $\sigma = \vec{\phi}_{p-1}^T \vec{b} \in \mathcal{T}(p - 1)$, (3.2) can be written as

$$(3.3) \quad \int_{-1}^1 \int_{-1}^1 (1 + \eta) \sigma \vec{\phi}_{p-1, /1}^T \vec{a}_{/1} \frac{1 - \eta}{2} d\xi d\eta = 0.$$

This shows that the function σ should be orthogonal (with respect to a weighting function) to the space $\mathcal{T}(p - 1)$ excluding the vertex 1 mode. This space is formed by the union of the two other vertex spaces, $\mathcal{V}_2 \cup \mathcal{V}_3$. The numerical results indicate that σ is only a function of η . Therefore only the η components of this equation can be considered. The basis for the space $\mathcal{V}_2 \cup \mathcal{V}_3$ consists of the two vertex modes, ϕ_{v2} and ϕ_{v3} , the side modes $\phi_{s1,m}$, $\phi_{s2,m}$, and $\phi_{s3,m}$ with $0 \leq m < p - 2$, and $\phi_{\text{int},m,n}$ with $0 \leq m < p - 3$, $0 \leq n < p - 3 - m$. All of these modes include the factor $\frac{1 - \eta}{2}$ and reach a maximum degree in η of $p - 1$, and thus the η component of any function of the space can be constructed from $(1 - \eta)\mathcal{P}(p - 2)$. The η component of the orthogonality constraint is then

$$(3.4) \quad \int_{-1}^1 \sigma \mathcal{P}(p - 2)(1 - \eta)^2(1 + \eta) d\eta = 0.$$

If σ is only a function of η , then $\sigma \in \mathcal{P}(p - 1)$. To satisfy this orthogonality requirement, σ must be the Jacobi polynomial $P_{p-1}^{(2,1)}(\eta)$. Because this polynomial is orthogonal to the functions $P_m^{(2,1)}(\eta)$ for $m \in [0, p - 2]$ and these functions span $\mathcal{P}(p - 2)$, this choice

satisfies (3.4), and thus (3.3) as well. The vertex 1 function that allows diagonal projection on the triangle is thus

$$\psi_{v1} = \frac{1+s}{2} \frac{P_{p-1}^{(2,1)}(s)}{P_{p-1}^{(2,1)}(1)},$$

where it has been normalized such that the value of the function at the vertex is 1. The other two vertex functions can be found by using the rotational symmetry of the triangle. For example, to find ψ_{v2} , one can substitute $-1-r-s$ for s to obtain

$$\psi_{v2} = \frac{-(r+s)}{2} \frac{P_{p-1}^{(2,1)}(-1-r-s)}{P_{p-1}^{(2,1)}(1)}.$$

For a GL rule to exist, these two functions should have the same roots along the adjacent side, $r = -1$. If $p - 1$ is even, this implies that the function should be an even function of s , and if $p - 1$ is odd, the function should be an odd function of s . Based on the fact that the Jacobi polynomials, $P_n^{(2,1)}(x)$, are orthogonal with respect to a nonsymmetric weighting function $(1-x)^2(1+x)$, it is fairly obvious that they are not symmetric. To be sure, the polynomial form given by

$$P_n^{(\alpha,\beta)}(x) = (1-x)^{-\alpha}(1+x)^{-\beta} \frac{d^n}{dx^n} \left[(1-x)^{(\alpha+n)}(1+x)^{(\beta+n)} \right]$$

is examined. Letting $\alpha = 2$ and $\beta = 1$, after some manipulation this can be rewritten as

$$P_n^{(2,1)}(x) = \frac{1}{1-x^2} \left[\frac{d}{dx} - \frac{n}{1-x} \right] \frac{d^{n-1}}{dx^{n-1}} (1-x^2)^{n+1}.$$

If n is even, this function must be even for a GL rule to exist, and if n is odd, it should be odd. For the case of n even, the function $\frac{d^{n-1}}{dx^{n-1}}(1-x^2)^{n+1}$ is odd. Denote it as $g(x)$. The above then becomes

$$P_n^{(2,1)}(x) = \frac{1}{1-x^2} \left[\frac{dg(x)}{dx} - \frac{n}{1-x}g(x) \right]$$

and

$$P_n^{(2,1)}(-x) = \frac{1}{1-x^2} \left[\frac{dg(x)}{dx} - \frac{n}{1+x}(-g(x)) \right].$$

For $P_n^{(2,1)}(x)$ to be even, $P_n^{(2,1)}(x) - P_n^{(2,1)}(-x)$ should equal 0. The above gives

$$P_n^{(2,1)}(x) - P_n^{(2,1)}(-x) = \frac{1}{1-x^2} \left[\frac{-2n}{1-x^2}g(x) \right].$$

$g(x)$ is not zero, so for any even n greater than zero the function is not even. A similar argument can be made for the case of odd n . The only case where the function is symmetric is $n = 0$.

Based on Theorem 3.3, because the roots of the diagonal projection vertex modes do not coincide along the side, a GL integration rule does not exist on the triangle for

$p > 1$. For $p = 1$, locating three Gauss points on the vertices does allow integration of the space $1, r, s$ exactly.

Although no GL integration rule exists on the triangle, the fact that a diagonal projection vertex function exists gives hope that a finite element method similar to the spectral element method on quadrilaterals can still be developed. The diagonal projection vertex mode is shown in Figure 3. The grayscale shows the values for vertex mode 1, which has the value 1 at the top of the triangle. The solid black contour lines are the zero contours for this function. The dashed contour lines are the zero contours for vertex mode 2 and 3, which are rotations of vertex mode 1. These lines are shown to further demonstrate that the zero intersection points do not coincide.

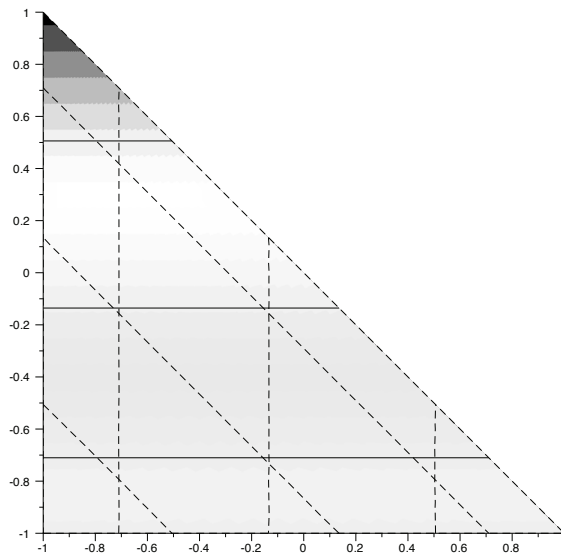


FIG. 3. Contours of the vertex 1 mode that allows diagonal projection. Dashed lines are the zero lines of the vertex 2 and vertex 3 modes.

The interesting thing about the function shown in Figure 3 is that it is localized near the vertex and close to 0 elsewhere. This is similar to the GLL vertex functions used in quadrilateral and hexahedral spectral element methods. The most important point is that such a function allows a diagonal approximation to the mass matrix that is accurate to order $p - 1$. On quadrilaterals this property allows optimal spatial convergence rates to be obtained by unsteady explicit simulations [20]. Thus on triangles, optimal explicit simulations using continuous high-order polynomial approximations may still be possible even though a GL rule does not exist. Our continuing work is to determine whether there exist side and interior modes which also have the diagonal projection property.

4. Conclusions. It has been proven that a Gauss-Lobatto (GL) integration rule for triangles that has characteristics similar to GL integration on line segments, quadrilaterals, and hexahedra does not exist. Specifically, there is no integration rule having a point at each triangle vertex, $p - 1$ points on each triangle side, and $(p - 1)(p - 2)/2$ points in the interior that is capable of exactly integrating the space $\mathcal{T}(2p - 1)$. This also implies that there is no equivalent to the spectral element GLL nodal basis on the triangle. However, the analysis also shows that there is a vertex mode that

allows a diagonal approximation to the mass matrix accurate to order $p - 1$. This function may be a key to developing explicit simulations using continuous high-order polynomial approximations on triangles.

REFERENCES

- [1] M. ABRAMOWITZ AND I. A. STEGUN, EDs., *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*, Dover Publications, New York, 1965.
- [2] M. G. BLYTH AND C. POZRIKIDIS, *A Lobatto interpolation grid over the triangle*, IMA J. Appl. Math., 71 (2006), pp. 153–169.
- [3] Q. CHEN AND I. BABUŠKA, *Approximate optimal points for polynomial interpolation of real functions in an interval and in a triangle*, Comput. Methods Appl. Mech. Engrg., 128 (1995), pp. 405–417.
- [4] M. J. S. CHIN-JOE-KONG, W. A. MULDER, AND M. V. VELDHUIZEN, *Higher-order triangular and tetrahedral finite elements with mass lumping for solving the wave equation*, J. Engrg. Math., 35 (1999), pp. 405–426.
- [5] G. COHEN, P. JOLY, J. E. ROBERTS, AND N. TORDJMAN, *Higher order triangular finite elements with mass lumping for the wave equation*, SIAM J. Numer. Anal., 38 (2001), pp. 2047–2078.
- [6] R. COOLS, *Constructing cubature formulae: The science behind the art*, Acta Numer., 6 (1997), pp. 1–54.
- [7] R. COOLS, *Monomial cubature rules since “Stroud”: A compilation. II. Numerical Evaluation of Integrals*, J. Comput. Appl. Math., 112 (1999), pp. 21–27.
- [8] R. COOLS, *Advances in multidimensional integration*, J. Comput. Appl. Math., 149 (2002), pp. 1–12.
- [9] R. COOLS, *An encyclopaedia of cubature formulas*, J. Complexity, 19 (2003), pp. 445–453.
- [10] R. COOLS, I. MYSOVSKIKH, AND H. SCHMID, *Cubature formulae and orthogonal polynomials*, J. Comput. Appl. Math., 127 (2001), pp. 121–152.
- [11] R. COOLS AND P. RABINOWITZ, *Monomial cubature rules since “Stroud”: A compilation*, J. Comput. Appl. Math., 48 (1993), pp. 309–326.
- [12] S. DEY, J. E. FLAHERTY, T. K. OHSUMI, AND M. S. SHEPHARD, *Integration by table look-up for p -version finite elements on curved tetrahedra*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 4532–4543.
- [13] M. DUBINER, *Spectral methods on triangles and other domains*, J. Sci. Comput., 6 (1991), pp. 345–390.
- [14] D. A. DUNAVANT, *High degree efficient symmetrical gaussian quadrature rules for the triangle*, Internat. J. Numer. Methods Engrg., 21 (1985), pp. 1129–1148.
- [15] J. S. HESTHAVEN, *From electrostatics to almost optimal nodal sets for polynomial interpolation in a simplex*, SIAM J. Numer. Anal., 35 (1998), pp. 655–676.
- [16] T. J. R. HUGHES, *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1987.
- [17] Y. LIU AND M. VINOKUR, *Exact integrations of polynomials and symmetric quadrature formulas over arbitrary polyhedral grids*, J. Comput. Phys., 140 (1998), pp. 122–147.
- [18] H. LUO AND C. POZRIKIDIS, *A Lobatto interpolation grid in the tetrahedron*, IMA J. Appl. Math., 71 (2006), pp. 298–313.
- [19] J. N. LYNNESS AND R. COOLS, *A survey of numerical cubature over triangles*, in Mathematics of Computation 1943–1993: A Half-Century of Computational Mathematics (Vancouver, BC, 1993), Proc. Sympos. Appl. Math. 48, AMS, Providence, RI, 1994, pp. 127–150.
- [20] Y. MADAY AND A. T. PATERA, *Spectral element methods for the incompressible Navier-Stokes equations*, in State-of-the-Art Surveys on Computational Mechanics, A. K. Noor and J. T. Oden, eds., The American Society of Mechanical Engineers, New York, 1989, pp. 71–143.
- [21] C. MAVRIPLIS AND J. VAN ROSENDALE, *Triangular spectral elements for incompressible fluid flow*, in Proceeding of the 11th AIAA Computational Fluid Dynamics Conference, Orlando, FL, 1993, paper AIAA-1993-3346.
- [22] I. P. MYSOVSKIKH, *Interpolyatsionnye kubaturnye formuly*, “Nauka,” Moscow, 1981.
- [23] R. PASQUETTI AND F. RAPETTI, *Spectral element methods on triangles and quadrilaterals: Comparisons and applications*, J. Comput. Phys., 198 (2004), pp. 349–362.
- [24] R. PASQUETTI AND F. RAPETTI, *Spectral element methods on unstructured meshes: Comparisons and recent advances*, J. Sci. Comput., 27 (2006), pp. 377–387.

- [25] H. T. RATHOD AND M. SHAJEDUL KARIM, *An explicit integration scheme based on recursion for the curved triangular finite elements*, *Comput. & Structures*, 80 (2002), pp. 43–76.
- [26] A. H. STROUD, *Approximate calculation of multiple integrals*, Prentice–Hall, Englewood Cliffs, NJ, 1971.
- [27] M. A. TAYLOR AND B. A. WINGATE, *A generalized diagonal mass matrix spectral element method for non-quadrilateral elements*, *Appl. Numer. Math.*, 33 (2000), pp. 259–265.
- [28] M. A. TAYLOR, B. A. WINGATE, AND L. P. BOS, *A cardinal function algorithm for computing multivariate quadrature points*, *SIAM J. Numer. Anal.*, 45 (2007), pp. 193–205.
- [29] S. WANDZURA AND H. XIAO, *Symmetric quadrature rules on a triangle*, *Comput. Math. Appl.*, 45 (2003), pp. 1829–1840.
- [30] T. WARBURTON, L. F. PAVARINO, AND J. S. HESTHAVEN, *A pseudo-spectral scheme for the incompressible Navier-Stokes equations using unstructured nodal elements*, *J. Comput. Phys.*, 164 (2000), pp. 1–21.

UNIFIED HYBRIDIZATION OF DISCONTINUOUS GALERKIN, MIXED, AND CONTINUOUS GALERKIN METHODS FOR SECOND ORDER ELLIPTIC PROBLEMS*

BERNARDO COCKBURN[†], JAYADEEP GOPALAKRISHNAN[‡], AND
RAYTCHO LAZAROV[§]

Abstract. We introduce a unifying framework for hybridization of finite element methods for second order elliptic problems. The methods fitting in the framework are a general class of mixed-dual finite element methods including hybridized mixed, continuous Galerkin, nonconforming, and a new, wide class of hybridizable discontinuous Galerkin methods. The distinctive feature of the methods in this framework is that the only globally coupled degrees of freedom are those of an approximation of the solution defined only on the boundaries of the elements. Since the associated matrix is sparse, symmetric, and positive definite, these methods can be efficiently implemented. Moreover, the framework allows, in a single implementation, the use of different methods in different elements or subdomains of the computational domain, which are then automatically coupled. Finally, the framework brings about a new point of view, thanks to which it is possible to see how to devise novel methods displaying very localized and simple mortaring techniques, as well as methods permitting an even further reduction of the number of globally coupled degrees of freedom.

Key words. discontinuous Galerkin methods, mixed methods, continuous methods, hybrid methods, elliptic problems

AMS subject classifications. 65N30, 65M60

DOI. 10.1137/070706616

1. Introduction. We introduce a new *unifying framework* for hybridization of finite element methods for second order elliptic problems. This framework is unifying in the sense that it includes as particular cases hybridized versions of mixed methods [4, 11, 26], the continuous Galerkin (CG) method [31], and a new, wide class of hybridizable discontinuous Galerkin (DG) methods. The unifying framework allows us to (i) significantly reduce the number of the globally coupled degrees of freedom of DG methods, (ii) use different methods in different parts of the computational domain and automatically couple them, and (iii) devise novel methods employing new mortaring techniques. We develop the unifying framework on the following model elliptic boundary value problem of second order written in mixed form:

$$\begin{aligned} (1.1a) \quad & \mathbf{q} + a \operatorname{grad} u = 0 && \text{on } \Omega, \\ (1.1b) \quad & \operatorname{div} \mathbf{q} + du = f && \text{on } \Omega, \\ (1.1c) \quad & u = g && \text{on } \partial\Omega. \end{aligned}$$

*Received by the editors October 29, 2007; accepted for publication (in revised form) November 7, 2008; published electronically February 25, 2009.

<http://www.siam.org/journals/sinum/47-2/70661.html>

[†]School of Mathematics, University of Minnesota, Minneapolis, MN 55455 (cockburn@math.umn.edu). This author's research was supported in part by the National Science Foundation (grant DMS-0411254) and by the University of Minnesota Supercomputing Institute.

[‡]Department of Mathematics, University of Florida, Gainesville, FL 32611–8105 (jayg@math.ufl.edu). This author's research was supported in part by the National Science Foundation (grants DMS-0410030, DMS-0713833, and SCREMS-0619080).

[§]Department of Mathematics, Texas A&M University, College Station, TX 77843–3368 (lazarov@math.tamu.edu). This author's research was supported in part by the National Science Foundation (grants NSF-DMS-0713829 and NSF-CNS-ITR-0540136).

Here $\Omega \subset \mathbb{R}^n$ is a polyhedral domain ($n \geq 2$), $d(\mathbf{x})$ is a scalar nonnegative function, and $a(\mathbf{x})$ is a matrix valued function that is symmetric and uniformly positive definite on Ω . In addition, we assume that the function g is the restriction of a smooth scalar function on $\partial\Omega$ and that the functions f , d , and a are smooth on $\overline{\Omega}$. These assumptions can be vastly generalized, but we take them for the sake of a transparent presentation of the design of our unifying framework.

1.1. The structure of the methods of the unifying framework. Let us begin the description of our results by arguing that what makes possible the construction of the unified framework is that all the numerical methods fitting in it are constructed by using a discrete version of a single property of the exact solution of problem (1.1). This property is a characterization of the values of the exact solution u on the interior boundaries of each of the elements K of any triangulation of the domain Ω , \mathcal{T}_h . Let us describe it.

If on the border of the element K , ∂K , we set $u = \lambda + g$, where

$$(1.2) \quad \lambda = \begin{cases} u & \text{on } \partial K \setminus \partial\Omega, \\ 0 & \text{on } \partial K \cap \partial\Omega, \end{cases} \quad \text{and} \quad g = \begin{cases} 0 & \text{on } \partial K \setminus \partial\Omega, \\ g & \text{on } \partial K \cap \partial\Omega, \end{cases}$$

by the *linearity* of the problem, we have that

$$(1.3) \quad (\mathbf{q}, u) = (\mathbf{Q}\lambda + \mathbf{Q}g + \mathbf{Q}f, \mathbf{U}\lambda + \mathbf{U}g + \mathbf{U}f) \quad \text{in } \Omega,$$

where the so-called *local solvers* $(\mathbf{Q}(\cdot), \mathbf{U}(\cdot))$ are defined on the element $K \in \mathcal{T}_h$ as follows. For any single-valued functions m on $L^2(\partial K)$ and f on $L^2(K)$, the functions $(\mathbf{Q}m, \mathbf{U}m)$ and $(\mathbf{Q}f, \mathbf{U}f)$ are the solutions of

$$(1.4a) \quad c \mathbf{Q}m + \mathbf{grad} \mathbf{U}m = 0, \quad \text{div } \mathbf{Q}m + d \mathbf{U}m = 0 \quad \text{on } K, \quad \mathbf{U}m = m \quad \text{on } \partial K,$$

$$(1.4b) \quad c \mathbf{Q}f + \mathbf{grad} \mathbf{U}f = 0, \quad \text{div } \mathbf{Q}f + d \mathbf{U}f = f \quad \text{on } K, \quad \mathbf{U}f = 0 \quad \text{on } \partial K,$$

where $c = a^{-1}$ for each element $K \in \mathcal{T}_h$.

Conversely, the above property holds if and only if (see, for example, [46]) the normal component of $\mathbf{Q}\lambda + \mathbf{Q}g + \mathbf{Q}f$ across interelement boundaries is continuous. We thus see that this *transmission* condition, which we formally express as

$$(1.5) \quad \llbracket \mathbf{Q}\lambda + \mathbf{Q}g + \mathbf{Q}f \rrbracket = 0,$$

completely *characterizes* the function λ . Here $\llbracket \cdot \rrbracket$ denotes the jump of the normal component of the a vector across ∂K .

The finite element methods of the unified framework are those that can be expressed as a discrete version of the above property. In this way, the only globally coupled degrees of freedom *are bound* to be those describing the approximation to λ . Thus, each of those method provides an approximate solution of the form

$$(1.6) \quad (\mathbf{q}_h, u_h) = (\mathbf{Q}\lambda_h + \mathbf{Q}g_h + \mathbf{Q}f, \mathbf{U}\lambda_h + \mathbf{U}g_h + \mathbf{U}f),$$

where λ_h , respectively, g_h , is an approximation in some finite-dimensional space M_h , respectively, M_h , of the values of u on the faces of the elements lying in the interior, respectively, in the border of Ω , and $(\mathbf{Q}m, \mathbf{U}m)$ and $(\mathbf{Q}f, \mathbf{U}f)$ are discrete versions of the exact local solvers (1.4)—we keep the same notation for the sake of simplicity. Moreover, the methods are such that λ_h *can* be determined by a discrete version of transmission condition (1.5), which we write as follows:

$$(1.7) \quad a_h(\lambda_h, \mu) = b_h(\mu) \quad \text{for all } \mu \in M_h.$$

In [26], where the hybridization of mixed methods was considered, the equation determining λ_h was called the *jump condition*. In our setting, it is called the *conservativity condition* to reflect the incorporation into the framework of DG and CG methods.

Note that all the methods in the unified framework provide approximations for (\mathbf{q}, u) in the interior of the elements $K \in \mathcal{T}_h$, (\mathbf{q}_h, u_h) , as well as an approximation of u on the interior border of the elements λ_h ; this is why they are called hybrid. This is in agreement with the definition of hybrid methods proposed in [22, p. 421]: “we may define more generally as a *hybrid* method any finite element method based on a formulation where one unknown is a function, or some of its derivatives, on the set Ω , and the other unknown is the trace of some of its derivatives of the same function, or the trace of the function itself, along the boundaries of the set K .” Here K denotes a typical element of the triangulation. A long list of hybrid methods can be found in [22, 12, 51].

Of course, not every finite element method displays the above roughly described structure; in particular, it might not even be a hybrid method. However, many such methods can be rewritten as hybrid methods; this process is what can be called the hybridization of a finite element method. We say that we can hybridize a given finite element method if we can find a *hybrid* method (part) of whose solution *coincides* with the solution of the given method. The original finite element method is called *hybridizable*, and the hybrid method is then said to be a *hybridization* of the original method; for short, we call it a hybridized method. Next, we give a brief overview of the hybridization techniques of relevance for our purposes.

1.2. Hybridization of finite element methods. The first hybridization of a finite element method was proposed in 1965 [39] for a numerical method for solving the equations of linear elasticity. Perhaps because it was then intended as an implementation technique, the distinction between hybridization and *static condensation*, a widely known algebraic manipulation for size reduction of *already* assembled matrices, is seldom made in the engineering literature. However, in 1985 [4], hybridization was shown to be more than an implementation trick as it was proven that the new unknown λ_h , also interpreted to be the *Lagrange multiplier* associated with a continuity condition on the approximate flux, contains *extra* information about the exact solution. This was used to enhance the accuracy of the approximation by means of a local postprocessing [4, 11, 35]; see also [10].

After yet another two decades, a new perspective on hybridization emerged [26], and the characterization of the approximate trace λ_h as the solution of weak formulation (1.7) was introduced; this was done in the setting of the hybridization of the Raviart–Thomas (RT) and Brezzi–Douglas–Marini (BDM) mixed methods of arbitrary degree. The special case of the lowest order RT method had been previously considered in [21] within the framework of a study of the equivalence of mixed and nonconforming methods. In [26], it was shown that formulation (1.7) not only simplifies the task of assembling the stiffness matrix for the multiplier but can be used to establish unsuspected links between apparently unrelated mixed methods. It was also shown that it allows the devising and analysis of new, variable degree versions of those methods [27].

This new hybridization approach was later extended to finite element methods for the stationary Stokes equations using spaces of exactly divergence-free velocities; it was intended as an effective technique to bypass the extremely difficult construction of such spaces. It was successfully applied to a DG method [15] and to a mixed method for Stokes flow [28, 29]. For a review of these results, see [30]. Recently [31],

this hybridization approach was applied to the CG method to pave the way for the computation of an $H(\text{div})$ -conforming approximation of the flux from the CG solution.

1.3. Hybridization of DG methods. In this paper, we continue this effort and show how to hybridize a large class of DG methods. Thus, we show that their approximate solution (\mathbf{q}_h, u_h) can be expressed as in (1.6) and that the approximate trace λ_h , which is nothing but the so-called numerical trace \hat{u}_h on the interelement boundaries (see [5]) satisfies weak formulation (1.7). In other words, we identify a class of DG methods whose globally coupled degrees of freedom are those of the numerical trace \hat{u}_h *only*; this results in an efficient implementation of these methods, as we argue below. In this way, the main disadvantages of DG methods for elliptic problems compared to other methods, namely, a higher number of globally coupled degrees of freedom for the same mesh and a lower sparsity of the corresponding stiffness matrices, are eliminated to a significant extent.

The simplest examples of such methods are obtained by using a DG method to define the local solvers and by taking what could be called the corresponding natural choice for the space M_h for the approximate trace λ_h . For example, we can use the local discontinuous Galerkin (LDG) method to define the local solvers and construct a hybridizable DG method. Surprisingly, it turns out that the resulting DG method is not an LDG method but one of the DG methods considered in [17]; see Corollary 3.2. A similar result holds for the hybridizable DG methods whose local solvers are the interior penalty (IP) method, that is, the resulting method is not the original IP method but the IP-like method considered in [38]; see Corollary 3.4. This is in sharp contrast with the RT, BDM, and CG methods, each of which can be hybridized by using as local solvers the RT, BDM, and CG methods, respectively.

It is interesting to note that the only known DG methods that turn out to be hybridizable by our technique are the following: a subset of the methods considered in [17], the minimal dissipation DG methods considered in [20], the minimal dissipation LDG method analyzed in [24], and the DG method considered in [38] and then rewritten as an IP method in [37]. With the exception of *some* LDG methods, *none* of the DG methods considered in the unified analysis of DG methods carried out in [5] is a hybridizable DG method. The reason is, roughly speaking, as follows. For all methods considered in [5], the variable \mathbf{q}_h is easily eliminated from the equations due to the fact that the numerical trace \hat{u}_h is *independent* of \mathbf{q}_h or $\mathbf{grad}u_h$; a *primal formulation* can then be found solely in terms of u_h . In contrast, in our approach, we eliminate *both* \mathbf{q}_h and u_h from the equations and obtain a formulation in terms of \hat{u}_h only, namely, (1.7). For this, it turns out that we need \hat{u}_h to be *dependent* on \mathbf{q}_h or $\mathbf{grad}u_h$, except for a few special LDG methods.

1.4. Properties of the algebraic system of hybridizable DG methods.

As pointed out above, since the degrees of freedom of the functions μ in the finite element space M_h are associated with the borders of the elements only, the stiffness matrix associated with weak formulation (1.7) of the numerical trace $\hat{u}_h = \lambda_h$ is significantly smaller than the one associated to the original variables (\mathbf{q}_h, u_h) . Moreover, the actual computation of the approximate solution of DG methods becomes competitive with that of hybridized mixed methods. For example, as we show below, on triangulations made of simplexes, the stiffness matrix associated with weak formulation (1.7) of any hybridizable DG method has the *same* size, block structure, and sparsity as the corresponding hybridized BDM [11] and RT [49] mixed methods; see [26] for details. Even more, it was recently proved (see [25, Property (iii) of Theorem 2.4]) that the stiffness matrices of the hybridized BDM and RT meth-

ods and the so-called *single face* hybridizable DG method are, in fact, *identical* provided $d = 0$.

1.5. New automatic coupling of different methods and mortaring techniques. One of the main features of the unified framework is that it allows for a single implementation of a vast class of finite element methods including DG, mixed, non-conforming, and CG methods and for their automatic coupling. Since it can be done even in the presence of nonmatching meshes, the unified framework provides a novel *coupling and mortaring* technique. This induces a *paradigm shift* in the way we view different finite element methods fitting in the framework, especially when considering adaptive algorithms. Indeed, since all these methods can be implemented within a single framework, the issue is now to investigate which method to use in what part of the domain in order to fully exploit its individual advantages. Let us briefly compare our new mortaring technique with the already established ones. Mortaring techniques (see the pioneering work [9]) were introduced to accommodate methods that can be defined in separate subdomains that could have been independently meshed. This technique introduces an auxiliary space for a Lagrange multiplier associated with a continuity constraint on the approximate solution. The resulting system could be written either as a saddle point problem, symmetric but indefinite [8], or as a non-conforming finite element approximation, which leads to a symmetric positive definite system; see, for example, [9, 42]. This classical mortaring is a powerful technique to achieve flexibility in the meshing and the choice of the finite element approximation. The work in this direction also includes coupling of mixed and CG [53], mixed and mixed finite element methods [2, 45], and DG and mixed methods [40].

However, this mortaring approach is very different from ours, since instead of enforcing the continuity of the approximation to u , we enforce a continuity condition on the approximation to the flux \mathbf{q} . The way of coupling and mortaring provided by the unified framework represents a simpler alternative to the above-mentioned mortaring techniques, as well as to earlier works on the coupling of CG and DG methods implicitly contained in [5] and explicitly emphasized in [48], as well as to the coupling of DG and mixed methods introduced in [23] and in [50].

1.6. Devising new methods. The unified framework provides a new point of view for constructing new methods. We provide three main examples of such methods. The first one is a family of methods well suited for *hp*-adaptivity and for dealing with nonmatching meshes. On each element $K \in \mathcal{T}_h$, it uses local solvers obtained from the RT, BDM, LDG, or CG methods by means of a suitable modification of the definition of the numerical trace of the flux of *some faces* of K only. For example, by modifying the numerical trace of the CG-H method on the element faces lying on the non-matching interface, we allow the method to handle nonmatching grids. This method represents an alternative to the coupling of DG and CG methods proposed in [48].

The second example is a variable-degree RT method that can be used on some classes of nonconforming meshes. The third example is called the embedded DG (EDG) method; it was introduced in the setting of shell problems in [43]. An EDG method is obtained from an already existing hybridizable method by simply modifying the space M_h . This capability can be used as a new mortaring technique for dealing with nonmatching meshes, as we are going to see. Moreover, some EDG methods give rise to a stiffness matrix whose size and sparsity is exactly equal to that of the *statically condensed* stiffness matrix of the CG method, while retaining the stabilization mechanisms typical of DG methods; see [43]. As a consequence, EDG methods can immediately be incorporated into existing commercial codes. Related to EDG meth-

ods are the so-called multiscale DG methods [44, 14], which were introduced with a similar intention but a different approach.

1.7. Possibilities and recent developments. The unified framework could be used to establish a single a priori and a single a posteriori error analysis of all the methods fitting in it. It could be used to compare different methods or to establish new relations between them just as the unsuspected relation between the RT and the BDM methods in [26] was recently uncovered by comparing their hybridized versions. The framework could also be used to further explore the relation between mixed and nonconforming methods like the relation between the RT method of lowest order and a nonconforming method established in [4] and exploited in [47]. This work was later generalized in [1], where links between a variety of mixed and nonconforming methods were established; see also the references therein. Finally, the unifying framework can be used to devise new preconditioners based on, for example, substructuring techniques. However, in this paper, none of the above-mentioned issues will be investigated.

On the other hand, several discoveries induced by the unifying framework have already taken place. In particular, new DG methods which are more accurate and efficient than any other known DG method have been uncovered. Indeed, by exploiting the structure of the unified framework, a new DG method called the *single face, hybridizable* (SFH) DG method was constructed, which lies *in between* the RT and BDM methods; see [25]. It is the first known DG method, using polynomials of degree k for both \mathbf{q}_h and u_h , proven to converge with order $k + 1$ in *both* variables; all other DG methods converge with order k in the flux only. Moreover, the SFH method shares with the RT and BDM methods their remarkable superconvergence properties; this allow for the element-by-element computation of a new approximation u_h^* converging with order $k + 2$. These results were then extended to other hybridizable DG methods in [33]. Therein, it was shown that in order to achieve the above-mentioned convergence properties, the interelement jumps of both unknowns have to be penalized essentially in the same way. This goes against the established belief that the interelement jumps of u_h need to be strongly penalized, while the interelement jumps of \mathbf{q}_h need not be.

Also recently, a study of EDG methods obtained from hybridizable DG methods by forcing the numerical trace to be *continuous* has been carried out in [32]. It was proven that these EDG methods *lose* the above-mentioned convergence properties because the numerical trace $\hat{\mathbf{q}}_h$ is not single valued. Moreover, numerical evidence was provided indicating that this loss of accuracy of the EDG method is not compensated by the computational advantage of having a reduced amount of globally coupled degrees of freedom. Hybridizable DG methods, with properly chosen penalization parameters, are thus more efficient than their EDG counterparts.

1.8. Organization of the paper. The paper is organized as follows. In section 2, we describe the general structure of the hybridized finite element methods and prove that the approximate trace λ_h is characterized as the solution of a weak formulation of the form (1.7); see Theorem 2.1. We then provide sufficient conditions for the existence and uniqueness of the solution λ_h ; see Theorem 2.4. Further in this section we give some implementation details and compare the memory requirements of hybridizable methods with those of some classical DG methods. In section 3, we give several examples of hybridizable finite element methods. These include mixed methods using RT and BDM finite element spaces, a large variety of DG, CG, and some nonconforming finite element methods. In section 4, we build on the results of

the previous section and construct the above-mentioned novel hybridizable methods. Finally, in section 5, we conclude the paper with a few extensions and some final remarks.

2. The general framework of hybridization. In this section, we display the structure of hybridized finite element methods for second order elliptic problem (1.1). We begin by presenting the exact definition of the linear forms appearing in the weak formulation of the form (1.7), determining the approximate trace λ_h . We then provide sufficient conditions for the existence and uniqueness of λ_h and show that the assembly of the corresponding matrix equation can be done in a typical finite element fashion. We end by describing the sparsity structure of the stiffness matrix and comparing it with that of the stiffness matrices of the hybridized RT, IP, and LDG methods.

2.1. Notation. We use the notation used in [5]; let us recall it. Let \mathcal{T}_h be a collection of disjoint elements that partition Ω . The shape of the elements is not important in this general framework. Moreover, triangulation \mathcal{T}_h need not be conforming (we say that a triangulation \mathcal{T}_h is conforming if whenever the intersection of the boundaries of any two elements has nonzero $(n - 1)$ -Lebesgue measure, the intersection is a face of each of the elements). So, \mathcal{T}_h can be a collection of simplices, quadrilaterals, cubes, or a mixture of them which are not required to align across element interfaces. An interior “face” of \mathcal{T}_h is any planar set e of positive $(n - 1)$ -dimensional measure of the form $e = \partial K^+ \cap \partial K^-$ for some two elements K^+ and K^- of the collection \mathcal{T}_h . (We use the word “face” even when $n = 2$.) We say that e is a boundary face if there is an element K of \mathcal{T}_h such that $e = \partial K \cap \partial\Omega$ and the $(n - 1)$ -Lebesgue measure of e is not zero. Let \mathcal{E}_h° and \mathcal{E}_h^∂ denote the set of interior and boundary faces of \mathcal{T}_h , respectively. We denote by \mathcal{E}_h the union of all the faces in \mathcal{E}_h° and \mathcal{E}_h^∂ . In all our examples, elements of \mathcal{E}_h° and \mathcal{E}_h^∂ are affine sets, although that is not required for the considerations in this section.

Finite element methods based on the mesh \mathcal{T}_h typically use some finite-dimensional polynomial approximation spaces on each element of \mathcal{T}_h . On an element K , we denote by $\mathbf{V}(K)$ the polynomial space in which the flux \mathbf{q} is approximated and by $W(K)$ the space in which the scalar solution u is approximated. The corresponding global finite element spaces are defined by

$$(2.1) \quad \mathbf{V}_h = \{\mathbf{v} : \mathbf{v}|_K \in \mathbf{V}(K)\} \quad \text{and} \quad W_h = \{w : w|_K \in W(K)\}.$$

On an interior face $e = \partial K^+ \cap \partial K^-$, we consider scalar and vector functions that are, in general, double valued. For any discontinuous (scalar or vector) function q in W_h or \mathbf{V}_h , the trace $q|_e$ is a double-valued function, whose two branches are denoted by $(q|_e)_{K^+}$ and $(q|_e)_{K^-}$. To simplify the notation, we often shorten these to q_{K^+} and q_{K^-} , respectively. These branches are defined by $q_{K^\pm}(\mathbf{x}) = \lim_{\epsilon \downarrow 0} q(\mathbf{x} - \epsilon \mathbf{n}_{K^\pm})$ for all \mathbf{x} in e . Here and elsewhere, \mathbf{n} denotes the double-valued function of unit normals on \mathcal{E}_h , so on any face $e \subseteq \partial K$, \mathbf{n}_K denotes the unit outward normal of K . The same notations are used for vector functions. For any double-valued vector function \mathbf{r} on an interior face e , we define the *jump* of its normal component across the face e by

$$[[\mathbf{r}]]_e := \mathbf{r}_{K^+} \cdot \mathbf{n}_{K^+} + \mathbf{r}_{K^-} \cdot \mathbf{n}_{K^-}.$$

On any face e of K lying on the boundary, we set

$$[[\mathbf{r}]]_e := \mathbf{r}_K \cdot \mathbf{n}_K.$$

To simplify the exposition, we use $\llbracket \mathbf{r} \rrbracket$ to denote the single-valued function on the entire set \mathcal{E}_h , which is equal to $\llbracket \mathbf{r} \rrbracket_e$ on every face $e \in \mathcal{E}_h$. Similarly, for any $e \in \mathcal{E}_h^\circ$, we define

$$\{\{\xi\}\}_e = \frac{1}{2}(\xi_{K^+} + \xi_{K^-}), \quad \{\{\mathbf{q}\}\}_e = \frac{1}{2}(\mathbf{q}_{K^+} + \mathbf{q}_{K^-}), \quad \llbracket \xi \rrbracket_e = \xi_{K^+} \mathbf{n}_{K^+} + \xi_{K^-} \mathbf{n}_{K^-}.$$

For a boundary face e in \mathcal{E}_h^∂ , the operator $\{\{\cdot\}\}_e$ is also considered to be the identity, so that we can put together local operators $\{\{\cdot\}\}_e$ to form a global operator $\{\{\cdot\}\}$ on \mathcal{E}_h , just as we did for $\llbracket \cdot \rrbracket$.

Our notation for inner products is standard: For functions u and v in $L^2(D)$, we write $(u, v)_D = \int_D uv \, dx$ if D is a domain of \mathbb{R}^n and $\langle u, v \rangle_D = \int_D uv \, dx$ if D is a domain of \mathbb{R}^{n-1} . To emphasize the mesh-dependent nature of certain integrals, we introduce the notation

$$(v, w)_{\mathcal{T}_h} = \sum_{K \in \mathcal{T}_h} (v, w)_K \quad \text{and} \quad \langle \mu, \lambda \rangle_{\mathcal{E}} = \sum_{e \in \mathcal{E}} \langle \mu, \lambda \rangle_e$$

for functions v, w and μ, λ defined on Ω and \mathcal{E}_h , respectively. Here \mathcal{E} is any subset of \mathcal{E}_h .

2.2. The general structure of the methods. To describe the structure of the methods fitting in the unified framework, we mimic the characterization of the exact solution given in the Introduction.

Thus, we begin by choosing the space M_h of approximate traces, by taking the approximation to λ, λ_h , in

$$(2.2) \quad M_h := \{\mu \in M_h : \mu = 0 \text{ on } \partial\Omega\}$$

and by setting $g_h = I_h g$, where I_h is a suitably defined interpolation operator with image in M_h . Recall that g is the extension by zero of the Dirichlet data on $\partial\Omega$ to \mathcal{E}_h° ; see (1.2).

Next, we introduce a discrete version of local solvers (1.4a) and (1.4b). The first local solver maps each function \mathbf{m} in M_h to the function $(\mathbf{Qm}, \mathcal{U}\mathbf{m})$ on Ω , whose restriction to any mesh element K is in $\mathbf{V}(K) \times W(K)$ and satisfies the following discretization of (1.4a):

$$(2.3a) \quad (c \mathbf{Qm}, \mathbf{v})_K - (\mathcal{U}\mathbf{m}, \operatorname{div} \mathbf{v})_K = -\langle \mathbf{m}, \mathbf{v} \cdot \mathbf{n} \rangle_{\partial K} \quad \text{for all } \mathbf{v} \in \mathbf{V}(K),$$

$$(2.3b) \quad -(\operatorname{grad} w, \mathbf{Qm})_K + \langle w, \widehat{\mathbf{Qm}} \cdot \mathbf{n} \rangle_{\partial K} + (d \mathcal{U}\mathbf{m}, w)_K = 0 \text{ for all } w \in W(K).$$

Here $\widehat{\mathbf{Qm}}$ represents the numerical trace of the flux, which is, in general, a double-valued function on \mathcal{E}_h° . In inner products involving $\widehat{\mathbf{Qm}}$ over a single simplex boundary ∂K , the integrand is assumed to be branch $(\widehat{\mathbf{Qm}})_K$ from that simplex. In all examples we consider in this paper, numerical flux $\widehat{\mathbf{Qm}}$ is either expressed explicitly in terms of $(\mathbf{Qm}, \mathcal{U}\mathbf{m})$ or is an unknown function. In the examples where the latter case arises, we introduce the space in which the unknown $\widehat{\mathbf{Qm}}$ lies and add new equations to render the resulting formulation uniquely solvable. At this point, however, the precise definition of $\widehat{\mathbf{Qm}}$ is not essential, as we are solely interested in displaying the structure of the method for any $\widehat{\mathbf{Qm}}$. Below, we formally require $\mathbf{m} \mapsto (\mathbf{Qm}, \widehat{\mathbf{Qm}}, \mathcal{U}\mathbf{m})$ to be a well-defined linear map; see Assumption 2.1.

The second local solver is a discretization of the second boundary value problem in (1.4b). It associates to any $f \in L^2(\Omega)$ the pair $(\mathbf{Q}f, \mathcal{U}f)$, whose restriction to each

element K is defined as the function in $\mathbf{V}(K) \times W(K)$ satisfying

(2.4a)

$$(c \mathbf{Q}f, \mathbf{v})_K - (\mathcal{U}f, \operatorname{div} \mathbf{v})_K = 0 \quad \text{for all } \mathbf{v} \in \mathbf{V}(K),$$

(2.4b)

$$-(\mathbf{grad} w, \mathbf{Q}f)_K + \left\langle w, \widehat{\mathbf{Q}}f \cdot \mathbf{n} \right\rangle_{\partial K} + (d \mathcal{U}f, w)_K = (f, w)_K \quad \text{for all } w \in W(K).$$

Just as for the first local solver, we leave undefined the numerical trace $\widehat{\mathbf{Q}}f$.

Obviously, while the functions $(\mathbf{Q}f, \mathcal{U}f)|_K$ and $(\mathbf{Q}\mathbf{m}, \mathcal{U}\mathbf{m})|_K$ are in $\mathbf{V}(K) \times W(K)$, the space in which $\widehat{\mathbf{Q}}f$ and $\widehat{\mathbf{Q}}\mathbf{m}$ lie will vary from example to example. Now we make our assumption about the local solvers.

Assumption 2.1 (existence and uniqueness of the local solvers). For every \mathbf{m} in \mathbf{M}_h , there is a unique set of functions of \mathbf{m} , $(\mathbf{Q}\mathbf{m}, \widehat{\mathbf{Q}}\mathbf{m}, \mathcal{U}\mathbf{m})$ depending linearly on \mathbf{m} and satisfying (2.3). Furthermore, for every f in $L^2(\Omega)$, there is a unique set of functions $(\mathbf{Q}f, \widehat{\mathbf{Q}}f, \mathcal{U}f)$ depending linearly on f and satisfying (2.4).

Each of the methods under consideration define an approximation to (\mathbf{q}, u) ,

$$(2.5) \quad (\mathbf{q}_h, u_h) = (\mathbf{Q}\lambda_h + \mathbf{Q}g_h + \mathbf{Q}f, \mathcal{U}\lambda_h + \mathcal{U}g_h + \mathcal{U}f) \in (\mathbf{V}_h \times W_h),$$

where λ_h is *assumed* to be *determined* by the following discrete version of transmission condition (1.5):

$$(2.6) \quad \left\langle \mu, \left[\widehat{\mathbf{Q}}\lambda_h + \widehat{\mathbf{Q}}g_h + \widehat{\mathbf{Q}}f \right] \right\rangle_{\mathcal{E}_h} = 0 \quad \text{for all } \mu \in M_h.$$

If we define the numerical flux by

$$(2.7) \quad \widehat{\mathbf{q}}_h := \widehat{\mathbf{Q}}\lambda_h + \widehat{\mathbf{Q}}g_h + \widehat{\mathbf{Q}}f,$$

and if the (extension by zero to \mathcal{E}_h of the) function $[[\widehat{\mathbf{q}}_h]]|_{\mathcal{E}_h^\circ}$ belongs to the space M_h , then condition (2.6) is simply stating that $[[\widehat{\mathbf{q}}_h]]|_{\mathcal{E}_h^\circ} = 0$ pointwise, that is, the normal component of the numerical trace $\widehat{\mathbf{q}}_h$ is single valued, or, adopting the terminology of [5], the function $\widehat{\mathbf{q}}_h$ is a *conservative* numerical flux. It is for this reason we call (2.6) the conservativity condition. If the function $[[\widehat{\mathbf{q}}_h]]|_{\mathcal{E}_h^\circ}$ does not belong to the space M_h , the conservativity condition imposes only the weak continuity of the normal component of the numerical trace $\widehat{\mathbf{q}}_h$, which, as a consequence, is not single valued.

It is worth noting that the method just described can be viewed as seeking the approximation $(\mathbf{q}_h, u_h, \lambda_h)$ in $\mathbf{V}_h \times W_h \times M_h$ satisfying

(2.8a)

$$(c \mathbf{q}_h, \mathbf{r})_{\mathcal{T}_h} - (u_h, \operatorname{div} \mathbf{r})_{\mathcal{T}_h} + \sum_{K \in \mathcal{T}_h} \langle \lambda_h, \mathbf{r} \cdot \mathbf{n} \rangle_{\partial K \setminus \partial \Omega} = -\langle g_h, \mathbf{r} \cdot \mathbf{n} \rangle_{\partial \Omega} \quad \text{for all } \mathbf{r} \in \mathbf{V}_h,$$

(2.8b)

$$-(\mathbf{q}_h, \mathbf{grad} w)_{\mathcal{T}_h} + \sum_{K \in \mathcal{T}_h} \langle \widehat{\mathbf{q}}_h \cdot \mathbf{n}, w \rangle_{\partial K} + (d u_h, w)_{\mathcal{T}_h} = (f, w)_{\mathcal{T}_h} \quad \text{for all } w \in W_h,$$

(2.8c)

$$\sum_{K \in \mathcal{T}_h} \langle \mu, \widehat{\mathbf{q}}_h \cdot \mathbf{n} \rangle_{\partial K} = 0 \quad \text{for all } \mu \in M_h.$$

Note that the first two equations are used to define local solvers (2.3) and (2.4), while the last is nothing but conservativity condition (2.6). This type of method is sometimes called a hybrid dual-mixed method. As pointed out in the Introduction, it is

called mixed because we seek approximations for the flux \mathbf{q}_h , as well as the potential u_h , on Ω . It is called hybrid dual because the approximate trace λ_h associated to the conservativity condition is an approximation for the trace of the potential u on the boundaries of the elements.

Many hybridized finite element methods admit this structure. For example, some classic hybridized mixed methods [4, 26] are obtained by an appropriate choice of the local spaces and by choosing $\widehat{\mathbf{Q}}(\cdot)$ in such a way that we have $\widehat{\mathbf{q}}_h = \mathbf{q}_h$. Many DG methods also fall into this form—although not all of them are hybridizable. Indeed, the schemes considered in the unified analysis of DG methods in [5] can be written in our notation as

$$\begin{aligned} (c \mathbf{q}_h, \mathbf{v})_{\mathcal{T}_h} - \sum_{K \in \mathcal{T}_h} (u_h, \operatorname{div} \mathbf{v})_K + \sum_{K \in \mathcal{T}_h} \langle \widehat{u}_h, \mathbf{v} \cdot \mathbf{n} \rangle_{\partial K \setminus \partial \Omega} &= -\langle g_h, \mathbf{v} \cdot \mathbf{n} \rangle_{\partial \Omega}, \\ -(\mathbf{grad} w, \mathbf{q}_h)_{\mathcal{T}_h} + \sum_{K \in \mathcal{T}_h} \langle w, \widehat{\mathbf{q}}_h \cdot \mathbf{n} \rangle_{\partial K} + (d u_h, w)_{\mathcal{T}_h} &= (f, w)_{\mathcal{T}_h}, \end{aligned}$$

where \widehat{u}_h and $\widehat{\mathbf{q}}_h$ are the so-called *numerical traces* of the DG method. Comparing these equations with (2.8) of our general framework, we immediately realize that $\widehat{u}_h = \lambda_h$ on \mathcal{E}_h° . We thus see that, for a finite element method to be hybridizable, its numerical trace \widehat{u}_h *must* be single valued. This implies, in particular, that the DG methods in [5] that are not *adjoint consistent* cannot be hybridized by using our technique. In contrast, the (normal component of the) numerical trace $\widehat{\mathbf{q}}_h$ is *not* required to be single valued, since conservativity condition (2.6) does not always ensure a single-valued numerical trace. Thanks to this flexibility, the CG method and the EDG methods turn out to be hybridizable.

This concludes the description of the general structure of the methods. Methods with this structure include a wide class of DG and hybridized mixed and CG methods, as we show in sections 3, 4, and 5.

2.3. The characterization of the variable λ_h . As we see next, the relevance of the methods fitting the previously described general structure resides in the fact that the λ_h can be characterized in terms of a simple weak formulation in which none of the other variables appear.

THEOREM 2.1. *Suppose Assumption 2.1 on the existence and uniqueness of the local solvers holds. Then $\lambda_h \in M_h$ satisfies conservativity condition (2.6) if and only if it satisfies*

$$(2.9) \quad a_h(\lambda_h, \mu) = b_h(\mu) \quad \text{for all } \mu \in M_h,$$

where

$$\begin{aligned} a_h(\eta, \mu) &= (c \mathbf{Q} \eta, \mathbf{Q} \mu)_{\mathcal{T}_h} + (d \mathcal{U} \eta, \mathcal{U} \mu)_{\mathcal{T}_h} + \left\langle 1, \left[(\mathcal{U} \mu - \mu) (\widehat{\mathbf{Q}} \eta - \mathbf{Q} \eta) \right] \right\rangle_{\mathcal{E}_h}, \\ b_h(\mu) &= \left\langle g_h, \left[\widehat{\mathbf{Q}} \mu \right] \right\rangle_{\mathcal{E}_h} + (f, \mathcal{U} \mu)_{\mathcal{T}_h} - \left\langle 1, \left[(\mathcal{U} \mu - \mu) (\widehat{\mathbf{Q}} f - \mathbf{Q} f) \right] \right\rangle_{\mathcal{E}_h} \\ &\quad + \left\langle 1, \left[\mathcal{U} f (\widehat{\mathbf{Q}} \mu - \mathbf{Q} \mu) \right] \right\rangle_{\mathcal{E}_h} \\ &\quad - \left\langle 1, \left[(\mathcal{U} \mu - \mu) (\widehat{\mathbf{Q}} g_h - \mathbf{Q} g_h) \right] \right\rangle_{\mathcal{E}_h} \\ &\quad + \left\langle 1, \left[(\mathcal{U} g_h - g) (\widehat{\mathbf{Q}} \mu - \mathbf{Q} \mu) \right] \right\rangle_{\mathcal{E}_h} \end{aligned}$$

for all η and $\mu \in M_h$.

Note that, since λ_h is an approximation of the function u on \mathcal{E}_h° , it is natural to expect bilinear form $a_h(\cdot, \cdot)$ to be symmetric. This motivates the following observation. Bilinear form $a_h(\cdot, \cdot)$ is symmetric if and only if numerical trace $\widehat{\mathbf{Q}}\cdot$ is such that

$$(2.10a) \quad \left\langle 1, \left[(\mathcal{U}\mu - \mu) (\widehat{\mathbf{Q}}\eta - \mathbf{Q}\eta) \right] \right\rangle_{\mathcal{E}_h} = \left\langle 1, \left[(\mathcal{U}\eta - \eta) (\widehat{\mathbf{Q}}\mu - \mathbf{Q}\mu) \right] \right\rangle_{\mathcal{E}_h}$$

for all $\eta, \mu \in M_h$. If we also have

$$(2.10b) \quad \left\langle 1, \left[(\mathcal{U}\mu - \mu) (\widehat{\mathbf{Q}}f - \mathbf{Q}f) \right] \right\rangle_{\mathcal{E}_h} = \left\langle 1, \left[\mathcal{U}f (\widehat{\mathbf{Q}}\mu - \mathbf{Q}\mu) \right] \right\rangle_{\mathcal{E}_h},$$

then

$$b_h(\mu) = \left\langle g_h, \left[\widehat{\mathbf{Q}}\mu \right] \right\rangle_{\mathcal{E}_h} + (f, \mathcal{U}\mu)_\Omega.$$

All the examples in this paper satisfy the above symmetry conditions.

Now we prove Theorem 2.1. Set

$$(2.11a) \quad a_h(\lambda_h, \mu) = - \left\langle \mu, \left[\widehat{\mathbf{Q}}\lambda_h \right] \right\rangle_{\mathcal{E}_h},$$

$$(2.11b) \quad b_h(\mu) = \left\langle \mu, \left[\widehat{\mathbf{Q}}g_h + \widehat{\mathbf{Q}}f \right] \right\rangle_{\mathcal{E}_h}$$

so that conservativity condition (2.6) takes the form (2.9). Theorem 2.1 then follows from the following result.

LEMMA 2.2 (elementary identities). *We have, for any $\mathbf{m}, \mu \in \mathbf{M}_h$ and $f \in L^2(\Omega)$,*

$$\begin{aligned} (i) \quad & - \left\langle \mu, \left[\widehat{\mathbf{Q}}\mathbf{m} \right] \right\rangle_{\mathcal{E}_h} = (c \mathbf{Q}\mathbf{m}, \mathbf{Q}\mu)_\Omega + (d \mathcal{U}\mathbf{m}, \mathcal{U}\mu)_\Omega \\ & \quad + \left\langle 1, \left[(\mathcal{U}\mu - \mu) (\widehat{\mathbf{Q}}\mathbf{m} - \mathbf{Q}\mathbf{m}) \right] \right\rangle_{\mathcal{E}_h}, \\ (ii) \quad & - \left\langle \mu, \left[\widehat{\mathbf{Q}}g_h \right] \right\rangle_{\mathcal{E}_h} = - \left\langle g_h, \left[\widehat{\mathbf{Q}}\mu \right] \right\rangle_{\mathcal{E}_h} \\ & \quad + \left\langle 1, \left[(\mathcal{U}\mu - \mu) (\widehat{\mathbf{Q}}g_h - \mathbf{Q}g_h) \right] \right\rangle_{\mathcal{E}_h} \\ & \quad - \left\langle 1, \left[(\mathcal{U}g_h - g_h) (\widehat{\mathbf{Q}}\mu - \mathbf{Q}\mu) \right] \right\rangle_{\mathcal{E}_h}, \\ (iii) \quad & - \left\langle \mu, \left[\widehat{\mathbf{Q}}f \right] \right\rangle_{\mathcal{E}_h} = - (f, \mathcal{U}\mu)_{\mathcal{T}_h} \\ & \quad + \left\langle 1, \left[(\mathcal{U}\mu - \mu) (\widehat{\mathbf{Q}}f - \mathbf{Q}f) \right] \right\rangle_{\mathcal{E}_h} \\ & \quad - \left\langle 1, \left[\mathcal{U}f (\widehat{\mathbf{Q}}\mu - \mathbf{Q}\mu) \right] \right\rangle_{\mathcal{E}_h}. \end{aligned}$$

To prove Lemma 2.2, we need some identities which follow from the equations defining the local solvers by integration by parts.

LEMMA 2.3 (relation between jumps and local residuals). *For any $\mathbf{m}, \mu \in \mathbf{M}_h$, $f \in L^2(\Omega)$, $\mathbf{v} \in \mathbf{V}_h$, and $w \in W_h$, the following identities hold:*

$$(2.12a) \quad (c \mathbf{Q}\mathbf{m} + \mathbf{grad} \mathcal{U}\mathbf{m}, \mathbf{v})_{\mathcal{T}_h} = + \left\langle 1, \left[(\mathcal{U}\mathbf{m} - \mathbf{m}) \mathbf{v} \right] \right\rangle_{\mathcal{E}_h},$$

$$(2.12b) \quad (\operatorname{div} \mathbf{Q}\mathbf{m} + d \mathcal{U}\mathbf{m}, w)_{\mathcal{T}_h} = - \left\langle 1, \left[w (\widehat{\mathbf{Q}}\mathbf{m} - \mathbf{Q}\mathbf{m}) \right] \right\rangle_{\mathcal{E}_h},$$

$$(2.12c) \quad (c \mathbf{Q}f + \mathbf{grad} \mathcal{U}f, \mathbf{v})_{\mathcal{T}_h} = + \left\langle 1, \left[\mathcal{U}f \mathbf{v} \right] \right\rangle_{\mathcal{E}_h},$$

$$(2.12d) \quad (\operatorname{div} \mathbf{Q}f + d \mathcal{U}f - f, w)_{\mathcal{T}_h} = - \left\langle 1, \left[w (\widehat{\mathbf{Q}}f - \mathbf{Q}f) \right] \right\rangle_{\mathcal{E}_h}.$$

Using these identities, we now prove Lemma 2.2.

Proof. Let us prove identity (i) of Lemma 2.2. We have

$$\begin{aligned}
 -\langle \mu, \llbracket \widehat{\mathbf{Q}}\mathbf{m} \rrbracket \rangle_{\mathcal{E}_h} &= -\langle \mu, \llbracket \mathbf{Q}\mathbf{m} \rrbracket \rangle_{\mathcal{E}_h} - \langle \mu, \llbracket (\widehat{\mathbf{Q}}\mathbf{m} - \mathbf{Q}\mathbf{m}) \rrbracket \rangle_{\mathcal{E}_h} \\
 &= (c \mathbf{Q}\mu, \mathbf{Q}\mathbf{m})_{\mathcal{T}_h} - (\mathcal{U}\mu, \operatorname{div} \mathbf{Q}\mathbf{m})_{\mathcal{T}_h} \\
 &\quad - \langle \mu, \llbracket (\widehat{\mathbf{Q}}\mathbf{m} - \mathbf{Q}\mathbf{m}) \rrbracket \rangle_{\mathcal{E}_h} \quad \text{by (2.3a),} \\
 &= (c \mathbf{Q}\mu, \mathbf{Q}\mathbf{m})_{\mathcal{T}_h} + (d \mathcal{U}\mathbf{m}, \mathcal{U}\mu)_{\mathcal{T}_h} \\
 &\quad + \langle 1, \llbracket \mathcal{U}\mu (\widehat{\mathbf{Q}}\mathbf{m} - \mathbf{Q}\mathbf{m}) \rrbracket \rangle_{\mathcal{E}_h} \\
 &\quad - \langle \mu, \llbracket (\widehat{\mathbf{Q}}\mathbf{m} - \mathbf{Q}\mathbf{m}) \rrbracket \rangle_{\mathcal{E}_h} \quad \text{by (2.12b).}
 \end{aligned}$$

This proves identity (i) of Lemma 2.2.

Now we prove identity (ii) of Lemma 2.2. To do that, note that, by identity (i) of Lemma 2.2, the bilinear form

$$B(\mathbf{m}, \mu) = \langle \mu, \llbracket \widehat{\mathbf{Q}}\mathbf{m} \rrbracket \rangle_{\mathcal{E}_h} + \langle 1, \llbracket (\mathcal{U}\mu - \mu) (\widehat{\mathbf{Q}}\mathbf{m} - \mathbf{Q}\mathbf{m}) \rrbracket \rangle_{\mathcal{E}_h}$$

is symmetric. As a consequence, identity (ii) of Lemma 2.2 follows from equality $B(\mu, g_h) = B(g_h, \mu)$.

Finally, we prove identity (iii) of Lemma 2.2. We have

$$\begin{aligned}
 -\langle \mu, \llbracket \widehat{\mathbf{Q}}f \rrbracket \rangle_{\mathcal{E}_h} &= -\langle \mu, \llbracket \mathbf{Q}f \rrbracket \rangle_{\mathcal{E}_h} - \langle \mu, \llbracket (\widehat{\mathbf{Q}}f - \mathbf{Q}f) \rrbracket \rangle_{\mathcal{E}_h} \\
 &= (c \mathbf{Q}\mu, \mathbf{Q}f)_{\mathcal{T}_h} - (\mathcal{U}\mu, \operatorname{div} \mathbf{Q}f)_{\mathcal{T}_h} \\
 &\quad - \langle \mu, \llbracket (\widehat{\mathbf{Q}}f - \mathbf{Q}f) \rrbracket \rangle_{\mathcal{E}_h} \quad \text{by (2.3a),} \\
 &= -(f, \mathcal{U}\mu)_{\mathcal{T}_h} + (c \mathbf{Q}\mu, \mathbf{Q}f)_{\mathcal{T}_h} + (d \mathcal{U}\mu, \mathcal{U}f)_{\mathcal{T}_h} \\
 &\quad + \langle 1, \llbracket (\mathcal{U}\mu - \mu) (\widehat{\mathbf{Q}}f - \mathbf{Q}f) \rrbracket \rangle_{\mathcal{E}_h} \quad \text{by (2.12d),} \\
 &= -(f, \mathcal{U}\mu)_{\mathcal{T}_h} + (\operatorname{div} \mathbf{Q}\mu, \mathcal{U}f)_{\mathcal{T}_h} + (d \mathcal{U}\mu, \mathcal{U}f)_{\mathcal{T}_h} \\
 &\quad + \langle 1, \llbracket (\mathcal{U}\mu - \mu) (\widehat{\mathbf{Q}}f - \mathbf{Q}f) \rrbracket \rangle_{\mathcal{E}_h} \quad \text{by (2.4a),} \\
 &= -(f, \mathcal{U}\mu)_{\mathcal{T}_h} - \langle 1, \llbracket \mathcal{U}f (\widehat{\mathbf{Q}}\mu - \mathbf{Q}\mu) \rrbracket \rangle_{\mathcal{E}_h} \\
 &\quad + \langle 1, \llbracket (\mathcal{U}\mu - \mu) (\widehat{\mathbf{Q}}f - \mathbf{Q}f) \rrbracket \rangle_{\mathcal{E}_h} \quad \text{by (2.12b).}
 \end{aligned}$$

This completes the proof of Lemma 2.2. \square

2.4. Sufficient conditions for the existence and uniqueness of λ_h . Next, we provide two conditions which are sufficient for the existence and uniqueness of λ_h . The first is a condition on the local solvers, and the second is a condition on the relation between the local solvers, on each element K of triangulation \mathcal{T}_h and the global space \mathbf{M}_h of approximate traces. It is worth emphasizing that, by guaranteeing the existence and uniqueness of λ_h , these simple conditions ensure the *automatic* coupling of the different local solvers even across nonmatching meshes. Note that no explicit conditions on triangulation \mathcal{T}_h are involved in these conditions.

Assumption 2.2 (on the positive semidefiniteness of the local solvers). The local solvers and the numerical flux traces in (2.3) and (2.4) are such that, for every $K \in \mathcal{T}_h$, the following holds:

$$(2.13a) \quad - \left\langle \mu, \widehat{\mathbf{Q}}\mu \cdot \mathbf{n} \right\rangle_{\partial K} \geq 0 \quad \text{for all } \mu \in M_h.$$

Moreover, there exists a space $M(\partial K)$ containing the set $\{\nu : \nu|_e \in \mathcal{P}_0(e) \text{ on each face } e \in \mathcal{E}_h^\circ \text{ lying on } \partial K\}$ such that

$$(2.13b) \quad \text{if } \left\langle \mu, \widehat{\mathbf{Q}}\mu \cdot \mathbf{n} \right\rangle_{\partial K} = 0 \text{ for some } \mu \in M_h, \text{ then } P_{\partial K}\mu = C_K$$

for some constant C_K , where $P_{\partial K}$ is the $L^2(\partial K)$ -orthogonal projection onto $M(\partial K)$.

Note that auxiliary space $M(\partial K)$ is not necessarily finite-dimensional. Its use is only theoretical; it is not used in practice in any way.

Let us argue that (2.13) is a reasonable condition on the positive semidefiniteness of the bilinear forms corresponding to the local solvers. Indeed, taking $\mathbf{v} := \mathbf{Q}\mu$ in (2.3a), $\mathbf{m} := \mu$ and $w := \mathcal{U}\mathbf{m}$ in (2.3b), and adding the equations, we get

$$(2.14) \quad \begin{aligned} - \left\langle \mathbf{m}, \widehat{\mathbf{Q}}\mu \cdot \mathbf{n} \right\rangle_{\partial K} &= (c \mathbf{Q}\mathbf{m}, \mathbf{Q}\mu)_K + (d \mathcal{U}\mathbf{m}, \mathcal{U}\mu)_K + \left\langle \left(\widehat{\mathbf{Q}}\mathbf{m} - \mathbf{Q}\mathbf{m} \right) \cdot \mathbf{n}, \mathcal{U}\mu - \mu \right\rangle_{\partial K} \\ &=: a_{h,K}(\mathbf{m}, \mu). \end{aligned}$$

Thus, (2.13a) ensures that bilinear form $a_{h,K}(\cdot, \cdot)$, which coincides with form $a_h(\cdot, \cdot)$ when Ω is single element K , is positive semidefinite. Further, condition (2.13b) states that those functions $\mathbf{m} \in M_h$ for which $a_{h,K}(\mathbf{m}, \mathbf{m}) = 0$ yield constants under an appropriate projection. This is a reasonable assumption, since it is a discrete version of a similar property of the exact solution. Indeed, for the exact solution, such a condition readily implies that $\mathbf{Q}\mathbf{m} = 0$ and, by (1.4a), that $\mathbf{m} = \mathcal{U}\mathbf{m} = \text{constant}$ on ∂K .

This argument suggests that it is reasonable to expect projection $P_{\partial K}$ to be strongly related to the *identity*, at least in parts of ∂K . The following assumption captures this property. It will allow us to establish a link between the different local solvers and, in so doing, to ensure the uniqueness of the solution of (1.7).

Assumption 2.3 (the “gluing condition”). *If $\mu \in M_h$, then on every interior face $e = \partial K^+ \cap \partial K^-$, either $\mu = P_{\partial K^+}\mu$ or $\mu = P_{\partial K^-}\mu$.*

We are now ready to state our result.

THEOREM 2.4 (existence and uniqueness of λ_h). *If Assumption 2.1 on the existence and the uniqueness of the local solvers, Assumption 2.2 on the positive semidefiniteness of the local solvers, and Assumption 2.3, the gluing condition, hold, then there is a unique solution λ_h of weak formulation (2.9).*

Proof. By Theorem 2.1, Assumption 2.1 guarantees the existence and the uniqueness of $\widehat{\mathbf{Q}}\lambda_h$. Therefore, system (2.9) is well defined. Since it is a square system, to prove the existence and the uniqueness of its solution, it is enough to show that if $a_h(\mu, \mu) = 0$ for some $\mu \in M_h$, we have that $\mu = 0$.

By Lemma 2.2,

$$a_h(\mu, \mu) = - \left\langle \mu, \left[\widehat{\mathbf{Q}}\mu \right] \right\rangle_{\mathcal{E}_h} = - \sum_{K \in \mathcal{T}_h} \left\langle \mu, \widehat{\mathbf{Q}}\mu \cdot \mathbf{n} \right\rangle_{\partial K}.$$

Now, since $a_h(\mu, \mu) = 0$, by (2.13a) of Assumption 2.2 on the positive semidefiniteness of the local solvers, each of the summands on the right-hand side must vanish. Thus,

$$\langle \mu, \widehat{\mathbf{Q}}\mu \cdot \mathbf{n} \rangle_{\partial K} = 0 \quad \text{for all } K \in \mathcal{T}_h.$$

By condition (2.13b), on any interior face $e = \partial K^+ \cap \partial K^-$, this implies

$$C_{K^+} = P_{\partial K^+}\mu = \frac{1}{|e|} \langle \mu, 1 \rangle_e = P_{\partial K^-}\mu = C_{K^-},$$

and by Assumption 2.3 (the gluing condition), we conclude that $C_{K^+} = \mu = C_{K^-}$ on the face e . This means that μ is a constant on \mathcal{E}_h . Since $\mu = 0$ on $\partial\Omega$, we see that μ is identically equal to zero on \mathcal{E}_h . This completes the proof. \square

2.5. The sparsity structure of the stiffness matrix for λ_h . Next, we comment on the sparsity structure of the stiffness matrix associated with weak formulation (1.7). For any given basis of the space of approximate traces M_h , we denote by $[\mu]$ the corresponding vector of coefficients of the representation of μ in a given basis of M_h . Then, weak formulation (2.9)

$$A[\lambda_h] = b,$$

where

$$[\mu]^t A[\lambda_h] = a_h(\lambda_h, \mu) \quad \text{and} \quad [\mu]^t b = b_h(\mu).$$

Now, by (2.11),

$$a_h(\eta, \mu) = - \sum_{K \in \mathcal{T}_h} \langle \mu, \widehat{\mathbf{Q}}\eta \cdot \mathbf{n} \rangle_{\partial K} \quad \text{and} \quad b_h(\mu) = \sum_{K \in \mathcal{T}_h} \langle \mu, (\widehat{\mathbf{Q}}f + \widehat{\mathbf{Q}}g_h) \cdot \mathbf{n} \rangle_{\partial K},$$

we have that

$$A = \sum_{K \in \mathcal{T}_h} A_K \quad \text{and} \quad b = \sum_{K \in \mathcal{T}_h} b_K,$$

where A_K and b_K are defined by

$$[\mu]^t A_K[\eta] = - \langle \mu, \widehat{\mathbf{Q}}\eta \cdot \mathbf{n} \rangle_{\partial K} \quad \text{and} \quad [\mu]^t b_K = \langle \mu, (\widehat{\mathbf{Q}}f + \widehat{\mathbf{Q}}g_h) \cdot \mathbf{n} \rangle_{\partial K}.$$

Thus, the matrix equations for the multiplier can be obtained in a typical finite element manner. Moreover, the sparsity of the matrices A_K and b_K can be deduced from the following result.

PROPOSITION 2.1. *Suppose Assumption 2.1 on the existence and the uniqueness of the local solvers holds. Then*

- (i) *if the support of $\mu \in M_h$ does not intersect ∂K , we have that $[\mu]^t b_K = 0$;*
- (ii) *if the support of $\mu \in M_h$ or the support of $\eta \in M_h$ does not intersect ∂K , we have that $[\mu]^t A_K[\eta] = 0$.*

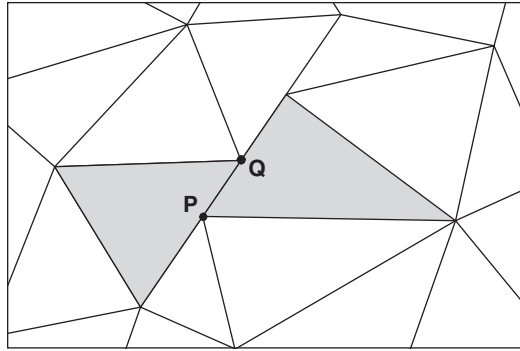


FIG. 2.1. Interior edge $e = \mathbf{PQ}$ and the support of local solver $(\mathcal{Q}\mathbf{m}, \mathcal{U}\mathbf{m})$ for any \mathbf{m} supported on e . Numerical trace $(\widehat{\mathcal{Q}\mathbf{m}})_K$ is generally nontrivial on the boundary of the two shadowed triangles K , but it vanishes on the boundary of other triangles.

Proof. That $[\mu]^t b_K = 0$ and $[\mu]^t A_K[\eta] = 0$ if the support of μ does not intersect ∂K follows immediately from the definition of b_K and A_K . Let us show that $[\mu]^t A_K[\eta] = 0$ if the support of η does not intersect ∂K . Since we are assuming that the local solvers are well defined, if the support of η does not intersect ∂K , we have, by Assumption 2.1, that $(\widehat{\mathcal{Q}\eta})_K = 0$ on ∂K , and the result follows. This completes the proof. \square

We emphasize that this result, illustrated in Figure 2.1, is possible due to the fact that numerical trace $\widehat{\mathcal{Q}\cdot}$ is *double valued* on all interior faces $e \in \mathcal{E}_h^\circ$. Indeed, take η as in the above proof and further assume that its support intersects $\partial K'$, where the intersection of ∂K and $\partial K'$ is a face e in \mathcal{E}_h° . Then $(\widehat{\mathcal{Q}\eta})_{K'}$ can be nontrivial on e , in general. However, this does not contradict the fact that $(\widehat{\mathcal{Q}\eta})_K = 0$ on e because the function $\widehat{\mathcal{Q}\eta}$ is double valued on e .

In the remainder of this subsection, we compare the number of globally coupled degrees of freedom and the number of nonzero entries of the stiffness matrix, restricting our attention to the case of a conforming triangulation \mathcal{T}_h (no hanging nodes). First, consider the case in which $\mathbf{M}_h := \mathcal{M}_{h,k}^c$, where

$$\mathcal{M}_{h,k}^c := \{ \mu \in \mathcal{C}(\mathcal{E}_h) : \mu|_e \in \mathcal{P}_k(e) \text{ for all faces } e \in \mathcal{E}_h \}.$$

Here, $\mathcal{C}(\mathcal{E}_h)$ denotes the space of continuous functions on \mathcal{E}_h and $\mathcal{P}_k(D)$ the set of polynomials of degree at most k on a domain D . Then the sparsity structure of the matrix A is exactly that of the statically condensed stiffness matrix of a CG method using approximations whose restriction to each simplex K is in $\mathcal{P}_k(K)$.

If, instead, we take $\mathbf{M}_h := \mathcal{M}_{h,k}$, where

$$(2.15) \quad \mathcal{M}_{h,k} = \{ \mu \in L^2(\mathcal{E}_h) : \mu|_e \in \mathcal{P}_k(e) \text{ for all faces } e \in \mathcal{E}_h^\circ \},$$

then by choosing basis functions whose support is always contained in a single face, we obtain matrix A , which has a block structure with square blocks of order equal to the dimension of $\mathcal{P}_k(e)$. The number of block rows and block columns is equal to the number of interior faces of triangulation $N_{i,f}$, and, on each block row, there are at most $(2n + 1)$ blocks that are not equal to zero. In other words, the size and sparsity structure of matrix A is precisely that of the stiffness matrix for the hybridized RT method using M_h as space of approximate traces; see [26]. This means that the order

TABLE 2.1

Comparison between hybridizable DG methods and two typical DG methods on simplicial meshes.

n	k	$R_{\text{d.o.f.}}$			n	k	$R_{\text{d.o.f.}}$		
		$R_{\text{d.o.f.}}$	R_{sparsity}	R_{sparsity}			$R_{\text{d.o.f.}}$	R_{sparsity}	R_{sparsity}
		$R_{\text{d.o.f.}}$					$R_{\text{d.o.f.}}$		
		R_{sparsity}					R_{sparsity}		
		IP	LDG		IP	LDG	IP	LDG	
2	1	1.00	1.20	3.00	3	1	0.67	0.63	2.16
	2	1.33	2.13	5.33	2	2	0.83	0.99	3.37
	3	1.67	3.33	8.33	3	3	1.00	1.42	4.86
	4	2.00	4.80	12.00	4	4	1.17	1.94	6.61

of matrix A , which is equal to the number of degrees of freedom of λ_h , is given by

$$N_{\text{d.o.f.}} = N_{\text{i.f.}} \dim \mathcal{P}_k(e)$$

and that the number of possibly nonvanishing entries of A is bounded by

$$N_{\text{sparsity}} = N_{\text{i.f.}} (2n + 1) (\dim \mathcal{P}_k(e))^2.$$

Let us now compare the size and sparsity structure of this stiffness matrix with those of the IP and the (Schur-complement matrix of the) LDG methods that use polynomials of degree k . The number of globally coupled degrees of freedom for both methods is

$$N_{\text{d.o.f.}}^{\text{IP}} = N_{\text{d.o.f.}}^{\text{LDG}} = N_s \dim \mathcal{P}_k(K),$$

where N_s denotes the number of simplexes of the triangulation. Moreover, the stiffness matrices in question have a block structure with square blocks of order equal to the dimension of $\mathcal{P}_k(K)$. On each block-row, the number of blocks that are not equal to zero are at most $(n+2)$ for the IP method and $((n+1)^2+1)$ for the LDG method; recall that, for the LDG method, the degrees of freedom of the neighbors of the neighbors are also involved. This means that the number of nonzero entries of the corresponding stiffness matrices are (bounded by)

$$N_{\text{sparsity}}^{\text{IP}} = N_s (n + 2) (\dim \mathcal{P}_k(K))^2, \quad N_{\text{sparsity}}^{\text{LDG}} = N_s ((n + 1)^2 + 1) (\dim \mathcal{P}_k(K))^2.$$

To compare with the hybridized methods, we consider the ratio of the number of globally coupled degrees of freedom $R_{\text{d.o.f.}}^{\text{DG}} := N_{\text{d.o.f.}}^{\text{DG}}/N_{\text{d.o.f.}}$ and the ratio of the number of entries different from zero $R_{\text{sparsity}}^{\text{IP}} := N_{\text{sparsity}}^{\text{IP}}/N_{\text{sparsity}}$ and $R_{\text{sparsity}}^{\text{LDG}} := N_{\text{sparsity}}^{\text{LDG}}/N_{\text{sparsity}}$. Since $N_s/N_{\text{i.f.}} \approx 2/(n+1)$ (up to a lower order term related to the faces on the boundary), then

$$R_{\text{sparsity}}^{\text{IP}} = \frac{2(n+2)}{(n+1)(2n+1)} \left(\frac{k}{n} + 1\right)^2, \quad R_{\text{sparsity}}^{\text{LDG}} = \frac{2((n+1)^2+1)}{(n+1)(2n+1)} \left(\frac{k}{n} + 1\right)^2.$$

In Table 2.1, we see that in two- or three-space dimensions, the hybridizable methods always have less degrees of freedom and have a stiffness matrix that is sparser than the corresponding LDG methods. The same is valid for the IP method in two-space dimensions and in three-space dimensions for $k \geq 3$. In three-space dimensions, the IP method with $k = 1$ is more advantageous than the corresponding hybridizable DG method; for $k = 2$, its advantages are, however, marginal.

It is interesting to extend the comparison with the IP method for which static condensation of the interior degrees of freedom has been carried out; of course, this

TABLE 2.2

Comparison between hybridizable and the statically-condensed IP methods on simplicial meshes.

n	k	$R_{\text{d.o.f.}}$	R_{sparsity}	n	k	$R_{\text{d.o.f.}}$	R_{sparsity}
2	3	1.50	2.70	3	4	1.13	1.86
	4	1.60	3.07		5	1.23	2.19
	5	1.67	3.33		6	1.32	2.49

can be done only if $k \geq n + 1$. In this case, the number of globally coupled degrees of freedom is

$$N_{\text{d.o.f.}}^{\text{sc-IP}} = N_s (\dim \mathcal{P}_k(K) - \dim \mathcal{P}_{k-n-1}(K)).$$

The stiffness matrix in question has again a block structure with square blocks of order equal to $(\dim \mathcal{P}_k(K) - \dim \mathcal{P}_{k-n-1}(K))$. On each block-row, the number of blocks that are not equal to zero are $n + 2$. Indeed, it can be shown that the interior degrees of freedom on a given simplex can be expressed in terms of the condensed degrees of freedom of the simplex and those of its neighbors, and that the condensed degrees of freedom can be expressed in terms of the interior degrees of freedom of the simplex *and* those of its neighbors. We then have

$$N_{\text{sparsity}}^{\text{sc-IP}} = N_s (n + 2) (\dim \mathcal{P}_k(K) - \dim \mathcal{P}_{k-n-1}(K))^2.$$

This implies that the corresponding ratios are

$$R_{\text{d.o.f.}}^{\text{sc-IP}} = \frac{2}{(n + 1)} \left(\frac{k}{n} + 1 \right) \left(1 - \prod_{j=1}^n \frac{k - j}{k + j} \right),$$

and

$$R_{\text{sparsity}}^{\text{sc-IP}} = \frac{2(n + 2)}{(n + 1)(2n + 1)} \left(\frac{k}{n} + 1 \right)^2 \left(1 - \prod_{j=1}^n \frac{k - j}{k + j} \right)^2.$$

We show some results in Table 2.2. We see that the hybridized methods produce smaller and more sparse matrices than the statically-condensed IP method.

The same argument could be made for DG methods on n -dimensional rectangular finite elements. In this case, the DG approximations could be based on polynomials of degree k (instead of polynomials of degree k in each variable in the case of continuous elements). Then the ratio between the degrees of freedom (and the sparsity) will be lower, since instead of the factor $N_s/N_{\text{i.f.}} \approx 2/(n + 1)$, we have the factor $N_r/N_{\text{i.f.}} \approx 2/2^n$.

A complete comparison of methods would require factoring in the costs of solving the algebraic problem. While greater sparsity or lesser number of degrees of freedom often yields faster solution methods, definitive conclusions can be made only after numerical experiments with specific direct or iterative methods; see [16] for such studies on older methods.

3. Examples of hybridizable methods. In this section, we give several examples of methods fitting the general structure described in the previous section. We restrict ourselves to methods that use the same local solver in all the elements K of triangulation \mathcal{T}_h . Throughout this section, we *assume* that \mathcal{T}_h is a *conforming simplicial triangulation*.

To define each of the methods, we have only to specify (1) the numerical trace of the flux $\widehat{\mathbf{Q}}$, (2) the local spaces $\mathbf{V}(K)$, $W(K)$, and (3) the space of approximate traces M_h . We then verify that the local solvers are well posed and discuss the conservativity condition by using Theorem 2.1. We use Theorem 2.4 to verify the existence and the uniqueness of the approximate trace λ_h and end by relating these results to relevant, earlier material.

Our examples are summarized Tables 3.1 and 3.2; some of them are schematically related in Figure 3.1. The first column of the tables consists of method names. We adopt the following convention: Suppose that we define the local solver on each element by using a numerical method previously known as the “N” method. Then we call the resulting hybridized formulation an “N-hybridizable method” or, in short, an “N-H” method. For example, if we use the well-known IP method to define the local solvers, then any hybridized formulation with such local solvers is denoted as IP-H. We also say that a finite element method is an N-H method if there is a hybridization of the method that is an N-H method.

In columns 2–4 of Table 3.1, we give the spaces of the local solvers and the approximate trace. In the fifth column, we indicate whether the method gives a single-valued flux trace $\widehat{\mathbf{q}}_h$ so the conservativity condition is satisfied in a strong form or $\widehat{\mathbf{q}}_h$ is double-valued so the methods leads to a weak conservativity condition. In the last two columns of Table 3.1, we define the numerical traces of the fluxes $\widehat{\mathbf{Q}}\mathbf{m}$ and $\widehat{\mathbf{Q}}f$. The weak formulations for the approximate traces obtained via Theorem 2.1 for each type of method are listed in Table 3.2.

3.1. The RT-H method. This method is obtained by using the RT method to define the local solvers. The three ingredients of the RT-H method are as follows:

1. For each $K \in \mathcal{T}_h$, we take

$$\widehat{\mathbf{Q}}\mathbf{m} = \mathbf{Q}\mathbf{m}, \quad \widehat{\mathbf{Q}}f = \mathbf{Q}f \quad \text{on } \partial K;$$

2. The finite element space $\mathbf{V}(K) \times W(K)$ is defined as Raviart–Thomas space of degree k :

$$\mathbf{V}(K) = \mathcal{P}_k(K)^n + \mathbf{x} \mathcal{P}_k(K), \quad W(K) = \mathcal{P}_k(K), \quad k \geq 0,$$

where $\mathcal{P}_k(K)^n$ denotes the set of vector functions whose components are in $\mathcal{P}_k(K)$;

3. We define the space of approximate traces as

$$M_h = \mathcal{M}_{h,k}.$$

The fact that the local solvers are well defined can be established by realizing that they are defined by using exactly the RT mixed finite element method. Indeed, if we insert the expression of numerical traces $\widehat{\mathbf{Q}}\mathbf{m}$ and $\widehat{\mathbf{Q}}f$ into the equations defining the local solvers, we see that they are nothing but the RT discretizations of exact local problems (1.4), as claimed. Since the RT method is well defined (see [49, 12]) local solvers $(\mathbf{Q}\mathbf{m}, \mathcal{U}\mathbf{m})$ and $(\mathbf{Q}f, \mathcal{U}f)$ are also well defined.

Note that conservativity condition (2.6) forces numerical trace $\widehat{\mathbf{q}}_h$ to be single valued. Indeed, because (extension by zero from \mathcal{E}_h° to \mathcal{E}_h of) $[\widehat{\mathbf{Q}}\lambda_h + \widehat{\mathbf{Q}}g_h + \widehat{\mathbf{Q}}f]$ and test functions μ belong to the same space, conservativity condition (2.6) forces equality

$$[\widehat{\mathbf{q}}_h] = [\mathbf{q}_h] = \llbracket \widehat{\mathbf{Q}}\lambda_h + \widehat{\mathbf{Q}}g_h + \widehat{\mathbf{Q}}f \rrbracket = 0 \quad \text{on } \mathcal{E}_h^\circ,$$

TABLE 3.1
Summary of the examples.

Method	$V(K)$	$W(K)$	M_h	Conservativity	$\widehat{\mathbf{Q}}_m$	$\widehat{\mathbf{Q}}_f$
RT-H	$\mathcal{P}_k(K)^n + \alpha \mathcal{P}_k(K)$	$\mathcal{P}_k(K)$	$\mathcal{M}_{h,k}$	strong	\mathbf{Q}_m	\mathbf{Q}_f
BDM-H	$\mathcal{P}_k(K)^n$	$\mathcal{P}_{k-1}(K)$	$\mathcal{M}_{h,k}$	strong	\mathbf{Q}_m	\mathbf{Q}_f
LDG-H	$\mathcal{P}_k(K)^n$	$\mathcal{P}_{k-1}(K)$	$\mathcal{M}_{h,k}$	strong	$\mathbf{Q}_m + \tau(\mathcal{U}m - m)n$	$\mathbf{Q}_f + \tau(\mathcal{U}f)n$
LDG-H	$\mathcal{P}_k(K)^n$	$\mathcal{P}_k(K)$	$\mathcal{M}_{h,k}$	strong	$\mathbf{Q}_m + \tau(\mathcal{U}m - m)n$	$\mathbf{Q}_f + \tau(\mathcal{U}f)n$
LDG-H	$\mathcal{P}_{k-1}(K)^n$	$\mathcal{P}_k(K)$	$\mathcal{M}_{h,k}$	strong	$\mathbf{Q}_m + \tau(\mathcal{U}m - m)n$	$\mathbf{Q}_f + \tau(\mathcal{U}f)n$
IP-H	$\mathcal{P}_k(K)^n$	$\mathcal{P}_k(K)$	$\mathcal{M}_{h,k}$	strong	$-\text{ograd } \mathcal{U}m + \tau(\mathcal{U}m - m)n$	$-\text{ograd } \mathcal{U}f + \tau(\mathcal{U}f)n$
NC-H	$\mathcal{P}_{k-1}(K)^2, k \text{ odd}$	$\mathcal{P}_k(K)$	$\mathcal{M}_{h,k-1}$	strong	a new unknown variable	a new unknown variable
CG-H	$\mathcal{P}_{k-1}(K)^n$	$\mathcal{P}_k(K)$	$\mathcal{M}_{h,k}^c$	weak	a new unknown variable	a new unknown variable

TABLE 3.2
Weak formulations for the approximate trace.

Method	$a_h(\eta, \mu)$	$b_h(\mu)$
RT-H	$(c \mathbf{Q}\eta, \mathbf{Q}\mu)_{\mathcal{T}_h} + (d \mathcal{U}\eta, \mathcal{U}\mu)_{\mathcal{T}_h}$	$(f, \mathcal{U}\mu)_{\mathcal{T}_h} + \langle g_h, \widehat{\mathbf{Q}}\mu \cdot \mathbf{n} \rangle_{\partial\Omega}$
BDM-H	$(c \mathbf{Q}\eta, \mathbf{Q}\mu)_{\mathcal{T}_h} + (d \mathcal{U}\eta, \mathcal{U}\mu)_{\mathcal{T}_h}$	$(f, \mathcal{U}\mu)_{\mathcal{T}_h} + \langle g_h, \mathbf{Q}\mu \cdot \mathbf{n} \rangle_{\partial\Omega}$
LDG-H	$(c \mathbf{Q}\eta, \mathbf{Q}\mu)_{\mathcal{T}_h} + (d \mathcal{U}\eta, \mathcal{U}\mu)_{\mathcal{T}_h} + \langle 1, [(\mathcal{U}\mu - \mu)(\tau(\mathcal{U}\eta - \eta)\mathbf{n})] \rangle_{\varepsilon_h}$	$(f, \mathcal{U}\mu)_{\mathcal{T}_h} + \langle g_h, \mathbf{Q}\mu \cdot \mathbf{n} + \tau \mathcal{U}\mu \rangle_{\partial\Omega}$
IP-H†	$(\text{ograd } \mathcal{U}\mu, \text{grad } \mathcal{U}\eta)_{\mathcal{T}_h} + (d \mathcal{U}\eta, \mathcal{U}\mu)_{\mathcal{T}_h} + \langle 1, [(\eta - \mathcal{U}\eta) \text{ograd } \mathcal{U}\mu + (\mu - \mathcal{U}\mu) \text{egrad } \mathcal{U}\eta] \rangle_{\varepsilon_h} + \langle 1, [(\mathcal{U}\mu - \mu)(\tau(\mathcal{U}\eta - \eta)\mathbf{n})] \rangle_{\varepsilon_h}$	$(f, \mathcal{U}\mu)_{\mathcal{T}_h} + \langle g_h, -\text{ograd } \mathcal{U}\mu \cdot \mathbf{n} + \tau \mathcal{U}\mu \rangle_{\partial\Omega}$
NC-H†	$(\text{ograd } \mathcal{U}\eta, \text{grad } \mathcal{U}\mu)_{\mathcal{T}_h} + (d \mathcal{U}\eta, \mathcal{U}\mu)_{\mathcal{T}_h}$	$(f, \mathcal{U}\mu)_{\mathcal{T}_h} + \langle g_h, \widehat{\mathbf{Q}}\mu \cdot \mathbf{n} \rangle_{\partial\Omega}$
CG-H†	$(\text{ograd } \mathcal{U}\eta, \text{grad } \mathcal{U}\mu)_{\mathcal{T}_h} + (d \mathcal{U}\eta, \mathcal{U}\mu)_{\mathcal{T}_h}$	$(f, \mathcal{U}\mu)_{\mathcal{T}_h} + \langle g_h, [\widehat{\mathbf{Q}}\mu] \rangle_{\varepsilon_h}$

†We assume that $a(\mathbf{x})$ is a constant on each element.

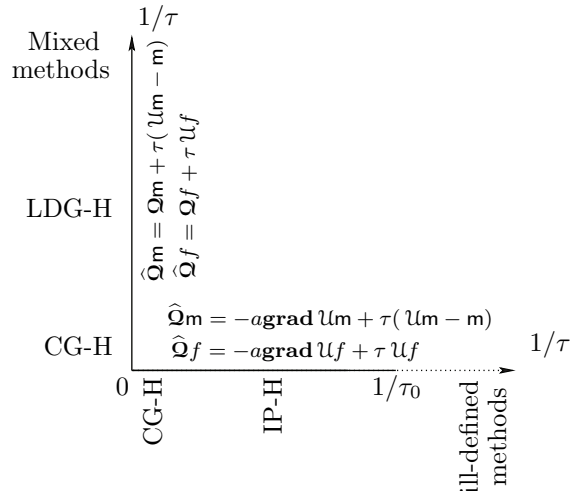


FIG. 3.1. Relations between some hybridizable methods in terms of stabilization parameter τ .

so the normal component of numerical trace $\hat{\mathbf{q}}_h$ is single-valued, and $\mathbf{q}_h \in H(\text{div}, \Omega)$. Moreover, Theorem 2.1 asserts that the conservativity condition is equivalent to (2.9) with

$$\begin{aligned} a_h(\eta, \mu) &= (c\mathbf{Q}\eta, \mathbf{Q}\mu)_{\mathcal{T}_h} + (d\mathcal{U}\eta, \mathcal{U}\mu)_{\mathcal{T}_h}, \\ b_h(\mu) &= \langle g_h, \mathbf{Q}\mu \cdot \mathbf{n} \rangle_{\partial\Omega} + (f, \mathcal{U}\mu)_{\mathcal{T}_h}, \end{aligned}$$

provided $g_h|_{\mathcal{E}_h^0} = 0$. This is, of course, a reasonable choice, since $g|_{\mathcal{E}_h^0} = 0$ and \mathbf{M}_h is a space of discontinuous functions.

These results appeared earlier in [26, Theorem 2.1], where the hybridized RT method of arbitrary order was considered; the case of the lowest order RT method was previously considered in [21]. We can thus conclude that the original RT method is an RT-H method. In [41], bilinear form $a_h(\cdot, \cdot)$ was shown to be positive definite; this implies that λ_h is uniquely determined. Next, we apply our general approach to this method and verify Assumption 2.2 on the positive semidefiniteness of the local solvers and Assumption 2.3, the gluing condition. By Theorem 2.4, this ensures the existence and the uniqueness of λ_h and hence that of approximation (\mathbf{q}_h, u_h) .

PROPOSITION 3.1. *Assumption 2.1 on the existence and the uniqueness of the local solvers, and Assumption 2.2 on the positive semidefiniteness of the local solvers hold for the RT-H method. Assumption 2.3, the gluing condition, also holds with*

$$M(\partial K) = \{\mu : \mu|_e \in \mathcal{P}_k(e) \text{ for all faces } e \text{ of } \partial K\}.$$

Proof. Assumption 2.1 obviously holds. Let us prove Assumption 2.2. To do that, we first show that condition (2.13a) holds. By identity (2.14) with $\mu := \mathbf{m}$, we have that

$$-\langle \mathbf{m}, \hat{\mathbf{Q}}\mathbf{m} \cdot \mathbf{n} \rangle_{\partial K} = (c\mathbf{Q}\mathbf{m}, \mathbf{Q}\mathbf{m})_K + (d\mathcal{U}\mathbf{m}, \mathcal{U}\mathbf{m})_K,$$

by the definition of $\hat{\mathbf{Q}}\mathbf{m}$. We thus see that condition (2.13a) is satisfied.

Now we verify condition (2.13b) with the given choice of $M(\partial K)$. If $\langle \mathbf{m}, \hat{\mathbf{Q}}\mathbf{m} \cdot \mathbf{n} \rangle_{\partial K} = 0$, we immediately obtain $\mathbf{Q}\mathbf{m}|_K = 0$. This implies that (2.3a) can be rewritten

as

$$(3.1) \quad (\mathbf{grad} \mathcal{U}m, \mathbf{v})_K - \langle \mathcal{U}m - m, \mathbf{v} \cdot \mathbf{n} \rangle_{\partial K} = 0 \quad \text{for all } \mathbf{v} \in \mathbf{V}(K).$$

It is well known (see, for example, [12]) that for a given $\mathbf{grad} \mathcal{U}m$ and $\mathcal{U}m - m$, there is a function $\mathbf{v} \in \mathbf{V}(K)$ such that

$$(3.2) \quad (\mathbf{v}, \mathbf{p}_{k-1})_K = (\mathbf{grad} \mathcal{U}m, \mathbf{p}_{k-1})_K \quad \text{for all } \mathbf{p}_{k-1} \in \mathcal{P}_{k-1}(K),$$

$$(3.3) \quad \langle \mathbf{v} \cdot \mathbf{n}, p_k \rangle_e = - \langle \mathcal{U}m - m, p_k \rangle_e \quad \text{for all } p_k \in \mathcal{P}_k(e)$$

for all faces e of K . Using this \mathbf{v} in (3.1), we find that

$$(\mathbf{grad} \mathcal{U}m, \mathbf{grad} \mathcal{U}m)_K + (\mathcal{U}m - m, \mathcal{U}m - m)_{\partial K} = 0.$$

This implies that $\mathcal{U}m$ is a constant on K , so m is constant on ∂K . This proves that condition (2.13b) is satisfied with $M(\partial K)$ as described.

It remains to verify Assumption 2.3. Since we are assuming that triangulation \mathcal{T}_h is conforming, each interior face $e = \partial K^+ \cap \partial K^-$ is *also* a face of both K^+ and K^- . Hence, since $\mu|_e \in \mathcal{P}_k(e)$, we have that $P_{\partial K^+} \mu = \mu = P_{\partial K^-} \mu$ on e . This completes the proof. \square

3.2. The BDM-H method. To obtain the BDM-H method, we use the BDM method to define the main three ingredients of the hybridization method:

1. For each $K \in \mathcal{T}_h$, we take

$$\widehat{\mathbf{Q}}m = \mathbf{Q}m, \quad \widehat{\mathbf{Q}}f = \mathbf{Q}f \quad \text{on } \partial K;$$

2. The finite element spaces are defined as

$$\mathbf{V}(K) = \mathcal{P}_k(K)^n, \quad W(K) = \mathcal{P}_{k-1}(K), \quad k \geq 1;$$

3. The space of approximate traces is defined as $M_h = \mathcal{M}_{h,k}$. This defines the BDM-H method.

Everything said about the RT-H method in the previous subsection applies to the BDM-H method. In particular, we have that the original BDM method is a BDM-H method; see [41].

3.3. The LDG-H methods. The LDG-H methods are obtained by using the LDG method to define the local solvers. The following specifications completely define the class of LDG-H methods:

1. The numerical traces

$$(3.4) \quad \widehat{\mathbf{Q}}m = \mathbf{Q}m + \tau_K(\mathcal{U}m - m)\mathbf{n}, \quad \widehat{\mathbf{Q}}f = \mathbf{Q}f + \tau_K(\mathcal{U}f)\mathbf{n} \quad \text{on } \partial K,$$

where τ_K is a function that can vary on ∂K .

2. The space $\mathbf{V}(K) \times W(K)$ as one of the following choices:

$$(3.5a) \quad \mathcal{P}_k(K)^n \times \mathcal{P}_{k-1}(K), \quad k \geq 1 \text{ and } \tau_K \geq 0 \text{ on } \partial K;$$

$$(3.5b) \quad \mathcal{P}_k(K)^n \times \mathcal{P}_k(K), \quad k \geq 0 \text{ and } \tau_K > 0 \text{ on at least one face of the simplex } K;$$

$$(3.5c) \quad \mathcal{P}_{k-1}(K)^n \times \mathcal{P}_k(K), \quad k \geq 1 \text{ and } \tau_K > 0 \text{ on } \partial K.$$

3. The space of approximate traces is

$$(3.6) \quad \mathbf{M}_h = \mathcal{M}_{h,k}.$$

Typically, the stabilization parameter τ of the LDG methods is a nonnegative constant on each face in \mathcal{E}_h . Here, we allow τ to be double valued on \mathcal{E}_h° , with two branches $\tau^- = \tau_{K^-}$ and $\tau^+ = \tau_{K^+}$ defined on the edge e shared by the finite elements K^- and K^+ . Now the functions $(\mathbf{Qm}, \mathcal{U}m)$ and $(\mathbf{Q}f, \mathcal{U}f)$ are the approximations given by the LDG method to exact solutions of (1.4) on each element, as claimed. As is well known (see [34, 17, 5]), the LDG method is uniquely solvable for $\tau_K > 0$. However, the above specifications define a wider class of LDG-H methods. We show that the existence and the uniqueness of the solution of the method can be guaranteed for each of choices (3.5).

PROPOSITION 3.2. *Assumption 2.1 on the existence and the uniqueness of the local solvers holds for the numerical traces given by (3.4) and with any of choices (3.5) for $\mathbf{V}(K) \times W(K)$.*

To prove this result for all the above-mentioned cases, we use the following auxiliary lemma.

LEMMA 3.1. *Let $\tau_K \geq 0$. With the choice of numerical traces in (3.4), local problems (2.3) and (2.4) are uniquely solvable if $\mathbf{V}(K) \times W(K)$ defined by (3.5) is such that whenever $w \in W(K)$ satisfies*

- (i) $\tau_K w = 0$ on ∂K , and
- (ii) $(w, \operatorname{div} \mathbf{v})_K = 0$ for all $\mathbf{v} \in \mathbf{V}(K)$,

we have that $w = 0$.

Proof. Let us prove the result for first local solver $(\mathbf{Qm}, \mathcal{U}m)$ defined by (2.3). The result for the other local mapping (2.4) is similar. It suffices to prove uniqueness, since this implies existence. To prove uniqueness, we must show that, when $\mathbf{m} = 0$, the only solution of (2.3) is the trivial one.

Taking $\mathbf{v} = \mathbf{Qm}$ and $w = \mathcal{U}m$ in (2.3) and adding the resulting equations, we get

$$(c \mathbf{Qm}, \mathbf{Qm})_K + \left\langle \mathcal{U}m, \left(\widehat{\mathbf{Qm}} - \mathbf{Qm} \right) \cdot \mathbf{n} \right\rangle_{\partial K} + (d \mathcal{U}m, \mathcal{U}m)_K = 0.$$

Inserting the definition of the numerical trace $\widehat{\mathbf{Qm}}$, we get

$$(c \mathbf{Qm}, \mathbf{Qm})_K + \langle \mathcal{U}m, \tau_K \mathcal{U}m \rangle_{\partial K} + (d \mathcal{U}m, \mathcal{U}m)_K = 0,$$

and since c is positive definite and symmetric, $d \geq 0$, and $\tau_K \geq 0$, we have that $\mathbf{Qm} = \mathbf{0}$.

It remains to show that $\mathcal{U}m = 0$. To do so, we note that the above equation implies that $(\tau \mathcal{U}m)_K = 0$ on ∂K . By (2.3a), we also have

$$(\mathcal{U}m, \operatorname{div} \mathbf{v})_K = 0 \quad \text{for all } \mathbf{v} \in \mathbf{V}(K).$$

By hypothesis (ii) of Lemma 3.1, this implies that $\mathcal{U}m = 0$. This completes the proof. \square

We are now ready to prove Proposition 3.2.

Proof. By Lemma 3.1, we have only to show that, for each of three choices (3.5), if $w \in W(K)$ satisfies $\tau_K w = 0$ on ∂K and $(w, \operatorname{div} \mathbf{v})_K = 0$ for all \mathbf{v} in $\mathbf{V}(K)$, then $w = 0$ on K .

Let us show that this is true for the spaces given by (3.5a). Since $\operatorname{div} : \mathbf{V}(K) \rightarrow W(K)$ is surjective, we know there is a \mathbf{v} in $\mathbf{V}(K)$ such that $\operatorname{div} \mathbf{v} = w$. This implies that $(w, w)_K = 0$ and hence that $w = 0$ on K .

Next, let us consider choice (3.5b). Since w must vanish on the face F where $\tau_K > 0$, we immediately have that $w = 0$ if $k = 0$. If $k \geq 1$, it can be factored as $w = \ell_F p_{k-1}$, with $p_{k-1} \in \mathcal{P}_{k-1}(K)$ and ℓ_F equal to the barycentric coordinate function of K that vanishes on F . Then, choosing \mathbf{v} in $\mathbf{V}(K) = \mathcal{P}_k(K)^n$ such that $\text{div } \mathbf{v} = p_{k-1}$, equation

$$0 = (\text{div } \mathbf{v}, w)_K = (\text{div } \mathbf{v}, \ell_F p_{k-1})_K = (p_{k-1}, \ell_F p_{k-1})_K$$

implies that p_{k-1} vanishes on K , so $w = 0$ on K .

Finally, let us consider choice (3.5c). Since $\tau_K > 0$ on ∂K , we have that $w = 0$ on ∂K , and a simple integration by parts gives that

$$(\mathbf{grad} w, \mathbf{v})_K = 0 \quad \text{for all } \mathbf{v} \in \mathbf{V}(K) = \mathcal{P}_{k-1}(K)^n.$$

Taking $\mathbf{v} = \mathbf{grad} w$ allows us to conclude that w is a constant on K and hence identically zero on K . This completes the proof. \square

Note that choices (3.4) of the numerical traces, (3.5) of the finite elements spaces $\mathbf{V}(K) \times W(K)$, and (3.6) for approximate trace space M_h clearly imply that, for all these LDG-H methods, conservativity condition (2.6) is satisfied strongly. Moreover, by Theorem 2.1, the conservativity condition is equivalent to $a_h(\lambda_h, \mu) = b_h(\mu)$ for all $\mu \in M_h$, where

$$\begin{aligned} a_h(\eta, \mu) &= (c \mathbf{Q} \eta, \mathbf{Q} \mu)_{\mathcal{T}_h} + (d \mathcal{U} \eta, \mathcal{U} \mu)_{\mathcal{T}_h} + \langle 1, [(\mathcal{U} \mu - \mu)(\tau(\mathcal{U} \eta - \eta) \mathbf{n})] \rangle_{\mathcal{E}_h}, \\ b_h(\mu) &= \langle g_h, \mathbf{Q} \mu \cdot \mathbf{n} + \tau \mathcal{U} \mu \rangle_{\partial \Omega} + (f, \mathcal{U} \mu)_{\mathcal{T}_h}, \end{aligned}$$

provided $g_h|_{\mathcal{E}_h^*} = 0$.

Form $a_h(\cdot, \cdot)$ is obviously symmetric. That it is also positive definite follows once Assumption 2.2 on the positive semidefiniteness of the local solvers is verified. Set

$$(3.7) \quad M(\partial K) = \{ \mu : \begin{aligned} &\mu|_e \in \mathcal{P}_k(e) \text{ for all faces } e \text{ where } \tau_K = 0, \text{ and} \\ &\mu|_e \in L^2(e) \text{ for all faces } e \text{ where } \tau_K > 0 \}. \end{aligned}$$

PROPOSITION 3.3. *Let the numerical traces be set by (3.4), the local spaces be as in any of choices (3.5), and the space of approximate traces be set by (3.6). Then, Assumption 2.2 on the positive semidefiniteness of the local solvers and Assumption 2.3, the gluing condition, are satisfied with $M(\partial K)$ defined by (3.7).*

Proof. We begin by showing that condition (2.13a) holds. By identity (2.14) with $\mu := \mathbf{m}$ and the definition of $\widehat{\mathbf{Q}} \mathbf{m}$, we have that

$$-\langle \mathbf{m}, \widehat{\mathbf{Q}} \mathbf{m} \cdot \mathbf{n} \rangle_{\partial K} = (c \mathbf{Q} \mathbf{m}, \mathbf{Q} \mathbf{m})_K + (d \mathcal{U} \mathbf{m}, \mathcal{U} \mathbf{m})_K + \langle \tau_K (\mathcal{U} \mathbf{m} - \mathbf{m}), \mathcal{U} \mathbf{m} - \mathbf{m} \rangle_{\partial K}.$$

Since $\tau_K \geq 0$ in all three cases (3.5), we see that condition (2.13a) is satisfied.

Now, let us verify condition (2.13b). If we assume that $\langle \mathbf{m}, \widehat{\mathbf{Q}} \mathbf{m} \cdot \mathbf{n} \rangle_{\partial K} = 0$, we immediately obtain that $\mathbf{Q} \mathbf{m}|_K = 0$ and $\tau(\mathcal{U} \mathbf{m} - \mathbf{m})|_{\partial K} = 0$. This implies that the first equation defining first local solver (2.3a) can be rewritten as

$$(3.8) \quad (\mathbf{grad } \mathcal{U} \mathbf{m}, \mathbf{v})_K - \langle \mathcal{U} \mathbf{m} - \mathbf{m}, \mathbf{v} \cdot \mathbf{n} \rangle_{\partial K} = 0 \quad \text{for all } \mathbf{v} \in \mathbf{V}(K).$$

We use this equation to show that in all three cases (3.5), condition (2.13b) is satisfied with $P_{\partial K}$ defined, on the face e of K , as the L^2 -projection into $\mathcal{P}_k(e)$ if $\tau|_e = 0$ and as the identity if $\tau|_e > 0$:

- (i) In case (3.5a), the result follows exactly as in the proof of Proposition 3.1.
- (ii) In case (3.5b), we know (see [24]) that there is a function $\mathbf{v} \in \mathcal{P}_k(K)^n$ such that

$$(3.9) \quad (\mathbf{v}, \mathbf{p}_{k-1})_K = (\mathbf{grad} \mathcal{U}m, \mathbf{p}_{k-1})_K \quad \text{for all } \mathbf{p}_{k-1} \in \mathcal{P}_{k-1}(K)^n,$$

$$(3.10) \quad \langle \mathbf{v} \cdot \mathbf{n}, p_k \rangle_e = - \langle \mathcal{U}m - m, p_k \rangle_e \quad \text{for all } p_k \in \mathcal{P}_k(e)$$

for all the faces e of K *except* one, say, the face e' on which $\tau > 0$. Setting this \mathbf{v} in (3.8) and using the fact that on e' we have that $m = \mathcal{U}m$, we obtain that $\mathcal{U}m$ is a constant on K and that $m = \mathcal{U}m$ on the remaining faces of ∂K . Thus, m is constant on ∂K and condition (2.13b) is verified. Assumption 2.3, the gluing condition, is trivially satisfied by virtue of the definition of $M(\partial K)$ in (3.7).

(iii) In case (3.5c), we immediately see that $m = \mathcal{U}m$ on ∂K . Now we take $\mathbf{v} = \mathbf{grad} \mathcal{U}m$ in (3.8) to get that $\mathcal{U}m$ is a constant. This verifies Assumption 2.2 as in the previous case. Assumption 2.3 obviously holds from the definition of $M(\partial K)$ in (3.7). \square

Our next result sheds light into the nature of numerical traces $\widehat{\mathbf{q}}_h$ and \widehat{u}_h of the LDG-H schemes.

PROPOSITION 3.4 (characterization of LDG-H methods). *Let the numerical traces be set by (3.4), the local spaces be as in any of choices (3.5), the space of approximate traces be set by (3.6), and (\mathbf{q}_h, u_h) be as defined in (2.5). Then conservativity condition (2.6) holds on \mathcal{E}_h° if and only if*

$$(3.11a) \quad \lambda_h = \widehat{u}_h = \left(\frac{\tau^+}{\tau^- + \tau^+} \right) u_h^+ + \left(\frac{\tau^-}{\tau^- + \tau^+} \right) u_h^- + \left(\frac{1}{\tau^+ + \tau^-} \right) \llbracket \mathbf{q}_h \rrbracket,$$

$$(3.11b) \quad \widehat{\mathbf{q}}_h = \left(\frac{\tau^-}{\tau^- + \tau^+} \right) \mathbf{q}_h^+ + \left(\frac{\tau^+}{\tau^- + \tau^+} \right) \mathbf{q}_h^- + \left(\frac{\tau^+ \tau^-}{\tau^- + \tau^+} \right) \llbracket u_h \rrbracket.$$

Proof. Suppose the conservativity condition holds. We need to prove (3.11a) and (3.11b). By the definition of $\widehat{\mathbf{q}}_h$ (see (2.7)) we have

$$\begin{aligned} \widehat{\mathbf{q}}_h &= \widehat{\mathbf{Q}}\lambda_h + \widehat{\mathbf{Q}}g_h + \widehat{\mathbf{Q}}f \\ &= (\mathbf{Q}\lambda_h + \mathbf{Q}g_h + \mathbf{Q}f) + \tau(\mathcal{U}\lambda_h + \mathcal{U}g_h + \mathcal{U}f - \lambda_h - g_h) \mathbf{n} \\ &= \mathbf{q}_h + \tau(u_h - \lambda_h - g_h) \mathbf{n}. \end{aligned}$$

Inserting this expression into the conservativity condition and taking g_h equal to zero on \mathcal{E}_h° , we obtain that, for any $\mu \in M_h$,

$$\langle \mu, \llbracket \widehat{\mathbf{q}}_h \rrbracket \rangle_{\mathcal{E}_h^\circ} = \langle \mu, \llbracket \mathbf{q}_h + \tau(u_h - \lambda_h) \mathbf{n} \rrbracket \rangle_{\mathcal{E}_h^\circ} = 0,$$

which implies, by our choice of spaces, that $\llbracket \widehat{\mathbf{q}}_h \rrbracket = 0$ on \mathcal{E}_h° or equivalently that

$$\llbracket \mathbf{q}_h \rrbracket + (\tau^+ u_h^+ + \tau^- u_h^-) - (\tau^+ + \tau^-) \lambda_h = 0 \text{ on } \mathcal{E}_h^\circ.$$

Solving for λ_h , we obtain (3.11a). To prove (3.11b), we simply insert the expression for λ_h into the identity

$$\widehat{\mathbf{q}}_h^+ \cdot \mathbf{n}^+ = \mathbf{q}_h^+ \cdot \mathbf{n}^+ + \tau^+ (u_h^+ - \lambda_h)$$

and perform a few algebraic manipulations.

The converse asserted by the proposition is trivial: If identities (3.11) hold, then the normal component of $\widehat{\mathbf{q}}_h$ is single valued on \mathcal{E}_h° and the conservativity condition is satisfied. This completes the proof. \square

COROLLARY 3.2. *The LDG method is not an LDG-H method for any finite τ .*

Proof. On any interior face $e \in \mathcal{E}_h^\circ$, the LDG method has a numerical trace \widehat{u}_h independent of \mathbf{q}_h ; see [34, 17, 5]. On the other hand, by Proposition 3.4, the LDG-H methods have numerical traces \widehat{u}_h that depend on $[\mathbf{q}_h]$. Since this dependence cannot be removed for any finite value of τ , we see that no LDG method is an LDG-H method. This completes the proof. \square

As known from [34, p. 2445] and [17, p. 1681], the independence of numerical trace \widehat{u}_h of the LDG methods of \mathbf{q}_h on interior faces \mathcal{E}_h° allows us to eliminate the unknown \mathbf{q}_h from the equations and to obtain a primal formulation involving only u_h . In contrast, in the LDG-H methods, \widehat{u}_h must depend on \mathbf{q}_h as well. Both approaches recover \mathbf{q}_h locally but using different mechanisms. Since the LDG-H methods lead to a formulation involving only numerical trace λ_h , they have fewer globally coupled unknowns than the LDG method for high order polynomials.

The LDG-H methods considered in this subsection were studied in [17] where it was proven, in particular, that the method is well defined for $\tau > 0$ on \mathcal{E}_h . Methods with $\tau = 0$ do not fit in the framework proposed in [5]; they have been recently studied in [24].

3.4. A limiting case of LDG-H methods. Here we consider hybridizable Galerkin methods that can be obtained formally considering limiting values of the penalty parameter in LDG-H methods. The motivation for doing this arises from the previous corollary (Corollary 3.2), whereby we know that the only chance for showing that an LDG method can be hybridized lies in cases where τ is allowed to be not finite.

We first examine how numerical traces of the previous LDG-H method change as we *formally* pass to a limit in τ . By letting τ^+ go to infinity on the interior face $e = \partial K^+ \cap \partial K^-$ while maintaining a fixed finite τ^- , we find that the expressions for the numerical traces obtained in Proposition 3.4 become

$$(3.12) \quad \widehat{u}_h = u_h^+ \quad \text{and} \quad \widehat{\mathbf{q}}_h = \mathbf{q}_h^- + \tau^- [u_h].$$

Note that the above expression for primal numerical trace \widehat{u}_h is independent of the fluxes, or, in other words, such traces will result in an LDG method. Indeed, the LDG method defined by these numerical traces have been thoroughly studied in the case $\tau^- > 0$; see [34, 17, 5].

In the special case $\tau^- = 0$, we get

$$\widehat{u}_h = u_h^+ \quad \text{and} \quad \widehat{\mathbf{q}}_h = \mathbf{q}_h^-,$$

which also defines a previously studied LDG method. For this scheme, the discontinuities of the approximate solution across interior interelement boundaries do not introduce any dissipation. The dissipative effect of the discontinuities is concentrated on the boundary of the domain and hence reduced to a “minimum,” which is the reason for its name, the *minimal dissipation* LDG method. Since this scheme does not fit the unified analysis in [5], it was studied in [20] and [24] for problems in one and several space dimensions, respectively.

The formal passage to limit solely in the expressions for numerical traces does not clarify if the limiting methods are hybridizable. In particular, we must explain

precisely what we mean by setting $\tau_K = \infty$ in the context of local solvers. To do so, let F_K be the union of one or more faces of the element K where we want to set the branch τ_K to ∞ . Since

$$\widehat{\mathbf{Q}}\mathbf{m} = \mathbf{Q}\mathbf{m} + \tau_K(\mathcal{U}\mathbf{m} - \mathbf{m})\mathbf{n},$$

we expect that in the formal limit of $\tau_K = \infty$, we should have $\mathcal{U}\mathbf{m} - \mathbf{m} = 0$. Then the value of $\widehat{\mathbf{Q}}\mathbf{m}$ on F_K becomes an unknown because the last term above is an unknown formal product of 0 with ∞ . Motivated by this, we now define the local solvers with $\widehat{\mathbf{Q}}\mathbf{m}$ and $\widehat{\mathbf{Q}}f$ as *new* unknowns. More precisely, setting

$$W(K) = \mathcal{P}_k(K), \quad \mathbf{V}(K) = \mathcal{P}_k(K)^n, \quad \mathbf{T}_K(F_K) = \{\mathbf{n}_K w|_{F_K} : w \in W(K)\},$$

we define local solution $(\mathbf{Q}\mathbf{m}, \mathcal{U}\mathbf{m}, (\widehat{\mathbf{Q}}\mathbf{m})_{F_K}) \in \mathbf{V}(K) \times W(K) \times \mathbf{T}_K(F_K)$ for any $\mathbf{m} \in \mathbf{M}_h$ by

$$(3.13a) \quad (c\mathbf{Q}\mathbf{m}, \mathbf{v})_K - (\mathcal{U}\mathbf{m}, \operatorname{div} \mathbf{v})_K = -\langle \mathbf{m}, \mathbf{v} \cdot \mathbf{n} \rangle_{\partial K} \quad \text{for all } \mathbf{v} \in \mathbf{V}(K),$$

$$(3.13b) \quad -(\mathbf{grad} w, \mathbf{Q}\mathbf{m})_K + \left\langle w, \widehat{\mathbf{Q}}\mathbf{m} \cdot \mathbf{n} \right\rangle_{\partial K} + (d\mathcal{U}\mathbf{m}, w)_K = 0 \quad \text{for all } w \in W(K),$$

$$(3.13c) \quad \mathcal{U}\mathbf{m} = \mathbf{m} \quad \text{on } F_K.$$

Here, just as for the LDG-H methods, we set

$$\widehat{\mathbf{Q}}\mathbf{m} = \mathbf{Q}\mathbf{m} + \tau_K(\mathcal{U}\mathbf{m} - \mathbf{m})\mathbf{n} \quad \text{on } \partial K \setminus F_K.$$

Similarly, we define $(\mathbf{Q}f, \mathcal{U}f, (\widehat{\mathbf{Q}}f)_{F_K})$ as the element of $\mathbf{V}(K) \times W(K) \times \mathbf{T}_K(F_K)$ such that

$$(3.14a) \quad (c\mathbf{Q}f, \mathbf{v})_K - (\mathcal{U}f, \operatorname{div} \mathbf{v})_K = 0 \quad \text{for all } \mathbf{v} \in \mathbf{V}(K),$$

$$(3.14b) \quad -(\mathbf{grad} w, \mathbf{Q}f)_K + \left\langle w, \widehat{\mathbf{Q}}f \cdot \mathbf{n} \right\rangle_{\partial K} + (d\mathcal{U}f, w)_K = (f, w) \quad \text{for all } w \in W(K),$$

$$(3.14c) \quad \mathcal{U}f = 0 \quad \text{on } F_K,$$

where

$$\widehat{\mathbf{Q}}f = \mathbf{Q}f + \tau_K(\mathcal{U}f)\mathbf{n} \quad \text{on } \partial K \setminus F_K.$$

We set the space of approximate traces by

$$(3.15) \quad \mathbf{M}_h = \{\mu \in \mathcal{M}_{h,k} : \mu|_{F_K} \text{ is continuous on } F_K \text{ for all } K \in \mathcal{T}_h\}.$$

Note that the continuity condition in the above definition reflects the fact that the local solvers satisfy strong Dirichlet boundary conditions on F_K for all $K \in \mathcal{T}_h$; see (3.13c) and (3.14c). This completes the definition of the *limiting case of the LDG-H method* when $\tau_K = \infty$ on F_K . From now on, the above modification of the LDG local solvers is tacitly understood whenever we say that a branch of τ is infinity on a face. It is easy to check, by arguments similar to that in Proposition 3.2, that local problems (3.13) and (3.14) are uniquely solvable for every \mathbf{m} in \mathbf{M}_h and every $f \in L^2(\Omega)$ provided, for each element $K \in \mathcal{T}_h$, τ_K is not identically equal to zero on ∂K whenever F_K is the empty set.

Note that, although the local solvers have been modified, Theorem 2.1 continues to apply because its proof only relies on the form of the first two equations in the

local problems. Indeed, (3.13a) and (3.13b) are identical in form to (2.3a) and (2.3b), respectively; a similar remark applies to the equation of the second local solvers. Therefore, Theorem 2.1 also holds in this case. In particular, we have that

$$a_h(\eta, \mu) = (c \mathbf{Q}\eta, \mathbf{Q}\mu)_{\mathcal{T}_h} + (d \mathcal{U}\eta, \mathcal{U}\mu)_{\mathcal{T}_h} + \sum_{K \in \mathcal{T}_h} \langle \tau(\mathcal{U}\eta - \eta), (\mathcal{U}\mu - \mu) \rangle_{\partial K \setminus F_K}.$$

Finally, it is not difficult to see that Proposition 3.3 also holds. By Theorem 2.4, bilinear form $a_h(\cdot, \cdot)$ is positive definite, and we can immediately see that λ_h is uniquely determined.

Note that, unlike all previous examples, conservativity condition (2.6) for these methods is only imposed weakly. This is because while the jumps of $\widehat{\mathbf{q}}_h$ lie in $\mathcal{M}_{h,k}$, the approximate traces μ are in the space \mathbf{M}_h , which is a strict subspace of $\mathcal{M}_{h,k}$. Since all LDG methods have single-valued numerical traces, this seems to suggest that no LDG method can be a limiting case of the LDG-H method. However, this is not the case, as we see next.

We consider the *one-sided limiting case of the LDG-H method*. This is the same as the above-defined limiting case of the LDG-H method but with the following additional assumption: For every interior face e in \mathcal{E}_h° , one branch of τ is infinity, and the other branch is finite-valued.

COROLLARY 3.3. *The one-sided limiting case of the LDG-H method coincides with the LDG method whose numerical traces on the interior faces are given by (3.12).*

Proof. Let λ_h^∞ denote the solution of the one-sided limiting case of the LDG-H method, and let

$$\mathbf{q}_h^\infty = \mathbf{Q}\lambda_h^\infty + \mathbf{Q}g_h + \mathbf{Q}f, \quad u_h^\infty = \mathcal{U}\lambda_h^\infty + \mathcal{U}g_h + \mathcal{U}f.$$

We will prove that \mathbf{q}_h^∞ and u_h^∞ coincide with the corresponding solution components $\mathbf{q}_h^{\text{LDG}}$ and u_h^{LDG} , respectively, of the LDG method with numerical traces set as in (3.12).

By the definition of the LDG method, $\mathbf{q}_h^{\text{LDG}}$ and u^{LDG} satisfy (2.8a)–(2.8b) with the λ_h and $\widehat{\mathbf{q}}_h$ therein set, respectively, to \widehat{u}_h and $\widehat{\mathbf{q}}_h$ of (3.12), which, for clarity, we will rewrite as $\widehat{u}_h^{\text{LDG}}$ and $\widehat{\mathbf{q}}_h^{\text{LDG}}$.

It suffices to show that \mathbf{q}_h^∞ and u_h^∞ satisfy the same equations as $\mathbf{q}_h^{\text{LDG}}$ and u_h^{LDG} . Adding local solver equations (3.13a) and (3.14a) over all elements, we find that \mathbf{q}_h^∞ and u_h^∞ satisfy the first equation of the LDG method with λ_h^∞ in place of $\widehat{u}_h^{\text{LDG}}$. But, since every interior edge has an infinite penalty branch and since

$$(3.16) \quad \lambda_h^\infty|_{F_K} = (u_h^\infty)_{F_K} \quad \text{for all elements } K,$$

we find that λ_h^∞ is in the same form as LDG numerical trace $\widehat{u}_h^{\text{LDG}}$.

Also, summing local solver equations (3.13b) and (3.14b) over all elements, we find that \mathbf{q}_h^∞ and u_h^∞ satisfy the second equation of the LDG methods, with $\widehat{\mathbf{q}}_h^\infty \equiv \widehat{\mathbf{Q}}\lambda_h^\infty + \widehat{\mathbf{Q}}g_h + \widehat{\mathbf{Q}}f$ in place of $\widehat{\mathbf{q}}_h^{\text{LDG}}$. We will now show that the second equation, in fact, holds with the LDG flux. For this, we use the fact that

$$(3.17) \quad \langle \llbracket \widehat{\mathbf{q}}_h^\infty \rrbracket, \mu \rangle_{\mathcal{E}_h} = 0$$

for all μ in the subspace M_h of functions in \mathbf{M}_h (defined by (3.15)), with $\mu|_{\partial\Omega} = 0$. Now, if w is any function in $W(K)$, then $w|_{F_K}$, extended by zero to \mathcal{E}_h , is in M_h . Therefore, (3.17) implies

$$\begin{aligned} \langle \widehat{\mathbf{q}}_h^\infty \cdot \mathbf{n}, w \rangle_{F_K} &= - \langle (\widehat{\mathbf{q}}_h^\infty)_{K^c} \cdot (\mathbf{n})_{K^c}, w \rangle_{F_K} \\ &= - \langle (\widehat{\mathbf{q}}_h^\infty)_{K^c} + (\tau)_{K^c} ((u_h^\infty)_{K^c} - \lambda_h^\infty) (\mathbf{n})_{K^c}, (\mathbf{n})_{K^c} w \rangle_{F_K}. \end{aligned}$$

Here, for notational convenience, we have denoted the branch of a multivalued function f from outside K by $(f)_{K^c}$. By (3.16), we can rewrite the right-hand side as

$$\langle \widehat{\mathbf{q}}_h^\infty \cdot \mathbf{n}, w \rangle_{F_K} = -\langle (\widehat{\mathbf{q}}_h^\infty)_{K^c} + (\tau)_{K^c} \llbracket u_h^\infty \rrbracket, (\mathbf{n})_{K^c} w \rangle_{F_K}$$

and conclude that

$$(3.18) \quad \sum_K \langle \widehat{\mathbf{q}}_h^\infty \cdot \mathbf{n}, w \rangle_{\partial K} = \sum_K \langle \widehat{\mathbf{q}}_h^{\text{LDG}} \cdot \mathbf{n}, w \rangle_{\partial K}.$$

Thus, \mathbf{q}_h^∞ and u_h^∞ satisfy the same equations as the LDG method with the same expressions for numerical traces as in the LDG case. \square

Note that in the above proof, $\widehat{\mathbf{q}}_h^\infty$ and $\widehat{\mathbf{q}}_h^{\text{LDG}}$ are not identical, in general, although (3.18) holds. This explains why the normal component of the limiting LDG-H numerical trace may not be single valued, although the numerical trace of its equivalent LDG method is single valued.

3.5. The CG-H method. The CG-H methods are obtained by using the CG method to define the local solvers. We are also going to see that they are also obtained from LDG-H methods by letting τ go to infinity everywhere.

Again, we need to specify the main ingredients of the local solvers. Similarly to the the limiting case of LDG-H methods, we need to give a new meaning of the local solvers since $\tau = \infty$. Since the numerical flux $\widehat{\mathbf{Q}} \cdot$ will be unknown, we need an appropriate space for its approximation.

1. For any $k \geq 1$ and any $K \in \mathcal{T}_h$, we define the finite element spaces by

$$(3.19) \quad \begin{aligned} \mathbf{V}(K) &= \mathcal{P}_{k-1}(K)^n, \quad W(K) = \mathcal{P}_k(K), \quad \text{and} \\ \mathbf{T}(\partial K) &:= \{\mathbf{n}_K w|_{\partial K} : w \in W(K)\}. \end{aligned}$$

2. The numerical traces of fluxes $\widehat{\mathbf{Q}} \cdot$ are unknown and will be determined by the modified local solvers as follows: $(\mathbf{Qm}, \mathcal{U}m, \widehat{\mathbf{Q}}m) \in \mathbf{V}(K) \times W(K) \times \mathbf{T}(\partial K)$ is a solution to the problem

$$(3.20a) \quad (c \mathbf{Qm}, \mathbf{v})_K - (\mathcal{U}m, \text{div } \mathbf{v})_K = -\langle \mathbf{m}, \mathbf{v} \cdot \mathbf{n} \rangle_{\partial K},$$

$$(3.20b) \quad -(\mathbf{grad} w, \mathbf{Qm})_K + \langle w, \widehat{\mathbf{Q}}m \cdot \mathbf{n} \rangle_{\partial K} + (d \mathcal{U}m, w)_K = 0,$$

$$(3.20c) \quad \mathcal{U}m = \mathbf{m} \quad \text{on } \partial K.$$

for all $\mathbf{v} \in \mathbf{V}(K)$ and $w \in W(K)$. Similarly, $(\mathbf{Q}f, \mathcal{U}f, \widehat{\mathbf{Q}}f) \in \mathbf{V}(K) \times W(K) \times \mathbf{T}(\partial K)$ is defined by

$$(3.21a) \quad (c \mathbf{Q}f, \mathbf{v})_K - (\mathcal{U}f, \text{div } \mathbf{v})_K = 0,$$

$$(3.21b) \quad -(\mathbf{grad} w, \mathbf{Q}f)_K + \langle w, \widehat{\mathbf{Q}}f \cdot \mathbf{n} \rangle_{\partial K} + (d \mathcal{U}f, w)_K = (f, w)_K,$$

$$(3.21c) \quad \mathcal{U}f = 0 \quad \text{on } \partial K.$$

for all $\mathbf{v} \in \mathbf{V}(K)$ and $w \in W(K)$,

3. For the space of approximate traces, we take

$$(3.22) \quad \mathbf{M}_h := \mathcal{M}_{h,k}^c.$$

We begin our discussion regarding the above CG-H method by verifying the assumptions required by Theorem 2.4.

PROPOSITION 3.5. *Assumption 2.1 on the existence and the uniqueness of the local solvers holds for the CG-H local solver. Assumption 2.2 on the positive semidefiniteness of the local solvers and Assumption 2.3, the gluing condition, hold with $M(\partial K) = L^2(\partial K)$.*

Proof. We prove the result for local solver $(\mathbf{Qm}, \mathcal{U}m, \widehat{\mathbf{Q}}m)$ defined by (3.20). The result for the local mapping defined by (3.21) is similar. Since the resulting system is square, we prove only uniqueness since this implies existence. Thus, we need to show that if $m = 0$, then the only solution is the trivial one.

Taking $v = \mathbf{Qm}$ in (3.20a) and $w = \mathcal{U}m$ in (3.20b) and adding the resulting equations, we get

$$(c\mathbf{Qm}, \mathbf{Qm})_K + \left\langle \mathcal{U}m, \left(\widehat{\mathbf{Q}}m - \mathbf{Qm} \right) \cdot \mathbf{n} \right\rangle_{\partial K} + (d\mathcal{U}m, \mathcal{U}m)_K = 0.$$

Since, by (3.20c), $\mathcal{U}m = 0$ on ∂K , we immediately obtain that $\mathbf{Qm} = \mathbf{0}$. This implies that (3.20a) can be rewritten as follows:

$$(\mathbf{grad}\mathcal{U}m, v)_K = 0 \quad \text{for all } v \in \mathbf{V}(K),$$

which implies that $\mathcal{U}m = 0$.

It remains to show that $\widehat{\mathbf{Q}}m = 0$. To do that, we use (3.20b) rewritten as

$$\left\langle w, \widehat{\mathbf{Q}}m \cdot \mathbf{n} \right\rangle_{\partial K} = 0 \quad \text{for all } w \in W(K).$$

By the definition of space $\mathbf{T}(\partial K)$, we can find a function $w \in W(K)$ such that $\widehat{\mathbf{Q}}m = w\mathbf{n}$. This readily implies that $\widehat{\mathbf{Q}}m = 0$. This completes the verification of Assumption 2.1.

Inequality (2.13a) of Assumption 2.2 can easily be seen to hold. The second part of Assumption 2.2 also holds, since $M(\partial K) = L^2(\partial K)$. Finally, Assumption 2.3 trivially holds. \square

Next, we discuss the conservativity condition. Flux approximation \mathbf{q}_h of the CG-H method is, in general, not in $H(\text{div}, \Omega)$. Nonetheless, it is interesting to observe that even the CG-H method has a weak conservativity property. This property holds for numerical flux trace $\widehat{\mathbf{q}}_h = \widehat{\mathbf{Q}}\lambda_h + \widehat{\mathbf{Q}}g_h + \widehat{\mathbf{Q}}f$, a quantity that is not present in the standard formulations of the CG methods but essential in our approach. Indeed, Theorem 2.1 asserts that $\widehat{\mathbf{q}}_h$ satisfies

$$\langle \mu, \llbracket \widehat{\mathbf{q}}_h \rrbracket \rangle_{\mathcal{E}_h^\circ} = 0 \quad \text{for all } \mu \in M_h,$$

which is a weak conservativity condition.

Observe that if a is a constant matrix on each element, by the definition of local solvers (3.20) and (3.21), we have that

$$(3.23) \quad \mathbf{Q}m = -a\mathbf{grad}\mathcal{U}m \quad \text{and} \quad \mathbf{Q}f = -a\mathbf{grad}\mathcal{U}f.$$

Hence, \mathbf{q}_h in (2.8a), being the sum of the local flux solutions, equals $-a\mathbf{grad}u_h$ on each element. Substituting this in (2.8b) and using the conservativity condition, we immediately see that u_h satisfies the standard CG equations. In addition, the boundary conditions defining local solvers (3.20c) and (3.21c) imply that u_h is continuous.

Thus, we conclude that this CG-H formulation coincides with the CG method whenever a is constant. In other words, the original CG method is a CG-H method when the matrix-valued function a is a constant on each element. In this case, we can also simplify the forms in (2.9) using (3.23) to

$$\begin{aligned} a_h(\eta, \mu) &= (a \mathbf{grad} \mathcal{U}\eta, \mathbf{grad} \mathcal{U}\mu)_{\mathcal{T}_h} + (d\mathcal{U}\eta, \mathcal{U}\mu)_{\mathcal{T}_h}, \\ b_h(\mu) &= \left\langle g_h, \left[\widehat{\mathcal{Q}}\mu \right] \right\rangle_{\mathcal{E}_h} + (f, \mathcal{U}\mu)_{\mathcal{T}_h}. \end{aligned}$$

Note that in our case, we do not necessarily have that $g_h|_{\mathcal{E}_h^\circ} = 0$. Hence, the corresponding integral cannot be performed only on $\partial\Omega$ as in the previous cases.

Formulation (2.9) is nothing but the weak formulation for the CG method with static condensation of its interior degrees of freedom. This hybridization approach for the CG methods of degree k is explored in [31], where, in particular, a postprocessing technique providing locally conservative flux approximations competitive with that given by the RT methods of degree $k - 1$ is introduced.

When the matrix-valued function a is not constant on each element, we cannot write (3.23) anymore. Instead, “ a ” has to be replaced by a function “ \mathbf{a} ,” which is, roughly speaking, the inverse of some local average of c , the inverse of a . In practice, however, we do not compute the matrix-valued function \mathbf{a} ; instead, we compute directly the functions $\mathcal{Q}\mathbf{m}$ and $\mathcal{Q}f$ by using the definition of the local solvers.

3.6. IP-H methods. The IP-H methods are obtained by using the numerical traces and the local solvers of the IP method. Thus,

1. the numerical traces are given by

$$(3.24) \quad \widehat{\mathcal{Q}}\mathbf{m} = -a \mathbf{grad} \mathcal{U}\mathbf{m} + \tau_K(\mathcal{U}\mathbf{m} - \mathbf{m}) \mathbf{n}, \widehat{\mathcal{Q}}f = -a \mathbf{grad} \mathcal{U}f + \tau_K(\mathcal{U}f) \mathbf{n}, \quad \text{on } \partial K;$$

2. the finite element space $\mathbf{V}(K) \times W(K)$ is defined for $k \geq 1$ as

$$(3.25) \quad \mathbf{V}(K) = \mathcal{P}_k(K)^n, \quad W(K) = \mathcal{P}_k(K);$$

3. the space of approximate traces is chosen as

$$(3.26) \quad \mathbf{M}_h := \mathcal{M}_{h,k}.$$

As before, τ is a double-valued function on \mathcal{E}_h° , with two branches $\tau^- = \tau_{K^-}$ and $\tau^+ = \tau_{K^+}$ defined on the edge e shared by the finite elements K^- and K^+ .

Note that IP methods can be defined by using a flux formulation, as the one employed here to define the local solvers or by means of a primal formulation; see [5]. These two IP methods, however, do coincide whenever the function a is a constant on each element $K \in \mathcal{T}_h$. For this reason, we are going to assume here that this is the case. All the results for this case, however, can be easily extended to the case in which a is not necessarily piecewise constant.

Next, we provide sufficient conditions for the IP-H method to be well defined. For simplicity, we assume that mesh \mathcal{T}_h is shape regular, that is, that there is a constant $\gamma > 0$ such that $h_K/\rho_K \leq \gamma$ for all simplexes $K \in \mathcal{T}_h$, where h_K is the diameter of K and ρ_K the diameter of the largest ball contained in K .

PROPOSITION 3.6. *Let the numerical traces be given by (3.24) and the local spaces by (3.25). Suppose $a(\mathbf{x})$ is a constant matrix on each element K . Then Assumption (2.1) on the existence and the uniqueness of the local solvers holds provided $\tau_K > c_0/h_K$ for some constant $c_0 > 0$ depending on γ and $a(\mathbf{x})$.*

For a proof, see [6, 3]. Having established that the local solvers are well defined, we can apply Theorem 2.1. We find that the conservativity condition implies that λ_h solves (2.9), with

$$\begin{aligned} a_h(\eta, \mu) &= (c \mathbf{Q}\eta, \mathbf{Q}\mu)_{\mathcal{T}_h} + (d \mathcal{U}\eta, \mathcal{U}\mu)_{\mathcal{T}_h} \\ &\quad + \langle 1, \llbracket (\mu - \mathcal{U}\mu)(a \mathbf{grad} \mathcal{U}\eta + \mathbf{Q}\eta) \rrbracket \rrbracket_{\mathcal{E}_h}, \\ &\quad + \langle 1, \llbracket (\mathcal{U}\mu - \mu)(\tau(\mathcal{U}\eta - \eta)\mathbf{n}) \rrbracket \rrbracket_{\mathcal{E}_h}, \\ b_h(\mu) &= \langle g_h, -a \mathbf{grad} \mathcal{U}\mu \cdot \mathbf{n} + \tau \mathcal{U}\mu \rangle_{\partial\Omega} + (f, \mathcal{U}\mu)_{\mathcal{T}_h}, \end{aligned}$$

provided $g_h|_{\mathcal{E}_h^o} = 0$. Using (2.12a) of Lemma 2.3 and the fact that $a(\mathbf{x})$ is constant on each K , we can simplify this expression as follows:

$$\begin{aligned} a_h(\eta, \mu) &= (c \mathbf{Q}\eta, \mathbf{Q}\mu)_{\mathcal{T}_h} - (c \mathbf{Q}\eta + \mathbf{grad} \mathcal{U}\eta, \mathbf{Q}\mu + a \mathbf{grad} \mathcal{U}\mu)_{\mathcal{T}_h} + (d \mathcal{U}\eta, \mathcal{U}\mu)_{\mathcal{T}_h} \\ &\quad + \langle 1, \llbracket (\mathcal{U}\mu - \mu)(\tau(\mathcal{U}\eta - \eta)\mathbf{n}) \rrbracket \rrbracket_{\mathcal{E}_h} \\ &= (a \mathbf{grad} \mathcal{U}\eta, \mathbf{grad} \mathcal{U}\mu)_{\mathcal{T}_h} + (d \mathcal{U}\eta, \mathcal{U}\mu)_{\mathcal{T}_h} \\ &\quad + \langle 1, \llbracket ((\mathcal{U}\eta - \eta) a \mathbf{grad} \mathcal{U}\mu + (\mathcal{U}\mu - \mu) a \mathbf{grad} \mathcal{U}\eta) \rrbracket \rrbracket_{\mathcal{E}_h} \\ &\quad + \langle 1, \llbracket (\mathcal{U}\mu - \mu)(\tau(\mathcal{U}\eta - \eta)\mathbf{n}) \rrbracket \rrbracket_{\mathcal{E}_h}. \end{aligned}$$

The positive definiteness of the form $a_h(\cdot, \cdot)$ can be proven as in the case of LDG-H methods. Indeed, this fact is an immediate consequence of Theorem 2.4 and the following result.

PROPOSITION 3.7. *Let the numerical traces of the fluxes be set by (3.24), the local spaces be defined by (3.25), and the space of approximate traces be set by (3.26). Suppose $a(\mathbf{x})$ is a constant matrix on each element K . Then Assumption 2.2 on the positive semidefiniteness of the local solvers and Assumption 2.3, the gluing condition, are satisfied with $M(\partial K) = \{\mu : \mu|_e \in \mathcal{P}_k(e) \text{ for all faces } e \in \partial K\}$ whenever $\tau_K > c_0/h_K$ for some constant $c_0 > 0$ depending on γ and $a(\mathbf{x})$.*

The proof of this result is similar to that of Proposition 3.3.

Just as for LDG-H methods, we can give a characterization of the IP-H methods. It is given in the proposition below, which is an analog of Proposition 3.4 for the LDG-H methods. Since the proof is similar, we omit it.

PROPOSITION 3.8 (characterization of IP-H methods). *Let the numerical traces be set by (3.24), the spaces be as in (3.25), and (\mathbf{q}_h, u_h) be as defined in (2.5). Then conservativity condition (2.6) holds if and only if on \mathcal{E}_h^o*

(3.27a)

$$\lambda_h = \widehat{u}_h = \left(\frac{\tau^+}{\tau^- + \tau^+} \right) u_h^+ + \left(\frac{\tau^-}{\tau^- + \tau^+} \right) u_h^- - \left(\frac{1}{\tau^+ + \tau^-} \right) \llbracket a \mathbf{grad} u_h \rrbracket,$$

(3.27b)

$$\widehat{\mathbf{q}}_h = - \left(\frac{\tau^-}{\tau^- + \tau^+} \right) a^+ \mathbf{grad} u_h^+ - \left(\frac{\tau^+}{\tau^- + \tau^+} \right) a^- \mathbf{grad} u_h^- + \left(\frac{\tau^+ \tau^-}{\tau^- + \tau^+} \right) \llbracket u_h \rrbracket.$$

We also have results analogous to Corollary 3.2.

COROLLARY 3.4. *The standard IP method is not an IP-H method for any finite τ .*

Proof. Comparing the numerical traces of the standard IP method (see [5, Table 3.1]), namely,

$$\widehat{u}_h^{\text{IP}} = \llbracket u_h \rrbracket \quad \text{and} \quad \widehat{\mathbf{q}}_h^{\text{IP}} = - \llbracket a \mathbf{grad} u_h \rrbracket + C \llbracket u_h \rrbracket,$$

with the expressions for the numerical traces in Proposition 3.8, we find that they cannot coincide for any value of τ . \square

In spite of this negative result, a stabilized DG finite element method introduced in [38] and rewritten in [37] as an IP method, turns out to be an IP-H method. To describe this scheme in a simple setting, assume that $d = 0$ and $g = 0$. The method, as presented in [38], does not use the function λ_h approximating $u|_{\mathcal{E}_h}$. Instead, it uses approximate fluxes ℓ_h approximating the normal component of $\mathbf{a}\mathbf{grad}u$. The space in which ℓ_h lies is the space of scalar double-valued functions defined by

$$L_h = \{q : q|_e \in \mathcal{P}_k(e) \text{ for all } e \in \mathcal{E}_h \text{ and } \overline{q_{K^+}} + q_{K^-} = 0 \text{ on } e = \partial K^+ \cap \partial K^-\}.$$

The DG method of [38] seeks $u_h \in W_h$, given by (2.1), with $W(K) = \mathcal{P}_k(K)$, and $\ell_h \in L_h$ such that

$$(3.28) \quad \sum_{K \in \mathcal{T}_h} \left\{ (\mathbf{a}\mathbf{grad}u_h, \mathbf{grad}v)_K - \langle \ell_h, v \rangle_{\partial K} - \langle \eta, u_h \rangle_{\partial K} \right\} - \alpha h \sum_{K \in \mathcal{T}_h} \langle \ell_h - \mathbf{a}\mathbf{grad}u_h \cdot \mathbf{n}_K, \eta - \mathbf{a}\mathbf{grad}v \cdot \mathbf{n}_K \rangle_{\partial K} = (f, v)$$

for all $v \in W_h$ and $\eta \in L_h$. Here, $\alpha > 0$ is a constant stabilization parameter, and $h = \max_{K \in \mathcal{T}_h} h_K$.

Taking $v \equiv 0$ and using that $\{\{\eta\}\} = 0$ on \mathcal{E}_h° , we get

$$(3.29) \quad \ell_h = \begin{cases} \{\{\mathbf{a}\mathbf{grad}u_h\}\} \cdot \mathbf{n} - \frac{1}{2\alpha h} \llbracket u_h \rrbracket \cdot \mathbf{n} & \text{on } \mathcal{E}_h^\circ, \\ \mathbf{a}\mathbf{grad}u_h \cdot \mathbf{n} - \frac{1}{\alpha h} u_h & \text{on } \mathcal{E}_h^\partial. \end{cases}$$

We see from the above equation that ℓ_h is indeed an approximation to the normal component of $\mathbf{a}\mathbf{grad}u$. Next, taking $\eta \equiv 0$ in (3.28) and substituting therein the expression for ℓ_h from (3.29), we get that $u_h \in W_h$ satisfies

$$(3.30) \quad \begin{aligned} & (\mathbf{a}\mathbf{grad}u_h, \mathbf{grad}v)_{\mathcal{T}_h} - \langle \{\{\mathbf{a}\mathbf{grad}v\}\}, \llbracket u_h \rrbracket \rangle_{\mathcal{E}_h} \\ & - \left\langle \{\{\mathbf{a}\mathbf{grad}u_h\}\} - \frac{1}{2\alpha h} \llbracket u_h \rrbracket, \llbracket v \rrbracket \right\rangle_{\mathcal{E}_h^\circ} \\ & - \left\langle \mathbf{a}\mathbf{grad}u_h - \frac{1}{\alpha h} u_h \mathbf{n}, v \mathbf{n} \right\rangle_{\mathcal{E}_h^\partial} \\ & - \left\langle \frac{\alpha h}{2} \llbracket \mathbf{a}\mathbf{grad}u_h \rrbracket, \llbracket \mathbf{a}\mathbf{grad}v \rrbracket \right\rangle_{\mathcal{E}_h^\circ} = (f, v) \end{aligned}$$

for all $v \in W_h$.

Now, we show that this is an IP-H method. Comparing the above formulation with the general primal formulation given by [5, equation (3.11)], we can easily verify that if we take

$$(3.31) \quad \begin{aligned} \widehat{u}_h &= \{\{u_h\}\} - \frac{\alpha h}{2} \llbracket \mathbf{a}\mathbf{grad}u_h \rrbracket \quad \text{on } \mathcal{E}_h^\circ, \\ \widehat{\mathbf{q}}_h &= \begin{cases} -\{\{\mathbf{a}\mathbf{grad}u_h\}\} + \frac{1}{2\alpha h} \llbracket u_h \rrbracket & \text{on } \mathcal{E}_h^\circ, \\ -\mathbf{a}\mathbf{grad}u_h + \frac{1}{\alpha h} u_h \mathbf{n} & \text{on } \mathcal{E}_h^\partial, \end{cases} \end{aligned}$$

we recover (3.30). Hence, the above numerical traces are exactly the numerical traces of the IP-H method given by Proposition 3.8 with $\tau^+ = \tau^- = (\alpha h)^{-1}$. This shows that the DG method proposed in [38] is an IP-H method. The correspondence between their flux approximation ℓ_h and our numerical flux trace follows immediately from (3.31) and (3.29):

$$\widehat{\mathbf{q}}_h \cdot \mathbf{n} = -\ell_h.$$

It also follows from Proposition 3.6 that the IP method of (3.30) is well defined when $\alpha > 0$ is sufficiently small; a result already established in [38].

Let us end by pointing out that other IP-H-like methods can be obtained. For example, we could take $\mathbf{V}(K) = \mathcal{P}_{k-1}(K)^n$.

3.7. The NC-H methods. We now consider nonconforming hybridizable (NC-H) methods and show that methods like the P_1 -nonconforming method introduced in [36] in the framework of the stationary Stokes equations, are, in fact, NC-H methods. Again the main components of the NC-H method are defined as follows:

1. For any $k \geq 1$, set

$$\begin{aligned} \mathbf{V}(K) &= \mathcal{P}_{k-1}(K)^n, & W(K) &= \mathcal{P}_k(K), \\ (3.32) \quad M(\partial K) &= \{q : q|_e \in \mathcal{P}_{k-1}(e) \text{ for every face } e \text{ of } K\}, \\ \mathbf{T}(\partial K) &= \{q\mathbf{n}_K : q|_e \in \mathcal{P}_{k-1}(e) \text{ for every face } e \text{ of } K\}. \end{aligned}$$

2. Define local solutions $(\mathbf{Qm}, \mathcal{U}m, (\widehat{\mathbf{Q}}\mathbf{m})_K)$ and $(\mathcal{Q}f, \mathcal{U}f, (\widehat{\mathbf{Q}}f)_K)$ as the elements of $\mathbf{V}(K) \times W(K) \times \mathbf{T}(\partial K)$ satisfying

$$(3.33a) \quad (c\mathbf{Qm}, \mathbf{v})_K - (\mathcal{U}m, \operatorname{div} \mathbf{v})_K = -\langle \mathbf{m}, \mathbf{v} \cdot \mathbf{n} \rangle_{\partial K},$$

$$(3.33b) \quad -(\operatorname{grad} w, \mathbf{Qm})_K + \left\langle w, \widehat{\mathbf{Q}}\mathbf{m} \cdot \mathbf{n} \right\rangle_{\partial K} + (d \mathcal{U}m, w)_K = 0,$$

$$(3.33c) \quad \langle \mathcal{U}m, \mu \rangle_{\partial K} = \langle \mathbf{m}, \mu \rangle_{\partial K},$$

for all $\mathbf{v} \in \mathbf{V}(K)$, $w \in W(K)$, and $\mu \in M(\partial K)$, and

$$(3.34a) \quad (c\mathcal{Q}f, \mathbf{v})_K - (\mathcal{U}f, \operatorname{div} \mathbf{v})_K = 0,$$

$$(3.34b) \quad -(\operatorname{grad} w, \mathcal{Q}f)_K + \left\langle w, \widehat{\mathbf{Q}}f \cdot \mathbf{n} \right\rangle_{\partial K} + (d \mathcal{U}f, w)_K = (f, w),$$

$$(3.34c) \quad \langle \mathcal{U}f, \mu \rangle_{\partial K} = 0,$$

$\mathbf{v} \in \mathbf{V}(K)$, $w \in W(K)$, and $\mu \in M(\partial K)$.

3. The space of approximate traces is given by $\mathbf{M}_h = \mathcal{M}_{h,k-1}$.

Having completed the definition of the main ingredients of the method, we now verify the assumptions of Theorem 2.4.

Sufficient conditions under which Assumption 2.1 on the existence and the uniqueness of the local solvers hold are given next.

PROPOSITION 3.9. *For $k = 1$ and arbitrary n and for odd $k > 1$ and $n = 2$, local solvers (3.33) and (3.34) have unique solutions.*

Proof. We prove only the result for the first local solver, since the other can be proven in a similar way. Since (3.33) is a square system, it suffices to prove that if $\mathbf{m} = 0$, then $\mathbf{Qm} = 0$, $\mathcal{U}m = 0$, and $\widehat{\mathbf{Q}}\mathbf{m} = 0$. Choosing $\mathbf{v} = \mathbf{Qm}$ and $w = \mathcal{U}m$,

adding (3.33a) and (3.33b), and integrating by parts, we get

$$(c\mathbf{Qm}, \mathbf{Qm})_K + (d\mathcal{U}m, \mathcal{U}m)_K + \left\langle \mathcal{U}m, \left(\widehat{\mathbf{Qm}} - \mathbf{Qm} \right) \cdot \mathbf{n} \right\rangle_{\partial K} = 0.$$

If $\mathbf{m} = 0$, (3.33c) implies that $\langle \mathcal{U}m, \mu \rangle_{\partial K} = 0$ for all $\mu \in M(\partial K)$. Since $(\widehat{\mathbf{Qm}} - \mathbf{Qm}) \cdot \mathbf{n} \in M(\partial K)$, then the last term on the left-hand side above is zero, and hence, $\mathbf{Qm} = 0$ and $d\mathcal{U}m = 0$. Substituting this into (3.33a), we have

$$0 = (\mathcal{U}m, \operatorname{div} \mathbf{v})_K = -(\mathbf{grad} \mathcal{U}m, \mathbf{v})_K \quad \text{for all } \mathbf{v} \in \mathcal{P}_{k-1}(K)^n,$$

where, while integrating by parts, we have again used that $\langle \mathcal{U}m, \mathbf{v} \cdot \mathbf{n} \rangle_{\partial K} = 0$. Thus, $\mathbf{grad} \mathcal{U}m$ vanishes, so $\mathcal{U}m$ is a constant function, and $\langle \mathcal{U}m, \mu \rangle_{\partial K} = 0$ implies that it vanishes identically.

It remains to show that $\widehat{\mathbf{Qm}} \cdot \mathbf{n}$ also vanishes. Since both \mathbf{Qm} and $d\mathcal{U}m$ vanish, (3.33b) implies that

$$(3.35) \quad \left\langle w, \widehat{\mathbf{Qm}} \cdot \mathbf{n} \right\rangle_{\partial K} = 0 \quad \text{for all } w \in \mathcal{P}_k(K).$$

For $k = 1$, that is, for Crouzeix–Raviart nonconforming finite elements, the result follows easily for any dimension $n \geq 2$. Indeed, let $\widehat{\mathbf{Qm}} \cdot \mathbf{n}|_{e_j} = a_j$ for some constants $a_j, j = 1, \dots, n + 1$. Let $w \in \mathcal{P}_1(K)$ be a linear function on K which takes values a_j at the centroids of the faces e_j of $K, j = 1, \dots, n + 1$. Then $0 = \langle w, \widehat{\mathbf{Qm}} \cdot \mathbf{n} \rangle_{\partial K} = \sum_{j=1}^{n+1} |e_j| a_j^2$ implies $a_j = 0$ for all faces, that is, $\widehat{\mathbf{Qm}} \cdot \mathbf{n} = 0$.

Finally, we show the same for k odd and $n = 2$. Let e_1, e_2 , and e_3 denote the three edges of K , and let $L_i^{(j)}$ denote the i th Legendre polynomial mapped affinely to e_j from $[-1, 1]$. Assume that the first vertex of the edge e_j is mapped to the point -1 , and that, as we go from its first to its second vertex, the triangle K is to our left. Since $\widehat{\mathbf{Qm}} \cdot \mathbf{n}|_{e_j} \in \mathcal{P}_{k-1}(e_j)$, we can write

$$\left(\widehat{\mathbf{Qm}} \right)_K \cdot \mathbf{n}_K|_{e_j} = \sum_{i=0}^{k-1} a_i^{(j)} L_i^{(j)}.$$

Note that when i is even, $L_i^{(j)}$ takes the same value at the endpoints of e_j . Therefore, for any even i , we can choose a w in (3.35) such that $w|_{e_1} = L_i^{(1)}, w|_{e_2} = -L_i^{(2)}$, and $w|_{e_3} = L_i^{(3)}$ (because with these choices $w|_{\partial K}$ is continuous). Then (3.35) implies that the coefficient $a_i^{(1)}$ vanishes. Repeating the argument for all edges, we find that $a_i^{(j)} = 0$ for all even i and $j = 1, 2, 3$. Next, for odd i , choose w such that $w|_{e_1} = L_i^{(1)}, w|_{e_2} = L_{i-1}^{(2)}$, and $w|_{e_3} = -L_i^{(3)}$. Since k is odd, these choices make $w|_{\partial K}$ continuous, so such a w can be found. With this w , (3.35) now gives that $a_i^{(1)} = 0$ for all odd i as well. Repeating this argument for other edges, we find all coefficients to be zero, so $\widehat{\mathbf{Qm}}$ vanishes. \square

Conservativity condition (2.6) with $M_h = \mathcal{M}_{h,k-1}$ clearly implies strong conservativity. Using Theorem 2.1 and noting that the unknown fluxes $\widehat{\mathbf{Q}}$ cancel off in weak formulation (2.9), by boundary condition (3.33c) for the local solver, we have that the bilinear form is symmetric:

$$a_h(\eta, \mu) = (c\mathbf{Q}\eta, \mathbf{Q}\mu)_{\mathcal{T}_h} + (d\mathcal{U}\eta, \mathcal{U}\mu)_{\mathcal{T}_h}.$$

Its positive definiteness will follow from Theorem 2.4 once Assumptions 2.2 and 2.3 are verified, which we do next.

PROPOSITION 3.10. *Assumption 2.2 on the positive semidefiniteness of the local solvers and Assumption 2.3, the gluing condition, are satisfied with $M(\partial K)$ defined as in (3.32).*

Proof. First, we show that condition (2.13a) holds. Taking $\mathbf{v} = \mathbf{Qm}$ in (3.33a), $w = \mathcal{U}m$ in (3.33b), and adding the equations, we get, after a few simple algebraic manipulations, that

$$\begin{aligned} -\langle \mathbf{m}, \widehat{\mathbf{Q}}\mathbf{m} \cdot \mathbf{n} \rangle_{\partial K} &= (c \mathbf{Qm}, \mathbf{Qm})_K + (d \mathcal{U}m, \mathcal{U}m)_K + \langle (\widehat{\mathbf{Q}}\mathbf{m} - \mathbf{Qm}) \cdot \mathbf{n}, \mathcal{U}m - m \rangle_{\partial K} \\ &= (c \mathbf{Qm}, \mathbf{Qm})_K + (d \mathcal{U}m, \mathcal{U}m)_K, \end{aligned}$$

by boundary condition (3.33c) for the local solver. This implies that (2.13a) of Assumption 2.2 is satisfied.

Now, we prove condition (2.13b). If $\langle \mathbf{m}, \widehat{\mathbf{Q}}\mathbf{m} \cdot \mathbf{n} \rangle_{\partial K} = 0$, then $\mathbf{Qm}|_K = 0$ and (3.33a) becomes

$$(\mathbf{grad} \mathcal{U}m, \mathbf{v})_K = \langle \mathcal{U}m - m, \mathbf{v} \cdot \mathbf{n} \rangle_{\partial K} = 0 \quad \text{for all } \mathbf{v} \in \mathbf{V}(K).$$

This implies that that $\mathcal{U}m$ is a constant. This shows that condition (2.13b) of Assumption 2.2 is satisfied.

Assumption 2.3 is trivially satisfied, and this completes the proof. \square

In Tables 3.1 and 3.2, we give the simplified weak formulation of the NC-H method under the further assumption that $c(\mathbf{x})$ is a constant matrix on each K in \mathcal{T}_h . In this case, we can show that the original NC method is an NC-H method. To see why, first observe that by summing up the last equation of the local solvers, we find that $u_h = \mathcal{U}\lambda_h + \mathcal{U}g_h + \mathcal{U}f$ satisfies

$$\langle \llbracket u_h \rrbracket, \mu \rangle_e = 0 \quad \text{for all } \mu \in \mathcal{P}_{k-1}(e)$$

for all interior faces e , so the weak continuity constraints of the discontinuous method are satisfied. Now, (2.12a) and (2.12c) become $(c \mathbf{Q}\lambda_h + \mathbf{grad} \mathcal{U}\lambda_h, \mathbf{v})_{\mathcal{T}_h} = 0$ and $(c \mathbf{Q}f + \mathbf{grad} \mathcal{U}f, \mathbf{v})_{\mathcal{T}_h} = 0$, which gives

$$\mathbf{q}_h = \mathbf{Q}\lambda_h + \mathbf{Q}g_h + \mathbf{Q}f = -a\mathbf{grad}(\mathcal{U}\lambda_h + \mathcal{U}g + \mathcal{U}f) = -a\mathbf{grad}u_h.$$

Then (2.8a) implies

$$(a\mathbf{grad}u_h, \mathbf{grad}v_h)_{\mathcal{T}_h} + (du_h, v_h)_{\mathcal{T}_h} = (f, v_h)$$

for all $v_h \in \{w : w \in W_h, \langle \llbracket w \rrbracket, \mu \rangle_{\mathcal{E}_h^\circ} = 0 \text{ for all } \mu \in M_h \text{ and } \langle w, \mathbf{m} \rangle_{\partial\Omega} = 0 \text{ for all } \mathbf{m} \in \mathbf{M}_h\}$, which is the familiar primal form of this nonconforming method. Note that although the information in g_h disappears from the right-hand side above, it is contained in u_h as $u_h = \mathcal{U}\lambda_h + \mathcal{U}g_h + \mathcal{U}f$.

Let us end this subsection by pointing out that, in the case of lowest order polynomials $k = 1$ and for the case in which $d = 0$ and both c and f are constant on each simplex K of triangulation \mathcal{T}_h , our hybridization framework allows us to recover a well-known relationship between the RT method of lowest degree and the nonconforming method [4, 47]. Let us sketch how to obtain it. In this case, we can easily show that local solver \mathbf{Qm} is the *same* for both this nonconforming method and that of the RT method of lowest degree; see the computation of the RT method in [26]. Since we also have that $\widehat{\mathbf{Q}}\mathbf{m} \cdot \mathbf{n} = \mathbf{Qm} \cdot \mathbf{n}$, we can conclude that the stiffness matrix associated with bilinear form $a_h(\cdot, \cdot)$ of both methods is also the same—if the degrees

of freedom for the numerical traces are the barycenters of the faces. Moreover, since the average on each simplex of the local solver $\mathcal{U}m$ coincides with the local solver $\mathcal{U}m$ of the RT method under consideration, the matrix associated with linear form $b_h(\cdot)$ is also the same for both methods. Of course, in both cases, we take g_h at the barycenter of each face $e \in \mathcal{E}_h^\partial$ to be the average of g on the face e . By Theorem 2.1, the degrees of freedom of the approximate traces are the same for both methods. The above-mentioned relation between the two methods now easily follows from the definition of approximate solutions (2.5).

4. Other novel methods. In this section, we build on the work done in the previous section and construct what are perhaps the three most important examples of methods of the unifying framework. The first is a class of methods employing different local solvers in different parts of the domain, which can easily deal with nonconforming meshes. The second is an RT method that can handle hanging nodes. The third is the family of EDG methods; they are constructed from already known hybridized methods in this unified framework in order to reduce their computational complexity. As for the examples of the previous section, we assume that the mesh is simplicial; however, we do not assume it to be necessarily conforming.

4.1. A class of hybridizable methods well suited for adaptivity. We introduce here a class of hybridizable methods able to use different local solvers in different elements and to easily handle nonconforming meshes. They are thus ideal to use with adaptive strategies. After introducing the methods, we prove that they are all well defined. We then discuss their main advantages and give several examples.

To define the methods, we need to specify the numerical fluxes, the local finite element spaces, and the space of approximate traces:

1. For any simplex $K \in \mathcal{T}_h$, we take

$$(4.1) \quad \widehat{\mathbf{Q}}m = \mathbf{Q}m + \tau_K(\mathcal{U}m - m)\mathbf{n}, \quad \widehat{\mathbf{Q}}f = \mathbf{Q}f + \tau_K(\mathcal{U}f)\mathbf{n} \quad \text{on } \partial K;$$

the function τ_K is allowed to change on ∂K .

2. The local space $\mathbf{V}(K) \times W(K)$ can be any of the following:

$$(4.2a) \quad (\mathcal{P}_{k(K)}(K)^n + \mathbf{x} \mathcal{P}_{k(K)}(K)) \times \mathcal{P}_{k(K)}(K),$$

where $k(K) \geq 0$ and $\tau_K \geq 0$ on ∂K ,

$$(4.2b) \quad \mathcal{P}_{k(K)}(K)^n \times \mathcal{P}_{k(K)-1}(K),$$

where $k(K) \geq 1$ and $\tau_K \geq 0$ on ∂K ,

$$(4.2c) \quad \mathcal{P}_{k(K)}(K)^n \times \mathcal{P}_{k(K)}(K),$$

where $k(K) \geq 0$ and $\tau_K > 0$ on at least one face e of K ,

$$(4.2d) \quad \mathcal{P}_{k(K)-1}(K)^n \times \mathcal{P}_{k(K)}(K),$$

where $k(K) \geq 1$ and $\tau_K > 0$ on ∂K .

3. The space of approximate traces is

$$(4.3a) \quad \mathbf{M}_h = \mathfrak{M}_h \cap \left\{ \mu : \mu|_{\partial K} \in \mathcal{C}(\overline{\{\mathbf{x} \in \partial K : \tau_K(\mathbf{x}) = \infty\}}}) \quad \forall K \in \mathcal{T}_h \right\},$$

where

$$(4.3b) \quad \mathfrak{M}_h := \{ \mu \in L^2(\mathcal{E}_h) : \mu|_e \in \mathcal{P}_{k(e)}(e) \text{ for all } e \in \mathcal{E}_h^\circ \}.$$

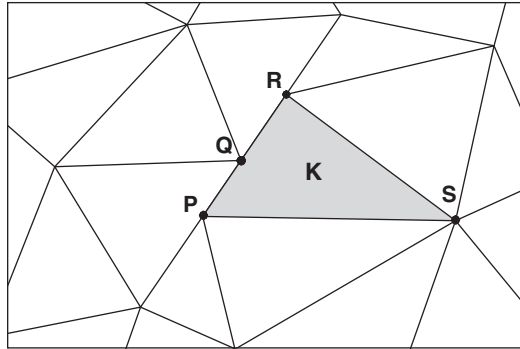


FIG. 4.1. The interior edges $e = \mathbf{PQ}$ and $e = \mathbf{QR}$ are contained in the face \mathbf{PR} of the element K . Assumption 4.1 is satisfied for this element if $\tau_K|_{\partial K} \in [0, \infty]$ and if $\tau_K|_{\mathbf{PQ}}$ and $\tau_K|_{\mathbf{QR}}$ are taken in $(0, \infty)$.

Here, if $e = \partial K^+ \cap \partial K^-$, we set

$$(4.3c) \quad k(e) := \begin{cases} \max\{k(K^+), k(K^-)\} & \text{if } \tau^+ < \infty \text{ and } \tau^- < \infty, \\ k(K^+) & \text{if } \tau^+ = \infty \text{ and } \tau^- < \infty, \\ k(K^-) & \text{if } \tau^+ < \infty \text{ and } \tau^- = \infty, \\ \min\{k(K^+), k(K^-)\} & \text{if } \tau^+ = \infty \text{ and } \tau^- = \infty, \end{cases}$$

and take $g_h = I_h g$ for some interpolation operator I_h into \mathbf{M}_h .

Note that the choice $\tau = \infty$ on some interior faces $e \in \mathcal{E}_h^\circ$ is allowed. We follow the convention that in this case, the definition of the local solvers has to be modified as for the limiting cases of LDG-H methods; see subsection 3.4. The definition of the methods is competed with the following assumption on the values of the stabilization parameter τ .

Assumption 4.1. For each element $K \in \mathcal{T}_h$ and each interior face $e \in \mathcal{E}_h^\circ$ on ∂K , $\tau_K|_e \in [0, \infty]$ and

$$(4.4) \quad \tau_K|_e \in (0, \infty) \text{ if } e \text{ is not a face of } K.$$

Let us briefly discuss this assumption. First, let us recall the difference between an interior face $e \in \mathcal{E}_h^\circ$ and the faces of the simplexes of the triangulation. Each simplex K in the partition \mathcal{T}_h has $n + 1$ faces determined by its vertices. On the other hand, if e is an interior face, we have that $e = \partial K^+ \cap \partial K^-$ for some elements K^+ and K^- in \mathcal{T}_h . We thus see that, for nonconforming meshes, although each interior face e is contained in a face of K^+ and a face of K^- , it is not necessarily a face of K^+ or K^- . See an example in Figure 4.1. The main motivation of the above assumption can now be easily seen. Indeed, take any $K \in \mathcal{T}_h$. If $e \subset \partial K$ is a face in \mathcal{E}_h° which is *not a face* of K , then the above assumption forces us to take the numerical trace corresponding to an LDG-H method; in this way, the nonconformity of the mesh can be dealt with in a very natural way. If, on the contrary, e is actually a face, the assumption allows us to take either $\tau_K = 0$, $\tau_K \in (0, \infty)$, or even $\tau_K = \infty$. In this way, the verification of Assumptions 2.1, 2.2, and 2.3 becomes extremely easy, as we are going to see next.

Next, we show that the approximate solution (\mathbf{q}_h, u_h) , (2.5), provided by this method is well defined.

PROPOSITION 4.1. Consider the method defined by (4.1), (4.2), and (4.3), and let Assumption 4.1 hold. Then Assumption 2.1 on the existence and the uniqueness of

the local solvers, Assumption 2.2 on the positive semidefiniteness of the local solvers, and Assumption 2.3, the gluing condition, hold with

$$(4.5) \quad M(\partial K) = \{ \mu : \text{ on any face } e \in \mathcal{E}_h^\circ \text{ on } \partial K, \mu|_e \in \mathcal{P}_{k(K)}(e) \text{ if } \tau_K|_e = 0, \\ \text{ and } \mu|_e \in L^2(e) \text{ if } \tau_K|_e > 0 \}.$$

Proof. Thanks to Theorem 2.4, we have only to satisfy Assumptions 2.1, 2.2, and 2.3. We begin by verifying Assumption 2.1 on the existence and the uniqueness of the local solvers. Let K be an arbitrary simplex of triangulation \mathcal{T}_h . Then, as discussed above, by condition (4.4), we have either $\tau_K = 0$, $\tau_K \in (0, \infty)$, or $\tau_K = \infty$ on each of the faces of each simplex K of triangulation \mathcal{T}_h . As a consequence, the fact that the local solvers are well defined can be easily obtained by a straightforward modification of the proofs of similar results for the LDG-H methods, Proposition 3.5, and the CG-H method, Proposition 3.5. For this reason, we do not present here the proof. However, let us note that whenever $\tau_K|_e = \infty$, we strongly impose a Dirichlet boundary condition, and so the space of approximate traces restricted to ∂K and local space $W(K)$ must satisfy the following compatibility condition:

$$\{ \mu|_S : \mu \in \mathbf{M}_h \} \subset \{ w|_S : w \in W(K) \}, \quad \text{where } S := \overline{\{ \mathbf{x} \in \partial K : \tau_K(\mathbf{x}) = \infty \}}.$$

This condition can be easily verified by noting that, if $\tau = \infty$ on the interior face $e \in \mathcal{E}_h^\circ$, then e must be a face of K by the conditions on the stabilization parameters (4.4), and since, by the definition of $k(e)$, (4.3c), we have that $k(e) \leq k(K)$.

Next, let us prove that Assumption 2.2 on the positive semidefiniteness of the local solvers is satisfied with $M(\partial K)$ as in (4.5). For choice (4.2a), it is easy to see that it follows from Proposition 3.1 and from the definition of $k(e)$, (4.3c). For the remaining choices, the result follows from Proposition 3.3 and the definition of $k(e)$, (4.3c).

Assumption 2.3, the gluing condition, also follows by using the arguments of the previous section. Indeed, for an interior face $e = \partial K^+ \cap \partial K^-$, if τ^+ or τ^- is positive, the result trivially follows from condition (4.3c) and the fact that on e , one of the projections $P_{\partial K^+}$ or $P_{\partial K^-}$ becomes the identity by the definition of $k(e)$, (4.3c). It remains to consider the case $\tau^+ = \tau^- = 0$. By (4.3c), either $k(K^+)$ or $k(K^-)$ equals $k(e)$, say, $k(e) = k(K^+)$. Then we immediately have that $P_{\partial K^+} \mu = \mu$. This completes the proof. \square

Next, let us discuss the main features of these methods.

(i) **Variable degree approximation spaces on conforming meshes.** The RT-H, BDM-H, and LDG-H methods considered in the previous section used a single local solver in each of the elements K of the conforming triangulation \mathcal{T}_h . A variable-degree version of each of these methods is a particular case of the class of methods presented here. Note that the case of the variable degree RT method, introduced and analyzed in [27], is exactly the variable-degree version of the method using the RT method as local solvers.

(ii) **Automatic coupling of different methods on conforming meshes.** The methods presented here allow for the use of different local solvers in different elements K of \mathcal{T}_h , which are then automatically coupled. For example, if we are working with the RT, LDG, and CG local solvers, the conservativity condition implicitly imposes

the following expressions for the numerical traces:

$$\begin{aligned} \widehat{u}_h &= u_h|_{\Omega_{LDG}} + \frac{1}{\tau_{LDG}} \llbracket \mathbf{q}_h \rrbracket, \quad \widehat{\mathbf{q}}_h = \mathbf{q}_h|_{\Omega_{RT}} && \text{(coupling RT and LDG),} \\ \widehat{u}_h &= u_h|_{\Omega_{CG}}, \quad \widehat{\mathbf{q}}_h = \mathbf{q}_h|_{\Omega_{LDG}} + \tau_{LDG} \llbracket u_h \rrbracket && \text{(coupling of LDG and CG),} \\ \widehat{u}_h &= u_h|_{\Omega_{CG}}, \quad \widehat{\mathbf{q}}_h = \mathbf{q}_h|_{\Omega_{RT}}. && \text{(coupling of CG and RT).} \end{aligned}$$

Note that this coupling holds even for nonconforming meshes.

It is interesting to compare the above couplings with other couplings in the available literature, namely,

$$\begin{aligned} \widehat{u}_h &= u_h|_{\Omega_{LDG}}, \quad \widehat{\mathbf{q}}_h = \mathbf{q}_h|_{\Omega_{RT}} + C_{11} \llbracket u_h \rrbracket && \text{(coupling of RT and LDG in [23]),} \\ \widehat{u}_h &= u_h|_{\Omega_{CG}}, \quad \widehat{\mathbf{q}}_h = \mathbf{q}_h|_{\Omega_{LDG}} + \tau_{LDG} \llbracket u_h \rrbracket && \text{(coupling of LDG and CG in [48]).} \end{aligned}$$

(iii) **Mortaring capabilities (for nonconforming meshes).** One of the advantageous features of DG methods is their ability to handle nonconforming meshes; see [52] for an application to structural mechanics. The methods under consideration incorporate this *mortaring* ability thanks to the very form that the numerical trace of the flux on ∂K takes on an interior face $e \in \mathcal{E}_h^\circ$ which is not a face of K , and thanks to the definition of the stabilization parameter τ therein. Let us give two examples.

If we have a conforming mesh, we can take the first choice of local spaces (4.2a) and set $\tau \equiv 0$. The resulting method, as we have seen, is nothing but the RT-H method. We can easily modify this method to handle nonconforming meshes by simply taking $\tau_K \in (0, \infty)$ on every interior face $e \in \mathcal{E}_h^\circ$ which is not a face of K , and otherwise, taking $\tau_K = 0$. Thus, the resulting method can be considered as a variation of the RT method, which is capable of handling nonconforming meshes.

We can do something similar with the CG method. Indeed, if the mesh is conforming, we can take the last choice of local spaces (4.2d) and set $\tau \equiv \infty$ to obtain the hybridized CG method. For nonconforming meshes, we can slightly modify the method by simply taking $\tau_K \in (0, \infty)$ on every interior face $e \in \mathcal{E}_h^\circ$ which is not a face of K , and otherwise, taking $\tau_K = \infty$. The resulting method is thus a variation of the CG method capable of handling nonconforming meshes. It constitutes an alternative to the coupling of the CG and the LDG methods proposed in [48] to deal with nonmatching meshes.

(iv) **The conservativity condition.** Let us end by noting that the stiffness matrix associated to the approximate trace λ_h is always symmetric and positive definite. Moreover, on the interior faces on which $\tau < \infty$, the conservativity condition is enforced strongly.

4.2. The RT method on meshes with hanging nodes. Consider the case of the variable degree RT-H method. The method is obtained by taking the numerical traces as in (4.1) with $\tau \equiv 0$, the local space (4.2a), and the multiplier space as in (4.3). This method does not belong to the family of methods described in the previous subsection because our choice of stabilization parameter does not satisfy condition (4.4). Thus, to ensure the existence and the uniqueness of the approximate solution, we have to impose special conditions on the meshes and *link* the definition of $k(K)$ to the structure of the mesh.

Let us illustrate how to do this in the two-dimensional case. The meshes \mathcal{T}_h we consider are constructed as follows. First, construct a conforming triangulation of Ω , $\mathcal{T}_h^{(0)} := \{K^{(0)}\}$. Then, take a subset of that triangulation $\mathcal{T}_h^{(0,1)}$ and divide each of its triangles into four congruent triangles; the set of those triangles is denoted by

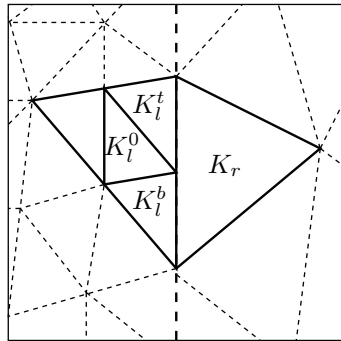


FIG. 4.2. In the presence of hanging nodes like the above, an RT-H method with the spaces on edges chosen to have the maximum degree from either side is well defined if $k(K_l^t) \geq k(K_r)$ and $k(K_l^b) \geq k(K_r)$. The degree $k(K_l^0)$ can be arbitrarily chosen.

$\mathcal{T}_h^{(1)}$. Next, for $j = 2, \dots, \ell$, given the set $\mathcal{T}_h^{(j-1)}$, pick the subset $\mathcal{T}_h^{(j-1,j)}$ and create the set of smaller triangles $\mathcal{T}_h^{(j)}$. The simple case $\ell = 1$ is illustrated in Figure 4.2. Finally, we establish a link between the mesh and the definition of the polynomial degree of the RT method on the triangle K , $k(K)$, as follows. If $e = \partial K^+ \cap \partial K^-$, we take $k(e) := \max\{k(K^+), k(K^-)\}$ and require that

$$(4.6) \quad \text{if } e \text{ is not an edge of } K^-, \text{ then } k(K^+) \geq k(K^-).$$

Next, we show that the method is well defined.

PROPOSITION 4.2. *The variable-degree RT-H method on meshes with hanging nodes as described above is uniquely solvable.*

Proof. If we proceed exactly as in Proposition 3.1, we can see that Assumption 2.1 on the existence and the uniqueness of the local solvers is verified and that Assumption 2.2 on the positive semidefiniteness of the local solvers is also verified *provided* we change the definition of the set $M(\partial K)$ to

$$M(\partial K) = \{\mu : \mu|_e \in \mathcal{P}_{k(e)}(e) \text{ for all edges } e \text{ of } \partial K\}.$$

The result follows if we prove that there is only one solution $\lambda_h \in M_h$ of weak formulation (1.7).

To do that, we proceed exactly as in the proof of Theorem 2.4. First, since Assumption 2.1 holds, we have that system (2.9) is well defined. Next, we show that $a_h(\mu, \mu) = 0$ for $\mu \in M_h$ implies $\mu = 0$. By Assumption 2.2, we readily obtain that, for any given $K \in \mathcal{T}_h$, we have that, on ∂K ,

$$C_K = P_{\partial K} \mu,$$

where $P_{\partial K}$ is the L^2 -projection into $M(\partial K)$ as defined above. It remains to show that this implies that μ is a constant on \mathcal{E}_h . To do that, we use the structure of the meshes and the definition of $k(K)$ for all $K \in \mathcal{T}_h$.

We proceed as follows. We claim that, for $j = \ell, \ell - 1, \dots, 0$, we have that $\mu|_{\partial K}$ is a constant for all $K \in \mathcal{T}_h^{(j)}$. This immediately implies that μ is a constant on \mathcal{E}_h , and since $\mu|_{\partial\Omega} = 0$, that $\mu = 0$ on \mathcal{E}_h .

It remains to prove the claim. We proceed by induction on j . Let us prove the inductive hypothesis for $j = \ell$. Let K be any triangle in $\mathcal{T}_h^{(\ell)}$ and pick any of its edges

e . If the edge e lies on $\partial\Omega$, we immediately have that $\mu = C_K = 0$. If $e = \partial K \cap \partial K'$ for some triangle $K' \in \mathcal{T}_h^{(\ell)}$, we proceed as in the proof of Theorem 2.4 to conclude that

$$\mu = C_K = C_{K'} \quad \text{on } e.$$

The only other remaining possibility, by construction of triangulation $\mathcal{T}_h^{(\ell)}$, is that $e = \partial K \cap \partial K'$ for some triangle $K' \in \mathcal{T}_h^{(\ell')}$, with $\ell' < \ell$. In this case, e is not an edge of K' , and, by condition (4.6) on $k(K)$, we have that $k(K) \geq k(K')$ and hence $k(e) := \max\{k(K), k(K')\} = k(K)$. This implies that

$$\mu = C_K \quad \text{on } e.$$

Since edge e was picked arbitrarily, we conclude that $\mu|_{\partial K}$ is a constant, as wanted.

Now, let us assume that the inductive hypothesis holds for $j = J$ and let us prove it also holds for $j = J - 1$. Let K be any triangle in $\mathcal{T}_h^{(J-1)}$ and pick any of its edges e . Since, by the inductive hypothesis, $\mu|_{\partial K}$ is a constant for all $K \in \mathcal{T}_h^{(J)}$, we have that $\mu|_{\partial K}$ is a constant for all $K \in \mathcal{T}_h^{(J-1, J)}$, since, by construction, each of the triangles in $\mathcal{T}_h^{(J-1, J)}$ is subdivided in four congruent triangles in $K \in \mathcal{T}_h^{(J)}$. Hence, $\mu = C_K$ on e if the edge e lies in the border of any triangle in $K \in \mathcal{T}_h^{(J-1, J)}$. To finish the proof, we need only to prove the same result in the remaining three cases: (i) if the edge e lies on $\partial\Omega$, (ii) if $e = \partial K \cap \partial K'$ for some triangle $K' \in \mathcal{T}_h^{(J-1)} \setminus \mathcal{T}_h^{(J-1, J)}$, and (iii) if $e = \partial K \cap \partial K'$ for some triangle $K' \in \mathcal{T}_h^{(J'-1)}$, with $J' < J$. This can be done exactly as in the previous step. \square

4.3. The EDG methods. Now we show that new methods [33] can be immediately generated from already existing hybridized methods by simply reducing the space of their approximate traces. The main interest of these EDG methods, introduced in the setting of shells problems in [43], stems from the further reduction in globally coupled unknowns achieved by reducing the approximate trace space M_h .

To construct such methods, we begin with selecting any method defined by uniquely solvable local problems (2.3), (2.4), and conservativity condition (2.6), yielding a unique approximate trace λ_h . Then, by Theorem 2.1, $\lambda_h \in M_h$ is the only solution of the weak formulation

$$(4.7) \quad a_h(\lambda_h, \mu) = b_h(I_h g; \mu) \quad \text{for all } \mu \in M_h,$$

where we are writing $b_h(I_h g; \mu)$ instead of just $b_h(\mu)$ in order to stress its dependency on $I_h g \in M_h$. We now define an EDG method by replacing the original approximate trace space M_h by a subspace \tilde{M}_h . This forces us to replace $I_h g \in M_h$ by $\tilde{I}_h g \in \tilde{M}_h$ and to change the conservativity condition, but the local solvers remain the same. Now, define the operator $\mathcal{J}_h : \tilde{M}_h \rightarrow M_h$ as the identity operator representing the natural *embedding* of \tilde{M}_h into M_h , hence, the name of these methods, and set

$$\tilde{M}_h := \left\{ \tilde{\mu} \in \tilde{M}_h : \tilde{\mu} = 0 \text{ on } \partial\Omega \right\}.$$

Then by Theorem 2.1, the new conservativity condition is equivalent to

$$(4.8) \quad a_h \left(\mathcal{J}_h \tilde{\lambda}_h, \mathcal{J}_h \tilde{\mu} \right) = b_h \left(\mathcal{J}_h \tilde{I}_h g; \mathcal{J}_h \tilde{\mu} \right) \quad \text{for all } \tilde{\mu} \in \tilde{M}_h,$$

where $\tilde{\lambda}_h \in \tilde{M}_h$ is the new approximate trace. Note that we have that $\mathcal{J}_h \tilde{\mu}|_{\partial\Omega} = 0$ for all $\tilde{\mu} \in \tilde{M}_h$.

To show that this EDG method is well defined, it suffices to prove that homogeneous equation (4.8) has only a trivial solution. For simplicity, let us assume that $a_h(\cdot, \cdot)$ is symmetric and positive definite. Thus, taking $\tilde{\mu} = \tilde{\lambda}_h$, we get $a_h(\mathcal{J}_h \tilde{\lambda}_h, \mathcal{J}_h \tilde{\lambda}_h) = 0$. By the positive definiteness of $a_h(\cdot, \cdot)$, we have $\mathcal{J}_h \tilde{\lambda}_h = 0$, which implies $\tilde{\lambda}_h = 0$. Hence, (4.8) is uniquely solvable.

Now, let us show that it is very easy to obtain the equations for the EDG method once those of the original method have been obtained. Denote by $[\lambda_h]$ the vector of the degrees of freedom of the function λ_h with respect to some basis in M_h . Similarly, denote by $[\tilde{\lambda}_h]$ the vector of degrees of freedom of the function $\tilde{\lambda}_h$ in \tilde{M}_h . Equation (4.7) can be written in a matrix form as $A[\lambda_h] = b(I_h g)$, and if $[\mathcal{J}_h \tilde{\lambda}_h] = T[\tilde{\lambda}_h]$, then the equation for $[\tilde{\lambda}_h]$ is $T^t A T [\tilde{\lambda}_h] = T^t b(\mathcal{J}_h I_h g)$. Here T is the rectangular matrix representing the basis of \tilde{M}_h with respect to basis of M_h . Since $\tilde{M}_h \subset M_h$, if we use the Lagrange basis functions, T is nothing but a connectivity matrix whose entries are zeroes and ones, so it is extremely easy to compute.

Note that the above considerations continue to hold if \mathcal{J}_h is any injective operator from \tilde{M}_h into M_h such that $\mathcal{J}_h \tilde{\mu}|_{\partial\Omega} = 0$ for all $\tilde{\mu} \in \tilde{M}_h$. Thus, new methods can also be created by using spaces \tilde{M}_h that are not necessarily subspaces of M_h . The main task here would be to find the matrix T which represents the basis of \tilde{M}_h with respect to basis of M_h .

Let us give some examples of EDG methods. The first example of an EDG method was proposed in [43]: It is obtained from an LDG-H method using approximations of degree k in each variable by forcing the continuity of the traces. Thus, whereas the functions in the space of approximate traces for the LDG-H method M_h are discontinuous on the borders of the elements, the functions of \tilde{M}_h are continuous therein. This allows the method to be immediately incorporated into commercial codes. On the other hand, this also results in the degradation of the conservativity properties of the EDG method, which hold only weakly. In some cases, this induces a *degradation* in the approximating properties of the method as recently proven in [32].

Indeed, in that paper, it was shown that when the stabilization parameter τ is taken to be of order one, the EDG method converges with order k for \mathbf{q} and order $k + 1$ for u . This has to be contrasted with the fact that the original LDG-H method converges with order $k + 1$ in *both* variables; see [33]. Moreover, in this case, the LDG-H has superconvergence properties that allow us to compute, in an element-by-element fashion, a new approximation to u converging with order $k + 2$; see also [33]. Such property does *not* hold for the corresponding EDG method. Even more, numerical experiments show that the computational advantage of the EDG method does not compensate for its loss of accuracy. On the other hand, if the stabilization parameter τ is taken to be of order h^{-1} , both the EDG and the LDG-H methods converge with the same orders, namely, k in \mathbf{q} and order $k + 1$ in u .

The second example is associated with the constructions of subspaces \tilde{M}_h of M_h that could be required to be very smooth. For example, we could ask that they be not only continuous on \mathcal{E}_h° but \mathcal{C}^1 -continuous. This might be reasonable to do if the solution is very smooth and varies slowly in Ω .

The third and last example is associated to methods for nonmatching grids. Suppose that Ω is divided into two domains Ω_1 and Ω_2 independently meshed, and that we are using the variation of the CG method to handle nonconforming methods described in the first subsection. Then, all the interior faces e lying on the interface $\Gamma := \partial\Omega_1 \cap \partial\Omega_2$ must be computed and used to define the space of traces M_h . (This can be done, although it is a computational geometry tour de force, especially in

three dimensions; see [52].) To reduce the size of the interface space on Γ , we can alternately find a subspace of M_h of functions which are polynomials on the interior faces determined by, say, the vertices of the triangulation of Ω_1 lying on Γ .

5. Extensions and generalizations. Although in all examples we have given, only simplicial elements have been considered, this is not essential. Obviously, quadrilateral, prismatic, and other elements could be handled easily by DG methods. Furthermore, our framework is applicable for mixed methods using other types of elements; see [12].

Note also that the considered DG, mixed, CG, and nonconforming finite element methods used to define the local solvers are not the only choices. Stabilized, Petrov–Galerkin methods, boundary element, and even (if possible) the exact solution can be used as local solvers. For example, the hybridization of the discontinuous Petrov–Galerkin method can be found in [18, 19].

In what follows, we sketch how to extend our results to include Neumann boundary conditions and interface transmission conditions. We also extend them to DG methods using other stabilization mechanisms.

5.1. Other boundary and transmission conditions. The hybridization method proposed here can be easily extended to other types of boundary and transmission conditions.

Neumann boundary condition. For example, the case when on part $\partial\Omega_N$ of the boundary $\partial\Omega$ the Neumann boundary condition $\mathbf{q} \cdot \mathbf{n} = q_N$ is specified can be incorporated easily in the hybridization procedure. We simply require that the approximate trace λ_h belongs to

$$M_h = \{\mu \in M_h : \mu = 0 \text{ on } \partial\Omega_D\},$$

where $\partial\Omega_D := \partial\Omega \setminus \partial\Omega_N$ is the Dirichlet boundary and replace conservativity condition (2.6) by

$$\left\langle \mu, \left[\widehat{\mathcal{Q}}\lambda_h + \widehat{\mathcal{Q}}g_h + \widehat{\mathcal{Q}}f \right] \right\rangle_{\mathcal{E}_h} = \langle \mu, q_N \rangle_{\partial\Omega_N} \quad \text{for all } \mu \in M_h.$$

Transmission condition. To handle transmission condition $[[\mathbf{q}]] = t$ on the $(n - 1)$ -dimensional surface Γ_t , we simply have to write

$$\left\langle \mu, \left[\widehat{\mathcal{Q}}\lambda_h + \widehat{\mathcal{Q}}g_h + \widehat{\mathcal{Q}}f \right] \right\rangle_{\mathcal{E}_h} = \langle \mu, q_N \rangle_{\partial\Omega_N} + \langle \mu, t \rangle_{\Gamma_t} \quad \text{for all } \mu \in M_h,$$

where we are assuming that $\Gamma_t \subset \mathcal{E}_h$. This case is equivalent to having a right-hand side that is a δ -function with a support on Γ_t .

Jump condition. Now, we can add jump condition $[[u]] = \mathbf{j}$ on the $(n - 1)$ -dimensional surface Γ_j , where $\mathbf{j} \cdot \mathbf{n}$ is given. Then we take triangulation \mathcal{T}_h such that $\Gamma_j \subset \mathcal{E}_h$ and proceed as follows. Since the exact solution is double valued on Γ_j , that is, since its traces on Γ_j are $u^\pm := \{u\} + 1/2 \mathbf{n}^\pm \cdot \mathbf{j}$, we take the approximation to these traces to be $\lambda_h + 1/2 \mathbf{n}^\pm \cdot \mathbf{j}$ on Γ_j and define the function $(\mathcal{Q}\mathbf{m}_j, \mathcal{U}\mathbf{m}_j)$ as the solution of local solver (2.3), with \mathbf{m}_j given by

$$\mathbf{m}_j = \begin{cases} \frac{1}{2} \mathbf{n}_K \cdot \mathbf{j} & \text{on } \partial K \cap \Gamma_j, \\ 0 & \text{elsewhere.} \end{cases}$$

Then, we simply rewrite the conservativity condition as

$$\left\langle \mu, \left[\widehat{\mathcal{Q}}\lambda_h + \widehat{\mathcal{Q}}\mathbf{m}_j + \widehat{\mathcal{Q}}g_h + \widehat{\mathcal{Q}}f \right] \right\rangle_{\mathcal{E}_h} = \langle \mu, q_N \rangle_{\partial\Omega_N} + \langle \mu, t \rangle_{\Gamma_t} \quad \text{for all } \mu \in M_h.$$

We see that the global system for λ_h has the same matrix and a right-hand side that incorporates the data related to the boundary and interface conditions. This particular example shows the ease with which the hybridizable methods can handle various types of boundary and transmission conditions for the differential equation.

5.2. Hybridizable DG methods with other stabilization mechanisms.

For each finite element $K \in \mathcal{T}_h$, the LDG-H method uses on ∂K the numerical trace $\widehat{\mathbf{q}}_h = \mathbf{q}_h + \tau(u_h - \lambda_h)\mathbf{n}$ and the IP-H uses the numerical trace $\widehat{\mathbf{q}}_h = -\mathbf{a}\mathbf{grad}u_h + \tau(u_h - \lambda_h)\mathbf{n}$. However, these are not the only choices for numerical traces we could use to generate stabilization through the difference between u_h and λ_h . Indeed, in the unified analysis of DG methods [5], we see that we can also take $\widehat{\mathbf{q}}_h = \mathbf{q}_h + \boldsymbol{\alpha}_r((u_h - \lambda_h)\mathbf{n})$ for the Brezzi–Manzini–Marini–Pietra–Russo (BMMPR) method [13] and $\widehat{\mathbf{q}}_h = -\mathbf{a}\mathbf{grad}u_h + \boldsymbol{\alpha}_r((u_h - \lambda_h)\mathbf{n})$ for the Bassi–Rebay–Mariotti–Pedinotti–Savini (BRMPS) method [7]. Here, for any $\boldsymbol{\varphi} \in \mathbf{L}^2(\partial K)$, the vector $\boldsymbol{\alpha}_r(\boldsymbol{\varphi})$ is the element of $\mathbf{V}(K)$ such that

$$\begin{aligned} \boldsymbol{\alpha}_r(\boldsymbol{\varphi}) &= -\tau \mathbf{r}_{e,K}(\boldsymbol{\varphi}) \text{ on each face } e \text{ of } K, \\ (\mathbf{r}_{e,K}(\boldsymbol{\varphi}), \mathbf{v})_K &= -\langle \boldsymbol{\varphi}, \mathbf{v} \rangle_e \text{ for all } \mathbf{v} \in \mathbf{V}(K). \end{aligned}$$

It is not difficult to verify that results similar to those obtained for the LDG-H and IP-H methods can also be obtained for similar BMMPR-H and BRMPS-H methods, respectively. Let us briefly comment on a couple of interesting details. To fix ideas, we consider the BMMPR-H methods. For these methods, Theorem 2.1 holds with

$$\begin{aligned} a_h(\eta, \mu) &= (c \boldsymbol{\Omega}\eta, \boldsymbol{\Omega}\mu)_{\mathcal{T}_h} + (d \mathcal{U}\eta, \mathcal{U}\mu)_{\mathcal{T}_h} + \langle 1, \llbracket (\mathcal{U}\mu - \mu)(\boldsymbol{\alpha}_r((\mathcal{U}\eta - \eta)\mathbf{n})) \rrbracket \rangle_{\mathcal{E}_h}, \\ b_h(\mu) &= \langle g_h, (\boldsymbol{\Omega}\mu + \boldsymbol{\alpha}_r(\mathcal{U}\mu\mathbf{n})) \cdot \mathbf{n} \rangle_{\partial\Omega} + (f, \mathcal{U}\mu)_{\mathcal{T}_h}, \end{aligned}$$

provided $g_h|_{\mathcal{E}_h^\circ} = 0$. It is not difficult to see that bilinear form $a_h(\cdot, \cdot)$ is symmetric. Indeed, we have that

$$\begin{aligned} &\langle (\mathcal{U}\mu - \mu)\mathbf{n}, \boldsymbol{\alpha}_r((\mathcal{U}\eta - \eta)\mathbf{n}) \rangle_{\partial K} \\ &= - \sum_{e \text{ face of } K} \tau_K|_e \langle (\mathcal{U}\mu - \mu)\mathbf{n}, \mathbf{r}_{e,K}((\mathcal{U}\eta - \eta)\mathbf{n}) \rangle_e \\ &= + \sum_{e \text{ face of } K} \tau_K|_e (\mathbf{r}_{e,K}((\mathcal{U}\mu - \mu)\mathbf{n}), \mathbf{r}_{e,K}((\mathcal{U}\eta - \eta)\mathbf{n}))_K. \end{aligned}$$

The fact that bilinear form $a_h(\cdot, \cdot)$ is positive definite follows from Theorem 2.4 and a slight modification of Proposition 3.3; in it, we take $M(\partial K) = \{v : v|_e \in \mathcal{P}_k(e), e \in \partial K\}$. Note that Assumption 2.2 is then satisfied, since $\mathbf{r}_{e,K}(\boldsymbol{\varphi}) = \mathbf{0}$, if and only if the L^2 -projection of $\boldsymbol{\varphi}|_e$ into $\mathcal{P}_k(e)$ is zero.

Finally, note that the conservativity condition is enforced strongly. In this case, however, we do not have an explicit expression of the approximate trace λ_h in terms of (\mathbf{q}_h, u_h) as we have for the LDG-H methods in Proposition 3.4. Instead, we have only the relation

$$\llbracket \boldsymbol{\alpha}_r(\lambda_h \mathbf{n}) \rrbracket = \llbracket \boldsymbol{\alpha}_r(u_h \mathbf{n}) \rrbracket + \llbracket \mathbf{q}_h \rrbracket \quad \text{on } \mathcal{E}_h^\circ.$$

Let us end by noting that extensions of this work to other problems arising in continuum mechanics, fluid dynamics, and electromagnetism constitutes the subject of ongoing work.

Acknowledgment. The first author would like to thank Martin Vohralík for bringing to his attention reference [21].

REFERENCES

- [1] T. ARBOGAST AND Z. CHEN, *On the implementation of mixed methods as nonconforming methods for second-order elliptic problems*, Math. Comp., 64 (1995), pp. 943–972.
- [2] T. ARBOGAST, L.C. COWSAR, M.F. WHEELER, AND I. YOTOV, *Mixed finite element methods on nonmatching multiblock grids*, SIAM J. Numer. Anal., 37 (2000), pp. 1295–1315.
- [3] D. N. ARNOLD, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.
- [4] D. N. ARNOLD AND F. BREZZI, *Mixed and nonconforming finite element methods: Implementation, postprocessing and error estimates*, RAIRO Modél. Math. Anal. Numér., 19 (1985), pp. 7–32.
- [5] D. N. ARNOLD, F. BREZZI, B. COCKBURN, AND L. D. MARINI, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.
- [6] G. A. BAKER, *Finite element methods for elliptic equations using nonconforming elements*, Math. Comp., 31 (1977), pp. 45–59.
- [7] F. BASSI, S. REBAY, G. MARIOTTI, S. PEDINOTTI, AND M. SAVINI, *A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows*, in Proceedings of the 2nd European Conference on Turbomachinery Fluid Dynamics and Thermodynamics, Antwerpen, Belgium, Technologisch Instituut, 1997, pp. 99–108.
- [8] F. BEN BELGACEM AND Y. MADAY, *The mortar element method for three dimensional finite elements*, M2AN Math. Model. Numer. Anal., 31 (1997), pp. 289–302.
- [9] C. BERNARDI, Y. MADAY, AND A. T. PATERA, *Domain Decomposition by the mortar element method*, in Asymptotic and Numerical Methods for Partial Differential Equations with Critical Parameters, H. G. Kaper and M. Garbey, eds., Kluwer Academic Publishers, Norwell, MA, 1993, pp. 269–286.
- [10] J. H. BRAMBLE AND J. XU, *A local post-processing technique for improving the accuracy in mixed finite-element approximations*, SIAM J. Numer. Anal., 26 (1989), pp. 1267–1275.
- [11] F. BREZZI, J. DOUGLAS, JR., AND L. D. MARINI, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.
- [12] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
- [13] F. BREZZI, G. MANZINI, L. D. MARINI, P. PIETRA, AND A. RUSSO, *Discontinuous finite elements for diffusion problems*, in Atti Convegno in onore di F. Brioschi, Istituto Lombardo, Accademia di Scienze e Lettere, Milan, 1999, pp. 197–217.
- [14] A. BUFFA, T. J. R. HUGHES, AND G. SANGALLI, *Analysis of a multiscale discontinuous Galerkin method for convection-diffusion problems*, SIAM J. Numer. Anal., 44 (2006), pp. 1420–1440.
- [15] J. CARRERO, B. COCKBURN, AND D. SCHÖTZAU, *Hybridized, globally divergence-free LDG methods. Part I: The Stokes problem*, Math. Comp., 75 (2006), pp. 533–563.
- [16] P. CASTILLO, *Performance of discontinuous Galerkin methods for elliptic PDEs*, SIAM J. Sci. Comput., 24 (2002), pp. 524–547.
- [17] P. CASTILLO, B. COCKBURN, I. PERUGIA, AND D. SCHÖTZAU, *An a priori error analysis of the local discontinuous Galerkin method for elliptic problems*, SIAM J. Numer. Anal., 38 (2000), pp. 1676–1706.
- [18] P. CAUSIN AND R. SACCO, *A discontinuous Petrov–Galerkin method with Lagrangian multipliers for second order elliptic problems*, SIAM J. Numer. Anal., 43 (2005), pp. 280–302.
- [19] P. CAUSIN AND R. SACCO, *Hierarchical mixed hybridized methods for elliptic problems*, Comput. Methods Appl. Mech. Engrg., 198 (2009), pp. 1061–1073.
- [20] F. CELIKER AND B. COCKBURN, *Superconvergence of the numerical traces of discontinuous Galerkin and hybridized mixed methods for convection-diffusion problems in one space dimension*, Math. Comp., 76 (2007), pp. 67–96.
- [21] Z. CHEN, *Equivalence between and multigrid algorithms for nonconforming and mixed methods for second-order elliptic problems*, East-West J. Numer. Math., 4 (1996), pp. 1–33.
- [22] P. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
- [23] B. COCKBURN AND C. DAWSON, *Approximation of the velocity by coupling discontinuous Galerkin and mixed finite element methods for flow problems*, Comput. Geosci. (Special issue: Locally Conservative Numerical Methods for Flow in Porous Media), 6 (2002), pp. 502–522.

- [24] B. COCKBURN AND B. DONG, *An analysis of the minimal dissipation local discontinuous Galerkin method for convection-diffusion problems*, J. Sci. Comput., 32 (2007), pp. 233–262.
- [25] B. COCKBURN, B. DONG, AND J. GUZMÁN, *A superconvergent LDG-hybridizable Galerkin method for second-order elliptic problems*, Math. Comp., to appear.
- [26] B. COCKBURN AND J. GOPALAKRISHNAN, *A characterization of hybridized mixed methods for second order elliptic problems*, SIAM J. Numer. Anal., 42 (2004), pp. 283–301.
- [27] B. COCKBURN AND J. GOPALAKRISHNAN, *Error analysis of variable degree mixed methods for elliptic problems via hybridization*, Math. Comp., 74 (2005), pp. 1653–1677.
- [28] B. COCKBURN AND J. GOPALAKRISHNAN, *Incompressible finite elements via hybridization. Part I: The Stokes system in two space dimensions*, SIAM J. Numer. Anal., 43 (2005), pp. 1627–1650.
- [29] B. COCKBURN AND J. GOPALAKRISHNAN, *Incompressible finite elements via hybridization. Part II: The Stokes system in three space dimensions*, SIAM J. Numer. Anal., 43 (2005), pp. 1651–1672.
- [30] B. COCKBURN AND J. GOPALAKRISHNAN, *New hybridization techniques*, GAMM Mitt. Ges. Angew. Math. Mech., 2 (2005), pp. 154–183.
- [31] B. COCKBURN, J. GOPALAKRISHNAN, AND H. WANG, *Locally conservative fluxes for the continuous Galerkin method*, SIAM J. Numer. Anal., 45 (2007), pp. 1742–1776.
- [32] B. COCKBURN, J. GUZMÁN, S.-C. SOON, AND H. STOLARSKI, *Analysis of the embedded discontinuous Galerkin method for second-order elliptic problems*, submitted.
- [33] B. COCKBURN, J. GUZMÁN, AND H. WANG, *Superconvergent discontinuous Galerkin methods for second-order elliptic problems*, Math. Comp., 78 (2009), pp. 1–24.
- [34] B. COCKBURN AND C.-W. SHU, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.
- [35] M. COMODI, *The Hellan-Herrmann-Johnson method: Some new error estimates and postprocessing*, Math. Comp., 52 (1989), pp. 17–29.
- [36] M. CROUZEIX AND P. A. RAVIART, *Conforming and nonconforming finite element methods for solving the stationary stokes equations*, RAIRO Modél. Math. Anal. Numér., 7 (1973), pp. 33–75.
- [37] V. DOBREV, R. LAZAROV, P. VASSILEVSKI, AND L. ZIKATANOV, *Two-level preconditioning of discontinuous Galerkin method for second order elliptic problems*, Numer. Linear Algebra Appl., 13 (2006), pp. 753–770.
- [38] R. EWING, J. WANG, AND Y. YANG, *A stabilized discontinuous finite element method for elliptic problems*, Numer. Linear Algebra Appl., 10 (2003), pp. 83–104.
- [39] B. M. FRAEJIS DE VEUBEKE, *Displacement and equilibrium models in the finite element method*, in Stress Analysis, O. Zienkiewicz and G. Holister, eds., Wiley, New York, 1977, pp. 145–197.
- [40] V. GIRAULT, S. SUN, M. F. WHEELER, AND I. YOTOV, *Coupling discontinuous Galerkin and mixed finite element discretizations using mortar finite elements*, SIAM J. Numer. Anal., 46 (2008), pp. 949–979.
- [41] J. GOPALAKRISHNAN, *A Schwarz preconditioner for a hybridized mixed method*, Comput. Methods Appl. Math., 3 (2003), pp. 116–134.
- [42] J. GOPALAKRISHNAN AND J. E. PASCIAK, *Multigrid for the mortar finite element method*, SIAM J. Numer. Anal., 37 (2000), pp. 1029–1052.
- [43] S. GÜZEY, B. COCKBURN, AND H. STOLARSKI, *The embedded discontinuous Galerkin methods: Application to linear shells problems*, Internat. J. Numer. Methods Engrg., 70 (2007), pp. 757–790.
- [44] J. T. R. HUGHES, G. SCOVAZZI, P. B. BOCHEV, AND A. BUFFA, *A multiscale discontinuous Galerkin method with the computational structure of a continuous Galerkin method*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 2761–2787.
- [45] Y. A. KUZNETSOV AND M.F. WHEELER, *Optimal order substructuring preconditioners for mixed finite element methods on nonmatching grids*, East-West J. Numer. Math., 3 (1995), pp. 127–143.
- [46] J. L. LIONS AND E. MAGENES, *Nonhomogeneous Boundary Value Problems and Applications*, Springer-Verlag, Berlin, 1972.
- [47] L. D. MARINI, *An inexpensive method for the evaluation of the solution of the lowest order Raviart–Thomas mixed method*, SIAM J. Numer. Anal., 22 (1985), pp. 493–496.
- [48] I. PERUGIA AND D. SCHÖTZAU, *On the coupling of local discontinuous Galerkin and conforming finite element methods*, J. Sci. Comput., 16 (2001), pp. 411–433.
- [49] P. A. RAVIART AND J. M. THOMAS, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of Finite Element Method, Lecture Notes in Math. 606, I. Galligani and E. Magenes, eds., Springer-Verlag, New York, 1977, pp. 292–315.

- [50] B. RIVIÈRE AND M. F. WHEELER, *Coupling locally conservative methods for single-phase flow*, Comput. Geosci. (Special issue: Locally Conservative Numerical Methods for Flow in Porous Media), 6 (2002), pp. 269–284.
- [51] J.-E. ROBERTS AND J.-M. THOMAS, *Mixed and hybrid methods*, in Finite Element Methods, Part 1, Handb. Numer. Anal. II, P. G. Ciarlet and J. L. Lions, eds., North-Holland, Amsterdam, 1991, pp. 523–639.
- [52] S. SIDDARTH, J. CARRERO, B. COCKBURN, K. TAMMA, AND R. KANAPADY, *The local discontinuous Galerkin method and component design integration for 3D elasticity*, in Proceedings of the Third M.I.T. Conference on Computational Fluid and Solid Mechanics, Cambridge, MA, 2005, pp. 492–494.
- [53] C. WIENERS AND B. WOHLMUTH, *The coupling of mixed and conforming finite element discretizations*, in Domain Decomposition Methods 10, Contemp. Math. 218, J. Mandel, C. Farhat, and X.-C. Cai, eds., American Mathematical Society, Providence, RI, 1997, pp. 453–459.

NUMERICAL DISPERSIVE SCHEMES FOR THE NONLINEAR SCHRÖDINGER EQUATION*

LIVIU I. IGNAT[†] AND ENRIQUE ZUAZUA[‡]

Abstract. We consider semidiscrete approximation schemes for the linear Schrödinger equation and analyze whether the classical dispersive properties of the continuous model hold for these approximations. For the conservative finite difference semidiscretization scheme we show that, as the mesh size tends to zero, the semidiscrete approximate solutions lose the dispersion property. This fact is proved by constructing solutions concentrated at the points of the spectrum where the second order derivatives of the symbol of the discrete Laplacian vanish. Therefore this phenomenon is due to the presence of numerical spurious high frequencies. To recover the dispersive properties of the solutions at the discrete level, we introduce two numerical remedies: Fourier filtering and a two-grid preconditioner. For each of them we prove Strichartz-like estimates and a local space smoothing effect, uniform in the mesh size. The methods we employ are based on classical estimates for oscillatory integrals. These estimates allow us to treat nonlinear problems with L^2 -initial data, without additional regularity hypotheses. We prove the convergence of the two-grid method for nonlinearities that cannot be handled by energy arguments and which, even in the continuous case, require Strichartz estimates.

Key words. finite differences, nonlinear Schrödinger equations, Strichartz estimates

AMS subject classifications. 65M12, 65T50, 35Q55

DOI. 10.1137/070683787

1. Introduction. Let us consider the linear (LSE) and the nonlinear (NSE) Schrödinger equations:

$$(1.1) \quad \begin{cases} iu_t + \Delta u = 0, & x \in \mathbb{R}^d, \quad t \neq 0, \\ u(0, x) = \varphi(x), & x \in \mathbb{R}^d, \end{cases}$$

and

$$(1.2) \quad \begin{cases} iu_t + \Delta u = F(u), & x \in \mathbb{R}^d, \quad t \neq 0, \\ u(0, x) = \varphi(x), & x \in \mathbb{R}^d, \end{cases}$$

respectively.

The linear equation (1.1) is solved by $u(t, x) = S(t)\varphi(x)$, where $S(t) = e^{it\Delta}$ is the free Schrödinger operator. The linear semigroup has two important properties. First, we have the conservation of the L^2 -norm

$$(1.3) \quad \|u(t)\|_{L^2(\mathbb{R}^d)} = \|\varphi\|_{L^2(\mathbb{R}^d)}$$

*Received by the editors February 28, 2007; accepted for publication (in revised form) November 10, 2008; published electronically February 25, 2009. This work has been supported by grant MTM2008-03541 of the Spanish MEC, the DOMINO Project CIT-370200-2005-10 in the PROFIT program, and the SIMUMAT project of the CAM (Spain).

<http://www.siam.org/journals/sinum/47-2/68378.html>

[†]Institute of Mathematics “Simion Stoilow” of the Romanian Academy, P.O. Box 1-764, RO-014700 Bucharest, Romania (liviu.ignat@gmail.com). This author’s research was also supported by reintegration grant RP-3, contract 4-01/10/2007 of CNCSIS Romania.

[‡]Basque Center for Applied Mathematics, Gran Via, 35 - 2, 48009 Bilbao, Spain (zuazua@bcamath.org).

and then a dispersive estimate of the form

$$(1.4) \quad |u(t, x)| = |S(t)\varphi(x)| \leq \frac{1}{(4\pi|t|)^{d/2}} \|\varphi\|_{L^1(\mathbb{R}^d)}, \quad x \in \mathbb{R}^d, \quad t \neq 0.$$

The space-time estimate

$$(1.5) \quad \|S(\cdot)\varphi\|_{L^{2+4/d}(\mathbb{R}, L^{2+4/d}(\mathbb{R}^d))} \leq C\|\varphi\|_{L^2(\mathbb{R}^d)},$$

due to Strichartz [27], is deeper. It guarantees that the solutions decay as t becomes large and that they gain some spatial integrability.

Inequality (1.5) was generalized by Ginibre and Velo [8]. They proved the mixed space-time estimate, well known as Strichartz’s estimate:

$$(1.6) \quad \|S(\cdot)\varphi\|_{L^q(\mathbb{R}, L^r(\mathbb{R}^d))} \leq C(q, r)\|\varphi\|_{L^2(\mathbb{R}^d)}$$

for the so-called $d/2$ -admissible pairs (q, r) . We recall that the exponent pair (q, r) is α -admissible (cf. [14]) if $2 \leq q, r \leq \infty$, $(q, r, \alpha) \neq (2, \infty, 1)$, and

$$(1.7) \quad \frac{1}{q} = \alpha \left(\frac{1}{2} - \frac{1}{r} \right).$$

The Strichartz estimates play an important role in the proof of the well-posedness of the NSE. Typically they are used when the energy methods fail to provide well-posedness results.

The nonlinear problem (1.2) with nonlinearity $F(u) = |u|^p u$, $p < 4/d$ and initial data in $L^2(\mathbb{R}^d)$ was first analyzed by Tsutsumi [30]. The author proved that, in this case, the NSE is globally well posed in $L^\infty(\mathbb{R}, L^2(\mathbb{R}^d)) \cap L^q_{loc}(\mathbb{R}, L^r(\mathbb{R}^d))$, where (q, r) is a $d/2$ -admissible pair depending on the nonlinearity F .

The Schrödinger equation has another remarkable property guaranteeing the gain of one half space derivative in $L^2_{x,t}$ (cf. [5] and [15]):

$$(1.8) \quad \sup_{x_0, R} \frac{1}{R} \int_{B(x_0, R)} \int_{-\infty}^{\infty} |(-\Delta)^{1/4} e^{it\Delta} \varphi|^2 dt dx \leq C\|\varphi\|_{L^2(\mathbb{R}^d)}^2.$$

It has played a crucial role in the study of the NSE with nonlinearities involving derivatives (see [16]). In particular, it is extremely useful when deriving compactness properties.

For other properties on the Schrödinger equation we refer the reader to [3] and [28].

In this paper we analyze whether semidiscrete schemes for the LSE have dispersive properties similar to (1.4), (1.6), and (1.8), uniform with respect to the mesh sizes. The study of these dispersion properties for these approximation schemes is relevant for introducing convergent schemes in the nonlinear context. Indeed, as mentioned above, the proof of the well-posedness of the NSE requires a fine use of the dispersion properties, and, consequently, it seems unlikely that the convergence of the numerical schemes could be proved if these dispersion properties are not verified at the numerical level.

Estimates similar to (1.6) for numerical solutions will allow proving uniform (on the mesh-size parameter) bounds on discrete versions of the space $L^\infty(\mathbb{R}, L^2(\mathbb{R}^d)) \cap L^q_{loc}(\mathbb{R}, L^r(\mathbb{R}^d))$. On the other hand, estimates similar to (1.8) on discrete solutions will give sufficient conditions to guarantee their compactness and thus the convergence towards the solution of the NSE (1.2).

However, as we shall see, standard numerical approximation schemes often fail to satisfy these dispersive estimates, uniformly in the mesh-size parameter, and important work needs to be done to develop numerical schemes that do fulfill these estimates uniformly.

To better illustrate the problems we shall address, let us first consider the conservative semidiscrete numerical scheme

$$(1.9) \quad \begin{cases} i \frac{du^h}{dt} + \Delta_h u^h = 0, & t > 0, \\ u^h(0) = \varphi^h. \end{cases}$$

Here u^h stands for the infinite unknown vector $\{u_j^h\}_{j \in \mathbb{Z}^d}$, $u_j(t)$ being the approximation of the solution at the node $x_j = \mathbf{j}h$, and Δ_h the classical second order finite difference approximation of Δ :

$$(1.10) \quad (\Delta_h u^h)_j = h^{-2} \sum_{k=1}^d (u_{j+e_k}^h + u_{j-e_k}^h - 2u_j^h).$$

In the one-dimensional (1-d) case, the lack of uniform dispersive estimates for the solutions of (1.9) has been observed by the authors in [12, 13]. The symbol of the Laplacian, ξ^2 , in the numerical scheme (1.9) is replaced by $4/h^2 \sin^2(\xi h/2)$ for the discrete Laplacian (1.10). The first and second derivatives of the latter vanish at the points $\pm\pi/h$ and $\pm\pi/2h$ of the spectrum. By building wave packets concentrated at the pathological spectral points $\pm\pi/2h$, it is possible to prove the lack of any uniform estimate of the type (1.4) or (1.6). Similar negative results can be shown to hold concerning (1.8) by building wave packets concentrated at $\pm\pi/h$.

The paper is organized as follows. In section 2 we analyze the conservative approximation scheme (1.9). We extend the 1-d results mentioned above and prove that this scheme does not ensure the gain of any uniform integrability or local smoothing property of the solutions with respect to the initial data. The behavior of the Fourier symbol of the numerical scheme provides a good insight to this pathological behavior. We then propose a Fourier filtering method allowing recovery of both the integrability and the local smoothing properties of the continuous model. The lack of dispersion properties for the linear scheme makes it of little use to approximate nonlinear problems. In fact, in subsection 2.5, by an explicit construction we see that the solutions of a cubic semidiscrete Schrödinger equation do not satisfy the dispersion property of the continuous one, uniformly in the mesh-size parameter.

We then introduce a numerical scheme for which the dispersion estimates are uniform. The proposed scheme involves a two-grid algorithm to precondition the initial data. Based on this numerical scheme for the LSE we build a convergent numerical scheme for the NSE in the class of $L^2(\mathbb{R}^d)$ -initial data.

Section 3 is dedicated to the analysis of the method based on the two-grid preconditioning of the initial data. We analyze the action of the linear semigroup $\exp(it\Delta_h)$ on the subspace of $l^2(h\mathbb{Z}^d)$ consisting of the slowly oscillating sequences generated by the two-grid method. Once we obtain Strichartz-like estimates in this subspace we apply them to approximate the NSE. The nonlinear term is approximated in such a way that it belongs to the class of slowly oscillating data which permits the use of the uniform Strichartz estimates.

The results in this paper should be compared to those in [25]. In that paper the authors analyze the Schrödinger equation on the lattice \mathbb{Z}^d without analyzing the

dependence on the mesh-size parameter h . They obtain Strichartz-like estimates in a class of exponents q and r larger than in the continuous one. But none of these results is uniform when working on the scaled lattice $h\mathbb{Z}^d$ and letting $h \rightarrow 0$ as our results in section 2 show.

In the context of equations on lattices we also mention [6, 19]. In these papers the authors analyze the dynamics of infinite harmonic lattices in the limit of the lattice distance ϵ tending to zero.

The analysis in this paper can be adapted to address fully discrete schemes. In [10] necessary and sufficient conditions are given guaranteeing uniform dispersion estimates for fully discrete schemes. The work of Nixon [20] is also worth mentioning. There the 1-d KdV equation is considered and space-time estimates are proved for the implicit Euler scheme.

2. A conservative scheme. In this section we analyze the conservative scheme (1.9). This scheme satisfies the classical properties of consistency and stability which imply L^2 -convergence. We construct pathological explicit solutions for (1.9) for which neither (1.6) nor (1.8) holds uniformly with respect to the mesh-size parameter h .

In our analysis we make use of the semidiscrete Fourier transform (SDFT) (we refer the reader to [29] for the main properties of the SDFT). For any $v^h \in l^2(h\mathbb{Z}^d)$ we define its SDFT at the scale h by

$$(2.11) \quad \widehat{v}^h(\xi) = (\mathcal{F}_h v^h)(\xi) = h^d \sum_{j \in \mathbb{Z}^d} e^{-i\xi \cdot jh} v_j^h, \quad \xi \in [-\pi/h, \pi/h]^d.$$

We will use the notation $A \lesssim B$ to report the inequality $A \leq \text{constant} \times B$, where the multiplicative constant is independent of h . The statement $A \simeq B$ is equivalent to $A \lesssim B$ and $B \lesssim A$.

Taking the SDFT in (1.9) we obtain that $u^h(t) = S^h(t)\varphi^h$ which is the solution of (1.9) satisfies

$$(2.12) \quad i\widehat{u}_t^h(t, \xi) + p_h(\xi)\widehat{u}^h(t, \xi) = 0, \quad t \in \mathbb{R}, \quad \xi \in [-\pi/h, \pi/h]^d,$$

where the function $p_h : [-\pi/h, \pi/h]^d \rightarrow \mathbb{R}$ is defined by

$$(2.13) \quad p_h(\xi) = \frac{4}{h^2} \sum_{k=1}^d \sin^2 \left(\frac{\xi_k h}{2} \right).$$

Solving the ODE (2.12) we obtain that the Fourier transform of u^h is given by

$$(2.14) \quad \widehat{u}^h(t, \xi) = e^{-itp_h(\xi)} \widehat{\varphi}^h(\xi), \quad \xi \in [-\pi/h, \pi/h]^d.$$

Observe that the new symbol $p_h(\xi)$ is different from the continuous one, $|\xi|^2$. In the 1-d case (see Figure 1), the symbol $p_h(\xi)$ changes convexity at the points $\xi = \pm\pi/2h$ and has critical points also at $\xi = \pm\pi/h$, two properties that the continuous symbol does not have. Using that

$$\inf_{\xi \in [-\pi/h, \pi/h]} |p_h''(\xi)| + |p_h'''(\xi)| > 0,$$

in [13] (see also [25] for $h = 1$) it has been proved that

$$(2.15) \quad \|u^h(t)\|_{l^\infty(h\mathbb{Z})} \lesssim \|\varphi^h\|_{l^1(h\mathbb{Z})} (|t|^{-1/2} + (|t|h)^{-1/3}), \quad t \neq 0.$$

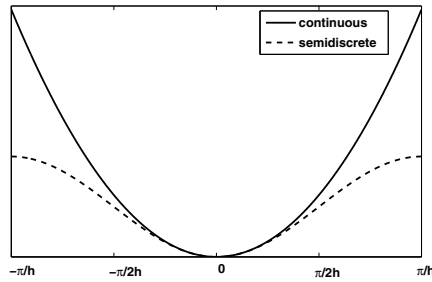


FIG. 1. The two symbols in dimension one.

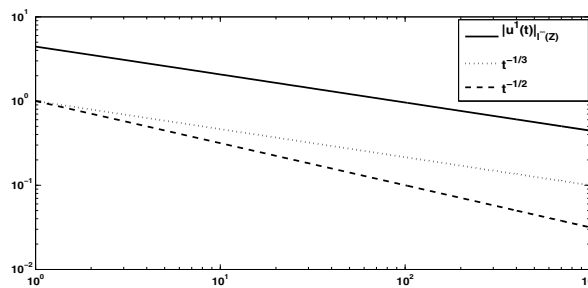


FIG. 2. Log-log plot of the time evolution of the $l^\infty(\mathbb{Z})$ -norm of the fundamental solution u^1 for (1.9).

Note that estimate (2.15) blows up as $h \rightarrow 0$. Therefore it does not yield uniform Strichartz estimates.

Figure 2 shows that (2.15) could not be improved for large time t . In fact when $h = 1$ and $\varphi^1 = \delta_0$ (δ_0 is the discrete Dirac function, where $(\delta_0)_0$ is one and zero otherwise) the solution $u^1(t)$ behaves as $t^{-1/3}$ for large time t instead of $t^{-1/2}$ in the case of the LSE.

In dimension d , similar results can be obtained in terms of the number of nonvanishing principal curvatures of the symbol and its gradient. Observe that, at the points $\xi = (\pm\pi/2h, \dots, \pm\pi/2h)$, all the eigenvalues of the Hessian matrix $H_{p_h} = (\partial_{ij}p_h)_{ij}$ vanish. Moreover, if k -components of the vector ξ coincide with $\pm\pi/2h$, the rank of H_{p_h} at this point is $d - k$ instead of d , as in the continuous case. This will imply that the solutions of (1.9), concentrated at these points of the spectrum, will behave as $t^{-(d-k)/2}(th)^{-k/3}$ instead of $t^{-d/2}$ as $t \rightarrow \infty$. This shows that there are no uniform estimates similar to (1.4) or (1.6) at the discrete level. But these inequalities are necessary to prove the uniform boundedness of the semidiscrete solutions in the nonlinear setting.

On the other hand, at the points $\xi = (\pm\pi/h, \dots, \pm\pi/h)$, the gradient of the symbol $p_h(\xi)$ vanishes. As we will see, these pathologies affect the dispersive properties of the semidiscrete scheme (1.9) and its solutions do not fulfill the regularizing property (1.8), uniformly in $h > 0$, which is needed to guarantee the compactness of the semidiscrete solutions. This constitutes an obstacle when passing to the limit as $h \rightarrow 0$ in the nonlinear semidiscrete models.

This section is organized as follows. Section 2.1 deals with the analysis of proper-

ties (1.4) and (1.6) for the solutions of (1.9). The local smoothing property is analyzed in section 2.2. In section 2.3 we prove uniform estimates similar to (1.4) and (1.8), uniformly with respect to the parameter h , in the class of initial data whose Fourier spectrum has been filtered conveniently. Strichartz-like estimates for filtered solutions are given in section 2.4.

In section 2.5 we analyze a numerical scheme for the 1-d cubic NSE based on the conservative approximation of the linear Schrödinger semigroup. We prove that its solutions do not remain uniformly bounded in any auxiliary space $L^q_{loc}(\mathbb{R}, L^r(h\mathbb{Z}))$.

2.1. Lack of uniform dispersive estimates. First, we construct explicit examples of solutions of (1.9) for which all the classical estimates of the continuous case (1.6) blow up.

THEOREM 2.1. *Let $T > 0$, $r_0 \geq 1$, and $r > r_0$. Then*

$$(2.16) \quad \sup_{h>0, \varphi^h \in l^{r_0}(h\mathbb{Z}^d)} \frac{\|S^h(T)\varphi^h\|_{l^r(h\mathbb{Z}^d)}}{\|\varphi^h\|_{l^{r_0}(h\mathbb{Z}^d)}} = \infty$$

and

$$(2.17) \quad \sup_{h>0, \varphi^h \in l^{r_0}(h\mathbb{Z}^d)} \frac{\|S^h(\cdot)\varphi^h\|_{L^1((0,T), l^r(h\mathbb{Z}^d))}}{\|\varphi^h\|_{l^{r_0}(h\mathbb{Z}^d)}} = \infty.$$

Remark 2.1. A finer analysis can be done. The same result holds if we take the supremum in (2.16) and (2.17) over the set of functions $\varphi^h \in l^{r_0}(h\mathbb{Z}^d)$ such that the support of their Fourier transform (2.11) contains at least one of the points of the set

$$(2.18) \quad \mathcal{M}_1^h = \left\{ \xi = (\xi_1, \dots, \xi_d) \in \left[-\frac{\pi}{h}, \frac{\pi}{h} \right]^d : \exists i \in \{1, \dots, d\} \text{ such that } \xi_i = \frac{\pi}{2h} \right\}.$$

Observe that at the above points the rank of the Hessian matrix H_{p^h} is at most $d - 1$.

Remark 2.2. Let \mathbf{P}^h be an interpolator, piecewise constant or linear. In view of Theorem 2.1, for any fixed $T > 0$, the uniform boundedness principle guarantees the existence of a function $\varphi \in L^2(\mathbb{R}^d)$ and a sequence φ^h such that $\mathbf{P}^h\varphi^h \rightarrow \varphi$ in $L^2(\mathbb{R}^d)$ and the corresponding solutions u^h of (1.9) satisfy $\|\mathbf{P}^h u^h\|_{L^1((0,T), L^r(\mathbb{R}^d))} \rightarrow \infty$.

Proof of Theorem 2.1. First, observe that it is sufficient to deal with the 1-d case. Indeed, for any sequence $\{\psi_j^h\}_{j \in \mathbb{Z}}$ set $\varphi_j^h = \psi_{j_1}^h \dots \psi_{j_d}^h$, where $\mathbf{j} = (j_1, j_2, \dots, j_d)$. We are thus considering discrete functions in separated variables. Then, for any t the following holds:

$$(S^h(t)\varphi^h)_{\mathbf{j}} = (S^{1,h}(t)\psi^h)_{j_1} (S^{1,h}(t)\psi^h)_{j_2} \dots (S^{1,h}(t)\psi^h)_{j_d},$$

where $S^{1,h}(t)$ is the linear semigroup generated by (1.9) in the 1-d case. Thus it is obvious that (2.16) and (2.17) hold in dimension $d \geq 2$, once we prove them in the 1-d case $d = 1$.

In the following we will consider the 1-d case $d = 1$ and prove (2.16), the other estimate (2.17) being similar. Using the properties of the SDFT it is easy to see that $(S^h(t)\varphi^h)_j = (S^1(t/h^2)\varphi^1)_j$, where $\varphi_j^1 = \varphi_j^h$, $j \in \mathbb{Z}$. A scaling argument in (2.16) shows that

$$(2.19) \quad \frac{\|S^h(T)\varphi^h\|_{l^q(h\mathbb{Z})}}{\|\varphi^h\|_{l^{q_0}(h\mathbb{Z})}} = h^{\frac{1}{q} - \frac{1}{q_0}} \frac{\|S^1(T/h^2)\varphi^1\|_{l^q(\mathbb{Z})}}{\|\varphi^1\|_{l^{q_0}(\mathbb{Z})}}.$$

Let us introduce the operator $S_1(t)$ defined by

$$(2.20) \quad (S_1(t)\varphi)(x) = \int_{-\pi}^{\pi} e^{-itp_1(\xi)} e^{ix\xi} \widehat{\varphi}(\xi) d\xi,$$

which is the extension of the semigroup generated by (1.9) for $h = 1$ to all $x \in \mathbb{R}$.

We point out that for any sequence $\{\varphi_j^1\}_{j \in \mathbb{Z}}$, $S_1(t)\varphi^1$ as in (2.20), which is defined for all $x \in \mathbb{R}$, is in fact the band-limited interpolator of the semidiscrete function $S^1(t)\varphi^1$. The results of Magyar, Stein, and Wainger [18] (see also Plancherel and Pólya [21]) on band-limited functions show that the following inequalities hold for any $q \geq 1$ and for all continuous functions $\widehat{\varphi}$ supported in $[-\pi, \pi]$:

$$c(q)\|\varphi\|_{l^q(\mathbb{Z})} \leq \|\varphi\|_{L^q(\mathbb{R})} \leq C(q)\|\varphi\|_{l^q(\mathbb{Z})}.$$

Thus for any $q > q_0 \geq 1$ the following holds for all functions φ^1 whose Fourier transform is supported in $[-\pi, \pi]$:

$$(2.21) \quad \frac{\|S^1(t)\varphi^1\|_{l^q(\mathbb{Z})}}{\|\varphi^1\|_{l^{q_0}(\mathbb{Z})}} \geq c(q, q_0) \frac{\|S_1(t)\varphi^1\|_{L^q(\mathbb{R})}}{\|\varphi^1\|_{L^{q_0}(\mathbb{R})}}.$$

In view of this property it is sufficient to deal with the operator $S_1(t)$.

Denoting $\tau = T/h^2$, by (2.19) the proof of (2.16) is reduced to the proof of the following fact about the new operator $S_1(t)$:

$$(2.22) \quad \lim_{\tau \rightarrow \infty} \tau^{\frac{1}{2}(\frac{1}{q_0} - \frac{1}{q})} \sup_{\text{supp}(\widehat{\varphi}) \subset [-\pi, \pi]} \frac{\|S_1(\tau)\varphi\|_{L^q(\mathbb{R})}}{\|\varphi\|_{L^{q_0}(\mathbb{R})}} = \infty.$$

The following lemma is the key point in the proof of the last estimate.

LEMMA 2.1. *There exists a positive constant c such that for all τ sufficiently large, there exists a function φ_τ such that $\|\varphi_\tau\|_{L^p(\mathbb{R})} \simeq \tau^{1/3p}$ for all $p \geq 1$ and*

$$(2.23) \quad |(S_1(t)\varphi_\tau)(x)| \geq \frac{1}{2}$$

for all $|t| \leq c\tau$ and $|x - tp'_1(\pi/2)| \leq c\tau^{1/3}$.

Remark 2.3. Lemma 2.1 shows a lack of dispersion in the semidiscrete setting when compared with the continuous one. In the latter, for any initial data φ_τ such that $\|\varphi_\tau\|_{L^1(\mathbb{R})} \simeq \tau^{1/3}$, the solution $S(t)\varphi_\tau$ of the LSE satisfies

$$\|S(t)\varphi_\tau\|_{L^\infty(\mathbb{R})} \lesssim \frac{\tau^{1/3}}{|t|^{1/2}} \lesssim \frac{1}{\tau^{1/6}}$$

for all $t \simeq \tau$, which is incompatible with (2.23).

The proof of Lemma 2.1 will be given later.

Assuming for the moment that Lemma 2.1 holds, we now prove (2.22). In view of Lemma 2.1, given $q > q_0 \geq 1$, for sufficiently large τ the following holds:

$$\sup_{\text{supp}(\widehat{\varphi}) \subset [-\pi, \pi]} \frac{\|S_1(\tau)\varphi\|_{L^q(\mathbb{R})}}{\|\varphi\|_{L^{q_0}(\mathbb{R})}} \gtrsim \tau^{\frac{1}{3q} - \frac{1}{3q_0}}.$$

Thus (2.22) holds and the proof is done. \square

Proof of Lemma 2.1. The techniques used below are similar to those used in [7] to get lower bounds on oscillatory integrals.

We define the relevant initial data through its Fourier transform. Let us first fix a positive function $\widehat{\varphi}$ supported on $(-1, 1)$ such that $\int_{-\pi}^{\pi} \widehat{\varphi} = 1$. For all positive τ , we set

$$\widehat{\varphi}_{\tau}(\xi) = \tau^{1/3} \widehat{\varphi}(\tau^{1/3}(\xi - \pi/2)).$$

We define φ_{τ} as the inverse Fourier transform of $\widehat{\varphi}_{\tau}$. Observe that $\widehat{\varphi}_{\tau}$ is supported in the interval $(\pi/2 - \tau^{-1/3}, \pi/2 + \tau^{-1/3})$ and $\int_{-\pi}^{\pi} \widehat{\varphi}_{\tau} = 1$. Also using that $\varphi_{\tau}(x) = \varphi_1(\tau^{-1/3}x)$ we get $\|\varphi_{\tau}\|_{L^p(\mathbb{R})} \simeq \tau^{1/3p}$ for any $p \geq 1$.

The mean value theorem applied to the integral occurring in the right-hand side of (2.20) shows that

$$(2.24) \quad |S_1(t)\varphi_{\tau}(x)| \geq \left(1 - 2\tau^{-1/3} \sup_{\xi \in \text{supp}(\widehat{\varphi}_{\tau})} |x - tp'_1(\xi)|\right) \int_{-\pi}^{\pi} \widehat{\varphi}_{\tau}(\xi) d\xi.$$

Using that the second derivative of p_1 vanishes at $\xi = \pi/2$ we obtain the existence of a positive constant c_1 such that

$$|x - tp'_1(\xi)| \leq |x - tp'_1(\pi/2)| + tc_1|\xi - \pi/2|^2, \quad \xi \simeq \pi/2.$$

In particular for all $\xi \in [\pi/2 - \tau^{-1/3}, \pi/2 + \tau^{-1/3}]$ the following holds:

$$|x - tp'_1(\xi)| \leq |x - tp'_1(\pi/2)| + tc_1\tau^{-2/3}.$$

Thus there exists a (small enough) positive constant c such that for all x and t satisfying $|x - tp'_1(\pi/2)| \leq c\tau^{1/3}$ and $t \leq c\tau$

$$2\tau^{-1/3} \sup_{\xi \in \text{supp}(\widehat{\varphi}_{\tau})} |x - tp'_1(\xi)| \leq \frac{1}{2}.$$

In view of (2.24) this yields (2.23) and finishes the proof. \square

2.2. Lack of uniform local smoothing effect. In order to analyze the local smoothing effect at the discrete level we introduce the discrete fractional derivatives on the lattice $h\mathbb{Z}^d$. We define, for any $s \geq 0$, the fractional derivative $(-\Delta_h)^{s/2}u^h$ at the scale h as

$$(2.25) \quad ((-\Delta_h)^{s/2}u^h)_j = \int_{[-\pi/h, \pi/h]^d} p_h^{s/2}(\xi) e^{i\mathbf{j}\cdot\xi h} \mathcal{F}_h(u^h)(\xi) d\xi, \quad \mathbf{j} \in \mathbb{Z}^d,$$

where $p_h(\cdot)$ is as in (2.13) and $\mathcal{F}_h(u^h)$ is the SDFFT of the sequence $\{u^h_{\mathbf{j}}\}_{\mathbf{j} \in \mathbb{Z}^d}$ at the scale h .

Concerning the local smoothing effect we have the following result.

THEOREM 2.2. *Let $T > 0$ and $s > 0$. Then*

$$(2.26) \quad \sup_{h>0, \varphi^h \in l^2(h\mathbb{Z}^d)} \frac{h^d \sum_{|\mathbf{j}|_h \leq 1} |((-\Delta_h)^{s/2}S^h(T)\varphi^h)_j|^2}{\|\varphi^h\|_{l^2(h\mathbb{Z}^d)}^2} = \infty$$

and

$$(2.27) \quad \sup_{h>0, \varphi^h \in l^2(h\mathbb{Z}^d)} \frac{h^d \sum_{|\mathbf{j}|_h \leq 1} \int_0^T |((-\Delta_h)^{s/2}S^h(t)\varphi^h)_j|^2 dt}{\|\varphi^h\|_{l^2(h\mathbb{Z}^d)}^2} = \infty.$$

Remark 2.4. The same result holds if we take the supremum in (2.26) and (2.27) over the set of functions $\varphi^h \in l^2(h\mathbb{Z}^d)$ such that the support of φ^h contains at least one of the points of the set

$$(2.28) \quad \mathcal{M}_2^h = \left\{ \xi = (\xi_1, \dots, \xi_d) \in \left[-\frac{\pi}{h}, \frac{\pi}{h} \right]^d : \xi_i = \pm \frac{\pi}{h}, i = 1, \dots, d \right\}.$$

Observe that at the above points the gradient of p_h vanishes.

In contrast with the proof of Theorem 2.1 we cannot reduce it to the 1-d case. This is due to the extra factor $p_h^{s/2}(\xi)$ which does not allow us to use separation of variables. The proof consists in reducing (2.26) and (2.27) to the case $h = 1$ and then using the following lemma.

LEMMA 2.2. *Let $s > 0$. There is a positive constant c such that for all τ sufficiently large there exists a function φ_τ^1 with $\|\varphi_\tau^1\|_{l^2(\mathbb{Z}^d)} = \tau^{d/2}$ and*

$$(2.29) \quad |((-\Delta_1)^{s/2} S^1(t)\varphi_\tau^1)_{\mathbf{j}}| \geq 1/2$$

for all $|t| \leq c\tau^2, |\mathbf{j}| \leq c\tau$.

We postpone the proof of Lemma 2.2 and proceed with the proof of Theorem 2.2.

Proof of Theorem 2.2. We prove (2.26), the other estimate (2.27) being similar. As in the previous section we reduce the proof to the case $h = 1$. By the definition of $(-\Delta_h)^{s/2}$ for any $\mathbf{j} \in \mathbb{Z}^d$ we have that

$$((-\Delta_h)^{s/2} S^h(t)\varphi^h)_{\mathbf{j}} = h^{-s}((-\Delta_1)^{s/2} S^1(t/h^2)\varphi^1)_{\mathbf{j}}, \quad \mathbf{j} \in \mathbb{Z}^d,$$

where $\varphi_{\mathbf{j}}^h = \varphi_{\mathbf{j}}^1, \mathbf{j} \in \mathbb{Z}^d$. Thus

$$\frac{h^d \sum_{|\mathbf{j}| \leq 1} |((-\Delta_h)^{s/2} S^h(T)\varphi^h)_{\mathbf{j}}|^2}{\|\varphi^h\|_{l^2(h\mathbb{Z}^d)}^2} = \frac{h^{-2s} \sum_{|\mathbf{j}| \leq 1/h} |((-\Delta_1)^{s/2} S^1(T/h^2)\varphi^1)_{\mathbf{j}}|^2}{\|\varphi^1\|_{l^2(\mathbb{Z}^d)}^2}.$$

With c and φ_τ given by Lemma 2.2 and τ such that $c\tau^2 = T/h^2$, i.e., $\tau = (T/c)^{1/2}h^{-1}$, we have $\|\varphi_\tau^1\|_{l^2(\mathbb{Z}^d)}^2 = \tau^d$ and

$$\lim_{\tau \rightarrow \infty} \frac{h^{-2s} \sum_{|\mathbf{j}| \leq 1/h} |((-\Delta_1)^{s/2} S^1(T/h^2)\varphi_\tau^1)_{\mathbf{j}}|^2}{\|\varphi_\tau^1\|_{l^2(\mathbb{Z}^d)}^2} \gtrsim \lim_{\tau \rightarrow \infty} \frac{\tau^{2s} \tau^d}{\tau^d} = \infty.$$

This finishes the proof. □

Proof of Lemma 2.2. We choose a positive function $\widehat{\varphi}$ supported in the unit ball with $\int_{\mathbb{R}^d} \widehat{\varphi} = 1$. Set for all $\tau \geq 1$ $\widehat{\varphi}_\tau^1(\xi) = \tau^d \widehat{\varphi}(\tau(\xi - \pi_d))$, where $\pi_d = (\pi, \dots, \pi)$. We define φ_τ^1 as the inverse Fourier transform at scale $h = 1$ of $\widehat{\varphi}_\tau^1$. Thus $\widehat{\varphi}_\tau^1$ is supported in $\{\xi : |\xi - \pi_d| \leq \tau^{-1}\}$, it has mass one, and $\|\varphi_\tau^1\|_{l^2(\mathbb{Z}^d)} \simeq \tau^{d/2}$. Applying the mean value theorem to the oscillatory integral occurring in the definition of $(-\Delta_1)^{s/2} S^1(t)\varphi_\tau^1$ and using that $p_1(\xi)$ behaves as a positive constant in the support of $\widehat{\varphi}_\tau^1$ we obtain that for some positive constant c_0

$$\begin{aligned} |((-\Delta_1)^{s/2} S^1(t)\varphi_\tau^1)_{\mathbf{j}}| &\geq \left(1 - 2\tau^{-1} \sup_{\xi \in \text{supp}(\widehat{\varphi}_\tau^1)} |\mathbf{j} - t\nabla p_1(\xi)| \right) \int_{[-\pi, \pi]^d} p_1^{s/2}(\xi) \widehat{\varphi}_\tau^1(\xi) d\xi \\ &\geq c_0 \left(1 - 2\tau^{-1} \sup_{\xi \in \text{supp}(\widehat{\varphi}_\tau^1)} |\mathbf{j} - t\nabla p_1(\xi)| \right) \int_{[-\pi, \pi]^d} \widehat{\varphi}_\tau^1(\xi) d\xi. \end{aligned}$$

Using that ∇p_1 vanishes at $\xi = \pi_d$ we obtain the existence of a positive constant c_1 such that

$$|\mathbf{j} - t\nabla p_1(\xi)| \leq |\mathbf{j}| + tc_1|\xi - \pi_d|, \quad \xi \sim \pi_d.$$

Then there exists a positive constant c such that for all \mathbf{j} and t satisfying $|\mathbf{j}| \leq c\tau$ and $t \leq c\tau^2$ the following holds:

$$2\tau^{-1} \sup_{\xi \in \text{supp}(\widehat{\varphi}_\tau)} |\mathbf{j} - t\nabla p_1(\xi)| \leq \frac{1}{2}.$$

Thus for all t and \mathbf{j} as above (2.29) holds. This finishes the proof. \square

2.3. Filtering of the initial data. As we have seen in the previous section the conservative scheme (1.9) does not reproduce the dispersive properties of the continuous LSE. In this section we prove that a suitable filtering of the initial data in the Fourier space provides uniform dispersive properties and a local smoothing effect. The key point to recover the decay rates (1.4) at the discrete level is to choose initial data with their SDFT supported away from the pathological points \mathcal{M}_1^h in (2.18). Similarly, the local smoothing property holds uniformly on h if the SDFT of the initial data is supported away from the points \mathcal{M}_2^h in (2.28).

For any positive $\epsilon < \pi/2$ we define Ω_ϵ^h , the set of all the points in the cube $[-\pi/h, \pi/h]^d$ whose distance is at least ϵ/h from the set in which some of the second order derivatives of $p_h(\xi)$ vanish:

$$\Omega_{\epsilon,d}^h = \left\{ \xi = (\xi_1, \dots, \xi_d) \in \left[-\frac{\pi}{h}, \frac{\pi}{h} \right]^d : \left| \xi_i \mp \frac{\pi}{2h} \right| \geq \frac{\epsilon}{h}, i = 1, \dots, d \right\}.$$

Let us define the class of functions $\mathcal{I}_{\epsilon,d}^h \subset l^2(h\mathbb{Z}^d)$, whose SDFT is supported on $\Omega_{\epsilon,d}^h$:

$$(2.30) \quad \mathcal{I}_{\epsilon,d}^h = \{ \varphi^h \in l^2(h\mathbb{Z}^d) : \text{supp}(\widehat{\varphi}^h) \subset \Omega_{\epsilon,d}^h \}.$$

We can view this subspace of initial data as a subclass of filtered data in the sense that the Fourier components corresponding to ξ such that $|\xi_i \pm \pi/2h| \leq \epsilon/h$ have been cut off or filtered out.

The following theorem shows that for initial data in this class the semigroup $S^h(t)$ has the same long time behavior as the continuous one, independently of h in what concerns the $l^{p'}(h\mathbb{Z}^d) - l^p(h\mathbb{Z}^d)$ decay property.

THEOREM 2.3. *Let $0 < \epsilon < \pi/2$ and $p \geq 2$. There exists a positive constant $C(\epsilon, p, d)$ such that*

$$(2.31) \quad \|S^h(t)\varphi^h\|_{l^p(h\mathbb{Z}^d)} \leq C(\epsilon, p, d)|t|^{-\frac{d}{2}(1-\frac{2}{p})} \|\varphi^h\|_{l^{p'}(h\mathbb{Z}^d)}, \quad t \neq 0,$$

holds for all $\varphi^h \in l^{p'}(h\mathbb{Z}^d) \cap \mathcal{I}_{\epsilon,d}^h$, uniformly on $h > 0$.

Proof. A scaling argument reduces the proof to the case $h = 1$. For any $\varphi^1 \in \mathcal{I}_{\epsilon,d}^1$ the solution of (1.9) is given by $S^1(t)\varphi^1 = K_{\epsilon,d}^1 * \varphi^1$, where

$$(2.32) \quad K_{\epsilon,d}^1(t, \mathbf{j}) = \int_{\Omega_{\epsilon,d}^1} e^{itp_1(\xi)} e^{i\mathbf{j}\cdot\xi} d\xi, \quad \mathbf{j} \in \mathbb{Z}^d.$$

As a consequence of Young's inequality it remains to prove that

$$(2.33) \quad \|K_{\epsilon,d}^1(t)\|_{l^p(\mathbb{Z}^d)} \leq C(\epsilon, p, d)|t|^{-d/2(1-1/p)}$$

for any $p \geq 2$ and for all $t \neq 0$. Observe that it is then sufficient to prove (2.33) in the 1-d case. Using that the second derivative of the function $\sin^2(\xi/2)$ is positive on $\Omega_{\epsilon,1}^1$ we obtain by the Van der Corput lemma (see [26, Prop. 2, Chap. 8, p. 332]) that $\|K_{\epsilon,1}^1(t)\|_{l^\infty(\mathbb{Z})} \leq c(\epsilon)|t|^{-1/2}$ which finishes the proof. \square

A similar result can be stated for the local smoothing effect. For a positive ϵ , let us define the set $\tilde{\Omega}_{\epsilon,d}^h$ of all points located at a distance of at least ϵ/h from the points $(\pm\pi/h)^d$:

$$\tilde{\Omega}_{\epsilon,d}^h = \left\{ \xi \in \left[-\frac{\pi}{h}, \frac{\pi}{h} \right]^d : \left| \xi_i \mp \frac{\pi}{h} \right| \geq \frac{\epsilon}{h}, i = 1, \dots, d \right\}.$$

Observe that on $\tilde{\Omega}_{\epsilon,d}^h$ the symbol $p_h(\xi)$ has no critical points other than $\xi = 0$. A similar argument as in [15] shows that the linear semigroup $S^h(t)$ gains one half space derivative in $L_{t,x}^2$ with respect to the initial datum filtered as above. More precisely, if \mathbf{P}_*^h denotes the band-limited interpolator (cf. [31, Chap. II])

$$(2.34) \quad (\mathbf{P}_*^h u^h)(x) = \int_{[-\pi/h, \pi/h]^d} \widehat{u}^h(\xi) e^{ix \cdot \xi} d\xi, \quad x \in \mathbb{R}^d,$$

the following holds.

THEOREM 2.4. *Let $\epsilon > 0$. There exists a positive constant $C(\epsilon, d)$ such that for any $R > 0$*

$$\int_{|x|>R} \int_{-\infty}^{\infty} |(-\Delta)^{1/4} \mathbf{P}_*^h e^{it\Delta_h} \varphi^h|^2 dt dx \leq C(\epsilon, d) R \|\varphi^h\|_{l^2(h\mathbb{Z}^d)}^2$$

holds for all $\varphi^h \in l^2(h\mathbb{Z}^d)$ with $\text{supp}(\widehat{\varphi}^h) \subset \tilde{\Omega}_{\epsilon,d}^h$, uniformly on $h > 0$.

To prove this result we make use of the following theorem.

THEOREM 2.5 (see [15, Theorem 4.1]). *Let \mathcal{O} be an open set in \mathbb{R}^d and ψ be a $C^1(\mathcal{O})$ function such that $\nabla\psi(\xi) \neq 0$ for any $\xi \in \mathcal{O}$. Assume that there is $N \in \mathbb{N}$ such that for any $(\xi_1, \dots, \xi_{d-1}) \in \mathbb{R}^{d-1}$ and $r \in \mathbb{R}$ the equations*

$$\psi(\xi_1, \dots, \xi_k, \underline{\xi}, \xi_{k+1}, \dots, \xi_{d-1}) = r, \quad k = 0, \dots, d-1,$$

have at most N solutions $\underline{\xi} \in \mathbb{R}$. For $a \in L^\infty(\mathbb{R}^d \times \mathbb{R})$ and $f \in \mathcal{S}(\mathbb{R}^d)$ define

$$W(t)f(x) = \int_{\mathcal{O}} e^{i(t\psi(\xi)+x \cdot \xi)} a(x, \psi(\xi)) \widehat{f}(\xi) d\xi.$$

Then for any $R > 0$

$$(2.35) \quad \int_{|x| \leq R} \int_{-\infty}^{\infty} |W(t)f(x)|^2 dt dx \leq cRN \int_{\mathcal{O}} \frac{|\widehat{f}(\xi)|^2}{|\nabla\psi(\xi)|} d\xi,$$

where c is independent of R and N and f .

Remark 2.5. The result remains true for domains \mathcal{O} where $|\nabla\psi|$ has zeros, provided that the right-hand side of (2.35) is finite.

Proof of Theorem 2.4. Observe that for any $\varphi^h \in l^2(h\mathbb{Z}^d)$ with $\text{supp}(\widehat{\varphi}^h) \subset \tilde{\Omega}_{\epsilon,d}^h$ we have

$$(\mathbf{P}_*^h e^{it\Delta_h} \varphi^h)(x) = \int_{\tilde{\Omega}_{\epsilon,d}^h} e^{itp_h(\xi)} e^{ix \cdot \xi} \widehat{\varphi}^h(\xi) d\xi, \quad x \in \mathbb{R}^d.$$

Applying Theorem 2.5 with $\mathcal{O} = \tilde{\Omega}_{\epsilon,d}^h$, $\psi = p_h(\xi)$, and $a \equiv 1$ and using that $|\nabla p_h(\xi)| \geq c(\epsilon, d)|\xi|$ for all $\xi \in \tilde{\Omega}_{\epsilon,d}^h$ we obtain that

$$\int_{|x|<R} \int_{-\infty}^{\infty} |(-\Delta)^{1/4} \mathbf{P}_*^h e^{it\Delta_h} \varphi^h|^2 dt dx \lesssim \int_{\tilde{\Omega}_{\epsilon,d}^h} \frac{|\widehat{\varphi}^h(\xi)|^2 |\xi|}{|\nabla p_h(\xi)|} d\xi \lesssim \|\varphi^h\|_{l^2(h\mathbb{Z}^d)}^2.$$

This finishes the proof. \square

2.4. Strichartz estimates for filtered data. In this section we are interested in deriving Strichartz-like estimates for the operator $S^h(t)$ when it acts on functions belonging to $\mathcal{I}_{\epsilon,d}^h$, the class of functions defined in (2.30).

The main ingredient in obtaining Strichartz estimates is the following result due to Keel and Tao [14].

THEOREM 2.6 (see [14, Theorem 1.2]). *Let H be a Hilbert space, (X, dx) be a measure space, and $U(t) : H \rightarrow L^2(X)$ be a one parameter family of mappings, which obey the energy estimate*

$$(2.36) \quad \|U(t)f\|_{L^2(X)} \leq C\|f\|_H$$

and the decay estimate

$$(2.37) \quad \|U(t)U(s)^*g\|_{L^\infty(X)} \leq C|t-s|^{-\sigma}\|g\|_{L^1(X)}$$

for some $\sigma > 0$. Then

$$\|U(t)f\|_{L^q(\mathbb{R}, L^r(X))} \leq C\|f\|_H \quad \forall f \in H,$$

(2.38)

$$\left\| \int_{\mathbb{R}} U(s)^*F(s, \cdot) ds \right\|_H \leq C\|F\|_{L^{q'}(\mathbb{R}, L^{r'}(X))} \quad \forall F \in L^{q'}(\mathbb{R}, L^{r'}(X)),$$

(2.39)

$$\left\| \int_0^t U(t)U(s)^*F(s, \cdot) ds \right\|_{L^q(\mathbb{R}, L^r(X))} \leq C\|F\|_{L^{\tilde{q}'}(\mathbb{R}, L^{\tilde{r}'}(X))} \quad \forall F \in L^{\tilde{q}'}(\mathbb{R}, L^{\tilde{r}'}(X))$$

for any σ -admissible pairs (q, r) and (\tilde{q}, \tilde{r}) .

Remark 2.6. With the same arguments as in [14], the following also holds for all (q, r) and (\tilde{q}, \tilde{r}) , σ -admissible pairs:

$$(2.40) \quad \left\| \int_0^t U(t-s)F(s, \cdot) ds \right\|_{L^q(\mathbb{R}, L^r(X))} \leq C\|F\|_{L^{\tilde{q}'}(\mathbb{R}, L^{\tilde{r}'}(X))}.$$

In the case of the Schrödinger semigroup, $S(t-s) = S(t)S(s)^*$, so (2.40) and (2.39) coincide. However, in our applications we will often deal with operators that do not satisfy $S(t-s) = S(t)S(s)^*$.

Let us choose $0 < \epsilon < \pi/2$, $K_d^{1,\epsilon}$ as in (2.32) and $U(t)\varphi^1 = K_d^{1,\epsilon} * \varphi^1$. We apply the above theorem to $U(t)$, with $X = \mathbb{Z}^d$, dx being the counting measure, and $H = l^2(\mathbb{Z}^d)$. In this way we obtain Strichartz estimates for the semigroup $S^1(t)$ when acting on $\mathcal{I}_{\epsilon,d}^1$, i.e., when $h = 1$. Then, by scaling, we obtain the following result in the class of filtered initial data.

THEOREM 2.7. *Let $0 < \epsilon < \pi/2$ and $(q, r), (\tilde{q}, \tilde{r})$ be two $d/2$ -admissible pairs.*

(i) *There exists a positive constant $C(d, r, \epsilon)$ such that*

$$(2.41) \quad \|S^h(\cdot)\varphi^h\|_{L^q(\mathbb{R}, l^r(h\mathbb{Z}^d))} \leq C(d, r, \epsilon)\|\varphi^h\|_{l^2(h\mathbb{Z}^d)}$$

holds for all functions $\varphi^h \in \mathcal{I}_{\epsilon, d}^h$ and for all $h > 0$.

(ii) *There exists a positive constant $C(d, r, \tilde{r}, \epsilon)$ such that*

$$(2.42) \quad \left\| \int_0^t S^h(t-s)f^h(s)ds \right\|_{L^q(\mathbb{R}, l^r(h\mathbb{Z}^d))} \leq C(d, r, \tilde{r}, \epsilon)\|f^h\|_{L^{\tilde{q}}(\mathbb{R}, l^{\tilde{r}}(h\mathbb{Z}^d))}$$

holds for all functions $f^h \in L^{\tilde{q}}(\mathbb{R}, l^{\tilde{r}}(h\mathbb{Z}^d))$ with $f(t) \in \mathcal{I}_{\epsilon, d}^h$ for a.e. $t \in \mathbb{R}$ and for all $h > 0$.

2.5. On the cubic NSE. In the previous sections we have seen that the linear semidiscrete scheme (1.9) does not satisfy uniform (with respect to h) dispersive estimates. Accordingly we cannot use it to get numerical approximations for the NSE with uniform bounds on spaces of the form $L^q((0, T), l^r(h\mathbb{Z}^d))$. However, one could agree that, even if a perturbation argument based on the variation of constants formula and the dispersive properties of the linear scheme does not provide uniform bounds for the nonlinear problem, these estimates could still be true.

In this section we give an explicit example showing that a numerical scheme for the cubic NSE based on the conservative scheme (1.9) does not satisfy uniform bounds in $L^q((0, T), l^r(h\mathbb{Z}^d))$. This shows that the conservative scheme (1.9) can be used neither for the LSE nor for the NSE within the $L^q((0, T), l^r(h\mathbb{Z}^d))$ -setting.

We consider an approximation scheme to the 1-d NSE with nonlinearity $2|u|^2u$:

$$(2.43) \quad i\partial_t u_n^h + (\Delta_h u^h)_n = |u_n^h|^2(u_{n+1}^h + u_{n-1}^h).$$

In what follows we shall refer to it as the Ablowitz–Ladik approximation [1] for the NSE.

As we shall see, this scheme possesses explicit solutions which blow up in any $L^q_{loc}(\mathbb{R}, l^r(h\mathbb{Z}))$ -norm with $r > 2$ and $q \geq 1$. We point out that this is compatible with the L^2 -convergence of the numerical scheme (2.43) for smooth initial data [1, 2].

Let us consider $\varphi \in L^2(\mathbb{R})$ as initial data for (1.2) with $F(u) = 2u|u|^2$. As initial condition for (2.43) we take $u^h(0) = \varphi^h$, φ^h being an approximation of φ . Let us assume the existence of a positive T such that for any $h > 0$, there exists $u^h \in L^\infty([0, T], l^2(h\mathbb{Z}))$ a solution of (2.43). The uniform boundedness of $\{u^h\}_{h>0}$ in $L^\infty([0, T], l^2(h\mathbb{Z}))$ does not suffice to prove its convergence to the solution of (1.2). One needs to analyze whether the solutions of (2.43) are uniformly bounded, with respect to h , in one of the auxiliary spaces $L^q_{loc}(\mathbb{R}, l^r(h\mathbb{Z}))$, a property that will guarantee that any possible limit point of $\{u^h\}_{h>0}$ belongs to $L^q((0, T), L^r(\mathbb{R}))$. We are going to show that these uniform estimates do not hold in general.

To do that we look for explicit travelling wave solutions of (2.43). By scaling, the problem can be reduced to the case $h = 1$. Indeed, u^h is a solution of (2.43) if the scaled function

$$u_n^1(t) = hu_n^h(th^2), \quad n \in \mathbb{Z}, \quad t \geq 0,$$

solves (2.43) for $h = 1$. In this case, $h = 1$, there are explicit solutions of (2.43) of the form

$$(2.44) \quad u_n^1(t) = A \exp(i(an - bt)) \operatorname{sech}(cn - dt)$$

for suitable constants A, a, b, c, d (for the explicit values we refer the reader to [2, p. 84]).

In view of the structure of u^1 it is easy to see that the solutions of (2.43), obtained from u^1 by scaling, are not uniformly bounded as $h \rightarrow 0$ in any auxiliary space $L^q((0, T), l^r(h\mathbb{Z}))$ with $r > 2$. Indeed, a scaling argument shows that

$$\frac{\|u^h\|_{L^q((0,T),l^r(h\mathbb{Z}))}}{\|u^h(0)\|_{l^2(h\mathbb{Z})}} = h^{\frac{1}{r} + \frac{2}{q} - \frac{1}{2}} \frac{\|u^1\|_{L^q((0,T/h^2),l^r(\mathbb{Z}))}}{\|u^1(0)\|_{l^2(\mathbb{Z})}}.$$

Observe that, for any $t > 0$, the $l^r(\mathbb{Z})$ -norm behaves as a constant:

$$\|u^1(t)\|_{l^r(\mathbb{Z})} \simeq \left(\int_{\mathbb{R}} \operatorname{sech}^r(cx - dt) dx \right)^{1/r} = \left(\int_{\mathbb{R}} \operatorname{sech}^r(cx) dx \right)^{1/r}.$$

Thus, for all $T > 0$ and $h > 0$ the solution u^1 satisfies

$$\|u^1\|_{L^q((0,T/h^2),l^r(\mathbb{Z}))} \simeq (Th^{-2})^{1/q}.$$

Consequently for any $r > 2$ the solution u^h on the lattice $h\mathbb{Z}$ satisfies

$$\frac{\|u^h\|_{L^q((0,T),l^r(h\mathbb{Z}))}}{\|u^h(0)\|_{l^2(h\mathbb{Z})}} \simeq h^{\frac{1}{r} - \frac{1}{2}} \rightarrow \infty, \quad h \rightarrow 0.$$

This example shows that, in order to deal with the nonlinear problem, the linear approximation scheme needs to be modified. In the following section we present a method that preserves the dispersion properties and that can be used successfully at the nonlinear level.

3. A two-grid algorithm. In this section we present a conservative scheme that preserves the dispersive properties we discuss in the previous sections. In fact, the scheme we shall consider is the standard one (1.9). But, this time, in order to avoid the lack of dispersive properties associated with the high frequency components, the scheme (1.9) will be restricted to the class of filtered data obtained by a two-grid algorithm. The advantage of this filtering method with respect to the Fourier one is that the filtering can be realized in the physical space.

The method, inspired by [9], that extends to several space variables the one introduced in [11], is roughly as follows. We consider two meshes: the coarse one of size $4h$, $4h\mathbb{Z}^d$, and the finer one, the computational one $h\mathbb{Z}^d$, of size $h > 0$. The method relies basically on solving the finite difference semidiscretization (1.9) on the fine mesh $h\mathbb{Z}^d$, but only for slowly oscillating data, interpolated from the coarse grid $4h\mathbb{Z}^d$. As we shall see, the 1/4 ratio between the two meshes is important to guarantee the convergence of the method. This particular structure of the data cancels the two pathologies of the discrete symbol mentioned in section 2. Indeed, a careful Fourier analysis of those initial data shows that their discrete Fourier transform vanishes quadratically in each variable at the points $\xi = (\pm\pi/2h)^d$ and $\xi = (\pm\pi/h)^d$. As we shall see, this suffices to recover at the discrete level the dispersive properties of the continuous model.

Once the discrete version of the dispersive properties has been proved, we explain how this method can be applied to a semidiscretization of the NSE with nonlinearity $f(u) = |u|^p u$. To do this, the nonlinearity has to be approximated in such a way that the approximate discrete nonlinearities belong to the subspace of filtered data as well.

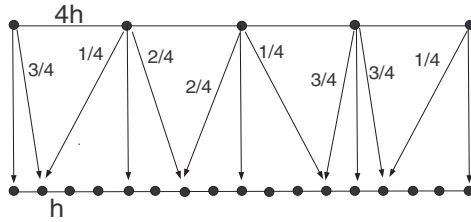


FIG. 3. The action of the operator $\tilde{\Pi}$ between the grids $4h\mathbb{Z}$.

3.1. The two-grid algorithm in the linear framework. To be more precise we introduce the following space of the slowly oscillating sequences. These sequences on the fine one $h\mathbb{Z}^d$ are those which are obtained from the coarse grid $4h\mathbb{Z}^d$ by an interpolation process. Note that, by scaling, any function defined on the lattice $h\mathbb{Z}^d$ can be viewed as a function on the lattice \mathbb{Z}^d . Thus it suffices to define this space for $h = 1$.

Let us consider the piecewise and continuous interpolator \mathbf{P}_1^1 acting on the coarse grid $4\mathbb{Z}^d$. We define the extension operator $\tilde{\Pi} : l^2(4\mathbb{Z}^d) \rightarrow l^2(\mathbb{Z}^d)$ (see Figure 3) by

$$(3.45) \quad (\tilde{\Pi}f)_j = (\mathbf{P}_1^1 f)_j, \quad \mathbf{j} \in \mathbb{Z}^d, \quad f : 4\mathbb{Z}^d \rightarrow \mathbb{C}.$$

We then define the space of the slowly oscillating sequences, $\tilde{\Pi}(4h\mathbb{Z}^d)$, as the image of the operator $\tilde{\Pi}$ acting on functions defined on $4h\mathbb{Z}^d$. We will also make use of $\tilde{\Pi}^* : l^2(h\mathbb{Z}^d) \rightarrow l^2(4h\mathbb{Z}^d)$, the adjoint of $\tilde{\Pi}$, defined by

$$(3.46) \quad (\tilde{\Pi}g_1^{4h}, g_2^h)_{l^2(h\mathbb{Z}^d)} = (g_1^{4h}, \tilde{\Pi}^*g_2^h)_{l^2(4h\mathbb{Z}^d)} \quad \forall g_1^{4h} \in l^2(4h\mathbb{Z}^d), \quad g_2^h \in l^2(h\mathbb{Z}^d),$$

where $(\cdot, \cdot)_{l^2(h\mathbb{Z}^d)}$ and $(\cdot, \cdot)_{l^2(4h\mathbb{Z}^d)}$ are the inner products on $l^2(h\mathbb{Z}^d)$ and $l^2(4h\mathbb{Z}^d)$, respectively.

In the 1-d case, the explicit expressions of $\tilde{\Pi}$ and $\tilde{\Pi}^*$ are given by

$$(\tilde{\Pi}g^{4h})_{4j+r} = \frac{4-r}{4}g_{4j}^{4h} + \frac{r}{4}g_{4j+4}^{4h}, \quad j \in \mathbb{Z}, \quad r \in \{0, 1, 2, 3\},$$

and

$$(\tilde{\Pi}^*g^h)_{4j} = \sum_{r=0}^3 \frac{4-r}{4}g_{4j+r}^h + \frac{r}{4}g_{4j-4+r}^h, \quad j \in \mathbb{Z}.$$

As we will see, $S^h(t)$ has appropriate decay properties when it acts on the subspace $\tilde{\Pi}(4h\mathbb{Z}^d)$, uniformly on $h > 0$. The main results concerning the gain of integrability are given in the following theorem.

THEOREM 3.1. *Let $p \geq 2$ and $(q, r), (\tilde{q}, \tilde{r})$ be two $d/2$ -admissible pairs. The following hold:*

(i) *There exists a positive constant $C(d, p)$ such that*

$$(3.47) \quad \|S^h(t)\tilde{\Pi}\varphi^{4h}\|_{l^p(h\mathbb{Z}^d)} \leq C(d, p)|t|^{-d(\frac{1}{2}-\frac{1}{p})}\|\tilde{\Pi}\varphi^{4h}\|_{l^{p'}(h\mathbb{Z}^d)}$$

for all $\varphi^{4h} \in l^{p'}(4h\mathbb{Z}^d)$, $h > 0$, and $t \neq 0$.

(ii) *There exists a positive constant $C(d, r)$ such that*

$$(3.48) \quad \|S^h(t)\tilde{\Pi}\varphi^{4h}\|_{L^q(\mathbb{R}, l^r(h\mathbb{Z}^d))} \leq C(d, r)\|\tilde{\Pi}\varphi^{4h}\|_{l^2(h\mathbb{Z}^d)}$$

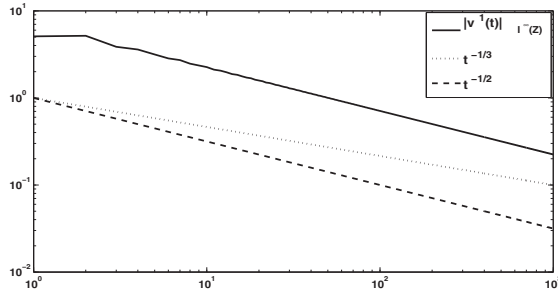


FIG. 4. Log-log plot of the time evolution of the $l^\infty(\mathbb{Z})$ -norm of $S^1(t)\tilde{\Pi}\delta_0$, where δ_0 is one in zero and vanishes otherwise.

for all $\varphi^{4h} \in l^2(4h\mathbb{Z}^d)$ and $h > 0$.

(iii) There exists a positive constant $C(d, r)$ such that

$$(3.49) \quad \left\| \int_{-\infty}^{\infty} S^h(t) * \tilde{\Pi} f^{4h}(s) ds \right\|_{l^2(h\mathbb{Z}^d)} \leq C(d, r) \|\tilde{\Pi} f^{4h}\|_{L^{q'}(\mathbb{R}, l^{r'}(h\mathbb{Z}^d))}$$

for all $f^{4h} \in L^{q'}(\mathbb{R}, l^{r'}(4h\mathbb{Z}^d))$ and $h > 0$.

(iv) There exists a positive constant $C(d, r, \tilde{r})$ such that

$$(3.50) \quad \left\| \int_0^t S^h(t-s) \tilde{\Pi} f^{4h}(s) ds \right\|_{L^q(\mathbb{R}, l^r(h\mathbb{Z}^d))} \leq C(d, r, \tilde{r}) \|\tilde{\Pi} f^{4h}\|_{L^{q'}(\mathbb{R}, l^{\tilde{r}'}(h\mathbb{Z}^d))}$$

for all $f^{4h} \in L^{q'}(\mathbb{R}, l^{\tilde{r}'}(4h\mathbb{Z}^d))$ and $h > 0$.

Remark 3.1. In the particular case $p = \infty$, estimate (3.47) shows that the solution of (1.9) with initial data in $\tilde{\Pi}(4h\mathbb{Z}^d)$ decays as $t^{-d/2}$ when t becomes large which agrees with the LSE. This can be seen in Figure 4, where the initial data has been chosen as $\tilde{\Pi}\delta_0$ (δ_0 being the discrete Dirac function defined on the coarse grid $4h\mathbb{Z}$). The solution behaves as $t^{-1/2}$ in contrast with the case presented in section 2, Figure 2, where the initial data was δ_0 (the discrete Dirac function defined on the fine grid $h\mathbb{Z}$) and the decay was as $t^{-1/3}$.

The following lemma gives a Fourier characterization of the data that are obtained by this two-grid algorithm involving the meshes $4h\mathbb{Z}^d$ and $h\mathbb{Z}^d$. Its proof uses only the definition of the discrete Fourier transform and we omit it.

LEMMA 3.1. Let $\psi^{4h} \in l^2(4h\mathbb{Z}^d)$. Then for all $\xi \in [-\pi/h, \pi/h]^d$

$$(3.51) \quad \widehat{\tilde{\Pi}\psi^{4h}}(\xi) = 4^d \widehat{\Pi\psi^{4h}}(\xi) \prod_{k=1}^d \cos^2(\xi_k h) \cos^2\left(\frac{\xi_k h}{2}\right),$$

where $(\Pi\psi^{4h})_{\mathbf{j}} = \psi_{\mathbf{j}}^{4h}$ if $\mathbf{j} \in 4\mathbb{Z}^d$ and vanishes elsewhere.

Remark 3.2. Observe that the right-hand side product in (3.51) vanishes (see the right of Figure 5 for the 1-d case) on the sets \mathcal{M}_1^h and \mathcal{M}_2^h defined in sections 2.1 and 2.2, respectively. This will allow us to recover the dispersive properties of the numerical scheme introduced in this section.

Remark 3.3. A simpler two-grid construction could be done by interpolating $2h\mathbb{Z}^d$ sequences. We would get for all $\psi^{2h} \in l^2(2h\mathbb{Z}^d)$ and $\xi \in [-\pi/h, \pi/h]^d$

$$\widehat{\tilde{\Pi}\psi^{2h}}(\xi) = 2^d \widehat{\Pi\psi^{2h}}(\xi) \prod_{k=1}^d \cos^2\left(\frac{\xi_k h}{2}\right),$$

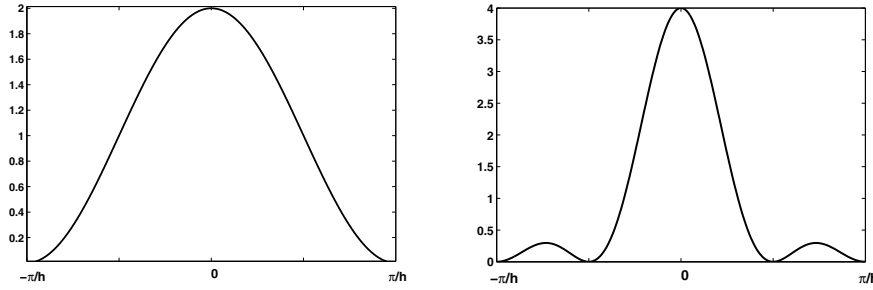


FIG. 5. Multiplicative factors introduced by the two-grid algorithm in dimension one in the case of mesh ratio 1/2 and 1/4.

where $(\Pi\psi^{2h})_{\mathbf{j}} = \psi_{\mathbf{j}}^{2h}$ if $\mathbf{j} \in 2\mathbb{Z}^d$ and vanishes elsewhere. In the 1-d case the multiplier introduced by this method is plotted in the left of Figure 5. This procedure would cancel the spurious numerical solutions at the frequencies \mathcal{M}_2^h but not at \mathcal{M}_1^h . In this case, as we proved in section 2, the Strichartz estimates would fail to be uniform on h . Thus we rather choose 1/4 as the ratio between the grids for the two-grid algorithm. We also point out that 4 is the smallest quotient of the grids for which the decay $l^1(h\mathbb{Z}^d) - l^\infty(h\mathbb{Z}^d)$ holds uniformly in the mesh parameter.

Proof of Theorem 3.1. Let us define the weighted operators $A_\beta^h(t) : l^2(h\mathbb{Z}^d) \rightarrow l^2(h\mathbb{Z}^d)$ by

$$(3.52) \quad (\widehat{A_\beta^h(t)\psi^h})(\xi) = e^{-itp_h(\xi)} |g(\xi h)|^\beta \widehat{\psi^h}(\xi), \quad \xi \in [-\pi/h, \pi/h],$$

where

$$g(\xi) = \prod_{k=1}^d \cos(\xi_k) \cos\left(\frac{\xi_k}{2}\right).$$

We will prove that for any $\beta \geq 1/4$, $A_\beta^h(t)$ satisfies the hypotheses of Theorem 2.6. Then, according to Lemma 3.1, observing that $S^h(t)\widetilde{\Pi}\varphi^{4h} = 4^d A_2^h(t)\Pi\varphi^{4h}$, we obtain (3.48), (3.49), and (3.50).

It is easy to see that $\|A_\beta^h(t)\psi^h\|_{l^2(h\mathbb{Z}^d)} \leq \|\psi^h\|_{l^2(h\mathbb{Z}^d)}$. According to this, it remains to prove that for any $\beta \geq 1/4$ and $t \neq s$ the following holds:

$$(3.53) \quad \|A_\beta^h(t)A_\beta^h(s)^* \psi^h\|_{l^\infty(h\mathbb{Z}^d)} \leq c(\beta, d) |t - s|^{-d/2} \|\psi^h\|_{l^1(h\mathbb{Z}^d)}.$$

A scaling argument reduces the proof to the case $h = 1$. We claim that (3.53) holds once

$$(3.54) \quad \|A_\gamma^1(t)\psi^1\|_{l^\infty(\mathbb{Z}^d)} \leq c(\gamma, d) |t|^{-d/2} \|\psi^1\|_{l^1(\mathbb{Z}^d)}$$

is satisfied for all $\gamma \geq 1/2$. Indeed, using that the operator $A_\alpha^1(t)$ satisfies $A_\alpha^1(t)^* = A_\alpha^1(-t)$ we obtain

$$\begin{aligned} \|A_\beta^1(t)A_\beta^1(s)^* \psi^1\|_{l^\infty(\mathbb{Z}^d)} &= \|A_\beta^1(t)A_\beta^1(-s)\psi^1\|_{l^\infty(\mathbb{Z}^d)} = \|A_{2\beta}^1(t-s)\psi^1\|_{l^\infty(\mathbb{Z}^d)} \\ &\lesssim |t-s|^{-d/2} \|\psi^1\|_{l^1(\mathbb{Z}^d)} \end{aligned}$$

for all $t \neq s$ and $\psi^1 \in l^1(\mathbb{Z}^d)$.

In the following we prove (3.54). We write $A_\gamma^1(t)$ as a convolution $A_\gamma^1(t)\psi^1 = K_{d,\gamma}^t * \psi^1$, where $\widehat{K_{d,\gamma}^t}(\xi) = e^{-itp_1(\xi)}|g(\xi)|^\gamma$. By Young's inequality it is sufficient to prove that for any $\gamma \geq 1/2$ and $t \neq 0$ the following holds:

$$(3.55) \quad \|K_{d,\gamma}^t\|_{l^\infty(\mathbb{Z}^d)} \leq c(\gamma, d)|t|^{-d/2}.$$

We observe that $K_{d,\gamma}^t$ can be written by separation of variables as

$$\widehat{K_{d,\gamma}^t}(\xi) = \prod_{k=1}^d e^{-4it \sin^2(\frac{\xi_k}{2})} \left| \cos(\xi_k) \cos\left(\frac{\xi_k}{2}\right) \right|^\gamma \prod_{j=1}^d \widehat{K_{1,\gamma}^t}(\xi_j).$$

It remains to prove that (3.55) holds in one space dimension. We make use of the following lemma.

LEMMA 3.2 (see [15, Corollary 2.9]). *Let $(a, b) \subset \mathbb{R}$ and $\psi \in C^3(a, b)$ be such that ψ'' changes monotonicity at finitely many points in the interval (a, b) . Then*

$$\left| \int_a^b e^{i(t\psi(\xi)-x\xi)} |\psi''(\xi)|^{1/2} \phi(\xi) d\xi \right| \leq c_\psi |t|^{-1/2} \left\{ \|\phi\|_{L^\infty(a,b)} + \int_a^b |\phi'(\xi)| d\xi \right\}$$

holds for all real numbers x and t .

Applying the above lemma with $\phi(\xi) = |\cos \xi|^{\gamma-1/2} |\cos(\xi/2)|^\gamma$, $\gamma \geq 1/2$, and $\psi(\xi) = -4 \sin^2(\xi/2)$, we obtain (3.55) for $d = 1$, which finishes the proof. \square

3.2. A conservative approximation of the NSE. We now build a convergent numerical scheme for the semilinear NSE equation in \mathbb{R}^d :

$$(3.56) \quad \begin{cases} iu_t + \Delta u = |u|^p u, & t \neq 0, \\ u(0, x) = \varphi(x), & x \in \mathbb{R}^d. \end{cases}$$

Our analysis applies for the nonlinearity $f(u) = -|u|^p u$ as well. In fact, the key point for the proof of the global existence of the solutions is that the L^2 -scalar product $(f(u), u)$ is a real number. All the results extend to more general nonlinearities $f(u)$ satisfying this condition under natural growth assumptions for L^2 -solutions (see [3, Chap. 4.6, p. 109]).

The first existence and uniqueness result for (3.56) with $L^2(\mathbb{R}^d)$ -initial data is as follows.

THEOREM 3.2 (global existence in $L^2(\mathbb{R}^d)$; see Tsutsumi [30]). *For $0 \leq p < 4/d$ and $\varphi \in L^2(\mathbb{R}^d)$, there exists a unique solution u in $C(\mathbb{R}, L^2(\mathbb{R}^d)) \cap L_{loc}^q(\mathbb{R}, L^{p+2}(\mathbb{R}^d))$ with $q = 4(p+1)/pd$ that satisfies the L^2 -norm conservation property and depends continuously on the initial condition in $L^2(\mathbb{R}^d)$.*

The proof uses standard arguments, the key ingredient being to work in the space $C(\mathbb{R}, L^2(\mathbb{R}^d)) \cap L_{loc}^q(\mathbb{R}, L^{p+2}(\mathbb{R}^d))$. This can only be done using Strichartz estimates. Local existence is proved by applying a fixed point argument to the integral formulation of (3.56) in that space. Global existence holds because of the $L^2(\mathbb{R}^d)$ -conservation property which excludes finite-time blow-up.

In order to introduce a numerical approximation of (3.56) it is convenient to give the definition of the weak solution of (3.56).

DEFINITION 3.1. *We say that u is a weak solution of (3.56) if the following hold:*
 (i) $u \in C(\mathbb{R}, L^2(\mathbb{R}^d)) \cap L_{loc}^q(\mathbb{R}, L^{p+2}(\mathbb{R}^d))$.

(ii) $u(0) = \varphi$ a.e. and

$$(3.57) \quad \int_{\mathbb{R}} \int_{\mathbb{R}^d} u(-i\psi_t + \Delta\psi) dx dt = \int_{\mathbb{R}} \int_{\mathbb{R}^d} |u|^p u \psi dx dt$$

for all $\psi \in \mathcal{D}(\mathbb{R}, H^2(\mathbb{R}^d))$, where p and q are as in the statement of Theorem 3.2.

In this section we consider the following numerical approximation scheme for (3.56):

$$(3.58) \quad i \frac{du^h}{dt} + \Delta_h u^h = \tilde{\Pi} f(\tilde{\Pi}^* u^h), \quad t \in \mathbb{R}; \quad u^h(0) = \tilde{\Pi} \varphi^{4h},$$

with $f(u) = |u|^p u$.

In order to prove the global existence of solutions of (3.58), we will need to guarantee the conservation of the $l^2(h\mathbb{Z}^d)$ -norm of solutions, a property that the solutions of the NSE satisfy. The choice $\tilde{\Pi} f(\tilde{\Pi}^* u^h)$ as an approximation of the nonlinear term $f(u)$ is motivated by the fact that

$$(3.59) \quad (\tilde{\Pi} f(\tilde{\Pi}^* u^h), u^h)_{l^2(h\mathbb{Z}^d)} = (f(\tilde{\Pi}^* u^h), \tilde{\Pi}^* u^h)_{l^2(4h\mathbb{Z}^d)} \in \mathbb{R},$$

that, as mentioned above, guarantees the conservation of the $l^2(h\mathbb{Z}^d)$ -norm.

The following holds.

THEOREM 3.3. *Let $p \in (0, 4/d)$ and $q = 4(p + 2)/dp$. Then for all $h > 0$ and for every $\varphi^{4h} \in l^2(4h\mathbb{Z}^d)$, there exists a unique global solution $u^h \in C(\mathbb{R}, l^2(h\mathbb{Z}^d)) \cap L^q_{loc}(\mathbb{R}, l^{p+2}(h\mathbb{Z}^d))$ of (3.58). Moreover, u^h satisfies*

$$(3.60) \quad \|u^h\|_{L^\infty(\mathbb{R}, l^2(h\mathbb{Z}^d))} \leq \|\tilde{\Pi} \varphi^{4h}\|_{l^2(h\mathbb{Z}^d)}$$

and for all finite interval I

$$(3.61) \quad \|u^h\|_{L^q(I, l^{p+2}(h\mathbb{Z}^d))} \leq c(I) \|\tilde{\Pi} \varphi^{4h}\|_{l^2(h\mathbb{Z}^d)},$$

where the above constants are independent of h .

Proof of Theorem 3.3. The local existence and uniqueness can be proved, as in the continuous case, by a combination of the Strichartz-like estimates in Theorem 3.1 and of a fixed point argument in the space $L^\infty((-T, T), l^2(h\mathbb{Z}^d)) \cap L^q((-T, T), l^{p+2}(h\mathbb{Z}^d))$, T being chosen small enough, depending on the initial data, but independent of h . Identity (3.59) guarantees the conservation of the l^2 -norm of the solutions, and, consequently, the lack of blow-up and the global existence of the solutions. \square

3.3. Convergence of the method. In what follows we use the piecewise constant interpolator \mathbf{P}_0^h . Given the initial datum $\varphi \in L^2(\mathbb{R}^d)$ for the PDE, we choose the approximating discrete data $(\varphi_j^{4h})_{j \in \mathbb{Z}^d}$ such that $\mathbf{P}_0^h \tilde{\Pi} \varphi^{4h}$ converges strongly to φ in $L^2(\mathbb{R}^d)$. Thus, in particular, $\|\mathbf{P}_0^h \tilde{\Pi} \varphi^{4h}\|_{L^2(\mathbb{R}^d)} \leq C(\|\varphi\|_{L^2(\mathbb{R}^d)})$.

The main convergence result is the following.

THEOREM 3.4. *Let p and q be as in Theorem 3.3 and u^h be the unique solution of (3.58) for the approximate initial data $\tilde{\Pi} \varphi^{4h}$ as above. Then the sequence $\mathbf{P}_0^h u^h$ satisfies*

$$(3.62) \quad \mathbf{P}_0^h u^h \overset{*}{\rightharpoonup} u \text{ in } L^\infty(\mathbb{R}, L^2(\mathbb{R}^d)), \quad \mathbf{P}_0^h u^h \rightharpoonup u \text{ in } L^q_{loc}(\mathbb{R}, L^{p+2}(\mathbb{R}^d)),$$

$$(3.63) \quad \mathbf{P}_0^h u^h \rightarrow u \text{ in } L^2_{loc}(\mathbb{R}^{d+1}), \quad \mathbf{P}_0^h \tilde{\Pi} f(\tilde{\Pi}^* u^h) \rightharpoonup |u|^p u \text{ in } L^q_{loc}(\mathbb{R}, L^{(p+2)'(\mathbb{R}^d)}),$$

where u is the unique solution of the NSE.

First, we sketch the main ideas of the proof. The main difficulty in the proof of Theorem 3.4 is the strong convergence $\mathbf{P}_0^h u^h \rightarrow u$ in $L^2_{loc}(\mathbb{R}^{d+1})$ which is needed to pass to the limit in the nonlinear term. Once it is obtained, the second convergence in (3.63) easily follows. Another technical difficulty comes from the fact that the interpolator \mathbf{P}_0^h is not compactly supported in the Fourier space. Thus we instead consider the band-limited interpolator \mathbf{P}_*^h introduced in (2.34) and prove the compactness for $\mathbf{P}_*^h u^h$. Once this is obtained, the L^2 -strong convergence of $\mathbf{P}_*^h u^h$ is transferred to $\mathbf{P}_0^h u^h$. This is a consequence of the following property of both interpolators (cf. [22, Thm. 3.4.2, p. 90]):

$$(3.64) \quad \|\mathbf{P}_0^h u^h(t) - \mathbf{P}_*^h u^h(t)\|_{L^2(\Omega)} \leq h \|\mathbf{P}_*^h u^h(t)\|_{H^1(\Omega)},$$

which holds for all real t and $\Omega \subset \mathbb{R}^d$.

To prove the L^2 -strong convergence of $\mathbf{P}_*^h u^h$ we will show that it is uniformly bounded in $L^2_{loc}(\mathbb{R}, H^{1/2}_{loc}(\mathbb{R}^d))$. We shall also obtain estimates in $L^2_{loc}(\mathbb{R}, H^1_{loc}(\mathbb{R}^d))$ which are not uniform on h but, according to (3.64), suffice to ensure that $\mathbf{P}_0^h u^h - \mathbf{P}_*^h u^h$ strongly converges to zero in $L^2_{loc}(\mathbb{R}^{d+1})$. The following lemma provides local estimates for $\mathbf{P}_*^h u^h$ in the H^s -norm.

LEMMA 3.3. *Let $s \geq 1/2$, let $I \subset \mathbb{R}$ be a bounded interval, and let $\chi \in C^\infty_c(\mathbb{R}^d)$. Then there is a constant $C(I, \chi)$, independent of h , such that*

$$(3.65) \quad \|\chi \mathbf{P}_*^h(S^h(t)\tilde{\Pi}\varphi^{4h})\|_{L^2(I, H^s(\mathbb{R}^d))} \leq \frac{C(I, \chi)}{h^{s-1/2}} \|\tilde{\Pi}\varphi^{4h}\|_{l^2(h\mathbb{Z}^d)}$$

holds for all functions $\varphi^{4h} \in l^2(4h\mathbb{Z}^d)$ and $h > 0$. Moreover, for any $d/2$ -admissible pair (q, r)

$$(3.66) \quad \left\| \chi \mathbf{P}_*^h \left(\int_0^t S^h(t-\tau)\tilde{\Pi}f^{4h}(\tau)d\tau \right) \right\|_{L^2(I, H^s(\mathbb{R}^d))} \leq \frac{C(I, \chi)}{h^{s-1/2}} \|\tilde{\Pi}f^{4h}\|_{L^{q'}(I, l^{r'}(h\mathbb{Z}^d))}$$

for all $f^{4h} \in L^{q'}(I, l^{r'}(4h\mathbb{Z}^d))$ and $h > 0$.

Proof. We divide the proof into two steps. The first one concerns the homogeneous estimate (3.65) and the second one (3.66).

Step 1. Regularity of the homogeneous term. To prove (3.65) it is sufficient to prove, for any $R > 0$, the existence of a positive constant $C(I, R)$ such that

$$\int_I \int_{|x|<R} |(-\Delta)^{s/2} \mathbf{P}_*^h(S^h(t)\tilde{\Pi}\varphi^{4h})|^2 dxdt \leq \frac{C(I, R)}{h^{2s-1}} \int_{[-\pi/h, \pi/h]^d} |\widehat{\varphi}^{4h}(\xi)|^2 d\xi.$$

Let us consider $\psi^h \in l^2(h\mathbb{Z}^d)$. Applying Theorem 2.5 to the function $\mathbf{P}_*^h(S^h(t)\psi^h)$ we obtain

$$(3.67) \quad \begin{aligned} \int_I \int_{|x|<R} |(-\Delta)^{s/2} \mathbf{P}_*^h(S^h(t)\psi^h)|^2 dxdt &\leq C(I, R) \int_{[-\pi/h, \pi/h]^d} \frac{|\xi|^{2s} |\widehat{\mathbf{P}_*^h \psi^h}(\xi)|^2 d\xi}{|\nabla p_h(\xi)|} \\ &\leq \frac{C(I, R)}{h^{2s-1}} \int_{[-\pi/h, \pi/h]^d} \frac{(\sum_{j=1}^d \xi_j^2)^{1/2} |\widehat{\psi^h}(\xi)|^2 d\xi}{(\sum_{j=1}^d \sin^2(\xi_j h)/h^2)^{1/2}} \\ &\lesssim \frac{C(I, R)}{h^{2s-1}} \int_{[-\pi/h, \pi/h]^d} \frac{|\widehat{\psi^h}(\xi)|^2 d\xi}{\prod_{j=1}^d |\cos(\xi_j h/2)|}, \end{aligned}$$

provided that all terms make sense. Note that this estimate holds for all $\psi \in l^2(h\mathbb{Z}^d)$. Observe, however, that the term in the denominator in the right-hand side integral may vanish for the high frequencies $\xi = (\pm\pi/h)^d$. In order to compensate this fact we consider initial data in the class of slowly oscillating sequences $\widetilde{\Pi}(4h\mathbb{Z}^d)$. Now, we apply the last estimates to $\psi^h = \widetilde{\Pi}\varphi^{4h}$. Thus

$$\begin{aligned} \int_I \int_{|x|<R} |(-\Delta)^{s/2} \mathbf{P}_*^h(S^h(t)\widetilde{\Pi}\varphi^{4h})|^2 dx dt &\leq \frac{C(I, R)}{h^{2s-1}} \int_{[-\pi/h, \pi/h]^d} \frac{|\widehat{\widetilde{\Pi}\varphi^{4h}}(\xi)|^2 d\xi}{\prod_{j=1}^d |\cos(\xi_j h/2)|} \\ &\leq \frac{C(I, R)}{h^{2s-1}} \int_{[-\pi/h, \pi/h]^d} |\widehat{\varphi}^{4h}(\xi)|^2 \prod_{j=1}^d |\cos(\xi_j h/2)|^3 d\xi \leq \frac{C(I, R)}{h^{2s-1}} \|\widetilde{\Pi}\varphi^{4h}\|_{l^2(h\mathbb{Z}^d)}. \end{aligned}$$

Step 2. Regularity of the inhomogeneous term. In the following we prove (3.66). This estimate will be reduced to the homogeneous one (3.65) by using the argument of Christ and Kiselev [4] (see also [24] in the context of the PDE). A simplified version, useful in PDE applications, is given in [24].

LEMMA 3.4. *Let X and Y be Banach spaces and assume that $K(t, s)$ is a continuous function taking its values in $B(X, Y)$, the space of bounded linear mappings from X to Y . Suppose that $-\infty \leq a < b \leq \infty$ and set*

$$Tf(t) = \int_a^b K(t, s)f(s)ds, \quad Wf(t) = \int_a^t K(t, s)f(s)ds.$$

Assume that $1 \leq p < q \leq \infty$ and $\|Tf\|_{L^q([a,b],Y)} \leq \|f\|_{L^p([a,b],X)}$. Then

$$\|Wf\|_{L^q([a,b],Y)} \leq \|f\|_{L^p([a,b],X)}.$$

Without loss of generality we can consider $I = [0, T]$. In view of the above lemma it is sufficient to prove that the operator

$$Tf^{4h}(t) = \chi \mathbf{P}_*^h \left(\int_0^T S^h(t-\tau)\widetilde{\Pi}f^{4h}(\tau)d\tau \right)$$

satisfies

$$\|Tf^{4h}\|_{L^2([0,T], H^s(\mathbb{R}^d))} \leq \frac{C(T, \chi)}{h^{s-1/2}} \|\widetilde{\Pi}f^{4h}\|_{L^{q'}([0,T], l^{r'}(h\mathbb{Z}^d))}.$$

We write Tf^{4h} as $Tf^{4h}(t) = \chi \mathbf{P}_*^h S^h(t)T_1f^{4h}(t)$, where

$$T_1f^{4h}(t) = \int_0^T S^h(s)^* \widetilde{\Pi}f^{4h}(s)ds.$$

Estimate (3.67) yields

$$\begin{aligned} \|Tf^{4h}\|_{L^2([0,T], H^s(\mathbb{R}^d))} &\leq \frac{C(I, \chi)}{h^{s-1/2}} \left\| \frac{\widehat{T_1f^{4h}}(\xi)}{\prod_{j=1}^d |\cos(\xi_j h/2)|^{1/2}} \right\|_{L^2([-\pi/h, \pi/h]^d)} \\ &\lesssim \frac{C(I, \chi)}{h^{s-1/2}} \left\| \frac{\widehat{T_1f^{4h}}(\xi)}{\prod_{j=1}^d |\cos(\xi_j h/2)|^{1/2} |\cos(\xi_j h)|^{1/2}} \right\|_{L^2([-\pi/h, \pi/h]^d)}, \end{aligned}$$

provided that all the above integrals are finite.

Explicit computations on $T_1 f^{4h}$ show that

$$\begin{aligned} & \frac{\widehat{T_1 f^{4h}}(\xi)}{\prod_{j=1}^d |\cos(\xi_j h/2)|^{1/2} |\cos(\xi_j h)|^{1/2}} \\ &= 4^d \int_0^T e^{i s p_h(\xi)} \prod_{j=1}^d \left| \cos\left(\frac{\xi_j h}{2}\right) \right|^{3/2} |\cos(\xi_j h)|^{3/2} \widehat{\Pi f^{4h}}(\xi, s) ds \\ &= 4^d \left(\int_0^T (A_{3/2}^h(s))^* \Pi f^{4h}(s) ds \right) \wedge(\xi), \end{aligned}$$

where the operator $A_{3/2}^h$ is defined in (3.52).

Applying Theorem 2.6 to the operator $A_{3/2}^h$ we obtain, by estimate (2.38), that

$$\left\| \int_0^T (A_{3/2}^h(s))^* \Pi f^{4h}(s) ds \right\|_{l^2(h\mathbb{Z}^d)} \lesssim \|\Pi f^{4h}\|_{L^{q'}([0,T], l^{r'}(h\mathbb{Z}^d))} \lesssim \|\widetilde{\Pi} f^{4h}\|_{L^{q'}([0,T], l^{r'}(h\mathbb{Z}^d))}.$$

The proof is now complete. \square

Proof of Theorem 3.4. Using (3.60) we obtain that $\mathbf{P}_0^h u^h$ is uniformly bounded in $L^\infty(\mathbb{R}, L^2(\mathbb{R}^d))$. This guarantees the existence of a function $u \in L^\infty(\mathbb{R}, L^2(\mathbb{R}^d))$ such that, up to a subsequence, $\mathbf{P}_0^h u^h \rightharpoonup^* u$ in $L^\infty(\mathbb{R}, L^2(\mathbb{R}^d))$. By (3.61) we obtain that $u \in L^q(I, L^{p+2}(\mathbb{R}^d))$ and, up to a subsequence, $\mathbf{P}_0^h u^h \rightharpoonup u$ in $L^q(I, L^{p+2}(\mathbb{R}^d))$.

In the following we prove the strong convergence of $\mathbf{P}_0^h u^h$. First, we prove that $\mathbf{P}_0^h u^h - \mathbf{P}_*^h u^h \rightarrow 0$ in $L_{loc}^2(\mathbb{R} \times \mathbb{R}^d)$. Second, we prove the compactness of $\mathbf{P}_*^h u^h$. Finally, we obtain that $\mathbf{P}_0^h u^h \rightarrow u$ in $L_{loc}^2(\mathbb{R} \times \mathbb{R}^d)$.

For any $\Omega \subset \mathbb{R}^d$, classical properties of the interpolator $\mathbf{P}_0^h u^h$ (see [22, Thm. 3.4.2, p. 90]) give us

$$\int_\Omega |\mathbf{P}_0^h u^h - \mathbf{P}_*^h u^h|^2 dx \leq h^2 \|\mathbf{P}_*^h u^h\|_{H^1(\Omega)}^2.$$

Applying Lemma 3.3 with $s = 1$ we obtain, for any $\chi \in C_c^\infty(\mathbb{R}^d)$,

$$\begin{aligned} \int_I \int_{\mathbb{R}^d} \chi^2 |\mathbf{P}_0^h u^h - \mathbf{P}_*^h u^h|^2 dx dt &\leq h^2 \int_I \int_{\mathbb{R}^d} \chi^2 |(I - \Delta)^{1/2} \mathbf{P}_*^h u^h|^2 dx dt \\ &\leq hC(I, \|\widetilde{\Pi} \varphi^{4h}\|_{l^2(h\mathbb{Z}^d)}^2) \rightarrow 0, \quad h \rightarrow 0. \end{aligned}$$

This shows that $\mathbf{P}_0^h u^h - \mathbf{P}_*^h u^h \rightarrow 0$ in $L_{loc}^2(\mathbb{R} \times \mathbb{R}^d)$.

Using Lemma 3.3 with $s = 1/2$ we obtain that for any smooth function χ , $\mathbf{P}_*^h u^h$ satisfies

$$\|\chi \mathbf{P}_*^h u^h\|_{L^2(I, H^{1/2}(\mathbb{R}^d))} \leq C(I, \chi, \|\widetilde{\Pi} \varphi^{4h}\|_{l^2(h\mathbb{Z}^d)}).$$

We can also prove the following uniform boundedness property of its time derivative:

$$\begin{aligned} \left\| \frac{d\mathbf{P}_*^h u^h}{dt} \right\|_{L^1(I, H^{-2}(\mathbb{R}^d))} &\leq \|\Delta_h \mathbf{P}_*^h u^h\|_{L^1(I, H^{-2}(\mathbb{R}^d))} + \|\mathbf{P}_*^h(|u^h|^p u^h)\|_{L^1(I, H^{-2}(\mathbb{R}^d))} \\ &\leq \|\mathbf{P}_*^h u^h\|_{L^1(I, L^2(\mathbb{R}^d))} + \|\mathbf{P}_*^h(|u^h|^p u^h)\|_{L^1(I, L^{(p+2)'(\mathbb{R}^d))} \leq C(I, \|\varphi\|_{L^2(\mathbb{R}^d)}). \end{aligned}$$

Using the embeddings $H^s(\Omega) \hookrightarrow_{comp} L^2(\Omega) \hookrightarrow H^{-2}(\Omega)$, $\Omega \subset \mathbb{R}^d$ being a bounded domain, and the compactness results of [23] we obtain the existence of a function v such that, up to subsequences, $\mathbf{P}_*^h u^h \rightarrow v$ in $L^2_{loc}(\mathbb{R} \times \mathbb{R}^d)$. Using the strong convergence of $\mathbf{P}_*^h u^h$ towards v we obtain that $v = u$ and $\mathbf{P}_0^h u^h \rightarrow u$ in $L^2_{loc}(\mathbb{R} \times \mathbb{R}^d)$.

Let $\Gamma \subset \mathbb{Z}^d$ be a finite set. Thus for any $s \in \Gamma$ we have $\mathbf{P}_0^h u^h(\cdot + sh) \rightarrow u$ in $L^2_{loc}(\mathbb{R} \times \mathbb{R}^d)$ and $\mathbf{P}_0^h u^h(\cdot + sh) \rightarrow u$ a.e. in $\mathbb{R} \times \mathbb{R}^d$. The operators $\tilde{\Pi}$ and $\tilde{\Pi}^*$ involve only a finite number of translations. Then $\mathbf{P}_0^h \tilde{\Pi} f(\tilde{\Pi}^* u^h) \rightarrow |u|^p u$ a.e. in $\mathbb{R} \times \mathbb{R}^d$ and $\mathbf{P}_0^h \tilde{\Pi} f(\tilde{\Pi}^* u^h) \rightharpoonup |u|^p u$ in $L^q(I, L^{(p+2)' }(\mathbb{R}^d))$.

Multiplying (3.58) by a function $\psi \in C_c^\infty(\mathbb{R}^{d+1})$, $\mathbf{P}_0^h u^h$ satisfies

$$(3.68) \quad \int_{\mathbb{R}} \int_{\mathbb{R}^d} \mathbf{P}_0^h u^h (-i\psi_t + \Delta^h \psi) dx dt = \int_{\mathbb{R}} \int_{\mathbb{R}^d} \mathbf{P}_0^h \tilde{\Pi} f(\tilde{\Pi}^* u^h) \psi dx dt.$$

All the above weak convergences of $\mathbf{P}_0^h u^h$ and (3.68) show that u satisfies (3.57).

It remains to prove that $u \in C(\mathbb{R}, L^2(\mathbb{R}^d))$ and $u(0) = \varphi$. To prove that $u \in C(\mathbb{R}, L^2(\mathbb{R}^d))$ we show its continuity at $t = 0$; the same argument works at any time t .

For any positive $0 \leq t \leq T < 1$, the Strichartz estimates in Theorem 3.1 and the Hölder inequality in time variable applied to the variation of constants formula give us

$$\begin{aligned} \|u^h(t) - S^h(t)\tilde{\Pi}\varphi^{4h}\|_{l^2(h\mathbb{Z}^d)} &\leq \left\| \int_0^t S^h(t-s)\tilde{\Pi}f(\tilde{\Pi}^*u^h)ds \right\|_{L^\infty([0,T], l^2(\mathbb{Z}^d))} \\ &\lesssim \| |u^h|^p u^h \|_{L^{q(h)' }([0,T], l^{(p+2)' }(h\mathbb{Z}^d))} \leq T^{(q-(p+2))/q} \|u^h\|_{L^q([0,T], l^{p+2}(h\mathbb{Z}^d))}^{p+1} \\ &\lesssim T^{1-pd/4} C(\|\varphi\|_{L^2(\mathbb{R}^d)}). \end{aligned}$$

Using that $\mathbf{P}_0^h u^h \rightharpoonup^* u$ and $\mathbf{P}_0^h S^h(\cdot)\varphi^h \rightharpoonup^* S(\cdot)\varphi$ in $L^\infty([0, T], L^2(\mathbb{R}^d))$ we get

$$\begin{aligned} \|u(t) - S(t)\varphi\|_{L^2(\mathbb{R}^d)} &\leq \liminf_{h \rightarrow 0} \|\mathbf{P}_0^h u^h(\cdot) - \mathbf{P}_0^h S^h(\cdot)\tilde{\Pi}\varphi^{4h}\|_{L^\infty([0,T], L^2(\mathbb{R}^d))} \\ &\lesssim T^{1-pd/4} C(\|\varphi\|_{L^2(\mathbb{R}^d)}). \end{aligned}$$

This proves that the solution u obtained as the limit of $\mathbf{P}_0^h u^h$ satisfies $u(t) \rightarrow \varphi$ in $L^2(\mathbb{R}^d)$ as $t \rightarrow 0$.

The uniqueness of the limit, a solution of the NSE (3.56), allows us to deduce that the whole sequence $\mathbf{P}_0^h u^h$ converges without extracting subsequences.

The proof of Theorem 3.4 is now complete. \square

3.4. The critical case $p = 4/d$. Our method works similarly in the critical case $p = 4/d$ for small initial data. More precisely, the following holds.

THEOREM 3.5. *There exists a constant ϵ , independent of h , such that for all initial data $\varphi^h \in \tilde{\Pi}(4h\mathbb{Z}^d)$ with $\|\varphi^h\|_{l^2(h\mathbb{Z}^d)} < \epsilon$, the semidiscrete critical equation (3.58) with $p = 4/d$ has a unique global solution $u^h \in C(\mathbb{R}, l^2(h\mathbb{Z}^d)) \cap L^{2+4/d}_{loc}(\mathbb{R}, l^{2+4/d}(h\mathbb{Z}^d))$. Moreover, for any $d/2$ -admissible pair (q, r) , $u^h \in L^q_{loc}(\mathbb{R}, l^r(h\mathbb{Z}^d))$ and*

$$\|u^h\|_{L^q(I, l^r(h\mathbb{Z}^d))} \leq C(q, I)\|\varphi^h\|_{l^2(h\mathbb{Z}^d)}$$

for all finite intervals I , uniformly on h .

With the same notation, as in the subcritical case, the following convergence result holds.

THEOREM 3.6. *Let $p = 4/d$. Under the smallness assumption of Theorem 3.5, the sequence $\mathbf{P}_0^h u^h$ satisfies*

$$\begin{aligned} \mathbf{P}_0^h u^h \overset{*}{\rightharpoonup} u \text{ in } L^\infty(\mathbb{R}, L^2(\mathbb{R}^d)), \quad \mathbf{P}_0^h u^h \rightharpoonup u \text{ in } L_{loc}^{4/d+2}(\mathbb{R}, L^{4/d+2}(\mathbb{R}^d)), \\ \mathbf{P}_0^h u^h \rightharpoonup u \text{ in } L_{loc}^2(\mathbb{R} \times \mathbb{R}^d), \quad \mathbf{P}_0^h \tilde{\Pi}(f(\tilde{\Pi}^* u^h)) \rightharpoonup |u|^{4/d} u \text{ in } L_{loc}^{(4/d+2)'}(\mathbb{R}, L^{(4/d+2)' }(\mathbb{R}^d)), \end{aligned}$$

where u is the unique weak solution of the critical NSE with $p = 4/d$.

In contrast with the viscous numerical scheme introduced in [12] this time we do not need to modify the exponent $4/d$ of the nonlinearity in the numerical scheme. In the present case, the class of Strichartz estimates for the linear semidiscrete semigroup hold for $d/2$ -admissible pairs and not for the some α -admissible pairs, $\alpha > d/2$. This allows us to use, for the numerical scheme based on the two-grid method, exactly the same nonlinearity as that given by the nonlinear problem after adapting it by means of extension and restriction operators $\tilde{\Pi}$ and $\tilde{\Pi}^*$ as in (3.58).

We have analyzed here the case of small L^2 -initial data. In the continuous case, the global well-posedness can be proved under a more general assumption:

$$(3.69) \quad \|e^{it\Delta} \varphi\|_{L^{2+4/d}(\mathbb{R}, L^{2+4/d}(\mathbb{R}^d))} \leq c_0$$

for some sufficiently small constant c_0 . Examples of φ satisfying (3.69) with large $L^2(\mathbb{R}^d)$ -norm are given in [17, Chap. 5, section 5.4, p. 108–109].

At the numerical level, condition (3.69) can be replaced by

$$(3.70) \quad \|S^h(t)\varphi^h\|_{L^{2+4/d}(\mathbb{R}, l^{2+4/d}(h\mathbb{Z}^d))} \leq c_1,$$

where c_1 is a positive, small enough constant and $\varphi^h \in \tilde{\Pi}(4h\mathbb{Z}^d)$. Clearly, for $\varphi^h \in \tilde{\Pi}(4h\mathbb{Z}^d)$ with small $l^2(h\mathbb{Z}^d)$ -norm, estimate (3.48) shows (3.70). The construction of $\varphi^h \in \tilde{\Pi}(4h\mathbb{Z}^d)$ with large $l^2(h\mathbb{Z}^d)$ -norm satisfying (3.70) is an open problem.

REFERENCES

- [1] M. J. ABLOWITZ AND J. F. LADIK, *Nonlinear differential-difference equations*, J. Math. Phys., 16 (1975), pp. 598–603.
- [2] M. J. ABLOWITZ, B. PRINARI, AND A. D. TRUBATCH, *Discrete and Continuous Nonlinear Schrödinger Systems*, London Math. Soc. Lecture Note Ser. 302, Cambridge University Press, Cambridge, UK, 2004.
- [3] T. CAZENAVE, *Semilinear Schrödinger Equations*, Courant Lect. Notes Math. 10, American Mathematical Society, Providence, RI, Courant Institute of Mathematical Sciences, New York, 2003.
- [4] M. CHRIST AND A. KISELEV, *Maximal functions associated to filtrations*, J. Funct. Anal., 179 (2001), pp. 409–425.
- [5] P. CONSTANTIN AND J. C. SAUT, *Local smoothing properties of dispersive equations*, J. Amer. Math. Soc., 1 (1988), pp. 413–439.
- [6] J. GIANNOULIS, M. HERRMANN, AND A. MIELKE, *Continuum descriptions for the dynamics in discrete lattices: Derivation and justification*, in Analysis, Modeling and Simulation of Multiscale Problems, Vol. 18, A. Mielke, ed., Springer, Berlin, 2006, pp. 435–466.
- [7] G. GIGANTE AND F. SORIA, *On a sharp estimate for oscillatory integrals associated with the Schrödinger equation*, Int. Math. Res. Not., no. 24 (2002), pp. 1275–1293.
- [8] J. GINIBRE AND G. VELO, *The global Cauchy problem for the nonlinear Schrödinger equation revisited*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 2 (1985), pp. 309–327.
- [9] R. GLOWINSKI, *Ensuring well-posedness by analogy: Stokes problem and boundary control for the wave equation*, J. Comput. Phys., 103 (1992), pp. 189–221.

- [10] L. I. IGNAT, *Fully discrete schemes for the Schrödinger equation. Dispersive properties*, Math. Models Methods Appl. Sci., 17 (2007), pp. 567–591.
- [11] L. I. IGNAT AND E. ZUAZUA, *A two-grid approximation scheme for nonlinear Schrödinger equations: Dispersive properties and convergence*, C. R. Acad. Sci. Paris, 341 (2005), pp. 381–386.
- [12] L. I. IGNAT AND E. ZUAZUA, *Dispersive properties of a viscous numerical scheme for the Schrödinger equation*, C. R. Acad. Sci. Paris, 340 (2005), pp. 529–534.
- [13] L. I. IGNAT AND E. ZUAZUA, *Dispersive properties of numerical schemes for nonlinear Schrödinger equations*, in Foundations of Computational Mathematics, Santander 2005, London Math. Soc. Lecture Note Ser. 331, L. M. Pardo et al., eds., Cambridge University Press, Cambridge, UK, 2006, pp. 181–207.
- [14] M. KEEL AND T. TAO, *Endpoint Strichartz estimates*, Amer. J. Math., 120 (1998), pp. 955–980.
- [15] C. E. KENIG, G. PONCE, AND L. VEGA, *Oscillatory integrals and regularity of dispersive equations*, Indiana Univ. Math. J., 40 (1991), pp. 33–69.
- [16] C. E. KENIG, G. PONCE, AND L. VEGA, *Small solutions to nonlinear Schrödinger equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 10 (1993), pp. 255–288.
- [17] F. LINARES AND G. PONCE, *Introduction to Nonlinear Dispersive Equations*, Publicações Matemáticas, IMPA, Rio de Janeiro, Brazil, 2004.
- [18] A. MAGYAR, E. M. STEIN, AND S. WAINGER, *Discrete analogues in harmonic analysis: Spherical averages*, Ann. of Math. (2), 155 (2002), pp. 189–208.
- [19] A. MIELKE, *Macroscopic behavior of microscopic oscillations in harmonic lattices via Wigner-Husimi transforms*, Arch. Ration. Mech. Anal., 181 (2006), pp. 401–448.
- [20] M. NIXON, *The discretized generalized Korteweg-de Vries equation with fourth order nonlinearity*, J. Comput. Anal. Appl., 5 (2003), pp. 369–397.
- [21] M. PLANCHEREL AND G. PÓLYA, *Fonctions entières et intégrales de Fourier multiples. II*, Comment. Math. Helv., 10 (1937), pp. 110–163.
- [22] A. QUARTERONI AND A. VALLI, *Numerical Approximation of Partial Differential Equations*, Springer Ser. Comput. Math. 23, Springer, Berlin, 1994.
- [23] J. SIMON, *Compact sets in the space $L^p(0, T; B)$* , Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65–96.
- [24] G. STAFFILANI AND D. TATARU, *Strichartz estimates for a Schrödinger operator with nonsmooth coefficients*, Comm. Partial Differential Equations, 27 (2002), pp. 1337–1372.
- [25] A. STEFANOV AND P. G. KEVREKIDIS, *Asymptotic behaviour of small solutions for the discrete nonlinear Schrödinger and Klein-Gordon equations*, Nonlinearity, 18 (2005), pp. 1841–1857.
- [26] E. M. STEIN, *Harmonic Analysis: Real-Variable Methods, Orthogonality, and Oscillatory Integrals*, Princeton Math. Ser. 43, Princeton University Press, Princeton, NJ, 1993.
- [27] R. S. STRICHARTZ, *Restrictions of Fourier transforms to quadratic surfaces and decay of solutions of wave equations*, Duke Math. J., 44 (1977), pp. 705–714.
- [28] T. TAO, *Nonlinear Dispersive Equations: Local and Global Analysis*, CBMS Regional Conf. Ser. in Math. 106, American Mathematical Society, Providence, RI, 2006.
- [29] L. N. TREFETHEN, *Spectral Methods in MATLAB*, Software Environ. Tools 10, SIAM, Philadelphia, 2000.
- [30] Y. TSUTSUMI, *L^2 -solutions for nonlinear Schrödinger equations and nonlinear groups*, Funkcial. Ekvac., 30 (1987), pp. 115–125.
- [31] R. VICHNEVETSKY AND J. B. BOWLES, *Fourier Analysis of Numerical Approximations of Hyperbolic Equations*, SIAM Stud. Appl. Math., 5, SIAM, Philadelphia, 1982.

DISCONTINUOUS GALERKIN METHODS FOR ADVECTION-DIFFUSION-REACTION PROBLEMS*

BLANCA AYUSO[†] AND L. DONATELLA MARINI[‡]

Abstract. We apply the weighted-residual approach recently introduced in [F. Brezzi et al., *Comput. Methods Appl. Mech. Engrg.*, 195 (2006), pp. 3293–3310] to derive discontinuous Galerkin formulations for advection-diffusion-reaction problems. We devise the basic ingredients to ensure stability and optimal error estimates in suitable norms, and propose two new methods.

Key words. discontinuous Galerkin, advection-diffusion-reaction, inf-sup condition

AMS subject classifications. 65N30, 65N12, 65G99, 76R99

DOI. 10.1137/080719583

1. Introduction. In recent years discontinuous Galerkin (DG) methods have become increasingly popular, and they have been used and analyzed for various kinds of applications: see, e.g., [2] for second order elliptic problems, [4], [3] for Reissner–Mindlin plates, and, for advection-diffusion problems, [13], [14], [23], [38], [20], [24], and [10].

Most DG methods for advection-diffusion or hyperbolic problems are constructed by specifying the numerical fluxes at the interelements, and, as far as we know, the advection field is mostly assumed to be either constant or divergence-free. In the present paper we follow a different path. On one hand, we derive DG formulations by applying the so-called weighted-residual approach of [6]. In this approach a DG method is written first in strong form, as a system of equations including the original PDE equation inside each element plus the necessary continuity conditions at interfaces. The variational form is then obtained by combining all these equations. In this way, the DG method establishes a linear relationship between the residual inside each element and the jumps across interelement boundaries. Such a linear relation permits us to recover DG methods proposed earlier in the literature, and at the same time provides a framework for devising new DG methods with the desired stability and consistency properties. As we shall show, this is possible, since stability and consistency can be ensured through a proper selection of the weights in the linear relationship, which in turn determines the DG method.

On the other hand, we deal with a variable reaction and a variable advection field which is not divergence-free. With respect to other papers treating variable coefficients (see, e.g., [17], [18], [11]) the novelty of the present paper is that we relax the usual coercivity condition relating advection and reaction (see condition (2.2) in section 2). To the best of our knowledge the weaker coercivity condition was assumed in [21], but there the advection field is constant, while for variable coefficients similar assumptions

*Received by the editors March 31, 2008; accepted for publication (in revised form) November 17, 2008; published electronically February 25, 2009.

<http://www.siam.org/journals/sinum/47-2/71958.html>

[†]Departamento de Matemáticas, Universidad Autónoma de Madrid, Madrid 28049, Spain (blanca.ayuso@uam.es). The work of this author was partially supported by MEC under project MTM2005-00714 and by CAM under project S0505/ESP-0158.

[‡]Dipartimento di Matematica, Università degli Studi di Pavia and IMATI del CNR, Via Ferrata 1, 27100 Pavia, Italy (marini@imati.cnr.it). The work of this author was partially supported by MIUR under project PRIN2006.

in a different context were used in [19]. Clearly, the weaker condition (2.2), together with variable coefficients, makes the analysis more complicated than usual, surely more complicated than one could expect at first sight, if one wants to take care of situations where advection and/or reaction dominate in different parts of the domain or, more generally, when diffusion is (comparatively) very small.

To ease the presentation we apply the weighted-residual approach to derive two DG methods proposed in the literature: the method introduced in [23], and that proposed in [24] and further analyzed in [10]. The former uses the nonsymmetric NIPG method for the diffusion terms and upwind for the convective part of the flux. In the latter the diffusion terms are treated with three different DG methods, and the whole physical flux is upwind. This makes the approach well suited for strongly advection-dominated problems (actually, the most interesting cases) but less adequate in the diffusion-dominated or intermediate regimes. We also introduce two new methods. One of them, that we refer to as *minimal choice*, contains the minimum number of terms needed to get stability and optimal order of convergence in all regimes. The other one is a more refined method, that contains as a particular case the method [20] and the *minimal choice*.

Our formulation allows us also to recover easily, for each of the methods analyzed, the corresponding SUPG-stabilized version. Many others methods could have been considered, but this would have made the paper practically unreadable. Moreover, our aim was not to compare the behavior of different schemes, but mostly to explore the possibilities and the ductility of the weighted-residual approach for designing and analyzing DG methods.

It is worth noticing that this approach seems to be particularly suited for understanding in a natural way which stabilization mechanisms are, hidden in each DG method, responsible for the behavior of the DG approximation in the different regimes of the problem. It also provides a way to perform stability and a priori error analysis in a unified framework. Furthermore, we think that it could be useful also for applications to a posteriori error analysis, a field which is well developed for conforming approximations but much less studied for DG approximations or even stabilized methods. This surely deserves some further and future research.

Throughout the paper we shall use standard notation for norms and seminorms in Sobolev spaces. To keep homogeneity of dimensions, we recall that on a domain Ω of diameter L we define

$$(1.1) \quad \|v\|_{k,\Omega}^2 := \sum_{s=0}^k L^{2s} |v|_{s,\Omega}^2, \quad v \in H^k(\Omega), \quad k \geq 0,$$

$$(1.2) \quad \|v\|_{k,\infty,\Omega} := \sum_{s=0}^k L^s |v|_{s,\infty,\Omega}, \quad v \in W^{k,\infty}(\Omega), \quad k \geq 0.$$

The outline of the paper is as follows. In section 2 we present the problem with all the assumptions necessary to the analysis, and we apply the weighted-residual approach. In section 3 we show examples of choices of the “weights,” leading to four methods: the methods of [23] and [24], and two new methods. In section 4 we deal with the approximation and prove stability in a suitable DG norm. We also prove stability in a norm of SUPG-type, thus providing control on the streamline derivative. Section 5 is devoted to a priori error analysis, and optimal convergence is proved in both norms. Finally, in section 6 we present an extensive set of numerical experiments to compare the methods and to validate our theoretical results.

2. Setting of the problem. To ease the presentation we shall restrict ourselves to the two-dimensional case, although the results here presented also hold in three dimensions. Let Ω be a bounded, convex, polygonal domain in \mathbf{R}^2 , and let $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ be the velocity vector field defined on $\overline{\Omega}$ with $\beta_i \in W^{1,\infty}(\Omega)$, $i = 1, 2$, $\gamma \in L^\infty(\Omega)$ the reaction coefficient, and ε a positive constant diffusivity coefficient. We define the *inflow* and *outflow* parts of $\Gamma = \partial\Omega$ in the usual fashion:

$$\begin{aligned} \Gamma^- &= \{x \in \Gamma : \boldsymbol{\beta}(x) \cdot \mathbf{n}(x) < 0\} = \text{inflow}, \\ \Gamma^+ &= \{x \in \Gamma : \boldsymbol{\beta}(x) \cdot \mathbf{n}(x) \geq 0\} = \text{outflow}, \end{aligned}$$

where $\mathbf{n}(x)$ denotes the unit outward normal vector to Γ at $x \in \Gamma$. Let $\Gamma_D \neq \emptyset$, and let Γ_N be the parts of the boundary Γ where Dirichlet and Neumann boundary conditions are assigned, so that $\Gamma = \overline{\Gamma_D} \cup \overline{\Gamma_N}$, $\Gamma_D \cap \Gamma_N = \emptyset$. Thus,

$$\Gamma_D^\pm = \Gamma_D \cap \Gamma^\pm, \quad \Gamma_N^\pm = \Gamma_N \cap \Gamma^\pm.$$

Let $f \in L^2(\Omega)$, $g_D \in H^{3/2}(\Gamma_D)$, $g_N \in H^{1/2}(\Gamma_N)$. Consider the advection-diffusion-reaction problem

$$(2.1) \quad \begin{aligned} \operatorname{div} \boldsymbol{\sigma}(u) + \gamma u &= f && \text{in } \Omega, \\ u &= g_D && \text{on } \Gamma_D, \\ (\boldsymbol{\beta} u \chi_{\Gamma_N^-} - \varepsilon \nabla u) \cdot \mathbf{n} &= g_N && \text{on } \Gamma_N, \end{aligned}$$

where $\boldsymbol{\sigma}(u)$ is the (physical) flux, given by

$$\boldsymbol{\sigma}(u) = -\varepsilon \nabla u + \boldsymbol{\beta} u,$$

and $\chi_{\Gamma_N^-}$ is the characteristic function of Γ_N^- . The meaning of the boundary conditions on Γ_N is that the total flux is imposed on Γ_N^- while on Γ_N^+ only the diffusive flux is specified (see [24]).

Since the first equation in (2.1) is equivalent to $-\varepsilon \Delta u + \boldsymbol{\beta} \cdot \nabla u + (\operatorname{div} \boldsymbol{\beta} + \gamma)u = f$, we introduce the “effective” reaction function $\varrho(x)$ and we make the assumption

$$(2.2) \quad \varrho(x) := \gamma(x) + \frac{1}{2} \operatorname{div} \boldsymbol{\beta}(x) \geq \varrho_0 \geq 0 \quad \forall x \in \Omega.$$

For the subsequent stability and error analysis we shall make the following assumptions on the coefficients: the advective field has neither closed curves nor stationary points, i.e.,

$$(2.3) \quad \boldsymbol{\beta} \text{ has no closed curves} \quad \text{and} \quad |\boldsymbol{\beta}(x)| \neq 0 \quad \forall x \in \Omega.$$

This implies, as we shall see later on (see Remark 2.1 and Appendix A), that

$$(H1) \quad \exists \eta \in W^{k+1,\infty}(\Omega) \quad \text{such that} \quad \boldsymbol{\beta} \cdot \nabla \eta \geq 2b_0 := 2 \frac{\|\boldsymbol{\beta}\|_{0,\infty,\Omega}}{L} \quad \text{in } \Omega.$$

Furthermore, we assume that

$$(H2) \quad \exists c_\beta > 0 \text{ such that } |\boldsymbol{\beta}(x)| \geq c_\beta \|\boldsymbol{\beta}\|_{1,\infty,\Omega} \quad \forall x \in \Omega,$$

and, given a shape-regular family \mathcal{T}_h of decompositions of Ω into triangles T ,

$$(H3) \quad \exists c_\varrho > 0 \text{ such that } \forall T \in \mathcal{T}_h \quad \|\varrho\|_{0,\infty,T} \leq c_\varrho (\min_T \varrho(x) + b_0).$$

Remark 2.1. Assumption (2.3), together with the regularity $\beta \in W^{1,\infty}(\Omega)$, ensures the well-posedness of the continuous problem in the pure hyperbolic limit ($\varepsilon = 0$). (See [16] and also [33] for details.) Condition (H1) is based on a result first established in [16, Lemma 2.3] under more regularity assumptions on β . Namely, for $\beta \in \mathcal{C}^k(\mathcal{U})$, $k \geq 1$ satisfying (2.3), \mathcal{U} being some neighborhood of $\overline{\Omega}$, the authors show the existence of $\eta \in \mathcal{C}^k(\mathcal{U})$ verifying $\beta \cdot \nabla \eta \geq b_0 > 0$ in Ω . However, by revisiting the proof in [16], it can be seen that the result holds true also if $\beta \in W^{1,\infty}(\Omega)$, provided it satisfies (2.3) (see Appendix A for details).

Assumption (H2) excludes undesirable situations of a small but highly oscillatory advection field and provides useful relations among norms. Indeed, from (1.2) we deduce

$$(2.4) \quad \begin{aligned} c_\beta \frac{\|\beta\|_{1,\infty,\Omega}}{L} &\leq b_0 := \frac{\|\beta\|_{0,\infty,\Omega}}{L} \leq \frac{\|\beta\|_{1,\infty,\Omega}}{L}, \\ \|\beta\|_{1,\infty,\Omega} &\leq \frac{\|\beta\|_{1,\infty,\Omega}}{L} \leq \frac{1}{c_\beta} \frac{\|\beta\|_{0,\infty,\Omega}}{L} = \frac{b_0}{c_\beta}. \end{aligned}$$

Hypothesis (H3) is always verified in the advection-dominated regime (it says nothing more than $\varrho \in L^\infty(\Omega)$). Instead, when the advection field is negligible, it forbids the problem to shift from reaction-dominated to diffusion-dominated within a single element. Note that, since we are interested in the case where the diffusion coefficient ε is very small, what we refer to as diffusion-dominated problem (that is, when both reaction and advection are also very small) has little practical interest.

Again let \mathcal{T}_h be a shape-regular family of decompositions of Ω into triangles T , such that each (open) boundary edge belongs either to Γ_D , or to Γ_N^+ or to Γ_N^- (in other words, we avoid edges that belong to two different types of boundaries). We denote by h_T the diameter of T , and we set $h = \max_{T \in \mathcal{T}_h} h_T$. Since we look for a solution of (2.1) a priori discontinuous, we need to recall the definition of typical tools such as *averages* and *jumps* on the edges for scalar- and vector-valued functions. Let T_1 and T_2 be two neighboring elements, let \mathbf{n}^1 and \mathbf{n}^2 be their outward normal unit vectors, and let φ^i and $\boldsymbol{\tau}^i$ be the restrictions of φ and $\boldsymbol{\tau}$ to T_i ($i = 1, 2$), respectively. Following [2] we set

$$(2.5) \quad \{\varphi\} = \frac{1}{2}(\varphi^1 + \varphi^2), \quad \llbracket \varphi \rrbracket = \varphi^1 \mathbf{n}^1 + \varphi^2 \mathbf{n}^2 \quad \text{on } e \in \mathcal{E}_h^\circ,$$

$$(2.6) \quad \{\boldsymbol{\tau}\} = \frac{1}{2}(\boldsymbol{\tau}^1 + \boldsymbol{\tau}^2), \quad \llbracket \boldsymbol{\tau} \rrbracket = \boldsymbol{\tau}^1 \cdot \mathbf{n}^1 + \boldsymbol{\tau}^2 \cdot \mathbf{n}^2 \quad \text{on } e \in \mathcal{E}_h^\circ,$$

where \mathcal{E}_h° is the set of interior edges e . For $e \in \mathcal{E}_h^\partial$, the set of boundary edges, we set

$$(2.7) \quad \llbracket \varphi \rrbracket = \varphi \mathbf{n}, \quad \{\varphi\} = \varphi, \quad \{\boldsymbol{\tau}\} = \boldsymbol{\tau}.$$

For future purposes we also introduce a weighted average, for both scalar- and vector-valued functions, as follows. With each internal edge e , shared by elements T_1 and T_2 , we associate two real nonnegative numbers α^1 and α^2 , with $\alpha^1 + \alpha^2 = 1$, and we define

$$(2.8) \quad \{\boldsymbol{\tau}\}_\alpha = \alpha^1 \boldsymbol{\tau}^1 + \alpha^2 \boldsymbol{\tau}^2 \quad \text{on internal edges.}$$

As shown, for instance, in [8] for a pure hyperbolic problem, a proper choice of α^1 and α^2 will introduce a stabilizing effect of upwind type into the scheme. We note that the

arithmetic average is obtained for $\alpha^1 = \alpha^2 = 1/2$, while the classical upwind flux is obtained when $\alpha^i = (\text{sign}(\boldsymbol{\beta} \cdot \mathbf{n}^i) + 1)/2$ for $i = 1, 2$ (where, as usual, $\text{sign}(x) = x/|x|$ for $x \neq 0$ and $\text{sign}(0) = 0$). Indeed, for vectors the following relation holds:

$$(2.9) \quad \{\boldsymbol{\tau}\}_\alpha \cdot \mathbf{n}_e = \left(\{\boldsymbol{\tau}\} + \frac{[\![\alpha]\!] }{2} [\![\boldsymbol{\tau}]\!] \right) \cdot \mathbf{n}_e,$$

whenever \mathbf{n}_e is orthogonal to e . Thus, if, for instance, T_1 is the upwind triangle, i.e., $\boldsymbol{\beta} \cdot \mathbf{n}^1 > 0$, then $\alpha = (1, 0)$ and

$$(2.10) \quad \begin{aligned} \{\boldsymbol{\tau}\}_\alpha \cdot \mathbf{n}^1 &= \left(\{\boldsymbol{\tau}\} + \frac{\mathbf{n}^1}{2} [\![\boldsymbol{\tau}]\!] \right) \cdot \mathbf{n}^1 = \boldsymbol{\tau}^1 \cdot \mathbf{n}^1 =: \{\boldsymbol{\tau}\}_{upw} \cdot \mathbf{n}^1, \\ \{\boldsymbol{\tau}\}_{1-\alpha} \cdot \mathbf{n}^2 &= \left(\{\boldsymbol{\tau}\} + \frac{\mathbf{n}^2}{2} [\![\boldsymbol{\tau}]\!] \right) \cdot \mathbf{n}^2 = \boldsymbol{\tau}^2 \cdot \mathbf{n}^2 =: \{\boldsymbol{\tau}\}_{dw} \cdot \mathbf{n}^2, \end{aligned}$$

while for scalar functions we obviously have

$$\{v\}_\alpha = v^1 =: \{v\}_{upw}, \quad \{v\}_{1-\alpha} = v^2 =: \{v\}_{dw}.$$

Taking $\alpha^i = 1/2 + t \text{sign}(\boldsymbol{\beta} \cdot \mathbf{n}^i)$ ($i = 1, 2$) will allow us, choosing t with $0 < t_0 \leq t \leq 1/2$ on each edge, to tune up the quantity of upwind.

We shall make extensive use of the identity [2, formula (3.3)]

$$(2.11) \quad \sum_{T \in \mathcal{T}_h} \int_{\partial T} \boldsymbol{\tau} \cdot \mathbf{n} \varphi = \sum_{e \in \mathcal{E}_h} \int_e \{\boldsymbol{\tau}\} \cdot [\![\varphi]\!] + \sum_{e \in \mathcal{E}_h^o} \int_e [\![\boldsymbol{\tau}]\!] \{\varphi\},$$

of the trace inequality [1], [2]

$$(2.12) \quad \|w\|_{0,e}^2 \leq C_t^2 (|e|^{-1} \|w\|_{0,T}^2 + |e| \|w\|_{1,T}^2), \quad e \subset \partial T, \quad w \in H^1(T),$$

with C_t a constant depending only on the minimum angle of T , and $|e| = \text{length of the edge } e$, and finally of the DG–Poincaré inequality [5]

$$(2.13) \quad \|v\|_{0,\Omega} \leq L C_P \left(|v|_{1,h}^2 + \sum_{e \notin \Gamma_N} \frac{1}{|e|} \|[\![v]\!]\|_{0,e}^2 \right)^{1/2},$$

where C_P is a positive constant depending on the minimum angle of \mathcal{T}_h , and $|\cdot|_{1,h}$ denotes the broken H^1 -seminorm. With the previous definitions, problem (2.1) is equivalent to

$$(2.14) \quad \begin{cases} \text{div} \boldsymbol{\sigma}(u) + \gamma u &= f & \text{in each } T \in \mathcal{T}_h, \\ [\![\boldsymbol{\sigma}(u)]\!] &= 0 & \text{on each } e \in \mathcal{E}_h^o, \\ [\![u]\!] &= 0 & \text{on each } e \in \mathcal{E}_h^o, \\ u &= g_D & \text{on each } e \in \Gamma_D, \\ (\boldsymbol{\beta} u \chi_{\Gamma_N^-} - \varepsilon \nabla u) \cdot \mathbf{n} &= g_N & \text{on each } e \in \Gamma_N. \end{cases}$$

Following the approach of [6], we shall introduce a variational formulation of (2.14) in which each of the equations above has the same relevance and is therefore treated in the same fashion. To do so, we introduce the space

$$V(\mathcal{T}_h) := \{v \in L^2(\Omega) \text{ such that } v|_T \in H^s(T) \ \forall T \in \mathcal{T}_h, \quad s > 3/2\},$$

and we assume that we have five operators $\mathcal{B}_0, \mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_1^D, \mathcal{B}_2^N$ from $V(\mathcal{T}_h)$ to $L^2(\Omega), \mathbf{L}^2(\mathcal{E}_h^\circ), L^2(\mathcal{E}_h^\circ), L^2(\Gamma_D), L^2(\Gamma_N)$, respectively. Then we consider the problem

$$(2.15) \quad \begin{cases} \text{Find } u \in V(\mathcal{T}_h) \text{ such that } \forall v \in V(\mathcal{T}_h) \\ \int_{\Omega} (\operatorname{div}_h \boldsymbol{\sigma}(u) + \gamma u - f) \mathcal{B}_0 v + \sum_{e \in \mathcal{E}_h^\circ} \int_e \llbracket u \rrbracket \cdot \mathcal{B}_1 v + \sum_{e \in \mathcal{E}_h^\circ} \int_e \llbracket \boldsymbol{\sigma}(u) \rrbracket \mathcal{B}_2 v \\ + \sum_{e \in \Gamma_D} \int_e (u - g_D) \mathcal{B}_1^D v + \sum_{e \in \Gamma_N} \int_e ((\boldsymbol{\beta} u \chi_{\Gamma_N^-} - \varepsilon \nabla u) \cdot \mathbf{n} - g_N) \mathcal{B}_2^N v = 0, \end{cases}$$

where div_h denotes the divergence element by element.

Different choices of the \mathcal{B} 's operators will give rise to different formulations. Since the solution of the original problem (2.1) is always a solution of (2.15), if we ensure uniqueness of the solution of (2.15), such a solution will coincide with the solution of the original problem. Sufficient conditions on the operators \mathcal{B} to guarantee uniqueness of the solution of (2.15) are given in [6, Theorem 1]. In the next section we shall present some choices of the operators verifying the hypotheses of the cited theorem.

3. Variational formulations. We will present four examples of different choices for the operators in (2.15). Two of them reproduce known formulations, while the other two will give rise to new methods.

Example 1. We set

$$(3.1) \quad \begin{aligned} \mathcal{B}_0 v|_T &= v \quad \forall T \in \mathcal{T}_h, & \mathcal{B}_1 v|_e &= c_e \frac{\varepsilon}{|e|} \llbracket v \rrbracket + \frac{\mathbf{n}^+}{2} \llbracket \boldsymbol{\beta} v \rrbracket \quad \forall e \in \mathcal{E}_h^\circ, \\ \mathcal{B}_2 v|_e &= -\{v\} \quad \forall e \in \mathcal{E}_h^\circ, \\ \mathcal{B}_1^D v|_e &= c_e \frac{\varepsilon}{|e|} \llbracket v \rrbracket \cdot \mathbf{n} - \boldsymbol{\beta} \cdot \mathbf{n} v \quad \forall e \in \Gamma_D^-, & \mathcal{B}_2^N v|_e &= -v \quad \forall e \in \Gamma_N^-. \end{aligned}$$

In (3.1) \mathbf{n}^+ is the normal to e such that $\boldsymbol{\beta} \cdot \mathbf{n}^+ \geq 0$, and c_e is a positive constant such that (see [2])

$$(3.2) \quad c_e \geq \eta_0 > 0 \quad \forall e \in \mathcal{E}_h.$$

We shall see that the definition of the operators on Γ^+ can be made arbitrary, without compromising the stability or consistency properties of the resulting methods. We can choose, for instance,

$$\mathcal{B}_1^D v = c_e \frac{\varepsilon}{|e|} v \quad \text{on } e \in \Gamma_D^+, \quad \mathcal{B}_2^N v = -v \quad \text{on } \Gamma_N^+.$$

With these choices, and setting

$$S_e = c_e \frac{\varepsilon}{|e|},$$

problem (2.15) reads

$$(3.3) \quad \begin{aligned} 0 &= \sum_{T \in \mathcal{T}_h} \int_T (\operatorname{div} \boldsymbol{\sigma}(u) + \gamma u - f) v + \sum_{e \in \mathcal{E}_h^\circ} \int_e \llbracket u \rrbracket \cdot \left(S_e \llbracket v \rrbracket + \frac{\mathbf{n}^+}{2} \llbracket \boldsymbol{\beta} v \rrbracket \right) \\ &- \sum_{e \in \mathcal{E}_h^\circ} \int_e \llbracket \boldsymbol{\sigma}(u) \rrbracket \{v\} + \sum_{e \in \Gamma_D^-} \int_e (u - g_D) \cdot (S_e \llbracket v \rrbracket - \boldsymbol{\beta} v) \cdot \mathbf{n} \\ &+ \sum_{e \in \Gamma_D^+} S_e \int_e (u - g_D) v - \sum_{e \in \Gamma_N} \int_e ((\boldsymbol{\beta} u \chi_{\Gamma_N^-} - \varepsilon \nabla_h u) \cdot \mathbf{n} - g_N) v. \end{aligned}$$

Using the identity (2.11) we have

$$\int_{\Omega} \operatorname{div}_h \boldsymbol{\sigma}(u) v = - \int_{\Omega} \boldsymbol{\sigma}(u) \cdot \nabla_h v + \sum_{e \in \mathcal{E}_h^\circ} \int_e \llbracket \boldsymbol{\sigma}(u) \rrbracket \{v\} + \sum_{e \in \mathcal{E}_h} \int_e \{\boldsymbol{\sigma}(u)\} \cdot \llbracket v \rrbracket.$$

Substituting in (3.3), and observing that the continuity of $\boldsymbol{\beta}$ and (2.10) implies

$$\begin{aligned} \sum_{e \in \mathcal{E}_h^\circ} \int_e \left(\{\boldsymbol{\beta}u\} \cdot \llbracket v \rrbracket + \llbracket u \rrbracket \cdot \frac{\mathbf{n}^+}{2} \llbracket \boldsymbol{\beta}v \rrbracket \right) &= \sum_{e \in \mathcal{E}_h^\circ} \int_e \left(\{\boldsymbol{\beta}u\} + \frac{\mathbf{n}^+}{2} \llbracket \boldsymbol{\beta}u \rrbracket \right) \cdot \llbracket v \rrbracket \\ &= \sum_{e \in \mathcal{E}_h^\circ} \int_e \{\boldsymbol{\beta}u\}_{upw} \cdot \mathbf{n}^+(v^+ - v^-) = \sum_{e \in \mathcal{E}_h^\circ} \int_e \{\boldsymbol{\beta}u\}_{upw} \cdot \llbracket v \rrbracket, \end{aligned}$$

we obtain the following formulation:

$$(3.4) \quad \left\{ \begin{array}{l} \text{Find } u \in V(\mathcal{T}_h) \text{ such that } \forall v \in V(\mathcal{T}_h) \\ \int_{\Omega} (\gamma uv - \boldsymbol{\sigma}(u) \cdot \nabla_h v) + \sum_{e \notin \Gamma_N} S_e \int_e \llbracket u \rrbracket \cdot \llbracket v \rrbracket + \sum_{e \in \mathcal{E}_h^\circ} \int_e \{\boldsymbol{\beta}u\}_{upw} \cdot \llbracket v \rrbracket \\ - \sum_{e \notin \Gamma_N} \int_e \{\varepsilon \nabla_h u\} \cdot \llbracket v \rrbracket + \int_{\Gamma^+} \boldsymbol{\beta} \cdot \mathbf{n} uv \\ = \sum_{T \in \mathcal{T}_h} \int_T f v + \sum_{e \in \Gamma_D} S_e \int_e g_D v - \sum_{e \in \Gamma_D^-} \int_e \boldsymbol{\beta} \cdot \mathbf{n} g_D v - \sum_{e \in \Gamma_N} \int_e g_N v. \end{array} \right.$$

We observe that for the diffusive part this method gives the so-called incomplete interior penalty Galerkin (IIPG) method proposed and analyzed in [36], while the advective part is upwinded through the operator \mathcal{B}_1 .

Example 2. We set

$$(3.5) \quad \begin{aligned} \mathcal{B}_0 v|_T &= v \quad \forall T \in \mathcal{T}_h, \\ \mathcal{B}_1 v|_e &= c_e \frac{\varepsilon}{|e|} \llbracket v \rrbracket + \{\varepsilon \nabla_h v\} + \frac{\mathbf{n}^+}{2} \llbracket \boldsymbol{\beta}v \rrbracket \quad \forall e \in \mathcal{E}_h^\circ, \\ \mathcal{B}_2 v|_e &= -\{v\} \quad \forall e \in \mathcal{E}_h^\circ, \quad \mathcal{B}_2^N v|_e = -v \quad \forall e \in \Gamma_N, \\ \mathcal{B}_1^D v|_e &= c_e \frac{\varepsilon}{|e|} v + (\varepsilon \nabla_h v - \boldsymbol{\beta}v \chi_{\Gamma_D^-}) \cdot \mathbf{n} \quad \forall e \in \Gamma_D. \end{aligned}$$

These choices reproduce the method introduced in [23] for the case $\gamma = 0$ and different boundary conditions. Indeed, in [23] the flux was not assigned at the inflow, and the boundary conditions were, with our notation,

$$u = g_D \quad \text{on } \Gamma_D \equiv \Gamma \setminus \Gamma_N^+, \quad (-\varepsilon \nabla u) \cdot \mathbf{n} = g_N \quad \text{on } \Gamma_N^+, \quad \Gamma_N^- = \emptyset.$$

In (3.5) the diffusive part corresponds to the NIPG method of [32], and the advective part is upwinded through \mathcal{B}_1 . Substituting (3.5) in (2.15), and using (2.10) and the continuity of β , leads to the problem

$$(3.6) \quad \left\{ \begin{array}{l} \text{Find } u \in V(\mathcal{T}_h) \quad \text{such that} \quad \forall v \in V(\mathcal{T}_h) \\ \int_{\Omega} (\gamma uv - \sigma(u) \cdot \nabla_h v) + \sum_{e \notin \Gamma_N} S_e \int_e [[u]] \cdot [v] + \sum_{e \in \mathcal{E}_h^o} \int_e \{\beta u\}_{upw} \cdot [v] \\ - \sum_{e \notin \Gamma_N} \int_e (\{\varepsilon \nabla_h u\} \cdot [v] - [u] \cdot \{\varepsilon \nabla_h v\}) + \sum_{e \in \Gamma^+} \int_e \beta \cdot \mathbf{n} uv \\ = \sum_{T \in \mathcal{T}_h} \int_T f v + \sum_{e \in \Gamma_D} \int_e g_D (S_e v + (\varepsilon \nabla_h v - \beta v \chi_{\Gamma_D^-}) \cdot \mathbf{n}) - \sum_{e \in \Gamma_N} \int_e g_N v. \end{array} \right.$$

Example 3. We set

$$\begin{aligned} \mathcal{B}_0 v|_T &= v \quad \forall T \in \mathcal{T}_h, \\ \mathcal{B}_1 v|_e &= c_e \frac{\varepsilon}{h} [[v]] - \theta \{\varepsilon \nabla v\}_{upw} \quad \forall e \in \mathcal{E}_h^o, \\ \mathcal{B}_2 v|_e &= -\{v\}_{dw} \quad \forall e \in \mathcal{E}_h^o, \quad \mathcal{B}_2^N v|_e = -v \quad \forall e \in \Gamma_N, \\ \mathcal{B}_1^D v|_e &= c_e \frac{\varepsilon}{h} v - (\theta \varepsilon \nabla v + \beta v \chi_{\Gamma_D^-}) \cdot \mathbf{n} \quad \forall e \in \Gamma_D^-, \end{aligned}$$

where θ is a parameter that allows us to include various formulations for treating the diffusive part: symmetric for $\theta = 1$, skew-symmetric for $\theta = -1$, and neutral for $\theta = 0$. This choice of the operators corresponds to the method introduced in [24] and analyzed in [10]. By substituting in (2.15), integrating by parts, and rearranging terms we obtain the following scheme:

$$(3.7) \quad \left\{ \begin{array}{l} \text{Find } u \in V(\mathcal{T}_h) \text{ such that } \forall v \in V(\mathcal{T}_h) \\ \int_{\Omega} (\gamma uv - \sigma(u) \cdot \nabla_h v) + \sum_{e \notin \Gamma_N} S_e \int_e [[u]] \cdot [v] + \sum_{e \in \mathcal{E}_h^o} \int_e \{\beta u\}_{upw} \cdot [v] \\ - \sum_{e \in \mathcal{E}_h^o} \int_e (\{\varepsilon \nabla_h u\}_{upw} \cdot [v] + \theta [u] \cdot \{\varepsilon \nabla_h v\}_{upw}) + \sum_{e \in \Gamma^+} \int_e \beta \cdot \mathbf{n} uv \\ - \sum_{e \in \Gamma_D} \int_e (\varepsilon \nabla_h u \cdot \mathbf{n} v + \theta u \varepsilon \nabla_h v \cdot \mathbf{n}) \\ = \sum_{T \in \mathcal{T}_h} \int_T f v + \sum_{e \in \Gamma_D} \int_e g_D (S_e [v] - \theta \varepsilon \nabla_h v - \beta v \chi_{\Gamma_D^-}) \cdot \mathbf{n} - \sum_{e \in \Gamma_N} \int_e g_N v. \end{array} \right.$$

In (3.7) the whole flux $\sigma(u)$ is upwinded through the operator \mathcal{B}_2 , but the upwind effect for the advective part is exactly the same as in methods (3.4) and (3.6).

Example 4. Let $\{\cdot\}_\alpha$ be the weighted average defined in (2.8)–(2.9). We set

$$\begin{aligned} \mathcal{B}_0 v|_T &= v \quad \forall T \in \mathcal{T}_h, \\ \mathcal{B}_1 v|_e &= c_e \frac{\varepsilon}{|e|} \llbracket v \rrbracket + \theta(\{\sigma(v)\}_\alpha - \{\beta v\}) \quad \forall e \in \mathcal{E}_h^\circ, \\ \mathcal{B}_2 v|_e &= -\{v\}_{1-\alpha} \quad \forall e \in \mathcal{E}_h^\circ, \quad \mathcal{B}_2^N v|_e = -v \quad \forall e \in \Gamma_N, \\ \mathcal{B}_1^D v|_e &= c_e \frac{\varepsilon}{h} v - (\theta \varepsilon \nabla_h v + \beta v \chi_{\Gamma_D^-}) \cdot \mathbf{n} \quad \forall e \in \Gamma_D. \end{aligned}$$

Substituting in (2.15) yields

$$(3.8) \quad \left\{ \begin{aligned} &\text{Find } u \in V(\mathcal{T}_h) \quad \text{such that} \quad \forall v \in V(\mathcal{T}_h) \\ &\int_\Omega \gamma uv - \sigma(u) \cdot \nabla_h v + \sum_{e \notin \Gamma_N} S_e \int_e \llbracket u \rrbracket \cdot \llbracket v \rrbracket - \theta \sum_{e \in \mathcal{E}_h^\circ} \int_e \llbracket u \rrbracket \cdot \{\beta v\} \\ &+ \sum_{e \in \mathcal{E}_h^\circ} \int_e (\{\sigma(u)\}_\alpha \cdot \llbracket v \rrbracket + \theta \llbracket u \rrbracket \cdot \{\sigma(v)\}_\alpha) + \sum_{e \in \Gamma^+} \int_e \beta \cdot \mathbf{n} uv \\ &- \sum_{e \in \Gamma_D} \int_e (\varepsilon \nabla_h u \cdot \mathbf{n} v + \theta u \varepsilon \nabla_h v \cdot \mathbf{n}) \\ &= \sum_{T \in \mathcal{T}_h} \int_T f v + \sum_{e \in \Gamma_D} g_D (S_e \llbracket v \rrbracket - \theta \varepsilon \nabla_h v - \beta v \chi_{\Gamma_D^-}) \cdot \mathbf{n} - \sum_{e \in \Gamma_N} \int_e g_N v. \end{aligned} \right.$$

In (3.8) θ again is a parameter that allows us to include different treatments of the diffusive part: symmetric for $\theta = 1$ SIPG(α) (see [35], [22]), nonsymmetric for $\theta = -1$, and neutral for $\theta = 0$. However, as we shall see in Remark 4.2, the case $\theta = -1$ gives rise to a formulation which is stable in a norm too weak, with a consequent loss of accuracy in the error estimates. Thus, it will not be further considered. The upwind is achieved in (3.8) through both operators \mathcal{B}_1 and \mathcal{B}_2 . Moreover, the use of the weighted average (2.8) should allow us to tune the amount of upwind on each edge. As a consequence, the formulation enjoys the nice feature of adapting easily from the advection-dominated to the diffusion-dominated regime.

All the above formulations share the common form

$$\left\{ \begin{aligned} &\text{Find } u \in V(\mathcal{T}_h) \quad \text{such that} \\ &a_h(u, v) = L(v) \quad \forall v \in V(\mathcal{T}_h). \end{aligned} \right.$$

Remark 3.1. In all cases, for obtaining the corresponding SUPG-stabilized DG formulations, one need only change the definition of the operator \mathcal{B}_0 into $\mathcal{B}_0 v = v + \mathbf{c}_T \beta \cdot \nabla v$ on each $T \in \mathcal{T}_h$, \mathbf{c}_T being a constant varying elementwise and depending on h_T and the coefficients of the problem $\beta, \varepsilon, \gamma$ (see [28], [25], and [24]).

4. Approximation. With any integer $k \geq 1$ we associate the finite element space of discontinuous piecewise polynomial functions

$$V_h^k = \{v \in L^2(\Omega) : v|_T \in \mathbf{P}^k(T) \quad \forall T \in \mathcal{T}_h\},$$

where, as usual, $P^k(T)$ is the space of polynomials of degree at most k on T . Replacing $V(\mathcal{T}_h)$ by V_h^k , we get the discrete problems, all sharing the form

$$(4.1) \quad \begin{cases} \text{Find } u_h \in V_h^k \text{ such that} \\ a_h(u_h, v_h) = L(v_h) \quad \forall v_h \in V_h^k. \end{cases}$$

Consistency. Consistency holds by construction in all the cases, so that

$$(4.2) \quad a_h(u - u_h, v_h) = 0 \quad \forall v_h \in V_h^k.$$

Stability. We shall prove stability in the norm

$$(4.3) \quad \|v\|^2 = \|v\|_d^2 + \|v\|_{rc}^2,$$

with

$$\begin{aligned} \|v\|_d^2 &:= \varepsilon|v|_{1,h}^2 + \varepsilon\|v\|_j^2 := \varepsilon|v|_{1,h}^2 + \sum_{e \notin \Gamma_N} \frac{\varepsilon}{|e|} \| [v] \|_{0,e}^2, \\ \|v\|_{rc}^2 &:= \|(\bar{\varrho} + b_0)^{1/2} v\|_{0,\Omega}^2 + \sum_{e \in \mathcal{E}_h} \| |\boldsymbol{\beta} \cdot \mathbf{n}|^{1/2} [v] \|_{0,e}^2, \end{aligned}$$

where $b_0 = \|\boldsymbol{\beta}\|_{0,\infty}/L$ is defined in (H1), and $\bar{\varrho}$ is the piecewise constant function defined as

$$(4.4) \quad \bar{\varrho}(x)|_T = \bar{\varrho}|_T, \quad \bar{\varrho}|_T = \min_{x \in T} \varrho(x) \quad \forall T \in \mathcal{T}_h.$$

Analogously, it will be useful to write the bilinear forms as

$$(4.5) \quad a_h(u, v) = a_h^d(u, v) + a_h^{rc}(u, v).$$

For simplicity, we start by considering the method (3.4), which corresponds to the “minimal choice” for the operators. Then we have

$$(4.6) \quad a_h^d(u, v) = \int_{\Omega} \varepsilon \nabla_h u \cdot \nabla_h v + \sum_{e \notin \Gamma_N} \int_e (S_e [u] - \{\varepsilon \nabla_h u\}) \cdot [v],$$

$$(4.7) \quad a_h^{rc}(u, v) = \int_{\Omega} (\gamma uv - u \boldsymbol{\beta} \cdot \nabla_h v) + \sum_{e \in \mathcal{E}_h^o} \int_e \{\boldsymbol{\beta} u\}_{upw} \cdot [v] + \int_{\Gamma^+} \boldsymbol{\beta} \cdot \mathbf{n} uv.$$

We note that, using (2.12) and arguing as in [2], we can easily see that there exists a (geometric) constant C_g , depending only on the degree of the polynomials and on the minimum angle of the decomposition such that

$$(4.8) \quad \sum_{e \notin \Gamma_N} \int_e \left| \{\varepsilon \nabla_h u\} [v] \right| \leq C_g \varepsilon |u|_{1,h} \|v\|_j \quad \forall u \in V_h^k, \forall v \in V(\mathcal{T}_h).$$

This implies that there exists a constant $C_d > 0$ such that

$$(4.9) \quad a_h^d(u, v) \leq C_d \|u\|_d \|v\|_d, \quad u \in V_h^k, v \in V(\mathcal{T}_h),$$

and, for η_0 in (3.2) verifying

$$(4.10) \quad \eta_0 > C_g^2/4,$$

there exists a positive constant α_d such that

$$(4.11) \quad a_h^d(v, v) \geq \alpha_d \|v\|_d^2, \quad v \in V_h^k.$$

We also note that, in general, one would rather require, say,

$$(4.12) \quad \eta_0 > \max\{C_g^2, 1\}$$

in order to have a quantifiable constant like $\alpha_d = 1/2$. In any case, the diffusive part alone would easily verify stability in all the methods. However, the technique of taking $v = u$, which is possibly the easiest way of proving stability, will not be sufficient when the reactive-advective part is also present, as it does not provide control on the L^2 -norm when advection dominates. Indeed, in all the cases we would have only

$$a_h^{rc}(v, v) \geq \|\bar{\varrho}^{1/2} v\|_{0,\Omega}^2 + \sum_{e \in \mathcal{E}_h} \|\beta \cdot \mathbf{n}\|^{1/2} \|v\|_{0,e}^2, \quad v \in V_h^k.$$

We will then prove stability in the norm (4.3) through an inf-sup condition. For that, following [28], we introduce the “weighting function” $\chi = \exp(-\eta)$, with η defined in (H1). The assumptions on η imply the existence of three positive constants $\chi_1^*, \chi_2^*, \chi_3^*$ such that

$$(4.13) \quad \chi_1^* \leq \chi \leq \chi_2^*, \quad |\nabla \chi| \leq \chi_3^*.$$

Our weighting function will be slightly different. Indeed, we shall take

$$(4.14) \quad \varphi = \chi + \kappa,$$

where κ is a constant such that

$$(4.15) \quad \chi_1^* + \kappa > 6 C_P L \chi_3^*, \quad \chi_1^* + \kappa > (\chi_2^* + \kappa)/2,$$

and C_P is the Poincaré constant appearing in (2.13).

The next lemma is a generalization to the case of variable β of that given in [26] for pure hyperbolic problems. See also [28] for the equivalent result for the SUPG-stabilized method and [34] for the conforming residual-free bubbles method. We point out, however, that here, thanks to the choice (4.14), we were able to remove the condition “ ε sufficiently small.”

LEMMA 4.1. *Let $a_h(\cdot, \cdot)$ be defined in (4.5)–(4.7), with*

$$(4.16) \quad \eta_0 > \max\{9C_g^2/4, 1\}.$$

Then, for every κ satisfying (4.15), the corresponding φ defined in (4.14) verifies

$$(4.17) \quad a_h^d(v_h, \varphi v_h) \geq \frac{\chi_1^* + \kappa}{6} \|v_h\|_d^2,$$

$$(4.18) \quad a_h^{rc}(v_h, \varphi v_h) \geq \frac{\chi_1^*}{2} \|v_h\|_{rc}^2,$$

$$(4.19) \quad \|\varphi v_h\| \leq \frac{\sqrt{145}}{6} (\chi_1^* + \kappa) \|v_h\|.$$

Proof. To simplify the notation we shall write

$$\alpha_1 = \chi_1^* + \kappa, \quad \alpha_2 = \chi_2^* + \kappa, \quad \alpha_3 \equiv \chi_3^*$$

so that

$$(4.20) \quad \alpha_1 \leq \varphi \leq \alpha_2, \quad |\nabla \varphi| \leq \alpha_3,$$

$$(4.21) \quad \text{(i) } \alpha_1 > 6 C_P L \alpha_3, \quad \text{(ii) } 2\alpha_1 > \alpha_2.$$

Conditions (4.8) and (4.20) give

$$\begin{aligned} a_h^d(v_h, \varphi v_h) &= \int_{\Omega} \varepsilon |\nabla_h v_h|^2 \varphi + \sum_{e \notin \Gamma_N} \int_e (S_e \llbracket v_h \rrbracket - \{\varepsilon \nabla_h v_h\}) \cdot \llbracket v \rrbracket \varphi + \int_{\Omega} \varepsilon \nabla_h v_h \cdot \nabla \varphi v_h \\ &\geq \varepsilon \left(\alpha_1 (\|v_h\|_{1,h}^2 + \eta_0 \|v_h\|_j^2) - \alpha_2 C_g |v_h|_{1,h} \|v_h\|_j - \alpha_3 |v_h|_{1,h} \|v_h\|_{0,\Omega} \right). \end{aligned}$$

This, using (4.21(ii)) and (4.16), then $\eta_0 \geq 1$ and (2.13), and finally (4.21(i)), gives easily

$$\begin{aligned} a_h^d(v_h, \varphi v_h) &\geq \varepsilon \left(\frac{\alpha_1}{3} (\|v_h\|_{1,h}^2 + \eta_0 \|v_h\|_j^2) - \alpha_3 |v_h|_{1,h} \|v_h\|_{0,\Omega} \right) \\ &\geq \varepsilon \frac{\alpha_1}{3} \left(\|v_h\|_{1,h}^2 + \|v_h\|_j^2 \right) - \alpha_3 C_P L \|v_h\|_d^2 \geq \frac{\alpha_1}{6} \|v_h\|_d^2, \end{aligned}$$

that is, (4.17). As regards the reactive-convective part, we observe that, after integration by parts, using (2.11) and the continuity of β and φ we get

$$\begin{aligned} - \int_{\Omega} \beta \cdot \nabla_h(\varphi v_h) v_h &= - \int_{\Omega} (\beta \cdot \nabla \varphi) v_h^2 - \frac{1}{2} \int_{\Omega} \beta \cdot \nabla_h(v_h^2) \varphi \\ (4.22) \quad &= - \frac{1}{2} \int_{\Omega} (\beta \cdot \nabla \varphi) v_h^2 + \frac{1}{2} \int_{\Omega} (\operatorname{div} \beta) \varphi v_h^2 - \frac{1}{2} \sum_{e \in \mathcal{E}_h} \int_e \{\beta \varphi\} \llbracket v_h^2 \rrbracket. \end{aligned}$$

Next, the continuity of β and φ easily imply that

$$\sum_{e \in \mathcal{E}_h^o} \int_e \{\beta v_h\} \cdot \llbracket \varphi v_h \rrbracket = \frac{1}{2} \sum_{e \in \mathcal{E}_h^o} \int_e \{\beta \varphi\} \cdot \llbracket v_h^2 \rrbracket.$$

From this and (2.10) we then have

$$(4.23) \quad \sum_{e \in \mathcal{E}_h^o} \int_e \{\beta v_h\}_{upw} \llbracket \varphi v_h \rrbracket = \frac{1}{2} \sum_{e \in \mathcal{E}_h^o} \int_e \{\beta \varphi\} \cdot \llbracket v_h^2 \rrbracket + \sum_{e \in \mathcal{E}_h^o} \int_e \frac{\beta \cdot \mathbf{n}^+}{2} \varphi \llbracket v_h \rrbracket^2.$$

By noting that (H1) and (4.20) imply

$$-\beta \cdot \nabla \varphi = (\beta \cdot \nabla \eta) \chi \geq 2b_0 \chi \geq 2b_0 \chi_1^*,$$

from (4.22)–(4.23), using (4.20), (2.2), and (4.4), we obtain

$$\begin{aligned} a_h^{rc}(v_h, \varphi v_h) &= \int_{\Omega} \left[\gamma + \frac{1}{2} (\operatorname{div} \beta) \right] \varphi v_h^2 - \frac{1}{2} \int_{\Omega} (\beta \cdot \nabla \varphi) v_h^2 \\ &\quad + \sum_{e \in \mathcal{E}_h^o} \int_e \frac{\beta \cdot \mathbf{n}^+}{2} \varphi \llbracket v_h \rrbracket^2 - \frac{1}{2} \int_{\Gamma^-} \beta \cdot \mathbf{n} \varphi v_h^2 + \frac{1}{2} \int_{\Gamma^+} \beta \cdot \mathbf{n} \varphi v_h^2 \\ &\geq \chi_1^* (\bar{\nu} + b_0)^{1/2} v_h \|_{0,\Omega}^2 + \frac{\alpha_1}{2} \sum_{e \in \mathcal{E}_h} \| |\beta \cdot \mathbf{n}|^{1/2} \llbracket v_h \rrbracket \|_{0,e}^2 \geq \frac{\chi_1^*}{2} \|v_h\|_{rc}^2, \end{aligned}$$

that is, (4.18). On the other hand, (4.19) again is an easy consequence of (2.13) and (4.20)–(4.21). \square

Remark 4.1. We point out that condition (4.16) has been taken in order to simplify the computation and to provide an easily quantifiable constant in (4.17) (very much in the spirit of (4.12) compared with the less demanding (4.10)). Looking at the proof, however, we see that we could stick to (4.10) (changing the conditions on κ in (4.15) in order to have α_2/α_1 as close to 1 as necessary). Hence, in some sense, the difficulty of finding “how big should η_0 be in practice” has not been worsened by the above trick.

Remark 4.2. Concerning the other three methods (3.6), (3.7), and (3.8), they exhibit essentially the same terms, with the only exception for the method (3.8), where the advective part contains

$$\sum_{e \in \mathcal{E}_h^\circ} \int_e ((\theta + 1)\{\beta v_h\}_\alpha - \theta\{\beta v_h\}) \llbracket \varphi v_h \rrbracket =: I_1,$$

instead of the left-hand term in (4.23). Using the definition (2.9) of the weighted average we obtain, instead of (4.23),

$$I_1 = \frac{1}{2} \sum_{e \in \mathcal{E}_h^\circ} \int_e \{\beta \varphi\} \cdot \llbracket v_h^2 \rrbracket + (\theta + 1) \sum_{e \in \mathcal{E}_h^\circ} \int_e \frac{\beta \cdot \llbracket \alpha \rrbracket}{2} \varphi \llbracket v_h \rrbracket^2,$$

where $\beta \cdot \llbracket \alpha \rrbracket = (2\alpha^+ - 1)\beta \cdot \mathbf{n}^+ > 0$ since α^+ , the weight associated with the upwind triangle, is $> 1/2$. Hence, (4.18) holds also for method (3.8) (possibly with a different constant) if $\theta > -1$. As already said, choosing $\theta = -1$ in (3.8) produces undesirable cancellations which lead to having stability in a norm too weak to ensure control on the advective part. Namely, we have

$$a_h(v_h, \varphi v_h) \geq C \left(\|(\bar{\nu} + b_0)^{1/2} v_h\|_{0,\Omega}^2 + \|v_h\|_d^2 + \sum_{e \in \Gamma} \|\beta \cdot \mathbf{n}\|^{1/2} \llbracket v_h \rrbracket_{0,e}^2 \right).$$

Suboptimal error estimates ($O(h^k)$) in this norm can be obtained, but the method is unstable in strongly advective regimes. Indeed, $\theta = -1$ gives rise to a method without any kind of upwind.

The following superapproximation results can be found in [29] and [37]. For convenience we briefly sketch the proof.

LEMMA 4.2. *Let $\varphi \in W^{k+1,\infty}(\Omega)$ be the function defined in (4.14). For $v_h \in V_h^k$, let $\widetilde{\varphi v_h}$ be the L^2 -projection of φv_h in V_h^k . Then*

$$(4.24) \quad \|\varphi v_h - \widetilde{\varphi v_h}\|_{0,\Omega} \leq C \frac{\|\chi\|_{k+1,\infty,\Omega}}{L} h \|v_h\|_{0,\Omega},$$

$$(4.25) \quad |\varphi v_h - \widetilde{\varphi v_h}|_{1,h} \leq C \frac{\|\chi\|_{k+1,\infty,\Omega}}{L} \|v_h\|_{0,\Omega},$$

$$(4.26) \quad \left(\sum_{e \in \mathcal{E}_h} \|\varphi v_h - \widetilde{\varphi v_h}\|_{0,e}^2 \right)^{1/2} \leq C \frac{\|\chi\|_{k+1,\infty,\Omega}}{L} h^{1/2} \|v_h\|_{0,\Omega},$$

where L is the diameter of Ω .

Proof. We shall deduce (4.24). Observe first that, since $\widetilde{\kappa v_h} \equiv \kappa v_h$,

$$\varphi v_h - \widetilde{\varphi v_h} \equiv \chi v_h - \widetilde{\chi v_h}.$$

Using classical interpolation results, the definition of the norm (1.2), the inverse inequality (see [12, Theorem 17.2, p. 135]), and $h < L$ we have

$$\begin{aligned} \|\varphi v_h - \widetilde{\varphi v_h}\|_{0,T} &\leq C h_T^{k+1} |\chi v_h|_{k+1,T} \leq C h_T^{k+1} \sum_{j=0}^k |\chi|_{k+1-j,\infty,T} |v_h|_{j,T} \\ &\leq C \|\chi\|_{k+1,\infty,\Omega} \sum_{j=0}^k \frac{h_T^{k+1} |v_h|_{j,T}}{L^{k+1-j}} \\ &\leq C C_{inv} \frac{\|\chi\|_{k+1,\infty,\Omega}}{L} \|v_h\|_{0,T} \sum_{j=0}^k \frac{h_T^{k+1-j}}{L^{k-j}} \\ &\leq C(k+1) h_T \frac{\|\chi\|_{k+1,\infty,\Omega}}{L} \|v_h\|_{0,T}. \end{aligned} \tag{4.27}$$

Hence, summing over all elements $T \in \mathcal{T}_h$ we reach (4.24). Exactly in the same way we prove (4.25), while (4.26) is a consequence of (4.24)–(4.25) via the trace inequality (2.12). \square

LEMMA 4.3. *In the hypotheses of Lemma 4.1, there exist two positive constants χ_4^*, χ_5^* such that, for any value of κ , the corresponding φ verifies*

$$a_h^d(v_h, \varphi v_h - \widetilde{\varphi v_h}) \leq \chi_4^* \|v_h\|_d^2 \quad \forall v_h \in V_h^k, \tag{4.28}$$

$$a_h^{rc}(v_h, \varphi v_h - \widetilde{\varphi v_h}) \leq \chi_5^* \left(\frac{h}{L}\right)^{1/2} \|v_h\|_{rc}^2 \quad \forall v_h \in V_h^k. \tag{4.29}$$

Proof. Using estimates (4.25)–(4.26) from Lemma 4.2, and then (2.13), we see that

$$\|\widetilde{\varphi v_h} - \varphi v_h\|_d \leq C \frac{\|\chi\|_{k+1,\infty,\Omega}}{L} \varepsilon^{1/2} \|v_h\|_{0,\Omega} \leq C C_P \|\chi\|_{k+1,\infty,\Omega} \|v_h\|_d.$$

Hence, from (4.9) we have

$$a_h^d(v_h, \widetilde{\varphi v_h} - \varphi v_h) \leq C_d \|v_h\|_d \|\widetilde{\varphi v_h} - \varphi v_h\|_d \leq C_d C C_P \|\chi\|_{k+1,\infty,\Omega} \|v_h\|_d^2,$$

that is, (4.28) with $\chi_4^* = C_d C C_P \|\chi\|_{k+1,\infty,\Omega}$. Before dealing with the reactive-convective part we observe that, if $P_h^0 \beta$ is the L^2 -projection of β onto constants, by definition of $\widetilde{\varphi v_h}$ it holds that

$$\int_{\Omega} P_h^0 \beta \cdot \nabla_h v_h (\varphi v_h - \widetilde{\varphi v_h}) = 0.$$

By integrating by parts and using (2.11) and (2.10) we then have

$$\begin{aligned} a_h^{rc}(v_h, \widetilde{\varphi v_h} - \varphi v_h) &= \int_{\Omega} [\gamma + \operatorname{div} \beta] v_h (\widetilde{\varphi v_h} - \varphi v_h) + \int_{\Omega} [\beta - P_h^0 \beta] \cdot \nabla_h v_h (\widetilde{\varphi v_h} - \varphi v_h) \\ &\quad - \sum_{e \notin \Gamma^+} \int_e \beta \cdot \llbracket v_h \rrbracket \{\widetilde{\varphi v_h} - \varphi v_h\} + \sum_{e \in \mathcal{E}_h^o} \int_e \frac{\beta \cdot \mathbf{n}^+}{2} \llbracket v_h \rrbracket \llbracket \widetilde{\varphi v_h} - \varphi v_h \rrbracket \\ &= I + II + III + IV. \end{aligned}$$

From (2.2), (H3), and (2.4) we have

$$\begin{aligned}
 I &= \int_{\Omega} \varrho v_h (\widetilde{\varphi v_h} - \varphi v_h) + \frac{1}{2} \int_{\Omega} \operatorname{div} \boldsymbol{\beta} v_h (\widetilde{\varphi v_h} - \varphi v_h) \\
 &\leq c_{\varrho} \|(\bar{\varrho} + b_0)^{1/2} v_h\|_{0,\Omega} \|(\bar{\varrho} + b_0)^{1/2} (\widetilde{\varphi v_h} - \varphi v_h)\|_{0,\Omega} + \frac{b_0}{2c_{\beta}} \|v_h\|_{0,\Omega} \|\widetilde{\varphi v_h} - \varphi v_h\|_{0,\Omega}.
 \end{aligned}$$

On the other hand, the definition (4.4) of $\bar{\varrho}$ and estimate (4.24) from Lemma 4.2 give

$$\begin{aligned}
 \|(\bar{\varrho} + b_0)^{1/2} (\widetilde{\varphi v_h} - \varphi v_h)\|_{0,\Omega}^2 &= \sum_{T \in \mathcal{T}_h} (\bar{\varrho}_T + b_0) \|(\widetilde{\varphi v_h} - \varphi v_h)\|_{0,T}^2 \\
 &\leq C \|\chi\|_{k+1,\infty,\Omega}^2 \left(\frac{h}{L}\right)^2 \sum_{T \in \mathcal{T}_h} (\bar{\varrho}_T + b_0) \|v_h\|_{0,T}^2 = C \|\chi\|_{k+1,\infty,\Omega}^2 \left(\frac{h}{L}\right)^2 \|(\bar{\varrho} + b_0)^{1/2} v_h\|_{0,\Omega}^2,
 \end{aligned}$$

so that

$$(4.30) \quad I \leq C \|\chi\|_{k+1,\infty,\Omega} \frac{h}{L} \|(\bar{\varrho} + b_0)^{1/2} v_h\|_{0,\Omega}^2.$$

Classical approximation results, (4.24), (2.4), and the inverse inequality give

$$(4.31) \quad II \leq Ch |\boldsymbol{\beta}|_{1,\infty,\Omega} |v_h|_{1,h} \frac{\|\chi\|_{k+1,\infty,\Omega} h}{L} \|v_h\|_{0,\Omega} \leq C \|\chi\|_{k+1,\infty,\Omega} \frac{h}{L} \frac{b_0}{c_{\beta}} \|v_h\|_{0,\Omega}^2.$$

Finally, from (4.26) we deduce

$$\begin{aligned}
 III + IV &\leq C \frac{h^{1/2}}{L} \|\boldsymbol{\beta}\|_{0,\infty,\Omega}^{1/2} \|v_h\|_{0,\Omega} \left(\sum_{e \in \mathcal{E}_h} \|\boldsymbol{\beta} \cdot \mathbf{n}\|^{1/2} \llbracket v_h \rrbracket \|_{0,e}^2 \right)^{1/2} \|\chi\|_{k+1,\infty,\Omega} \\
 (4.32) \quad &\leq C \left(\frac{h}{L}\right)^{1/2} \left(b_0 \|v_h\|_{0,\Omega}^2 + \sum_{e \in \mathcal{E}_h} \|\boldsymbol{\beta} \cdot \mathbf{n}\|^{1/2} \llbracket v_h \rrbracket \|_{0,e}^2 \right) \|\chi\|_{k+1,\infty,\Omega}.
 \end{aligned}$$

Collecting (4.30)–(4.32) we then get

$$a_h^{rc}(v_h, \widetilde{\varphi v_h} - \varphi v_h) \leq C \|\chi\|_{k+1,\infty,\Omega} \left(\frac{h}{L}\right)^{1/2} \|v_h\|_{rc}^2,$$

that is, (4.29) with $\chi_5^* = C \|\chi\|_{k+1,\infty,\Omega}$. \square

The next theorem provides the first stability result for the variational formulations presented in section 3.

THEOREM 4.4. *In the hypotheses of Lemma 4.1, there exists a positive constant $\alpha_S = \alpha_S(\boldsymbol{\beta}, \Omega)$, and $h_0 = h_0(\boldsymbol{\beta}) > 0$, such that, for $h < h_0$,*

$$\sup_{v_h \in V_h^k} \frac{a_h(u_h, v_h)}{\|v_h\|} \geq \alpha_S \|u_h\| \quad \forall u_h \in V_h^k.$$

Proof. For $u_h \in V_h^k$, let $v_h = \widetilde{\varphi u_h} \in V_h^k$ be the L^2 -projection of φu_h as defined previously. We shall prove that

$$(4.33) \quad \|v_h\| \leq c_1 \|u_h\|,$$

$$(4.34) \quad a_h(u_h, v_h) \geq c_2 \|u_h\|^2.$$

Adding and subtracting φu_h , from (4.17) we have first

$$\begin{aligned} a_h^d(u_h, \widetilde{\varphi u_h}) &= a_h^d(u_h, \widetilde{\varphi u_h} - \varphi u_h) + a_h^d(u_h, \varphi u_h) \\ &\geq a_h^d(u_h, \widetilde{\varphi u_h} - \varphi u_h) + \frac{\chi_1^* + \kappa}{6} \|u_h\|_d^2. \end{aligned}$$

Using estimate (4.28) we then have easily that for $\chi_1^* + \kappa$ bigger than $12\chi_4^*$ we find

$$a_h^d(u_h, \widetilde{\varphi u_h}) \geq \chi_4^* \|u_h\|_d^2.$$

In a similar way, from (4.29) and (4.18) one has, for $h < h_0$,

$$a_h^{rc}(u_h, \widetilde{\varphi u_h}) \geq C \|u_h\|_{rc}^2,$$

with C depending only on χ_1^*, χ_5^* . On the other hand, using (4.19) and Lemma 4.2, we have easily

$$\|\widetilde{\varphi u_h}\| \leq c_1 \|u_h\|,$$

that is, (4.33), with c_1 depending on χ_1^* and $\|\chi\|_{k+1, \Omega}$. □

Stability in a stronger norm. In a strongly advection-dominated regime it is desirable to have a control also on the streamline derivative; that is, it is necessary to have in (4.3) a term of SUPG type. We set

$$(4.35) \quad \|v\|_{DG}^2 := \|v\|^2 + \|v\|_S^2, \quad \|v\|_S^2 := \sum_{T \in \mathcal{T}_h} \frac{h_T}{\|\beta\|_{0, \infty, T}} \|P_h^k(\beta \cdot \nabla v)\|_{0, T}^2,$$

where P_h^k again is the L^2 -projection on V_h^k .

Remark 4.3. The presence of the projection in (4.35) is due to the fact that we assumed β to be a variable function, and hence $\beta \cdot \nabla_h u_h \notin V_h^k$. Clearly, whenever $\beta \cdot \nabla_h u_h \in V_h^k$, that is, if β is either constant (see [25], [20], [10]) or piecewise linear (see [23]), the projection can be removed.

Stability in the norm (4.35) can again be achieved through an inf-sup condition.

LEMMA 4.5. *There exists a constant $C_S > 0$, independent of $h, \varepsilon, \beta, \gamma$, such that*

$$(4.36) \quad \sup_{v_h \in V_h^k} \frac{a_h(u_h, v_h)}{\|v_h\|} \geq C_S (\|u_h\|_S - \|u_h\|) \quad \forall u_h \in V_h^k.$$

Proof. For $u_h \in V_h^k$, let $P_h^k(\beta \cdot \nabla_h u_h) \in V_h^k$ be the L^2 -projection on V_h^k of $\beta \cdot \nabla_h u_h$, for which the following estimates hold:

$$(4.37) \quad \forall T \in \mathcal{T}_h : |P_h^k(\beta \cdot \nabla_h u_h)|_{1, T} \leq C_{inv} h_T^{-1} \|P_h^k(\beta \cdot \nabla_h u_h)\|_{0, T},$$

and, for any edge e , shared by two elements T^+ and T^- ,

$$(4.38) \quad \begin{aligned} \|[P_h^k(\beta \cdot \nabla_h u_h)]\|_{0, e}^2 &\leq C |e|^{-1} \|P_h^k(\beta \cdot \nabla_h u_h)\|_{0, T^+ \cup T^-}^2, \\ \|[P_h^k(\beta \cdot \nabla_h u_h)]\|_{0, e}^2 &\leq C |e|^{-1} \|P_h^k(\beta \cdot \nabla_h u_h)\|_{0, T^+ \cup T^-}^2. \end{aligned}$$

Inequality (4.37) is the usual inverse inequality, while (4.38) is deduced through the trace inequality (2.12) and (4.37). We then set $v_h = \sum_{T \in \mathcal{T}_h} c_T (P_h^k(\beta \cdot \nabla_h u_h))|_T$, where

$$c_T = \begin{cases} \frac{h_T}{\|\beta\|_{0, \infty, T}} & \text{if advection dominates in } T, \\ 0 & \text{otherwise.} \end{cases}$$

We shall prove that

$$(4.39) \quad \|v_h\| \leq C_1 \|u_h\|_S,$$

$$(4.40) \quad a_h(u_h, v_h) \geq C_2 (\|u_h\|_S^2 - \|v_h\| \|u_h\|_S).$$

We prove first (4.39), having in mind that, if advection dominates, then

$$(4.41) \quad \varepsilon < h_T \|\boldsymbol{\beta}\|_{0,\infty,T}/2, \quad \|\gamma + \operatorname{div}\boldsymbol{\beta}\|_{0,\infty,T} < \|\boldsymbol{\beta}\|_{0,\infty,T}/h_T \quad \forall T \in \mathcal{T}_h.$$

From (4.37) and (4.41) we deduce

$$(4.42) \quad \varepsilon |v_h|_{1,h}^2 = \sum_{T \in \mathcal{T}_h} \varepsilon \left(\frac{h_T}{\|\boldsymbol{\beta}\|_{0,\infty,T}} \right)^2 |P_h^k(\boldsymbol{\beta} \cdot \nabla_h u_h)|_{1,T}^2 \leq C \|u_h\|_S^2.$$

Similarly, from (4.38) and (4.41) we have

$$(4.43) \quad \sum_{e \notin \Gamma_N} S_e \|v_h\|_{0,e}^2 = \sum_{e \notin \Gamma_N} c_e \frac{\varepsilon}{|e|} \|c_T P_h^k(\boldsymbol{\beta} \cdot \nabla_h u_h)\|_{0,e}^2 \leq C \|u_h\|_S^2$$

and

$$(4.44) \quad \sum_{e \in \mathcal{E}_h} \|\boldsymbol{\beta} \cdot \mathbf{n}\|^{1/2} \|v_h\|_{0,e}^2 = \sum_{e \in \mathcal{E}_h} \|\boldsymbol{\beta} \cdot \mathbf{n}\|^{1/2} \|c_T P_h^k(\boldsymbol{\beta} \cdot \nabla_h u_h)\|_{0,e}^2 \leq C \|u_h\|_S^2.$$

Since $\varrho = (\gamma + \operatorname{div}\boldsymbol{\beta}) - \frac{1}{2}\operatorname{div}\boldsymbol{\beta}$, in view of (4.41) and (2.4) we deduce

$$\|\varrho\|_{0,\infty,T} \leq \|\gamma + \operatorname{div}\boldsymbol{\beta}\|_{0,\infty,T} + \frac{1}{2} \|\operatorname{div}\boldsymbol{\beta}\|_{0,\infty,T} \leq \frac{\|\boldsymbol{\beta}\|_{0,\infty,T}}{h_T} + \frac{\|\boldsymbol{\beta}\|_{1,\infty,\Omega}}{2L}.$$

Hence, from (H2) and since $h_T \leq h < L$ we deduce

$$c_T \|\varrho\|_{0,\infty,T} \leq 1 + \frac{h_T}{2Lc_\beta} \leq 1 + \frac{1}{2c_\beta}.$$

Consequently,

$$(4.45) \quad \|\bar{\varrho}^{1/2} v_h\|_{0,\Omega}^2 \leq \sum_{T \in \mathcal{T}_h} \|\varrho\|_{0,\infty,T} c_T^2 \|P_h^k(\boldsymbol{\beta} \cdot \nabla_h u_h)\|_{0,T}^2 \leq C \|u_h\|_S^2.$$

Finally, always from (H2),

$$(4.46) \quad \|v_h\|_{0,\Omega}^2 = \sum_{T \in \mathcal{T}_h} \left(\frac{h_T}{\|\boldsymbol{\beta}\|_{0,\infty,T}} \right)^2 \|P_h^k(\boldsymbol{\beta} \cdot \nabla_h u_h)\|_{0,T}^2 \leq \frac{h}{c_\beta \|\boldsymbol{\beta}\|_{1,\infty,\Omega}} \|u_h\|_S^2,$$

and then, since $b_0 = \|\boldsymbol{\beta}\|_{0,\infty,\Omega}/L$, $\|\boldsymbol{\beta}\|_{0,\infty,\Omega} \leq \|\boldsymbol{\beta}\|_{1,\infty,\Omega}$, and $h < L$,

$$b_0 \|v_h\|_{0,\Omega}^2 \leq \frac{1}{c_\beta} \|u_h\|_S^2.$$

This and (4.45) can be written as

$$(4.47) \quad \|(\bar{\varrho} + b_0)^{1/2} v_h\|_{0,\Omega}^2 \leq C \|u_h\|_S^2,$$

and (4.39) is proved. We turn now to prove (4.40), again referring to formulation (3.4). For the diffusive part we have, via the Cauchy–Schwarz inequality and (4.42),

$$\int_{\Omega} \varepsilon \nabla_h u_h \nabla_h v_h \leq \varepsilon^{1/2} |u_h|_{1,h} \varepsilon^{1/2} |v_h|_{1,h} \leq C \varepsilon^{1/2} |u_h|_{1,h} \|u_h\|_S.$$

For the integrals on the edges, the Cauchy–Schwarz inequality and (4.43) give

$$\sum_{e \notin \Gamma_N} S_e \int_e \llbracket u_h \rrbracket \llbracket v_h \rrbracket \leq C \left(\sum_{e \notin \Gamma_N} S_e \|u_h\|_{0,e}^2 \right)^{1/2} \|u_h\|_S \leq C \|u_h\|_j \|u_h\|_S.$$

In an analogous way, the Cauchy–Schwarz inequality, trace inequality (2.12), the inverse inequality, and (4.43) give

$$\sum_{e \notin \Gamma_N} \int_e \{\varepsilon \nabla_h u_h\} \cdot \llbracket v_h \rrbracket \leq C \varepsilon^{1/2} |u_h|_{1,h} \|u_h\|_S,$$

so that

$$(4.48) \quad a_h^d(u_h, w_h) \leq C \|u_h\| \|w_h\|_S.$$

For the reactive and advective terms, integration by parts, formula (2.11), and the definition of the upwind average (2.10) give

$$\begin{aligned} a_h^r(u_h, v_h) &= \int_{\Omega} \varrho u_h v_h + \int_{\Omega} (\beta \cdot \nabla_h u_h) v_h + \frac{1}{2} \int_{\Omega} \operatorname{div} \beta u_h v_h \\ &\quad + \sum_{e \in \mathcal{E}_h^o} \int_e \frac{\beta \cdot \mathbf{n}^+}{2} \llbracket u_h \rrbracket \llbracket v_h \rrbracket - \sum_{e \notin \Gamma^+} \int_e \beta \cdot \llbracket u_h \rrbracket \{v_h\}. \end{aligned}$$

By definition of projection we have

$$(4.49) \quad \int_{\Omega} (\beta \cdot \nabla_h u_h) v_h = \int_{\Omega} P_h^k(\beta \cdot \nabla_h u_h) v_h = \|u_h\|_S^2,$$

and by the Cauchy–Schwarz inequality, (H2), and (4.47)

$$(4.50) \quad \int_{\Omega} \varrho u_h v_h \leq c_{\varrho} \|(\bar{\varrho} + b_0)^{1/2} u_h\|_{0,\Omega} \|(\bar{\varrho} + b_0)^{1/2} v_h\|_{0,\Omega} \leq C \|(\bar{\varrho} + b_0)^{1/2} u_h\|_{0,\Omega} \|u_h\|_S.$$

Using (2.4), (4.46), and (H2) we obtain

$$(4.51) \quad \begin{aligned} \int_{\Omega} \operatorname{div} \beta u_h v_h &\leq \left(\frac{\|\beta\|_{1,\infty,\Omega}}{L} \right) \|u_h\|_{0,\Omega} \left(\frac{h}{c_{\beta} \|\beta\|_{1,\infty,\Omega}} \right)^{1/2} \|u_h\|_S \\ &\leq \frac{b_0^{1/2}}{c_{\beta}} \left(\frac{h}{L} \right)^{1/2} \|u_h\|_{0,\Omega} \|u_h\|_S \leq C \|u_h\| \|u_h\|_S. \end{aligned}$$

Finally, from the Cauchy–Schwarz inequality and (4.44) we easily obtain

$$(4.52) \quad \sum_{e \in \mathcal{E}_h} \int_e \frac{\beta \cdot \mathbf{n}^+}{2} \llbracket u_h \rrbracket \llbracket v_h \rrbracket \leq C \left(\sum_{e \in \mathcal{E}_h^o} \|\beta \cdot \mathbf{n}\|^{1/2} \llbracket u_h \rrbracket_{0,e}^2 \right)^{1/2} \|u_h\|_S.$$

Collecting (4.49), (4.50), (4.51), and (4.52) we obtain

$$a_h^{rc}(u_h, v_h) \geq \|u_h\|_S^2 - C\|u_h\|\|u_h\|_S.$$

From (4.48) and the above estimate we then have

$$a_h(u_h, v_h) \geq \|u_h\|_S^2 - C\|u_h\|\|u_h\|_S,$$

which, together with (4.39), gives (4.36). \square

THEOREM 4.6. *There exists a constant $C_S = C_S(\beta, \Omega) > 0$, and $h_0 = h_0(\beta) > 0$, such that, for $h < h_0$,*

$$\sup_{v_h \in V_h^k} \frac{a_h(u_h, v_h)}{\|v_h\|} \geq C_S \|u_h\|_{DG} \quad \forall u_h \in V_h^k.$$

Proof. The result follows from Theorem 4.4 and Lemma 4.5. \square

We finally conclude by proving a result which provides stability in a norm of SUPG type but without the projection. However, this requires stronger regularity assumptions on β , dictated by the polynomial degree. More precisely, when using V_h^k , we can prove stability in the norm

$$(4.53) \quad \|u_h\|_{SS}^2 := \|u_h\|^2 + \|u_h\|_\beta^2, \quad \text{with } \|u_h\|_\beta^2 = \sum_{T \in \mathcal{T}_h} \frac{h_T}{\|\beta\|_{0,\infty,T}} \|\beta \cdot \nabla u_h\|_{0,T}^2,$$

only if $\beta \in W^{k,\infty}(\Omega)$. In other words, our initial assumption $\beta \in W^{1,\infty}(\Omega)$ guarantees stability in the norm (4.53) only for piecewise linear approximations.

THEOREM 4.7. *Let $\beta \in W^{k,\infty}(\Omega)$, $k \geq 1$ being the polynomial degree of V_h^k . Assume that*

$$(H2a) \quad \exists c_\beta > 0 \text{ such that } |\beta(x)| \geq c_\beta \|\beta\|_{k,\infty,\Omega} \quad \forall x \in \Omega.$$

Then, there exists a constant $C_{ss} = C_{ss}(\beta, \Omega) > 0$, and $h_0 = h_0(\beta) > 0$, such that, for $h < h_0$,

$$(4.54) \quad \sup_{v_h \in V_h^k} \frac{a_h(u_h, v_h)}{\|v_h\|} \geq C_{ss} \|u_h\|_{SS} \quad \forall u_h \in V_h^k.$$

Proof. The proof is accomplished by proceeding similarly as for Theorem 4.6, and we omit the details. Indeed, the only step that needs to be modified is (4.49), as all the others hold with the norm $\|\cdot\|_S$ replaced by $\|\cdot\|_\beta$, by simply using the stability of the L^2 -projection. By adding and subtracting $\sum_{T \in \mathcal{T}_h} c_T (\beta \cdot \nabla u_h)|_T$ we find

$$\begin{aligned} \int_\Omega (\beta \cdot \nabla_h u_h) v_h &= \|u_h\|_\beta^2 + \int_\Omega c_T (\beta \cdot \nabla_h u_h) [P_h^k(\beta \cdot \nabla_h u_h) - \beta \cdot \nabla_h u_h] \\ &\geq \|u_h\|_\beta^2 - \|u_h\|_\beta \left(\sum_{T \in \mathcal{T}_h} c_T \|P_h^k(\beta \cdot \nabla_h u_h) - \beta \cdot \nabla_h u_h\|_{0,T}^2 \right)^{1/2}. \end{aligned}$$

To estimate the second term, note that the regularity of β allows us to use the superapproximation property (4.27) (with β now playing the role of φ , and ∇u_h playing the role of v_h). This plus inverse inequality and (H2a) give

$$\begin{aligned} \|P_h^k(\beta \cdot \nabla_h u_h) - \beta \cdot \nabla_h u_h\|_{0,T} &\leq Ch_T^k |\beta \cdot \nabla_h u_h|_{k,T} \leq Ck \frac{\|\beta\|_{k,\infty,\Omega}}{L} h_T \|\nabla_h u_h\|_{0,T} \\ &\leq C \frac{\|\beta\|_{k,\infty,\Omega}}{L} \|u_h\|_{0,T} \leq C \frac{\|\beta\|_{0,\infty,T}}{c_\beta L} \|u_h\|_{0,T}. \end{aligned}$$

Since $h < L$ we then have

$$\sum_{T \in \mathcal{T}_h} c_T \|P_h^k(\beta \cdot \nabla u_h) - \beta \cdot \nabla u_h\|_{0,T}^2 \leq C \sum_{T \in \mathcal{T}_h} \left(\frac{h_T}{L}\right) \frac{\|\beta\|_{0,\infty,T}}{c_\beta^2 L} \|u_h\|_{0,T}^2 \leq \frac{C}{c_\beta^2} b_0 \|u_h\|_{0,\Omega}^2.$$

Thus,

$$\int_{\Omega} (\beta \cdot \nabla_h u_h) v_h \geq \|u_h\|_{\beta}^2 - C \|u_h\|_{\beta} \|u_h\|.$$

Then, the result (4.54) follows. \square

5. A priori error estimates. We next show a priori error estimates in the norms (4.3) and (4.35) for the methods presented. Let P_h^k be the L^2 -projection in V_h^k , for which the following local approximation property holds:

$$(5.1) \quad \|u - P_h^k u\|_{r,T} \leq Ch^{k+1-r} |u|_{k+1,T}, \quad r = 0, 1, 2, \quad T \in \mathcal{T}_h,$$

$$(5.2) \quad \|u - P_h^k u\|_{r,p,T} \leq Ch^{k+1-r} |u|_{k+1,p,T}, \quad 1 \leq p \leq \infty, \quad r = 0, 1, \quad T \in \mathcal{T}_h.$$

Moreover, from (5.1) and (2.12) we deduce that

$$(5.3) \quad \|u - P_h^k u\|_{0,e} \leq Ch_T^{k+1/2} |u|_{k+1,T} \quad \forall e \in \mathcal{E}_h.$$

THEOREM 5.1. *Let u be the solution of (2.1), and let u_h be the solution of the discrete problems (4.1). There exists a constant $C_0 = C_0(\Omega)$, depending on the domain Ω , the shape regularity of \mathcal{T}_h , and the polynomial degree (but independent of h and the coefficients of the problem), such that*

$$(5.4) \quad \|u - u_h\| \leq C_0(\Omega) h^k \left(\varepsilon^{1/2} + \|\beta\|_{0,\infty,\Omega}^{1/2} h^{1/2} + \|\varrho\|_{0,\infty,\Omega}^{1/2} h \right).$$

Proof. We define

$$\eta = u - P_h^k u, \quad \delta = u_h - P_h^k u.$$

From Theorem 4.4 and Galerkin orthogonality (4.2) we have

$$(5.5) \quad \alpha_S \|\delta\| \leq \frac{a_h(\delta, v_h)}{\|v_h\|} = \frac{a_h(\eta, v_h)}{\|v_h\|}.$$

The diffusive part is standard and can be easily estimated through the trace inequality (2.12), (5.1), and (5.3):

$$(5.6) \quad a_h^d(\eta, v_h) \leq Ch^k \varepsilon^{1/2} |u|_{k+1,\Omega} \|v_h\|_d.$$

Regarding the advective part, since $P_h^0 \beta \cdot \nabla_h v_h \in V_h^k$, by definition of projection

$$\int_{\Omega} P_h^0 \beta \cdot \nabla_h v_h \eta = 0.$$

From this, the Cauchy–Schwarz inequality, (5.2), the inverse inequality, (2.4), and (5.1) we have

$$\begin{aligned} \int_{\Omega} -(\beta \cdot \nabla_h v_h) \eta &= \int_{\Omega} (P_h^0 \beta - \beta) \cdot \nabla_h v_h \eta \leq Ch |\beta|_{1,\infty,\Omega} |v_h|_{1,h} \|\eta\|_{0,\Omega} \\ (5.7) \quad &\leq C \frac{\|\beta\|_{1,\infty,\Omega}}{L} \|v_h\|_{0,\Omega} \|\eta\|_{0,\Omega} \leq C \frac{b_0}{c_\beta} \|v_h\|_{0,\Omega} h^{k+1} |u|_{k+1,\Omega} \\ &\leq Ch^{k+1} b_0^{1/2} |u|_{k+1,\Omega} \|v_h\| = C \left(\frac{\|\beta\|_{0,\infty,\Omega}}{L} \right)^{1/2} h^{k+1} |u|_{k+1,\Omega} \|v_h\|. \end{aligned}$$

Using (5.3) we obtain

$$\begin{aligned}
 \sum_e \int_e \{\beta \eta\} \cdot [v_h] &\leq \|\beta\|_{0,\infty,\Omega}^{1/2} \sum_e \|\{\eta\}\|_{0,e} \|\beta \cdot \mathbf{n}\|^{1/2} [v_h]_{0,e} \\
 (5.8) \qquad \qquad \qquad &\leq C \|\beta\|_{0,\infty,\Omega}^{1/2} |u|_{k+1,\Omega} \|v_h\|,
 \end{aligned}$$

and, arguing similarly, we have

$$(5.9) \qquad \sum_e \int_e \frac{\beta \cdot \mathbf{n}^+}{2} [\eta] \cdot [v_h] \leq C \|\beta\|_{0,\infty,\Omega}^{1/2} |u|_{k+1,\Omega} \|v_h\|.$$

Finally, by writing $\gamma = \varrho - \text{div}\beta/2$, using (H3), (2.4), and (5.1) we obtain

$$\begin{aligned}
 \int_{\Omega} \gamma \eta v_h &\leq \|\varrho\|_{0,\infty,\Omega}^{1/2} \|\eta\|_{0,\Omega} c_p^{1/2} \|(\bar{\varrho} + b_0)^{1/2} v_h\|_{0,\Omega} + \frac{b_0^{1/2}}{c_\beta} \|\eta\|_{0,\Omega} b_0^{1/2} \|v_h\|_{0,\Omega} \\
 (5.10) \qquad \qquad &\leq C h^{k+1} \left(\|\varrho\|_{0,\infty,\Omega}^{1/2} + \left(\frac{\|\beta\|_{0,\infty,\Omega}}{L} \right)^{1/2} \right) |u|_{k+1,\Omega} \|v_h\|.
 \end{aligned}$$

Then collecting (5.6)–(5.10) and using $h/L < 1$ we obtain

$$a_h(\eta, v_h) \leq C h^k \left(\varepsilon^{1/2} + \|\beta\|_{0,\infty,\Omega}^{1/2} h^{1/2} + \|\varrho\|_{0,\infty,\Omega}^{1/2} h \right) |u|_{k+1,\Omega} \|v_h\|.$$

Hence, substituting this estimate into (5.5) gives

$$\|\delta\| \leq C(\Omega) h^k \left(\varepsilon^{1/2} + \|\beta\|_{0,\infty,\Omega}^{1/2} h^{1/2} + \|\varrho\|_{0,\infty,\Omega}^{1/2} h \right) |u|_{k+1,\Omega}.$$

The result (5.4) then follows by the triangle inequality. \square

THEOREM 5.2. *Let u be the solution of (2.1), and let u_h be the solution of the discrete problems (4.1). There exists a constant $C_1 = C_1(\Omega)$, depending on Ω , the shape regularity of \mathcal{T}_h , and the polynomial degree (but independent of $\gamma, \beta, \varepsilon$, and h), such that*

$$\|u - u_h\|_{DG} \leq C_1(\Omega) h^k \left(\varepsilon^{1/2} + \|\beta\|_{0,\infty,\Omega}^{1/2} h^{1/2} + \|\varrho\|_{0,\infty,\Omega}^{1/2} h \right) |u|_{k+1,\Omega}.$$

Proof. The proof follows the same steps of Theorem 5.1, using the stability result of Theorem 4.6. Hence we omit the details. \square

Remark 5.1. The same error estimates hold in the norm $\|\cdot\|_{SS}$ under the assumption $\beta \in W^{k,\infty}(\Omega)$.

Remark 5.2. Theorems 5.1 and 5.2 provide robust a priori error estimates, which are optimal in all regimes. More precisely, we have

$$\|u - u_h\|, \quad \|u - u_h\|_{DG} \simeq \begin{cases} O(h^{k+1/2}) & \text{if advection dominates,} \\ O(h^k) & \text{if diffusion dominates,} \\ O(h^{k+1}) & \text{if reaction dominates.} \end{cases}$$

COROLLARY 5.3. *As a direct consequence of our error analysis we have the following result:*

$$(5.11) \qquad \|u - u_h\|_{0,\Omega} \leq C_2 |u|_{k+1,\Omega} \begin{cases} h^{k+1/2} & \text{if advection dominates,} \\ h^k & \text{if diffusion dominates,} \\ h^{k+1} & \text{if reaction dominates,} \end{cases}$$

where C_2 depends on the domain Ω , the shape regularity of \mathcal{T}_h , the polynomial degree, and the coefficients of the problem γ, β , and ε (but is independent of h).

Remark 5.3. Estimate (5.11) is suboptimal in the diffusion-dominated regime, since it was simply obtained through (2.13) and (5.4). In the advection-dominated regime, although suboptimal of $1/2$, it is the best that one can expect for a regular triangulation without any further assumption on the construction-orientation of the mesh (see [30] for a counterexample in the pure hyperbolic case). Improved estimates in the case of β constant have been rigorously shown in [31] (for the pure hyperbolic case) under certain restrictions on the mesh and, more recently in [15], under milder assumptions on the grid. The techniques used in these papers rely strongly on the hypothesis that β is constant and do not seem to be easily extendable to the case of variable β . However, as we shall see in the next section, in many test cases optimal order of convergence in L^2 is attained for quite general mesh partitions.

6. Numerical experiments. In this section we compare on various test problems the methods analyzed in the previous sections. All the experiments were performed on the unit square $\Omega = (0, 1)^2$, using piecewise linear approximations on triangular grids, structured and unstructured. In all the graphics, method (3.4) is represented by $-\cdot-\star-\cdot-$; method (3.6) with $-\square-$; method (3.7) with $\cdots\circ\cdots$; and method (3.8) with $-x-$. For formulations (3.7) and (3.8) we report the results corresponding to $\theta = 1$, i.e., the symmetric treatment of the diffusive part. All the computations were done in MATLAB7, on a Powerbook 1.5 with 2GB of Ram memory.

Example 1: Case of smooth solution. We take $\beta = [1, 1]^T$ and $\gamma = 0$, and we vary the diffusion coefficient $\varepsilon = 1, 10^{-3}, 10^{-9}$. The forcing term f is chosen so that the analytical solution of (2.1), with Dirichlet boundary conditions, is given by $u(x, y) = \sin(2\pi x)\sin(2\pi y)$. Figures 6.1 and 6.2 represent, on a log-log scale, the convergence diagrams in the norm $\|\cdot\|_{DG}$ (and $\|\cdot\|$, resp.) versus the mesh size $h = \max_T h_T \approx 1/5, 1/9, 1/18, 1/36$. Clearly, the convergence rates are the same for all the methods, in agreement with the theory of section 5: first order accuracy when diffusion dominates and order $3/2$ in the convection-dominated regime. Figure 6.3 depicts in a log-log scale the convergence diagrams in the L^2 -norm with respect to the mesh size, $h = 2^{-2}, 2^{-3}, 2^{-4}, 2^{-5}, 2^{-6}$, on structured grids. Similar results, although not reported here, were obtained on unstructured grids. Observe that, due to smoothness of the solution, second order convergence is attained in all regimes for all the methods but method (3.7), which is only first order accurate when diffusion dominates. This is due to the fact that in the method (3.7) upwind is done on the whole flux. In method (3.8) the whole flux is also upwind, but the use of the weighted average (2.9) allows

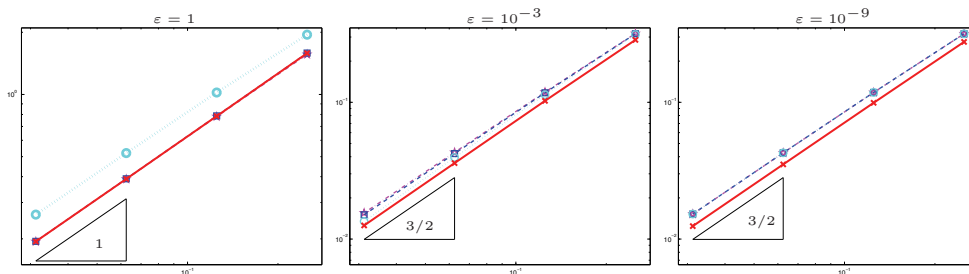


FIG. 6.1. Example 1. Convergence diagrams in the $\|\cdot\|_{DG}$ -norm. Unstructured grids.

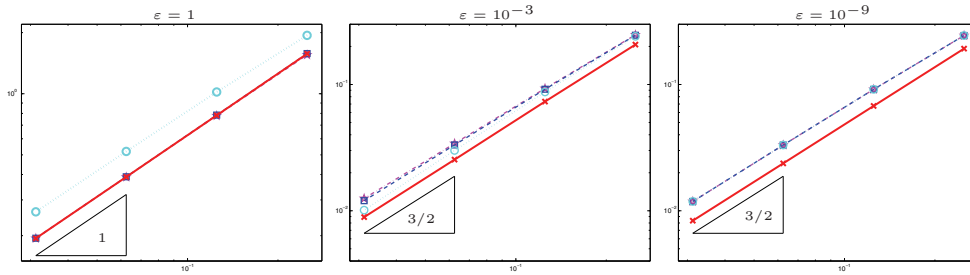


FIG. 6.2. Example 1. Convergence diagrams in the $\|\cdot\|$ -norm. Unstructured grids.

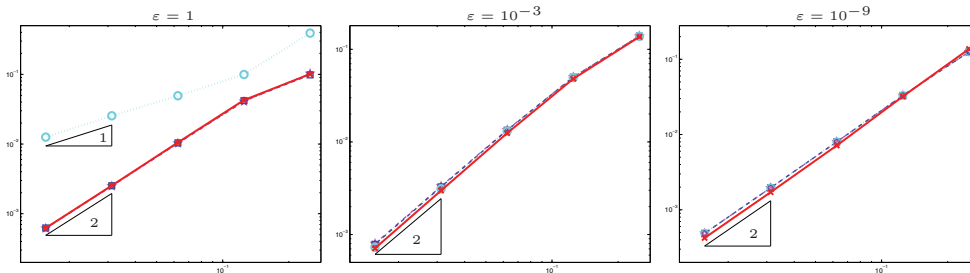


FIG. 6.3. Example 1. Convergence diagrams in the L^2 -norm. Structured grids.

us to tune the amount of upwind as a function of the data. It would be worth devising an automatic tuning. We did not yet, and found numerically the following “optimal values”: $(\alpha^1, \alpha^2) = (0.55, 0.45)$ for $\epsilon = 1$, $(\alpha^1, \alpha^2) = (0.64, 0.36)$ for $\epsilon = 10^{-3}$, and $(\alpha^1, \alpha^2) = (0.9, 0.1)$ for $\epsilon \leq 10^{-5}$.

Example 2: Rotating flow. This example is taken from [24]. The data are $\gamma = 0$, $\beta = [y - 1/2, 1/2 - x]^T$, and no external forces act on the system. The solution u is prescribed along the slit $1/2 \times [0, 1/2]$ as follows:

$$u(1/2, y) = \sin^2(2\pi y), \quad y \in [0, 1/2] .$$

In Figure 6.4, for $\epsilon = 10^{-9}$, we have represented the approximate solution obtained with the four methods on a structured triangular grid of 512 elements. As can be seen, all the methods perform similarly, and no significant differences can be appreciated. An important feature of all the methods is the absence of crosswind diffusion which occurs with stabilized conforming methods (see, e.g., [9], [7]). To better assess this feature of the methods, we have plotted in Figure 6.5 the profile of the approximate solutions at $y = 1/2$.

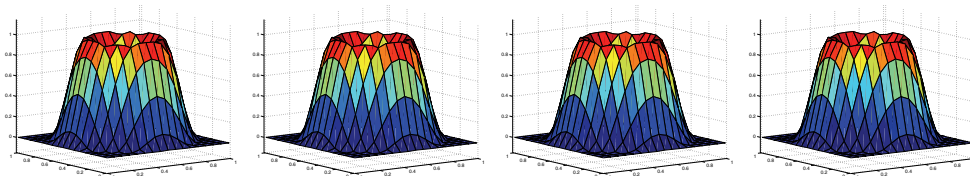


FIG. 6.4. Example 2. Approximate solutions for $\epsilon = 10^{-9}$ on structured grids. From left to right: methods (3.4), (3.6), (3.7), and (3.8).

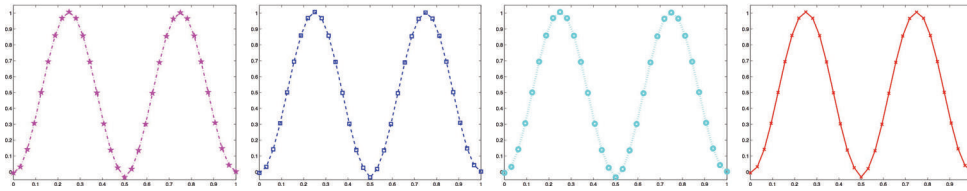


FIG. 6.5. *Example 2. Profile of the approximate solutions at $y = 1/2$; $\varepsilon = 1e - 07$. From left to right: methods (3.4), (3.6), (3.7), and (3.8).*

Example 3. Internal layers. The next example is devoted to assessing the performance of the methods in the presence of interior layers. We set $\gamma = 0$, $\beta = [1/2, \sqrt{3}/2]^T$, and Dirichlet boundary conditions as follows:

$$u = \begin{cases} 1 & \text{on } \{y = 0, 0 \leq x \leq 1\}, \\ 1 & \text{on } \{x = 0, y \leq 1/5\}, \\ 0 & \text{elsewhere.} \end{cases}$$

The diffusion coefficient is varied from $\varepsilon = 10^{-3}$ to the limit case $\varepsilon = 0$ (pure hyperbolic case). In Figure 6.6 we represent the approximate solutions obtained on structured grids of 512 triangles with all methods for $\varepsilon = 10^{-3}$. They all behave poorly in the intermediate regimes, as they produce wiggles close to the boundary. These oscillations disappear in the strongly advection-dominated regime (see Figure 6.7), and the internal layer is sharply captured, with very small overshooting/undershooting. This can be better observed in Figure 6.8, where we have represented the profiles of the solutions at $x = 0$. Similar results were observed for the profiles at $y = 0.5$. We notice that the boundary layers on the outflow are missed in all the methods. This is a known drawback of DG approximations: as soon as advection dominates they behave as if the problem were purely hyperbolic. See also the next example for a similar behavior.

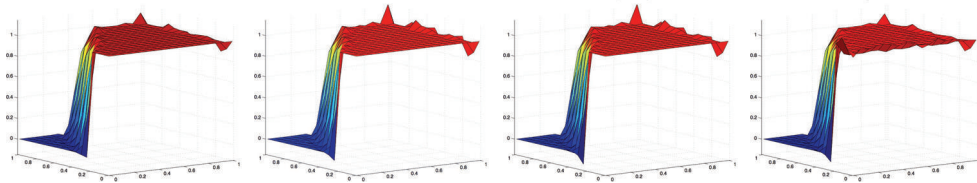


FIG. 6.6. *Example 3. Approximate solutions for $\varepsilon = 10^{-3}$ on unstructured grids. From left to right: methods (3.4), (3.6), (3.7), and (3.8).*

Example 4. Boundary layers. In this example we apply the methods to a boundary layer problem taken from [23]. The data are $\gamma = 0$ and $\beta = [1, 1]^T$, and we again vary the diffusion coefficient ε . The forcing term f is chosen so that the exact solution is given by

$$u(x, y) = x + y(1 - x) + \frac{e^{-1/\varepsilon} - e^{-(1-x)(1-y)/\varepsilon}}{1 - e^{-1/\varepsilon}}, \quad (x, y) \in \Omega.$$

This problem can be regarded as a multidimensional variant of the one-dimensional problem considered by Melenk and Schwab in [27]. Unlike the classical test case [38],

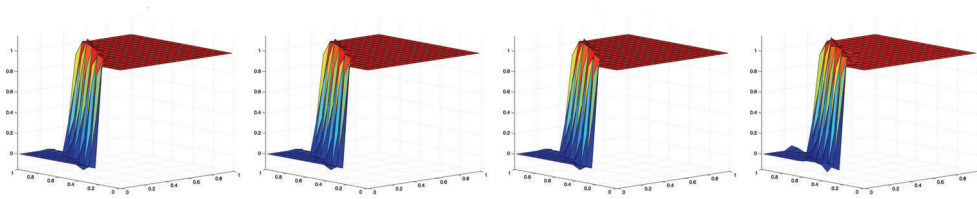


FIG. 6.7. Example 3. Approximate solutions for $\varepsilon = 10^{-9}$ on unstructured grids. From left to right: methods (3.4), (3.6), (3.7), and (3.8).

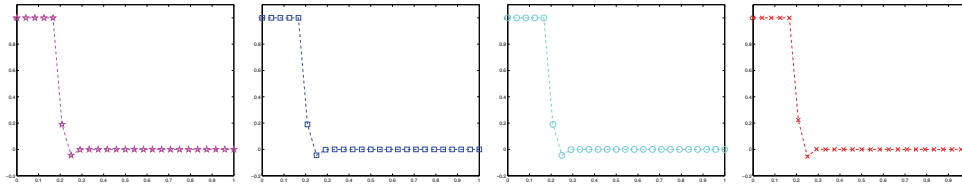


FIG. 6.8. Example 3. Profile of the approximate solutions at $x = 0$; $\varepsilon = 1e - 09$. From left to right: methods (3.4), (3.6), (3.7), and (3.8).

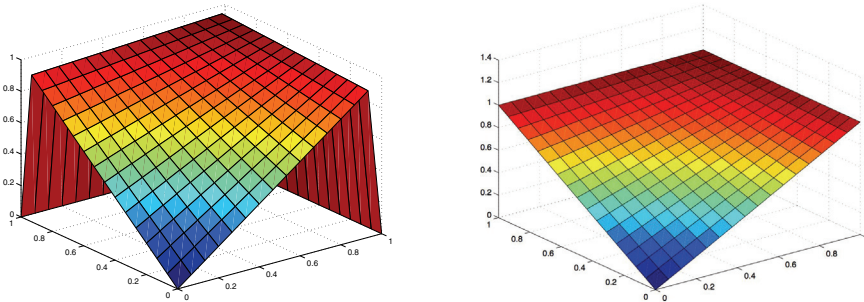


FIG. 6.9. Example 4. Exact solution (left), approximate solution with method (3.7) (right); $\varepsilon = 10^{-9}$.

u does not reduce, in the hyperbolic limit case, to a linear function in the interior of the domain, as shown in Figure 6.9(left), for $\varepsilon = 10^{-9}$. In Figure 6.9(right) only the solution obtained with the method (3.7) is represented, as all the methods do not exhibit visible differences in the strongly advective regime. Notice that, since boundary conditions are imposed in a weak way, the boundary layer is not captured by the DG approximations, although the solution is free of spurious oscillations. In Figure 6.10 we compare the methods for $\varepsilon = 10^{-3}$ and structured grids with $24 \times 24 \times 2$ triangles. Again, no substantial differences can be observed, except for small oscillations in the method (3.7) (third plot in the figure), probably due to the upwind treatment of the diffusive part of the flux. For this test case we chose not to plot convergence diagrams in the norms (4.3) or (4.35) since, due to the weak approximation of the boundary conditions, the main contribution to the error comes from the error in the boundary layer, which is $O(1)$, as can be seen in Figures 6.9 and 6.10. Figure 6.11 represents the convergence diagrams in the L^1 -norm for $\varepsilon = 10^{-3}$ and $h = 1/5, 1/9, 1/18, 1/36$. Note that as we would expect in this regime, and since we are measuring global errors, first order convergence is achieved. Although there are no great differences

between the methods, it seems that in this case method (3.4) gives the most accurate approximation. This can also be checked from Figure 6.10. Finally, Figure 6.12 shows the convergence diagrams in terms of $h = 1/5, 1/9, 1/18, 1/36$ on unstructured grids for $\varepsilon = 10^{-9}$ in the L^2 -norm (left), the $\|\cdot\|_d$ -norm in the interior of the domain (i.e., without the contribution of the boundary elements) (center), and in the norm $\|\cdot\|_S$ defined in (4.35) (right). Note that all the methods give optimal order of convergence in L^2 in the advection-dominated regime (see Remark 5.3).

Example 5. Compressible advection-diffusion problem. We conclude with a test where the advection field is not divergence-free. We set $\gamma = 0$ and $\beta = [yx^2 + 1, xy^2 + 1]^T$. The flow enters the computational domain Ω from two sides of Γ , namely $\{x = 0\}$ and $\{y = 0\}$. The forcing term is chosen as

$$f = \begin{cases} 0 & \text{on } 0 \leq x \leq 1/2, \quad 0 \leq y \leq 1/3, \\ -1 & \text{on } 1/2 < x \leq 1, \quad 1/3 < y \leq 1. \end{cases}$$

Nonhomogeneous Dirichlet boundary conditions were imposed on Γ^- :

$$u = \begin{cases} 2x & \text{on } 0 \leq x \leq 1, \quad y = 0, \\ 3y & \text{on } x = 0, \quad 0 \leq y \leq 1, \end{cases}$$

and homogeneous Neumann conditions on $\Gamma^+ = \{x = 1, 0 < y < 1\} \cup \{y = 1, 0 < x < 1\}$. Figure 6.13 shows a vector diagram of the advection field (left) and two different views of the approximate solution obtained with method (3.8) for $\varepsilon = 10^{-9}$ on a structured triangular mesh with $h = 1/16$. In Figure 6.14 we represent the approximate solutions obtained on structured grids of 512 triangles ($h = 1/16$) with

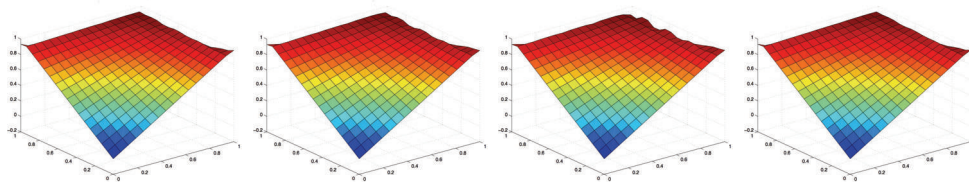


FIG. 6.10. Example 4. Approximate solutions for $\varepsilon = 10^{-3}$. From left to right: methods (3.4), (3.6), (3.7), and (3.8).

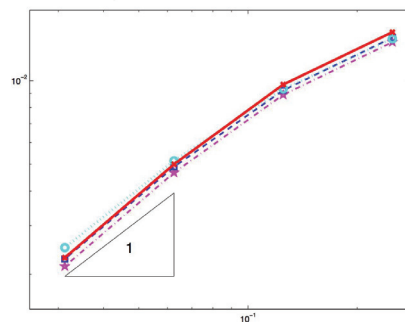


FIG. 6.11. Example 4. Convergence diagrams in the L^1 -norm; $\varepsilon = 10^{-3}$.

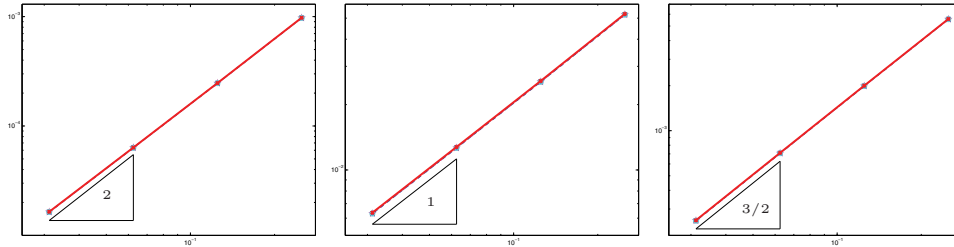


FIG. 6.12. Example 4. Convergence diagrams in the norms L^2 (left), interior $\|\cdot\|_d$ (center), and $\|\cdot\|_S$ (right); $\varepsilon = 10^{-9}$. Unstructured grids.

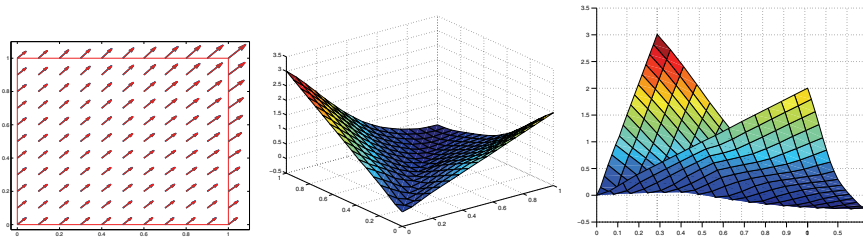


FIG. 6.13. Example 5. Left: vector diagram of advection field. Center and right: two views of the approximate solution obtained with method (3.8) for $\varepsilon = 10^{-9}$ on a structured mesh.

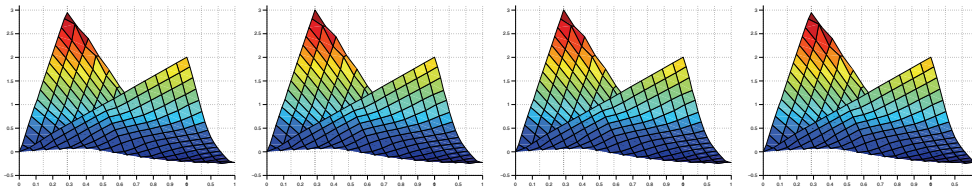


FIG. 6.14. Example 5. Approximate solutions for $\varepsilon = 10^{-3}$ on structured meshes. From left to right: methods (3.4), (3.6), (3.7), and (3.8).

all methods for $\varepsilon = 10^{-3}$. From the graphics we observe that all methods behave similarly and provide a good approximate solution also when $\text{div } \beta \neq 0$.

Remark 6.1. In general, it is neither easy to compare the performance of different DG methods, nor to design a relevant test. For advection-diffusion-reaction problems it is even more complicated, if one wants to take care of all the possible regimes and of the variety of stabilizations. From the tests that we performed so far it seems that all the methods presented in the paper behave similarly, at least in the strongly advection-dominated case. Some differences appear in the intermediate regimes but not enough to draw definite conclusions. From the computational point of view method (3.4) is simpler than the others. On the other hand, method (3.8) seems promising to adjust to varying regimes, provided a sound automatic tuning of the upwind could be found.

7. Conclusions. By using the weighted-residual approach of [6] we set a unified framework for deriving and analyzing various methods for advection-diffusion-reaction problems. The analysis carried out applies to the case of variable convection and reaction fields, and shows that optimal estimates in DG norms are achieved. In particular, we relaxed the usual coercivity condition (see assumption (2.2)), thus allowing for taking care of a variety of situations, if one wants to allow cases of a (comparatively)

very small diffusion. All the methods considered in this paper seem to have the same stability and accuracy properties, in all regimes. This is also confirmed numerically, though the method (3.8) seems to be more flexible in the intermediate regimes, thanks to the possibility of tuning the amount of upwind.

Appendix A. We briefly sketch how the function $\eta \in W^{k+1,\infty}(\Omega)$ in (H1) can be constructed. Arguing as in [16] we can guarantee that, for β satisfying (2.3),

$$(A.1) \quad \text{if } \beta \in [W^{1,\infty}(\Omega)]^2 \implies \exists \tilde{\eta} \in W^{1,\infty}(\Omega) \quad \text{s.t.} \quad \beta \cdot \nabla \tilde{\eta} \geq 2b_0 > 0 \quad \text{in } \Omega.$$

We next show how from this function η_0 the more regular η in (H1) can be constructed. Let $\{\mathcal{U}_\alpha^+\}_\alpha$ be a finite open covering of Ω such that each \mathcal{U}_α^+ enjoys the following property: there exists some $\varepsilon_1 > 0$ (to be chosen later) such that

$$(A.2) \quad \text{if } x, y \in \mathcal{U}_\alpha^+ \implies \|\beta(x) - \beta(y)\|_{0,\infty} < \varepsilon_1$$

and

$$(A.3) \quad \forall x, y \in \mathcal{U}_\alpha^+ \quad \beta(x) \cdot \nabla \tilde{\eta}(y) \geq b_0.$$

Inequality (A.3) is actually a consequence of (A.2) and (A.1). Indeed,

$$\beta(x) \cdot \nabla \tilde{\eta}(y) = \beta(y) \cdot \nabla \tilde{\eta}(y) + [\beta(x) - \beta(y)] \cdot \nabla \tilde{\eta}(y) \geq 2b_0 - \varepsilon_1 \|\nabla \tilde{\eta}\|_{0,\infty}.$$

Hence, by taking $\varepsilon_1 = b_0/\|\nabla \tilde{\eta}\|_{0,\infty}$ one can guarantee (A.3). Let $\mathcal{U}_\alpha^- \subset \mathcal{U}_\alpha^+$ be such that (A.2) holds with such choice of ε_1 (so that (A.3) is valid for all x and $y \in \mathcal{U}_\alpha^-$), and such that $\{\mathcal{U}_{\alpha'}^-\}_{\alpha'}$ is still an open covering of Ω . Next, on each $\mathcal{U}_{\alpha'}^-$ we mollify $\tilde{\eta}$ by convolution with some ρ_δ mollifier; $\eta_{\alpha'}^\delta = \tilde{\eta} * \rho_\delta$ in $\mathcal{U}_{\alpha'}^-$. Then, by taking a partition of unity $\{\phi_{\alpha'}\}_{\alpha'}$ associated with the covering $\{\mathcal{U}_{\alpha'}^-\}_{\alpha'}$ we can construct η as in (H1) by gluing the mollified $\eta_{\alpha'}^\delta$, that is, $\eta = \sum_{\alpha'} \eta_{\alpha'}^\delta \cdot \phi_{\alpha'}$. Thus, the existence of η sufficiently smooth satisfying (H1) is guaranteed.

REFERENCES

- [1] D. N. ARNOLD, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.
- [2] D. N. ARNOLD, F. BREZZI, B. COCKBURN, AND L. D. MARINI, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.
- [3] D. N. ARNOLD, F. BREZZI, R. FALK, AND L. D. MARINI, *Locking-free Reissner-Mindlin elements without reduced integration*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 3660–3671.
- [4] D. N. ARNOLD, F. BREZZI, AND L. D. MARINI, *A family of discontinuous Galerkin finite elements for the Reissner-Mindlin plate*, J. Sci. Comput., 22/23 (2005), pp. 25–45.
- [5] S. C. BRENNER, *Poincaré–Friedrichs inequalities for piecewise H^1 functions*, SIAM J. Numer. Anal., 41 (2003), pp. 306–324.
- [6] F. BREZZI, B. COCKBURN, L. D. MARINI, AND E. SÜLI, *Stabilization mechanisms in discontinuous Galerkin finite element methods*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 3293–3310.
- [7] F. BREZZI, L. D. MARINI, AND A. RUSSO, *On the choice of a stabilizing subgrid for convection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 127–148.
- [8] F. BREZZI, L. D. MARINI, AND E. SÜLI, *Discontinuous Galerkin methods for first-order hyperbolic problems*, Math. Models Methods Appl. Sci., 14 (2004), pp. 1893–1903.
- [9] A. N. BROOKS AND T. J. R. HUGHES, *Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 32 (1982), pp. 199–259.

- [10] A. BUFFA, T. J. R. HUGHES, AND G. SANGALLI, *Analysis of a multiscale discontinuous Galerkin method for convection-diffusion problems*, SIAM J. Numer. Anal., 44 (2006), pp. 1420–1440.
- [11] E. BURMAN AND P. ZUNINO, *A domain decomposition method based on weighted interior penalties for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 44 (2006), pp. 1612–1638.
- [12] P. G. CIARLET, *Basic error estimates for elliptic problems*, in Handbook of Numerical Analysis, Vol. II, Handb. Numer. Anal. II, North-Holland, Amsterdam, 1991, pp. 17–351.
- [13] B. COCKBURN, *Discontinuous Galerkin methods for convection-dominated problems*, in High-Order Methods for Computational Physics, Lect. Notes Comput. Sci. Eng. 9, Springer-Verlag, Berlin, 1999, pp. 69–224.
- [14] B. COCKBURN AND C. DAWSON, *Some extensions of the local discontinuous Galerkin method for convection-diffusion equations in multidimensions*, in The Mathematics of Finite Elements and Applications, X, MAFELAP 1999 (Uxbridge), Elsevier, Oxford, UK, 2000, pp. 225–238.
- [15] B. COCKBURN, B. DONG, AND J. GUZMÁN, *Optimal convergence of the original DG method for the transport-reaction equation on special meshes*, SIAM J. Numer. Anal., 46 (2008), pp. 1250–1265.
- [16] A. DEVINATZ, R. ELLIS, AND A. FRIEDMAN, *The asymptotic behavior of the first real eigenvalue of second order elliptic operators with a small parameter in the highest derivatives. II*, Indiana Univ. Math. J., 23 (1973–1974), pp. 991–1011.
- [17] A. ERN AND J.-L. GUERMOND, *Discontinuous Galerkin methods for Friedrichs’ systems. I. General theory*, SIAM J. Numer. Anal., 44 (2006), pp. 753–778.
- [18] A. ERN AND J.-L. GUERMOND, *Discontinuous Galerkin methods for Friedrichs’ systems. II. Second-order elliptic PDEs*, SIAM J. Numer. Anal., 44 (2006), pp. 2363–2388.
- [19] A. ERN AND J.-L. GUERMOND, *Discontinuous Galerkin methods for Friedrichs’ systems. Part III. Multifield theories with partial coercivity*, SIAM J. Numer. Anal., 46 (2008), pp. 776–804.
- [20] J. GOPALAKRISHNAN AND G. KANSCHAT, *A multilevel discontinuous Galerkin method*, Numer. Math., 95 (2003), pp. 527–550.
- [21] J. GUZMÁN, *Local analysis of discontinuous Galerkin methods applied to singularly perturbed problems*, J. Numer. Math., 14 (2006), pp. 41–56.
- [22] B. HEINRICH AND K. PIETSCH, *Nitsche type mortaring for some elliptic problem with corner singularities*, Computing, 68 (2002), pp. 217–238.
- [23] P. HOUSTON, C. SCHWAB, AND E. SÜLI, *Discontinuous hp-finite element methods for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 39 (2002), pp. 2133–2163.
- [24] T. J. R. HUGHES, G. SCOVAZZI, P. B. BOCHEV, AND A. BUFFA, *A multiscale discontinuous Galerkin method with the computational structure of a continuous Galerkin method*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 2761–2787.
- [25] C. JOHNSON, U. NÄVERT, AND J. PITKÄRANTA, *Finite element methods for linear hyperbolic problems*, Comput. Methods Appl. Mech. Engrg., 45 (1984), pp. 285–312.
- [26] C. JOHNSON AND J. PITKÄRANTA, *An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation*, Math. Comp., 46 (1986), pp. 1–26.
- [27] J. M. MELENK AND C. SCHWAB, *An hp finite element method for convection-diffusion problems in one dimension*, IMA J. Numer. Anal., 19 (1999), pp. 425–453.
- [28] U. NÄVERT, *A Finite Element Method for Convection-Diffusion Problems*, Ph.D. thesis, Department of Computer Science, Chalmers University of Technology, Göteborg, Sweden, 1982.
- [29] J. NITSCHKE AND A. SCHATZ, *On local approximation properties of L_2 -projection on spline-subspaces*, Applicable Anal., 2 (1972), pp. 161–168.
- [30] T. E. PETERSON, *A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation*, SIAM J. Numer. Anal., 28 (1991), pp. 133–140.
- [31] G. R. RICHTER, *An optimal-order error estimate for the discontinuous Galerkin method*, Math. Comp., 50 (1988), pp. 75–88.
- [32] B. RIVIÈRE, M. F. WHEELER, AND V. GIRAULT, *Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems. I*, Comput. Geosci., 3 (1999), pp. 337–360 (2000).
- [33] H.-G. ROOS, M. STYNES, AND L. TOBISKA, *Numerical Methods for Singularly Perturbed Differential Equations: Convection-Diffusion and Flow Problems*, Springer Ser. Comput. Math. 24, Springer-Verlag, Berlin, 1996.
- [34] G. SANGALLI, *Global and local error analysis for the residual-free bubbles method applied to advection-dominated problems*, SIAM J. Numer. Anal., 38 (2000), pp. 1496–1522.

- [35] R. STENBERG, *Mortaring by a method of J. A. Nitsche*, in Computational Mechanics (Buenos Aires, 1998), CD-ROM file, Centro Internac. Métodos Numér. Ing., Barcelona, Spain, 1998.
- [36] S. SUN AND M. F. WHEELER, *Symmetric and nonsymmetric discontinuous Galerkin methods for reactive transport in porous media*, SIAM J. Numer. Anal., 43 (2005), pp. 195–219.
- [37] L. B. WAHLBIN, *Superconvergence in Galerkin Finite Element Methods*, Lecture Notes in Math. 1605, Springer-Verlag, Berlin, 1995.
- [38] H. ZARIN AND H.-G. ROOS, *Interior penalty discontinuous approximations of convection-diffusion problems with parabolic layers*, Numer. Math., 100 (2005), pp. 735–759.

ON MESH GEOMETRY AND STIFFNESS MATRIX CONDITIONING FOR GENERAL FINITE ELEMENT SPACES*

QIANG DU[†], DESHENG WANG[‡], AND LIYONG ZHU[§]

Abstract. The performance of finite element computation depends strongly on the quality of the geometric mesh and the efficiency of the numerical solution of the linear systems resulting from the discretization of partial differential equation (PDE) models. It is common knowledge that mesh geometry affects not only the approximation error of the finite element solution but also the spectral properties of the corresponding stiffness matrix. In this paper, for typical second-order elliptic problems, some refined relationships between the spectral condition number of the stiffness matrix and the mesh geometry are established for general finite element spaces defined on simplicial meshes. The derivation of such relations for general high-order elements is based on a new trace formula for the element stiffness matrix. It is shown that a few universal geometric quantities have the same dominant effect on the stiffness matrix conditioning for different finite element spaces. These results provide guidance to the studies of both linear algebraic solvers and the unstructured geometric meshing.

Key words. condition number, mesh quality, finite element method, unstructured mesh

AMS subject classifications. 65N30, 65F10

DOI. 10.1137/080718486

1. Introduction. The finite element solution of partial differential equations (PDEs) often involves mesh generation and optimization, the assembly of discrete algebraic systems using the finite element basis, and the solution of these systems by some algebraic solvers. Traditionally, the different components have often been studied separately, so as to maximize the independence between the various software components and to make the finite element method a versatile and popular methodology for many applications. In recent years, the finite element community has been paying increasing attention to an integrated adaptive solution strategy. It thus becomes important to understand the interplay between the various components in order to improve the overall performance of finite element simulations.

A major objective of the study we have undertaken recently is to explore the relations among the mesh geometry, the efficiency of the linear solver for the resulting finite element linear system of equations, and the interpolation (or discretization) error. While it has been common in the meshing community to examine the quality of mesh with respect to various geometric measures, there have also been a number of works relating mesh quality to interpolation or discretization errors; see, for instance, [5, 7, 8, 10, 12, 33, 37, 39] and the references cited therein. Connections between the performance of the algebraic solvers and general unstructured meshes have also been made, but with much less rigor and generality. Perhaps the most widely known

*Received by the editors March 14, 2008; accepted for publication (in revised form) October 23, 2008; published electronically March 13, 2009.

<http://www.siam.org/journals/sinum/47-2/71848.html>

[†]Department of Mathematics, Pennsylvania State University, University Park, PA 16802 (qdu@math.psu.edu). The work of this author is supported in part by NSF DMS-0712744.

[‡]Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore (desheng@ntu.edu.sg). The work of this author is supported in part by grants NTU start-up M58110011, ARC 29/07 T207B2202, and NRF 2007 IDM-IDM002-010.

[§]Laboratory of Computational Physics, Institute of Applied Physics and Computational Mathematics, Beijing, China (zhu.liyong@iapcm.ac.cn).

facts in this direction were based on the vast experiences in the application of finite element technology, such as the belief that poorly shaped elements can give rise to ill-conditioned matrices, which tend to slow down or even prevent the convergence of iterative solvers. Even with the increasing popularity of the unstructured simplicial meshing in finite element simulations, there were relatively few attempts at general discussions on the precise connections between the solver performances and the qualities of unstructured meshing. From among the notable works we recall [36], in which the effect of the unstructured irregular grids on the performance of algebraic solvers and preconditioners has been examined through numerical examples. In [6, 18], the trade-offs associated with the cost of mesh improvement in terms of solution efficiency have been analyzed numerically. In [37], comprehensive discussions have been made on mesh quality measures, and in particular, on how a *good* element for resolving the discretization error may at the same time be *good* for the efficient solution of the resulting algebraic systems. More recently in [15], a mesh and solver co-adaptation strategy has been studied in the context of finite element methods for anisotropic problems.

In a more general arena, but closely related to our objective, the exploration of the properties of the stiffness matrix resulting from the finite element discretizations in relation to the underlying geometric meshes has remained a continuing theme in the finite element literature for half a century. Precise and explicit descriptions of the relations between mesh geometry and the spectral condition numbers are naturally helpful to the understanding of the whole finite element solution process. Yet the current understanding of such relations remains largely incomplete despite a number of existing investigations [2, 21, 37, 38].

In this work, we are able to establish a precise relation between the mesh geometry and the spectral condition number of the stiffness matrix for some typical second-order elliptic equations discretized by general finite element methods based on unstructured simplicial meshes in any space dimension. An important conclusion following from our analysis is that the effect of the element geometry on the conditioning of the stiffness matrices for more general finite element methods is similar to that of the conforming linear Lagrange finite element. Consequently, a simplicial mesh that makes the stiffness matrices less ill-conditioned for the linear element tends to do the same for high-order elements as well. Results of such generality, to the best of our knowledge, have not been presented before in the literature. They bring new understanding to mesh generation and optimization and the solution of discrete algebraic systems.

Our analysis is based on the derivation of an explicit trace formula for the element stiffness matrix corresponding to the finite element approximation to the Laplace operator (presented in section 2). While requiring only routine calculations, the trace formula appears to be new and quite elegant. It helps us to derive, in section 3, more precise estimates on the extreme eigenvalues of element stiffness matrices for general finite element spaces in terms of the element and mesh geometries, using an earlier framework on the estimation of stiffness matrix conditioning in [20, 21]. Some known calculations in the literature on the linear Lagrange finite element are also presented there as comparisons. The new estimate not only makes some of the classical works (such as those in [19, 20, 21]) more precise but also makes some observations for special cases (such as those in [37]) more general. In addition, we specialize to various cases and consider the relevant extensions (in section 4). The theoretical analysis is also complemented by numerical experiments which serve as further validation.

2. Finite element approximation and a new trace formula. In this section, we first derive a new trace formula for the element stiffness matrix for the

Laplace operator using general finite element methods. We then recall briefly the abstract framework on the condition number estimate for general symmetric second-order elliptic equations given in [21] and the discussion on the linear Lagrange finite element given in [37]. These results form the basis of discussions on the condition number estimation for general high-order elements on general unstructured simplicial meshes.

2.1. Basic finite element terminology. Given an open bounded convex domain $\Omega \in R^d$ with a Lipschitz-continuous boundary, we consider the following general self-adjoint linear second-order elliptic boundary value problem:

$$(2.1) \quad \begin{cases} - \sum_{i,j=1}^d \frac{\partial}{\partial x_i} \left(a_{ij} \frac{\partial u}{\partial x_j} \right) + a_0 u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

where the coefficient matrix $\tilde{A} = (a_{ij})_{i,j=1}^d$ is symmetric positive definite everywhere in Ω and $a_0 \geq 0$ in Ω . Both \tilde{A} and a_0 are assumed to be smooth and uniformly bounded for simplicity. In addition, we let $f \in L^2(\Omega)$. The corresponding variational weak form is as follows: Find $u \in H_0^1(\Omega)$ such that

$$(2.2) \quad a_\Omega(u, v) = \int_\Omega \sum_{i,j=1}^d \left(a_{ij} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} \right) \, d\mathbf{x} + \int_\Omega a_0 uv \, d\mathbf{x} = \int_\Omega f v \, d\mathbf{x} \quad \forall v \in H_0^1(\Omega).$$

It is well known that the above weak variational form (2.2) has a unique solution in $H_0^1(\Omega)$ [11]. Let τ denote the finite element mesh (a triangulation, or equivalently, a simplicial mesh for much of our discussion). Appropriate finite element spaces with suitably chosen nodal basis functions $\{\phi_j\}_{j=1}^N$ may then be employed to discretize the continuous problem (2.2), resulting in algebraic systems associated with the finite element approximations. For any (simplicial) element $t \in \tau$, we assume that the nodal basis, when restricted to t , is given by a canonical transformation from a nodal basis defined on a reference simplex described by the barycentric coordinates

$$(2.3) \quad t_0 = \left\{ (b_1, b_2, \dots, b_{d+1}) \mid b_i \geq 0, \sum b_j = 1 \right\}.$$

Concerning the finite element space, we make an additional assumption that the nodal basis on t_0 is *invariant* with respect to the permutation of the vertices, a property that is satisfied by most of the finite element spaces.

Let K and M be the $N \times N$ stiffness and mass matrices, respectively, generated by the finite element methods, that is,

$$K = (k_{ij}), \quad k_{ij} = a_\Omega(\phi_i, \phi_j) \quad \text{and} \quad M = (m_{ij}), \quad m_{ij} = \int_\Omega \rho \phi_i \phi_j \, d\mathbf{x}.$$

Here, as in [20], a positive density function $\rho = \rho(x)$ is introduced into the mass matrix. While for much of the discussion we focus on the case when $\rho = 1$ is a constant, a nonuniform density can be very useful in dealing with highly nonuniform meshes. Without further complicating the discussion, we assume that ρ remains positive and smooth in the domain of interest.

Obviously, both K and M are symmetric, with M being positive definite and K being either positive or nonnegative definite. Denote the element matrices corresponding to K and M by K_t and M_t , respectively, for any (simplicial) element $t \in \tau$.

We use n to denote the dimension of K_t and M_t , which corresponds to the degree of freedom or the number of nodal basis functions for the element t .

The eigenvalues of K and M are denoted by $\{\lambda_i^K\}_{i=1}^N$ and $\{\lambda_i^M\}_{i=1}^N$, which are ordered by

$$\lambda_1^K \leq \lambda_2^K \leq \dots \leq \lambda_N^K, \quad \lambda_1^M \leq \lambda_2^M \leq \dots \leq \lambda_N^M .$$

In this notation, λ_1^K and λ_1^M are the minimal eigenvalues of K and M , and λ_N^K and λ_N^M are the maximal eigenvalues, respectively. Similarly, we use $\{\lambda_i^{K_t}\}_{i=1}^n$ and $\{\lambda_i^{M_t}\}_{i=1}^n$ to denote the eigenvalues of K_t and M_t , respectively, which are also ordered by

$$\lambda_1^{K_t} \leq \lambda_2^{K_t} \leq \dots \leq \lambda_n^{K_t}, \quad \lambda_1^{M_t} \leq \lambda_2^{M_t} \leq \dots \leq \lambda_n^{M_t} .$$

For the case of a conforming linear finite element, the nodal basis on the element t is simply given by the coordinates $\{b_1, b_2, \dots, b_{d+1}\}$. Let $\{z_j\}_1^{d+1}$ be the vertices of t with z_j having corresponding barycentric coordinates $b_j = 1$ and $b_i = 0$ for $i \neq j$. It is well known that for each i , $b_i = b_i(x)$ is a linear function of $x \in t$, representing the ratio of the volume formed by the simplex with vertices $x \cup \{b_j, j \neq i\}$ and the volume of t . Moreover,

$$x = \sum_j b_j z_j .$$

It is also trivial to see that ∇b_i gives the normal direction of the $(d - 1)$ -dimensional face A_i of t , opposite to the vertex z_i , and $|\nabla b_i|$ is the reciprocal of the height of the simplex t corresponding to the vertex z_i . Equivalently, we have [11]

$$(2.4) \quad |\nabla b_i| = \frac{|A_i|}{d|t|}$$

with $|t|$ denoting the volume of t and $|A_i|$ being the area of the face A_i for each $1 \leq i \leq d + 1$.

2.2. A trace formula for the element stiffness matrix. We now derive a new trace formula for the stiffness matrix associated with the Laplace operator discretized by general simplicial finite element spaces. We adopt the notation introduced in the previous subsection but specialize to the case of

$$(2.5) \quad a_\Omega(u, v) = \int_\Omega \nabla u \cdot \nabla v, dx$$

for any u, v in $H^1(\Omega)$. In this case, (2.2) corresponds to the Poisson equation with a homogeneous Dirichlet boundary condition if we take $u, v \in H_0^1(\Omega)$.

Given a simplex t , we use $\{L_i(\{b_j\})\}_{i=1}^n$ to denote a general form of the nodal basis functions on t , and the finite element approximation is given by functions whose restrictions on t are linear combinations of $\{L_i\}$.

Notice that it is assumed that the set of basis functions remains invariant under any permutation to vertices, and thus under any permutation of the barycentric coordinates.

Consider the element stiffness matrix K_t . Its (k, l) th entry is now given by

$$a_t(L_k, L_l) = \int_t \nabla L_k \cdot \nabla L_l dx .$$

In particular, we have the m th diagonal entry given by

$$\begin{aligned} a_t(L_m, L_m) &= \int_t \nabla L_m \cdot \nabla L_m dx \\ &= \sum_{i=1}^d \int_t \left(\sum_{j=1}^{d+1} \frac{\partial L_m}{\partial b_j} \frac{\partial b_j}{\partial x_i} \right)^2 dx \\ &= \sum_{i=1}^d \sum_{j=1}^{d+1} \sum_{k=1}^{d+1} \frac{\partial b_j}{\partial x_i} \frac{\partial b_k}{\partial x_i} \int_t \frac{\partial L_m}{\partial b_j} \frac{\partial L_m}{\partial b_k} dx . \end{aligned}$$

Here, we have used the fact that $\frac{\partial b_j}{\partial x_i}$ and $\frac{\partial b_k}{\partial x_i}$ are constants on t . Now, we sum over m to get the trace of K_t ,

$$\begin{aligned} \text{Tr}(K_t) &= \sum_{m=1}^n a_t(L_m, L_m) \\ &= \sum_{m=1}^n \sum_{i=1}^d \sum_{j=1}^{d+1} \sum_{k=1}^{d+1} \frac{\partial b_j}{\partial x_i} \frac{\partial b_k}{\partial x_i} \int_t \frac{\partial L_m}{\partial b_j} \frac{\partial L_m}{\partial b_k} dx \\ (2.6) \quad &= \sum_{i=1}^d \sum_{j=1}^{d+1} \sum_{k=1}^{d+1} \frac{\partial b_j}{\partial x_i} \frac{\partial b_k}{\partial x_i} \int_t \sum_{m=1}^n \frac{\partial L_m}{\partial b_j} \frac{\partial L_m}{\partial b_k} dx . \end{aligned}$$

By the invariance of the set of basis functions under the permutation of the barycentric coordinates, we see that there are two constants α_n^d and β_n^d such that

$$(2.7) \quad \int_t \sum_{m=1}^n \left(\frac{\partial L_m}{\partial b_j} \right)^2 dx = \alpha_n^d |t| \quad \forall j ,$$

$$(2.8) \quad \int_t \sum_{m=1}^n \frac{\partial L_m}{\partial b_j} \frac{\partial L_m}{\partial b_k} dx = \beta_n^d |t| \quad \forall j \neq k .$$

Thus, we may use (2.7) and (2.8) for the cases $k = j$ and $k \neq j$, respectively, to complete the sum in (2.6) over the index m first. This leads to

$$\text{Tr}(K_t) = \alpha_n^d |t| \sum_{i=1}^d \sum_{j=1}^{d+1} \left(\frac{\partial b_j}{\partial x_i} \right)^2 + \beta_n^d |t| \sum_{i=1}^d \sum_{j=1}^{d+1} \sum_{k \neq j}^{d+1} \frac{\partial b_j}{\partial x_i} \frac{\partial b_k}{\partial x_i} .$$

Noticing from the definition of $\{b_j\}$ that

$$\nabla \left(\sum_j b_j \right) \cdot \nabla \left(\sum_j b_j \right) = 0 ,$$

we then further obtain

$$\begin{aligned} \text{Tr}(K_t) &= (\alpha_n^d - \beta_n^d) |t| \sum_{i=1}^d \sum_{j=1}^{d+1} \left(\frac{\partial b_j}{\partial x_i} \right)^2 \\ (2.9) \quad &= (\alpha_n^d - \beta_n^d) |t| \sum_{j=1}^{d+1} |\nabla b_j|^2 = (\alpha_n^d - \beta_n^d) d^{-2} Q_d(t) , \end{aligned}$$

where, according to (2.4), the term $\mathcal{Q}_d(t)$ is given by

$$(2.10) \quad \mathcal{Q}_d(t) = \frac{1}{|t|} \sum_{i=1}^{d+1} |A_i|^2$$

for any d -dimensional simplex t with $|t|$ being its volume and $\{A_i\}_{i=1}^{d+1}$ being the areas (volumes) of its $(d - 1)$ -dimensional faces.

Now, if we let $\gamma_n^d = (\alpha_n^d - \beta_n^d)d^{-2}$, then by a symmetry consideration, we can get the following equivalent form of γ_n^d :

$$\gamma_n^d = \frac{1}{d^3(d+1)|t|} \sum_{m=1}^n \sum_{j=1}^{d+1} \sum_{k=1}^{d+1} \int_t \left(\frac{\partial L_m}{\partial b_j} - \frac{\partial L_m}{\partial b_k} \right)^2 dx .$$

Moreover, with a change of variable in the integral, we get a geometry-independent form of γ_n^d as follows:

$$(2.11) \quad \gamma_n^d = \frac{1}{d^3(d+1)|t_0|} \sum_{m=1}^n \sum_{j=1}^{d+1} \sum_{k=1}^{d+1} \int_{t_0} \left(\frac{\partial L_m}{\partial b_j} - \frac{\partial L_m}{\partial b_k} \right)^2 dx ,$$

where t_0 is the standard reference simplex defined in (2.3).

We thus arrive at the following theorem.

THEOREM 2.1 (a new trace formula). *For any general finite element spaces defined on a simplicial mesh τ with the nodal basis on any d -dimensional simplex $t \in \tau$ satisfying the invariance property specified above, the element stiffness matrix for (2.5) has the trace formula*

$$(2.12) \quad \text{Tr}(K_t) = \gamma_n^d \mathcal{Q}_d(t),$$

where n is the cardinality of the set of local nodal basis functions, γ_n^d is the positive constant defined by (2.11), and $\mathcal{Q}_d(t)$ is as defined by (2.10).

It is important to note that γ_n^d is a positive constant that depends only on the corresponding basis functions on the reference simplex t_0 and is independent of the geometry of the particular element t . Thus, we see the elegance of the above trace formula: it implies that the trace of the element stiffness matrix for general finite element spaces (with an invariant basis) is a product of two factors, with one being γ_n^d , which is completely independent of the element t , and the other being $\mathcal{Q}_d(t)$, the trace of K_t corresponding to the linear nodal basis consisting of $\{b_j\}_{j=1}^{d+1}$, which is completely independent of the choice of the finite element spaces (as long as they take some invariant basis). While the calculation of the special case for the linear element is widely known in standard finite element texts [2, 11, 38], to the best of our knowledge, the more general cases have not been presented in the literature. Our derivation of the results is indeed for general finite element spaces on simplicial meshes that include the classical standard Lagrange finite element spaces of any order, and other exotic spaces, such as the enrichment of the conforming linear element with bubble functions or stabilized finite element spaces [1].

As a corollary, using the nonnegativeness of K_t , we can get an estimate for the maximum eigenvalue $\lambda_n^{K_t}$ of the element stiffness matrix K_t .

COROLLARY 2.1. *Under the above conditions, we have*

$$(2.13) \quad \frac{\gamma_n^d}{n-1} \mathcal{Q}_d(t) \leq \lambda_n^{K_t} \leq \gamma_n^d \mathcal{Q}_d(t) .$$

Though the upper and lower bounds in (2.13) differ by a factor of $n - 1$, the above estimate does provide a precise control on the contribution due to the mesh geometry on the largest eigenvalue of the element stiffness matrix. To be discussed later, this is crucial to the application of the framework developed in [20, 21] for estimating the condition number of the assembled global stiffness matrix K on the whole domain.

Naturally, by summing over all elements, we may also get a trace formula for the global stiffness matrix using the result of the above theorem. Let us consider

$$(2.14) \quad -\nabla \cdot (\mu \nabla u) = f \quad \text{in } \Omega,$$

with a diffusion coefficient $\mu = \mu(x)$ and Neumann boundary condition $\frac{\partial u}{\partial n} = g$ on $\partial\Omega$. Assume that the f and g are compatible so that the equation is solvable.

COROLLARY 2.2. *For any general finite element spaces defined on a simplicial mesh τ with the nodal basis on any d -dimensional simplex $t \in \tau$ satisfying the invariance property specified above, let K_μ be the global stiffness matrix of (2.14) with a Neumann boundary condition. If μ remains a constant on t for any $t \in \tau$, then K has the trace formula*

$$(2.15) \quad \text{Tr}(K_\mu) = \gamma_n^d \sum_{t \in \tau} \mu_t Q_d(t),$$

where μ_t denotes the value of μ on $t \in \tau$.

The trace formulae can be extended to more general cases, where μ is not necessarily a constant on t but remains invariant under the transformation of permuting the vertices. For example, in two dimensions, μ on an element t can take on a function of the form $c_1 + c_2 b_1 b_2 b_3$, with $\{b_i\}$ being the barycentric coordinates on t and c_1, c_2 being some constants.

Note that for Dirichlet boundary conditions, contributions from the basis functions corresponding to the boundary nodes are not normally assembled into the stiffness matrix, which thus may lead to a minor alteration of the trace formula. We note that in the literature, it has been suggested that the minimization of the trace of the stiffness matrix can be used to optimize finite element grids; we see from (2.15) that the dependence of the trace on the mesh geometry is in fact the same for general finite element spaces.

In practical implementation of the finite element methods, especially with the use of high-order finite element spaces, the assembly of the stiffness and mass matrices is often done with the help of numerical integration. With enough precision in the numerical quadrature, the order of accuracy of the finite element methods can be preserved [38].

It is then natural to ask if the use of quadrature affects the discussions in this paper and thus the relation between the mesh geometry and the conditioning of the stiffness and mass matrices.

Let us consider first a simplex t which is mapped via an affine transform F to the reference element t_0 . Let $\{w_m, y_m \in t\}$ be a quadrature formula on t_0 , that is,

$$\int_{t_0} g(y) dy \sim \sum_m w_m |t_0| g(y_m).$$

Notice that a factor t_0 is added in the quadrature so that a normalization condition $\sum_m w_m = 1$ is satisfied. We assume in addition that $\{w_m, y_m \in t\}$ gives an *invariant quadrature*; that is, it is invariant with respect to a permutation of the vertices of

t_0 , which is satisfied, for instance, by the one point quadrature at the barycenter, the midside rule, and other invariant high-order Gaussian quadratures [41]. For the entries of the element stiffness matrix for the Poisson equation $a_t(L_k, L_l)$, the integral on t is approximated by

$$(2.16) \quad \int_t g(x)dx \sim \sum_m w_m |t| g(F^{-1}y_m).$$

Now, define the modified bilinear form as

$$\hat{a}_t(\phi, \psi) = \sum_m w_m |t| \nabla \phi(x_m) \cdot \nabla \psi(x_m)$$

for any polynomials ϕ and ψ defined on t and $\{x_m = F^{-1}y_m\}$. We then can follow a similar derivation given above to compute the trace of the modified element stiffness matrix $\hat{K}_t = (\hat{a}_t(L_k, L_j))$ to get the following.

THEOREM 2.2. *For any general finite element spaces defined on a simplicial mesh τ with the nodal basis on any d -dimensional simplex $t \in \tau$ satisfying the invariance property specified above, we have the following trace formula for the modified element stiffness matrix \hat{K}^t for (2.5) computed using an invariant numerical quadrature:*

$$(2.17) \quad \text{Tr}(\hat{K}^t) = \hat{\gamma}_n^d \mathcal{Q}_d(t),$$

where n and $\mathcal{Q}_d(t)$ are as defined before, F is the affine map that maps t to t_0 , and $\hat{\gamma}_n^d$ is a positive constant defined by

$$(2.18) \quad \hat{\gamma}_n^d = \frac{1}{d^3(d+1)} \sum_{i=1}^n \sum_{j=1}^{d+1} \sum_{k=1}^{d+1} \sum_m w_m \left(\frac{\partial L_i}{\partial b_j}(y_m) - \frac{\partial L_i}{\partial b_k}(y_m) \right)^2.$$

The significance of Theorem 2.2 lies in the fact that the only geometric factor affecting the trace remains to be $\mathcal{Q}_d(t)$ even with the use of a numerical integration. Of course, the assumption that the quadrature is invariant is crucial for the observation to hold.

Before we conclude the discussion on the trace formula, we make a few comments on the constant $\hat{\gamma}_n^d$. First of all, it is possible to get some explicit estimates of $\hat{\gamma}_n^d$. For instance, as seen before, for a linear finite element in any dimension, we have $\hat{\gamma}_{d+1}^1 = 1/d^2$. Naturally, it would be interesting to investigate the asymptotic behavior of $\hat{\gamma}_n^d$ as n gets larger. This would be of interest for the case of very high order Lagrange elements and p or $h-p$ finite element spaces. Such a behavior will be studied in future works.

3. Mesh-dependent condition number estimates. In this section, we first discuss some detailed computations given in [37] on the relation between condition numbers of the stiffness matrices and the mesh geometry in some special cases. These results provide insight into the type of estimates we can expect in general. Afterwards, we recall some earlier estimates on the condition number of the stiffness matrices presented in [21]. We then use the trace formula derived in the previous section to reveal the detailed dependence of the condition numbers on the mesh geometry in the more general settings.

3.1. Some known results on the linear Lagrange finite element. We first focus on a special case corresponding to the Poisson equation with a homogeneous boundary condition:

$$(3.1) \quad \begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

and its equivalent variational weak form: Find $u \in H_0^1(\Omega)$ such that

$$(3.2) \quad \int_{\Omega} \nabla u \cdot \nabla v \, d\mathbf{x} = \int_{\Omega} f v \, d\mathbf{x} \quad \forall v \in H_0^1(\Omega).$$

While the explicit forms of the element stiffness matrices for linear triangular and tetrahedral elements can be found in many standard finite element texts, a detailed calculation can be found in [37], where careful discussions on the bounds of the eigenvalues of element stiffness matrices are also presented with respect to the mesh quality corresponding to the linear Lagrange finite element. Here, we briefly recall the results presented in [37]. Similar calculations have been given in many other works; see, for example, [20, 38, 40]. In the two space dimension, let $\{l_i, \theta_i\}$ ($i = 1, 2, 3$) be the edge lengths and internal angles of a triangle $t \in \tau$ with area $|t|$. Then the element stiffness matrix on the triangle t is precisely [37]

$$(3.3) \quad \begin{aligned} K_t &= |t| (\nabla b_i \cdot \nabla b_j) = \frac{1}{8|t|} \begin{pmatrix} 2l_1^2 & l_3^2 - l_1^2 - l_2^2 & l_2^2 - l_1^2 - l_3^2 \\ l_3^2 - l_1^2 - l_2^2 & 2l_2^2 & l_1^2 - l_2^2 - l_3^2 \\ l_2^2 - l_1^2 - l_3^2 & l_1^2 - l_2^2 - l_3^2 & 2l_3^2 \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} \cot(\theta_2) + \cot(\theta_3) & -\cot(\theta_3) & -\cot(\theta_2) \\ -\cot(\theta_3) & \cot(\theta_1) + \cot(\theta_3) & -\cot(\theta_1) \\ -\cot(\theta_2) & -\cot(\theta_1) & \cot(\theta_1) + \cot(\theta_2) \end{pmatrix}. \end{aligned}$$

In [37], the roots of its characteristic polynomial are computed as $\lambda_1 = 0$ and

$$(3.4) \quad \lambda_{2,3} = \frac{1}{8|t|} \left(l_1^2 + l_2^2 + l_3^2 \pm \sqrt{(l_1^2 + l_2^2 + l_3^2)^2 - 48|t|^2} \right).$$

The largest root $\lambda_3^{K_t}$ is a scale-invariant indicator of the quality of the triangle's shape in terms of (3.4). Similar calculations can be found in other works as well; see, for example, [36], where eigenvalues of the diagonally preconditioned element stiffness matrix have also been explicitly computed. Note that the eigenvalues are nonnegative and $\lambda_1 = 0$, so

$$(3.5) \quad \frac{1}{8|t|} \sum_j l_j^2 = \frac{1}{2} \sum \cot(\theta_j) \leq \lambda_3^{K_t} \leq \sum \cot(\theta_j) = \frac{1}{4|t|} \sum_j l_j^2.$$

The above equation is a special case of (2.13), and as explained in [37], it also shows that if any of the angles approaches 0 or π , it would lead to large $\lambda_3^{K_t}$, thus affecting the conditioning of the stiffness matrix. These angle conditions, as pointed out in [36], reflect the common knowledge of minimizing the element distortion, a principle behind the Delaunay triangulation [22, 37], and are compatible with the angle conditions for guaranteeing the uniform finite element approximations of derivatives [3, 38].

Similarly, the element stiffness matrix for the linear Lagrange element on a three-dimensional tetrahedron t can be written as [37]

$$(3.6) \quad K_t = \frac{1}{6} \begin{pmatrix} k_{11} & -l_{34} \cot(\theta_{34}) & -l_{24} \cot(\theta_{24}) & -l_{23} \cot(\theta_{23}) \\ -l_{34} \cot(\theta_{34}) & k_{22} & -l_{14} \cot(\theta_{14}) & -l_{13} \cot(\theta_{13}) \\ -l_{24} \cot(\theta_{24}) & -l_{14} \cot(\theta_{14}) & k_{33} & -l_{12} \cot(\theta_{12}) \\ -l_{23} \cot(\theta_{23}) & -l_{13} \cot(\theta_{13}) & -l_{12} \cot(\theta_{12}) & k_{44} \end{pmatrix},$$

where l_{ij} is the edge of t with a corresponding dihedral angle θ_{ij} , and the diagonal entries $\{k_{ii}\}$ are such that the row sums are all identically zero.

In [37], the characteristic polynomial of K_t is calculated as

$$(3.7) \quad p(\lambda) = \lambda^4 - \frac{1}{9|t|} \sum_{i=1}^4 |A_i|^2 \lambda^3 + \frac{1}{36} \sum_{1 \leq j < k \leq 4} l_{jk}^2 \lambda^2 - \frac{|t|}{9} \lambda.$$

From (3.6), we can see that if one of the dihedral angles approaches 0, its cotangent approaches infinity, and so does $\lambda_4^{K_t}$, the maximum eigenvalue of K_t . For a tetrahedron, it is possible for one dihedral angle to be arbitrarily close to π without any dihedral angle of the tetrahedron being small (see [37]). Although an angle approaching π has a cotangent approaching negative infinity, surprisingly, such a tetrahedron does not induce a large eigenvalue in K_t because each entry on the diagonal of K_t is nonnegative and has the form $\sum_{i,j} l_{ij} \cot \theta_{ij}$. Therefore, if t has no dihedral angle close to 0, the diagonal entries of K_t are bounded from the above, and thus so is $\lambda_4^{K_t}$. This observation does not depend on whether t has planar angles near 0.

For $\lambda_4^{K_t}$ of a tetrahedron t , the following equation holds [37]:

$$(3.8) \quad \frac{\mathcal{Q}_3(t)}{27} \leq \lambda_4^{K_t} \leq \frac{\mathcal{Q}_3(t)}{9},$$

where $\mathcal{Q}_3(t)$ is as given in (2.10). This is again a special case of our general estimates (2.13) for the linear tetrahedral element (with $d = 3$ and $n = 4$). It shows that $\lambda_4^{K_t}$ (and thus λ_N^K) is not scale-invariant so that $\lambda_4^{K_t}$ grows linearly with the longest edge, as pointed out in [37].

These calculations give some insight into how the conditioning of the element stiffness matrix for the linear element might be dependent on the mesh geometry. For a general higher-order finite element, it is not always possible to analytically solve for the eigenvalues of element stiffness matrices. Instead, the trace formula developed in the previous section can help establishing the link between the mesh and the element stiffness matrices for general finite element spaces. The only key step that remains to be worked out is to see how the global stiffness matrix condition number is related to that of the element stiffness matrix. This is to be addressed next.

3.2. Some known condition number estimates. As stated before, we are interested in studying the stiffness matrix conditioning for general self-adjoint elliptic equations discretized by general finite element spaces on unstructured simplicial meshes. In [20, 21], a general estimate on the spectral properties of the global stiffness matrix in relation to that of the element stiffness matrix was given:

$$(3.9) \quad \max_{t \in \tau} (\lambda_n^{K_t}) \leq \lambda_N^K \leq P_* \max_{t \in \tau} (\lambda_n^{K_t}),$$

$$(3.10) \quad \lambda_1^* \min_{t \in \tau} (\lambda_1^{M_t}) \leq \lambda_1^K \leq \lambda_1^* P_* \max_{t \in \tau} (\lambda_n^{M_t}),$$

where P_* is the maximal number of elements in τ meeting at a nodal point, and λ_1^* is the smallest eigenvalue of the following elliptic eigenproblem:

$$(3.11) \quad \begin{cases} - \sum_{i,j=1}^d \frac{\partial}{\partial x_i} \left(a_{ij} \frac{\partial u}{\partial x_j} \right) = \lambda \rho u & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

where λ denotes any of the eigenvalues and u denotes a corresponding nonzero eigenfunction. The density ρ can be taken to be the unit constant in most cases, but for highly nonuniform meshes, a nonuniform density tends to give sharper estimates. From (3.9) and (3.10), the spectral condition number of the stiffness matrix K , $\text{Cond}(K)$, satisfies the following inequalities [21]:

$$(3.12) \quad \frac{\max_{t \in \tau}(\lambda_n^{K_t})}{\lambda_1^* P_* \max_{t \in \tau}(\lambda_n^{M_t})} \leq \text{Cond}(K) \leq \frac{P_* \max_{t \in \tau}(\lambda_n^{K_t})}{\lambda_1^* \min_{t \in \tau}(\lambda_1^{M_t})}.$$

The lowest exact eigenvalue λ_1^* can be regarded as a constant that depends only on the intrinsic properties of the continuous problem but does not depend on the discretization parameters. In this paper, we always consider those meshes with uniformly bounded P_* .

It is certainly interesting to examine the sharpness of the estimates (3.12), or rather, the corresponding estimates on the extreme eigenvalues in (3.9)–(3.10). As our interests are to explore the connection between the mesh geometry and the condition number estimates, it can be seen from (3.9) that the estimate on the largest eigenvalue of the stiffness matrix is sharp up to at most a mesh-independent constant factor. But the lower and upper bounds in (3.10) can be different by orders of magnitude in an unstructured grid for a constant density ρ . A nonuniform density that matches the element volumes can help make the bounds sharper, as shown in [20]. This issue will be revisited in later sections. We note here that in cases where (3.12) is sharp, it remains to find good estimates on the extreme eigenvalues of the element stiffness and mass matrices.

3.3. Condition number estimates for general finite element spaces. For a given PDE, the relation between mesh geometry and stiffness matrix conditioning may vary with respect to different finite element spaces. To be able to utilize the trace formula and the estimates on the maximum eigenvalues of the element stiffness matrix established in the previous section, we again focus on the model problem (3.1) with an appropriate finite element space. Hence, in this subsection, K denotes the global stiffness matrix corresponding only to (3.1).

By (3.12), to bound the condition number of K , we need estimates on $\lambda_1^{M_t}$, $\lambda_n^{M_t}$, and $\lambda_n^{K_t}$ for the element mass and stiffness matrices corresponding to (3.1).

The dependence of the spectral properties of the mass matrices on the mesh geometry has been previously studied. Some detailed computation can be found, for example, in [40]. For the element mass matrices, the computation is even simpler. Given a general finite element basis function $\psi = \psi(x)$ of the form

$$\psi(x) = \sum_{|\vec{k}|_1 \leq n} \alpha_{\vec{k}} b_1^{k_1} b_2^{k_2} \dots b_{d+1}^{k_{d+1}},$$

where $\{b_i\}$ are the barycentric coordinates, \vec{k} is a $(d + 1)$ -dimensional multi-index with $|\vec{k}|_1$ being the l_1 norm, and the coefficients $\alpha_{\vec{k}}$ depend only on the finite element

space chosen, but are independent of mesh geometry. For the uniform density $\rho = 1$, using a change of variable to the reference element t_0 , it is easy to get the following.

LEMMA 3.1. *For the model bilinear form a_Ω in (2.5), for the constant density $\rho = 1$, the element mass matrix M_t on the element t satisfies*

$$(3.13) \quad M_t = \frac{|t|}{|t_0|} M_{t_0} .$$

Consequently,

$$(3.14) \quad \min_{t \in \tau} \lambda_1^{M_t} = \delta_n \min_{t \in \tau} |t| , \quad \max_{t \in \tau} \lambda_n^{M_t} = \sigma_n \max_{t \in \tau} |t| ,$$

where δ_n and σ_n are two constants given by

$$(3.15) \quad \delta_n = \frac{1}{|t_0|} \lambda_1^{M_{t_0}} , \quad \sigma_n = \frac{1}{|t_0|} \lambda_n^{M_{t_0}} .$$

Note that the constants δ_n and σ_n are independent of the element t but only on t_0 and the corresponding local finite element basis. For a nonuniform density ρ , we have the following.

LEMMA 3.2. *For the model bilinear form a_Ω in (2.5), we have*

$$(3.16) \quad \delta_n \min_{t \in \tau} \{\rho_{\min}^t |t|\} \leq \min_{t \in \tau} \lambda_1^{M_t} \leq \delta_n \max_{t \in \tau} \{\rho_{\max}^t |t|\} ,$$

$$(3.17) \quad \sigma_n \max_{t \in \tau} \{\rho_{\min}^t |t|\} \leq \max_{t \in \tau} \lambda_n^{M_t} \leq \sigma_n \max_{t \in \tau} \{\rho_{\max}^t |t|\} ,$$

where ρ_{\min}^t and ρ_{\max}^t are the minimum and maximum values of ρ on the element t .

Proof. Given any $\vec{z} = (z_1, \dots, z_n)^T$, we define the function $\phi = \sum_{k=1}^n z_k L_k$, where $\{L_k\}$ is the nodal basis of the finite element space on t . Obviously, we have

$$\rho_{\min}^t \int_t \phi^2 dt \leq \int_t \rho \phi^2 dt \leq \rho_{\max}^t \int_t \phi^2 dt .$$

By the definition of the element mass matrices, it is then easy to see that

$$\rho_{\min}^t \vec{z}^T M_t^1 \vec{z} \leq \vec{z}^T M_t \vec{z} \leq \rho_{\max}^t \vec{z}^T M_t^1 \vec{z} ,$$

where M_t^1 corresponds to the element mass matrix with the constant density $\rho = 1$. Then by Lemma 3.1 and the variational definitions of the extreme eigenvalues, we immediately get the results in (3.16) and (3.17). \square

The above lemmas are valid for general finite element spaces, and it simply implies that, by (3.10), a lower bound for the smallest eigenvalue λ_1^K of the global stiffness matrix is proportional to the volume of the smallest element, while an upper bound is proportional to the volume of the largest element; that is, see the following.

LEMMA 3.3. *Under the conditions on the finite element spaces described earlier, for the model bilinear form a_Ω in (2.5), the smallest eigenvalue of the global stiffness matrix satisfies*

$$(3.18) \quad \lambda_1^* \delta_n \min_{t \in \tau} \{\rho_{\min}^t |t|\} \leq \lambda_1^K \leq \lambda_1^* P_* \sigma_n \max_{t \in \tau} \{\rho_{\max}^t |t|\} ,$$

where δ_n and σ_n are two constants defined in (3.15).

We note that the lower and upper bounds in (3.18) remain nearly on the same order for meshes with quasi-uniform element volumes in terms of the dependence on the mesh geometry. Thus, we expect that (3.18) may be less effective in highly graded or adapted meshes containing elements of very different sizes. We will revisit this in later discussions.

Now to complete the condition number estimate, we need only bound the largest eigenvalue λ_N^K . From (3.9), we know that λ_N^K is related to the largest eigenvalues of the element stiffness matrices. Given the bounds on the largest eigenvalues of the element stiffness matrices in (2.13), bounds of λ_N^K may be derived.

LEMMA 3.4. *Under the conditions on the finite element spaces described earlier, for the model bilinear form a_Ω in (2.5), the largest eigenvalue of the global stiffness matrix satisfies*

$$(3.19) \quad \frac{\gamma_n^d}{(n-1)} \max_{t \in \tau} \{Q_d(t)\} \leq \lambda_N^K \leq \gamma_n^d P_* \max_{t \in \tau} \{Q_d(t)\}.$$

Combining the results of Lemmas 3.3 and 3.4, we get the following.

THEOREM 3.1. *Under the assumptions on the finite element spaces made earlier, we have, for the model bilinear form a_Ω in (2.5), the following condition number estimate:*

$$(3.20) \quad \frac{\gamma_n^d \max_{t \in \tau} \{Q_d(t)\}}{(n-1) \lambda_1^* P_* \sigma_n \max_{t \in \tau} \{\rho_{\max}^t |t|\}} \leq \text{Cond}(K) \leq \frac{\gamma_n^d P_* \max_{t \in \tau} \{Q_d(t)\}}{\lambda_1^* \delta_n \min_{t \in \tau} \{\rho_{\min}^t |t|\}}.$$

The proof of the theorem simply follows directly from the application of Lemma 3.1, Corollary 2.1, and estimates (3.9) and (3.10).

The above result is for the Poisson equation (3.1), and the results for general diffusion equations can also be derived. As our objective is to explore the mesh dependence, we do not intend to get the optimal estimates with respect to all the quantities and parameters involved. Instead, we focus on results that have precise dependence on the geometric factors of the simplicial meshes. This can be easily achieved. For example, let us consider the following diffusion equation with a variable diffusion coefficient

$$(3.21) \quad \begin{cases} -\nabla(A(x)\nabla u) = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega \end{cases}$$

with $A = A(x)$ a $d \times d$ symmetric positive definite tensor satisfying

$$(3.22) \quad 0 < \beta_1 I \leq A(x) \leq \beta_2 I$$

uniformly for $x \in \Omega$ for some positive constants β_1 and β_2 .

We use K_A to denote the stiffness matrix associated with the finite element discretization of (3.21) to differentiate from the notation K , which is reserved to denote the stiffness matrix for the Poisson equation (3.1) in this subsection. Then it is easy to check that for any $\vec{y} \in R^N$, we have

$$\vec{y}^T K_A \vec{y} = \int_{\Omega} (\nabla u_h)^T A(x) \nabla u_h dx \leq \beta_2 \int_{\Omega} (\nabla u_h)^T \nabla u_h dx = \beta_2 \vec{y}^T K \vec{y},$$

and similarly,

$$\vec{y}^T K_A \vec{y} \geq \beta_1 \vec{y}^T K \vec{y}.$$

Thus, using the standard variational characterization of the extreme eigenvalues, we immediately get the result of the following theorem.

THEOREM 3.2. *Under the assumptions on the finite element spaces made earlier, we have, for the stiffness matrix K_A corresponding to (3.21), the estimate for the smallest eigenvalue,*

$$(3.23) \quad \beta_1 \lambda_1^* \delta_n \min_{t \in \tau} \{\rho_{\min}^t |t|\} \leq \lambda_1^{K_A} \leq \beta_2 \lambda_1^* P_* \sigma_n \max_{t \in \tau} \{\rho_{\max}^t |t|\},$$

and the estimate for the largest eigenvalue,

$$(3.24) \quad \frac{\beta_1 \gamma_n^d}{(n-1)} \max_{t \in \tau} \{Q_d(t)\} \leq \lambda_N^{K_A} \leq \beta_2 \gamma_n^d P_* \max_{t \in \tau} \{Q_d(t)\}.$$

Consequently, we also have the following condition number estimates:

$$(3.25) \quad \frac{\beta_1 \gamma_n^d \max_{t \in \tau} \{Q_d(t)\}}{(n-1) \beta_2 \lambda_1^* P_* \sigma_n \max_{t \in \tau} \{\rho_{\max}^t |t|\}} \leq \text{Cond}(K_A) \leq \frac{\beta_2 \gamma_n^d P_* \max_{t \in \tau} \{Q_d(t)\}}{\beta_1 \lambda_1^* \delta_n \min_{t \in \tau} \{\rho_{\min}^t |t|\}}.$$

The above theorem is very general and is valid in any space dimension for a general diffusion equation and for a general and possibly high-order finite element space (with an invariant nodal basis) defined on a general unstructured simplicial mesh. Despite the appearance of many terms in the estimate (3.25), a very precise relation between the conditioning of the global stiffness matrix and the mesh geometry is revealed by the bounds. Indeed, the most relevant quantities in (3.25) to the meshing qualities are simply the two ratios

$$\max_{t \in \tau} \{Q_d(t)\} / \max_{t \in \tau} \{\rho_{\max}^t |t|\} \quad \text{and} \quad \max_{t \in \tau} \{Q_d(t)\} / \min_{t \in \tau} \{\rho_{\min}^t |t|\},$$

assuming that P_* , the maximal number of elements meeting at a nodal point, is under control. We note that for highly anisotropic problems or problems with strong inhomogeneous coefficients, the difference between β_1/β_2 and β_2/β_1 can be large. This issue is to be visited in later sections.

3.4. Mesh geometry and stiffness matrix conditioning. Based on Theorem 3.2, it can be said that, at least for problems that are not highly anisotropic, the most important geometric quantities that affect the conditioning of the global finite element stiffness matrix are the scaled volume $\rho_{\min}^t |t|$ of each element t (or $\rho_{\max}^t |t|$, as we anticipate that ρ_{\max}^t and ρ_{\min}^t are of the same order for a given t), and the corresponding value of $Q_d(t)$. This is a rather universal property that is valid for general finite element spaces and general model equations. An effective control on these quantities in the meshing procedure may bear significance on the control of the conditioning of the linear systems coming from the finite element approximations. In the two-dimensional case, we know that $Q_2(t)$ corresponds to $\sum \cot \theta_i$ with $\{\theta_i\}_{i=1}^3$ being the angles of the triangle t ; thus, avoiding small angles in the triangulation is always preferred, as in the case of the Delaunay triangulation [22, 34, 37]. In fact, $Q_d(t)$ (for $d = 2$ or 3) has also been used as a mesh quality measure in many earlier studies on unstructured triangular meshes [5, 27, 30]. It has been labeled as a (smooth) conditioning quality measure in [37] based on the explicit calculation quoted earlier for the special case of the Poisson equation with a piecewise linear element. Relations between $Q_d(t)$ and other mesh quantity measures (see a nice summary in [37]) can

also be established. For example, let $r_{in}(t)$ be the radius of the largest inner-sphere of t ; then

$$\mathcal{Q}_d(t)r_{in}^2(t) \leq |t|^{-1} \sum_{i=1}^{d+1} |A_i|^2 \hat{h}_i^2(t) \leq (d+1)^3 |t|,$$

where $\{\hat{h}_i(t)\}$ are the heights of the simplex t corresponding to the faces $\{A_i\}$. Similarly, letting $r_{mc}(t)$ be the radius of the smallest containment sphere of t (the min-containment radius [37]), we have

$$4\mathcal{Q}_d(t)r_{mc}^2(t) \geq |t|^{-1} \sum_{i=1}^{d+1} |A_i|^2 \hat{h}_i^2(t) \geq (d+1)^3 |t|.$$

These inequalities imply that

$$(3.26) \quad \frac{(d+1)^3 |t|}{4r_{mc}^2(t)} \leq \mathcal{Q}_d(t) \leq \frac{(d+1)^3 |t|}{r_{in}^2(t)}.$$

Thus, how $\mathcal{Q}_d(t)$ varies with respect to a scaled volume is very much related to the traditional characterization of the dependence of $r_{in}(t)$ and $r_{mc}(t)$ on the volume. We leave more discussions along this line for future work.

4. Numerical validation and applications. We now apply the general estimates obtained in the previous section to various special cases. Some of these are widely known and are consistent with the popular understanding in the finite element and meshing community, while others are interesting on their own. Numerical examples are provided to assess whether the estimates are sharp.

4.1. Two-dimensional uniform triangular element. As a special case, we consider a two-dimensional rectangular domain with a uniform triangular mesh consisting of right triangles, but with different aspect ratios; see Figure 4.1 for an illustration. We take $\rho = 1$ in this case. Let h be the length of the diagonal of each right triangle, and let θ and $\pi/2 - \theta$ be the two acute angles. Theorem 3.1 implies the following.

COROLLARY 4.1. *Given the uniform triangular mesh described above, and under the assumptions on the finite element spaces made earlier, for the model bilinear form a_Ω in (2.5), we have the condition number estimate*

$$(4.1) \quad \frac{c_1}{h^2 \sin^2(2\theta)} \leq \text{Cond}(K) \leq \frac{c_2}{h^2 \sin^2(2\theta)}$$

for some positive constants c_1 and c_2 , independent of h and θ .

Proof. It follows from a simple calculation that for each triangle t , we have $|t| = h^2 \sin(2\theta)/4$ and $\mathcal{Q}_2(t) = 8/\sin(2\theta)$. Substituting into the inequality (3.20), we get (4.1) immediately. \square

The result in Corollary 4.1 is widely known in the finite element and meshing community [2, 38]. It is in fact quite sharp. In Tables 4.1 and 4.2, we present some numerical results computed on such uniform triangular meshes with the total number of elements being fixed ($= 8192$), but with different values for the angle θ . Thus, we get meshes of varying degrees of aspect ratio, and h^2 is proportional to $\sin^{-1}(2\theta)$. The estimate (4.1) in Corollary 4.1 predicts that the condition number is proportional

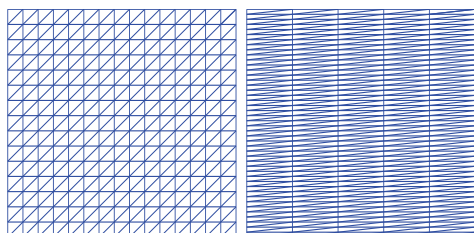


FIG. 4.1. Uniform triangular meshes with isosceles right triangles (left) and right triangles with small angles (right).

TABLE 4.1
The linear element case for the uniform triangular mesh.

Mesh	λ_{\max}^K	λ_{\max}^K	λ_{\min}^K	Condition number
4×1024	256.00	1024.011	0.004699	$2.179372e+5$
8×512	64.004	256.0577	0.004788	$5.347539e+4$
16×256	16.016	64.24519	0.004811	$1.335275e+4$
32×128	4.0655	16.99518	0.004817	$3.528104e+3$
64×64	1.5000	7.995182	0.004818	$1.659380e+3$

TABLE 4.2
The quadratic element case for uniform triangular mesh.

Mesh	λ_{\max}^K	λ_{\max}^K	λ_{\min}^K	Condition number
4×1024	682.67	1365.352	0.001198	$1.140045e+6$
8×512	170.69	341.4147	0.001204	$2.836673e+5$
16×256	42.750	85.66466	0.001205	$7.108828e+4$
32×128	11.023	22.66466	0.001205	$1.880261e+4$
64×64	4.7420	10.66466	0.001205	$8.846961e+3$

to $\sin^{-1}(2\theta)$, regardless of the order of the finite element spaces used. The same proportionality is true for λ_{\max}^K as predicted by Lemma 3.4 and the computation above. In Tables 4.1 and 4.2, for each mesh we report the corresponding largest eigenvalue of the element stiffness matrix, the extreme eigenvalues, and the condition number of global stiffness matrix. The quantity λ_{\min}^K is nearly unchanged, which is consistent with the theoretical prediction in Lemma 3.3 since the elements have a constant volume. Meanwhile, λ_{\max}^K and $\text{Cond}(K)$ both grow when θ approaches 0 or $\pi/2$, while their minimum values are attained for $\theta = \pi/4$ corresponding to the 64×64 mesh.

In Figure 4.2, we plot with respect to $\sin^{-1}(2\theta)$ (the horizontal axis) the curves of the largest eigenvalue and the condition number, respectively, for both the linear and the quadratic elements. The condition number for the quadratic case is normalized by a factor of 5.12 so as to fit into the same plot range. The perfect linear behavior verifies the theoretical prediction.

4.2. Finite element on quasi-volume-uniform, shape-regular meshes.

The previous example focuses on the effect of the shape regularity on the condition number with a uniform element size (volume). We now discuss some effect of the element size on the condition number when the shapes of the elements remain regular.

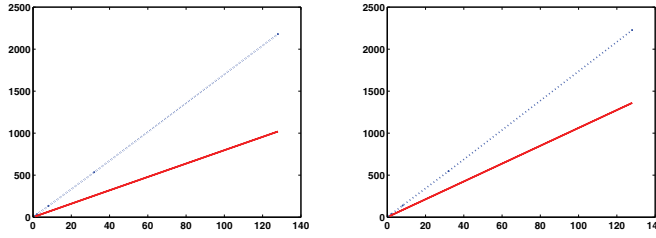


FIG. 4.2. Plots against $\sin^{-1}(2\theta)$ (the horizontal axis) of $\text{Cond}(K)$ and $\lambda_{\max}(K)$ for the linear element (left) and $\text{Cond}(K)/5.12$ and $\lambda_{\max}(K)$ for the quadratic element (right). Here, the solid lines represent $\lambda_{\max}(K)$.

In this subsection, we consider a simplicial mesh τ with simplices $t \in \tau$ satisfying

$$(4.2) \quad \rho_1 |t|^{2-2/d} \leq \sum_i A_i^2 \leq \rho_2 |t|^{2-2/d} \quad \forall t \in \tau$$

for some positive constants ρ_1 and ρ_2 , independent of t . We refer to such meshes as *shape regular*. In light of (3.26), to assure (4.2), it is sufficient to assume that

$$r_{mc}(t) \leq \rho_3 r_{in}(t) \quad \forall t \in \tau$$

for some constant ρ_3 . Note that the latter condition is consistent with the traditional meaning of shape regularity given in standard texts (see, e.g., [11]).

Meanwhile, we refer to a simplicial mesh τ as being *quasi volume-uniform* if

$$(4.3) \quad \min_{t \in \tau} |t| \geq \rho_3 \max_{t \in \tau} |t| \quad \forall t \in \tau$$

holds for some positive constant ρ_3 , independent of t . Note also that this is somewhat different from the traditional notion of a quasi-uniform mesh, which is measured using the diameters of the elements rather than the volumes [11].

First of all, we take $d = 2$ and consider the conforming linear element space on a quasi-volume-uniform and shape-regular triangulation. Theorem 3.1 implies the following.

COROLLARY 4.2. *For the model bilinear form a_Ω in (2.5) with a two-dimensional linear triangular element space defined on a quasi-volume-uniform and shape-regular triangulation, if h is the mesh parameter (diameter of the largest triangle), then*

$$c_1 h^{-2} \leq \text{Cond}(K) \leq c_2 h^{-2}$$

for some constants c_1 and c_2 , independent of h .

Proof. We notice that under the assumption on the triangulation, for each triangle t , $\mathcal{Q}_2(t) = |t|^{-1} \sum_{i=1}^3 |A_i|^2$ and $|t|h^{-2}$ remains uniformly bounded below and above by positive constants. Substituting into the inequality (3.20) with $\rho = 1$, we get the corollary immediately. \square

While the above corollary is widely known, a lesser-known version about general Lagrange triangular finite element spaces remains true [2].

COROLLARY 4.3. *For the model bilinear form a_Ω in (2.5) discretized by a finite element space with an invariant basis defined on a quasi-volume-uniform d -dimensional*

simplicial mesh with h being the mesh parameter (diameter of the largest simplex), if we further assume that all the simplices are shape regular in the sense of (4.2), then

$$c_1^{(n,d)} h^{-2} \leq \text{Cond}(K) \leq c_2^{(n,d)} h^{-2}$$

for some constants $c_1^{(n,d)}$ and $c_2^{(n,d)}$, which are dependent on the finite element basis on the reference element t_0 and dimension d , but are independent of h .

The proof follows from the same line of argument as in the two-dimensional linear element case. We note that these corollaries can of course be derived in other ways, for instance, with the use of inverse inequality [11].

4.3. Finite element on nonuniform shape-regular meshes. We now consider the case of shape-regular meshes, as specified by (4.2), but without the quasi-volume-uniform assumption. Thus, the element sizes $|t|$ are allowed to vary in a very large range. We then have the following.

COROLLARY 4.4. *For the model bilinear form a_Ω in (2.5) with a general simplicial finite element space satisfying the conditions given in Theorem 3.1, corresponding to a d -dimensional simplicial mesh τ satisfying condition (4.2), we have*

$$(4.4) \quad \frac{c_1 \max_{t \in \tau} |t|^{1-2/d}}{\max_{t \in \tau} \{\rho_{\max}^t |t|\}} \leq \text{Cond}(K) \leq \frac{c_2 \max_{t \in \tau} |t|^{1-2/d}}{\min_{t \in \tau} \{\rho_{\min}^t |t|\}}$$

for some constants c_1 and c_2 which are dependent on the finite element space but are independent of mesh geometry.

The above result is interesting, for example, in the context of adaptive finite element simplicial meshes satisfying (4.2) but containing elements with considerable variations in their sizes. Preserving shape regularity is often implemented in the local mesh refinement procedure so that it is reasonable to expect that (4.2) is satisfied.

Let h_{\min} be the diameter of the smallest element in an adaptive finite element mesh satisfying (4.2). In both one and two space dimensions, we see that the use of a constant density $\rho = 1$ would yield an upper bound proportional to h_{\min}^{-2} , which is about the same order for the condition number of the linear system resulting from a uniform mesh of the mesh size h_{\min} , though the degree of freedom (and thus the dimension of the global stiffness matrix) may be much smaller in the adaptive case than in the uniform case. Yet, this is generally not sharp. In [20], it was shown that with the element size distribution being inversely proportional to the nonuniform density, that is,

$$(4.5) \quad c_1 N^{-1} \leq \rho_{\min}^t |t| \leq \rho_{\max}^t |t| \leq c_2 N^{-1},$$

where N is the number of elements in τ , and c_1 and c_2 are some positive constants, the sharper upper bound

$$(4.6) \quad \text{Cond}(K) \leq cN h_{\min}$$

holds. This is naturally consistent with the estimate given in (4.4). The sharper estimate indicates a much better condition number, and thus further demonstrates the greater efficiency of the adaptive mesh in both representing the PDE solutions and improving the conditioning of the resulting linear systems. For the inequalities (4.5) to hold for a smoothly defined density function, the variation in the element sizes needs to be properly controlled. Yet, we present a simple numerical example to

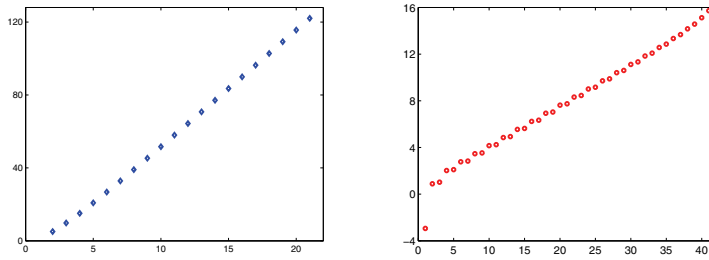


FIG. 4.3. Plots of $\text{Cond}(K)h_{\min}$ (left) with respect to different m (horizontal axis) and the logarithms of all 41 eigenvalues of K for $m = 21$ (right).

illustrate that the bound (4.6) remains quite accurate even for highly graded meshes. We take a two point boundary value problem,

$$(4.7) \quad -u'' = f \quad \text{on } (-1, 1), \quad \text{and} \quad u(-1) = u(1) = 0,$$

and consider the discretization with the linear finite element on a geometrically graded mesh $x_i = \text{sgn}(i - m)2^{|i-m|-m}$ for $0 \leq i \leq 2m$. For this mesh, $h_{\min} = 2^{1-m}$ and $N = 2m - 1$. We note first that the bounds on the largest eigenvalue in Lemma 3.4 gives the sharp estimate

$$\lambda_{2m-1}^K = O(h_{\min}^{-1}) = O(2^m).$$

In Figure 4.3, we plot, with respect to m (the horizontal axis in the left figure), the product of h_{\min} and $\text{Cond}(K)$. The near linear scaling in m of $\text{Cond}(K)h_{\min}$ implies that $\text{Cond}(K)$ grows on the order of Nh_{\min}^{-1} rather than $O(h_{\min}^{-2})$, which is consistent with the sharper estimate (4.6).

As mentioned in [37], a few small elements in a largely uniform mesh tend to produce large condition numbers, but in fact, they may only lead to a few outliers in the eigenvalue distributions and can thus be treated effectively. We also plot in Figure 4.3 the distribution of the logarithm of all 41 eigenvalues for the stiffness matrix corresponding to $m = 21$. It shows that, rather than giving only a few outliers, the geometrically (exponentially) graded meshes produced nearly exponentially distributed eigenvalues. The same numerical results can also be reproduced for two-dimensional analogues as well.

Similar numerical examples can be constructed for two-dimensional problems. In dimensions three or higher, if we take $\rho = 1$, then the upper bound of the condition number estimate in Corollary 4.4 shows the dependence on h_{\min}^{-d} , which is even worse than the dimension-independent estimate $O(h_{\min}^{-2})$ in Corollary 4.2 for a quasi-volume-uniform shape-regular mesh with mesh size h_{\min} . One may expect that it might be possible to get sharper bounds using a nonuniform density. We will examine these issues in greater detail in the future.

4.4. Finite element with three-dimensional tetrahedral meshes. Finite element methods are very popular for many large-scale three-dimensional problems. Three-dimensional unstructured tetrahedral mesh generation and optimization have also attracted much attention. For most mesh generators, a mesh sizing measure is introduced so that a mesh with suitably distributed sizing measure can be produced. Yet, controlling the shape regularity of the elements in spaces of three and higher dimensions remains a challenging task [17, 24, 31, 37].

TABLE 4.3

Extreme eigenvalues for Poisson equation with three-dimensional coarser meshes.

N_τ	$\max_{t \in \tau} \mathcal{Q}_3(t)$	$\min_{t \in \tau} t $	$\max_{t \in \tau} t $	$\lambda_{\min}^{K_1}$	$\lambda_{\max}^{K_1}$	$\lambda_{\min}^{K_2}$	$\lambda_{\max}^{K_2}$
7553	934	0.00097	0.236	0.240	110.42	0.03	209
7838	472	.00193	0.269	0.253	60.64	0.03	107.3
7879	311	.00298	0.282	0.254	36.15	0.03	72
7837	181	.00491	0.264	0.254	26.51	0.03	42.5
7532	160	.00515	0.208	0.2401	17.30	0.03	37.51
7737	118	.00594	0.242	0.247	18.24	0.03	28.67
7584	108	.0103	0.235	0.240	16.68	0.03	25.72
7545	95	0.00932	0.202	0.240	13.61	0.03	23.02

TABLE 4.4

Extreme eigenvalues for Poisson equation with a three-dimensional finer mesh.

N_τ	$\max_{t \in \tau} \mathcal{Q}_3(t)$	$\min_{t \in \tau} t $	$\max_{t \in \tau} t $	$\lambda_{\min}^{K_1}$	$\lambda_{\max}^{K_1}$	$\lambda_{\min}^{K_2}$	$\lambda_{\max}^{K_2}$
21244	2013	1.43e-4	0.0702	0.0836	227.8	0.01	450.11
22180	1103	1.58e-4	0.105	0.0831	113.37	0.01	247
22098	599	4.34e-4	0.101	0.0833	59.53	0.01	137.1
22060	401	5.99e-4	0.0927	0.0833	49.25	0.01	91.17
22065	335	8.09e-4	0.0958	0.0833	43.32	0.01	75.81
21460	259	8.66e-4	0.0727	0.0833	33.51	0.01	58.67
21522	257	7.51e-4	0.0743	0.0834	33.92	0.01	59.4
21710	255	4.73e-4	0.0854	0.0835	25.15	0.01	58.10
21638	121	2.25e-3	0.0837	0.0834	17.23	0.01	28.13
21575	115	2.18e-3	0.0784	0.0834	18.79	0.01	26.93
21315	114	2.34e-4	0.0749	0.0834	17.05	0.01	26.4
21273	84	2.73e-3	0.0696	0.0835	13.75	0.01	20.33

We now present some examples of the condition numbers of the stiffness matrix for the Poisson equation (2.5) in a cubic box $[0, 10]^3$ with a homogeneous Dirichlet boundary condition. The equation is solved based on some unstructured tetrahedral meshes generated with a uniform sizing measure. For detailed discussions on the related mesh generation procedures, we refer to [13, 14, 16, 17, 25] and the references cited therein. In our numerical results, computations are performed on meshes having two levels of resolution with the coarser meshes having element numbers ranging from 7500 to 7900 and with the finer meshes having element numbers ranging from 21200 to 22100. The results of the corresponding extreme eigenvalues of the global stiffness matrices denoted by $\{\lambda_{\min}^{K_i}, \lambda_{\max}^{K_i}\}_{i=1}^2$ for the linear and quadratic elements, respectively, are reported in Tables 4.3 and 4.4 for the various meshes.

It is of course straightforward to get the condition numbers from the ratios of the extreme eigenvalues. In each case, we also list the number of elements (N_τ) in the mesh τ , the maximum ($\max_{t \in \tau} |t|$) and minimum ($\min_{t \in \tau} |t|$) values of the element volumes, and the maximum value of $\mathcal{Q}_3(t)$ for $t \in \tau$. We may see from the tables that the smallest eigenvalues $\{\lambda_{\min}^{K_i}\}_{i=1}^2$ remain nearly constant for meshes at the same level with a ratio of nearly factor 8 between the linear and quadratic elements. Notice that the smallest and the largest element volumes do vary between meshes at the same level, so the lower and upper bounds in (3.18) are not tight in this case. Meanwhile, for the largest eigenvalues, they follow proportionally to the values of $\mathcal{Q}_d(t)$ as predicted by estimate (3.19). More extensive computational studies for more general equations and geometric domains are currently under investigation.

4.5. Effect of anisotropy. Diffusion equations with highly anisotropic coefficients have wide applications in many practical problems. In [37], some discussions have been given for the linear finite element corresponding to an anisotropic Poisson equation of the form (3.21) with $A = A(x)$ being replaced by a constant matrix B .

To simplify the notation, we take the two-dimensional case as an example. Let v_1, v_2 denote the orthogonal unit eigenvectors of B , and let ξ_1, ξ_2 be the corresponding eigenvalues. Then, $B = \xi_1 v_1 v_1^T + \xi_2 v_2 v_2^T$. Let

$$G = \frac{1}{\sqrt{\xi_1}} v_1 v_1^T + \frac{1}{\sqrt{\xi_2}} v_2 v_2^T, \quad \text{or equivalently, } G = B^{-\frac{1}{2}}.$$

Define the change of variable $(\tilde{x}, \tilde{y})^T = G(x, y)^T$ and $\tilde{f}(\tilde{x}, \tilde{y}) = f(G^{-1}(\tilde{x}, \tilde{y}))$. Let $\tilde{\Omega}$ denote the image of Ω and \tilde{t} denote the image of an element t for any $t \in \tau$. With the above change of variable, (3.21) with the constant coefficient matrix B becomes (2.5) for variables $(\tilde{x}, \tilde{y}) \in \tilde{\Omega}$ with unknown solution \tilde{u} and right-hand side \tilde{f} .

When the linear Lagrange finite element method is employed to solve the problem (3.21) with coefficient matrix B , since $G^{-1} \nabla b_i = \tilde{\nabla} \tilde{b}_i$ ($\tilde{\nabla}$ and $\{\tilde{b}_i\}$ are the gradient operator and the linear Lagrange basis in the new variable, respectively), we see that the element stiffness matrix K_t on a triangle t is identical to the element stiffness matrix for \tilde{t} corresponding to the Poisson equation (2.5). Consequently, equilateral elements may not necessarily lead to good conditioning for stiffness matrices of anisotropic equations [37]. In [33], it is argued that an optimal uniform triangular mesh is equilateral with respect to the metric, which is the inverse of the coefficient matrix, which is consistent to the computation given in [37].

With the help of transformation G , the computations given in [37] can be readily applied to the case of more general finite element spaces, following similar discussions given in the earlier sections. It can thus be seen that the important geometric factors affecting the stiffness matrix conditioning, for highly anisotropic problems, are $\rho_{\min}^{\tilde{t}} |\tilde{t}|$ and $\rho_{\max}^{\tilde{t}} |\tilde{t}|$ of the transformed element \tilde{t} , and the corresponding value of $|\tilde{t}|^{-1} \sum_{i=1}^{d+1} |\tilde{A}_i|^2$.

For instance, consider the two-dimensional case with B being a diagonal tensor with diagonal entries 81 and 1. In this case, G is also diagonal with entries 1/9 and 1. Hence, thin triangles with an aspect ratio of roughly nine, oriented parallel to the x -axis, ideally provide the optimal stiffness matrix conditioning. The numerical results in Table 4.5 are obtained by solving the anisotropic equation in a unit square with a linear finite element on meshes shown in Figure 4.1. As predicted, the condition number corresponding to the triangulation of the 27×243 rectangular mesh is the smallest.

TABLE 4.5
The linear element case for the anisotropic Poisson problem.

Mesh	$\lambda_{\max}^{K_t}$	λ_{\max}^K	λ_{\min}^K	Condition number
3×2187	7.290278e+2	2.916332e+3	0.112615	2.589641e+4
9×729	81.252329	3.278779e+2	0.122119	2.684904e+3
27×243	13.500000	71.876786	0.123214	5.833500e+2
81×81	81.252329	3.278767e+2	0.123336	2.658407e+3
243×27	7.290278e+2	2.916321e+3	0.123348	2.364306e+4

5. Conclusion and future work. In this paper, the relations between the spectral condition number of stiffness matrix and mesh geometry are systematically explored. Our main results are rigorously derived and yet applicable to very general equations, finite element spaces, and geometric meshes. They may lead to more work in the following two directions: better understanding of the effect of geometry on the matrix conditioning can lead to the development of better iterative solvers; at the same time, better mesh generation and optimization strategies and mesh quality measures can be devised to generate meshes on which a compromise between the efficiency of the solver and the discretization error can be reached so that optimal performance of finite element computations can be obtained.

There remain many interesting issues to be investigated in the future; for instance, preconditioning can greatly improve the performance of the linear algebraic solvers, and for many practical applications, the discrete algebraic problems can be tractable only if effective preconditioners are used. It will thus be interesting to study the precise dependence of the condition number estimates on the mesh geometry for preconditioned stiffness matrices [36]. Also, it is well known that the stiffness matrix conditioning will be different when different basis functions are employed [4, 9]. Comparisons of different basis selections for high-order elements remain to be investigated. This is particularly important for the p -version or $h-p$ version finite element methods [23, 35]. In addition, we have not considered equations involving convection terms, which may be solved by stabilized finite elements; the streamline-upwind Petrov/Galerkin methods; and the residual-free bubbles methods. Such discussions may become more complex due to the possible lack of symmetry in the stiffness matrix and the loss of variational structure. Extensions to other interesting physical models such as the elasticity equations and Stokes equations, and to nonsimplicial meshes such as quadrilateral and hexahedral meshes (see [32, 28]), can also be considered. While many more issues remain to be examined, the present work complements existing work in the literature, and together, they provide a rigorous and systematic foundation for future studies.

Acknowledgment. The authors thank the support of Lab for Scientific and Engineering Computing, Chinese Academy of Sciences, where this work was first initiated.

REFERENCES

- [1] D. ARNOLD, F. BREZZI, AND M. FORTIN, *A stable finite element for the Stokes equations*, *Calcolo*, 12 (1984), pp. 337–344.
- [2] O. AXELSSON AND V. BARKER, *Finite Element Solution of Boundary Value Problems*, Academic Press, London, 1983; reprinted as *Classics Appl. Math.* 35, SIAM, Philadelphia, 2001.
- [3] I. BABUŠKA AND A. K. AZIZ, *On the angle condition in the finite element method*, *SIAM J. Numer. Anal.*, 13 (1976), pp. 214–226.
- [4] E. BARRAGY AND G. CAREY, *Preconditioners for high degree elements*, *Comput. Methods Appl. Mech. Engrg.*, 93 (1991), pp. 97–110.
- [5] R. E. BANK AND R. K. SMITH, *Mesh smoothing using a posteriori error estimates*, *SIAM J. Numer. Anal.*, 34 (1997), pp. 979–997.
- [6] M. BATDORF, L. FREITAG, AND C. OLLIVIER-GOOCH, *Computational study of the effect of unstructured and mesh quality on solution efficiency*, in *Proceedings of the 13th CFD Conference*, AIAA, Reston, VA, 1997.
- [7] M. BERZINS, *Mesh quality: A function of geometry, error estimates or both?*, *Engineering with Computers*, 15 (1999), pp. 236–247.
- [8] M. BERZINS, *A solution-based triangular and tetrahedral mesh quality indicator*, *SIAM J. Sci. Comput.*, 19 (1998), pp. 2051–2060.

- [9] G. CAREY AND E. BARRAGY, *Basis function selection and precondition high degree finite element and spectral methods*, BIT, 29 (1989), pp. 794–804.
- [10] W. CAO, *On the error of linear interpolation and orientation, aspect ratio, and internal angles of a triangle*, SIAM J. Numer. Anal., 43 (2005), pp. 19–40.
- [11] P. CIARLET, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978; reprinted as Classics in Appl. Math. 40, SIAM, Philadelphia, 2002.
- [12] M. DELFOUR, G. PAYRE, AND J. ZOLESIO, *An optimal triangulation for second-order elliptic problems*, Comput. Methods Appl. Mech. Engrg., 50 (1985), pp. 231–261.
- [13] Q. DU, V. FABER, AND M. GUNZBURGER, *Centroidal Voronoi tessellations: Applications and algorithms*, SIAM Rev., 41 (1999), pp. 637–676.
- [14] Q. DU AND M. GUNZBURGER, *Grid generation and optimization based on centroidal Voronoi tessellations*, Appl. Comput. Math., 133 (2002), pp. 591–607.
- [15] Q. DU, Z. HUANG, AND D. WANG, *Mesh and solver co-adaptation in finite element methods for anisotropic problems*, Numer. Methods Partial Differential Equations, 21 (2005), pp. 859–874.
- [16] Q. DU AND D. WANG, *Tetrahedral mesh generation and optimization based on centroidal Voronoi tessellations*, Internat. J. Numer. Methods Engrg., 56 (2003), pp. 1355–1373.
- [17] Q. DU AND D. WANG, *Recent progress in robust and quality mesh generation*, J. Comput. Appl. Math., 195 (2006), pp. 8–23.
- [18] L. FREITAG AND C. OLLIVIER-GOOCH, *A cost/benefit analysis of simplicial mesh improvement techniques as measured by solution efficiency*, Internat. J. Comput. Geom. Appl., 10 (2000), pp. 361–382.
- [19] I. FRIED, *Condition of finite element matrices generated from nonuniform meshes*, AIAA Journal, 10 (1972), pp. 219–221.
- [20] I. FRIED, *Bounds on the spectral and maximum norms of the finite element stiffness, flexibility and mass matrices*, Int. J. Solids Structures, 9 (1973), pp. 1013–1034.
- [21] I. FRIED, *Numerical Solution of Differential Equations*, Academic Press, New York, 1979.
- [22] P. GEORGE AND H. BOROUCAKI, *Delaunay Triangulation and Meshing, Application to Finite Elements*, Hermès, Paris, 1998.
- [23] N. HU, X. GUO, AND I. KATZ, *Bounds for eigenvalues and condition number in the p -version of the finite element methods*, Math. Comp., 67 (1998), pp. 1423–1450.
- [24] L. JU, *Conforming centroidal Voronoi Delaunay triangulation for quality mesh generation*, Intern. J. Numer. Anal. Model., 4 (2007), pp. 531–547.
- [25] L. JU, M. GUNZBURGER, AND W. ZHAO, *Adaptive finite element methods for elliptic PDEs based on conforming centroidal Voronoi–Delaunay triangulations*, SIAM J. Sci. Comput., 28 (2006), pp. 2023–2053.
- [26] M. KITTUR, R. HUSTON, AND F. OSWALD, *Finite-Element Grid Improvement by Minimization of Stiffness Matrix Trace*, NASA Tech. Report 87-C-4, Glenn Research Center, Cleveland, OH, 1987.
- [27] P. KNUPP, *Matrix norms and the condition number: A general framework to improve mesh quality via node-movement*, in Proceedings of the 8th International Meshing Roundtable (Lake Tahoe, CA), 1999, pp. 13–22.
- [28] P. KNUPP, *Hexahedral and tetrahedral mesh shape optimization*, Internat. J. Numer. Methods Engrg., 58 (2003), pp. 319–332.
- [29] M. KRÍŽEK, *On the maximum angle condition for linear tetrahedral elements*, SIAM J. Numer. Anal., 29 (1992), pp. 513–520.
- [30] A. LIU AND B. JOE, *Relationship between tetrahedron shape measures*, BIT, 34 (1994), pp. 268–287.
- [31] G. MILLER, D. TALMOR, S.-H. TENG, AND N. WALKINGTON, *A Delaunay based numerical method for three dimensions: Generation, formulation and partition*, in Proceedings of the 27th ACM Symposium on Theory of Computing, ACM, New York, 1995, pp. 683–692.
- [32] P. MING AND Z. SHI, *Quadrilateral mesh*, Chin. Ann. Math. Ser. B, 23 (2002), pp. 235–252.
- [33] S. OH AND J. YIM, *Optimal finite element mesh for elliptic equation of divergence form*, Appl. Math. Comput., 162 (2005), pp. 969–989.
- [34] A. OKABE, B. BOOTS, K. SUGIHARA, AND S. CHIU, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, Wiley, Chichester, UK, 2000.
- [35] E. OLSEN AND J. DOUGLAS, *Bounds on spectral condition numbers of matrices arising in the p -version of the finite element method*, Numer. Math., 69 (1995), pp. 333–352.
- [36] A. RAMAGE AND A. J. WATHEN, *On preconditioning for finite element equations on irregular grids*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 909–921.

- [37] J. SHEWCHUK, *What is a Good Linear Finite Element? Interpolation, Conditioning, Anisotropy and Quality Measures*, Tech. report, Department of Computer Science, University of California, Berkeley, CA, 2003.
- [38] G. STRANG AND G. FIX, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
- [39] I. TSUKERMAN, *A general accuracy criterion for finite element approximation*, IEEE Trans. Magnetics, 35 (1998), pp. 1–4.
- [40] A. WATHEN, *Realistic eigenvalue bounds for the Galerkin mass matrix*, IMA J. Numer. Anal., 7 (1987), pp. 449–457.
- [41] Y. XU, *Orthogonal polynomials and cubature formulae on balls, simplices, and spheres*, J. Comput. Appl. Math., 127 (2001), pp. 349–368.

DYNAMICAL SYSTEMS AND NON-HERMITIAN ITERATIVE EIGENSOLVERS*

MARK EMBREE[†] AND RICHARD B. LEHOUCQ[‡]

Abstract. Simple preconditioned iterations can provide an efficient alternative to more elaborate eigenvalue algorithms. We observe that these simple methods can be viewed as forward Euler discretizations of well-known autonomous differential equations that enjoy appealing geometric properties. This connection facilitates novel results describing convergence of a class of preconditioned eigensolvers to the leftmost eigenvalue, provides insight into the role of orthogonality and biorthogonality, and suggests the development of new methods and analyses based on more sophisticated discretizations. These results also highlight the effect of preconditioning on the convergence and stability of the continuous-time system and its discretization.

Key words. eigenvalues, dynamical systems, inverse iteration, preconditioned eigensolvers, geometric invariants

AMS subject classifications. 15A18, 37C10, 65F15, 65L20

DOI. 10.1137/07070187X

1. Introduction. Suppose we seek a small number of eigenvalues (and the associated eigenspace) of the non-Hermitian matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$, having at our disposal a nonsingular matrix $\mathbf{N} \in \mathbb{C}^{n \times n}$ that approximates \mathbf{A} . Given a starting vector $\mathbf{p}_0 \in \mathbb{C}^n$, compute

$$(1.1) \quad \mathbf{p}_{j+1} = \mathbf{p}_j + \mathbf{N}^{-1}(\theta_j - \mathbf{A})\mathbf{p}_j,$$

where $\theta_j - \mathbf{A}$ is shorthand for $\mathbf{I}\theta_j - \mathbf{A}$, and

$$\theta_j = \frac{(\mathbf{A}\mathbf{p}_j, \mathbf{p}_j)}{(\mathbf{p}_j, \mathbf{p}_j)}$$

for some inner product (\cdot, \cdot) . Knyazev, Neymeyr, and others have studied this iteration for Hermitian positive definite \mathbf{A} ; see [21, 22] and references therein for convergence analysis and numerical experiments.

Clearly the choice of \mathbf{N} will influence the behavior of this iteration. With $\mathbf{N} = \mathbf{A}$, the method (1.1) reduces to (scaled) inverse iteration:

$$\mathbf{p}_{j+1} = \mathbf{A}^{-1}\mathbf{p}_j\theta_j.$$

We are interested in the case where \mathbf{N} approximates \mathbf{A} , yet one can apply \mathbf{N}^{-1} to a vector much more efficiently than \mathbf{A}^{-1} itself. Such a \mathbf{N} acts as a preconditioner for \mathbf{A} , and, hence, (1.1) represents a preconditioned iteration.

*Received by the editors September 4, 2007; accepted for publication (in revised form) November 7, 2008; published electronically March 13, 2009.

<http://www.siam.org/journals/sinum/47-2/70187.html>

[†]Department of Computational and Applied Mathematics, Rice University, 6100 Main Street – MS 134, Houston, TX 77005-1892 (embree@rice.edu). This author's research supported by U.S. Department of Energy grant DE-FG03-02ER25531 and National Science Foundation grant DMS-CAREER-0449973.

[‡]Sandia National Laboratories, P.O. Box 5800, MS 1110, Albuquerque, NM 87185-1110 (rblehou@sandia.gov). Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the U.S. Department of Energy under contract DE-AC04-94AL85000.

This method contrasts with a different class of algorithms, based on inverse iteration (or the shift-invert Arnoldi algorithm), that apply a preconditioner to accelerate an “inner iteration” that approximates the solution to a linear system at each step; see, e.g., [24, 13, 16] and [6, Chapter 11]. For numerous practical large-scale non-Hermitian eigenvalue problems, such as those described in [25, 41], these inner iterations can be extremely expensive and highly dependent on the quality of the preconditioner. In contrast, as we shall see, the iteration (1.1) can converge to a leftmost eigenpair even when \mathbf{N} is a suitable multiple of the identity.

This paper provides a rigorous convergence theory that establishes sufficient conditions for (1.1) to converge to the leftmost eigenpair for non-Hermitian \mathbf{A} . We obtain these results by viewing this iteration as the forward Euler discretization of the autonomous nonlinear differential equation

$$(1.2) \quad \dot{\mathbf{p}} = \mathbf{N}^{-1} \left(\mathbf{p} \frac{(\mathbf{A}\mathbf{p}, \mathbf{p})}{(\mathbf{p}, \mathbf{p})} - \mathbf{A}\mathbf{p} \right)$$

with a unit step size. Here \mathbf{A} and \mathbf{N} are fixed but \mathbf{p} depends on a parameter, t ; $\dot{\mathbf{p}}$ denotes differentiation with respect to t . In the absence of preconditioning, the differential equation (1.2) has been studied in connection with power iteration [10, 29], as described in more detail below. The nonzero steady-states of this system correspond to (right) eigenvectors of \mathbf{A} , and, hence, one might attempt to compute eigenvalues by driving this differential equation to steady-state as swiftly as possible. Properties of the preconditioner determine which of the eigenvectors is an attracting steady-state.

The differential equation (1.2) enjoys a distinguished property, observed, for example, in [10, 29] with $\mathbf{N} = \mathbf{I}$. Suppose that \mathbf{p} solves (1.2), $\theta = (\mathbf{p}, \mathbf{p})^{-1}(\mathbf{A}\mathbf{p}, \mathbf{p})$, and \mathbf{N} is self-adjoint and invertible (\mathbf{A} may be non-self-adjoint). Then for all t ,

$$(1.3) \quad \begin{aligned} \frac{d}{dt}(\mathbf{p}, \mathbf{N}\mathbf{p}) &= (\mathbf{N}^{-1}(\mathbf{p}\theta - \mathbf{A}\mathbf{p}), \mathbf{N}\mathbf{p}) + (\mathbf{p}, \mathbf{N}\mathbf{N}^{-1}(\mathbf{p}\theta - \mathbf{A}\mathbf{p})) \\ &= (\mathbf{p}\theta, \mathbf{p}) - (\mathbf{A}\mathbf{p}, \mathbf{p}) + (\mathbf{p}, \mathbf{p}\theta) - (\mathbf{p}, \mathbf{A}\mathbf{p}) \\ &= 0. \end{aligned}$$

Thus, $(\mathbf{p}, \mathbf{N}\mathbf{p})$ is an *invariant* (or *first integral*), as its value is independent of time; see [19, section 1.3] for a discussion of the unpreconditioned case ($\mathbf{N} = \mathbf{I}$), and, e.g., [4, 18] for a general introduction to invariant theory and geometric integration.

The invariant describes a manifold in n -dimensional space, $(\mathbf{p}, \mathbf{N}\mathbf{p}) = (\mathbf{p}_0, \mathbf{N}\mathbf{p}_0)$, on which the solution to the differential equation with $\mathbf{p}(0) = \mathbf{p}_0$ must fall. Simple discretizations, such as Euler’s method (1.1), do not typically respect such invariants, giving approximate solutions that drift from the manifold. Invariant-preserving alternatives (see, e.g., [18, 26]) generally require significantly more computation per step (though a tractable method for the unpreconditioned, Hermitian case has been proposed by Nakamura, Kajiwara, and Shiotani [28]). Our goal is to explain the relationship between convergence and stability of the continuous and discrete dynamical systems. In particular, the quadratic invariant is a crucial property of the continuous system, and plays an important role in the convergence theory of the corresponding discretization, even when that iteration does not preserve the invariant.

For a non-Hermitian problem, one naturally wonders how (1.1) can be modified to incorporate estimates of both left and right eigenvectors. In this case, we obtain

the coupled iteration (given here without preconditioning)

$$(1.4) \quad \begin{cases} \dot{\mathbf{p}} = \mathbf{p}\theta - \mathbf{A}\mathbf{p}, \\ \dot{\mathbf{q}} = \mathbf{q}\bar{\theta} - \mathbf{A}^*\mathbf{q}, \end{cases} \quad \theta = \frac{(\mathbf{A}\mathbf{p}, \mathbf{q})}{(\mathbf{p}, \mathbf{q})},$$

and a simple derivation reveals that (\mathbf{p}, \mathbf{q}) is invariant. Our analysis demonstrates that this two-sided dynamical system often suffers from finite-time blowup; in the discrete scheme this is tantamount to incurable breakdown, a well-known ailment of oblique projection methods (see [5] for a discussion and references to the literature within the context of non-Hermitian Lanczos methods).

A longstanding association exists between eigenvalue iterations and differential equations [1, 2, 3, 10, 11, 15, 19], often involving the observation that iterates of a particular eigenvalue algorithm are *exactly* discrete-time samples of some underlying continuous-time system. Notable examples include Rayleigh quotient gradient flow [10, 27], connections between the QR algorithm for dense eigenproblems and Toda flow [29, 39], and more general “isospectral flows” [42]. For example, Chu notes that the iterates of the standard power method can be obtained as integer-time samples of the solution to the system (1.2) with $\mathbf{N} = \mathbf{I}$ and \mathbf{A} replaced by $\log \mathbf{A}$ [10, eq. (2.7)].

The present study draws upon this body of work, but takes a different perspective: we seek a better understanding of iterations such as (1.1) that provide only *approximate* solutions (with a truncation error due to discretization) to continuous time systems such as (1.2). The distinction is significant: for example, a continuous-time generalization of the power method will converge, with mild caveats, to the largest magnitude eigenvalue, whereas the related systems we study can potentially converge to the leftmost eigenvalue at a shift-independent rate with little more work per iteration than the power method; see Theorems 4.4 and 6.3.

The connection between eigensolvers and continuous-time dynamical systems also arises in applications. For example, the Car–Parrinello method [8] determines the Kohn–Sham eigenstates from a second-order ordinary differential equation, Newton’s equations of motion (see [34, p. 1086] for a formulation using (1.2) with no preconditioning). The heavy ball optimization method [35] also formulates the minimum of the Rayleigh quotient via a second order ordinary differential equation. In [7], the ground state solution of Bose–Einstein condensates are determined via a normalized gradient flow discretized by several time integration schemes. (Both the Kohn–Sham eigenstates and Bose–Einstein condensates give rise to self-adjoint nonlinear eigenvalue problems.)

We begin our investigation with a study of various unpreconditioned iterations ($\mathbf{N} = \mathbf{I}$). Section 2 introduces basic differential equations for computation of invariant subspaces of matrix pencils, and then identifies parameter choices that yield invariant-preserving iterations. Near steady states, the solutions to these systems can be viewed as exact invariant subspaces for nearby matrices, as observed in section 3. From this point we focus on single vector iterations for standard eigenvalue problems. Section 4 describes exact solution formulas for two unpreconditioned continuous-time systems, one-sided and two-sided methods. As such exact solutions for the preconditioned case are elusive, we analyze such systems asymptotically using center manifold theory in section 5. These two sections provide the foundation for the main result of section 6, the development of sufficient conditions for convergence of (1.1) for non-Hermitian matrices.

2. Dynamical systems and invariant manifolds. We first examine properties of the dynamical system (1.2) and various generalizations suitable for computing

eigenvalues of non-Hermitian matrix pencils. Let $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{n \times n}$ be general matrices with fixed (time-invariant) entries. For the generalized eigenvalue problem $\mathbf{A}\mathbf{x} = \mathbf{B}\mathbf{x}\lambda$ with $\mathbf{N} = \mathbf{I}$, the system (1.2) expands to

$$\dot{\mathbf{p}} = \mathbf{B}\mathbf{p}\theta - \mathbf{A}\mathbf{p}$$

for appropriate $\theta = \theta(t)$. This equation suggests a generalization from a system with the single vector $\mathbf{p} \in \mathbb{C}^n$ to a system that evolves an entire subspace, given by the range of a matrix $\mathbf{P} \in \mathbb{C}^{n \times k}$:

$$\dot{\mathbf{P}} = \mathbf{B}\mathbf{P}\mathbf{L} - \mathbf{A}\mathbf{P},$$

where differentiation is still with respect to the autonomous variable t ; we shall address the choice of $\mathbf{L}(t) \in \mathbb{C}^{k \times k}$ momentarily. (Quantities such as \mathbf{L} are t -dependent unless explicitly stated otherwise; we typically suppress the t argument to simplify notation.)

For non-Hermitian problems one might simultaneously evolve an equation for the adjoint to obtain approximations to the left eigenspace, which suggests the system

$$(2.1) \quad \begin{aligned} \dot{\mathbf{P}} &= \mathbf{B}\mathbf{P}\mathbf{L} - \mathbf{A}\mathbf{P} \\ \dot{\mathbf{Q}} &= \mathbf{B}^*\mathbf{Q}\mathbf{M}^* - \mathbf{A}^*\mathbf{Q}, \end{aligned}$$

with initial conditions $\mathbf{P}(0) = \mathbf{P}_0$ and $\mathbf{Q}(0) = \mathbf{Q}_0$, where $\mathbf{P}, \mathbf{Q} \in \mathbb{C}^{n \times k}$, and $\mathbf{L}, \mathbf{M} \in \mathbb{C}^{k \times k}$. The choice we make for the time-dependent $\mathbf{L}, \mathbf{M} \in \mathbb{C}^{k \times k}$ can potentially couple \mathbf{P} and \mathbf{Q} as introduced in (1.4). Here \cdot^* denotes the conjugate transpose and (\cdot, \cdot) the standard Euclidean inner product (though this analysis generalizes readily to arbitrary inner products). If this system is at a steady state, i.e., $\dot{\mathbf{P}} = \dot{\mathbf{Q}} = \mathbf{0}$, then

$$(2.2) \quad \mathbf{B}\mathbf{P}\mathbf{L} = \mathbf{A}\mathbf{P}, \quad \mathbf{B}^*\mathbf{Q}\mathbf{M}^* = \mathbf{A}^*\mathbf{Q},$$

and, hence, provided \mathbf{P} and \mathbf{Q} have full column rank, the eigenvalues of \mathbf{L} and \mathbf{M} are included in the spectrum of the pencil $\mathbf{A} - \lambda\mathbf{B}$, while the columns of \mathbf{P} and \mathbf{Q} span right- and left-invariant subspaces of the same pencil. We shall motivate the choice of \mathbf{L} and \mathbf{M} through generalizations of the invariant discussed in the introduction. The following notation facilitates the analysis of these subspace iterations.

DEFINITION 2.1. *Given $\mathbf{P}, \mathbf{Q} \in \mathbb{C}^{n \times k}$, define $(\mathbf{P}, \mathbf{Q}) = \mathbf{Q}^*\mathbf{P} \in \mathbb{C}^{k \times k}$; i.e., the (i, j) entry of (\mathbf{P}, \mathbf{Q}) satisfies $(\mathbf{P}, \mathbf{Q})_{i,j} := (\mathbf{P}\mathbf{e}_j, \mathbf{Q}\mathbf{e}_i)$, where \mathbf{e}_ℓ denotes the ℓ th column of the $k \times k$ identity matrix.*

In this notation, we have the homogeneity property $(\mathbf{P}\mathbf{L}, \mathbf{Q}) = \mathbf{Q}^*\mathbf{P}\mathbf{L} = (\mathbf{P}, \mathbf{Q})\mathbf{L}$.

Consider the pairs of (time-dependent) functions

$$(2.3) \quad (\mathbf{Q}, \mathbf{P}), \quad (\mathbf{P}, \mathbf{Q}) \quad \text{and} \quad (\mathbf{P}, \mathbf{P}), \quad (\mathbf{Q}, \mathbf{Q})$$

with derivatives

$$\frac{d}{dt}(\mathbf{Q}, \mathbf{P}) = (\dot{\mathbf{Q}}, \mathbf{P}) + (\mathbf{Q}, \dot{\mathbf{P}}), \quad \frac{d}{dt}(\mathbf{P}, \mathbf{Q}) = (\dot{\mathbf{P}}, \mathbf{Q}) + (\mathbf{P}, \dot{\mathbf{Q}}),$$

and

$$\frac{d}{dt}(\mathbf{P}, \mathbf{P}) = (\dot{\mathbf{P}}, \mathbf{P}) + (\mathbf{P}, \dot{\mathbf{P}}), \quad \frac{d}{dt}(\mathbf{Q}, \mathbf{Q}) = (\dot{\mathbf{Q}}, \mathbf{Q}) + (\mathbf{Q}, \dot{\mathbf{Q}}).$$

Inspired by (1.3), we next investigate how best to choose \mathbf{L} and \mathbf{M} to make either pair in (2.3) invariant under the system (2.1).

THEOREM 2.2. *For the system of ordinary differential equations (2.1) with initial conditions $\mathbf{P}(0) = \mathbf{P}_0 \in \mathbb{C}^{n \times k}$ and $\mathbf{Q}(0) = \mathbf{Q}_0 \in \mathbb{C}^{n \times k}$, the choices*

$$(2.4) \quad \mathbf{L} = (\mathbf{B}\mathbf{P}, \mathbf{Q})^{-1}(\mathbf{A}\mathbf{P}, \mathbf{Q}), \quad \mathbf{M}^* = (\mathbf{Q}, \mathbf{B}\mathbf{P})^{-1}(\mathbf{Q}, \mathbf{A}\mathbf{P})$$

give

$$\frac{d}{dt}(\mathbf{P}, \mathbf{Q}) = \frac{d}{dt}(\mathbf{Q}, \mathbf{P}) = \mathbf{0},$$

and, hence, $(\mathbf{P}, \mathbf{Q}) = (\mathbf{P}_0, \mathbf{Q}_0)$ and $(\mathbf{Q}, \mathbf{P}) = (\mathbf{Q}_0, \mathbf{P}_0)$ hold for all t .

Proof. Note that

$$\begin{aligned} \frac{d}{dt}(\mathbf{P}, \mathbf{Q}) &= (\dot{\mathbf{P}}, \mathbf{Q}) + (\mathbf{P}, \dot{\mathbf{Q}}) \\ &= (\mathbf{B}\mathbf{P}, \mathbf{Q})\mathbf{L} - (\mathbf{A}\mathbf{P}, \mathbf{Q}) + \mathbf{M}(\mathbf{P}, \mathbf{B}^*\mathbf{Q}) - (\mathbf{P}, \mathbf{A}^*\mathbf{Q}) \\ \left(\frac{d}{dt}(\mathbf{Q}, \mathbf{P})\right)^* &= (\mathbf{P}, \dot{\mathbf{Q}}) + (\dot{\mathbf{P}}, \mathbf{Q}) \\ &= \mathbf{M}(\mathbf{P}, \mathbf{B}^*\mathbf{Q}) - (\mathbf{P}, \mathbf{A}^*\mathbf{Q}) + (\mathbf{B}\mathbf{P}, \mathbf{Q})\mathbf{L} - (\mathbf{A}\mathbf{P}, \mathbf{Q}), \end{aligned}$$

where we have used (2.1) and the homogeneity property. We can force $(d/dt)(\mathbf{P}, \mathbf{Q})$ to zero by setting \mathbf{L} and \mathbf{M} as in (2.4). \square

The next result is a direct analogue of Theorem 2.2 for the second pair in (2.3). We omit the proof, a minor adaptation of the last one.

THEOREM 2.3. *For the system of ordinary differential equations (2.1) with initial conditions $\mathbf{P}(0) = \mathbf{P}_0 \in \mathbb{C}^{n \times k}$ and $\mathbf{Q}(0) = \mathbf{Q}_0 \in \mathbb{C}^{n \times k}$, the choices*

$$\mathbf{L} = (\mathbf{B}\mathbf{P}, \mathbf{P})^{-1}(\mathbf{A}\mathbf{P}, \mathbf{P}), \quad \mathbf{M}^* = (\mathbf{Q}, \mathbf{B}\mathbf{Q})^{-1}(\mathbf{Q}, \mathbf{A}\mathbf{Q})$$

give

$$\frac{d}{dt}(\mathbf{P}, \mathbf{P}) = \frac{d}{dt}(\mathbf{Q}, \mathbf{Q}) = \mathbf{0},$$

and, hence, $(\mathbf{P}, \mathbf{P}) = (\mathbf{P}_0, \mathbf{P}_0)$ and $(\mathbf{Q}, \mathbf{Q}) = (\mathbf{Q}_0, \mathbf{Q}_0)$ for all t .

The formulations for \mathbf{L} and \mathbf{M} given in Theorems 2.2 and 2.3 are known as *generalized Rayleigh quotients* [38]. With these values of \mathbf{L} and \mathbf{M} , we refer to (2.1) as the *two-sided* and *one-sided* dynamical systems. Theorem 2.2 shows that if $\mathbf{P}_0^* \mathbf{Q}_0 = \mathbf{I}$, then the two-sided solutions will preserve this property (allowing for biorthogonal bases for left and right invariant subspaces), though possibly at the expense of growing $\|\mathbf{P}\|$ or $\|\mathbf{Q}\|$. Theorem 2.3, on the other hand, shows that the one-sided iteration maintains $\|\mathbf{P}\|$ and $\|\mathbf{Q}\|$, though biorthogonality will generally be lost. From the invariants we also see that the system preserves the rank of solutions to both one- and two-sided equations—provided they exist (see section 4). Since (\mathbf{P}, \mathbf{P}) is fixed for the one-sided system, so too are all singular values (and, thus, the rank) of \mathbf{P} . For the two-sided system, if $(\mathbf{P}_0, \mathbf{Q}_0)$ is full rank, (\mathbf{P}, \mathbf{Q}) must always be as well, and, hence, \mathbf{P} and \mathbf{Q} individually have full rank.

We denote the dynamical systems (2.1) given the generalized Rayleigh quotients of Theorems 2.2 and 2.3 as “two-sided” and “one-sided”, respectively. We refer to the ensuing schemes that result from discretizing (2.1) as “two-sided” and “one-sided” iterations.

3. Invariants and backward stability. We saw in (2.2) that, at a steady state, the eigenvalues of \mathbf{L} and \mathbf{M} are exact eigenvalues of the pencil $\mathbf{A} - \lambda\mathbf{B}$. As the system *approaches* a steady state, how well do the eigenvalues of the invariant-preserving choices for \mathbf{L} and \mathbf{M} approximate the eigenvalues of the pencil?

First, consider the one-sided system, with \mathbf{L} as given in Theorem 2.3 and \mathbf{P} full rank. The first part of (2.1) can then be written as

$$\mathbf{0} = \mathbf{BPL} - \left(\mathbf{A} + \dot{\mathbf{P}}(\mathbf{P}, \mathbf{P})^{-1}\mathbf{P}^* \right) \mathbf{P},$$

from which we see that the eigenvalues of \mathbf{L} form a subset of the spectrum of the perturbed pencil $(\mathbf{A} + \dot{\mathbf{P}}(\mathbf{P}, \mathbf{P})^{-1}\mathbf{P}^*) - \lambda\mathbf{B}$. How large can such perturbations be? Note that $(\mathbf{P}, \mathbf{P})^{-1}\mathbf{P}^* = \mathbf{P}^+$ is the pseudoinverse of \mathbf{P} , and so

$$\left\| \dot{\mathbf{P}}(\mathbf{P}, \mathbf{P})^{-1}\mathbf{P}^* \right\| \leq \left\| \dot{\mathbf{P}} \right\| \left\| \mathbf{P}^+ \right\| = \frac{\left\| \dot{\mathbf{P}} \right\|}{\sigma_k},$$

where σ_k is the smallest singular value of $\mathbf{P} \in \mathbb{C}^{n \times k}$. As discussed at the end of section 2, the choice of \mathbf{L} in Theorem 2.3 that makes (\mathbf{P}, \mathbf{P}) invariant also makes σ_k invariant. Thus, when $\|\dot{\mathbf{P}}\|$ is small, i.e., near a steady state, we conclude that the eigenvalues of \mathbf{L} are the exact eigenvalues of a nearby pencil, with σ_k^{-1} acting as a condition number does in a backward error bound; that condition number can be set to one simply by taking $(\mathbf{P}_0, \mathbf{P}_0) = \mathbf{I}$. (This is related to an error bound for Rayleigh–Ritz eigenvalue estimates for a Hermitian matrix using a nonorthogonal basis; see [32, Theorem 11.10.1].) This analysis suggests that a departure from orthogonality in a numerical integration of the differential equation is reflected in degrading accuracy of the approximate eigenvalues.

Now consider the two-sided system with \mathbf{L} and \mathbf{M} as given by Theorem 2.2 with nonsingular $(\mathbf{BP}, \mathbf{Q})$. We wish to rewrite (2.1) in the form

$$\begin{aligned} \mathbf{0} &= \mathbf{BPL} - (\mathbf{A} + \mathbf{E})\mathbf{P} \\ \mathbf{0} &= \mathbf{B}^*\mathbf{QM}^* - (\mathbf{A}^* + \mathbf{E}^*)\mathbf{Q} \end{aligned}$$

for the same \mathbf{E} in both iterations. Lemma 1 of [20] implies that such a perturbation \mathbf{E} exists if and only if

$$(\mathbf{BP}, \mathbf{Q})\mathbf{L} = \mathbf{M}(\mathbf{BP}, \mathbf{Q}),$$

which holds for the choice of \mathbf{L} and \mathbf{M} given in Theorem 2.2. The perturbation \mathbf{E} is not unique, but $\mathbf{EP} = \dot{\mathbf{P}}$ and $\mathbf{E}^*\mathbf{Q} = \dot{\mathbf{Q}}$. Moreover, the “main theorem” of [20] gives

$$\min \|\mathbf{E}\|_2 = \max \left\{ \|\dot{\mathbf{P}}\|_2, \|\dot{\mathbf{Q}}\|_2 \right\}$$

if $(\mathbf{P}, \mathbf{P}) = \mathbf{I}_k$ and $(\mathbf{Q}, \mathbf{Q}) = \mathbf{I}_k$. However, as the authors of [20] explain, a small $\|\mathbf{E}\|_2$ is irrelevant unless $\|(\mathbf{P}, \mathbf{Q})^{-1}\|_2$ is also small. In particular, when \mathbf{P} is orthogonal to \mathbf{Q} , $\min \|\mathbf{E}\|_2$ is undefined. The discussion following Theorem 4.1 in subsection 4.1 explains that a large (or undefined) $\|(\mathbf{P}, \mathbf{Q})^{-1}\|_2$ is equivalent to near breakdown (or serious breakdown) of the two-sided dynamical system.

We caution the reader that backward stability alone does not provide information on forward error, or accuracy, of the steady-states when $\mathbf{A} \neq \mathbf{A}^*$. The relevance of backward stability is that the solution of our one- and two-sided systems are, at all times, steady-states for a related dynamical system. The distance to this related perturbed system depends upon the norm of the residuals.

4. Convergence analysis. At least for single-vector iterations (i.e., $k = 1$), the analysis of the one- and two-sided dynamical systems follows readily from the remarkable fact that, in many cases, simple formulas give the exact solutions of these nonlinear differential equations. This observation, inspired by a lemma of Nanda [29], informs convergence analysis of the eigeniterations that result from the discretization of these equations. Though expressed for the standard eigenvalue problem, these results can naturally be adapted to the generalized case by replacing \mathbf{A} with $\mathbf{B}^{-1}\mathbf{A}$. We discuss the solution operators for two-sided systems, followed by one-sided systems.

4.1. Two-sided systems. The following result generalizes a result of Nanda [29, Lemma 1.4] for the two-sided dynamical system.

THEOREM 4.1. *Consider the partitioned set of ordinary differential equations*

$$(4.1) \quad \begin{aligned} \dot{\mathbf{p}} &= \mathbf{p}\theta - \mathbf{A}\mathbf{p} \\ \dot{\mathbf{q}} &= \mathbf{q}\bar{\theta} - \mathbf{A}^*\mathbf{q}, \end{aligned}$$

with $\mathbf{p}(0) = \mathbf{p}_0$ and $\mathbf{q}(0) = \mathbf{q}_0$, where $\mathbf{p}, \mathbf{q} \in \mathbb{C}^n$, $(\mathbf{p}_0, \mathbf{q}_0) \neq 0$, and

$$\theta = \frac{(\mathbf{A}\mathbf{p}, \mathbf{q})}{(\mathbf{p}, \mathbf{q})}.$$

Then there exists some $t_f > 0$ such that for all $t \in [0, t_f)$,

$$\mathbf{p}(t) = e^{-\mathbf{A}t}\mathbf{p}_0\pi(t), \quad \mathbf{q}(t) = e^{-\mathbf{A}^*t}\mathbf{q}_0\overline{\pi(t)},$$

where

$$(4.2) \quad \pi(t) = \sqrt{\frac{(\mathbf{p}_0, \mathbf{q}_0)}{(e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}^*t}\mathbf{q}_0)}}.$$

Proof. We define $\mathbf{p}(t) = e^{-\mathbf{A}t}\mathbf{p}_0\pi(t)$ and $\mathbf{q}(t) = e^{-\mathbf{A}^*t}\mathbf{q}_0\overline{\pi(t)}$, and will show that these formulas satisfy the system (4.1). Note that

$$\begin{aligned} \dot{\pi} &= \frac{\pi}{2} \frac{((\mathbf{A}e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}^*t}\mathbf{q}_0) + (e^{-\mathbf{A}t}\mathbf{p}_0, \mathbf{A}^*e^{-\mathbf{A}^*t}\mathbf{q}_0))}{(e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}^*t}\mathbf{q}_0)} \\ &= \pi \frac{(\mathbf{A}e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}^*t}\mathbf{q}_0)}{(e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}^*t}\mathbf{q}_0)} \\ &= \pi \frac{(\mathbf{A}e^{-\mathbf{A}t}\mathbf{p}_0\pi, e^{-\mathbf{A}^*t}\mathbf{q}_0\bar{\pi})}{(e^{-\mathbf{A}t}\mathbf{p}_0\pi, e^{-\mathbf{A}^*t}\mathbf{q}_0\bar{\pi})} = \pi \frac{(\mathbf{A}\mathbf{p}, \mathbf{q})}{(\mathbf{p}, \mathbf{q})} = \pi\theta. \end{aligned}$$

Differentiating the formulas for \mathbf{p} and \mathbf{q} , thus, gives

$$\begin{aligned} \dot{\mathbf{p}} &= -\mathbf{A}e^{-\mathbf{A}t}\mathbf{p}_0\pi + e^{-\mathbf{A}t}\mathbf{p}_0\dot{\pi} = -\mathbf{A}\mathbf{p} + \theta\mathbf{p} \\ \dot{\mathbf{q}} &= -\mathbf{A}^*e^{-\mathbf{A}^*t}\mathbf{q}_0\bar{\pi} + e^{-\mathbf{A}^*t}\mathbf{q}_0\dot{\bar{\pi}} = -\mathbf{A}^*\mathbf{q} + \bar{\theta}\mathbf{q}, \end{aligned}$$

as required. The hypothesis that $(\mathbf{p}_0, \mathbf{q}_0) \neq 0$ ensures the existence of the solution at time $t = 0$. The formula will hold for all $t > 0$, until potentially

$$(4.3) \quad (e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}^*t}\mathbf{q}_0) = 0.$$

We define t_f to be the smallest positive t for which (4.3) holds. If no such positive t exists, the solution exists for all $t > 0$ and we can take $t_f = \infty$ in the statement of the theorem. \square

Theorem 4.1 gives $(\mathbf{p}, \mathbf{q}) = (\mathbf{p}_0, \mathbf{q}_0)$, precisely as Theorem 2.2 indicates. Under the conditions of Theorem 4.1, solutions of the two-sided single-vector equations (4.1) have the same direction as solutions of the simpler linear systems $\dot{\mathbf{x}} = -\mathbf{A}\mathbf{x}$, $\mathbf{x}(0) = \mathbf{p}_0$ and $\dot{\mathbf{y}} = -\mathbf{A}^*\mathbf{y}$, $\mathbf{y}(0) = \mathbf{q}_0$, but the magnitudes of \mathbf{p} and \mathbf{q} vary nonlinearly with (4.2). In particular, the inner product of \mathbf{p} and \mathbf{q} can be zero—even with both \mathbf{p} and \mathbf{q} nonzero—leading to finite time blow-up of (4.1). Note that if

$$\left(\frac{e^{-\mathbf{A}t}\mathbf{p}_0}{\sqrt{(\mathbf{p}_0, \mathbf{q}_0)}}, \frac{e^{-\mathbf{A}^*t}\mathbf{q}_0}{\sqrt{(\mathbf{q}_0, \mathbf{p}_0)}} \right) = 0,$$

then $\pi(t)$ is undefined. Hence, finite time blow-up is analogous to *serious breakdown* [43, p. 389], a problem endemic to oblique projection methods (see, e.g., [5]). This ratio will be nonzero but small in the vicinity of blow-up (or *near-breakdown*), a situation that commonly occurs in discretizations of these equations. The salient issue is that \mathbf{p} and \mathbf{q} are nearly orthogonal and so

$$(4.4) \quad \frac{(\mathbf{p}, \mathbf{q})}{\|\mathbf{p}\| \|\mathbf{q}\|} = \left(\frac{e^{-\mathbf{A}t}\mathbf{p}_0}{\|e^{-\mathbf{A}t}\mathbf{p}_0\|}, \frac{e^{-\mathbf{A}^*t}\mathbf{q}_0}{\|e^{-\mathbf{A}^*t}\mathbf{q}_0\|} \right)$$

is a useful quantity to measure. This number is small when the secant of the angle between \mathbf{p} and \mathbf{q} is large. In section 6 we shall see the important consequences of these observations for eigensolvers derived from the discretization of (4.1).

One can avoid breakdown altogether by using starting vectors \mathbf{p}_0 and \mathbf{q}_0 that are sufficiently accurate approximations to the right and left eigenvectors of \mathbf{A} associated with the leftmost eigenvalue. Suppose \mathbf{A} is diagonalizable with a simple leftmost eigenvalue λ_1 , and all other eigenvalues strictly to the right of λ_1 . Thus, there exists invertible \mathbf{X} and diagonal $\mathbf{\Lambda}$ such that

$$\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$$

with $\mathbf{\Lambda}_{1,1} = \lambda_1$. Write $\lambda_j = \mathbf{\Lambda}_{j,j}$, so that $\text{Re } \lambda_j > \text{Re } \lambda_1$ for $j = 2, \dots, n$. Define $\mathbf{r} = \mathbf{X}^{-1}\mathbf{p}_0$ and $\mathbf{s} = \mathbf{X}^*\mathbf{q}_0$; i.e., \mathbf{r} and \mathbf{s} are the expansions of the starting vectors in biorthogonal bases of right and left eigenvectors of \mathbf{A} .

THEOREM 4.2. *Under the setting established in the last paragraph, the condition*

$$|r_1 s_1| > \sum_{j=2}^n |r_j s_j|$$

is sufficient to ensure that the dynamical system (4.1) has a solution for all $t \geq 0$ given by Theorem 4.1; i.e., no incurable breakdown occurs.

Proof. First note that

$$(e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}^*t}\mathbf{q}_0) = (\mathbf{X}e^{-\mathbf{\Lambda}t}\mathbf{X}^{-1}\mathbf{p}_0, \mathbf{X}^{-*}e^{-\mathbf{\Lambda}^*t}\mathbf{X}^*\mathbf{q}_0) = (e^{-2\mathbf{\Lambda}t}\mathbf{r}, \mathbf{s}) = \sum_{j=1}^n r_j \bar{s}_j e^{-2\lambda_j t}.$$

Since $\text{Re } \lambda_1 < \text{Re } \lambda_j$ for $j > 2$, we have $|e^{-2\lambda_1 t}| \geq |e^{-2\lambda_j t}|$ for all $t \geq 0$. The hypothesis involving \mathbf{r} and \mathbf{s} , thus, implies, for $t \geq 0$, that

$$|r_1 s_1 e^{-2\lambda_1 t}| \geq \sum_{j=2}^n |r_j s_j e^{-2\lambda_j t}|.$$

Given this expression, we can twice apply the triangle inequality to conclude

$$\begin{aligned} 0 &< |r_1 \bar{s}_1 e^{-2\lambda_1 t}| - \sum_{j=2}^n |r_j \bar{s}_j e^{-2\lambda_j t}| \\ &\leq |r_1 \bar{s}_1 e^{-2\lambda_1 t}| - \left| \sum_{j=2}^n r_j \bar{s}_j e^{-2\lambda_j t} \right| \leq \left| \sum_{j=1}^n r_j \bar{s}_j e^{-2\lambda_j t} \right| = \left| \left(e^{-\mathbf{A}t} \mathbf{p}_0, e^{-\mathbf{A}^*t} \mathbf{q}_0 \right) \right|. \end{aligned}$$

Hence, $\pi(t)$ in Theorem 4.1 is finite for all $t \geq 0$, ensuring that the solution to the dynamical system (4.1) does not blow up at finite time. \square

Theorem 4.2 implies that finite-time blow-up (or serious breakdown) is not generic for (4.1). However, the sufficient condition provided suggests that excellent initial approximations to the leftmost (left and right) eigenvectors are needed.

4.2. One-sided systems. The single vector one-sided system possesses a similar exact solution, which has been studied in the context of gradient flows associated with Rayleigh quotient iteration. We shall see that finite-time blow-up is never a concern for such systems. The following is a modest restatement of a result of Nanda [29, Lemma 1.4] (who considers the differential equation acting on the unit ball in \mathbb{R}^n).

THEOREM 4.3. *Consider the ordinary differential equation*

$$(4.5) \quad \dot{\mathbf{p}} = \mathbf{p}\theta - \mathbf{A}\mathbf{p},$$

with $\mathbf{A} \in \mathbb{R}^{n \times n}$ and initial condition $\mathbf{p}(0) = \mathbf{p}_0 \in \mathbb{R}^n$, where $\mathbf{p}_0 \neq \mathbf{0}$ and

$$\theta = \frac{(\mathbf{A}\mathbf{p}, \mathbf{p})}{(\mathbf{p}, \mathbf{p})}.$$

Then for all $t \geq 0$, (4.5) has the exact solution

$$\mathbf{p}(t) = e^{-\mathbf{A}t} \mathbf{p}_0 \omega(t),$$

where

$$\omega(t) = \sqrt{\frac{(\mathbf{p}_0, \mathbf{p}_0)}{(e^{-\mathbf{A}t} \mathbf{p}_0, e^{-\mathbf{A}t} \mathbf{p}_0)}}.$$

We omit the proof of this result, which closely mimics that of Theorem 4.1. Of course, a similar formula can be written for the one-sided equation for $\mathbf{q}(t)$. The restriction to real matrices guarantees that $(\mathbf{A}e^{-\mathbf{A}t} \mathbf{p}_0, e^{-\mathbf{A}t} \mathbf{p}_0) = (e^{-\mathbf{A}t} \mathbf{p}_0, \mathbf{A}e^{-\mathbf{A}t} \mathbf{p}_0)$; the result also holds for complex Hermitian \mathbf{A} .

As before, \mathbf{p} has the same direction as the solution to the dynamical system $\dot{\mathbf{x}} = -\mathbf{A}\mathbf{x}$ with $\mathbf{x}(0) = \mathbf{p}_0$, but the magnitude is scaled by the nonlinear scalar ω . Provided $\mathbf{p}_0 \neq \mathbf{0}$, the one-sided system (4.5) cannot blow up in finite time, since $(\mathbf{p}, \mathbf{p}) \neq 0$, in stark contrast to the two-sided iteration. This collinearity implies that the \mathbf{p} vectors produced by the one- and two-sided systems provide equally accurate approximations to the desired eigenvector, at least until the latter breaks down.

When \mathbf{A} has a unique simple eigenvalue of smallest real part and the hypotheses of Theorem 4.1 or 4.3 are met, the asymptotic analysis of the associated dynamical system readily follows; cf. [19, section 1.3] for a generic asymptotic linear stability

analysis of the one-sided iteration. In fact, one can develop explicit bounds on the sine of the angle between \mathbf{p} and the desired eigenvector \mathbf{x}_1 , defined as

$$\sin \angle(\mathbf{p}, \mathbf{x}_1) := \min_{\alpha \in \mathbb{C}} \frac{\|\alpha \mathbf{p} - \mathbf{x}_1\|}{\|\mathbf{x}_1\|}.$$

THEOREM 4.4. *Suppose \mathbf{A} can be diagonalized, $\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$, and the eigenvalues of \mathbf{A} can be ordered as*

$$\text{Real}(\lambda_1) < \text{Real}(\lambda_2) \leq \dots \leq \text{Real}(\lambda_n).$$

Let \mathbf{x}_1 and \mathbf{y}_1 denote right and left eigenvectors associated with λ_1 , with $\|\mathbf{x}_1\| = 1$ and $\mathbf{y}_1^ \mathbf{x}_1 = 1$. Then the solution $\mathbf{p}(t)$ to both systems (4.1) and (4.5) satisfies*

$$\sin \angle(\mathbf{p}(t), \mathbf{x}_1) \leq \|\mathbf{X}\| \|\mathbf{X}^{-1}\| \frac{\|\mathbf{p}_0\|}{|\mathbf{y}_1^* \mathbf{p}_0|} e^{\text{Re}(\lambda_1 - \lambda_2)t}$$

for all $t \geq 0$ in the case of (4.5), and for all $t \in [0, t_f]$ in the case of (4.1).

Proof. Since \mathbf{x}_1 is a unit vector, we can write

$$\sin \angle(\mathbf{p}(t), \mathbf{x}_1) = \min_{\alpha \in \mathbb{C}} \|\alpha \mathbf{p}(t) - \mathbf{x}_1\|.$$

In both (4.5) and (4.1), $\mathbf{p}(t)$ is collinear with $e^{-\mathbf{A}t} \mathbf{p}_0$, so we can proceed with

$$\begin{aligned} \sin \angle(\mathbf{p}(t), \mathbf{x}_1) &= \min_{\alpha \in \mathbb{C}} \|\alpha \mathbf{X} e^{-\mathbf{A}t} \mathbf{X}^{-1} \mathbf{p}_0 - \mathbf{x}_1\| \\ &\leq \left\| \frac{e^{\lambda_1 t}}{\mathbf{y}_1^* \mathbf{p}_0} \mathbf{X} e^{-\mathbf{A}t} \mathbf{X}^{-1} \mathbf{p}_0 - \mathbf{x}_1 \right\| \leq \|\mathbf{X}\| \|\mathbf{X}^{-1}\| \frac{\|\mathbf{p}_0\|}{|\mathbf{y}_1^* \mathbf{p}_0|} e^{\text{Re}(\lambda_1 - \lambda_2)t}. \end{aligned}$$

The first inequality follows from choosing a (suboptimal) value of α that cancels the terms in the \mathbf{x}_1 direction. (For similar analysis of the Arnoldi eigenvalue iteration, see [37, Proposition 2.1].) \square

An analogous bound could be developed for the convergence of \mathbf{q} to the left eigenvector \mathbf{y}_1 . When \mathbf{A} is far from normal, one typically observes a transient stage of convergence that could be better described via analysis that avoids the diagonalization of \mathbf{A} ; see, e.g., [40, section 28], which includes similar analysis for the power method.

The two-sided iteration converges to left and right eigenvectors of \mathbf{A} associated with the leftmost eigenvalue, *provided the method does not breakdown on the way to this limit*. Several natural questions arise: How common is breakdown? How well do discretizations mimic this dynamical system? Before investigating these issues in section 6, we first address how preconditioning can accelerate—and complicate—the convergence of these continuous-time systems.

5. Preconditioned dynamical systems. What does it mean to precondition the eigenvalue problem? Several different strategies have been proposed in the literature (see especially the discussion in [21, pp. 109–110]); here we shall investigate analogous approaches for our continuous time dynamical systems, and the implications such modifications have on the convergence behavior described in the last section.

One might first consider applying to the generalized eigenvalue problem

$$\mathbf{A}\mathbf{p} = \mathbf{B}\mathbf{p}\lambda,$$

left and right preconditioners \mathbf{M} and \mathbf{N} , so as to obtain the equivalent pencil

$$(5.1) \quad (\mathbf{M}^{-1}\mathbf{A}\mathbf{N}) (\mathbf{N}^{-1}\mathbf{p}) = (\mathbf{M}^{-1}\mathbf{B}\mathbf{N}) (\mathbf{N}^{-1}\mathbf{p}) \lambda.$$

Provided \mathbf{B} is invertible, one could then define

$$\begin{aligned}\widehat{\mathbf{A}} &:= (\mathbf{M}^{-1}\mathbf{B}\mathbf{N})^{-1} (\mathbf{M}^{-1}\mathbf{A}\mathbf{N}) = \mathbf{N}^{-1}\mathbf{B}^{-1}\mathbf{A}\mathbf{N} \\ \widehat{\mathbf{p}} &:= \mathbf{N}^{-1}\mathbf{p},\end{aligned}$$

then apply the concepts from the preceding sections to the standard eigenvalue problem $\widehat{\mathbf{A}}\widehat{\mathbf{p}} = \widehat{\mathbf{p}}\lambda$. For example, we could seek the leftmost eigenpair of $\widehat{\mathbf{A}}$ by evolving the dynamical system

$$\dot{\widehat{\mathbf{p}}} = \widehat{\mathbf{p}}\widehat{\theta} - \widehat{\mathbf{A}}\widehat{\mathbf{p}},$$

with the (preconditioned) Rayleigh quotient

$$\widehat{\theta} = \frac{(\widehat{\mathbf{A}}\widehat{\mathbf{p}}, \widehat{\mathbf{p}})}{(\widehat{\mathbf{p}}, \widehat{\mathbf{p}})} = \frac{(\mathbf{N}^{-1}\mathbf{B}^{-1}\mathbf{A}\mathbf{p}, \mathbf{N}^{-1}\mathbf{p})}{(\mathbf{N}^{-1}\mathbf{p}, \mathbf{N}^{-1}\mathbf{p})}.$$

Note that $\widehat{\mathbf{A}}$ and $\mathbf{B}^{-1}\mathbf{A}$ share the same spectrum because they are similar, and, hence, the asymptotic rate in Theorem 4.4 is immune to the preconditioner. The application of \mathbf{N} could affect the system’s transient behavior, but \mathbf{M} exerts no influence at all.¹

Several choices for \mathbf{N} are interesting. Taking $\mathbf{N} = \mathbf{A}^{-1}$ gives $\widehat{\mathbf{A}} = \mathbf{A}\mathbf{B}^{-1}$, an alternative to the $\mathbf{B}^{-1}\mathbf{A}$ form suggested by the original problem. Similarity transformations can also be used to *balance* a matrix to improve the conditioning of the eigenvalue problem [31, 33], in which case \mathbf{N} is constructed as a diagonal matrix that reduces the norm of $\widehat{\mathbf{A}}$. Such balancing tends to decrease the departure from normality associated with the largest magnitude eigenvalues. In fact, in the 1960 article that introduced this idea, Osborne refers to this procedure as “pre-conditioning” [31]. A more extreme—if impractical—approach takes \mathbf{N} to be a matrix that diagonalizes $\mathbf{B}^{-1}\mathbf{A}$ (provided such a matrix exists), a choice that minimizes the constant $\|\mathbf{X}\|\|\mathbf{X}^{-1}\|$ that describes the departure from normality in Theorem 4.4.

As useful as such improvements might be, these strategies fail to alter the asymptotic convergence rate described in Theorem 4.4. To potentially improve this rate, one can apply the preconditioner \mathbf{N}^{-1} directly to the residual $\mathbf{p}\theta - \mathbf{A}\mathbf{p}$. Consider the dynamical system

$$(5.2) \quad \dot{\mathbf{p}} = \mathbf{N}^{-1}(\mathbf{p}\theta - \mathbf{A}\mathbf{p}),$$

where θ refers to the usual (unpreconditioned) Rayleigh quotient $\theta = (\mathbf{A}\mathbf{p}, \mathbf{p})/(\mathbf{p}, \mathbf{p})$. Discretization of this system results in the familiar preconditioned eigensolver described in (1.1). For this case, a generalization of Theorem 4.3 has proved elusive; we have found no closed form for the exact solution. Indeed, as we shall next see, the choice of preconditioner can even complicate the system’s local behavior.

Let \mathbf{x}_1 denote a unit eigenvector of \mathbf{A} associated with the eigenvalue λ_1 . Note that \mathbf{x}_1 is a steady-state of (5.2), linearizing about which gives the Jacobian

$$(5.3) \quad \mathbf{J} = \mathbf{N}^{-1}(\mathbf{I} - \mathbf{x}_1\mathbf{x}_1^*)(\lambda_1 - \mathbf{A}).$$

As $\mathbf{J}\mathbf{x}_1 = \mathbf{0}$, the Jacobian \mathbf{J} always has a zero eigenvalue, adding complexity to conventional linear stability analysis. The challenge can be magnified by a poor

¹Alternatively, by substituting $(\mathbf{M}^{-1}\mathbf{B}\mathbf{N})^{-1}\widehat{\mathbf{p}} := \mathbf{N}^{-1}\mathbf{p}$ in (5.1), we obtain a system driven by $\widetilde{\mathbf{A}} = \mathbf{M}^{-1}\mathbf{A}\mathbf{B}^{-1}\mathbf{M}$ that is independent of \mathbf{N} .

choice for \mathbf{N} . For example, suppose

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad \mathbf{N} = \mathbf{N}^{-1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \lambda_1 = 1,$$

so that

$$\mathbf{J} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix};$$

i.e., the Jacobian is a Jordan block with a double eigenvalue at zero.

To obtain a rough impression of the behavior of the continuous system when θ is in the vicinity of λ_1 , consider the constant-coefficient equation $\dot{\mathbf{p}} = \mathbf{N}^{-1}(\mathbf{p}\lambda_1 - \mathbf{A}\mathbf{p})$, whose solution obeys the simple formula

$$\mathbf{p}(t) = e^{\mathbf{N}^{-1}(\lambda_1 - \mathbf{A})t} \mathbf{p}(0).$$

Hence, the asymptotic behavior of \mathbf{p} is controlled by the spectrum of $\mathbf{N}^{-1}(\lambda_1 - \mathbf{A})$. Assuming that $\mathbf{N}^{-1}(\lambda_1 - \mathbf{A})$ has a simple zero eigenvalue, the convergence of this system to the dominant eigenvector depends on the nonzero eigenvalues of $\mathbf{N}^{-1}(\lambda_1 - \mathbf{A})$: if this matrix has any other eigenvalues in the closed right half plane, the system will not generically converge; if all nonzero eigenvalues are in the open left half plane, then the convergence rate will be determined by the rightmost of them.

Specific choices for \mathbf{N}^{-1} will naturally depend significantly on the application problem at hand; in our general setting we seek to characterize basic traits of effective preconditioners. From the perspective of the convergence rate of the continuous dynamical system, we seek a preconditioner \mathbf{N}^{-1} such that the nonzero eigenvalues of $\mathbf{N}^{-1}(\lambda_1 - \mathbf{A})$ are as far to the left as possible. While the leftmost eigenvalues of $\mathbf{N}^{-1}(\lambda_1 - \mathbf{A})$ do not much affect the behavior of the continuous system, they can have a significant effect on the stability of the discretized difference equation, i.e., the related eigensolvers. For example, if $\mathbf{N}^{-1}(\lambda_1 - \mathbf{A})$ moves all nonzero eigenvalues into the left half plane, then replacing \mathbf{N} by $\frac{1}{2}\mathbf{N}$ doubles the convergence rate of the continuous system. (We shall see on page 1461 that there is “no free lunch” for practical computations: the improved convergence rate of the continuous system is counter-balanced by the need to use a smaller step size in the discretized system.)

To rigorously analyze the local behavior of the fully nonlinear system when \mathbf{p} approximates the eigenvector \mathbf{x}_1 , we shall apply the center manifold theorem [9, 17], a tool for studying a dynamical system whose Jacobian has an eigenvalue on the imaginary axis. (Alternatively, we could restrict the system to the unit sphere in \mathbb{R}^n .) We assume that $\mathbf{A} \in \mathbb{R}^{n \times n}$. Without loss of generality, assume that $\lambda_1 = 0$, so that the Jacobian at \mathbf{x}_1 (5.3) takes the form $\mathbf{J} = -\mathbf{N}^{-1}(\mathbf{I} - \mathbf{x}_1 \mathbf{x}_1^*) \mathbf{A}$. Thus, for \mathbf{p} near \mathbf{x}_1 we have

$$\dot{\mathbf{p}} = \mathbf{J}\mathbf{p} + \mathbf{F}(\mathbf{p})$$

for the nonlinear function $\mathbf{F}(\mathbf{p}) = \mathbf{N}^{-1}(\theta(\mathbf{p})\mathbf{p} - (\mathbf{A}\mathbf{p}, \mathbf{x}_1)\mathbf{x}_1)$ that, by definition of the Jacobian, satisfies $\|\mathbf{F}(\mathbf{p})\| = o(\|\mathbf{p} - \mathbf{x}_1\|)$.

Suppose that \mathbf{J} has a simple zero eigenvalue, and the rest of its spectrum is in the open left half plane. There exists some invertible (real, if \mathbf{J} is real) matrix \mathbf{S} with first column \mathbf{x}_1 and

$$\mathbf{S}^{-1}\mathbf{J}\mathbf{S} = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{bmatrix}$$

for some $\mathbf{C} \in \mathbb{R}^{(n-1) \times (n-1)}$ whose spectrum is in the open left half plane.

We now transform coordinates into a form in which the center manifold theorem can most readily be applied. Define

$$\mathbf{r}(t) = \mathbf{S}^{-1}(\mathbf{p}(t) - \mathbf{x}_1),$$

so that

$$\dot{\mathbf{r}} = (\mathbf{S}^{-1}\mathbf{J}\mathbf{S})\mathbf{S}^{-1}(\mathbf{p} - \mathbf{x}_1) + \mathbf{S}^{-1}\mathbf{F}(\mathbf{p}) = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{bmatrix} \mathbf{r} + \mathbf{G}(\mathbf{r}),$$

where $\mathbf{G}(\mathbf{r}) := \mathbf{S}^{-1}\mathbf{F}(\mathbf{S}\mathbf{r} + \mathbf{x}_1) = \mathbf{S}^{-1}\mathbf{F}(\mathbf{p})$. By design, $\mathbf{S}^{-1}\mathbf{x}_1 = \mathbf{e}_1$; hence, $\mathbf{G}(\mathbf{r})$ satisfies

$$(5.4) \quad \mathbf{G}(\mathbf{r}) = \mathbf{S}^{-1}\mathbf{N}^{-1}\mathbf{S} \left(\left(\frac{(\mathbf{A}\mathbf{S}\mathbf{r}, \mathbf{S}\mathbf{r}) + (\mathbf{A}\mathbf{S}\mathbf{r}, \mathbf{x}_1)}{(\mathbf{S}\mathbf{r}, \mathbf{S}\mathbf{r}) + 2(\mathbf{x}_1, \mathbf{S}\mathbf{r}) + 1} \right) (\mathbf{r} + \mathbf{e}_1) - (\mathbf{A}\mathbf{S}\mathbf{r}, \mathbf{x}_1)\mathbf{e}_1 \right).$$

Now we are prepared to cast this diagonalized problem into the conventional setting for center manifold theory. We write

$$\mathbf{r} = \begin{bmatrix} \alpha \\ \mathbf{b} \end{bmatrix}$$

for $\alpha \in \mathbb{R}$ and $\mathbf{b} \in \mathbb{R}^{n-1}$. Using MATLAB index notation for convenience, the \mathbf{r} system is simply

$$\begin{bmatrix} \dot{\alpha} \\ \dot{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{bmatrix} \begin{bmatrix} \alpha \\ \mathbf{b} \end{bmatrix} + \begin{bmatrix} \mathbf{G}([\alpha; \mathbf{b}])_1 \\ \mathbf{G}([\alpha; \mathbf{b}])_{2:n} \end{bmatrix},$$

that is,

$$\dot{\alpha} = \mathbf{G}([\alpha; \mathbf{b}])_1, \quad \dot{\mathbf{b}} = \mathbf{C}\mathbf{b} + \mathbf{G}([\alpha; \mathbf{b}])_{2:n}.$$

Notice that the component α only figures in the nonlinear terms; we wish to determine how that contribution affects the magnitude of the \mathbf{b} component—that is, the portion of the solution that we hope decays as $t \rightarrow \infty$. Notice that $\mathbf{b} = \mathbf{0}$ corresponds to the case when \mathbf{p} is collinear with \mathbf{x}_1 . In this case \mathbf{p} may differ from the unit eigenvector \mathbf{x}_1 , but regardless it is a fixed point of the dynamical system, and provided $\mathbf{p} \neq \mathbf{0}$ we are content. In particular, if $\mathbf{b} = \mathbf{0}$, then $\mathbf{A}\mathbf{S}\mathbf{r} = \mathbf{0}$ too (recall that $\lambda = 0$), and we can see from (5.4) that $\mathbf{G}(\mathbf{r}) = \mathbf{0}$. In this case

$$\dot{\alpha} = \mathbf{G}([\alpha; \mathbf{0}])_1 = 0, \quad \dot{\mathbf{b}} = \mathbf{C}\mathbf{0} + \mathbf{G}([\alpha; \mathbf{0}])_{2:n} = \mathbf{0},$$

so any such \mathbf{r} is a fixed point of the dynamical system. We can put this in grander language: there exists some $\delta > 0$ such that if

$$\mathbf{r}_0 \in \left\{ \begin{bmatrix} \alpha \\ \mathbf{0} \end{bmatrix} : |\alpha| < \delta \right\} =: \mathcal{M},$$

then the dynamical system with $\mathbf{r}(0) = \mathbf{r}_0$ satisfies $\mathbf{r}(t) \in \mathcal{M}$ for all $t > 0$. (In particular, $\mathbf{r}(t) = \mathbf{r}(0) \in \mathcal{M}$.) The set \mathcal{M} is called a *local invariant manifold*. We can define this manifold (locally) by the requirement that

$$\mathbf{b} = \mathbf{g}(\alpha) := \mathbf{0},$$

which trivially satisfies $\mathbf{g}(0) = \mathbf{0}$ and the Jacobian of \mathbf{g} at $\alpha = 0$ is $D\mathbf{g}(0) = \mathbf{0}$; furthermore, \mathbf{g} is arbitrarily smooth near $\alpha = 0$. Together, these properties ensure that \mathcal{M} is a *center manifold* of the dynamical system. (We are fortunate in this case to have an explicit, trivial expression for this manifold.)

All that remains is to apply Theorem 2 from Carr [9, p. 4]. Consider the equation

$$\dot{u} = \mathbf{G}([u; \mathbf{g}(u)])_1 = \mathbf{G}([u; \mathbf{0}])_1 = 0.$$

The solution $u(t) = 0$ is clearly stable—if $u(t) = \varepsilon$, then $|u(t) - 0| = |\varepsilon|$ is bounded for all $t > 0$ —and, thus, Theorem 2(a) from [9] implies that the solution $\mathbf{r}(t) = \mathbf{0}$ is a stable solution of the system

$$\dot{\mathbf{r}} = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{bmatrix} \mathbf{r} + \mathbf{G}(\mathbf{r}).$$

Note that the solution $u(t) = 0$ is not *asymptotically stable*, that is, we do not have $u(t) \rightarrow 0$ if $u(0) = \varepsilon$ for small, nonzero ε . Were this the case, then we would be able to conclude that the \mathbf{r} system was asymptotically stable. This would contradict our expectation that the original dynamical system will converge to something in $\text{span}\{\mathbf{x}_1\}$, not necessarily to \mathbf{x}_1 itself. In particular, if \mathbf{N} is self-adjoint, then $(\mathbf{N}\mathbf{p}, \mathbf{p})$ is an invariant of the system, and so we expect that $\mathbf{p}(t) \rightarrow \xi\mathbf{x}_1$ for ξ determined by

$$|\xi|^2 = \frac{(\mathbf{N}\mathbf{p}, \mathbf{p})}{(\mathbf{N}\mathbf{x}_1, \mathbf{x}_1)}.$$

We now have stability of the zero state of the \mathbf{r} system, but that only means that solutions sufficiently close to $\mathbf{r} = \mathbf{0}$ do not diverge. To say more—to say that the solutions actually converge to the center manifold—we can apply Theorem 2(b) of [9], which we slightly paraphrase here. Since the zero solution of the \mathbf{r} equation is stable, for $\|[\alpha(0); \mathbf{b}(0)]\|$ sufficiently small, there exists some solution $u(t)$ of the equation $\dot{u}(t) = \mathbf{G}([u; \mathbf{g}(u)])_1 = 0$ and positive constant γ such that

$$\alpha(t) = u(t) + O(e^{-\gamma t}), \quad \mathbf{b}(t) = \mathbf{g}(u(t)) + O(e^{-\gamma t}).$$

In particular, in our setting such solutions $u(t)$ will be constant: $u(t) = c$, and so there exist

$$\alpha(t) = c + O(e^{-\gamma t}), \quad \mathbf{b}(t) = O(e^{-\gamma t}),$$

and, in particular, $\|\mathbf{b}(t)\| \rightarrow 0$ as $t \rightarrow \infty$. Thus, for $\|\mathbf{r}_0\|$ sufficiently small,

$$\mathbf{r}(t) = \begin{bmatrix} c \\ \mathbf{0} \end{bmatrix} + O(e^{-\gamma t}),$$

so that $\mathbf{p}(t) = \mathbf{S}\mathbf{r}(t) + \mathbf{x}_1 = (1 + c)\mathbf{x}_1 + O(e^{-\gamma t})$. The preceding discussion is summarized in the following result.

THEOREM 5.1. *If $\|\mathbf{p}(0) - \mathbf{x}_1\|$ is sufficiently small and $\mathbf{N}^{-1}(\mathbf{I} - \mathbf{x}_1\mathbf{x}_1^*)(\lambda - \mathbf{A})$ has a simple zero eigenvalue with all other eigenvalues in the open left half plane, then there exists $\gamma > 0$ and $\xi \in \mathbb{R}$ such that, as $t \rightarrow \infty$,*

$$\|\mathbf{p}(t) - \xi\mathbf{x}_1\| = O(e^{-\gamma t}).$$

In the case of self-adjoint, invertible \mathbf{N} , $|\xi| = |(\mathbf{p}_0, \mathbf{N}\mathbf{p}_0)|$.

Note that if \mathbf{N} is Hermitian and invertible but indefinite, then there always exists some unit vector \mathbf{p}_0 such that $(\mathbf{p}_0, \mathbf{N}\mathbf{p}_0) = 0$. If this starting vector is sufficiently close to the unit eigenvector \mathbf{x}_1 of \mathbf{A} , then we have not ruled out the possibility that the system converges to the zero vector, rather than a desired eigenvector.

6. Discrete dynamical systems. The previous sections have addressed the quadratic invariant and convergence behavior of the continuous-time, one- and two-sided dynamical systems. For purposes of computation, one naturally wonders how closely such properties are mimicked by the solutions to discretizations of these systems. The present section considers the convergence and preservation of the quadratic invariant by the discrete flow under a forward Euler time integration. We focus on this canonical integrator for three reasons: (1) this discretization leads to the algorithm (1.1) proposed in the literature; (2) analysis for forward Euler serves as a first step toward understanding more sophisticated algorithms; (3) more elaborate methods are not always practical. For example, the implicit midpoint rule will preserve the quadratic invariant $(\mathbf{p}, \mathbf{N}\mathbf{p})$ [18, IV.2.1] of the one-sided system (1.2), but since this method takes the form

$$\mathbf{p}_{j+1} = \mathbf{p}_j + h\mathbf{N}^{-1} \left(\theta_{j+1} \left(\frac{\mathbf{p}_j + \mathbf{p}_{j+1}}{2} \right) - \mathbf{A} \left(\frac{\mathbf{p}_j + \mathbf{p}_{j+1}}{2} \right) \right)$$

$$\theta_{j+1} = \frac{(\mathbf{p}_j + \mathbf{p}_{j+1})^T \mathbf{A} (\mathbf{p}_j + \mathbf{p}_{j+1})}{(\mathbf{p}_j + \mathbf{p}_{j+1})^T (\mathbf{p}_j + \mathbf{p}_{j+1})},$$

its implementation requires the solution of a (nonlinear) system of equations at each step: a far more expensive proposition (per step) than the humble forward Euler method. (For a more sophisticated discretization in the unpreconditioned Hermitian case, along with a cautionary note about use of large step-size in the forward Euler method, see [28].)

6.1. Departure from the manifold. Given $\mathbf{A} \in \mathbb{R}^{n \times n}$, for notational convenience we rewrite the two-sided system in the form

$$(6.1) \quad \begin{aligned} \dot{\mathbf{p}} &= \mathbf{p}\theta - \mathbf{A}\mathbf{p} =: \mathbf{f}(\mathbf{p}, \mathbf{q}) \\ \dot{\mathbf{q}} &= \mathbf{q}\theta - \mathbf{A}^T \mathbf{q} =: \mathbf{g}(\mathbf{p}, \mathbf{q}), \end{aligned}$$

with $\theta = (\mathbf{q}^T \mathbf{p})^{-1} \mathbf{q}^T \mathbf{A} \mathbf{p} = \theta^T$ and initial conditions $\mathbf{p}(0) = \mathbf{p}_0 \in \mathbb{R}^n$ and $\mathbf{q}(0) = \mathbf{q}_0 \in \mathbb{R}^n$. Similarly, the one-sided system (now including preconditioning) is

$$(6.2) \quad \dot{\mathbf{p}} = \mathbf{N}^{-1}(\mathbf{p}\theta - \mathbf{A}\mathbf{p}) =: \mathbf{N}^{-1}\mathbf{f}(\mathbf{p}, \mathbf{p}),$$

with $\theta = (\mathbf{p}^T \mathbf{p})^{-1} \mathbf{p}^T \mathbf{A} \mathbf{p} = \theta^T$ and $\mathbf{p}(0) = \mathbf{p}_0 \in \mathbb{R}^n$.

In section 2 we showed that this system preserves the quadratic invariant $\mathbf{q}^T \mathbf{p}$. To what extent do discretizations respect such conservation, and what are the implications of any drift from this manifold? To understand the role of discrete quadratic invariants, we consider the error when using a forward Euler time integrator.

We begin with the two-sided iteration. The finite-time blow-up established in Theorem 4.1 is a strike against this method. Before abandoning it altogether, we wish to investigate the consequences of the blow-up on the discrete two-sided eigensolver. The forward Euler applied to (6.1) leads to the iteration

$$(6.3) \quad \mathbf{p}_{j+1} = \mathbf{p}_j + h\mathbf{f}_j$$

$$(6.4) \quad \mathbf{q}_{j+1} = \mathbf{q}_j + h\mathbf{g}_j,$$

where $\mathbf{f}_j := \mathbf{f}(\mathbf{p}_j, \mathbf{q}_j)$ and $\mathbf{g}_j := \mathbf{g}(\mathbf{p}_j, \mathbf{q}_j)$. With the mild caveat that $\mathbf{q}_j^T \mathbf{p}_j \neq 0$, the form of the Rayleigh quotient gives

$$\mathbf{q}_j^T \mathbf{f}_j = 0 = \mathbf{p}_j^T \mathbf{g}_j.$$

This simple observation is critical to understanding the drift of the forward Euler iterates from the invariant manifold. It implies, for example, that the first iteration of (6.3)–(6.4) produces an iterate that is quadratically close to the manifold:

$$\mathbf{q}_1^T \mathbf{p}_1 = \mathbf{q}_0^T \mathbf{p}_0 + h^2 (\mathbf{g}_0^T \mathbf{f}_0),$$

which is perhaps surprising given the forward Euler method’s $O(h)$ accuracy. Writing the departure from the manifold as

$$d_j = \mathbf{q}_j^T \mathbf{p}_j - \mathbf{q}_0^T \mathbf{p}_0,$$

we, thus, have $d_1 = h^2(\mathbf{g}_0^T \mathbf{f}_0)$. From this we can compute

$$d_2 = (\mathbf{q}_2^T \mathbf{p}_2 - \mathbf{q}_1^T \mathbf{p}_1) + d_1 = h^2 (\mathbf{g}_1^T \mathbf{f}_1 + \mathbf{g}_0^T \mathbf{f}_0)$$

and, in general, $d_{j+1} = h^2 \sum_{k=0}^j \mathbf{g}_k^T \mathbf{f}_k$. (This result is a special case of one derived in [18] for partitioned Runge–Kutta systems.) Thus, we can bound the relative drift from the manifold as

$$(6.5) \quad \frac{|\mathbf{q}_{j+1}^T \mathbf{p}_{j+1} - \mathbf{q}_0^T \mathbf{p}_0|}{|\mathbf{q}_0^T \mathbf{p}_0|} \leq h^2 \sum_{k=0}^j \frac{\|\mathbf{f}_k\| \|\mathbf{g}_k\|}{|\mathbf{q}_0^T \mathbf{p}_0|}.$$

The definitions of $\mathbf{f}(\mathbf{p}, \mathbf{q})$ and $\mathbf{g}(\mathbf{p}, \mathbf{q})$ imply

$$\begin{aligned} \|\mathbf{f}_k\| &\leq (|\theta_k| + \|\mathbf{A}\|) \|\mathbf{p}_k\| \leq \left(1 + \frac{\|\mathbf{q}_k\| \|\mathbf{p}_k\|}{|\mathbf{q}_k^T \mathbf{p}_k|}\right) \|\mathbf{A}\| \|\mathbf{p}_k\| \\ \|\mathbf{g}_k\| &\leq (|\theta_k| + \|\mathbf{A}\|) \|\mathbf{q}_k\| \leq \left(1 + \frac{\|\mathbf{p}_k\| \|\mathbf{q}_k\|}{|\mathbf{p}_k^T \mathbf{q}_k|}\right) \|\mathbf{A}\| \|\mathbf{q}_k\|. \end{aligned}$$

Substituting these formulas into (6.5), we arrive at the following result.

THEOREM 6.1. *The forward Euler iterates (6.3)–(6.4) for the two-sided dynamical system (6.1) satisfy*

$$(6.6) \quad \frac{|\mathbf{q}_{j+1}^T \mathbf{p}_{j+1} - \mathbf{q}_0^T \mathbf{p}_0|}{|\mathbf{q}_0^T \mathbf{p}_0|} \leq h^2 \frac{\|\mathbf{A}\|^2}{|\mathbf{q}_0^T \mathbf{p}_0|} \sum_{k=0}^j \left(1 + \frac{\|\mathbf{q}_k\| \|\mathbf{p}_k\|}{|\mathbf{q}_k^T \mathbf{p}_k|}\right)^2 \|\mathbf{q}_k\| \|\mathbf{p}_k\|.$$

This bound implies that the departure from the manifold is proportional to the square of the step size, and involves the secants of the angles formed by \mathbf{q}_k and \mathbf{p}_k , $k = 0, \dots, j$, as well as the norms of \mathbf{q}_k and \mathbf{p}_k . Moreover, unless the cosines of the angles between \mathbf{q}_k and \mathbf{p}_k are bounded away from zero, there does not exist a step size h such that all iterates remain near the quadratic manifold. The proof of the theorem demonstrates that the secant of the angle is at least as large as the normalized residuals. Numerical experiments indicate that these bounds are descriptive; see the first example in section 6.3. A conclusion is that serious breakdown (as discussed after Theorem 4.1) leads to *incurable breakdown* of the two-sided iteration because forward Euler mimics the continuous solution and cannot “step-over” the point of blow-up.

Given the shortcomings of the two-sided iteration, we shall, henceforth, focus on the one-sided dynamical system, and also include preconditioning (6.2). The associated forward Euler discretization takes the form

$$(6.7) \quad \mathbf{p}_{j+1} = \mathbf{p}_j + h\mathbf{N}^{-1}\mathbf{f}_j,$$

where now $\mathbf{f}_j = \mathbf{f}(\mathbf{p}_j, \mathbf{p}_j)$. (Here we see that the time-step h directly multiplies the preconditioner \mathbf{N} , so that the effect of scaling \mathbf{N} to improve the convergence rate of the continuous-time system, as discussed on page 1456, is equivalent to choosing a smaller time-step in the discrete setting.)

The following analysis will play a useful role in our main convergence result, Theorem 6.3. For the rest of the paper we assume that \mathbf{N} is symmetric and invertible, which, as seen in the Introduction, ensures that solutions of the continuous system reside on an invariant manifold $\mathbf{p}^T \mathbf{N} \mathbf{p} = \text{constant}$. At each time step, the discrete iteration incurs a local departure from that manifold of

$$e_{j+1} := \mathbf{p}_{j+1}^T \mathbf{N} \mathbf{p}_{j+1} - \mathbf{p}_j^T \mathbf{N} \mathbf{p}_j = h^2 \mathbf{f}_j^T \mathbf{N}^{-1} \mathbf{f}_j.$$

Hence, if \mathbf{N}^{-1} is additionally positive definite (e.g., $\mathbf{N}^{-1} = \mathbf{I}$), the drift is monotone increasing—an important property for the forthcoming convergence theory.

When \mathbf{N} is positive definite, we can define vector norms

$$\|\mathbf{z}\|_{\mathbf{N}^{-1}}^2 := \mathbf{z}^T \mathbf{N}^{-1} \mathbf{z}, \quad \|\mathbf{z}\|_{\mathbf{N}}^2 := \mathbf{z}^T \mathbf{N} \mathbf{z}$$

(which in turn induce matrix norms), with $\|\mathbf{z}\|_{\mathbf{N}^{-1}} \leq \|\mathbf{N}^{-1}\| \|\mathbf{z}\|_{\mathbf{N}}$. Thus, we write

$$e_{j+1} = h^2 \|\mathbf{f}_j\|_{\mathbf{N}^{-1}}^2 \leq h^2 \|\mathbf{N}^{-1}\|^2 \|\mathbf{f}_j\|_{\mathbf{N}}^2 = h^2 \|\mathbf{N}^{-1}\|^2 \|\mathbf{r}_j\|_{\mathbf{N}}^2 \|\mathbf{p}_j\|_{\mathbf{N}}^2,$$

where we use the normalized residual $\mathbf{r}_j := \mathbf{f}_j / \|\mathbf{p}_j\|_{\mathbf{N}} = (\theta_j - \mathbf{A})\mathbf{p}_j / \|\mathbf{p}_j\|_{\mathbf{N}}$. Now consider the aggregate, global drift from the manifold:

$$\begin{aligned} d_{j+1} &:= \mathbf{p}_{j+1}^T \mathbf{N} \mathbf{p}_{j+1} - \mathbf{p}_0^T \mathbf{N} \mathbf{p}_0 \\ &= \sum_{k=1}^{j+1} e_k \leq h^2 \|\mathbf{N}^{-1}\|^2 \sum_{k=0}^j \|\mathbf{r}_k\|_{\mathbf{N}}^2 (d_k + \|\mathbf{p}_0\|_{\mathbf{N}}^2). \end{aligned}$$

In particular, d_{j+1} is determined by the step size, the residual norms, and the growth in the norm of the iterates. For further simplification, choose some $M > 0$ such that $\|\mathbf{r}_k\|_{\mathbf{N}}^2 \leq M$ for all $k = 0, \dots, j$. One coarse (but j -independent) possibility is

$$(6.8) \quad M := \inf_{s \in \mathbb{R}} 4\|\mathbf{A} - s\|_{\mathbf{N}}^2 \geq \inf_{s \in \mathbb{R}} \|(\mathbf{A} - s) - (\theta_k - s)\|_{\mathbf{N}}^2 \geq \|\mathbf{r}_k\|_{\mathbf{N}}^2,$$

which is invariant to shifts in \mathbf{A} . (In terms of the Euclidean norm, we, thus, have $M \leq 4\kappa(\mathbf{N}) \inf_{s \in \mathbb{R}} \|\mathbf{A} - s\|^2$, where $\kappa(\mathbf{N}) = \|\mathbf{N}\| \|\mathbf{N}^{-1}\|$.) Hence,

$$d_{j+1} \leq h^2 M \|\mathbf{N}^{-1}\|^2 \sum_{k=0}^j (d_k + \|\mathbf{p}_0\|_{\mathbf{N}})^2 = h^2 M \|\mathbf{N}^{-1}\|^2 \left((j+1)\|\mathbf{p}_0\|_{\mathbf{N}}^2 + \sum_{k=1}^j d_k \right)$$

(since $d_0 = 0$). Thus, if we define the sequence $\{\widehat{d}_k\}$ by

$$(6.9) \quad \widehat{d}_{j+1} = h^2 M \|\mathbf{N}^{-1}\|^2 \left((j+1) + \sum_{k=1}^j \widehat{d}_k \right),$$

then the departure from the manifold obeys $d_{j+1} \leq \widehat{d}_{j+1} \|\mathbf{p}_0\|_{\mathbf{N}}^2$. Equation (6.9) is a binomial recurrence whose solution can be written explicitly:

$$\widehat{d}_{j+1} = \sum_{k=1}^{j+1} \binom{j+1}{k} (h^2 M \|\mathbf{N}^{-1}\|^2)^k = (1 + h^2 M \|\mathbf{N}^{-1}\|^2)^{j+1} - 1.$$

THEOREM 6.2. *Let $\mathbf{N} \in \mathbb{R}^{n \times n}$ be symmetric and positive definite, and define M by (6.8). Then the forward Euler iterates (6.7) for the preconditioned one-sided dynamical system (6.2) satisfy*

$$(6.10) \quad 0 \leq \frac{\mathbf{p}_{j+1}^T \mathbf{N} \mathbf{p}_{j+1} - \mathbf{p}_0^T \mathbf{N} \mathbf{p}_0}{\mathbf{p}_0^T \mathbf{N} \mathbf{p}_0} \leq (1 + h^2 M \|\mathbf{N}^{-1}\|^2)^{j+1} - 1,$$

the upper bound being asymptotic to $(j+1)h^2 \|\mathbf{N}^{-1}\|^2 M$ as $h \rightarrow 0$.

Note that a small eigenvalue of \mathbf{N} results in a small time-step h . The bound also provides an estimate of a critical time-step

$$h\sqrt{j+1} \lesssim \frac{1}{\|\mathbf{N}^{-1}\| \sqrt{M}}$$

for forward Euler, limiting the departure from the quadratic manifold. Highly non-normal problems for which $\|\mathbf{A} - s\| \gg \max_k |\lambda_k - s|$ also result in tiny time-steps.

Theorem 6.2 leads to an interesting observation—despite the fact that the forward Euler method generally incurs an $O(h)$ truncation error and the global error grows exponentially in j for fixed h (see (6.12) and, e.g., [14, section 1.3]), for a one-sided iteration the drift from the quadratic manifold is $O(h^2)$ and both linear and nondecreasing in j for all starting vectors, under mild restrictions. This monotone departure from the manifold is exploited in the discrete convergence analysis to follow. So, although explicit Runge–Kutta methods (such as forward Euler) do not preserve quadratic invariants (see [18, Chapter IV]), the forward Euler iterates for the one-sided systems remain nearby. The reader is referred to [18, Chapter IV] for further information and references, including the use of projection to remain on the quadratic manifold.

6.2. Discrete convergence theory. Just as the local drift from the manifold at each iteration contributes to the global drift, so local truncation errors committed by each step of an ODE solver aggregate into a global error. How does this accumulated error affect convergence of the discrete method as we compute \mathbf{p}_j with $j \rightarrow \infty$?

In this section, we seek conditions that will ensure that the *discrete preconditioned one-sided iteration* (6.7) converges to the same eigenvector as the continuous system.

First, we establish the setting that will be used through this rest of this section. Suppose $\mathbf{A} \in \mathbb{R}^{n \times n}$ has a simple eigenvalue λ_1 strictly to the left of all other eigenvalues (and, hence, real). Without loss of generality (via a unitary similarity transformation) we can assume that \mathbf{A} takes the form

$$(6.11) \quad \mathbf{A} = \begin{bmatrix} \lambda_1 & \mathbf{d}^T \\ \mathbf{0} & \mathbf{C} \end{bmatrix}.$$

Let \mathbf{x}_1 and \mathbf{y}_1 denote unit-length right and left eigenvectors associated with λ_1 ; in these coordinates we can take $\mathbf{x}_1 = [1, 0, \dots, 0]^T$. Theorems 4.3, 4.4, and 5.1 provide conditions under which the solution $\mathbf{p}(t)$ of the continuous system converges in angle to the eigenvector \mathbf{x}_1 (e.g., if $\mathbf{N} = \mathbf{I}$ and $\mathbf{y}_1^T \mathbf{p}_0 \neq 0$).

Before beginning the convergence analysis, one should appreciate that the conditions established in the last paragraph are not sufficient to guarantee convergence of the discrete iteration. Consider the following example. When $\mathbf{N} = \mathbf{I}$, the forward Euler iterate of the one-sided system at step k can be written as

$$\mathbf{p}_k = \prod_{j=0}^{k-1} \varphi_j(\mathbf{A}) \mathbf{p}_0$$

for linear factors $\varphi_j(z) = 1 + h(\theta_j - z)$. If any of these factors has λ_1 as a root, then \mathbf{p}_k will have no component in the direction of the eigenvector \mathbf{x}_1 , and so λ_1 and \mathbf{x}_1 will not influence the iteration: convergence of \mathbf{p}_k to \mathbf{x}_1 is impossible. Concrete matrices that exhibit such behavior are simple to construct. For any *fixed* $h > 0$, set

$$\mathbf{A} = \begin{bmatrix} 0 & -1 - 2/h \\ 0 & 1 \end{bmatrix}, \quad \mathbf{p}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Theorem 4.3 guarantees that the continuous one-sided system will converge for this \mathbf{A} and \mathbf{p}_0 . At the first step of the forward Euler method $\theta_0 = -1/h$, so that $\varphi_0(0) = 0$ and $\mathbf{p}_1 = [h+2, -h]^T$ is an eigenvector for $\lambda_2 = 1$, and \mathbf{p}_k will never have a component in the \mathbf{x}_1 direction for any $k \geq 1$. (Note that $\varphi_j(\lambda_1) = 1 + h(\theta_j - \lambda_1) = 0$ implies that $\theta_j - \lambda_1 = -1/h < 0$, and this is impossible if \mathbf{A} is normal. As h is reduced, complete deflation requires an increasing departure from normality.) The more sophisticated restarted Arnoldi algorithm exhibits a similar phenomenon; see [12].

Under what circumstances can we guarantee convergence? To answer this question, we first review the conventional global error analysis for the forward Euler method; for details, see, e.g., [14, section 1.3]. The first step begins with the exact solution at time $t = 0$: $\mathbf{p}_0 = \mathbf{p}(0)$. Each subsequent step introduces a local truncation error, while also magnifying the global error aggregated at previous steps. Suppose we wish to integrate for $t \in [0, \tau]$ with $\tau = kh$ for some integer k . With the local truncation error at each step bounded by

$$T_h := \max_{0 \leq t \leq \tau} \frac{1}{2} h \|\dot{\mathbf{p}}(t)\|,$$

one can show that

$$(6.12) \quad \|\mathbf{p}_k - \mathbf{p}(\tau)\| \leq \frac{T_h}{L} (e^{\tau L} - 1),$$

where L is a Lipschitz constant for our differential equation; in Appendix A we show that $L = 10\|\mathbf{N}^{-1}\|\|\mathbf{A}\|$ will suffice. This expression for the global error captures an essential feature: for fixed τ , the fact that $T_h = O(h)$ implies that we can always select $h > 0$ sufficiently small as to make the difference between the forward Euler iterate $\mathbf{p}_{\tau/h}$ and the exact solution $\mathbf{p}(\tau)$ arbitrarily small. However, if we increase k with $h > 0$ *fixed*, the bound indicates an *exponential* growth in the error. To show that \mathbf{p}_k converges (in angle) to an eigenvector as $k \rightarrow \infty$, further work is required. In this effort, the preservation of the quadratic invariant characterized in Theorem 6.2 plays an essential role.

Preconditioning significantly complicates the convergence theory. For simplicity, our analysis imposes the stringent requirement that, in the coordinates in which \mathbf{A} takes the form (6.11), we have

$$(6.13) \quad \mathbf{N}^{-1} = \begin{bmatrix} \eta & \mathbf{0} \\ \mathbf{0} & \mathbf{M} \end{bmatrix}$$

in addition to the requirement that \mathbf{N}^{-1} be symmetric and positive definite. The trivial off-diagonal blocks prevent the preconditioner from using the growing component of \mathbf{p}_k in \mathbf{x}_1 to enlarge the component in the unwanted eigenspace.

A crucial ingredient in our convergence analysis is the constant

$$\gamma := \|\Pi_1(\mathbf{I} + h\mathbf{N}(\lambda_1 - \mathbf{A}))\| = \|\mathbf{I} + h\mathbf{M}(\lambda_1 - \mathbf{C})\|,$$

where $\mathbf{\Pi}_1 := \mathbf{I} - \mathbf{x}_1 \mathbf{x}_1^T$ is a projector onto the complement of the desired invariant subspace. This constant γ , a function of h , measures the potency of the preconditioner: the smaller, the better. For example, in the ideal case that $\mathbf{M} = (\mathbf{C} - \lambda_1)^{-1}$, we have $\gamma = |1 - h|$, giving $\gamma = 0$ for the large step size $h = 1$, and that $\gamma \rightarrow 1$ as $h \rightarrow 0$.

With γ in hand, we are prepared to state our convergence result. Here, $\kappa(\mathbf{N}) = \|\mathbf{N}\| \|\mathbf{N}^{-1}\|$ denotes the condition number of the preconditioner.

THEOREM 6.3. *Given (6.11), (6.13), and assumptions on λ_1 , \mathbf{x}_1 , and \mathbf{N} established in the previous paragraphs, suppose that \mathbf{p}_0 is chosen so that the continuous dynamical system converges in angle to an eigenvector associated with the distinct, simple leftmost eigenvalue λ_1 (e.g., $\mathbf{y}_1^T \mathbf{p}_0 \neq 0$ suffices if $\mathbf{N} = \mathbf{I}$). Furthermore, suppose there exists $h > 0$ for which*

$$(6.14) \quad \gamma \in [0, 1/\sqrt{\kappa(\mathbf{N})}).$$

Then after preliminary iteration with a sufficiently small time-step h_0 , the forward Euler method with time-step h will converge (in angle) to the desired eigenvector:

$$(6.15) \quad \sin(\angle(\mathbf{p}_k, \mathbf{x}_1)) = O(\gamma^k).$$

Asymptotically, the Rayleigh quotient converges to λ at the same rate:

$$(6.16) \quad |\theta_k - \lambda| = O(\gamma^k),$$

which in the case $\mathbf{d} = \mathbf{0}$ improves to $|\theta_k - \lambda| = O(\gamma^{2k})$.

Proof. Denote the k th iterate by

$$\mathbf{p}_k = \begin{bmatrix} \alpha_k \\ \mathbf{b}_k \end{bmatrix}.$$

- *Convergence of the forward Euler method to the continuous solution, and convergence of the continuous solution to the eigenvector, together ensure that preliminary forward Euler steps will get close to the eigenvector.* To show that $\sin(\angle(\mathbf{p}_k, \mathbf{x}_1)) \rightarrow 0$ as $k \rightarrow \infty$, we will show that $\|\mathbf{b}_k\| \rightarrow 0$ while $|\alpha_k|$ is bounded away from zero. The convergence of the forward Euler method at a fixed time $\tau \geq 0$ (see (6.12)), with the assumption that the continuous system converges for the given \mathbf{p}_0 (as described in sections 4–5), ensures that we can run the forward Euler iteration with a sufficiently small time-step that, after $k \geq 0$ iterations, $\|\mathbf{b}_k\|$ is sufficiently small that

$$(6.17) \quad \frac{\|\mathbf{b}_k\|^2 \|\lambda_1 - \mathbf{C}\|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} + \frac{\|\mathbf{b}_k\| \|\mathbf{d}\|}{\sqrt{\alpha_k^2 + \|\mathbf{b}_k\|^2}} \leq \frac{\varepsilon}{h \|\mathbf{M}\|}$$

for some $\varepsilon \in [0, 1/\sqrt{\kappa(\mathbf{N})} - \gamma]$; here $\gamma \in [0, 1/\sqrt{\kappa(\mathbf{N})})$ and $h > 0$ are as in the statement of the theorem. Note that the left-hand side of (6.17) will get small when $\|\mathbf{b}_k\|$ is small, since $|\alpha_k|$ is bounded away from zero. This follows from Theorem 6.2 (monotonic drift of the invariant) and the fact that \mathbf{N} is symmetric positive definite, which imply that for any j ,

$$(6.18) \quad \|\mathbf{p}_j\|^2 \geq \frac{1}{\|\mathbf{N}\|} \mathbf{p}_j^T \mathbf{N} \mathbf{p}_j \geq \frac{1}{\|\mathbf{N}\|} \mathbf{p}_{j-1}^T \mathbf{N} \mathbf{p}_{j-1} \geq \frac{1}{\kappa(\mathbf{N})} \|\mathbf{p}_{j-1}\|^2.$$

- *Condition (6.17) ensures that θ_k is close to λ_1 .* Since

$$\theta_k = \frac{\lambda_1 \alpha_k^2 + \alpha_k \mathbf{d}^T \mathbf{b}_k + \mathbf{b}_k^T \mathbf{C} \mathbf{b}_k}{\alpha_k^2 + \|\mathbf{b}_k\|^2},$$

we have

$$\begin{aligned}
 |\theta_k - \lambda_1| &= \frac{|\lambda_1 \alpha_k^2 + \alpha_k \mathbf{d}^T \mathbf{b}_k + \mathbf{b}_k^T \mathbf{C} \mathbf{b}_k - \lambda_1 (\alpha_k^2 + \mathbf{b}_k^T \mathbf{b}_k)|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} \\
 &\leq \frac{|\mathbf{b}_k^T (\mathbf{C} - \lambda_1) \mathbf{b}_k|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} + \frac{|\alpha_k| \|\mathbf{b}_k\| \|\mathbf{d}\|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} \\
 (6.19) \quad &\leq \frac{\|\mathbf{b}_k\|^2 \|\mathbf{C} - \lambda_1\|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} + \frac{\|\mathbf{b}_k\| \|\mathbf{d}\|}{\sqrt{\alpha_k^2 + \|\mathbf{b}_k\|^2}},
 \end{aligned}$$

where the last inequality uses the fact that $|\alpha_k| \leq \sqrt{\alpha_k^2 + \|\mathbf{b}_k\|^2}$. Now condition (6.17) implies that the Rayleigh quotient θ_k is sufficiently close to the eigenvalue λ_1 :

$$(6.20) \quad |\theta_k - \lambda_1| \leq \frac{\varepsilon}{h \|\mathbf{M}\|}.$$

The next step of the iteration, with time-step $h > 0$ specified in the statement of the theorem, produces

$$\begin{bmatrix} \alpha_{k+1} \\ \mathbf{b}_{k+1} \end{bmatrix} = \mathbf{p}_{k+1} = \mathbf{p}_k + h \mathbf{N}^{-1} (\theta_k - \mathbf{A}) \mathbf{p}_k = \begin{bmatrix} \alpha_k + \eta h ((\theta_k - \lambda_1) \alpha_k - \mathbf{d}^T \mathbf{b}_k) \\ (\mathbf{I} + h \mathbf{M} (\theta_k - \mathbf{C})) \mathbf{b}_k \end{bmatrix}.$$

Adding zero in a convenient way gives

$$\begin{aligned}
 \|\mathbf{b}_{k+1}\| &= \|(\mathbf{I} + h \mathbf{M} (\lambda_1 - \mathbf{C})) \mathbf{b}_k + h (\theta_k - \lambda_1) \mathbf{M} \mathbf{b}_k\| \\
 &\leq \|\mathbf{I} + h \mathbf{M} (\lambda_1 - \mathbf{C})\| \|\mathbf{b}_k\| + h |\lambda_1 - \theta_k| \|\mathbf{M}\| \|\mathbf{b}_k\| \\
 (6.21) \quad &\leq (\gamma + \varepsilon) \|\mathbf{b}_k\|.
 \end{aligned}$$

In particular, since $0 \leq \gamma + \varepsilon < 1/\kappa(\mathbf{N}) \leq 1$, this guarantees a fixed reduction in the component of the forward Euler iterate in the unwanted eigenspace. (The second inequality follows from condition (6.14) and bound (6.20).) After checking a few details, we shall see that this condition is the key to convergence.

• *Subsequent Rayleigh quotients must also remain close to λ_1 .* We now show that the new Rayleigh quotient, θ_{k+1} , automatically satisfies the requirement (6.20) with the same $\varepsilon > 0$ and time-step. Repeating the calculation that culminated in (6.19), we obtain

$$|\theta_{k+1} - \lambda_1| \leq \frac{\|\mathbf{b}_{k+1}\|^2 \|\mathbf{C} - \lambda_1\|}{\alpha_{k+1}^2 + \|\mathbf{b}_{k+1}\|^2} + \frac{\|\mathbf{d}\| \|\mathbf{b}_{k+1}\|}{\sqrt{\alpha_{k+1}^2 + \|\mathbf{b}_{k+1}\|^2}}.$$

Now we use (6.18), a consequence of the monotonic drift from the invariant manifold, to deduce that

$$\begin{aligned}
 |\theta_{k+1} - \lambda_1| &\leq \frac{\kappa(\mathbf{N})(\gamma + \varepsilon)^2 \|\mathbf{b}_k\|^2 \|\mathbf{C} - \lambda_1\|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} + \frac{\sqrt{\kappa(\mathbf{N})(\gamma + \varepsilon)} \|\mathbf{d}\| \|\mathbf{b}_k\|}{\sqrt{\alpha_k^2 + \|\mathbf{b}_k\|^2}} \\
 &\leq \frac{\|\mathbf{b}_k\|^2 \|\mathbf{C} - \lambda_1\|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} + \frac{\|\mathbf{d}\| \|\mathbf{b}_k\|}{\sqrt{\alpha_k^2 + \|\mathbf{b}_k\|^2}},
 \end{aligned}$$

since $\gamma + \varepsilon < 1/\sqrt{\kappa(\mathbf{N})}$. The condition (6.17) then implies that

$$|\theta_{k+1} - \lambda_1| \leq \frac{\varepsilon}{h \|\mathbf{M}\|},$$

which guarantees that the Rayleigh quotient cannot wander too far from λ_1 .

• *Subsequent iterates and Rayleigh quotients must eventually converge.* The bound on $|\theta_{k+1} - \lambda_1|$ just established allows us to repeat the argument resulting in (6.21) at future steps, giving

$$\|\mathbf{b}_{k+m}\| \leq (\gamma + \varepsilon)^m \|\mathbf{b}_k\|$$

along with, via a slight modification of (6.18),

$$\begin{aligned} (6.22) \quad |\theta_{k+m} - \lambda_1| &\leq \frac{\kappa(\mathbf{N})(\gamma + \varepsilon)^{2m} \|\mathbf{b}_k\|^2 \|\mathbf{C} - \lambda_1\|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} + \frac{\sqrt{\kappa(\mathbf{N})(\gamma + \varepsilon)^m \|\mathbf{d}\|} \|\mathbf{b}_k\|}{\sqrt{\alpha_k^2 + \|\mathbf{b}_k\|^2}} \\ &\leq \frac{\|\mathbf{b}_k\|^2 \|\mathbf{C} - \lambda_1\|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} + \frac{\|\mathbf{d}\| \|\mathbf{b}_k\|}{\sqrt{\alpha_k^2 + \|\mathbf{b}_k\|^2}}. \end{aligned}$$

Thus, $|\theta_{k+m} - \lambda_1| \leq \varepsilon/(h\|\mathbf{M}\|)$ for all $m \geq 1$. As $\|\mathbf{b}_{k+m}\| \rightarrow 0$, the component in the desired eigenvector does not vanish, as again a generalization of (6.18) gives

$$\|\mathbf{p}_{k+m}\| \geq \frac{1}{\sqrt{\kappa(\mathbf{N})}} \|\mathbf{p}_0\|.$$

Thus, with $\mathbf{x}_1 = \mathbf{e}_1$, we have

$$\begin{aligned} \sin \angle(\mathbf{p}_{k+m}, \mathbf{x}_1) &= \min_{\xi} \frac{\|\xi \mathbf{p}_{k+m} - \mathbf{x}_1\|}{\|\mathbf{x}_1\|} = \min_{\xi} \left\| \begin{bmatrix} \xi \alpha_{k+m} - 1 \\ \xi \mathbf{b}_{k+m} \end{bmatrix} \right\| \\ &\leq \frac{\|\mathbf{b}_{k+m}\|}{|\alpha_{k+m}|} \leq (\gamma + \varepsilon)^m \frac{\|\mathbf{b}_k\|}{|\alpha_{k+m}|}, \end{aligned}$$

where we have taken $\xi = \alpha_{k+m}^{-1}$ for the first inequality. As $|\alpha_{k+m}|$ is bounded away from zero, we have $\sin \angle(\mathbf{p}_{k+m}, \mathbf{x}_1) = O((\gamma + \varepsilon)^m)$ as $m \rightarrow \infty$. Since $\|\mathbf{b}_{k+m}\| \rightarrow 0$ as $m \rightarrow \infty$, we can take the ε used in (6.19) to be arbitrarily small as the iterations progress, giving the asymptotic rate given in (6.15). Similarly, from (6.22) we observe that the Rayleigh quotient converges as in (6.16). The $O(\gamma^m)$ term in that bound falls out if $\mathbf{d} = \mathbf{0}$. \square

We now make several remarks concerning Theorem 6.14 and its proof. (1) As \mathbf{N} becomes increasingly ill-conditioned, the hypothesis (6.14) in the theorem becomes more and more difficult to satisfy. We can only guarantee convergence for an ill-conditioned preconditioner if that preconditioner gives a small value of γ , i.e., if it gives a rapid convergence rate. (2) A curiosity of condition (6.17) is that the requirement is more strict when convergence is slower, i.e., when γ is near $\kappa(\mathbf{N})^{-1/2}$. (3) One does not in general know whether θ_k falls to the left or right of λ_1 . If \mathbf{A} is normal, then as θ_k must fall the convex hull of its spectrum, and so $\theta_k \geq \lambda_1$; for nonnormal \mathbf{A} , it is possible that $\theta_k < \lambda_1$. (4) The proof of the theorem exploits the monotonic drift from the manifold described by Theorem 6.2. This drift is easily monitored, so providing a useful (and cheap) check on convergence of the iteration during computation. If this drift reaches a point where it is not small, projection to the quadratic manifold is easily undertaken; see [18, Chapter IV] for further information.

Theorem 6.3 considers the general case of nonsymmetric \mathbf{A} and a somewhat stringent notion of preconditioning. For the important special case of symmetric positive definite \mathbf{A} , Knyazev and Neymeyr [23] provide convergence estimates (and review much literature) for the one-sided forward Euler discretization (6.3). They provide

rates of convergence given a symmetric positive definite preconditioner \mathbf{N} for \mathbf{A} . However, a connection with dynamical systems is not made and instead optimization is applied to the Rayleigh quotient.

If $\mathbf{M} = \mathbf{I}$, and \mathbf{C} is normal (which is possible even if \mathbf{A} itself is not normal due to $\mathbf{d} \neq \mathbf{0}$) with spectrum given by $\sigma(\mathbf{C}) = \{\lambda_2, \dots, \lambda_n\}$, we can estimate an optimal time-step as follows. We wish to minimize

$$\gamma = \max_{i=2, \dots, n} |1 + h(\lambda_1 - \lambda_i)|,$$

a simple minimax approximation problem on a discrete set; see, e.g., [36, section 8.5]. In particular, if all the eigenvalues are real (i.e., \mathbf{C} is symmetric) and $\lambda_2 \leq \lambda_3 \leq \dots \leq \lambda_n$, then the best h must give

$$1 + h(\lambda_1 - \lambda_2) = -1 - h(\lambda_1 - \lambda_n).$$

This can be solved to obtain $h = 2/(\lambda_2 + \lambda_n - 2\lambda_1)$, from which we compute

$$\gamma = \frac{\lambda_n - \lambda_2}{\lambda_n + \lambda_2 - 2\lambda_1}.$$

Notice that this agrees with the convergence rate of the power method applied to $\mathbf{A} - \sigma\mathbf{I}$ for the optimal shift $\sigma = \frac{1}{2}(\lambda_2 + \lambda_n)$ to the leftmost eigenvector \mathbf{x}_1 ; see, e.g., [43, p. 572]. With the optimal choice of h , the forward Euler method recovers the convergence rate of an optimally shifted power method to \mathbf{x}_1 .

Again, suppose that $\mathbf{M} = \mathbf{I}$, so that $\gamma = \gamma(h) \rightarrow 1$ as $h \rightarrow 0$. However, this limit need not be approached from below; that is, for some matrices \mathbf{C} we will have $\gamma(h) > 1$ for all h sufficiently small.² The behavior of γ in this limit bears a close connection to the *logarithmic norm* of $\lambda_1 - \mathbf{C}$, which is defined as

$$\beta(\lambda_1 - \mathbf{C}) := \lim_{h \downarrow 0} \frac{\|\mathbf{I} + h(\lambda_1 - \mathbf{C})\| - 1}{h};$$

see, e.g., [30], [40, Chapter 17]. In particular, $\gamma(h) < 1$ for all sufficiently small $h > 0$ provided $\beta(\lambda_1 - \mathbf{C}) < 0$. One can show that the logarithmic norm of a matrix coincides with the numerical abscissa, that is, the real part of the rightmost point in the numerical range:

$$\begin{aligned} \beta(\lambda_1 - \mathbf{C}) &= \max_{\mathbf{v} \in \mathbb{C}^{n-1}, \|\mathbf{v}\|=1} \operatorname{Re} \mathbf{v}^*(\lambda_1 - \mathbf{C})\mathbf{v} \\ &= \max \left\{ \eta : \eta \in \sigma\left(\frac{1}{2}((\lambda_1 - \mathbf{C}) + (\lambda_1 - \mathbf{C}^T))\right) \right\}; \end{aligned}$$

see, e.g., [40, Theorem 17.4]. When is $\gamma(h) > 1$? That is, for what matrices can we not apply our convergence theory by taking h arbitrarily small? We can answer this question by finding requirements on \mathbf{C} that ensure $\beta(\lambda_1 - \mathbf{C}) < 0$. From the above analysis we see that

$$\beta(\lambda_1 - \mathbf{C}) = \lambda_1 - \min_{\mathbf{v} \in \mathbb{C}^{n-1}, \|\mathbf{v}\|=1} \operatorname{Re} \mathbf{v}^* \mathbf{C} \mathbf{v}.$$

Since \mathbf{C} is essentially the restriction $\mathbf{A}|_{\mathbf{x}_1^\perp}$ of \mathbf{A} to the orthogonal complement of the eigenvector \mathbf{x}_1 , we can summarize as follows.

LEMMA 6.4. *Suppose $\mathbf{N} = \mathbf{I}$. Then $\gamma < 1$ for all h sufficiently small if and only if λ_1 is not in the numerical range of $\mathbf{A}|_{\mathbf{x}_1^\perp}$ (equivalently, \mathbf{C}).*

²In this case the matrix \mathbf{A} does not satisfy the hypotheses of the theorem; convergence is still possible. Experiments with a small example gave convergence after a bit of initial irregularity.

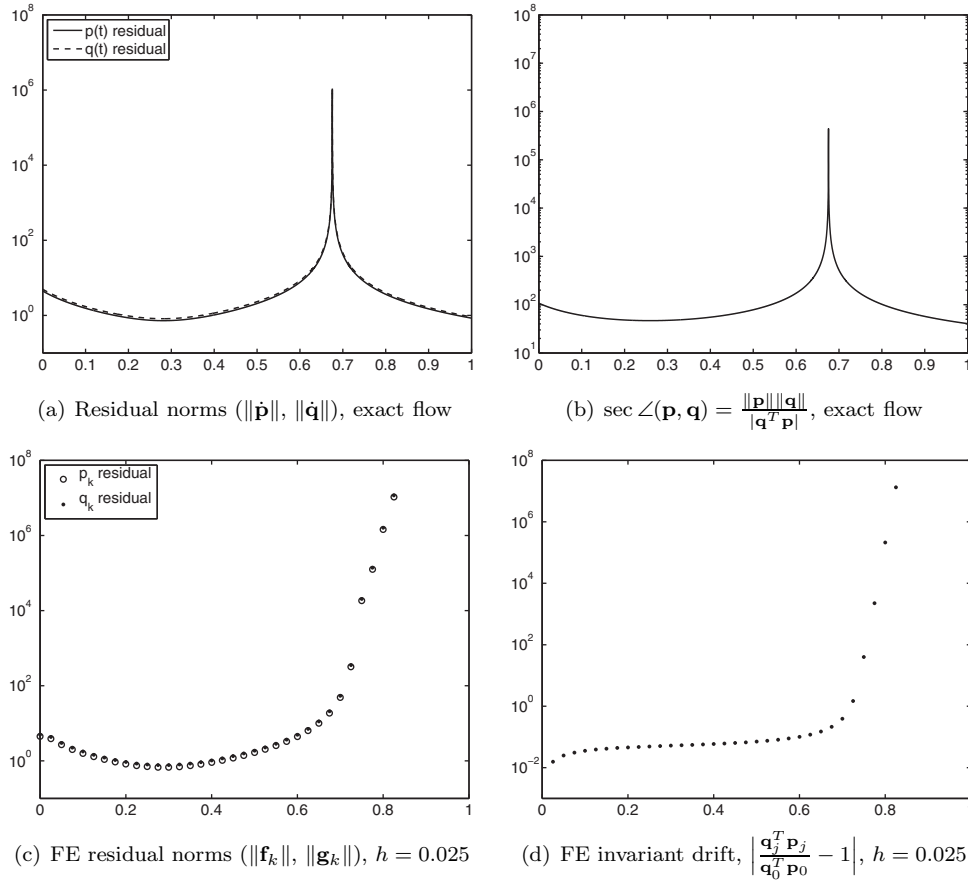


FIG. 6.1. Sampled flow and forward Euler (FE) approximations for the two-sided system with \mathbf{T}_ρ^{100} and $\rho = 1/(20 \cdot 101)$. The horizontal axis denotes time. Note the blow-up of the exact solution near $t = 0.675$, and the consequences of this behavior for the discretized method.

6.3. Numerical experiments. In this section we investigate Theorems 4.1, 6.1, and 6.3 through several computational examples. Our first experiment applies to the tridiagonal matrix

$$\mathbf{T}_\rho^n \equiv \begin{bmatrix} 2 & -1 + \rho & & 0 \\ -1 - \rho & 2 & \ddots & \\ & \ddots & \ddots & -1 + \rho \\ 0 & & -1 - \rho & 2 \end{bmatrix} \in \mathbb{R}^{n \times n},$$

where $n = 100$ and $\rho = 1/(20(n + 1))$. The eigenvalues are all real and the condition number of the matrix of eigenvectors is modest. All computations in Figure 6.1 use the same starting vectors \mathbf{p}_0 and \mathbf{q}_0 , which are taken to be (different) random vectors. (Results vary with the other choices for these vectors.)

Figures 6.1(a) and 6.1(b) show the exact solution to the two-sided unpreconditioned system, as given by Theorem 4.1. The residuals $\|\cdot \mathbf{p}\| = \|\mathbf{p}\theta - \mathbf{A}\mathbf{p}\|$ and $\|\cdot \mathbf{q}\| = \|\mathbf{q}\bar{\theta} - \mathbf{A}^*\mathbf{q}\|$ begin to decrease, but then rise as t approaches a critical point

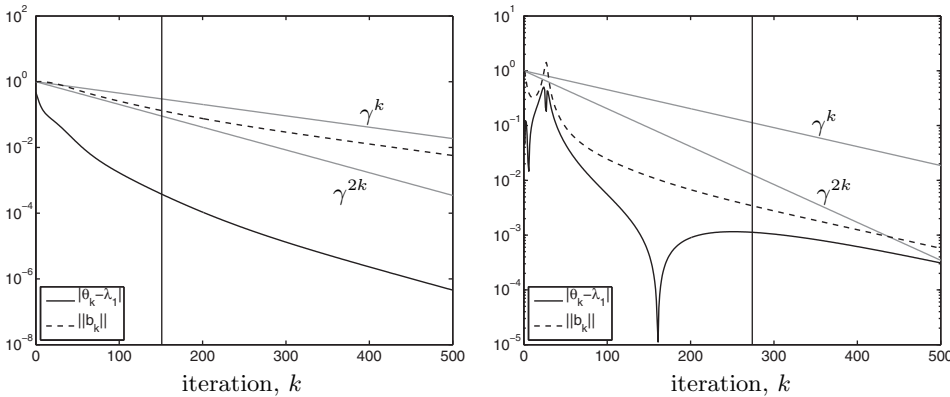


FIG. 6.2. Computational confirmation of Theorem 6.3 for a normal matrix (left) and a non-normal matrix (right), both with $\mathbf{N} = \mathbf{I}$. In the normal case, the residual $|\theta_k - \lambda|$ converges like γ^{2k} , while in the nonnormal case $|\theta_k - \lambda|$ only converges like γ^k . The vertical lines denote the point at which the hypotheses of the convergence theorem hold.

near $t = 0.675$, where cusps develop, indicating that a pole as given by $\pi(t)$ of Theorem 4.1 is encountered by the flow. The same behavior is seen in a plot of the secant of the angle between \mathbf{p} and \mathbf{q} . Figures 6.1(c) and 6.1(d) display the discrete flow associated with a forward Euler time integrator with a time step of $h = 0.025$. As expected, when the iterates depart from the quadratic manifold, the residuals explode in size, as in the exact solution. One can also show that the secant of the angle between \mathbf{p}_j and \mathbf{q}_j , and the norms of \mathbf{p}_j and \mathbf{q}_j , also begin to grow near $t \approx .675$, consistent with Theorem 6.1.

Decreasing the time-step h does not avoid the blow-up—in fact, the time at which the explosive growth occurs is largely independent of the time-step because of the onset of incurable breakdown associated with the continuous dynamical system. In contrast to the latter, the discrete dynamical system cannot simply step over the pole associated with continuous dynamical system. Aside from special cases such as the one described by Theorem 4.2, these results appear to be common and do not significantly depend on specially engineered starting vectors (though breakdown will occur at different points in time, of course). We also implemented the symplectic Euler method (that preserves quadratic invariants) for this class of matrices and observed behavior consistent with the forward Euler method combined with a projection. In contrast, the one-sided discretized forward Euler iterations converge to the left eigenvalue and associated eigenvector.

Next, we investigate the convergence analysis described in Theorem 6.3 for a simple example with $\mathbf{N} = \mathbf{I}$. Let \mathbf{A} be the matrix with $a_{j,j} = (j - 1)/(N - 1)$ for $j = 1, \dots, N$, and all other entries equal to zero except perhaps for the vector \mathbf{d}^T in entries 2 through N of the first row; cf. (6.11). The plots in Figure 6.2 use $N = 64$, comparing $\mathbf{d}^T = \mathbf{0}$ (left) and $\mathbf{d}^T = [1, \dots, 1]$ (right). In both cases we take $h = 1/2$, for which (6.14) gives $\gamma = 0.992 \dots \in [0, 1]$ as required. We take \mathbf{p}_0 to be the same randomly generated unit vector in both cases. This initial vector does not satisfy (6.17), but this condition is eventually met after a number of iterations, denoted by the vertical line in each plot. For the normal case in the left plot, $\|\mathbf{b}_k\|$ converges like γ^k , while the error in the Rayleigh quotient $|\theta_k - \lambda_1|$ converges like γ^{2k} as predicted. The nonnormality induced by the \mathbf{d} vector spoils this convergence for the Rayleigh quotient, as seen in the right plot; now both $\|\mathbf{b}_k\|$ and $|\theta_k - \lambda_1|$ converge

like γ^k , consistent with Theorem 6.3. The spikes in the latter plot correspond to points where the Rayleigh quotient θ_k crossed over the desired eigenvalue λ_1 , something only possible for nonnormal iterations.

7. Summary. This paper demonstrates the fruitful relationship between several nonlinear dynamical systems and certain simple preconditioned eigensolvers for non-symmetric eigenvalue problems. Properties of the continuous-time systems, such as system invariants and the asymptotic behavior of the exact solution, can inform the convergence theory for practical algorithms derived from discretizations, as we illustrate with Theorem 6.1 for the forward Euler discretization. Generalizations to more sophisticated discretizations, along with relaxation of the stringent requirements on the preconditioner in Theorem 6.1, are natural avenues for future research.

Appendix A. Lipschitz constant for Euler's method. To apply the standard convergence theory for the forward Euler method applied to the system

$$\dot{\mathbf{p}} = \mathbf{N}^{-1}(\theta(\mathbf{p})\mathbf{p} - \mathbf{A}\mathbf{p}),$$

we seek a constant $L > 0$ such that

$$\|\mathbf{N}^{-1}(\theta(\mathbf{u})\mathbf{u} - \mathbf{A}\mathbf{u}) - \mathbf{N}^{-1}(\theta(\mathbf{v})\mathbf{v} - \mathbf{A}\mathbf{v})\| \leq L\|\mathbf{u} - \mathbf{v}\|$$

for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. First we note that

$$\|(\theta(\mathbf{u})\mathbf{u} - \mathbf{A}\mathbf{u}) - (\theta(\mathbf{v})\mathbf{v} - \mathbf{A}\mathbf{v})\| \leq \|\theta(\mathbf{u})\mathbf{u} - \theta(\mathbf{v})\mathbf{v}\| + \|\mathbf{A}\|\|\mathbf{u} - \mathbf{v}\|.$$

We focus attention on the first term on the right:

$$\begin{aligned} \|\theta(\mathbf{u})\mathbf{u} - \theta(\mathbf{v})\mathbf{v}\| &\leq \|\theta(\mathbf{u})\mathbf{u} - \theta(\mathbf{v})\mathbf{u} + \theta(\mathbf{v})\mathbf{u} - \theta(\mathbf{v})\mathbf{v}\| \\ &\leq |\theta(\mathbf{u}) - \theta(\mathbf{v})|\|\mathbf{u}\| + |\theta(\mathbf{v})|\|\mathbf{u} - \mathbf{v}\| \\ (A.1) \qquad \qquad \qquad &\leq |\theta(\mathbf{u}) - \theta(\mathbf{v})|\|\mathbf{u}\| + \|\mathbf{A}\|\|\mathbf{u} - \mathbf{v}\|. \end{aligned}$$

(In this last inequality and others that follow, we neglect the opportunity to take tighter bounds that would lead to smaller constants but greater analytical complexity.)

Next, we need to bound $|\theta(\mathbf{u}) - \theta(\mathbf{v})|\|\mathbf{u}\|$ in terms of $\|\mathbf{u} - \mathbf{v}\|$. For convenience (assuming neither \mathbf{u} nor \mathbf{v} is zero), define the unit vectors $\hat{\mathbf{u}} = \mathbf{u}/\|\mathbf{u}\|$ and $\hat{\mathbf{v}} = \mathbf{v}/\|\mathbf{v}\|$, with $\boldsymbol{\varepsilon} = \hat{\mathbf{v}} - \hat{\mathbf{u}}$, so that

$$\begin{aligned} |\theta(\mathbf{u}) - \theta(\mathbf{v})| &= |\hat{\mathbf{u}}^T \mathbf{A} \hat{\mathbf{u}} - \hat{\mathbf{v}}^T \mathbf{A} \hat{\mathbf{v}}| \\ &= |\hat{\mathbf{u}}^T \mathbf{A} \hat{\mathbf{u}} - \hat{\mathbf{u}}^T \mathbf{A} \hat{\mathbf{u}} - \boldsymbol{\varepsilon}^T \mathbf{A} \hat{\mathbf{u}} - \hat{\mathbf{u}}^T \mathbf{A} \boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T \mathbf{A} \boldsymbol{\varepsilon}| \\ (A.2) \qquad \qquad \qquad &\leq 2\|\boldsymbol{\varepsilon}\|\|\mathbf{A}\| + \|\boldsymbol{\varepsilon}\|^2\|\mathbf{A}\|. \end{aligned}$$

Now note that

$$\|\boldsymbol{\varepsilon}\| = \|\hat{\mathbf{v}} - \hat{\mathbf{u}}\| = \frac{\|\|\mathbf{u}\|\mathbf{v} - \|\mathbf{v}\|\mathbf{u} + \|\mathbf{v}\|\mathbf{v} - \|\mathbf{v}\|\mathbf{u}\|}{\|\mathbf{u}\|\|\mathbf{v}\|} \leq \frac{\|\|\mathbf{u}\| - \|\mathbf{v}\|\|}{\|\mathbf{u}\|} + \frac{\|\mathbf{u} - \mathbf{v}\|}{\|\mathbf{u}\|}.$$

Apply the triangle inequality to obtain $\|\|\mathbf{u}\| - \|\mathbf{v}\|\| \leq \|\mathbf{u} - \mathbf{v}\|$, from which we conclude

$$(A.3) \qquad \qquad \qquad \|\boldsymbol{\varepsilon}\| \leq \frac{2}{\|\mathbf{u}\|}\|\mathbf{u} - \mathbf{v}\|.$$

Since $\widehat{\mathbf{u}}$ and $\widehat{\mathbf{v}}$ are unit vectors, we alternatively have the coarse bound $\|\boldsymbol{\varepsilon}\| = \|\widehat{\mathbf{u}} - \widehat{\mathbf{v}}\| \leq 2$, which we can apply to (A.2) to obtain

$$\begin{aligned} |\theta(\mathbf{u}) - \theta(\mathbf{v})| &\leq 2\|\boldsymbol{\varepsilon}\|\|\mathbf{A}\| + \|\boldsymbol{\varepsilon}\|^2\|\mathbf{A}\| \\ &\leq 2\|\boldsymbol{\varepsilon}\|\|\mathbf{A}\| + 2\|\boldsymbol{\varepsilon}\|\|\mathbf{A}\| = 4\|\mathbf{A}\|\|\boldsymbol{\varepsilon}\|. \end{aligned}$$

Now using (A.3), the bound first bound on $\|\boldsymbol{\varepsilon}\|$,

$$|\theta(\mathbf{u}) - \theta(\mathbf{v})| \leq 8 \frac{\|\mathbf{A}\|}{\|\mathbf{u}\|} \|\mathbf{u} - \mathbf{v}\|.$$

Substituting this bound into (A.1) gives

$$\|\theta(\mathbf{u})\mathbf{u} - \theta(\mathbf{v})\mathbf{v}\| \leq 9\|\mathbf{A}\|\|\mathbf{u} - \mathbf{v}\|,$$

and, finally, we arrive at the Lipschitz constant

$$\|\mathbf{N}^{-1}(\theta(\mathbf{u})\mathbf{u} - \mathbf{A}\mathbf{u}) - \mathbf{N}^{-1}(\theta(\mathbf{v})\mathbf{v} - \mathbf{A}\mathbf{v})\| \leq 10 \|\mathbf{N}^{-1}\| \|\mathbf{A}\| \|\mathbf{u} - \mathbf{v}\|.$$

Thus, we define

$$(A.4) \quad L = 10 \|\mathbf{N}^{-1}\| \|\mathbf{A}\|.$$

The Rayleigh quotient $\theta(\mathbf{p})$ is undefined in the case that $\mathbf{p} = \mathbf{0}$. However, as $\|\mathbf{p}\| \rightarrow 0$, we have that $\|\theta(\mathbf{p})\mathbf{p} - \mathbf{A}\mathbf{p}\| \rightarrow 0$, and this motivates the definition that $\theta(\mathbf{p})\mathbf{p} - \mathbf{A}\mathbf{p} = \mathbf{0}$ if $\mathbf{p} = \mathbf{0}$.

The above analysis excludes the case that $\mathbf{u} = \mathbf{0}$ and/or $\mathbf{v} = \mathbf{0}$, but with our definition of this singular case we have, e.g., if $\mathbf{u} = \mathbf{0}$, that

$$\|(\theta(\mathbf{u})\mathbf{u} - \mathbf{A}\mathbf{u}) - (\theta(\mathbf{v})\mathbf{v} - \mathbf{A}\mathbf{v})\| = \|(\theta(\mathbf{v})\mathbf{v} - \mathbf{A}\mathbf{v})\| \leq 2\|\mathbf{A}\|\|\mathbf{v}\| \leq 10\|\mathbf{A}\|\|\mathbf{u} - \mathbf{v}\|,$$

and obviously if $\mathbf{u} = \mathbf{v} = \mathbf{0}$, we have

$$\|(\theta(\mathbf{u})\mathbf{u} - \mathbf{A}\mathbf{u}) - (\theta(\mathbf{v})\mathbf{v} - \mathbf{A}\mathbf{v})\| = 0 = 10\|\mathbf{A}\|\|\mathbf{u} - \mathbf{v}\|.$$

Hence, the Lipschitz constant (A.4) holds for all \mathbf{u} and \mathbf{v} .

Acknowledgments. We thank Pierre-Antoine Absil, Moody Chu, Kyle Gallivan, Anthony Kellems, Christian Lubich, and Qiang Ye, and anonymous referees for their numerous helpful suggestions concerning this work and its presentation.

REFERENCES

- [1] P.-A. ABSIL, *Continuous-time systems that solve computational problems*, Int. J. Uncov. Comput., 2 (2006), pp. 291–304.
- [2] P.-A. ABSIL, R. MAHONY, AND R. SEPULCHRE, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ, 2008.
- [3] P.-A. ABSIL, R. SEPULCHRE, AND R. MAHONY, *Continuous-time subspace flows related to the symmetric eigenproblem*, Pacific J. Optim., 4 (2008), pp. 179–194.
- [4] V. I. ARNOLD, *Ordinary Differential Equations*, 3rd ed., Springer-Verlag, Berlin, 1992.
- [5] Z. BAI, D. DAY, AND Q. YE, *ABLE: An adaptive block Lanczos method for non-Hermitian eigenvalue problems*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 1060–1082.
- [6] Z. BAI, J. DEMMEL, J. DONGARRA, A. RUHE, AND H. VAN DER VORST, EDS., *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, SIAM, Philadelphia, 2000.

- [7] W. BAO AND Q. DU, *Computing the ground state solution of Bose–Einstein condensates by a normalized gradient flow*, SIAM J. Sci. Comp., 25 (2004), pp. 1674–1697.
- [8] R. CAR AND M. PARRINELLO, *Unified approach for molecular dynamics and density functional theory*, Phys. Rev. Lett., 55 (1985), pp. 2471–2474.
- [9] J. CARR, *Applications of Centre Manifold Theory*, Springer-Verlag, New York, 1981.
- [10] M. T. CHU, *Curves on s^{n-1} that lead to eigenvalues or their means of a matrix*, SIAM J. Alg. Disc. Math., 7 (1986), pp. 425–432.
- [11] M. T. CHU, *On the continuous realization of iterative processes*, SIAM Rev., 30 (1988), pp. 375–387.
- [12] M. EMBREE, *The Arnoldi eigenvalue iteration with exact shifts can fail*, SIAM J. Matrix Anal. Appl., to appear.
- [13] M. A. FREITAG AND A. SPENCE, *Convergence theory for inexact inverse iteration applied to the generalised nonsymmetric eigenvalue problem*, Electron. Trans. Numer. Anal., 28 (2007), pp. 40–64.
- [14] C. W. GEAR, *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
- [15] G. H. GOLUB AND L.-Z. LIAO, *Continuous methods for extreme and interior eigenvalue problems*, Linear Algebra Appl., 415 (2006), pp. 31–51.
- [16] G. H. GOLUB AND Q. YE, *Inexact inverse iteration for generalized eigenvalue problems*, BIT, (2000), pp. 671–684.
- [17] J. GUCKENHEIMER AND P. HOLMES, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.
- [18] E. HAIRER, C. LUBICH, AND G. WANNER, *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, 2nd ed., Springer-Verlag, Berlin, 2006.
- [19] U. HELMKE AND J. B. MOORE, *Optimization and Dynamical Systems*, Springer, London, 1994.
- [20] W. KAHAN, B. N. PARLETT, AND E. JIANG, *Residual bounds on approximate eigensystems of nonnormal matrices*, SIAM J. Num. Anal., 19 (1982), pp. 470–484.
- [21] A. V. KNYAZEV, *Preconditioned eigensolvers—an oxymoron?*, Elec. Trans. Numer. Anal., 7 (1998), pp. 104–123.
- [22] A. V. KNYAZEV AND K. NEYMEYR, *Efficient solution of symmetric eigenvalue problems using multigrid preconditioners in the locally optimal block conjugate gradient method*, Elec. Trans. Numer. Anal., 7 (2003), pp. 38–55.
- [23] A. V. KNYAZEV AND K. NEYMEYR, *A geometric theory for preconditioned inverse iteration. III: A short and sharp convergence estimate for generalized eigenvalue problems*, Linear Algebra Appl., 358 (2003), pp. 95–114.
- [24] Y.-L. LAI, K.-Y. LIN, AND W.-W. LIN, *An inexact inverse iteration for large sparse eigenvalue problems*, Numer. Linear Algebra Appl., 1 (1997), pp. 1–13.
- [25] R. B. LEHOUCQ AND A. J. SALINGER, *Large-scale eigenvalue calculations for stability analysis of steady flows on massively parallel computers*, Internat. J. Numer. Methods Fluids, 36 (2001), pp. 309–327.
- [26] B. LEIMKUHNER AND S. REICH, *Simulating Hamiltonian Dynamics*, Cambridge University Press, Cambridge, 2005.
- [27] R. MAHONY AND P.-A. ABSIL, *The continuous time Rayleigh quotient flow on the sphere*, Linear Algebra Appl., 368 (2003), pp. 343–357.
- [28] Y. NAKAMURA, K. KAJIWARA, AND H. SHIOTANI, *On an integrable discretization of the Rayleigh quotient gradient system and the power method with a shift*, J. Comput. Appl. Math., 96 (1998), pp. 77–90.
- [29] T. NANDA, *Differential equations and the QR algorithm*, SIAM J. Numer. Anal., 22 (1985), pp. 310–321.
- [30] O. NEVANLINNA, *Convergence of Iterations for Linear Equations*, Birkhäuser, Basel, 1993.
- [31] E. E. OSBORNE, *On pre-conditioning of matrices*, J. ACM, 7 (1960), pp. 338–345.
- [32] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, no. 20 in Classics in Applied Mathematics, SIAM, Philadelphia, 1998. Amended reprint of 1980 Prentice-Hall edition.
- [33] B. N. PARLETT AND C. REINSCH, *Balancing a matrix for calculation of eigenvalues and eigenvectors*, Numer. Math., 13 (1969), pp. 293–304.
- [34] M. C. PAYNE, M. P. TEETER, D. C. ALLAN, T. ARIAS, AND J. JOANNOPOULOS, *Iterative minimization techniques for ab initio total-energy calculations: Molecular dynamics and conjugate gradients*, Rev. Mod. Phys, 64 (1992), pp. 1045–1097.
- [35] B. T. POLYAK, *Introduction to Optimization*, Translation Series in Mathematics and Engineering, Optimization Software, Inc., New York, 1987.
- [36] M. J. D. POWELL, *Approximation Theory and Methods*, Cambridge University Press, Cambridge, 1981.

- [37] Y. SAAD, *Variations on Arnoldi's method for computing eigenlements of large unsymmetric matrices*, *Linear Algebra Appl.*, 34 (1980), pp. 269–295.
- [38] G. W. STEWART AND J.-G. SUN, *Matrix Perturbation Theory*, Academic Press, San Diego, CA, 1990.
- [39] W. W. SYMES, *The QR algorithm and scattering for the finite nonperiodic Toda lattice*, *Physica D*, 4 (1982), pp. 275–280.
- [40] L. N. TREFETHEN AND M. EMBREE, *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*, Princeton University Press, Princeton, NJ, 2005.
- [41] J. S. WARSA, T. A. WAREING, J. E. MOREL, J. M. MCGHEE, AND R. B. LEHOUCQ, *Krylov subspace iterations for deterministic k -eigenvalue calculations*, *Nuc. Sci. Engrg.*, 147 (2004), pp. 26–42.
- [42] D. S. WATKINS, *Isospectral flows*, *SIAM Rev.*, 26 (1984), pp. 379–391.
- [43] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, 1965.

A NEW FICTITIOUS DOMAIN APPROACH INSPIRED BY THE EXTENDED FINITE ELEMENT METHOD*

JAROSLAV HASLINGER[†] AND YVES RENARD[‡]

Abstract. The purpose of this paper is to present a new fictitious domain approach inspired by the extended finite element method introduced by Moës, Dolbow, and Belytschko in [*Internat. J. Numer. Methods Engrg.*, 46 (1999), pp. 131–150]. An optimal method is obtained thanks to an additional stabilization technique. Some a priori estimates are established and numerical experiments illustrate different aspects of the method. The presentation is made on a simple Poisson problem with mixed Neumann and Dirichlet boundary conditions. The extension to other problems or boundary conditions is quite straightforward.

Key words. fictitious domain, Xfem, approximation of elliptic problems, stabilization technique

AMS subject classifications. 65N30, 65N15

DOI. 10.1137/070704435

1. Introduction. The extended finite element method (Xfem) was introduced by Moës, Dolbow, and Belytschko in [18] and developed in many papers such as [5, 16, 19, 23, 28]. The first application of Xfem was done in structural mechanics when dealing with cracked domains. The specificity of the method is that it combines a level-set representation of the geometry of the crack (introduced in [25]) with an enrichment of a finite element space by singular and discontinuous functions. The enrichment of a finite element space with a singular function has been studied earlier by Strang and Fix in [26]. The originality of Xfem consists in a particular way of defining the enrichment via the multiplication by a partition of unity provided by basis functions of a Lagrange finite element method. Several strategies can be considered in order to extend or improve the original Xfem. Some of these strategies are presented in [16]. An a priori error estimate of a variant of Xfem for cracked domains is presented in [5].

In this work we adapt the techniques of Xfem to develop a new method allowing computations in domains whose boundaries are independent of the mesh. A similar attempt was done in [17, 27]. Our goal is to develop a fully optimal method. It can be considered as a fictitious domain-type method. Its advantage, compared to existing ones (see, for instance, [11, 13]), is its ability to easily treat complex boundary conditions. The elementary matrices, however, have to be computed taking into account the geometry of the real boundary (in a nonlinear framework this disadvantage disappears since the tangent stiffness matrix has to be frequently recomputed).

Therefore, this method can be of interest for computational domains having moving boundaries or boundaries with a complex geometry and various conditions on them (Dirichlet, Neumann, Signorini, ...). In this paper, only Dirichlet and Neu-

*Received by the editors October 5, 2007; accepted for publication (in revised form) November 18, 2008; published electronically March 25, 2009. This work was supported by “l’Agence Nationale de la Recherche,” project ANR-05-JCJC-0182-01.

<http://www.siam.org/journals/sinum/47-2/70443.html>

[†]Department of Numerical Mathematics, Faculty of Mathematics and Physics, Sokolovská 83, 18675 Praha 8, Czech Republic (Jaroslav.Haslinger@mff.cuni.cz). This author’s research was supported by grant MSM0021620839 of the Czech Ministry of Education and IAA100750802 of GAAV CR.

[‡]Université de Lyon, CNRS, INSA-Lyon, ICJ UMR5208, LaMCoS UMR5259, F-69621, Villeurbanne, France (Yves.Renard@insa-lyon.fr).

mann boundary conditions are considered. An extension to more complex boundary data is straightforward, at least from the implementation point of view.

The outline of this paper is as follows. In section 1, we introduce the model problem which is represented by a simple Poisson equation with Neumann and Dirichlet boundary conditions. In section 2 we describe the new method for a model problem without any stabilization. Section 3 is devoted to a convergence analysis of this approach. An abstract result is obtained which gives a convergence rate of order \sqrt{h} under reasonable regularity assumptions on the solution even for high order finite elements. The main part of this paper is section 4 where a new stabilized method is introduced. Under appropriate assumptions we prove the stability of this formulation as well as optimal error estimates. In section 5 we briefly mention details on the computational implementation. Numerical experiments for a model example with different choices of finite element spaces are presented in section 6. The paper is completed with three appendices with proofs of trace theorems needed in the text.

2. Setting of the problem. We present a new approach for numerical realization of elliptic problems. The theoretical presentation is made for a two or three-dimensional simply connected bounded domain Ω with a sufficiently smooth boundary. Let $\tilde{\Omega} \subset \mathbb{R}^d$ ($d = 2$ or $d = 3$) be a rectangular or parallelepiped domain (the fictitious domain) containing Ω in its interior. We consider that the boundary Γ of Ω is split into two parts Γ_N and Γ_D (see Figure 1). It is assumed that Γ_D has a nonzero $(d - 1)$ -dimensional Lebesgue measure.

Let us consider the following problem in Ω :

Find $u : \Omega \mapsto \mathbb{R}$ such that

- (1) $-\Delta u = f$ in Ω ,
- (2) $u = 0$ on Γ_D ,
- (3) $\partial_n u = g$ on Γ_N ,

where $f \in L^2(\Omega)$, $g \in L^2(\Gamma_N)$ are given data and n is the outward unit normal vector to Γ . The weak formulation of such a problem is well known and reads as follows:

$$(4) \quad \begin{cases} \text{Find } u \in V_0 \text{ such that} \\ a(u, v) = l(v) \quad \forall v \in V_0, \end{cases}$$

where

$$V = H^1(\Omega), \quad V_0 = \{v \in V : v = 0 \text{ on } \Gamma_D\},$$

$$a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v d\Omega, \quad l(v) = \int_{\Omega} f v d\Omega + \int_{\Gamma_N} g v d\Gamma.$$

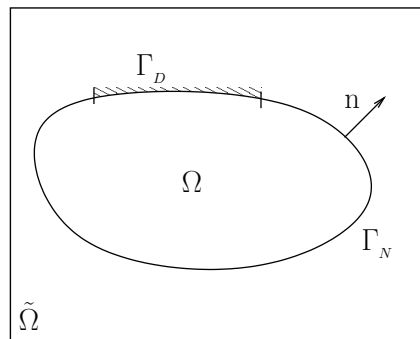


FIG. 1. Fictitious and real domains.

It is also well known that this problem can be expressed by means of the following mixed formulation:

$$(5) \quad \begin{cases} \text{Find } u \in V \text{ and } \lambda \in W \text{ such that} \\ a(u, v) + \langle \lambda, v \rangle_{W, X} = l(v) \quad \forall v \in V, \\ \langle \mu, u \rangle_{W, X} = 0 \quad \forall \mu \in W, \end{cases}$$

where $X = \{w \in L^2(\Gamma_D) : \exists v \in V \text{ such that } w = v|_{\Gamma_D}\}$, $W = X'$, and $\langle \mu, v \rangle_{W, X}$ denotes the duality pairing between W and X . Let

$$V_0^\# = \left\{ v \in V : \int_{\Gamma_D} v d\Gamma = 0 \right\}.$$

Then $a(\cdot, \cdot)$ is coercive on $V_0^\#$ (a direct consequence of Peetre–Tartar lemma, see [10] for instance), i.e., there exists $\alpha > 0$ such that

$$(6) \quad a(v, v) \geq \alpha \|v\|_V^2 \quad \forall v \in V_0^\#.$$

From this, the existence and uniqueness of a solution to Problem (5) follows. In addition, $\lambda = -\partial_n u$ on Γ_D . Problem (5) is also equivalent to the problem of finding a saddle point of the following Lagrangian on $V \times W$:

$$(7) \quad \mathcal{L}(v, \mu) = \frac{1}{2} a(v, v) + \langle \mu, v \rangle_{W, X} - l(v).$$

3. The new fictitious domain method. The new fictitious domain approach which will be studied in this paper requires the introduction of two finite dimensional finite element spaces $\tilde{V}^h \subset H^1(\tilde{\Omega})$ and $\tilde{W}^h \subset L^2(\tilde{\Omega})$ on the fictitious domain $\tilde{\Omega}$. As $\tilde{\Omega}$ can be a rectangular or parallelepiped domain, the ones can be defined on the same structured mesh \mathcal{T}^h (see Figure 2). Note that in the following, we only use the fact that the family of meshes is quasi-uniform (in the classical sense of Ciarlet [6, 7]). Next we shall suppose that

$$(8) \quad \tilde{V}^h = \left\{ v^h \in \mathcal{C}(\tilde{\Omega}) : v^h|_T \in P(T) \quad \forall T \in \mathcal{T}^h \right\},$$

where $P(T)$ is a finite dimensional space of regular functions such that $P(T) \supseteq P_k(T)$ for some $k \geq 1$ integer. The mesh parameter h stands for $h = \max_{T \in \mathcal{T}^h} h_T$ where h_T is the diameter of T .

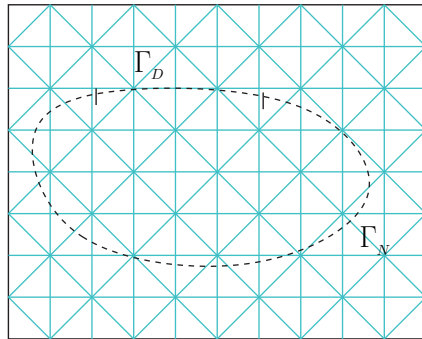


FIG. 2. Example of a structured mesh.

Then one can build

$$V^h := \widetilde{V}^h|_{\Omega}, \quad \text{and} \quad W^h := \widetilde{W}^h|_{\Gamma_D},$$

which are natural discretizations of V and W , respectively. An approximation of Problem (5) is defined as follows:

$$(9) \quad \begin{cases} \text{Find } u^h \in V^h \text{ and } \lambda^h \in W^h \text{ such that} \\ a(u^h, v^h) + \int_{\Gamma_D} \lambda^h v^h d\Gamma = l(v^h) \quad \forall v^h \in V^h, \\ \int_{\Gamma_D} \mu^h u^h d\Gamma = 0 \quad \forall \mu^h \in W^h. \end{cases}$$

Similarly to Xfem, where the shape functions of the finite element space are multiplied with a Heaviside function, this corresponds here to the multiplication of the shape functions with the characteristic function of Ω .

4. Convergence analysis. Let us define the following space:

$$(10) \quad V_0^h = \left\{ v^h \in V^h : \int_{\Gamma_D} \mu^h v^h d\Gamma = 0 \quad \forall \mu^h \in W^h \right\}.$$

This space can be viewed to be a (nonconforming) discretization of V_0 . In addition, we shall suppose that \widetilde{W}^h and \widetilde{V}^h are chosen in such a way that the following two conditions are satisfied for every $h > 0$:

$$(11) \quad 1|_{\Gamma_D} \in W^h,$$

$$(12) \quad \bar{\mu}^h \in W^h : \int_{\Gamma_D} \bar{\mu}^h v^h d\Gamma = 0 \quad \forall v^h \in V^h \implies \bar{\mu}^h = 0.$$

LEMMA 1. *The bilinear form $a(\cdot, \cdot)$ is uniformly V_0^h -elliptic; i.e., there exists $\alpha > 0$ independent of h such that*

$$a(v^h, v^h) \geq \alpha \|v^h\|_V \quad \forall v^h \in V_0^h.$$

Proof. It follows from the fact that $V_0^h \subset V_0^{\#}$. \square

PROPOSITION 1. *Suppose that (11) and (12) are satisfied. Then the solution (u^h, λ^h) to Problem (9) is unique and there exists a constant $C > 0$ independent of \widetilde{V}^h and \widetilde{W}^h such that¹*

$$\|u^h\|_V \leq C \|l\|_{H^{-1}(\Omega)}.$$

Proof. Since $1|_{\Gamma_D} \in W^h$, it follows from the last equality in (9) that $u^h \in V_0^{\#}$.

The existence and uniqueness of (u^h, λ^h) now follows from (12) and Lemma 1. The announced estimate comes from the fact that $a(u^h, u^h) = l(u^h)$. \square

We prove now the following abstract result (the extension of Cea’s lemma).

¹In what follows, the symbol C will be used to denote a generic positive constant which does not depend on h and which can take different values at different places of its appearance.

LEMMA 2. Let (u, λ) and (u^h, λ^h) be the solution to Problems (5) and (9), respectively. Suppose that (11) and (12) are satisfied. Then there exists a constant $C > 0$ independent of \tilde{V}^h and \tilde{W}^h such that

$$\|u - u^h\|_V \leq C \left(\inf_{v^h \in V_0^h} \|u - v^h\|_V + \sup_{v^h \in V_0^h, v^h \neq 0} \frac{|a(u, v^h) - l(v^h)|}{\|v^h\|_V} \right).$$

Proof. For a given function $v^h \in V_0^h$ one has

$$\begin{aligned} \alpha \|u^h - v^h\|_V^2 &\leq a(u^h - v^h, u^h - v^h) \\ &= a(u - v^h, u^h - v^h) + l(u^h - v^h) - a(u, u^h - v^h). \end{aligned}$$

Thus,

$$\|u^h - v^h\|_V \leq C \|u - v^h\|_V + \sup_{w^h \in V_0^h, w^h \neq 0} \frac{|a(u, w^h) - l(w^h)|}{\|w^h\|_V}.$$

From the triangle inequality $\|u - u^h\|_V \leq \|u - v^h\|_V + \|u^h - v^h\|_V$ we obtain the result. \square

Remark 1. The term $\sup_{v^h \in V_0^h, v^h \neq 0} \frac{|a(u, v^h) - l(v^h)|}{\|v^h\|_V}$ is called a consistency error.

COROLLARY 1. Under the assumptions of Lemma 2, there exists a constant $C > 0$ independent of \tilde{V}^h and \tilde{W}^h such that

$$(13) \quad \|u - u^h\|_V \leq C \left(\inf_{v^h \in V_0^h} \|u - v^h\|_V + \inf_{\mu^h \in W^h} \|\lambda - \mu^h\|_W \right).$$

Proof. Since u is a solution to Problem (5) one has

$$a(u, v^h) = l(v^h) - \langle \lambda, v^h \rangle_{W, X} \quad \forall v^h \in V_0^h.$$

The definition of V_0^h yields

$$a(u, v^h) - l(v^h) = -\langle \lambda, v^h \rangle_{W, X} = \langle \mu^h - \lambda, v^h \rangle_{W, X} \quad \forall v^h \in V_0^h \quad \forall \mu^h \in W^h,$$

so that

$$|a(u, v^h) - l(v^h)| \leq \inf_{\mu^h \in W^h} \|\lambda - \mu^h\|_W \|v^h\|_V \quad \forall v^h \in V_0^h.$$

This, together with Lemma 2, gives (13). \square

We establish now the following convergence result.

PROPOSITION 2. Suppose that (11) and (12) are satisfied and, in addition, let the system $\{V_0^h\}, \{W^h\}, h \rightarrow 0+$ be dense in V_0 and $L^2(\Gamma_D)$, respectively. Then

$$u^h \rightarrow u \quad \text{in } V, \quad h \rightarrow 0+,$$

where u and u^h are the first components of the solution to (5) and (9), respectively.

Proof. From Proposition 1 it follows that

$$\|u^h\|_V \leq C \quad \forall h > 0.$$

Thus, there exists a subsequence, still denoted by the same symbol and an element $\bar{u} \in V$ such that

$$(14) \quad u^h \rightharpoonup \bar{u} \quad \text{in } V, \quad h \rightarrow 0+.$$

Since $\{W^h\}$ is dense in $L^2(\Gamma_D)$, for any $\mu \in L^2(\Gamma_D)$ there exists a sequence $\{\mu^h\}$, $\mu^h \in W^h$ such that

$$(15) \quad \mu^h \rightarrow \mu \quad \text{in } L^2(\Gamma_D), \quad h \rightarrow 0+.$$

Passing to the limit in the last equality in (9), using (14) and (15) we see that

$$\int_{\Gamma_D} \mu \bar{u} d\Gamma = 0 \quad \forall \mu \in L^2(\Gamma_D),$$

which is equivalent to $\bar{u} \in V_0$. Let $\bar{v} \in V_0$ be given. Then, by the assumption there exists a sequence $\{\bar{v}^h\}$, $\bar{v}^h \in V_0^h$ such that

$$(16) \quad \bar{v}^h \rightarrow \bar{v} \quad \text{in } V, \quad h \rightarrow 0+.$$

Since u^h solves (9) we have

$$a(u^h, \bar{v}^h) = l(\bar{v}^h).$$

From this, (14), and (16) we see that

$$a(\bar{u}, \bar{v}) = l(\bar{v}) \quad \forall \bar{v} \in V_0,$$

i.e., $u := \bar{u}$ solves the original problem. As u is unique, the whole sequence $\{u^h\}$ tends weakly to u in V . Strong convergence of $\{u^h\}$ to u follows from the fact that

$$|u^h|_{1,\Omega} \rightarrow |u|_{1,\Omega},$$

which is easy to verify. \square

In what follows, we shall estimate the first term on the right of (13). To simplify our presentation we shall consider a purely homogeneous Dirichlet problem, i.e., with $\Gamma_D = \Gamma$ and such that its solution u belongs to $H^{1+d/2+\varepsilon}(\Omega) \cap H_0^1(\Omega)$ for some $\varepsilon > 0$ ($\Omega \subset \mathbb{R}^d$). From the embedding theorem it immediately follows that

$$(17) \quad u \in C^1(\bar{\Omega}).$$

For $\delta > 0$ given, we denote by Ω_δ the subset of Ω :

$$\Omega_\delta = \{x \in \Omega : \text{dist}(x, \Gamma) > \delta\}.$$

Let η_h be a sufficiently smooth cutoff function:

$$\eta_h = \begin{cases} 1 & \text{in } \Omega \setminus \Omega_{2h}, \\ 0 & \text{in } \Omega_{3h}. \end{cases}$$

In $\Omega_{2h} \setminus \Omega_{3h}$ the function η_h is defined in such a way that

$$(18) \quad \|\nabla^j \eta_h\|_{C(\bar{\Omega})} \leq \frac{C}{h^j}, \quad j = 1, 2.$$

The solution u can be split and written in the form

$$u = \eta_h u + (1 - \eta_h)u.$$

Next, we show that

$$(19) \quad \|\eta_h u\|_V \leq C\sqrt{h}, \quad h \rightarrow 0+.$$

Indeed,

$$(20) \quad \|\eta_h u\|_V^2 = \|u\|_{1,\Omega \setminus \Omega_{2h}}^2 + \|\eta_h u\|_{1,\Omega_{2h} \setminus \Omega_{3h}}^2.$$

From (17) it immediately follows that

$$\|u\|_{1,\Omega \setminus \Omega_{2h}}^2 \leq Ch, \quad h \rightarrow 0+.$$

To get the estimate of the second term on the right of (20) it is sufficient to estimate the respective seminorm. It holds:

$$(21) \quad \begin{aligned} |\eta_h u|_{1,\Omega_{2h} \setminus \Omega_{3h}}^2 &\leq C \left(\int_{\Omega_{2h} \setminus \Omega_{3h}} |\nabla \eta_h|^2 u^2 d\Omega + \int_{\Omega_{2h} \setminus \Omega_{3h}} \eta_h^2 |\nabla u|^2 d\Omega \right) \\ &\leq Ch, \quad h \rightarrow 0+, \end{aligned}$$

making use of (18) and the elementary estimate

$$(22) \quad \max_{x \in \Omega_{2h} \setminus \Omega_{3h}} |u(x)| \leq Ch,$$

which holds in view of the fact that $u = 0$ on Γ . From (21) and (22) we obtain (19).

Let V_{00}^h be a subset of V_h containing functions vanishing in a vicinity of Γ . More precisely,

$$V_{00}^h = \{v^h \in V^h : v^h(a) = 0 \quad \forall a \in \mathcal{N}^h\},$$

where \mathcal{N}^h is the set of those nodes of \mathcal{T}^h which lie in $\Omega \setminus \Omega_{3h/2}$. Observe that $V_{00}^h \subset V_0^h$.

By $\Pi_T v$ we denote the standard P -Lagrange interpolate of v on an element $T \in \mathcal{T}^h$, $T \subset \Omega$. Since $P \supseteq P_k$ ($k \geq 1$) we know that

$$(23) \quad \|v - \Pi_T v\|_{1,T} \leq Ch_T \|v\|_{2,T}$$

holds for any $v \in H^2(T)$, $T \in \mathcal{T}^h$ and $T \subset \Omega$.

PROPOSITION 3. *Suppose that \tilde{V}^h is defined by (8), let (11) and (12) be satisfied, and, in addition,*

$$(24) \quad \inf_{\mu^h \in W^h} \|\lambda - \mu^h\|_W \leq Ch^\beta, \quad \text{for some } \beta \geq 1/2.$$

Let the solution u of (4) with $\Gamma = \Gamma_D$ be such that $u \in H^{1+d/2+\varepsilon}(\Omega) \cap H_0^1(\Omega)$, $\varepsilon > 0$. Then

$$\|u - u^h\|_V \leq C\sqrt{h}, \quad h \rightarrow 0+.$$

Proof. It is sufficient to estimate the first term on the right of (13). It holds:

$$\begin{aligned} \inf_{v^h \in V_0^h} \|u - v^h\|_V &\leq \inf_{v^h \in V_{00}^h} \|u - v^h\|_V = \inf_{v^h \in V_{00}^h} \|\eta_h u + (1 - \eta_h)u - v^h\|_V \\ &\leq \|\eta_h u\|_V + \|(1 - \eta_h)u - v^h\|_V \quad \forall v^h \in V_{00}^h. \end{aligned}$$

We construct v^h as follows:

$$v^h|_T = \Pi_T \left((1 - \eta_h)u|_T \right) \quad \text{if } T \subset \Omega,$$

otherwise, we set $v^h = 0$. It is readily seen that $v^h \in V_{00}^h$ and from (23) it follows that

$$(25) \quad \|(1 - \eta_h)u - v^h\|_V \leq Ch\|(1 - \eta_h)u\|_{2,\Omega} \leq Ch\|u\|_{2,\Omega} + Ch\|\eta_h u\|_{2,\Omega}.$$

A direct computation shows that

$$(26) \quad \|\eta_h u\|_{2,\Omega} \leq \frac{C}{\sqrt{h}}, \quad h \rightarrow 0+.$$

Indeed, the $H^2(\Omega)$ -seminorm can be estimated by

$$|\eta_h u|_{2,\Omega}^2 \leq C \left(\|\nabla^2 \eta_h\|_{C(\bar{\Omega})}^2 \int_{\Omega_{2h} \setminus \Omega_{3h}} u^2 d\Omega + \int_{\Omega_{2h} \setminus \Omega_{3h}} |\nabla \eta_h|^2 |\nabla u|^2 d\Omega + |u|_{2,\Omega}^2 \right) \leq \frac{C}{h},$$

as follows from (18) and (22). Using (26) in (25) we see that

$$\|(1 - \eta_h)u - v^h\|_V \leq C\sqrt{h}, \quad h \rightarrow 0+.$$

From this and (19) we finally arrive at

$$\inf_{v^h \in V_0^h} \|u - v^h\|_V \leq C\sqrt{h}. \quad \square$$

The convergence rate given by the previous proposition is only of order \sqrt{h} . The numerical experiments of section 7 show that this result, based on the classical formulation, is optimal, in general. The aim of the next section is to propose a stabilization technique to overcome this limitation.

5. A stabilized formulation. In this section we adapt a stabilization technique presented by Barbosa and Hughes in [2, 3] in order to recover an optimal rate of convergence. Note that the link between this stabilization technique and the former Nitsche’s method [20] has been established in [24]. Moreover, it has been recently used to interface problems with nonmatching meshes in [4] and to elastostatic contact problems in [14]. We present its symmetric version although the nonsymmetric one can be considered in the same way. This technique is based on the addition of a supplementary term involving the normal derivative on Γ_D . In fact, we need a little bit more general definition. Let us suppose that we have at our disposal an operator

$$R^h : V^h \longrightarrow L^2(\Gamma_D),$$

which approximates the normal derivative on Γ_D , (i.e., for $v^h \in V^h$ converging to a sufficiently smooth function v , $R^h(v^h)$ tends to $\partial_n v$ in an appropriate sense). Several

choices of R^h will be proposed later. We suppose that the following estimate holds for this operator:

$$(27) \quad h^{1/2} \|R^h(v^h)\|_{0,\Gamma_D} \leq C \|\nabla v^h\|_{0,\Omega} \quad \forall v^h \in V^h, \forall h > 0.$$

To obtain the stabilized problem we replace the Lagrangian (7) by the following one:

$$\mathcal{L}_h(v^h, \mu^h) = \mathcal{L}(v^h, \mu^h) - \frac{\gamma}{2} \int_{\Gamma_D} (\mu^h + R^h(v^h))^2 d\Gamma, \quad v^h \in V^h, \mu^h \in W^h,$$

where for the sake of simplicity $\gamma := h\gamma_0$ is chosen to be a positive constant over Ω (for nonuniform meshes, an element dependent parameter $\gamma = h_T\gamma_0$ is a better choice). The corresponding discrete problem reads as follows:

$$(28) \quad \begin{cases} \text{Find } u^h \in V^h \text{ and } \lambda^h \in W^h \text{ such that} \\ a(u^h, v^h) + \int_{\Gamma_D} \lambda^h v^h d\Gamma - \gamma \int_{\Gamma_D} (\lambda^h + R^h(u^h)) R^h(v^h) d\Gamma = l(v^h) \quad \forall v^h \in V^h, \\ \int_{\Gamma_D} \mu^h u^h d\Gamma - \gamma \int_{\Gamma_D} (\lambda^h + R^h(u^h)) \mu^h d\Gamma = 0 \quad \forall \mu^h \in W^h. \end{cases}$$

As in [2], let us define the form $\mathcal{B}_h : (V^h \times W^h)^2 \rightarrow \mathbb{R}$ by

$$\begin{aligned} \mathcal{B}_h(u^h, \lambda^h; v^h, \mu^h) := & a(u^h, v^h) + \int_{\Gamma_D} \lambda^h v^h d\Gamma + \int_{\Gamma_D} \mu^h u^h d\Gamma \\ & - \gamma \int_{\Gamma_D} (\lambda^h + R^h(u^h)) (\mu^h + R^h(v^h)) d\Gamma. \end{aligned}$$

Then, (28) is equivalent to

$$(29) \quad \begin{cases} \text{Find } u^h \in V^h \text{ and } \lambda^h \in W^h \text{ such that} \\ \mathcal{B}_h(u^h, \lambda^h; v^h, \mu^h) = l(v^h), \quad \forall (v^h, \mu^h) \in V^h \times W^h. \end{cases}$$

Moreover, this formulation is consistent in the sense that the solution (u, λ) to problem (5) satisfies

$$(30) \quad \bar{\mathcal{B}}_h(u, \lambda; v^h, \mu^h) = l(v^h), \quad \forall v^h \in V^h, \forall \mu^h \in W^h,$$

provided that $\lambda \in L^2(\Gamma_D)$ with $\bar{\mathcal{B}}_h$ having the same definition as \mathcal{B}_h but replacing $R^h(u)$ by $\partial_n u$.

The following hypothesis on the approximation property of W^h will be needed to get an abstract result. Let $P^h : L^2(\Gamma_D) \rightarrow W^h$ be the L^2 -projection on W^h . We suppose that there exists a constant $C > 0$ independent of h such that

$$(31) \quad \|P^h v - v\|_{0,\Gamma_D} \leq Ch^{1/2} \|v\|_{1/2,\Gamma_D}, \quad \forall v \in H^{1/2}(\Gamma_D).$$

This allows one to establish the following “inf-sup” property of \mathcal{B}_h .

LEMMA 3. *Let hypotheses (11), (27), and (31) be satisfied. Then for $\gamma_0 > 0$ sufficiently small there exists a constant $C > 0$ independent of h such that*

$$(32) \quad \sup_{(0,0) \neq (z^h, \eta^h) \in V^h \times W^h} \frac{\mathcal{B}_h(v^h, \mu^h; z^h, \eta^h)}{\| (z^h, \eta^h) \|} \geq C \| (v^h, \mu^h) \|,$$

where $\| (z^h, \eta^h) \|^2 := \|z^h\|_V^2 + h^{-1} \|z^h\|_{0,\Gamma_D}^2 + h \|\eta^h\|_{0,\Gamma_D}^2$.

Proof. The proof is an adaptation of the one in [24], Lemma 5. First of all, for $(v^h, \mu^h) \in V^h \times W^h$ arbitrary, $\gamma_0 > 0$ sufficiently small, and from (27) one has

$$\begin{aligned} \mathcal{B}_h(v^h, \mu^h; v^h, -\mu^h) &= \|\nabla v^h\|_{0,\Omega}^2 + \gamma_0 h \|\mu^h\|_{0,\Gamma_D}^2 - \gamma_0 h \|R^h(v^h)\|_{0,\Gamma_D}^2 \\ (33) \qquad \qquad \qquad &\geq C \left(\|\nabla v^h\|_{0,\Omega}^2 + h \|\mu^h\|_{0,\Gamma_D}^2 \right). \end{aligned}$$

Next, from (27) and the Young inequality we get for $\bar{\mu}^h := h^{-1} P^h v^h$:

$$\begin{aligned} \mathcal{B}_h(v^h, \mu^h; 0, \bar{\mu}^h) &= \int_{\Gamma_D} \bar{\mu}^h v^h d\Gamma - \gamma \int_{\Gamma_D} (\mu^h + R^h(v^h)) \bar{\mu}^h d\Gamma \\ &\geq h^{-1} \|P^h v^h\|_{0,\Gamma_D}^2 \\ &\quad - C \left(\|\nabla v^h\|_{0,\Omega} + h^{1/2} \|\mu^h\|_{0,\Gamma_D} \right) h^{-1/2} \|P^h v^h\|_{0,\Gamma_D} \\ &\geq h^{-1} \|P^h v^h\|_{0,\Gamma_D}^2 \\ &\quad - \frac{C^2}{2} \left(\|\nabla v^h\|_{0,\Omega} + h^{1/2} \|\mu^h\|_{0,\Gamma_D} \right)^2 - \frac{h^{-1}}{2} \|P^h v^h\|_{0,\Gamma_D}^2 \\ (34) \qquad \qquad \qquad &\geq \frac{h^{-1}}{2} \|P^h v^h\|_{0,\Gamma_D}^2 - \bar{C} \left(\|\nabla v^h\|_{0,\Omega}^2 + h \|\mu^h\|_{0,\Gamma_D}^2 \right). \end{aligned}$$

We now take $(z^h, \eta^h) = (v^h, -\mu^h + \delta \bar{\mu}^h)$ in (32) with $\delta > 0$. Using (33), (34), and δ sufficiently small one has

$$\begin{aligned} \mathcal{B}_h(v^h, \mu^h; z^h, \eta^h) &= \mathcal{B}_h(v^h, \mu^h, v^h, -\mu^h) + \delta \mathcal{B}_h(v^h, \mu^h, 0, \bar{\mu}^h) \\ (35) \qquad \qquad \qquad &\geq C \left(\|\nabla v^h\|_{0,\Omega}^2 + h^{-1} \|P^h v^h\|_{0,\Gamma_D}^2 + h \|\mu^h\|_{0,\Gamma_D}^2 \right). \end{aligned}$$

Since $\{1\} \subset W^h$, then for the L^2 -projection of v^h on $\{1\}$ we obtain

$$(36) \qquad \|P^h v^h\|_{0,\Gamma_D}^2 \geq \int_{\Gamma_D} \left(\frac{1}{|\Gamma_D|} \int_{\Gamma_D} v^h d\Gamma \right)^2 d\Gamma = \frac{1}{|\Gamma_D|} \left(\int_{\Gamma_D} v^h d\Gamma \right)^2.$$

Let $\beta > 0$ be sufficiently small. Then it holds:

$$\begin{aligned} (37) \quad \|\nabla v^h\|_{0,\Omega}^2 + h^{-1} \|P^h v^h\|_{0,\Gamma_D}^2 &= \|\nabla v^h\|_{0,\Omega}^2 + (1 - \beta) h^{-1} \|P^h v^h\|_{0,\Gamma_D}^2 \\ &\quad + \beta h^{-1} \|P^h v^h - v^h + v^h\|_{0,\Gamma_D}^2 \\ (38) \qquad \qquad \qquad &\geq \|\nabla v^h\|_{0,\Omega}^2 + (1 - \beta) \frac{1}{|\Gamma_D| \text{diam}(\Omega)} \left(\int_{\Gamma_D} v^h d\Gamma \right)^2 \\ &\quad + \beta h^{-1} \left(\|v^h\|_{0,\Gamma_D}^2 - \|P^h v^h - v^h\|_{0,\Gamma_D}^2 \right) \\ &\geq C \left(\|v^h\|_V^2 + \beta h^{-1} \left(\|v^h\|_{0,\Gamma_D}^2 - h \|v^h\|_{1/2,\Gamma_D}^2 \right) \right) \\ (39) \qquad \qquad \qquad &\geq C \left(\|v^h\|_V^2 + h^{-1} \|v^h\|_{0,\Gamma_D}^2 \right), \end{aligned}$$

where we used (31), the fact that $(\|\nabla v^h\|_{0,\Omega}^2 + (1 - \beta) \frac{1}{|\Gamma_D| \text{diam}(\Omega)} (\int_{\Gamma_D} v^h d\Gamma)^2)^{1/2}$ is an equivalent norm on V and the trace theorem. Finally, one obtains (32) combining (35) and (39) together with the fact that $\|(z^h, \eta^h)\| \leq C \|(v^h, \mu^h)\|$. \square

Remark 2. The inf-sup condition straightforwardly ensures the existence and uniqueness of a solution to the discrete problem (28) for $\gamma_0 > 0$ sufficiently small.

Now, we can prove the following abstract error estimate.

THEOREM 1. *Let (11), (27), and (31) be satisfied and $\gamma_0 > 0$ be sufficiently small. If (u, λ) is the solution to Problem (5) such that $\lambda \in L^2(\Gamma_D)$, then there exists a constant $C > 0$ independent of h and (u, λ) such that the following estimate holds:*

$$\begin{aligned} & \left\| \left\| (u - u^h, \lambda - \lambda^h) \right\| \right\| \\ & \leq C \inf_{v^h \in V^h, \mu^h \in W^h} \left(\left\| \left\| (u - v^h, \lambda - \mu^h) \right\| \right\| + h^{1/2} \|R^h(v^h) - \partial_n u\|_{0, \Gamma_D} \right). \end{aligned}$$

Proof. From (30) it follows that

$$\overline{\mathcal{B}}_h(u, \lambda, z^h, \eta^h) = \mathcal{B}_h(u^h, \lambda^h, z^h, \eta^h) \quad \forall (z^h, \eta^h) \in V^h \times W^h.$$

Thus, for any $(v^h, \mu^h) \in V^h \times W^h$ one has

$$\begin{aligned} & \mathcal{B}_h(v^h, \mu^h, z^h, \eta^h) - \overline{\mathcal{B}}_h(u, \lambda, z^h, \eta^h) \\ & = \mathcal{B}_h(v^h - u^h, \mu^h - \lambda^h, z^h, \eta^h) \quad \forall (z^h, \eta^h) \in V^h \times W^h. \end{aligned}$$

A direct computation leads to

$$\begin{aligned} \mathcal{B}_h(v^h, \mu^h; z^h, \eta^h) - \overline{\mathcal{B}}_h(u, \lambda; z^h, \eta^h) & \leq C \left(\left\| \left\| (u - v^h, \lambda - \mu^h) \right\| \right\| \right. \\ & \quad \left. + h^{1/2} \|R^h(v^h) - \partial_n u\|_{0, \Gamma_D} \right) \left\| \left\| (z^h, \eta^h) \right\| \right\|. \end{aligned}$$

Further,

$$\begin{aligned} \left\| \left\| (u - u^h, \lambda - \lambda^h) \right\| \right\| & \leq \left\| \left\| (u - v^h, \lambda - \mu^h) \right\| \right\| + \left\| \left\| (v^h - u^h, \mu^h - \lambda^h) \right\| \right\| \\ & \leq \left\| \left\| (u - v^h, \lambda - \mu^h) \right\| \right\| \\ & \quad + C \sup_{(0,0) \neq (z^h, \eta^h) \in V^h \times W^h} \frac{\mathcal{B}_h(v^h - u^h, \mu^h - \lambda^h; z^h, \eta^h)}{\left\| \left\| (z^h, \eta^h) \right\| \right\|} \\ & \leq C \left(\left\| \left\| (u - v^h, \lambda - \mu^h) \right\| \right\| + h^{1/2} \|R^h(v^h) - \partial_n u\|_{0, \Gamma_D} \right) \end{aligned}$$

holds for any $(v^h, \mu^h) \in V^h \times W^h$. \square

In the rest of this section we show how to use the abstract result of Theorem 1 to establish an optimal a priori error estimate for the following standard finite element spaces:

$$(40) \quad \widetilde{V}^h = \left\{ v^h \in \mathcal{C}(\widetilde{\Omega}) : v^h|_T \in P_{k_u}(T) \quad \forall T \in \mathcal{T}^h \right\}, \quad k_u \geq 1,$$

$$(41) \quad \widetilde{W}^h = \left\{ \mu^h \in L^2(\widetilde{\Omega}) : \mu^h|_T \in P_{k_\lambda}(T) \quad \forall T \in \mathcal{T}^h \right\}, \quad k_\lambda \geq 0.$$

In order to estimate the boundary terms, we shall need the following classical estimate which is satisfied for any $T \in \mathcal{T}^h$ and any $w \in H^1(T)$ provided that Γ_D is smooth enough (see Appendix A for the proof):

$$(42) \quad \|w\|_{0, \Gamma_D \cap T}^2 \leq C (h_T^{-1} \|w\|_{0, T}^2 + h_T \|w\|_{1, T}^2).$$

Let $k = \min(k_u, k_\lambda + 1)$ and consider two continuous extension operators:

$$\begin{aligned} T_u^k &: H^{k+1}(\Omega) \longrightarrow H^{k+1}(\tilde{\Omega}), \\ T_\lambda^k &: H^{k-1/2}(\Gamma_D) \longrightarrow H^k(\tilde{\Omega}), \end{aligned}$$

where $H^{k-1/2}(\Gamma_D)$ stands for the space of traces on Γ_D of functions from $H^k(\Omega)$. Due to Calderón’s extension theorem, it is always possible to build such operators provided that the domain Ω has the uniform cone property (see [1], for instance). This allows us to define the following interpolation operators on \tilde{V}^h and \tilde{W}^h :

$$\begin{aligned} \tilde{\Pi}_u^{k,h}(v) &:= \Pi^{k,h}(T_u^k(v)) \quad \forall v \in H^{k+1}(\Omega), \\ \tilde{\Pi}_\lambda^{k,h}(\mu) &:= \Pi^{k-1,h}(T_\lambda^k(\mu)) \quad \forall \mu \in H^{k-1/2}(\Gamma_D), \end{aligned}$$

where $\Pi^{k,h}$ stands for the standard Lagrange interpolation operator by piecewise polynomial functions of degree less or equal k defined on the mesh \mathcal{T}^h . An exception has to be done for $k = 1$ when the Lagrange interpolation operator will be replaced by Clément’s one for the interpolation of the multiplier since functions from $H^1(\tilde{\Omega})$ are not generally continuous (see [8]). Due to the known approximation properties of these operators on regular families of meshes (see [7] and [8]), one has for any $v \in H^{k+1}(\Omega)$:

$$\begin{aligned} \|\tilde{\Pi}_u^{k,h}(v) - v\|_V &\leq \|\tilde{\Pi}_u^{k,h}(v) - T_u^k(v)\|_{1,\tilde{\Omega}} \\ &\leq Ch^k \|T_u^k(v)\|_{k+1,\tilde{\Omega}} \leq Ch^k \|v\|_{k+1,\Omega}, \end{aligned}$$

and for any $\mu \in H^{k-1/2}(\Gamma_D)$ taking into account (42):

$$\begin{aligned} \|\tilde{\Pi}_\lambda^{k,h}(\mu) - \mu\|_{0,\Gamma_D}^2 &\leq C \sum_{T \in \mathcal{T}^h} \left(h^{-1} \|\tilde{\Pi}_\lambda^{k,h}(\mu) - T_\lambda^k(\mu)\|_{0,T}^2 + h \|\tilde{\Pi}_\lambda^{k,h}(\mu) - T_\lambda^k(\mu)\|_{1,T}^2 \right) \\ &\leq Ch^{2k-1} \|T_\lambda^k(\mu)\|_{k,\tilde{\Omega}}^2 \leq Ch^{2k-1} \|\mu\|_{k-1/2,\Gamma_D}^2. \end{aligned}$$

In the same way one can derive the estimate $\|\tilde{\Pi}_u^{k,h}(v) - v\|_{0,\Gamma_D}$ for $v \in H^{k+1}(\Omega)$ and also obtain the estimate (31) (using Clément’s interpolation operator). Thus, an a priori error estimate can be derived provided that the following approximation property of R^h holds:

$$(43) \quad \left\| R^h \left(\tilde{\Pi}_u^{k,h}(v) \right) - \partial_n v \right\|_{0,\Gamma_D} \leq Ch^{k-1/2} \|v\|_{k+1,\Omega}.$$

THEOREM 2. *Let \tilde{V}^h and \tilde{W}^h be defined by (40) and (41), respectively. Let (u, λ) be the solution to Problem (5) such that $u \in H^{k+1}(\Omega)$ and $\lambda \in H^{k-1/2}(\Gamma_D)$ for $k = \min\{k_u, k_\lambda + 1\}$. Assume that (27) and (43) are satisfied. Then the following estimate holds:*

$$\| |(u - u^h, \lambda - \lambda^h)| \| \leq Ch^k \|u\|_{k+1,\Omega},$$

where (u^h, λ^h) is the solution to Problem (28).

Remark 3. Note that for $k_\lambda \geq 1$ the use of $\tilde{W}^h \cap \mathcal{C}(\tilde{\Omega})$ instead of \tilde{W}^h does not change the result. Note also that the definition of the norm $\| |(u - u^h, \lambda - \lambda^h)| \|$ involves

a standard error estimate for $\|u - u^h\|_V$. However, it does not provide an estimate of $\|\lambda - \lambda^h\|_{-1/2, \Gamma_D}$ but the one of $h^{1/2}\|\lambda - \lambda^h\|_{0, \Gamma_D}$. An additional optimal estimate of $\|u - u^h\|_{0, \Gamma_D}$ is also available without supplementary regularity assumptions. This is due to the use of the Pitkäranta technique [21]. Error estimates with natural norms instead of mesh dependent norms are also possible for the stabilized problem (see [3]).

5.1. Case $R^h(v^h) = \partial_n v^h$ and an additional condition on the mesh. A natural choice for the operator R^h is of course

$$R^h(v^h) = \partial_n v^h \text{ on } \Gamma_D,$$

which corresponds to the original method of Barbosa and Hughes. In this case, unfortunately, the stability condition (27) is verified only under an additional regularity assumption on the intersection of the mesh with Ω . We denote by \hat{T} a reference element such that $T = \tau_T(\hat{T})$ for all $T \in \mathcal{T}^h$, where τ_T is a regular affine transformation in \mathbb{R}^d . The assumption on the mesh can be expressed as follows (see [21] for a similar one):

(44) There exists a radius $\hat{\rho} > 0$ independent of h such that for any $T \in \mathcal{T}^h, T \cap \Omega \neq \emptyset$ the reference element \hat{T} contains a ball $B(\hat{y}_T, \hat{\rho})$ which satisfies $B(\hat{y}_T, \hat{\rho}) \subset \tau_T^{-1}(T \cap \Omega)$.

Under this assumption, inequality (27) is satisfied for \tilde{V}^h defined by (40) (see the proof in Appendix B). Moreover, the following lemma says that (43) is also satisfied.

LEMMA 4. Let \tilde{V}^h be defined by (40), $R^h(v^h) = \partial_n v^h$ on Γ_D and assume that (44) is satisfied. Then (43) is satisfied as well.

Proof. Recall that $k = \min(k_u, k_\lambda + 1)$. Using (42) and standard interpolation error estimates one has for any $v \in H^{k+1}(\Omega)$:

$$\begin{aligned} \|R^h(\tilde{\Pi}_u^{k,h}(v)) - \partial_n v\|_{0, \Gamma_D}^2 &\leq \sum_{T \in \mathcal{T}^h} \|\nabla \tilde{\Pi}_u^{k,h}(v) - \nabla v\|_{0, \Gamma_D \cap T}^2 \\ &\leq C \sum_{T \in \mathcal{T}^h} \left(h^{-1} \|\nabla \tilde{\Pi}_u^{k,h}(v) - \nabla T_u^k(v)\|_{0, T}^2 \right. \\ &\quad \left. + h \|\nabla \tilde{\Pi}_u^{k,h}(v) - \nabla T_u^k(v)\|_{1, T}^2 \right) \\ &\leq C \sum_{T \in \mathcal{T}^h} \left(h^{-1} \left(h^k \|T_u^k(v)\|_{k+1, T} \right)^2 \right. \\ &\quad \left. + h \left(h^{k-1} \|T_u^k(v)\|_{k+1, T} \right)^2 \right) \\ &\leq Ch^{2k-1} \|v\|_{k+1, \Omega}^2. \quad \square \end{aligned}$$

We can deduce that if $R^h(v^h) = \partial_n v^h$ on Γ_D , the estimate of Theorem 2 holds provided that (44) is satisfied. This assumption, however, restricts the use of our fictitious domain approach. Indeed if, for instance, one wants to approximate an evolving boundary, the intersection of elements with the real domain will be arbitrary. The aim of the next section is to introduce an operator R^h with a reinforced stability, enabling us to work with an arbitrary domain.

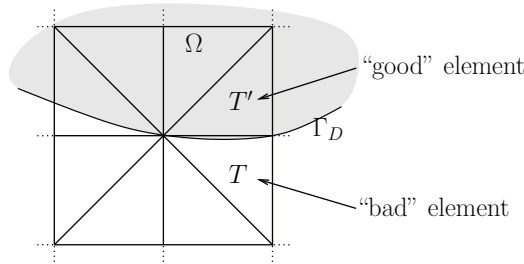


FIG. 3. The choice of T' for an element T having a small intersection with Ω . In this case, it is more stable to evaluate the normal derivative from a natural extension of v^h from T' on T because smaller is the thickness of this intersection; poorer approximation of the normal derivative on $T \cap \partial\Omega$ is obtained using $v^h|_T$.

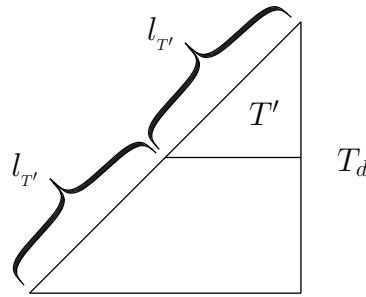


FIG. 4. Prolongation of T' .

5.2. Operator R^h with a reinforced stability. We give here an example of how to construct an operator R^h ensuring both the approximation property (43) as well as the stability property (27) for an arbitrary intersection of the mesh \mathcal{T}^h with the domain Ω . The proposed construction is only local and quite simple to implement.

Let $\hat{\rho} > 0$ be an a priori given small radius ($\hat{\rho} \ll 1$). For each element $T \in \mathcal{T}^h$ such that $T \cap \Omega \neq \emptyset$, we will designate by T' either the element T itself if there is a ball $B(\hat{y}_T, \hat{\rho}) \subset \tau_T^{-1}(T \cap \Omega)$ (a “good” element) or any neighbor element possessing this property if T itself does not satisfy it (T is a “bad” element).

The proposed operator R^h will simply be equal to $\partial_n \bar{v}_{T',T}^h$ where $\bar{v}_{T',T}^h$ is either $v^h|_T$ if $T' = T$ or the natural extension of $v^h|_{T'}$ onto T if $T' \neq T$. Of course, $\hat{\rho} > 0$ has to be sufficiently small such that T' always exists, which is not a big constraint.

It is not difficult to see that the stability condition (27) is satisfied with such a choice of the operator R^h (see Appendix C for the sketch of the proof). The following lemma establishes that (43) is also satisfied so that the estimate of Theorem 2 holds, again.

LEMMA 5. Let \tilde{V}^h be defined by (40), and $R^h(v^h) := \partial_n \bar{v}_{T',T}^h$ on Γ_D . Then (43) is satisfied.

Proof. Suppose that T is a “bad” element, i.e., $T \cap \Omega$ is “thin” and let T' be a “good” neighbor element as described above (see also Figure 3). We prolong T' and construct the new element T_d as shown in Figure 4. The interpolation on T_d is defined by the interpolation on T' . More precisely:

$$\text{let } v \in H_{\text{loc}}^{k+1}(\mathbb{R}^d) \text{ and } v_{T'} := v|_{T'}.$$

By $\Pi_{T'} v_{T'}$ we denote the P_k -Lagrange interpolant of $v_{T'}$ constructed on T' (i.e., using degrees of freedom in T') but with the domain of definition being the whole \mathbb{R}^d . The interpolation of v on T_d is defined as

$$\Pi_{T_d} v := \Pi_{T'} v_{T'}|_{T_d}.$$

Classical arguments based on the fact that $v - \Pi_{T_d} v$ vanishes for all polynomials of degree less or equal k lead to the following approximation property (see [6], for instance):

$$\|v - \Pi_{T_d} v\|_{m, T_d} \leq C h_{T_d}^{k+1-m} \|v\|_{k+1, T_d} \quad (h_{T_d} \leq 2h_{T'}).$$

Analogically to Lemma 6 (see Appendix A) it holds:

$$\|v\|_{0, \Gamma_D \cap T_d}^2 \leq C \left(h_{T_d}^{-1} \|v\|_{0, T_d}^2 + h_{T_d} \|v\|_{1, T_d}^2 \right).$$

To get (43) we proceed as in Lemma 4. Only we have to sort all elements into “good” and “bad” ones and to use either Π_T or Π_{T_d} . \square

6. Some practical details for implementation. The implementation of the proposed method requires one to overcome a certain number of difficulties. First of all, one has to select bases of the spaces V^h and W^h from the ones of \widetilde{V}^h and \widetilde{W}^h . As far as V^h is concerned, the task is rather easy because it suffices to select the basis functions among the ones of \widetilde{V}^h which are not identically equal to zero in Ω (one can eventually remove those for which the intersection of their support with Ω is too small). It is a little more difficult to find a basis of the space W^h . Indeed, the traces on Γ_D of basis functions of \widetilde{W}^h may be linearly dependent, especially if Γ_D is rectilinear. A possible way to overcome this difficulty is to eliminate the redundant functions by analyzing the elementary mass matrices whose components are $\int_{\Gamma_D \cap T} \psi_i \psi_j d\Gamma$, where $\{\psi_i\}$ are the shape functions of \widetilde{W}^h .

Another difficulty concerns the numerical integration: one needs to build integration formulas on the intersection of elements with the domain Ω as well as on the intersection of elements with Γ_D . Our finite element library, Getfem++ [22], uses splitting of elements into simplices in a conformal way with respect to $\partial\Omega$ and then it applies a standard integration formula on each subelement. If $\partial\Omega$ is curved, then some curved subelements can be used. One obtains an integration formula on Γ_D by considering the faces of the subelements lying on Γ_D .

The natural extension of functions on “bad” elements which is needed to obtain the fully stabilized method described in section 5.2 consists in seeking information in a “good” nearby element. This can be a handicap for certain finite element codes where calculations are done only elementwise. A possible remedy is to precompute a global discrete extension operator which gives the solution extended to “bad” elements from the original one. Then, the matrices involving $R^h(v_h)$ are obtained as a composition of classical matrices with this discrete extension operator.

The Xfem method is often associated with the use of some level-sets of functions defined on the mesh. This is particularly useful when, for instance, one needs to represent an evolving interface. In our case such a level-set can be utilized to represent the boundary of Ω . The implementation in Getfem++ uses this strategy. Generally, this involves an additional approximation of Ω . In our numerical tests presented in the next section, the level-set functions are piecewise second degree polynomials. In this case the level-set approach has no influence on the rate of convergence of the used finite element methods which are of the first and second order.

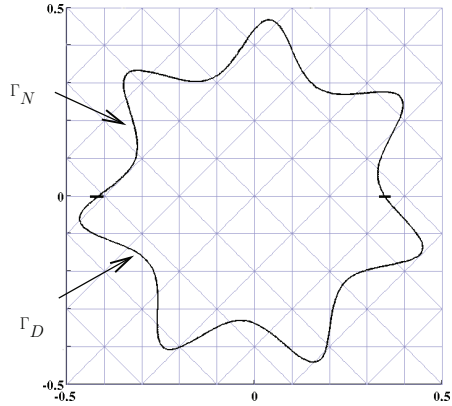


FIG. 5. Test domain and a triangular structured mesh.

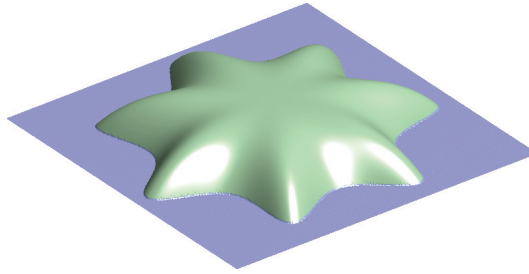


FIG. 6. Exact solution.

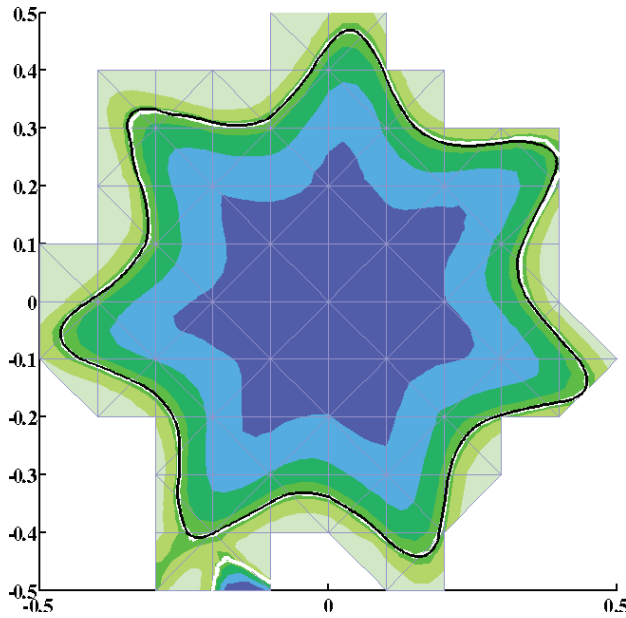


FIG. 7. Approximated solution on a rough mesh with the P_2/P_1 method. Only the elements intersecting Ω are depicted. The black curve is the boundary of Ω and the white curve is the zero level-set of the approximated solution.

7. Numerical experiments. In this section, we present 2D-numerical tests. The fictitious domain is $\tilde{\Omega} =]-1/2, 1/2[^2$. The exact solution is $u(x) = R^4 - |x|^4(5 + 3 \sin(7\theta + \frac{7\pi}{36}))/2$, where $R = 0.47$ and $\theta(x) = \arctan(x_2/x_1)$. The real domain is $\Omega = \{x \in \mathbb{R}^2 : u(x) < 0\}$, and the Dirichlet and Neumann boundary conditions are defined on $\Gamma_D = \Gamma \cap \{x \in \mathbb{R}^2 : x_2 < 0\}$ and $\Gamma_N = \Gamma \cap \{x \in \mathbb{R}^2 : x_2 > 0\}$.

The domain Ω is represented in Figure 5 with an example of a triangular structured mesh. The exact solution is shown in Figure 6 while a computed solution on a rough mesh is depicted in Figure 7.

7.1. Without stabilization. First, we present numerical tests without any stabilization. We tested several choices of the finite element spaces \widetilde{V}_h and \widetilde{W}_h .

In order to avoid the locking phenomena, the couple of selected finite element spaces should satisfy as much as possible a discrete mesh independent inf-sup condition since the stabilization is not used. For instance, it is known that the P_1/P_0 method for the discretization of u, λ , respectively, does not satisfy such a condition. The linear system to be solved is of the form

$$(45) \quad \begin{pmatrix} K & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} U \\ \Lambda \end{pmatrix} = \begin{pmatrix} L \\ 0 \end{pmatrix},$$

where U and Λ are the degrees of freedom of u^h and λ^h , respectively, and the components of K, B , and L are

$$K_{ij} = \int_{\Omega} \nabla \varphi_i \cdot \nabla \varphi_j d\Omega, \quad B_{ij} = \int_{\Gamma_D} \psi_i \varphi_j d\Gamma, \quad L_i = \int_{\Omega} f \varphi_i d\Omega + \int_{\Gamma_N} g \varphi_i d\Gamma,$$

with $\{\varphi_i\}, \{\psi_j\}$ being the selected basis functions of $\widetilde{V}_h, \widetilde{W}_h$, respectively. In our experiments, this system is solved using the library Superlu [9] (a direct LU solver for sparse matrices).

The test program can be downloaded on the Getfem++ web site [22]. It allows one to test many other couples of elements and to treat also 3D problems.

The couples of spaces tested are the following: $P_1/P_0, P_1+/P_0$ (a standard continuous P_1 element for u enriched by a cubic bubble function and a standard P_0 element for the multiplier), Q_1/Q_0 (standard continuous Q_1 and discontinuous Q_0 elements on quadrilaterals), $P_2/P_1, P_2/P_0$, and Q_2/Q_1 .

Rates of convergence are presented in Figure 8. One can see that in all experiments the rate of convergence in the $H^1(\Omega)$ -norm is better than the theoretical one given by Proposition 3 except for the P_1/P_0 case which is a little bit slower than $h^{1/2}$. The choice P_1/P_0 suffers of course from the non-satisfaction of a mesh-independent inf-sup condition. It has to be stressed that in all the experiments without stabilization, and particularly for the P_1/P_0 case, a singular linear system can be obtained. However, in all examples, presented here, we selected some cases with a non-singular linear system. It is also seen that convergence of the multiplier is not generally obtained, especially for degree one methods. Figure 9 illustrates a poor quality of the multiplier for the P_1/P_0 method. The P_2/P_1 method gives slightly better results (see Figure 10 still with some oscillations in parts where the intersection of the element with the domain Ω is very small).

7.2. The stabilized method with $R^h(v^h) = \partial_n v^h$. The numerical experiments are now done using the standard Barbosa–Hughes stabilization technique (with $\gamma = 0.1$). It has been proven in section 5.1 that this method is optimal whenever the intersection of elements with the domain Ω is not too small. This is not easy to satisfy in computations. Of course, one way to avoid small intersections would be to move a little bit some mesh nodes, at least when a structured mesh is not required. We did not test this possibility.

Unlike (45), the linear system to be solved is now of the form

$$\begin{pmatrix} K_\gamma & B_\gamma^T \\ B_\gamma & -M_\gamma \end{pmatrix} \begin{pmatrix} U \\ \Lambda \end{pmatrix} = \begin{pmatrix} L \\ 0 \end{pmatrix},$$

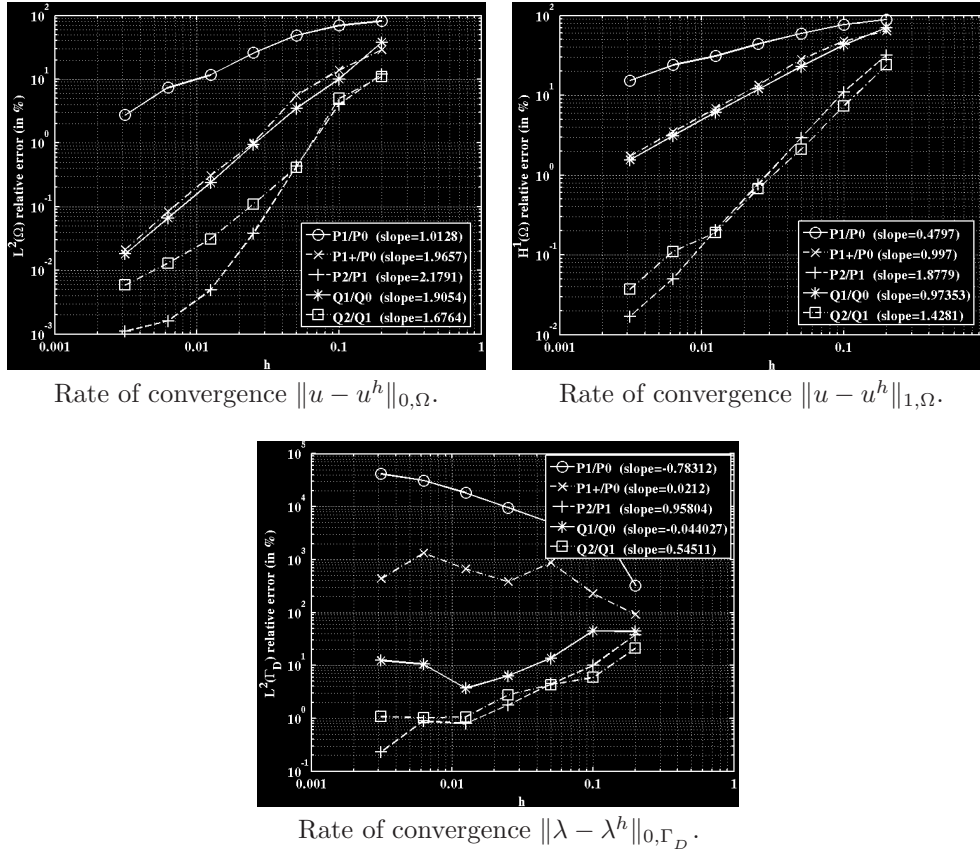


FIG. 8. Rates of convergence for some couples of finite element spaces with no stabilization.

where the components of K_γ , B_γ , and M_γ are

$$\begin{aligned}
 (K_\gamma)_{ij} &= \int_{\Omega} \nabla \varphi_i \cdot \nabla \varphi_j d\Omega - \gamma \int_{\Gamma_D} R^h(\varphi_i) R^h(\varphi_j) d\Gamma, \\
 (B_\gamma)_{ij} &= \int_{\Gamma_D} \psi_i (\varphi_j - \gamma R^h(\varphi_j)) d\Gamma, \\
 (M_\gamma)_{ij} &= \gamma \int_{\Gamma_D} \psi_i \psi_j d\Gamma,
 \end{aligned}$$

respectively. Note that K_γ is invertible provided that γ is sufficiently small. The whole matrix of the system is invertible as well whatever is B_γ .

Rates of convergence are presented in Figure 11 for the same couples of elements as in the previous section. The stabilization significantly improves the convergence of the P_1/P_0 choice (the stabilization with bubble functions is no longer necessary) and the convergence of quadratic elements. Moreover, the linear system is guaranteed to be invertible. Figure 12 shows that also the approximation of the multiplier is considerably improved. The convergence rate is improved by the stabilization, but some problems remain with too small intersections of elements with Ω even for degree two methods (see Figure 13).

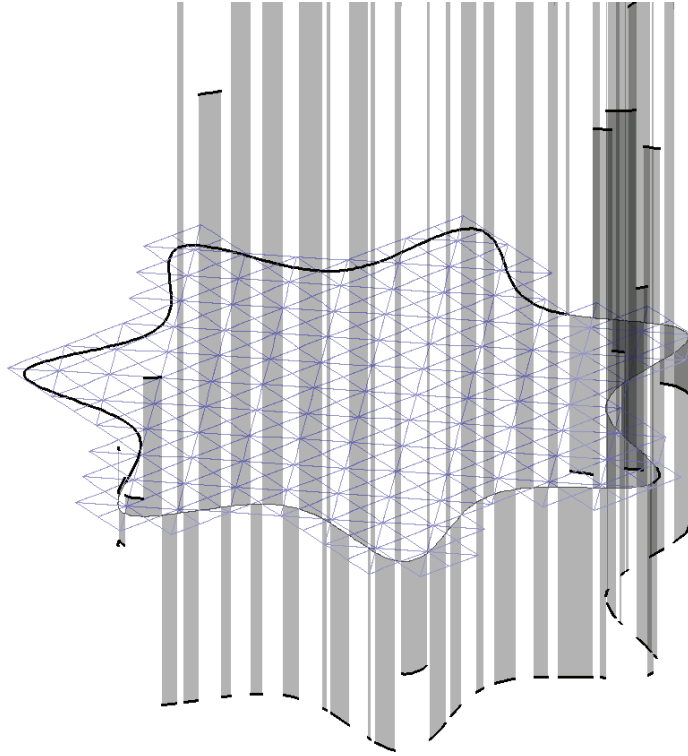


FIG. 9. Multiplier on Γ_D with no stabilization for the P_1/P_0 method ($h = 0.05$).

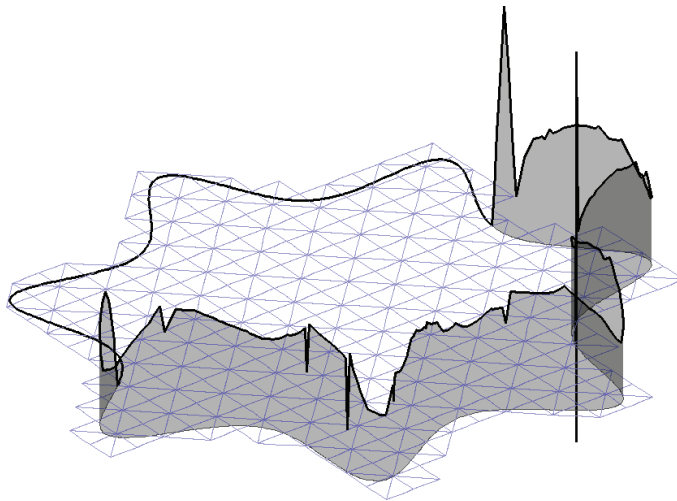
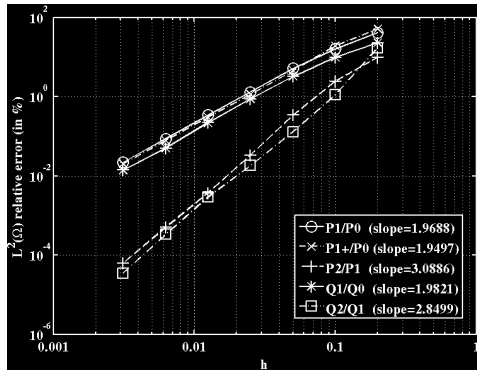
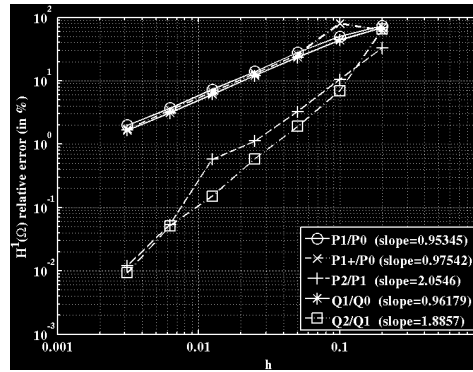


FIG. 10. Multiplier on Γ_D with no stabilization for the P_2/P_1 method ($h = 0.05$).

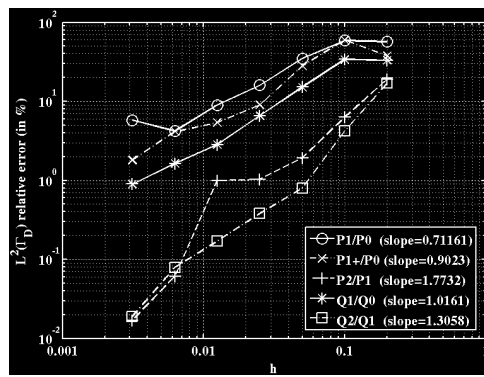
7.3. The fully stabilized method. We now consider the fully stabilized method described in section 5.2. An element T is considered to be “bad” when $|T \cap \Omega|$ is less than one percent of $|T|$. The convergence curves given in Figure 14 are rather the same than with the standard Barbosa–Hughes stabilization used in the previous sec-



Rate of convergence $\|u - u^h\|_{0,\Omega}$.



Rate of convergence $\|u - u^h\|_{1,\Omega}$.



Rate of convergence $\|\lambda - \lambda^h\|_{0,\Gamma_D}$.

FIG. 11. Rates of convergence for some couples of finite element spaces with the Barbosa–Hughes stabilization.

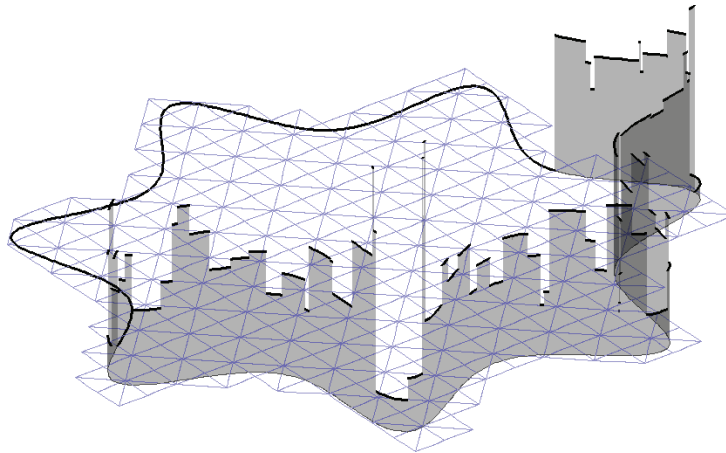


FIG. 12. Multiplier on Γ_D with the Barbosa–Hughes stabilization for the P_1/P_0 method ($h = 0.05$).

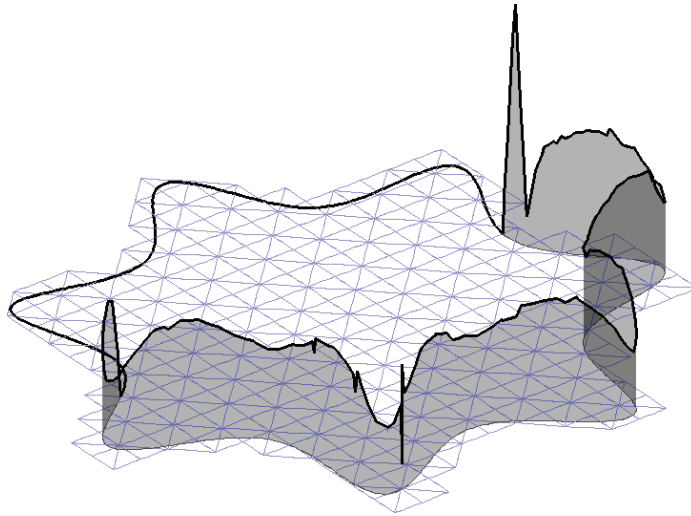
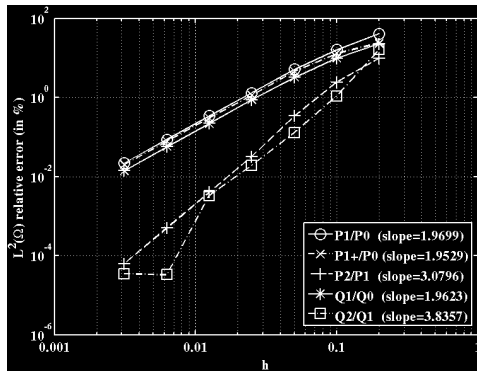
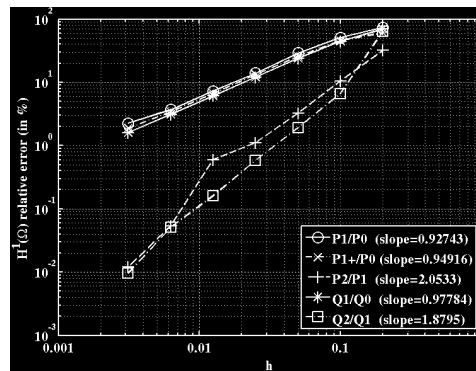


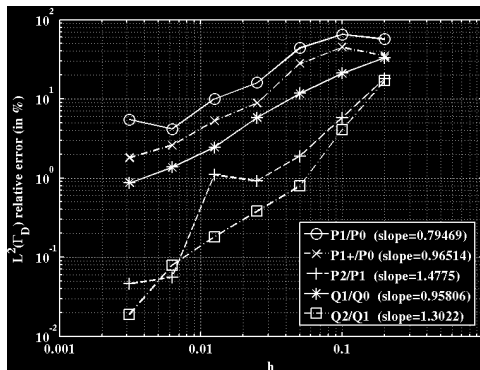
FIG. 13. Multiplier on Γ_D with the Barbosa-Hughes stabilization for the P_2/P_1 method ($h = 0.05$).



Rate of convergence $\|u - u^h\|_{0,\Omega}$.



Rate of convergence $\|u - u^h\|_{1,\Omega}$.



Rate of convergence $\|\lambda - \lambda^h\|_{0,\Gamma_D}$.

FIG. 14. Rates of convergence for some couples of finite element spaces with the fully stabilized method.

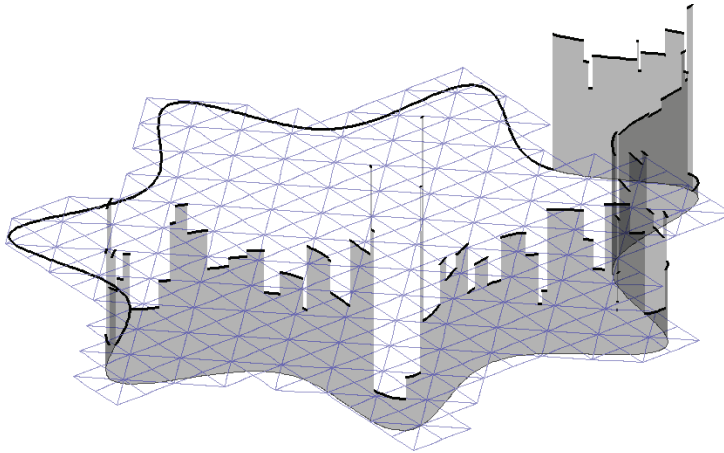


FIG. 15. Multiplier on Γ_D with the fully stabilized method for the P_1/P_0 method ($h = 0.05$).

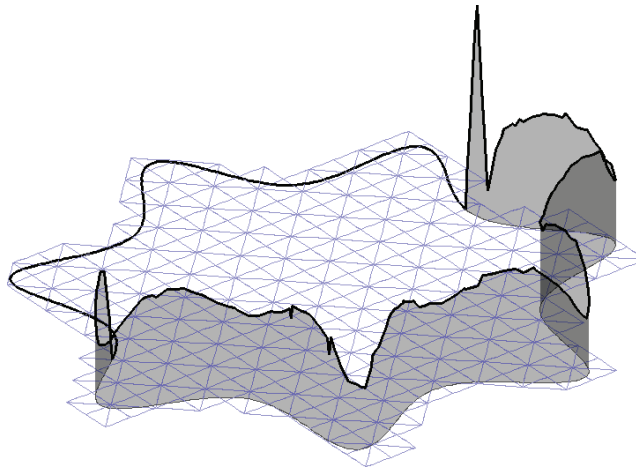


FIG. 16. Multiplier on Γ_D with the fully stabilized method for the P_2/P_1 method ($h = 0.05$).

tion. However, we see that the multipliers behave in a more regular way than before (see Figures 15 and 16). The difference lies only on the elements having a too small intersection with the domain.

8. Concluding remarks. In this paper, we combined the Xfem approach together with the Barbosa–Hughes stabilized formulation to get a new fictitious domain method. This method is quite simple to implement since all the variables (multipliers and primal variables) are defined on a single mesh independent of the computational domain. Moreover, it potentially allows one to treat complex boundary conditions (such as contact and friction).

The fully stabilized method introduced in section 5.2 leads to a robust method in the sense that it converges whatever is the intersection of the domain with the mesh. This is not the case if the Barbosa–Hughes stabilization technique is used alone, for which the quality of the approximation of the multiplier cannot be guaranteed on the elements having a too small intersection with the domain. Note that in [21] a similar

approach is presented. However, the error estimate is given under the assumption (44) and the definition of multipliers requires the construction of a quasi-uniform family of meshes on the boundary.

Appendix A. In this appendix we prove the trace inequality (42). For a proof in a more classical framework see, for instance, [12]. The proof is done by scaling with respect to a reference element \hat{T} .

We recall that for all $T \in \mathcal{T}^h$ one has $T = \tau_T(\hat{T})$, where τ_T is an affine and invertible mapping in \mathbb{R}^d . We make the following hypotheses:

- (a) Γ_D is a Lipschitz-continuous boundary.
- (b) there exists a constant $C_2 > 0$ independent of h and $T \in \mathcal{T}^h$ such that $\|\nabla\tau_T\|_{\infty,T} \leq C_2 h_T$ and $\|\nabla\tau_T^{-1}\|_{\infty,T} \leq C_2 h_T^{-1}$.

These two hypotheses are satisfied for regular families of meshes provided that Γ_D is Lipschitz-continuous.

LEMMA 6. *Let (a) and (b) be satisfied. Then there exists a constant $C > 0$ independent of h and $T \in \mathcal{T}^h$ such that*

$$\|v\|_{0,\Gamma_D \cap T}^2 \leq C (h_T^{-1} \|v\|_{0,T}^2 + h_T \|v\|_{1,T}^2), \quad \forall v \in H^1(T).$$

Proof. Since $C^\infty(T)$ is dense in $H^1(T)$ one can confine to functions $v \in C^\infty(T)$. Denoting $\hat{\Gamma}_D = \tau_T^{-1}(\Gamma_D \cap T)$ and $\hat{\mathbf{n}}$ a unit normal vector to $\hat{\Gamma}_D$, one has

$$\int_{\Gamma_D \cap T} v^2 d\Gamma = \int_{\hat{\Gamma}_D} \hat{v}^2 |\det(\nabla\tau_T)| \|\nabla\tau_T^{-1}\hat{\mathbf{n}}\| d\hat{\Gamma} \leq C h_T^{d-1} \int_{\hat{\Gamma}_D} \hat{v}^2 d\hat{\Gamma},$$

where $\hat{v} = v \circ \tau_T$. Let us prove now that the following trace inequality:

$$(46) \quad \int_{\hat{\Gamma}_D} \hat{v}^2 d\hat{\Gamma} \leq C_3 \|\hat{v}\|_{1,\hat{T}}^2 \quad \forall \hat{v} \in C^\infty(\hat{T}),$$

is such that the constant $C_3 > 0$ does not depend on the position of $\hat{\Gamma}_D$ inside of \hat{T} . This has been proved for a straight intersection in [11]. Let us consider the case $\hat{\Gamma}_D$ curved. For a sufficiently small mesh parameter h the curve $\hat{\Gamma}_D$ is a graph of a function over a segment \hat{l} contained in \hat{T} . Without loss of generality we may assume that \hat{l} coincides with the \hat{x} -axis (after appropriate shift and rotation of \hat{T}). Then Γ_D can be parametrized by the mean of a function $(\hat{x}, a(\hat{x}))$, $\hat{x} \in \hat{l}$ and one has

$$\hat{v}(\hat{x}, a(\hat{x})) = \hat{v}(\hat{x}, 0) + \int_0^{a(\hat{x})} \frac{\partial}{\partial y} \hat{v}(\hat{x}, \tau) d\tau.$$

Thus,

$$\hat{v}^2(\hat{x}, a(\hat{x})) \leq C \left(\hat{v}^2(\hat{x}, 0) + \int_0^{a(\hat{x})} \left(\frac{\partial}{\partial y} \hat{v}(\hat{x}, \tau) \right)^2 d\tau \right),$$

where $C > 0$ is an absolute constant. Integrating over \hat{l} we obtain

$$\begin{aligned} \int_{\hat{l}} \hat{v}^2(\hat{x}, a(\hat{x})) d\hat{x} &\leq C \left(\int_{\hat{l}} \hat{v}^2(\hat{x}, 0) d\hat{x} + \int_{\hat{T}} \left(\frac{\partial}{\partial y} \hat{v}(\hat{x}, \tau) \right)^2 d\tau d\hat{x} \right) \\ &\leq C \|v\|_{1,\hat{T}} \end{aligned}$$

using the result for the straight segment \hat{l} . Now, we can conclude by the fact that

$$\begin{aligned} \int_{\hat{\Gamma}_D} \hat{v}^2 d\hat{\Gamma} &= \int_{\hat{l}} \hat{v}^2(\hat{x}, a(\hat{x})) \sqrt{1 + (a'(\hat{x}))^2} d\hat{x} \\ &\leq \max_{\hat{l}} \sqrt{1 + (a'(\hat{x}))^2} \int_{\hat{l}} \hat{v}^2(\hat{x}, a(\hat{x})) d\hat{x}, \end{aligned}$$

since Γ_D is assumed to be Lipschitz-continuous.

Using now (46) and $\|\nabla \hat{v}\|_{\infty, \hat{T}} \leq Ch_T \|\nabla v\|_{\infty, T}$ we can establish the estimate of the lemma:

$$\begin{aligned} \|v\|_{0, \Gamma_D \cap T}^2 &\leq Ch_T^{d-1} \int_{\hat{T}} \left(\hat{v}^2 + |\hat{\nabla} \hat{v}|^2 \right) d\hat{x} \\ &\leq Ch_T^{-1} \int_{\hat{T}} \left(\hat{v}^2 + |\hat{\nabla} \hat{v}|^2 \right) |\det(\nabla \tau_T)| d\hat{x} \\ &\leq Ch_T^{-1} \int_T v^2 dx + Ch_T \int_T |\nabla v|^2 dx. \quad \square \end{aligned}$$

Appendix B. We prove the discrete trace inequality (27) when $R^h(v^h) = \partial_n v^h$ provided that (44) is satisfied under the same hypotheses on the family of meshes and on Γ_D as in Appendix A. First we prove the following auxiliary result.

LEMMA 7. Let v^h be defined on $T \in \mathcal{T}^h$ by $v^h(x) := \hat{v}(\tau_T^{-1}(x))$ with $\hat{v} \in P_k(\mathbb{R}^d)$ and suppose that (44) is satisfied. Then there exists a constant $C > 0$ independent of h, T , and \hat{v} such that

$$\int_{\Gamma_D \cap T} (v^h)^2 d\Gamma \leq Ch_T^{-1} \int_{\Omega \cap T} (v^h)^2 dx.$$

Proof. Because of the equivalence of norms on $P_k(\mathbb{R}^d)$, one has

$$\begin{aligned} \|\hat{v}\|_{\infty, \hat{T}}^2 &\leq \|\hat{v}\|_{\infty, B(\hat{y}_T, 2)}^2 = \|\hat{v} \circ t_{(-\hat{y}_T)}\|_{\infty, B(0, 2)}^2 \\ &\leq C \|\hat{v} \circ t_{(-\hat{y}_T)}\|_{0, B(0, \hat{\rho})}^2 = C \|\hat{v}\|_{0, B(\hat{y}_T, \hat{\rho})}^2 \leq C \int_{\tau_T^{-1}(T \cap \Omega)} \hat{v}^2 d\hat{x}, \end{aligned}$$

where $t_{(-\hat{y}_T)}$ is the translation defined by $t_{(-\hat{y}_T)}(x) = x - \hat{y}_T$. Thus, still with notations of Appendix A:

$$\begin{aligned} \int_{\Gamma_D \cap T} (v^h)^2 d\Gamma &= \int_{\hat{\Gamma}_D} \hat{v}^2 |\det(\nabla \tau_T)| \|\nabla \tau_T^{-1} \hat{\mathbf{n}}\| d\hat{\Gamma} \leq Ch_T^{d-1} \|\hat{v}\|_{\infty, \hat{T}}^2 |\hat{\Gamma}_D| \\ &\leq C \frac{h_T^{d-1}}{h_T^d} \int_{\tau_T^{-1}(T \cap \Omega)} \hat{v}^2 |\det(\nabla \tau_T)| d\hat{x} = Ch_T^{-1} \int_{T \cap \Omega} (v^h)^2 dx. \quad \square \end{aligned}$$

Now, summing up the previous estimate over elements of \mathcal{T}^h one obtains the following result.

LEMMA 8. Let v^h be defined on Ω by $v^h(x)|_T = \hat{v}_T(\tau_T^{-1}(x))$, $\hat{v}_T \in P_k(\mathbb{R}^d)$, $T \in \mathcal{T}^h$, and suppose that (44) is satisfied. Then the following estimate holds with a constant $C > 0$ independent of h and v^h :

$$h \int_{\Gamma_D} (v^h)^2 d\Gamma \leq C \int_{\Omega} (v^h)^2 dx.$$

The discrete trace inequality (27) can be now easily deduced since $\|\partial_n v^h\|_{0,\Gamma_D} \leq \|\nabla v^h\|_{0,\Gamma_D}$, and for a quasi-uniform family of meshes the previous lemma can be applied to ∇v^h componentwise.

Appendix C. We now adapt the proof of Appendix B to the operator $R^h(v^h)$ defined in section 5.2. The difference comes from those elements $T \in \mathcal{T}^h$ having a too small intersection with Ω (“bad” elements) and for which a neighbor element T' has been selected to make a natural extension of functions. For such an element, the proof of Appendix B has to be modified because we evaluate the polynomial on a larger zone than $\hat{T} = \tau_{T'}^{-1}(T')$, namely, on $\hat{T}_{T,T'} = \tau_{T'}^{-1}(T' \cup (T \cap \Omega))$. With the quasi-uniform assumption for the meshes, it is readily seen that this zone is included in $\hat{T}_{\rho_R} = \{x \in \mathbb{R}^d : \text{dist}(x, \hat{T}) \leq \rho_R\}$ for some $\rho_R > 0$ independent of h , T and T' . Lemma 7 can be easily adapted remarking that there exists a constant $C > 0$ independent of h such that

$$\|\hat{v}\|_{\infty, \hat{T}_{\rho_R}} \leq C \|\hat{v}\|_{\infty, \hat{T}} \quad \forall \hat{v} \in P_k(\mathbb{R}^d),$$

using again that all norms are equivalent in $P_k(\mathbb{R}^d)$. From this the estimate

$$\int_{\Gamma_D \cap T} (v^h)^2 d\Gamma \leq Ch_T^{-1} \int_{\Omega \cap T'} (v^h)^2 dx,$$

where $v^h(x) := \hat{v}(\tau_{T'}^{-1}(x))$, $x \in \mathbb{R}^d$ follows. Thus, (27) can be established remarking that the element T' can be selected as a neighbor element only a finite number times independently of h still due to the quasi-uniform property of the meshes.

Acknowledgment. Many thanks to Julien Pommier for his participation to obtain nice numerical experiments.

REFERENCES

- [1] R.A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
- [2] H.J.C. BARBOSA AND T.J.R. HUGHES, *The finite element method with Lagrange multipliers on the boundary: Circumventing the Babuška-Brezzi condition*, *Comput. Methods Appl. Mech. Engrg.*, 85 (1991), pp. 109–128.
- [3] H.J.C. BARBOSA AND T.J.R. HUGHES, *Boundary Lagrange multipliers in finite element methods: Error analysis in natural norms*, *Numer. Math.*, 62 (1992), pp. 1–15.
- [4] P. HANSBO, C. LOVADINA, I. PERUGIA, AND G. SANGALLI, *A Lagrange multiplier method for the finite element solution of elliptic interface problems using nonmatching meshes*, *Numer. Math.*, 100 (2005), pp. 91–115.
- [5] E. CHAHINE, P. LABORDE, AND Y. RENARD, *Crack-tip enrichment in the Xfem method using a cut-off function*, *Internat. J. Numer. Methods Engrg.*, 75 (2008), pp. 629–646.
- [6] P.G. CIARLET, *The finite element method for elliptic problems*, *Studies in Mathematics and its Applications* 4, North-Holland, Amsterdam, 1978.
- [7] P.G. CIARLET, *The finite element method for elliptic problems*, in *Handbook of Numerical Analysis*, Volume II, Part 1, P.G. Ciarlet and J.L. Lions, eds., North-Holland, Amsterdam, 1991, pp. 17–352.
- [8] P. CLÉMENT, *Approximation by finite elements functions using local regularization*, *RAIRO, Anal. Numer.*, 9 (1975), pp. 77–84.
- [9] J.W. DEMMEL, J.R. GILBERT, AND X.S. LI, *A general purpose library for the direct solution of large, sparse, nonsymmetric systems*, <http://crd.lbl.gov/~xiaoye/SuperLU/>.
- [10] A. ERN AND J.-L. GUERMOND, *Theory and practice of finite elements*, *Appl. Math. Sci.*, 159 (2004).
- [11] V. GIRAULT AND R. GLOWINSKI, *Error analysis of a fictitious domain method applied to a Dirichlet problem*, *Japan J. Indust. Appl. Math.*, 12 (1995), pp. 487–514.
- [12] P. GRISVARD, *Elliptic problems in nonsmooth domains. Monographs and Studies in Mathematics*, Pitman (Advanced Publishing Program), Boston, MA, 1985.

- [13] J. HASLINGER AND A. KLARBRING, *Fictitious domain/mixed finite element approach for a class of optimal shape design problems*, Math. Model. Numer. Anal. (M2AN), 4 (1995), pp. 435–450.
- [14] P. HILD AND Y. RENARD, *A stabilized Lagrange multiplier method for the finite element approximation of contact problems in elastostatics*, submitted.
- [15] T. HUGHES AND L.P. FRANCA, *A new finite element formulation for computational fluid dynamics. VII. The Stokes problem with various well-posed boundary conditions: Symmetric formulations that converge for all velocity/pressure spaces*, Comput. Methods Appl. Mech. Engrg., 65 (1987), pp. 85–96.
- [16] P. LABORDE, J. POMMIER, Y. RENARD, AND M. SALAÜN, *High order extended finite element method for cracked domains*, Internat. J. Numer. Methods Engrg., 64 (2005), pp. 354–381.
- [17] N. MOËS, E. BÉCHET, AND M. TOURBIER, *Imposing Dirichlet boundary conditions in the eXtended Finite Element Method*, Internat. J. Numer. Methods Engrg., 12 (2006), pp. 354–381.
- [18] N. MOËS, J. DOLBOW, AND T. BELYTSCHKO, *A finite element method for crack growth without remeshing*, Internat. J. Numer. Methods Engrg., 46 (1999), pp. 131–150.
- [19] N. MOËS, A. GRAVOUIL, AND T. BELYTSCHKO, *Non-planar 3D crack growth by the extended finite element and level sets, Part I: Mechanical model*, Internat. J. Numer. Methods Engrg., 11 (2002), pp. 2549–2568.
- [20] J. NITSCHKE, *Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind*, Abh. Math. Univ. Hamburg, 36 (1971), pp. 9–15.
- [21] J. PITKÄRANTA, *Local stability conditions for the Babuška method of Lagrange multipliers*, Math. Comput., 35 (1980), pp. 1113–1129.
- [22] Y. RENARD AND J. POMMIER, *Getfem++*. An open source generic C++ library for finite element methods, <http://home.gna.org/getfem/>.
- [23] F.L. STAZI, E. BUDYN, J. CHESSA, AND T. BELYTSCHKO, *An extended finite element method with higher-order elements for curved cracks*, Comput. Mech., 31 (2003), pp. 38–48.
- [24] R. STENBERG, *On some techniques for approximating boundary conditions in the finite element method*, J. Comput. Appl. Math., 63 (1995), pp. 139–148.
- [25] M. STOLARSKA, D.L. CHOPP, N. MOËS, AND T. BELYTSCHKO, *Modelling crack growth by level sets in the extended finite element method*, Internat. J. Numer. Methods Engrg., 51 (2001), pp. 943–960.
- [26] G. STRANG AND G.J. FIX, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
- [27] N. SUKUMAR, D.L. CHOPP, N. MOËS, AND T. BELYTSCHKO, *Modeling holes and inclusions by level sets in the extended finite element method*, Comput. Methods Appl. Mech. Eng., 46 (2001), pp. 6183–6200.
- [28] N. SUKUMAR, N. MOËS, B. MORAN, AND T. BELYTSCHKO, *Extended finite element method for three dimensional crack modelling*, Internat. J. Numer. Methods Engrg., 48 (2000), pp. 1549–1570.

A SADDLE POINT APPROACH TO THE COMPUTATION OF HARMONIC MAPS*

QIYA HU[†], XUE-CHENG TAI[‡], AND RAGNAR WINTHER[§]

Abstract. In this paper we consider numerical approximations of a constraint minimization problem, where the object function is a quadratic Dirichlet functional for vector fields and the interior constraint is given by a convex function. The solutions of this problem are usually referred to as harmonic maps. The solution is characterized by a nonlinear saddle point problem, and the corresponding linearized problem is well-posed near strict local minima. The main contribution of the present paper is to establish a corresponding result for a proper finite element discretization in the case of two space dimensions. Iterative schemes of Newton type for the discrete nonlinear saddle point problems are investigated, and mesh independent preconditioners for the iterative methods are proposed.

Key words. harmonic maps, nonlinear constraints, saddle point problems, error estimates

AMS subject classifications. 35A40, 65C20, 65N30

DOI. 10.1137/060675575

1. Introduction. The solutions of many systems of linear partial differential equations can be characterized as minimizers of quadratic functionals over a set of linear constraints. Examples of such systems are the linear Stokes system for fluid flow, the Reissner–Mindlin plate model, and the so-called mixed formulation of second order elliptic equations. The discretizations of these systems lead to linear systems with a saddle point structure, and conditioning of the systems deteriorates as the mesh becomes finer. As a consequence, substantial research on preconditioned iterative methods for the corresponding discrete systems has taken place; cf., for example, [2, 3] or [18, Chapter 6]. The purpose of the present paper is to perform a corresponding analysis for a nonlinear problem. We will study a simple variant of the problem characterizing harmonic maps with respect to a compact manifold. In particular, we will focus on stability and error estimates for the discretization and on preconditioning of the linear saddle point systems arising in a Newton iteration.

For a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, we shall consider the problem of finding local minima of a constrained minimization problem of the form

$$(1.1) \quad \min_{\mathbf{v} \in \mathbf{H}_g^1(\Omega; \mathcal{M})} \mathcal{E}(\mathbf{v}) = \frac{1}{2} \int_{\Omega} |\nabla \mathbf{v}|^2 dx.$$

*Received by the editors November 21, 2006; accepted for publication (in revised form) December 8, 2008; published electronically March 25, 2009. The work was supported by the Norwegian Research Council, LSEC (Laboratory of Scientific and Engineering Computing) at the Chinese Academy of Sciences, the Key Project of the Natural Science Foundation of China G10531080, the National Basic Research Program of China 2005CB321702, and the Natural Science Foundation of China G10771178.

<http://www.siam.org/journals/sinum/47-2/67557.html>

[†]LSEC, Institute of Computational Mathematics and Scientific Engineering Computing, Chinese Academy of Sciences, Beijing 100080, China (hqy@lsec.cc.ac.cn).

[‡]Department of Mathematics, University of Bergen, Johannes Brunsgate 12, Bergen, 5008, Norway and Division of Mathematical Science, School of Physical & Mathematical Sciences, Nanyang Technological University, 21 Nanyang Link, Singapore, 637371, Singapore (tai@mi.uib.no, xctai@ntu.edu.sg).

[§]Centre of Mathematics for Applications and Department of Informatics, University of Oslo, P.B. 1053, Blindern, Oslo, Norway (ragnar.winther@cma.uio.no).

Here $\mathbf{H}_{\mathbf{g}}^1(\Omega; \mathcal{M})$ is the set of vector fields with values in a smooth compact manifold \mathcal{M} in \mathbb{R}^d , with function values and first derivatives in $L^2(\Omega)$, and such that the elements \mathbf{v} of $\mathbf{H}_{\mathbf{g}}^1(\Omega; \mathcal{M})$ satisfies $\mathbf{v}|_{\partial\Omega} = \mathbf{g}$ for fixed vector field \mathbf{g} defined on the boundary $\partial\Omega$. We will further assume that the target manifold \mathcal{M} is implicitly given in the form

$$\mathcal{M} = \{\mathbf{v} \in \mathbb{R}^d \mid F(\mathbf{v}) = 0\},$$

where the function $F : \mathbb{R}^d \rightarrow \mathbb{R}^k$ is a smooth function, and it will be assumed that the compatibility condition $F(\mathbf{g}) = 0$ holds. More specific assumptions on F and the boundary data \mathbf{g} will be given later. Problems of the form (1.1) arise, for example, in liquid crystal and superconductor simulations. The solutions of the problem (1.1) are frequently referred to as harmonic maps [7]. In the present paper we will restrict our study to the case $k = 1$, i.e., \mathcal{M} is of dimension $d - 1$. We will focus on a nonlinear saddle point approach to compute the solutions of the problem (1.1).

For a review of results on the continuous harmonic map problem, we refer to [7, 24, 29, 30]. The purpose of the present paper is to discuss a finite element method for approximating the constraint minimization problem (1.1). For the simplest case of (1.1), with interior constraint given by $|\mathbf{v}| = 1$, several numerical approaches have been discussed; cf., for example, [1, 4, 5, 13, 14, 15, 16, 20, 21, 25, 26, 32]. Variants of the projection method are proposed and analyzed in [1, 5, 16]. However, the standard projection method applies only to the simplest model. Moreover, it was illustrated in [5] that the projection method converges only for very special regular and quasi-uniform triangulations for the discretized harmonic map problem. The relaxation method of [13, 21, 25] is using point relaxation with the constraint required at each grid point. Both convergence analysis and numerical experiments are supplied in [25]. An advantage with the relaxation method is that it is very easy to implement. However, disadvantages are that the relaxation parameter has to be chosen properly to obtain convergence and that the convergence of such fixed point iterations is slow. Another commonly used approach for harmonic map problems is to use penalization methods; cf. [4, 14, 15, 16, 20]. It is even often combined with the gradient decent method, which produces some time evolution equations; cf. [4, 11, 12, 14, 15, 16, 20]. The approach and analysis given in [4] even work for general p -harmonic problems, with p close to 1. The analysis of [14, 15] is also valid for problems coupling harmonic maps with Navier–Stokes equations.

The main contribution of the present paper is to discuss the use of a saddle point approach for the construction of numerical methods for the constraint minimization problem (1.1). We will show that the corresponding saddle point problem is stable near exact local minima. This is achieved by verifying the standard stability conditions for linear saddle point problems. This verification has the extra difficulty in that the coercivity condition will not hold, in general, but only on the kernel of the linearized constraint. Using the standard stability conditions for the corresponding discrete saddle point problem, we will construct finite element methods such that the corresponding discrete solutions admit an optimal error estimate in the energy norm. Due to some technical difficulties, caused by the use of inverse inequalities to handle some nonlinear terms, this analysis of the finite element discretization is restricted to two space dimensions, i.e., $d = 2$. In this case we also establish that any critical point of the functional \mathcal{E} with respect to $\mathbf{H}_{\mathbf{g}}^1(\Omega; \mathcal{M})$ is indeed a local minimum. Compared with other approaches [4, 11, 14, 15], our estimates do not depend on extra artificial parameters like a weight parameter for the penalty method or a step size for a gradient flow. We will also study Newton’s method for the discrete nonlinear saddle

point problem and propose a simple and efficient preconditioner for the linear systems arising during the iterations. Numerical tests will be given to show the efficiency of the proposed method.

The outline of the paper is as follows. In section 2, the notations and assumption will be specified. In section 3, the continuous problem is studied. The problem (1.1) is formally transformed to a saddle point problem, and stability results will be proved for the continuous model. In section 4 we first describe a finite element discretization for (1.1), and then the discrete stability conditions are established. Using these stability conditions, the existence, local uniqueness, and the error estimates are derived in section 5. Variants of Newton's method are analyzed in section 6, while numerical experiments are presented in section 7.

2. Notation and preliminaries. Throughout this paper we will use c and C to denote generic positive constants, not necessarily the same at different occurrences. It is assumed that the constants are independent of the mesh size h , which will be introduced later. For vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$, we use $\mathbf{v} \cdot \mathbf{w}$ to denote the Euclidian inner product, while the notation $\mathbf{A} : \mathbf{B}$ is used to denote the Frobenius inner product of two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d \times d}$. The corresponding norms are given by $|\mathbf{v}|$ and $|\mathbf{A}|$, respectively. For a vector or matrix \mathbf{A} , \mathbf{A}^t is the transpose of \mathbf{A} . In the special case of vectors $\mathbf{v} = (v_1, v_2)$ in \mathbb{R}^2 we will use $\mathbf{v}^\perp = (-v_2, v_1)$ to denote the corresponding vector obtained by a rotation of 90 degrees.

For $m \geq 0$, we will use $H^m = H^m(K)$ to denote the real valued L^2 -based Sobolev spaces on domain $K \subset \mathbb{R}^d$, the corresponding norm by $\|\cdot\|_{m,K}$, and $|\cdot|_{m,K}$ is the seminorm involving only the m th order derivatives. The subspace H_0^m is the closure in H^m of $C_0^\infty(K)$, while H^{-m} is the dual of H_0^m with respect to an extension of the L^2 inner product $\langle \cdot, \cdot \rangle$. The corresponding L^∞ -based Sobolev spaces are denoted $W^{m,\infty}(K)$, with associated norm $\|\cdot\|_{m,\infty,K}$. For all the Sobolev norms, we will omit K in case $K = \Omega$. In general, we will use boldface symbols for vector or matrix valued functions. The gradient operator with respect to the spatial variable $\mathbf{x} = (x_1, x_2, \dots, x_d)$ is denoted $\nabla = (\partial/\partial x_1, \partial/\partial x_2, \dots, \partial/\partial x_d)^t$. Furthermore, the gradient of a vector valued function $\mathbf{v} = (v_1, v_2, \dots, v_d)^t$, $\nabla \mathbf{v}$, is the matrix valued function obtained by taking the gradient rowwise, i.e., $(\nabla \mathbf{v})_{ij} = \partial v_i / \partial x_j$.

In order to specify the properties of the constraint functional $F : \mathbb{R}^d \rightarrow \mathbb{R}$, defining the constraint manifold \mathcal{M} , we will use $\mathbf{D}F$ to denote the gradient of F , i.e., $\mathbf{D}F(\mathbf{v}) = (\partial F / \partial v_1, \dots, \partial F / \partial v_d)^t$ and the corresponding Hessian by $\mathbf{D}^2 F(\mathbf{v}) = (\partial^2 F / \partial v_i \partial v_j)_{i,j=1}^d$. Throughout this paper we will assume that the constraint functional F satisfies the following:

- (i) F is convex and smooth. Furthermore, there exist constants c_0 and c_1 such that

$$(2.1) \quad c_0 |\mathbf{v}|^2 \leq \mathbf{D}^2 F(\xi) \mathbf{v} \cdot \mathbf{v} \leq c_1 |\mathbf{v}|^2, \quad \xi, \mathbf{v} \in \mathbb{R}^d.$$

- (ii) $F(\mathbf{0}) < 0$ and $\mathbf{D}F(\mathbf{0}) = 0$.

- (iii) There exists an $\ell > 0$ such that the matrix function $\mathbf{D}^2 F$ satisfies

$$(2.2) \quad |\mathbf{D}^2 F(\xi_1) - \mathbf{D}^2 F(\xi_2)| \leq \ell |\xi_1 - \xi_2|, \quad \xi_1, \xi_2 \in \mathbb{R}^d.$$

The analysis below will still hold if the assumptions (2.1) and (2.2) are only valid for all ξ, ξ_1, ξ_2 in a neighborhood of a continuous true solution.

For the boundary function \mathbf{g} of (1.1), we assume that it has been extended into the interior of Ω such that $\mathbf{g} \in \mathbf{H}^1(\Omega)$. Corresponding to \mathbf{g} , we let

$$\mathbf{H}_{\mathbf{g}}^1(\Omega) = \{\mathbf{v} \in \mathbf{H}^1(\Omega) : \mathbf{v} = \mathbf{g} \text{ on } \partial\Omega\}.$$

If $\mathbf{v} : \Omega \rightarrow \mathbb{R}^d$ is a smooth vector field, then it follows from the chain rule that

$$(2.3) \quad \nabla F(\mathbf{v}) = (\nabla \mathbf{v})^t \mathbf{D}F(\mathbf{v}),$$

where the product on the right-hand side is the ordinary matrix-vector product. Furthermore, we have

$$(2.4) \quad \nabla \mathbf{D}F(\mathbf{v}) = \mathbf{D}^2 F(\mathbf{v}) \nabla \mathbf{v}.$$

From assumptions (i)–(ii) and the Taylor expansion we obtain the following estimate:

$$(2.5) \quad 2c_1^{-1}|F(\mathbf{0})| \leq |\mathbf{v}(\mathbf{x})|^2 \leq 2c_0^{-1}|F(\mathbf{0})|, \quad \mathbf{x} \in \Omega$$

for any \mathbf{v} satisfying $F(\mathbf{v}) \equiv 0$ in Ω . Similarly, we derive

$$(2.6) \quad |\mathbf{D}F(\mathbf{v})| \geq c_0|\mathbf{v}|$$

for any \mathbf{v} , and hence $|\mathbf{D}F(\mathbf{v}(\mathbf{x}))| > 0$ if $\mathbf{v}(\mathbf{x}) \in \mathcal{M}$.

Let us note that the interior constraint in (1.1), given by $\mathbf{v}(\mathbf{x}) \in \mathcal{M}$, implies that a local minimum of (1.1) satisfies $\mathbf{u} \in \mathbf{H}_{\mathbf{g}}^1(\Omega) \cap \mathbf{L}^\infty(\Omega)$. In fact, if we restrict the analysis to the case $d = 2$, with the manifold \mathcal{M} taken to be the unit circle \mathbf{S}^1 , and we assume that the boundary $\partial\Omega$ and the boundary data \mathbf{g} are sufficiently regular, then there is a unique smooth global minimizer of (1.1) under the condition that the degree of \mathbf{g} is zero; cf. [7, Theorem 12] and [22]. However, this result is not true for more general harmonic map problems [30, 24].

We will first consider the characterization of critical points of the functional \mathcal{E} over $\mathbf{H}_{\mathbf{g}}^1(\Omega; \mathcal{M})$. The outline below follows a standard Lagrange multiplier approach to constrained optimization; cf., for example, [6] for the finite-dimensional case or [17, 19] in the infinite-dimensional case. A vector field $\mathbf{u} \in \mathbf{H}_{\mathbf{g}}^1(\Omega; \mathcal{M})$ is such a critical point if it satisfies

$$(2.7) \quad \langle \nabla \mathbf{u}, \nabla \mathbf{v} \rangle = 0$$

for any \mathbf{v} in the tangent space of $\mathbf{H}_{\mathbf{g}}^1(\Omega; \mathcal{M})$ at \mathbf{u} , i.e., for any $\mathbf{v} \in \mathbf{H}_0^1(\Omega)$ such that $\mathbf{D}F(\mathbf{u}) \cdot \mathbf{v} \equiv 0$. In the saddle point approach which we shall consider here we will view the critical points \mathbf{u} as elements of the larger space $\mathbf{H}_{\mathbf{g}}^1(\Omega)$. Assume that \mathbf{u} has the extra regularity property that

$$(2.8) \quad \mathbf{u} \in \mathbf{H}_{\mathbf{g}}^1(\Omega) \cap \mathbf{W}^{1,\infty}(\Omega).$$

Then any such \mathbf{u} is a critical point if and only if there is a $\lambda \in L^2(\Omega)$ such that the pair (\mathbf{u}, λ) satisfies the first order conditions

$$(2.9) \quad \begin{aligned} \langle \nabla \mathbf{u}, \nabla \mathbf{v} \rangle + \langle \mathbf{D}F(\mathbf{u}) \cdot \mathbf{v}, \lambda \rangle &= 0, & \mathbf{v} \in \mathbf{H}_0^1(\Omega), \\ \langle F(\mathbf{u}), \mu \rangle &= 0, & \mu \in H^{-1}(\Omega). \end{aligned}$$

To see this we assume that \mathbf{u} is a critical point satisfying (2.8), and let $\mathbf{z} = \mathbf{D}F(\mathbf{u})/|\mathbf{D}F(\mathbf{u})|$. For any $\mathbf{v} \in \mathbf{H}_0^1(\Omega)$, let $\mathbf{v}_\tau = \mathbf{v} - (\mathbf{v} \cdot \mathbf{z})\mathbf{z}$. As a consequence $\mathbf{D}F(\mathbf{u}) \cdot \mathbf{v}_\tau = 0$, and, by (2.7),

$$0 = \langle \nabla \mathbf{u}, \nabla \mathbf{v}_\tau \rangle = \langle \nabla \mathbf{u}, \nabla \mathbf{v} \rangle - \langle \nabla \mathbf{u}, \nabla(\mathbf{v} \cdot \mathbf{z})\mathbf{z} \rangle.$$

From (2.3), the constraint implies that $(\nabla \mathbf{u})^t \mathbf{z} = 0$. Therefore, the final inner product above can be rewritten as

$$\langle \nabla \mathbf{u}, \nabla(\mathbf{v} \cdot \mathbf{z})\mathbf{z} \rangle = \langle \nabla \mathbf{u} : \nabla \mathbf{z}, \mathbf{v} \cdot \mathbf{z} \rangle.$$

Hence, the system (2.9) holds with

$$(2.10) \quad \lambda = -\nabla \mathbf{u} : \nabla \mathbf{z} / |\mathbf{D}F(\mathbf{u})| = -\nabla \mathbf{u} : \nabla \mathbf{D}F(\mathbf{u}) / |\mathbf{D}F(\mathbf{u})|^2,$$

where the last identity again is a consequence of the constraint. Note that it follows from (2.8) that the multiplier λ is actually in $L^\infty(\Omega)$.

The variational problem (2.9) is the Euler–Lagrangian equation for the constrained minimization problem (1.1), and the system is a weak formulation of the problem

$$(2.11) \quad \begin{aligned} -\Delta \mathbf{u} + \lambda \mathbf{D}F(\mathbf{u}) &= 0 && \text{in } \Omega, \\ F(\mathbf{u}) &= 0 && \text{in } \Omega. \end{aligned}$$

In the simplest case when $\mathcal{M} = \mathbf{S}^{d-1}$, i.e., the unit disc in \mathbb{R}^d , we have $\lambda = -|\nabla \mathbf{u}|^2$ and

$$-\Delta \mathbf{u} - |\nabla \mathbf{u}|^2 \mathbf{u} = 0 \text{ in } \Omega \quad \mathbf{u} = \mathbf{g} \text{ on } \partial\Omega.$$

In the present paper we will restrict our attention to the critical points \mathbf{u} of \mathcal{E} over $\mathbf{H}_{\mathbf{g}}^1(\Omega; \mathcal{M})$ that are local minimizers. So assume that the pair (\mathbf{u}, λ) is a solution of (2.9), satisfying the regularity property (2.8), and let $\mathbf{w} = \mathbf{w}(t)$ be a smooth curve in $\mathbf{H}_{\mathbf{g}}^1(\Omega; \mathcal{M})$, defined for t in a neighborhood of the origin such that $\mathbf{w}(0) = \mathbf{u}$ and $\mathbf{w}'(0) = \mathbf{v}$. Hence, since $F(\mathbf{w}(t)) \equiv 0$, we must have $\mathbf{D}F(\mathbf{u}) \cdot \mathbf{v} = 0$, and

$$(2.12) \quad \mathbf{D}F(\mathbf{u}) \cdot \mathbf{w}''(0) = -\mathbf{D}^2F(\mathbf{u})\mathbf{v} \cdot \mathbf{v}.$$

Furthermore, if we define a real valued function $\phi = \phi(t)$ by

$$\phi(t) = \mathcal{E}(\mathbf{w}(t)) = \frac{1}{2} \langle \nabla \mathbf{w}(t), \nabla \mathbf{w}(t) \rangle,$$

then

$$\phi'(t) = \langle \nabla \mathbf{w}(t), \nabla \mathbf{w}'(t) \rangle \quad \text{and} \quad \phi''(t) = \langle \nabla \mathbf{w}'(t), \nabla \mathbf{w}'(t) \rangle + \langle \nabla \mathbf{w}(t), \nabla \mathbf{w}''(t) \rangle.$$

Hence, it follows from the system (2.9) that $\phi'(0) = \langle \nabla \mathbf{u}, \nabla \mathbf{v} \rangle = 0$, and if \mathbf{u} corresponds to a local minimum of \mathcal{E} over $\mathbf{H}_{\mathbf{g}}^1(\Omega; \mathcal{M})$, then the second order condition

$$\phi''(0) = \langle \nabla \mathbf{v}, \nabla \mathbf{v} \rangle + \langle \nabla \mathbf{u}, \nabla \mathbf{w}''(0) \rangle \geq 0$$

must hold. However, by using the system (2.9) and (2.12), we obtain that

$$\langle \nabla \mathbf{u}, \nabla \mathbf{w}''(0) \rangle = -\langle \mathbf{D}F(\mathbf{u}) \cdot \nabla \mathbf{w}''(0), \lambda \rangle = \langle \mathbf{D}^2F(\mathbf{u})\mathbf{v} \cdot \mathbf{v}, \lambda \rangle.$$

Therefore, the second order condition takes the form

$$(2.13) \quad \phi''(0) = \langle \nabla \mathbf{v}, \nabla \mathbf{v} \rangle + \langle \mathbf{D}^2F(\mathbf{u})\mathbf{v} \cdot \mathbf{v}, \lambda \rangle \geq 0.$$

In fact, let us refer to a local minimum \mathbf{u} of \mathcal{E} over $\mathbf{H}_{\mathbf{g}}^1(\Omega; \mathcal{M})$ as a *strict local minimum* if there is a positive constant β such that

$$\frac{d^2}{dt^2} \mathcal{E}(\mathbf{w}(t))|_{t=0} \geq \beta \|\mathbf{v}\|_1^2$$

for any smooth curve $\mathbf{w} = \mathbf{w}(t)$ in $\mathbf{H}_{\mathbf{g}}^1(\Omega; \mathcal{M})$ satisfying $\mathbf{w}(0) = \mathbf{u}$ and $\mathbf{w}'(0) = \mathbf{v}$. It follows from the calculation above that the function $\phi(t) = \mathcal{E}(\mathbf{w}(t))$ satisfies

$$(2.14) \quad \phi''(0) = \langle \nabla \mathbf{v}, \nabla \mathbf{v} \rangle + \langle \mathbf{D}^2F(\mathbf{u})\mathbf{v} \cdot \mathbf{v}, \lambda \rangle \geq \beta \|\mathbf{v}\|_1^2$$

for all $\mathbf{v} \in \mathbf{H}_0^1(\Omega)$ satisfying $\mathbf{D}F(\mathbf{u}) \cdot \mathbf{v} = 0$. As we shall see below this condition is closely tied to a stability condition for a linearization of the system (2.9).

The saddle point approach can be regarded as the limiting case of the penalty method. In the commonly used penalty approach, cf. [4, 14, 15, 16, 20], one is seeking a local minimizer of the following regularized problem:

$$\min_{\mathbf{v} \in \mathbf{H}_g^1(\Omega)} \mathcal{E}(\mathbf{v}) + \frac{1}{2\epsilon} \int_{\Omega} |F(\mathbf{v})|^2 dx,$$

where the penalty parameter $\epsilon > 0$ has to be properly chosen. The saddle point system (2.9) is formally obtained in the limit as ϵ tends to zero. The advantage of the saddle point approach is that the standard mixed finite element theory, cf. [9], tells us how to choose the finite element spaces properly to avoid possible instabilities, and there is no need to choose a penalty parameter.

3. Stability of the linearized problem. Throughout the rest of this paper we will assume that the pair (\mathbf{u}, λ) is a solution of the system (2.9), corresponding to a local minimum of \mathcal{E} over $\mathbf{H}_g^1(\Omega; \mathcal{M})$ and satisfying the regularity property

$$(3.1) \quad \mathbf{u} \in \mathbf{H}_g^1(\Omega) \cap \mathbf{W}^{1,\infty}(\Omega), \quad \lambda \in L^\infty(\Omega).$$

In particular, \mathbf{u} and λ are related by (2.10), and the second order condition (2.13) holds, i.e.,

$$a(\mathbf{u}, \lambda; \mathbf{v}, \mathbf{v}) \geq 0$$

for all $\mathbf{v} \in \mathbf{Z}_u$, where the bilinear form $a(\mathbf{u}, \lambda; \cdot, \cdot)$ is given by

$$a(\mathbf{u}, \lambda; \mathbf{v}, \hat{\mathbf{v}}) = \langle \nabla \mathbf{v}, \nabla \hat{\mathbf{v}} \rangle + \langle \mathbf{D}^2 F(\mathbf{u}) \mathbf{v} \cdot \hat{\mathbf{v}}, \lambda \rangle$$

and

$$\mathbf{Z}_u = \{ \mathbf{v} \in \mathbf{H}_0^1(\Omega) : \langle \mathbf{D}F(\mathbf{u}) \cdot \mathbf{v}, \mu \rangle = 0, \quad \mu \in L^2(\Omega) \}.$$

For the analysis below, it will be useful to consider linearization of the saddle point system (2.9). More precisely, we consider systems of the following form:

Find $(\mathbf{v}, \mu) \in \mathbf{H}_0^1(\Omega) \times H^{-1}(\Omega)$ such that

$$(3.2) \quad \begin{aligned} a(\mathbf{u}, \lambda; \mathbf{v}, \hat{\mathbf{v}}) + \langle \mathbf{D}F(\mathbf{u}) \cdot \hat{\mathbf{v}}, \mu \rangle &= \langle \mathbf{f}, \mathbf{v} \rangle, & \hat{\mathbf{v}} &\in \mathbf{H}_0^1(\Omega), \\ \langle \mathbf{D}F(\mathbf{u}) \cdot \mathbf{v}, \hat{\mu} \rangle &= \langle \sigma, \mu \rangle, & \hat{\mu} &\in H^{-1}(\Omega), \end{aligned}$$

where (\mathbf{u}, λ) is the exact solution of (2.9) satisfying (3.1). Here $\mathbf{f} \in \mathbf{H}^{-1}(\Omega)$ and $\sigma \in H_0^1(\Omega)$ represent data.

Our goal is to show that this linear system is well-posed, i.e., we will show that the map

$$(\mathbf{f}, \sigma) \in \mathbf{H}^{-1}(\Omega) \times H_0^1(\Omega) \mapsto (\mathbf{v}, \mu) \in \mathbf{H}_0^1(\Omega) \times H^{-1}(\Omega)$$

is well defined and bounded. This will be established by verifying the standard stability conditions for saddle points systems; cf. [8] or [9]. We will first establish the so-called inf-sup condition.

THEOREM 3.1. *Let (\mathbf{u}, λ) satisfy (3.1) and be related by (2.10). Then there is a positive constant β_1 , depending on \mathbf{u} , such that*

$$(3.3) \quad \inf_{\mu \in H^{-1}(\Omega)} \sup_{\mathbf{v} \in \mathbf{H}_0^1(\Omega)} \frac{\langle \mathbf{D}F(\mathbf{u}) \cdot \mathbf{v}, \mu \rangle}{\|\mathbf{v}\|_1 \|\mu\|_{-1}} \geq \beta_1.$$

Proof. For any $\mu \in H^{-1}(\Omega)$, there exists a $\varphi \in H_0^1(\Omega)$ such that

$$(3.4) \quad \frac{\langle \mu, \varphi \rangle}{\|\varphi\|_1} = \|\mu\|_{-1}.$$

Define $\mathbf{v} = \varphi \frac{\mathbf{w}}{|\mathbf{w}|^2}$, where $\mathbf{w} = \mathbf{D}F(\mathbf{u})$. Then, by Leibniz' rule, there exists a $c > 0$, depending on \mathbf{u} , such that

$$\|\nabla \mathbf{v}\|_0 \leq c\|\varphi\|_1.$$

Furthermore,

$$\langle \mathbf{D}F(\mathbf{u}) \cdot \mathbf{v}, \mu \rangle = \langle \varphi, \mu \rangle = \|\varphi\|_1 \|\mu\|_{-1}.$$

Hence, the desired inequality holds with $\beta_1 = 1/c$. \square

Next we need to consider the properties of the bilinear form $a(\mathbf{u}, \lambda; \cdot, \cdot)$. It is straightforward to check that this bilinear form is bounded in the sense that

$$(3.5) \quad a(\mathbf{u}, \lambda; \mathbf{v}, \hat{\mathbf{v}}) \leq C(\mathbf{u}, \lambda) |\mathbf{v}|_1 |\hat{\mathbf{v}}|_1, \quad \mathbf{v}, \hat{\mathbf{v}} \in \mathbf{H}_0^1(\Omega),$$

where the constant $C(\mathbf{u}, \lambda)$ depends on the norms of \mathbf{u} and λ indicated by (3.1).

The final key property for the stability analysis of the linear system (3.2) is the requirement that the bilinear form $a(\mathbf{u}, \lambda; \cdot, \cdot)$ is coercive on the linearized constraint space $\mathbf{Z}_{\mathbf{u}}$. It should be noted that this bilinear form is, in general, not coercive on the entire space $\mathbf{H}_0^1(\Omega)$. For example, in the simplest case when $\mathcal{M} = \mathbf{S}^{d-1}$, we have

$$a(\mathbf{u}, \lambda; \mathbf{v}, \mathbf{v}) = \int_{\Omega} (|\nabla \mathbf{v}|^2 - |\nabla \mathbf{u}|^2 |\mathbf{v}|^2) \, d\mathbf{x}.$$

On the other hand, the stability theory of [8] requires only that

$$(3.6) \quad a(\mathbf{u}, \lambda; \mathbf{v}, \mathbf{v}) \geq \beta \|\mathbf{v}\|_1^2, \quad \mathbf{v} \in \mathbf{Z}_{\mathbf{u}}$$

for a suitable positive constant β , and this is exactly the strict minimum condition (2.14). Therefore, if \mathbf{u} is a strict local minimum, then the linear system (3.2) is well-posed.

Furthermore, if we restrict to two space dimensions, i.e. $d = 2$, then the coercivity condition (3.6) always holds. This is a consequence of the following theorem, which implies that in this case every critical point (\mathbf{u}, λ) satisfying (3.1) is a strict local minimum, and the corresponding problem (3.2) is well-posed.

THEOREM 3.2. *Assume that $d = 2$. Let (\mathbf{u}, λ) satisfy (3.1) and be related by (2.10). Then there is a positive constant β_2 , depending on \mathbf{u} , such that*

$$(3.7) \quad a(\mathbf{u}, \lambda; \mathbf{v}, \mathbf{v}) = \langle \nabla \mathbf{v}, \nabla \mathbf{v} \rangle + \langle \mathbf{D}^2 F(\mathbf{u}) \mathbf{v} \cdot \mathbf{v}, \lambda \rangle \geq \beta_2 \|\mathbf{v}\|_1^2, \quad \mathbf{v} \in \mathbf{Z}_{\mathbf{u}}.$$

Remark 3.1. The result of this theorem will not be true, in general, if the target manifold \mathcal{M} is of higher dimension. However, in [23] a sufficient condition on \mathbf{u} and \mathcal{M} , referred to as the ‘‘cut locus condition,’’ is given, which ensures that the operator associated with the bilinear form $a(\mathbf{u}, \lambda; \cdot, \cdot)$, restricted to the tangent space $\mathbf{Z}_{\mathbf{u}}$, is invertible, and hence the linear system (3.2) will be well-posed.

Before we give the proof of the theorem we will establish an auxiliary result.

LEMMA 3.1. *Assume that the conditions given in Theorem 3.2 hold and define $\mathbf{w} = (w_1, w_2)^t = \mathbf{D}F(\mathbf{u})$. Then,*

$$\lambda \mathbf{D}^2 F(\mathbf{u}) \mathbf{w}^\perp \cdot \mathbf{w}^\perp = -\frac{w_1^2 |\nabla w_2|^2 + w_2^2 |\nabla w_1|^2 - 2w_1 w_2 \nabla w_1 \cdot \nabla w_2}{|\mathbf{w}|^2}.$$

Proof. It follows from (2.10) that the multiplier λ can be expressed as $\lambda = -\nabla \mathbf{u} : \nabla \mathbf{w} / |\mathbf{w}|^2$. Hence,

$$(3.8) \quad \lambda \mathbf{D}^2 F(\mathbf{u}) \mathbf{w}^\perp \cdot \mathbf{w}^\perp = \frac{\nabla \mathbf{u} : \nabla \mathbf{w}}{|\mathbf{w}|^2} (F_{11} w_2^2 + F_{22} w_1^2 - 2F_{12} w_1 w_2),$$

where $F_{ij} = \partial^2 F / \partial u_i \partial u_j$. Furthermore, since $\nabla F(\mathbf{u}) \equiv 0$, we have from (2.3) that

$$w_1 \nabla \mathbf{u}_1 + w_2 \nabla \mathbf{u}_2 = 0,$$

while (2.4) implies that

$$\nabla w_i = F_{i1} \nabla u_1 + F_{i2} \nabla u_2.$$

By combining these identities, we obtain

$$\begin{aligned} & (F_{11} w_2^2 + F_{22} w_1^2 - 2F_{12} w_1 w_2) \nabla u_1 \cdot \nabla w_1 \\ &= w_2^2 (F_{11} \nabla u_1 + F_{12} \nabla u_2) \cdot \nabla w_1 - w_1 w_2 (F_{22} \nabla u_2 + F_{12} \nabla u_1) \cdot \nabla w_1 \\ &= w_2^2 |\nabla w_1|^2 - w_1 w_2 \nabla w_1 \cdot \nabla w_2. \end{aligned}$$

A similar argument shows that

$$(F_{11} w_2^2 + F_{22} w_1^2 - 2F_{12} w_1 w_2) \nabla u_2 \cdot \nabla w_2 = w_1^2 |\nabla w_2|^2 - w_1 w_2 \nabla w_1 \cdot \nabla w_2,$$

and hence the desired identity follows from (3.8). \square

Proof of Theorem 3.2. As above we let $\mathbf{w} = \mathbf{D}F(\mathbf{u})$. For any $\mathbf{v} \in \mathbf{Z}_{\mathbf{u}}$, there exists an α such that $\mathbf{v} = \alpha \mathbf{w}^\perp$. In fact, we have

$$(3.9) \quad \alpha = \frac{\mathbf{v} \cdot \mathbf{w}^\perp}{|\mathbf{w}|^2}.$$

From the estimates (2.5)–(2.6) and condition (3.1), we see that $\alpha \in H_0^1(\Omega)$. The key identity we will use is the pointwise relation

$$(3.10) \quad |\nabla \mathbf{v}|^2 + \lambda \mathbf{D}^2 F(\mathbf{u}) \mathbf{v} \cdot \mathbf{v} = |\nabla(\alpha |\mathbf{w}|)|^2.$$

In order to verify this identity note that

$$\nabla(\alpha |\mathbf{w}|) = |\mathbf{w}| \nabla \alpha + \frac{\alpha}{|\mathbf{w}|} (w_1 \nabla w_1 + w_2 \nabla w_2).$$

Hence,

$$\begin{aligned} |\nabla(\alpha |\mathbf{w}|)|^2 &= |\mathbf{w}|^2 |\nabla \alpha|^2 + \frac{|\alpha|^2}{|\mathbf{w}|^2} |w_1 \nabla w_1 + w_2 \nabla w_2|^2 \\ &\quad + 2\alpha (w_1 \nabla \alpha \cdot \nabla w_1 + w_2 \nabla \alpha \cdot \nabla w_2). \end{aligned}$$

On the other hand,

$$|\nabla \mathbf{v}|^2 = |\mathbf{w}|^2 |\nabla \alpha|^2 + \alpha^2 |\nabla \mathbf{w}|^2 + 2\alpha(w_1 \nabla \alpha \cdot \nabla w_1 + w_2 \nabla \alpha \cdot \nabla w_2).$$

Therefore,

$$\begin{aligned} |\nabla \mathbf{v}|^2 - |\nabla(\alpha|\mathbf{w}|)|^2 &= \alpha^2 \left(|\nabla \mathbf{w}|^2 - \frac{|w_1 \nabla w_1 + w_2 \nabla w_2|^2}{|\mathbf{w}|^2} \right) \\ &= \frac{\alpha^2}{|\mathbf{w}|^2} (w_1^2 |\nabla w_2|^2 + w_2^2 |\nabla w_1|^2 - 2w_1 w_2 \nabla w_1 \cdot \nabla w_2) \\ &= -\lambda \mathbf{D}^2 F(\mathbf{u}) \mathbf{v} \cdot \mathbf{v}, \end{aligned}$$

where the last identity follows from Lemma 3.1. Hence, we have verified (3.10).

Let $\mu = \alpha|\mathbf{w}|$. Then $\mathbf{v} = \frac{\mu}{|\mathbf{w}|} \mathbf{w}^\perp$, and hence

$$\nabla \mathbf{v} = \frac{1}{|\mathbf{w}|} \mathbf{w}^\perp \cdot \nabla \mu + \mu \nabla \left(\frac{\mathbf{w}^\perp}{|\mathbf{w}|} \right).$$

Therefore, since \mathbf{u} satisfies (3.1), Poincaré’s inequality implies that

$$\|\nabla \mathbf{v}\|_0 \leq c(\|\nabla \mu\|_0 + \|\mu\|_0) \leq c\|\nabla(\alpha|\mathbf{w}|)\|_0,$$

where the constant c depends on \mathbf{u} . Together with (3.10) this implies the desired inequality of the theorem. \square

4. A stable discretization. The purpose of this section is to analyze a finite element discretization of the constrained minimization problem (1.1). Due to some technical difficulties caused by the use of inverse inequalities to treat some nonlinear terms, cf. (4.3) below, the analysis given here is restricted to the case $d = 2$. As a consequence, the bilinear form $a(\mathbf{u}, \lambda; \cdot, \cdot)$ will satisfy the coercivity bound given in Theorem 3.2.

So, for the rest of the paper, we assume that $d = 2$ and that $\Omega \subset \mathbb{R}^2$ is a polygonal domain. Given a shape regular and quasi-uniform family of triangulation $\{\mathcal{T}_h\}$ of Ω , with a mesh size $h < 1$, let \mathcal{N}_h denote the set of nodes associated with \mathcal{T}_h . We use V_h to denote the space of continuous piecewise linear functions and $V_{h,0} = V_h \cap H_0^1(\Omega)$. The notations \mathbf{V}_h and $\mathbf{V}_{h,0}$ will be used for the vector version of the corresponding spaces. We will use π_h to denote the usual nodal interpolation operators onto the spaces V_h and \mathbf{V}_h . Standard approximation properties of spaces of piecewise linear functions will be used below. In particular, we will use the estimates

$$(4.1) \quad \|(I - \pi_h)v\|_1 \leq Ch\|v\|_2, \quad v \in H^2(\Omega),$$

and

$$(4.2) \quad \|(I - P_h)v\|_{-1} \leq Ch\|v\|_0, \quad v \in L^2(\Omega).$$

Here, $P_h : L^2(\Omega) \rightarrow V_{h,0}$ is the L^2 projection. Due to the quasi-uniformity of the mesh, the operator P_h can be extended to a uniformly bounded operator on H^{-1} . Moreover, the following inverse inequalities hold:

$$(4.3) \quad \|v\|_\infty \leq C \log(h^{-1})\|v\|_1, \quad \|v\|_1 \leq Ch^{-1}\|v\|_0, \quad v \in V_h.$$

Set $\mathbf{g}_h = \pi_h \mathbf{g}$ (on $\partial\Omega$). We define

$$\mathbf{V}_{h,\mathbf{g}} = \{\mathbf{v} \in \mathbf{V}_h : \mathbf{v}|_{\partial\Omega} = \mathbf{g}_h\}.$$

We will consider the following discretized minimization problem:

$$(4.4) \quad \min_{\mathbf{v} \in \mathbf{V}_{h,\mathbf{g}}} \mathcal{E}(\mathbf{v}) \text{ subject to } F(\mathbf{v}) = 0 \text{ on } \mathcal{N}_h.$$

The Lagrange functional $L : \mathbf{V}_{h,\mathbf{g}} \times V_{h,0} \mapsto \mathbb{R}$ is

$$(4.5) \quad L(\mathbf{v}, \mu) = \mathcal{E}(\mathbf{v}) + \int_{\Omega} \mu \pi_h F(\mathbf{v}) d\mathbf{x} \quad (\mathbf{v}, \mu) \in \mathbf{V}_{h,\mathbf{g}} \times V_{h,0}.$$

The first order condition defining the critical points of L leads to the following discrete counterpart of the nonlinear saddle point problem (2.9):

Find $(\mathbf{u}_h, \lambda_h) \in \mathbf{V}_{h,\mathbf{g}} \times V_{h,0}$ such that

$$(4.6) \quad \begin{aligned} \langle \nabla \mathbf{u}_h, \nabla \mathbf{v} \rangle + \langle \pi_h [\mathbf{D}F(\mathbf{u}_h) \cdot \mathbf{v}], \lambda_h \rangle &= 0, & \mathbf{v} \in \mathbf{V}_{h,0}, \\ \langle \pi_h F(\mathbf{u}_h), \mu \rangle &= 0, & \mu \in V_{h,0}. \end{aligned}$$

However, we shall first analyze the discrete counterpart of the linearized system (3.2). For a given $(\hat{\mathbf{u}}, \hat{\lambda}) \in \mathbf{V}_{h,\mathbf{g}} \times V_{h,0}$, let us define the bilinear form $a_h(\hat{\mathbf{u}}, \hat{\lambda}; \cdot, \cdot)$ to be

$$a_h(\hat{\mathbf{u}}, \hat{\lambda}; \mathbf{v}, \hat{\mathbf{v}}) = \langle \nabla \mathbf{v}, \nabla \hat{\mathbf{v}} \rangle + \langle \pi_h [\mathbf{D}^2 F(\hat{\mathbf{u}}) \mathbf{v} \cdot \hat{\mathbf{v}}], \hat{\lambda} \rangle.$$

Similarly, as in (3.2) for the continuous problem, the linearized problem for (4.6) is to find $(\mathbf{v}, \mu) \in \mathbf{V}_{h,0} \times V_{h,0}$ such that

$$(4.7) \quad \begin{aligned} a_h(\hat{\mathbf{u}}, \hat{\lambda}; \mathbf{v}, \hat{\mathbf{v}}) + \langle \pi_h [\mathbf{D}F(\hat{\mathbf{u}}) \cdot \hat{\mathbf{v}}], \mu \rangle &= \langle \mathbf{f}, \hat{\mathbf{v}} \rangle, & \hat{\mathbf{v}} \in \mathbf{V}_{h,0} \\ \langle \pi_h [\mathbf{D}F(\hat{\mathbf{u}}) \cdot \mathbf{v}], \hat{\mu} \rangle &= \langle \sigma, \hat{\mu} \rangle, & \hat{\mu} \in V_{h,0}. \end{aligned}$$

For a given $\hat{\mathbf{u}} \in \mathbf{V}_{h,\mathbf{g}}$, define

$$Z_{h,\hat{\mathbf{u}}} = \{\mathbf{v} \in \mathbf{V}_{h,0} : \mathbf{D}F(\hat{\mathbf{u}}) \cdot \mathbf{v} = 0 \text{ on } \mathcal{N}_h\}.$$

LEMMA 4.1. *Let $\Phi : \mathbb{R}^2 \times \mathbb{R}^2 \times \dots \times \mathbb{R}^2 \mapsto \mathbb{R}^2$ be a smooth function. Then we have the following estimates for all $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k \in \mathbf{V}_h$:*

$$(4.8) \quad |\pi_h \Phi(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k)|_1 \leq C \sum_{i=1}^k \|\mathbf{D}_{\mathbf{v}_i} \Phi\|_{0,\infty} |\mathbf{v}_i|_1;$$

$$(4.9) \quad \|(\pi_h - I)\Phi(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k)\|_0 \leq Ch \sum_{i=1}^k \|\mathbf{D}_{\mathbf{v}_i} \Phi\|_{0,\infty} |\mathbf{v}_i|_1.$$

Above, the constant C is independent of h , Φ , and \mathbf{v}_i . The norm $\|\mathbf{D}_{\mathbf{v}_i} \Phi\|_{0,\infty}$ stands for $\|\mathbf{D}_{\mathbf{v}_i} \Phi(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k)\|_{0,\infty}$, with $\mathbf{D}_{\mathbf{v}_i} \Phi(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k) = \partial \Phi(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k) / \partial \mathbf{v}_i$.

Proof. For clarity, we shall only give the proof for $k = 2$. The extension of the proof for general cases is straightforward.

For an element $e \in \mathcal{T}_h$, let $p_i, i = 1, 2, 3$ be the vertices of e . Under the condition that the finite element mesh \mathcal{T}_h is regular and quasi-uniform, we have the following equivalent H^1 norms for $\mathbf{v} \in \mathbf{V}_h$:

$$(4.10) \quad |\mathbf{v}|_{1,e} \cong \sum_{i,j=1}^3 |\mathbf{v}(p_i) - \mathbf{v}(p_j)|^2, \quad \mathbf{v} \in \mathbf{V}_h, e \in \mathcal{T}_h.$$

In particular,

$$|\pi_h \Phi(\mathbf{v}_1, \mathbf{v}_2)|_{1,e}^2 \leq \sum_{i,j=1}^3 |\Phi(\mathbf{v}_1(p_i), \mathbf{v}_2(p_i)) - \Phi(\mathbf{v}_1(p_j), \mathbf{v}_2(p_j))|^2.$$

Thus, we get (4.8) from the following estimate:

$$\begin{aligned} |\pi_h \Phi(\mathbf{v}_1, \mathbf{v}_2)|_{1,e}^2 &\leq 2 \sum_{i,j=1}^3 \left(|\Phi(\mathbf{v}_1(p_i), \mathbf{v}_2(p_i)) - \Phi(\mathbf{v}_1(p_j), \mathbf{v}_2(p_i))|^2 \right. \\ &\quad \left. + |\Phi(\mathbf{v}_1(p_j), \mathbf{v}_2(p_i)) - \Phi(\mathbf{v}_1(p_j), \mathbf{v}_2(p_j))|^2 \right) \\ &\leq 2 \sum_{i,j=1}^3 \left(\|\mathbf{D}_{\mathbf{v}_1} \Phi\|_{0,\infty,e}^2 |\mathbf{v}_1(p_i) - \mathbf{v}_1(p_j)|^2 + \|\mathbf{D}_{\mathbf{v}_2} \Phi\|_{0,\infty,e}^2 |\mathbf{v}_2(p_i) - \mathbf{v}_2(p_j)|^2 \right). \end{aligned}$$

Next, we estimate (4.9). By the definition of the interpolation operator π_h , we have

$$(\pi_h - I)\Phi(\mathbf{v}_1, \mathbf{v}_2)(p) = \sum_{i=1}^3 [\Phi(\mathbf{v}_1(p_i), \mathbf{v}_2(p_i)) - \Phi(\mathbf{v}_1(p), \mathbf{v}_2(p))] \chi_i(p) \quad p \in e,$$

where $\{\chi_i\}_{i=1}^3$ are the barycentric coordinates on e . From this, we see that

$$\begin{aligned} \|(\pi_h - I)\Phi(\mathbf{v}_1, \mathbf{v}_2)\|_{0,e}^2 &\leq C \sum_{i=1}^3 \int_e |(\Phi(\mathbf{v}_1(p_i), \mathbf{v}_2(p_i)) - \Phi(\mathbf{v}_1, \mathbf{v}_2)) \chi_i|^2 \\ (4.11) \quad &\leq C \sum_{i,j=1}^3 \int_e \left(\|\mathbf{D}_{\mathbf{v}_1} \Phi\|_{0,\infty,e}^2 |\mathbf{v}_1(p_i) - \mathbf{v}_1(p_j)|^2 + \|\mathbf{D}_{\mathbf{v}_2} \Phi\|_{0,\infty,e}^2 |\mathbf{v}_2(p_i) - \mathbf{v}_2(p_j)|^2 \right) \\ &\leq Ch^2 \sum_{i,j=1}^3 \left(\|\mathbf{D}_{\mathbf{v}_1} \Phi\|_{0,\infty,e}^2 |\mathbf{v}_1|_{1,e}^2 + \|\mathbf{D}_{\mathbf{v}_2} \Phi\|_{0,\infty,e}^2 |\mathbf{v}_2|_{1,e}^2 \right). \end{aligned}$$

Thus, the estimate (4.9) is verified. \square

For the lemma above, it is essential that the functions \mathbf{v}_i are finite element functions. If $\mathbf{v}_1 \in \mathbf{W}^{1,\infty}(\Omega)$ and $\mathbf{v}_2 \in \mathbf{V}_h$, then we obtain

$$(4.12) \quad \|(\pi_h - I)\Phi(\mathbf{v}_1, \mathbf{v}_2)\|_0 \leq Ch(\|\mathbf{D}_{\mathbf{v}_1} \Phi\|_{0,\infty} |\mathbf{v}_1|_{1,\infty} + \|\mathbf{D}_{\mathbf{v}_2} \Phi\|_{0,\infty} |\mathbf{v}_2|_1).$$

The next result, which is essential for our analysis, is a discrete version of Theorem 3.2. As in the previous section, (\mathbf{u}, λ) is a solution of (2.9) satisfying (3.1).

THEOREM 4.1. *There exists positive constants γ_0 and h_0 such that, for $(\hat{\mathbf{u}}, \hat{\lambda}) \in \mathbf{V}_{h,\mathbf{g}} \times V_{h,0}$ satisfying*

$$(4.13) \quad \|\hat{\mathbf{u}} - \pi_h \mathbf{u}\|_1 + \|\hat{\lambda} - P_h \lambda\|_{-1} \leq \gamma / \log^2(h^{-1})$$

with $h \leq h_0$ and $\gamma \leq \gamma_0$, we have

$$(4.14) \quad a_h(\hat{\mathbf{u}}, \hat{\lambda}; \mathbf{v}, \mathbf{v}) \geq \beta_3 \|\mathbf{v}\|_1^2, \quad \mathbf{v} \in Z_{h,\hat{\mathbf{u}}}.$$

Here the constants γ_0, h_0, β_3 depend on \mathbf{u} .

In order to prove the above theorem, we need to derive some auxiliary results. The main idea is to relate (4.14) to the continuous problem, and then use Theorem 3.2 and some approximate properties of the operators π_h and P_h . As before, we shall use $\mathbf{w} = \mathbf{D}F(\mathbf{u})$, with \mathbf{u} being the true solution; see (3.1). Given a $(\hat{\mathbf{u}}, \hat{\lambda})$ satisfying (4.13), we define $\hat{\mathbf{w}} = \mathbf{D}F(\hat{\mathbf{u}})$. For any $\mathbf{v} \in \mathbf{Z}_{h,\hat{\mathbf{u}}}$, let us define

$$(4.15) \quad \alpha(p_i) = \frac{\mathbf{v}(p_i) \cdot \hat{\mathbf{w}}^\perp(p_i)}{|\hat{\mathbf{w}}(p_i)|^2}, \quad p_i \in \mathcal{N}_h.$$

From the above definition, it is clear that

$$\alpha = \pi_h \left(\frac{\mathbf{v} \cdot \hat{\mathbf{w}}^\perp}{|\hat{\mathbf{w}}|^2} \right) \in V_{h,0}, \quad \mathbf{v} = \pi_h(\alpha \hat{\mathbf{w}}^\perp).$$

We have used the relation $\hat{\mathbf{w}} \cdot \mathbf{v} = 0$ on \mathcal{N}_h in getting the last equality. Corresponding to the true solution \mathbf{u} and a given $\hat{\mathbf{u}} \in \mathbf{Z}_{h,\hat{\mathbf{u}}}$, let $\varepsilon_h \in \mathbf{H}_0^1(\Omega)$ be the function given by $\varepsilon_h = \alpha \mathbf{w}^\perp - \mathbf{v}$. We see clearly that

$$(4.16) \quad \varepsilon_h + \mathbf{v} \in \mathbf{Z}_{\mathbf{u}}.$$

For a given $\hat{\mathbf{u}}$ satisfying (4.13), one can verify by assumption (i) on the constraint function F , cf. (2.1), and the inverse estimate (4.3) that

$$|\mathbf{w}(p) - \hat{\mathbf{w}}(p)| = |\mathbf{D}F(\hat{\mathbf{u}}(p)) - \mathbf{D}F(\pi_h \mathbf{u}(p))| \leq c_1 \gamma, \quad p \in \mathcal{N}_h.$$

Thus, by choosing γ small enough, one can guarantee that

$$(4.17) \quad 0 < c|\mathbf{w}(p)| \leq |\hat{\mathbf{w}}(p)| \leq C|\mathbf{w}(p)|, \quad p \in \mathcal{N}_h.$$

Hence, we conclude that (4.13) implies that there is a constant C , depending only on u , such that

$$(4.18) \quad \|\hat{\mathbf{u}}\|_1, \|\hat{\mathbf{u}}\|_{0,\infty} \leq C.$$

LEMMA 4.2. *Let $(\hat{\mathbf{u}}, \hat{\lambda}) \in \mathbf{V}_{h,\mathbf{g}} \times V_{h,0}$ satisfy (4.13). Then we have the estimate*

$$\left| \pi_h \left(\varphi \frac{\hat{\mathbf{w}}}{|\hat{\mathbf{w}}|^2} \right) \right|_1 \leq C|\varphi|_1, \quad \varphi \in V_{h,0},$$

where the constant C depends on \mathbf{u} .

Proof. Let $\psi = \pi_h(\varphi \frac{\hat{\mathbf{w}}}{|\hat{\mathbf{w}}|^2})$. Using (4.10), we see that

$$(4.19) \quad \begin{aligned} |\psi|_{1,e}^2 &\leq C \sum_{i,j} \left| \varphi(p_i) \frac{\hat{\mathbf{w}}(p_i)}{|\hat{\mathbf{w}}(p_i)|^2} - \varphi(p_j) \frac{\hat{\mathbf{w}}(p_j)}{|\hat{\mathbf{w}}(p_j)|^2} \right|^2 \\ &\leq C \sum_{i,j} \left[\frac{|\varphi(p_i) - \varphi(p_j)|^2}{|\hat{\mathbf{w}}(p_i)|^2} + |\varphi(p_j)|^2 \cdot \left| \frac{\hat{\mathbf{w}}(p_i)}{|\hat{\mathbf{w}}(p_i)|^2} - \frac{\hat{\mathbf{w}}(p_j)}{|\hat{\mathbf{w}}(p_j)|^2} \right|^2 \right]. \end{aligned}$$

It follows from (4.10) and (4.17)–(4.18) that

$$(4.20) \quad \sum_{i,j} \frac{|\varphi(p_i) - \varphi(p_j)|^2}{|\hat{\mathbf{w}}(p_i)|^2} \leq C|\varphi|_{1,e}^2.$$

On the other hand, we have by (4.17)–(4.18) and assumption (iii) on the constraint function F , cf. (2.2),

$$\begin{aligned} \left| \frac{\hat{\mathbf{w}}(p_i)}{|\hat{\mathbf{w}}(p_i)|^2} - \frac{\hat{\mathbf{w}}(p_j)}{|\hat{\mathbf{w}}(p_j)|^2} \right|^2 &\leq C|\hat{\mathbf{w}}(p_i) - \hat{\mathbf{w}}(p_j)|^2 \leq C|\hat{\mathbf{u}}(p_i) - \hat{\mathbf{u}}(p_j)|^2 \\ &\leq C(|\hat{\mathbf{u}} - \pi_h \mathbf{u}(p_i) - (\hat{\mathbf{u}} - \pi_h \mathbf{u}(p_j))|^2 + |\pi_h \mathbf{u}(p_i) - \pi_h \mathbf{u}(p_j)|^2). \end{aligned}$$

Thus, we get by the inverse estimate (4.3) and (4.13) that

$$\begin{aligned} (4.21) \quad &\sum_{i,j} \left[|\varphi(p_j)|^2 \cdot \left| \frac{\hat{\mathbf{w}}(p_i)}{|\hat{\mathbf{w}}(p_i)|^2} - \frac{\hat{\mathbf{w}}(p_j)}{|\hat{\mathbf{w}}(p_j)|^2} \right|^2 \right] \\ &\leq C\|\varphi\|_{0,\infty,e}^2 \cdot \|\hat{\mathbf{u}} - \pi_h \mathbf{u}\|_{1,e}^2 + \|\varphi\|_{0,e}^2 \cdot \|\pi_h \mathbf{u}\|_{1,\infty,e}^2 \\ &\leq C(\gamma^2 + \|\mathbf{u}\|_{1,\infty,e}^2)\|\varphi\|_{1,e}^2. \end{aligned}$$

Substituting (4.20)–(4.21) into (4.19), we obtain the desired bound. \square

Remark 4.1. If we apply Lemma 4.1 on the function ψ defined by $\psi = \pi_h(\varphi \frac{\hat{\mathbf{w}}}{|\hat{\mathbf{w}}|^2})$, we will get that

$$|\psi|_1 \leq C \log(h^{-1})|\varphi|_1.$$

The result we are getting here is better. We have removed the factor $\log(h^{-1})$.

LEMMA 4.3. *Let $(\hat{\mathbf{u}}, \hat{\lambda}) \in \mathbf{V}_{h,\mathbf{g}} \times V_{h,0}$ satisfy (4.13). Then, there exist positive constants h_0 and γ_0 , depending on \mathbf{u} , such that*

$$a(\mathbf{u}, \lambda; \mathbf{v}, \mathbf{v}) \geq \frac{\beta_2}{2}\|\mathbf{v}\|_1^2, \quad \mathbf{v} \in \mathbf{Z}_{h,\hat{\mathbf{u}}}$$

for $0 < h \leq h_0$ and $0 < \gamma \leq \gamma_0$.

Proof. For any $\mathbf{v} \in \mathbf{Z}_{h,\hat{\mathbf{u}}}$, let α and ε_h be defined as in (4.15) and (4.16). From $\pi_h(\alpha\pi_h \mathbf{w}^\perp) = \pi_h(\alpha \mathbf{w}^\perp)$, we have

$$(4.22) \quad \varepsilon_h = (I - \pi_h)(\alpha \mathbf{w}^\perp) + \pi_h[\alpha\pi_h(\mathbf{w} - \hat{\mathbf{w}})^\perp].$$

From (4.12) and also using the inverse inequality (4.3), we get that

$$\begin{aligned} (4.23) \quad &|(I - \pi_h)(\alpha \mathbf{w}^\perp)|_1^2 \leq Ch^2(\|\mathbf{w}^\perp\|_{0,\infty}^2|\alpha|_1^2 + \|\alpha\|_{0,\infty}^2\|\mathbf{w}^\perp\|_{1,\infty}^2) \\ &\leq Ch^2 \log^2(h^{-1})\|\mathbf{u}\|_{1,\infty}^2|\alpha|_1^2. \end{aligned}$$

Note that there exists a ξ such that

$$\pi_h[\alpha\pi_h(\mathbf{w} - \hat{\mathbf{w}})^\perp] = \pi_h\left[\alpha\pi_h\left(\pi_h \mathbf{D}^2 F(\xi)(\pi_h \mathbf{u} - \hat{\mathbf{u}})\right)^\perp\right].$$

A repeated application of (4.8) and (4.3) gives

$$(4.24) \quad |\pi_h[\alpha\pi_h(\mathbf{w} - \hat{\mathbf{w}})^\perp]|_1^2 \leq C \log^4(h^{-1})|\alpha|_1^2|\pi_h \mathbf{u} - \hat{\mathbf{u}}|_1^2.$$

From Lemma 4.2, we see that

$$(4.25) \quad |\alpha|_1 \leq C|\mathbf{v}|_1.$$

Combining (4.23)–(4.25) with (4.13), we see that

$$(4.26) \quad |\varepsilon_h|_1^2 \leq C(h^2 \log^2(h^{-1})\|\mathbf{u}\|_{1,\infty}^2 + \gamma^2)|\alpha|_1^2 \leq C(h^2 \log^2(h^{-1})\|\mathbf{u}\|_{1,\infty}^2 + \gamma^2)|\mathbf{v}|_1^2.$$

The following estimate follows from (3.5) and (3.7):

$$(4.27) \quad \begin{aligned} a(\mathbf{u}, \lambda; \mathbf{v}, \mathbf{v}) &= a(\mathbf{u}, \lambda; \mathbf{v} + \varepsilon_h, \mathbf{v} + \varepsilon_h) - a(\mathbf{u}, \lambda; \mathbf{v}, \varepsilon_h) + a(\mathbf{u}, \lambda; \varepsilon_h, \varepsilon_h) \\ &\geq C(\beta_2 \|\mathbf{v} + \varepsilon_h\|_1^2 - |\mathbf{v}|_1 |\varepsilon_h|_1 - |\varepsilon_h|_1^2). \end{aligned}$$

Choosing h and γ small enough, we obtain the desired result from (4.26) and (4.27). \square

Proof of Theorem 4.1. In the proof, we always assume that h and γ are small. Note that

$$(4.28) \quad \begin{aligned} a_h(\hat{\mathbf{u}}, \hat{\lambda}; \mathbf{v}, \mathbf{v}) - a(\mathbf{u}, \lambda; \mathbf{v}, \mathbf{v}) &= \langle \pi_h[\mathbf{D}^2 F(\hat{\mathbf{u}})\mathbf{v} \cdot \mathbf{v}], \hat{\lambda} \rangle - \langle \mathbf{D}^2 F(\mathbf{u})\mathbf{v} \cdot \mathbf{v}, \lambda \rangle \\ &= \langle \pi_h[\mathbf{D}^2 F(\hat{\mathbf{u}})\mathbf{v} \cdot \mathbf{v}], \hat{\lambda} - \lambda \rangle + \langle (\pi_h - I)[\mathbf{D}^2 F(\hat{\mathbf{u}})\mathbf{v} \cdot \mathbf{v}], \lambda \rangle \\ &\quad + \langle (\mathbf{D}^2 F(\hat{\mathbf{u}}) - \mathbf{D}^2 F(\mathbf{u}))\mathbf{v} \cdot \mathbf{v}, \lambda \rangle = I_1 + I_2 + I_3. \end{aligned}$$

The meaning of I_i is self-explainable. Since $\lambda \in L^2(\Omega)$, we obtain from (4.13) that

$$\begin{aligned} \|\hat{\lambda}_h - \lambda\|_{-1} &\leq \|\hat{\lambda}_h - P_h \lambda\|_{-1} + \|P_h \lambda - \lambda\|_{-1} \\ &\leq \gamma / \log^2(h^{-1}) + Ch \|\lambda\|_0. \end{aligned}$$

Using Lemma 4.1, we see that

$$\begin{aligned} |\pi_h[\mathbf{D}^2 F(\hat{\mathbf{u}})\mathbf{v} \cdot \mathbf{v}]|_1 &\leq C(\|\mathbf{D}^2 F(\hat{\mathbf{u}}) \cdot \mathbf{v}\|_{0,\infty} |\mathbf{v}|_1 + \|\mathbf{v}\|_{0,\infty}^2 \|\mathbf{D}^3 F(\hat{\mathbf{u}})\|_{0,\infty} |\hat{\mathbf{u}}|_1) \\ &\leq C \log^2(h^{-1}) \|\mathbf{v}\|_1^2. \end{aligned}$$

For a small h , a combination of the above two inequalities leads to

$$|I_1| = |(\pi_h[\mathbf{D}^2 F(\hat{\mathbf{u}})\mathbf{v} \cdot \mathbf{v}], \hat{\lambda}_h - \lambda)| \leq C \log^2(h^{-1}) \|\mathbf{v}\|_1^2 (\gamma / \log^2(h^{-1}) + Ch \|\lambda\|_0) \leq C\gamma \|\mathbf{v}\|_1^2.$$

Similarly, we use Lemma 4.1 to prove that

$$\begin{aligned} |I_2| &= |((\pi_h - I)[\mathbf{D}^2 F(\hat{\mathbf{u}})\mathbf{v} \cdot \mathbf{v}], \lambda)| \\ &\leq \|(\pi_h - I)[\mathbf{D}^2 F(\hat{\mathbf{u}})\mathbf{v} \cdot \mathbf{v}]\|_0 \cdot \|\lambda\|_0 \leq Ch \log^2(h^{-1}) \|\mathbf{v}\|_1^2 \end{aligned}$$

and

$$\begin{aligned} |I_3| &= |((\mathbf{D}^2 F(\hat{\mathbf{u}}) - \mathbf{D}^2 F(\mathbf{u}))\mathbf{v} \cdot \mathbf{v}, \lambda)| \\ &\leq \|(\mathbf{D}^2 F(\hat{\mathbf{u}}) - \mathbf{D}^2 F(\mathbf{u}))\mathbf{v} \cdot \mathbf{v}\|_0 \cdot \|\lambda\|_0 \leq C\gamma \|\mathbf{v}\|_1^2. \end{aligned}$$

Choosing h and γ small enough, we obtain the desired result from Lemma 4.3 and the estimates above of the three terms appearing in (4.28). \square

THEOREM 4.2. *Assume that $(\hat{\mathbf{u}}, \hat{\lambda}) \in \mathbf{V}_{h,\mathbf{g}} \times V_{h,0}$ satisfies the condition (4.13). There exists a constant $\beta_4 > 0$, which depends on \mathbf{u} , such that*

$$(4.29) \quad \inf_{\mu \in V_{h,0}} \sup_{\mathbf{v} \in \mathbf{V}_{h,0}} \frac{\langle \pi_h[\mathbf{D}F(\hat{\mathbf{u}}) \cdot \mathbf{v}], \mu \rangle}{\|\mu\|_{-1} \|\mathbf{v}\|_1} \geq \beta_4.$$

Proof. For the φ given in (3.4), let $\varphi_h = P_h \varphi$. Then, we see that

$$\frac{\langle \mu_h, \varphi_h \rangle}{\|\varphi_h\|_1} \geq \beta_1 \|\mu_h\|_{-1}.$$

Define $\mathbf{v}_h = \pi_h[\varphi_h \frac{\mathbf{D}F(\hat{\mathbf{u}})}{|\mathbf{D}F(\hat{\mathbf{u}})|^2}]$. Then,

$$\langle \pi_h[\mathbf{D}F(\hat{\mathbf{u}}) \cdot \mathbf{v}_h], \mu_h \rangle = \langle \mu_h, \varphi_h \rangle.$$

From Lemma 4.2, one gets that $|\mathbf{v}_h|_1 \leq C|\varphi_h|_1$. By collecting these estimates, the theorem is established. \square

Together with the Theorems 4.1 and 4.2, the saddle point theory given in [8] or [9] assures existence, stability, and uniqueness of the solution of the linearized saddle point system (4.7), as long as $(\hat{\mathbf{u}}, \hat{\lambda})$ satisfies (4.13). In the next section, we shall use these properties to prove some results for the corresponding nonlinear systems.

Remark 4.2. If we replace $V_{h,0}$ by V_h in (4.29), the inf-sup condition (4.29) may not be satisfied. This is why we use the $V_{h,0}$, instead of V_h , as finite element space for the Lagrange multiplier.

5. The discrete nonlinear problem. The main purpose of this section is to establish existence and uniqueness of solutions of the discretized nonlinear saddle point problem (4.6) in a neighborhood of a continuous solution (\mathbf{u}, λ) of the system (2.9). As above, we assume that (\mathbf{u}, λ) corresponds to a local minimum of the functional \mathcal{E} over $\mathbf{H}_{\mathbf{g}}^1(\Omega; \mathcal{M})$ and that the regularity assumption (3.1) holds. Furthermore, we will show that the discrete solutions converge to the continuous solution with a linear rate with respect to the mesh parameter h . However, we start by summarizing some properties of the linearized saddle point system.

For notational simplicity, we shall use X , X_h , and $X_{h,\mathbf{g}}$ defined by $X = \mathbf{H}_0^1(\Omega) \times H^{-1}(\Omega)$, $X_h = \mathbf{V}_{h,0} \times V_{h,0}$, and $X_{h,\mathbf{g}} = \mathbf{V}_{h,\mathbf{g}} \times V_{h,0}$. Let $\|\cdot\|_X$ denote the norm on the product space $\mathbf{H}_0^1(\Omega) \times H^{-1}(\Omega)$, and let $\|\cdot\|_{X^*}$ denote the norm on the dual space $X^* = \mathbf{H}^{-1}(\Omega) \times H_0^1(\Omega)$. The norm $\|\cdot\|_{L(X,X^*)}$ will be used to denote the norm of a bounded linear operator from X to X^* . The spaces X_h and $X_{h,\mathbf{g}}$ are equipped with the norm of X , while X_h^* is equal to X_h as a set, but equipped with the dual norm of X with respect to the L^2 inner products. Similarly, the norm $\|\cdot\|_{L(X_h,X_h^*)}$ is the associated operator norm.

Let $x = (\mathbf{u}, \lambda)$ be a solution of (2.9). Corresponding to x , let $G(x) \in X^*$ be given by

$$\langle G(x), y \rangle = \langle \nabla \mathbf{u}, \nabla \mathbf{v} \rangle + \langle \mathbf{D}F(\mathbf{u}) \cdot \mathbf{v}, \lambda \rangle + \langle F(\mathbf{u}), \mu \rangle, \quad y = (\mathbf{v}, \mu) \in X.$$

As usual, $\langle \cdot, \cdot \rangle$ is the duality pairing which extends the standard L^2 inner product. Associated with G , we define a mapping $G'(x) : X \rightarrow X^*$ by

$$(5.1) \quad \langle G'(x) \cdot y, \hat{y} \rangle = a(\mathbf{u}, \lambda; \mathbf{v}, \hat{\mathbf{v}}) + \langle \mathbf{D}F(\mathbf{u}) \cdot \hat{\mathbf{v}}, \mu \rangle + \langle \mathbf{D}F(\mathbf{u}) \cdot \mathbf{v}, \hat{\mu} \rangle$$

for all $y = (\mathbf{v}, \mu), \hat{y} = (\hat{\mathbf{v}}, \hat{\mu}) \in X = \mathbf{H}_0^1(\Omega) \times H^{-1}(\Omega)$. The operator $G'(x)$ is formally the Fréchet differential of G at x .

Recall from the saddle point theory given in [8, 9] that Theorems 3.2–3.1 imply that the system (3.2) has a unique solution (\mathbf{v}, μ) , which depends continuously on $(\mathbf{f}, \sigma) \in X^*$. Thus we have the following result.

THEOREM 5.1. *If (\mathbf{u}, λ) satisfies the regularity assumption (3.1), then the map $G'(x)$ defined by (5.1) is an isomorphism from $X = \mathbf{H}_0^1(\Omega) \times H^{-1}(\Omega)$ to $X^* = \mathbf{H}^{-1}(\Omega) \times H_0^1(\Omega)$.*

For the discretized saddle point problem, let $G_h : X_{h,\mathbf{g}} \rightarrow X_h^*$ be the map defined by (4.6). For any $\hat{x} = (\hat{\mathbf{u}}, \hat{\lambda}) \in X_{h,\mathbf{g}}$, $G_h(\hat{x})$ is the operator that satisfies

$$\langle G_h(\hat{x}), \hat{y} \rangle = \langle \nabla \hat{\mathbf{u}}, \nabla \hat{\mathbf{v}} \rangle + \langle \pi_h[\mathbf{D}F(\hat{\mathbf{u}}) \cdot \hat{\mathbf{v}}], \hat{\lambda} \rangle + \langle \pi_h F(\hat{\mathbf{u}}), \hat{\mu} \rangle, \quad \hat{y} = (\hat{\mathbf{v}}, \hat{\mu}) \in X_h.$$

Thus, problem (4.6) is, in fact, to find $x_h = (\mathbf{u}_h, \lambda_h) \in X_{h,\mathbf{g}}$ such that

$$(5.2) \quad \langle G_h(x_h), y \rangle = 0, \quad y = (\hat{\mathbf{v}}, \hat{\mu}) \in X_h.$$

Let $G'_h(\hat{x})$ be the Fréchet derivative of G_h at $\hat{x} = (\hat{\mathbf{u}}, \hat{\lambda}) \in X_{h,\mathbf{g}}$. Then, $G'_h(\hat{x}) : X_h \rightarrow X_h^*$ is the linear operator given by

$$(5.3) \quad \begin{aligned} \langle G'_h(\hat{x})y, \hat{y} \rangle &= a_h(\hat{\mathbf{u}}, \hat{\lambda}; \mathbf{v}, \hat{\mathbf{v}}) + \langle \pi_h[\mathbf{D}F(\hat{\mathbf{u}}) \cdot \hat{\mathbf{v}}], \mu \rangle + \langle \pi_h[\mathbf{D}F(\hat{\mathbf{u}}) \cdot \mathbf{v}], \hat{\mu} \rangle, \\ y &= (\mathbf{v}, \mu) \in X_h, \quad \hat{y} = (\hat{\mathbf{v}}, \hat{\mu}) \in X_h. \end{aligned}$$

By Theorems 4.1–4.2, the following result is a consequence of the theory given in [8, 9].

THEOREM 5.2. *Assume that $\hat{x} = (\hat{\mathbf{u}}, \hat{\lambda}) \in X_{h,\mathbf{g}}$ satisfies the condition (4.13). For sufficiently small h and γ , the map $G'_h(\hat{x})$ is an isomorphism from X_h to X_h^* . Moreover,*

$$(5.4) \quad \|G'_h(\hat{x})^{-1}\|_{L(X_h^*, X_h)} \leq M,$$

where M is a constant independent of h and $\hat{x} = (\hat{\mathbf{u}}, \hat{\lambda})$.

Define $x_* = (\pi_h \mathbf{u}, P_h \lambda)$, and set $y_* = G'_h(x_*)$. We can use similar techniques as for Theorems 4.1 to prove the following lemma.

LEMMA 5.1. *For any $\hat{x} = (\hat{\mathbf{u}}, \hat{\lambda}) \in X_{h,\mathbf{g}}$ satisfying (4.13), we have*

$$\|G'_h(\hat{x}) - G'_h(x_*)\|_{L(X_h, X_h^*)} \leq C \log(h^{-1}) \|\hat{x} - x_*\|_X,$$

where C depends on \mathbf{u} and λ .

Proof. By the definition of G'_h , we have, for any $y = (\mathbf{v}, \mu) \in X_h$ and $\hat{y} = (\hat{\mathbf{v}}, \hat{\mu}) \in X_h$,

$$(5.5) \quad \begin{aligned} \langle (G'_h(\hat{x}) - G'_h(x_*))y, \hat{y} \rangle &= \langle \pi_h[\mathbf{D}^2F(\hat{\mathbf{u}})\mathbf{v} \cdot \hat{\mathbf{v}}], \hat{\lambda} - P_h \lambda \rangle \\ &+ \langle \pi_h[(\mathbf{D}^2F(\hat{\mathbf{u}}) - \mathbf{D}^2F(\pi_h \mathbf{u}))\mathbf{v} \cdot \hat{\mathbf{v}}], P_h \lambda \rangle \\ &+ \langle \pi_h[(\mathbf{D}F(\hat{\mathbf{u}}) - \mathbf{D}F(\pi_h \mathbf{u})) \cdot \hat{\mathbf{v}}], \mu \rangle \\ &+ \langle \pi_h[(\mathbf{D}F(\hat{\mathbf{u}}) - \mathbf{D}F(\pi_h \mathbf{u})) \cdot \mathbf{v}], \hat{\mu} \rangle. \end{aligned}$$

It is clear that

$$(5.6) \quad \langle \pi_h[\mathbf{D}^2F(\hat{\mathbf{u}})\mathbf{v} \cdot \hat{\mathbf{v}}], \hat{\lambda} - P_h \lambda \rangle \leq \|\pi_h[\mathbf{D}^2F(\hat{\mathbf{u}})\mathbf{v} \cdot \hat{\mathbf{v}}]\|_1 \|\hat{\lambda} - P_h \lambda\|_{-1}.$$

As in the proof of Lemma 4.1, we deduce

$$\begin{aligned} \|\pi_h[\mathbf{D}^2F(\hat{\mathbf{u}})\mathbf{v} \cdot \hat{\mathbf{v}}]\|_1 &\leq C \|\mathbf{D}^2F(\hat{\mathbf{u}})\mathbf{v}\|_{0,\infty} \cdot \|\hat{\mathbf{v}}\|_1 \\ &+ C \|\mathbf{D}^2F(\hat{\mathbf{u}})\|_{0,\infty} \cdot \|\mathbf{v}\|_1 \cdot \|\hat{\mathbf{v}}\|_{0,\infty} \\ &+ C \|\mathbf{D}^2F(\hat{\mathbf{u}})\|_{0,\infty} \cdot \|\mathbf{v}\|_{0,\infty} \cdot \|\hat{\mathbf{v}}\|_{0,\infty}. \end{aligned}$$

Then, we further get by the inverse inequality (4.3)

$$\|\pi_h[\mathbf{D}^2F(\hat{\mathbf{u}})\mathbf{v} \cdot \hat{\mathbf{v}}]\|_1 \leq C \log^3(h^{-1}) \|\mathbf{v}\|_1 \cdot \|\hat{\mathbf{v}}\|_1.$$

Plugging this in (5.6), together with (4.13), leads to

$$\langle \pi_h[\mathbf{D}^2F(\hat{\mathbf{u}})\mathbf{v} \cdot \hat{\mathbf{v}}], \hat{\lambda} - P_h \lambda \rangle \leq C \gamma \log(h^{-1}) \|\mathbf{v}\|_1 \|\hat{\mathbf{v}}\|_1.$$

Similarly, we deduce by (2.2), the inverse inequality (4.3), and (4.13)

$$\begin{aligned} & \|\pi_h[(\mathbf{D}^2 F(\hat{\mathbf{u}}) - \mathbf{D}^2 F(\pi_h \mathbf{u}))\mathbf{v} \cdot \hat{\mathbf{v}}]\|_1 \\ & \leq C\ell \log^3(h^{-1})\|\hat{\mathbf{u}} - \pi_h \mathbf{u}\|_1 \cdot \|\mathbf{v}\|_1 \cdot \|\hat{\mathbf{v}}\|_1 \\ & \leq C\ell\gamma \log(h^{-1})\|\mathbf{v}\|_1\|\hat{\mathbf{v}}\|_1. \end{aligned}$$

Estimating the last two terms in (5.5) by Lemma 4.1, (4.3), and (4.13), we obtain the result. The constants C in the estimates depend on \mathbf{u} and λ . \square

At this point, we need to recall the implicit function theorem as, for example, given in Lemma 1 of [10]. From the implicit function theorem, we can conclude that if there is a $\delta > 0$ such that

$$(5.7) \quad \hat{x} \in X_h, \|\hat{x} - x_*\|_X \leq \delta \text{ implies } \|G'_h(\hat{x}) - G'_h(x_*)\|_{L(X_h, X_h^*)} \leq \frac{1}{2M},$$

then the equation

$$(5.8) \quad G_h(\hat{x}) = \hat{y}$$

has a unique solution for all \hat{y} satisfying

$$\|\hat{y} - y_*\|_{X^*} \leq \frac{\delta}{2M}.$$

Here M is the positive constant appearing in Theorem 5.2. From Lemma 5.1, we see that the condition (5.7) is fulfilled if we choose $\delta = 1/(2MC \log(h^{-1}))$. Hence, we have that (5.8) has a unique solution \hat{x} satisfying

$$\|\hat{x} - x_*\|_X \leq \frac{1}{2MC \log(h^{-1})}$$

for all \hat{y} such that

$$\|\hat{y} - y_*\|_{X^*} \leq \frac{1}{4M^2C \log(h^{-1})}.$$

Furthermore, we can conclude from Lemma 1 of [10] that

$$(5.9) \quad \|\hat{x} - x_*\|_X \leq 2M\|\hat{y} - y_*\|_{X^*}.$$

Note that our desired equation is $G_h(x) = 0$. Thus, if we can verify that

$$(5.10) \quad \|G_h(x_*)\|_{X^*} = \|y_*\|_{X^*} \leq \frac{1}{4M^2C \log(h^{-1})},$$

we can conclude existence and uniqueness of the solution of this equation. If we assume more smoothness on \mathbf{u} , this is a consequence of the following lemma.

LEMMA 5.2. *Assume that $\mathbf{u} \in \mathbf{H}^2(\Omega) \cap \mathbf{W}^{1,\infty}(\Omega)$. Then we have*

$$\|G_h(x_*)\|_{X^*} \leq Ch, \text{ with } x_* = (\pi_h \mathbf{u}, P_h \lambda).$$

Proof. It suffices to prove that

$$(5.11) \quad |\langle G_h(x_*), \hat{x} \rangle| \leq Ch\|\hat{x}\|_X, \quad \hat{x} = (\mathbf{v}, \mu) \in X_h.$$

We have by (2.9) and the definition of G_h

$$(5.12) \quad \begin{aligned} \langle G_h(x_*) , \hat{x} \rangle &= \langle \nabla(\pi_h \mathbf{u} - \mathbf{u}), \nabla \mathbf{v} \rangle + \langle \pi_h F(\pi_h \mathbf{u}), \mu \rangle - \langle F(\mathbf{u}), \mu \rangle \\ &\quad + \langle \pi_h [\mathbf{D}F(\pi_h \mathbf{u}) \cdot \mathbf{v}], P_h \lambda \rangle - \langle \mathbf{D}F(\mathbf{u}) \cdot \mathbf{v}, \lambda \rangle. \end{aligned}$$

It is clear that

$$(5.13) \quad |\langle \nabla(\pi_h \mathbf{u} - \mathbf{u}), \nabla \mathbf{v} \rangle| \leq |\pi_h \mathbf{u} - \mathbf{u}|_1 \cdot |\mathbf{v}|_1 \leq Ch \|\mathbf{u}\|_2 \cdot |\mathbf{v}|_1.$$

Note that since $\pi_h F(\pi_h \mathbf{u}) = \pi_h F(\mathbf{u})$, we obtain from (4.1) that

$$(5.14) \quad \begin{aligned} |\langle \pi_h F(\pi_h \mathbf{u}), \mu \rangle - \langle F(\mathbf{u}), \mu \rangle| &= |\langle \pi_h - I \rangle F(\mathbf{u}), \mu| \\ &\leq \|(\pi_h - I)F(\mathbf{u})\|_1 \cdot \|\mu\|_{-1} \leq Ch \|F(\mathbf{u})\|_2 \cdot \|\mu\|_{-1}. \end{aligned}$$

Furthermore, by the assumptions on F and the estimates (4.1), (4.2), and (4.12), we get

$$(5.15) \quad \begin{aligned} &|\langle \pi_h [\mathbf{D}F(\pi_h \mathbf{u}) \cdot \mathbf{v}], P_h \lambda \rangle - \langle \mathbf{D}F(\mathbf{u}) \cdot \mathbf{v}, \lambda \rangle| \\ &\leq | \langle (\pi_h - I) [\mathbf{D}F(\mathbf{u}) \cdot \mathbf{v}], P_h \lambda \rangle | + | \langle \mathbf{D}F(\mathbf{u}) \cdot \mathbf{v}, P_h \lambda - \lambda \rangle | \\ &\leq \|(\pi_h - I) [\mathbf{D}F(\mathbf{u}) \cdot \mathbf{v}]\|_0 \cdot \|P_h \lambda\|_0 + \|\mathbf{D}F(\mathbf{u}) \cdot \mathbf{v}\|_1 \cdot \|P_h \lambda - \lambda\|_{-1} \\ &\leq Ch \|\mathbf{D}F(\mathbf{u}) \cdot \mathbf{v}\|_1 \cdot \|\lambda\|_0 \leq Ch \|\mathbf{D}F(\mathbf{u})\|_{1,\infty} \cdot \|\lambda\|_0 \cdot \|\mathbf{v}\|_1. \end{aligned}$$

Substituting (5.13)–(5.15) into (5.12), gives (5.11). \square

From this lemma, we see that y_* satisfies (5.10) for small h . Thus, there exists a unique solution for (4.6). Moreover, the solution satisfies the estimate (5.9). We state this conclusion more clearly in the following theorem.

THEOREM 5.3. *Assume that $\mathbf{u} \in \mathbf{H}^2(\Omega) \cap \mathbf{W}^{1,\infty}(\Omega)$. Then, for sufficiently small h , there exists a unique saddle point $(\mathbf{u}_h, \lambda_h) \in X_h$ for (4.6) in a small neighborhood of $(\pi_h \mathbf{u}, P_h \lambda)$. Moreover, the following error estimate holds:*

$$\|\mathbf{u}_h - \mathbf{u}\|_1 + \|\lambda_h - \lambda\|_{-1} \leq Ch.$$

6. Preconditioned iterative methods. We shall combine a preconditioning technique with the classical Newton’s method; cf., for example [27, chapter 7], to solve the nonlinear saddle point problem (4.6) or equivalently (5.2). Of course, Newton’s method will only converge if the initial value is close enough to the true solution. Therefore, in practical computations, it is often necessary to use another global method to obtain an appropriate initial value. A systematic study of such techniques is beyond the scope the present work. However, some alternatives to supply a good initial value are given in the example in section 7.2 below.

Let $x_0 = (\mathbf{u}_h^0, \lambda_h^0) \in X_h$ be a suitable initial guess. The Newton iteration is given by

$$(6.1) \quad x_{n+1} = x_n - G'_h(x_n)^{-1} G_h(x_n), \quad n = 0, 1, \dots$$

Assume that the initial guess $(\mathbf{u}_h^0, \lambda_h^0)$ satisfies (4.13), with a small γ . Using Theorem 5.2, combined with Lemma 5.1 and the standard properties of Newton’s method, it follows that all $x_n = (\mathbf{u}_h^n, \lambda_h^n)$ satisfy (4.13), with the same γ , and all the operators $G'_h(x_n)$ are invertible. Moreover, the sequence $\{(\mathbf{u}_h^n, \lambda_h^n)\}$ converges with almost order 2, i.e.,

$$\|\mathbf{u}_h^{n+1} - \mathbf{u}_h\|_1 + \|\lambda_h^{n+1} - \lambda_h\|_{-1} \leq C \log^2(h^{-1}) (\|\mathbf{u}_h^n - \mathbf{u}_h\|_1 + \|\lambda_h^n - \lambda_h\|_{-1})^2.$$

For the iteration (6.1), we need to invert $G'_h(x_n)$, i.e., we need to solve the system

$$(6.2) \quad G'_h(x_n)(x_{n+1} - x_n) = -G_h(x_n).$$

From Theorem 5.2, we obtain that $G'_h(x_n)$ is an isomorphism from X_h to X_h^* . Moreover, $\|G'_h(x_n)\|_{L(X_h, X_h^*)}$ is bounded, and the bound is independent of h and n if the initial value is chosen close enough to the true solution. Hence, based on preconditioning theory as in [2, 3], we see that any isomorphism from X_h^* to X_h is an optimal preconditioner for system (6.2). Due to this, we can construct some efficient preconditioners for (6.2). Let $\mathbf{\Delta}_h$ and Δ_h be the finite element discretizations for the vector and scalar Laplacian operators $\mathbf{\Delta}$ and Δ on $\mathbf{V}_{h,0}$ and $V_{h,0}$, respectively. To be precise, $\mathbf{\Delta}_h : \mathbf{V}_{h,0} \mapsto \mathbf{V}_{h,0}$ is the mapping defined by

$$\langle \mathbf{\Delta}_h \mathbf{u}_h, \mathbf{v} \rangle = -\langle \nabla \mathbf{u}_h, \nabla \mathbf{v} \rangle, \quad \mathbf{v} \in \mathbf{V}_{h,0}.$$

Then the operator

$$T_h = \begin{pmatrix} -\mathbf{\Delta}_h^{-1} & 0 \\ 0 & -\Delta_h \end{pmatrix}$$

is an isomorphism from X_h^* to X_h , with associated operator norm bounded independently of h . Thus, $T_h \circ G'_h(x_n)$ maps X_h to X_h , with condition numbers bounded independently of h and n . However, in order to make the preconditioner efficient, it is necessary to simplify the evaluation of the operator T_h . We therefore replace $\mathbf{\Delta}_h^{-1}$ by another spectral equivalent operator, i.e., by a preconditioner for the discrete Laplacian using domain decomposition or multigrid methods [31, 33]. The linear system (6.2) is then solved by the preconditioned minimum residual method, with the modified T_h operator \tilde{T}_h as the preconditioner; cf. [28] or [18, Chapter 6]. Since the condition number of the operator $\tilde{T}_h \circ G'_h(x_n)$ is bounded independent of h and n , so is the convergence of the iteration.

7. Numerical experiments. Numerical experiments for the harmonic map problem with $\mathcal{M} = \mathbf{S}^1$, i.e., the unit circle, will be done. The domain Ω is always a square. The domain is triangulated by first dividing it into $h \times h$ squares. Then, each square is divided into two triangles by the diagonal with a negative slope of Ω , which is further divided into triangles by the diagonal with a negative slope. The finite element problem (4.6) is to find $(\mathbf{u}_h, \lambda_h) \in \mathbf{V}_{h,\mathbf{g}} \times V_{h,0}$ such that

$$(7.1) \quad \begin{aligned} \langle \nabla \mathbf{u}_h, \nabla \hat{\mathbf{v}}_h \rangle + \langle \pi_h(\mathbf{u}_h \cdot \hat{\mathbf{v}}_h), \lambda_h \rangle &= 0, & \hat{\mathbf{v}}_h &\in \mathbf{V}_{h,0}, \\ \langle \pi_h(|\mathbf{u}_h|^2 - 1), \hat{\mu}_h \rangle &= 0, & \hat{\mu}_h &\in V_{h,0}. \end{aligned}$$

For the finite element method, we need to integrate over each element $e \in \mathcal{T}_h$. If we use the three vertices of e as the integration points, then the mass matrix reduces to a diagonal matrix. Correspondingly, the system (7.1) reduces to

$$(7.2) \quad \begin{aligned} -\mathbf{L}_h \mathbf{u}_h + \lambda_h \mathbf{u}_h &= \mathbf{0} & \text{on } \mathcal{N}_h, \\ |\mathbf{u}_h|^2 - 1 &= 0 & \text{on } \mathcal{N}_h. \end{aligned}$$

Above \mathbf{L}_h is the standard five-point finite difference discrete Laplacian approximation. For the Newton iteration (6.1), we need to solve the system

$$(7.3) \quad \begin{pmatrix} -\mathbf{L}_h + \mathbf{\Lambda}_n & \text{diag}(\mathbf{u}_n) \\ \text{diag}(\mathbf{u}_n)^t & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{u}_{n+1} - \mathbf{u}_n \\ \lambda_{n+1} - \lambda_n \end{pmatrix} = \begin{pmatrix} \mathbf{L}_h \mathbf{u}_n - \lambda_n \mathbf{u}_n \\ (1 - |\mathbf{u}_n|^2)/2 \end{pmatrix} \text{ on } \mathcal{N}_h.$$

Here and below, we use the simplified notation $(\mathbf{u}_n, \lambda_n)$ instead of $(\mathbf{u}_h^n, \lambda_h^n)$. Furthermore, \mathbf{A}_n and $\mathbf{diag}(\mathbf{u}_n)$ are the matrix representations of the operators $\mathbf{v} \mapsto \pi_h(\lambda_n \mathbf{v})$ and $\mu \mapsto \pi_h(\mu \mathbf{u}_n)$, respectively. From Theorem 5.2, it is interesting to observe that the block-diagonal matrix $T_h = \mathbf{diag}(\mathbf{L}_h^{-1}, L_h)$ is a uniform preconditioner for the matrix of system (7.3).

For the Newton iteration (7.3) with the preconditioner

$$T_h = \mathbf{diag}(\mathbf{L}_h^{-1}, L_h),$$

the matrix \mathbf{L}_h^{-1} in T_h is replaced by a symmetric and spectrally equivalent multi-grid operator, while the matrix L_h is simply a discrete Laplacian with homogeneous Dirichlet boundary conditions. By doing so, no matrix needs to be inverted during the iterations. The cost per iteration is $O(N)$, where N is the degree of freedom for the discretization.

In the following, we will investigate if it is possible to replace Newton’s method with a modified method where the linear system (6.2) is only solved to a given accuracy. More precisely, we shall compare the behavior of the exact and an inexact Newton solver:

- The exact Newton solver: This refers to the scheme where we solve the linear system (6.2) with a preconditioned minimum residual method, which is terminated when the residual is reduced by a factor of 10^{10} .
- The inexact Newton solver: This refers to the scheme where the Newton iterations (6.2) are terminated when the residual is reduced by a factor of 10^2 .

In the tables, we show the numerical errors e_n versus the iteration number n , where e_n is defined as

$$(7.4) \quad e_n = \|\mathbf{u}_h^n - \mathbf{u}_h\|_{\mathbf{H}_h^1} + \|\lambda_h^n - \lambda_h\|_{H_h^{-1}},$$

where $\|x_h\|_{\mathbf{H}_h^1}^2 = (\pi_h x_h)^t (I - \mathbf{L}_h) \pi_h x_h$ and $\|y_h\|_{H_h^{-1}}^2 = (\pi_h y_h)^t (I - L_h)^{-1} \pi_h y_h$.

7.1. A smooth harmonic map. In the first example we consider a smooth harmonic map

$$\mathbf{u} = (\sin(\theta(x, y)), \cos(\theta(x, y))),$$

with $\theta = k \log(\sqrt{(x - a)^2 + (y - b)^2})$ and $\lambda = -|\nabla \mathbf{u}|^2$ on $\Omega = [0, 1] \times [0, 1]$. We have used $a = b = -0.1$ and $k = 3$. The initial guess was $\mathbf{u}_0 = 2(\pi_h \mathbf{u} + \epsilon)$, where ϵ is a random noise vector field with values between -0.3 and 0.3 and $\lambda_0 = 0$.

When using the inexact Newton solver, the stop criterion is obtained in less than 20 iterations, with a few exceptions in the first nonlinear iterations where the maximum was 80. For the exact Newton solver, the stop criterion is obtained in less than 50 iterations with a few exceptions in the first nonlinear iterations where as much as 300 iterations were required on the finest mesh. Hence, except for the first iterations, the required number of iterations seems to be bounded independent of the mesh size. This is due to the property of the preconditioner.

In Table 1 we estimate the convergence of the L^2 and H^1 norms of the error of $\mathbf{u} - \mathbf{u}_h$ in terms of h . We observe linear convergence in \mathbf{H}^1 and quadratic convergence in L^2 , respectively. The convergence in \mathbf{H}^1 is in accordance with the error estimate of Theorem 5.1. The improved rate of convergence in L^2 has not been justified in this paper, but this effect is in agreement with standard linear theory. Also, in the

TABLE 1
The L_2 and H^1 error of \mathbf{u} and the L_2 error of λ with respect to h .

h	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}
$\ \mathbf{u} - \mathbf{u}_h\ _0$	6.7e-1	3.6e-2	9.4e-3	2.4e-3	6.0e-4
$\ \mathbf{u} - \mathbf{u}_h\ _1$	4.6	1.1	5.7e-1	2.9e-1	1.4e-1
$\ \lambda - \lambda_h\ _0$	4.2e-1	2.2e-2	1.6e-3	1.5e-4	1.2e-5

TABLE 2
Convergence for the exact and inexact Newton solver with $h = 2^{-4}$.

	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8
<i>Exact</i>	3.2e+1	9.3	1.7	2.3e-1	4.0e-3	3.4e-6	2.6e-9	-
<i>Inexact</i>	3.2e+1	9.5	1.7	2.4e-1	3.5e-3	1.1e-5	1.0e-7	2.7e-9

TABLE 3
Convergence for the the inexact Newton solver.

$h \setminus it.$	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8
2^{-2}	9.2	2.6	4.7e-1	2.8e-2	1.9e-4	9.9e-7	7.7e-9	7.6e-10
2^{-3}	1.6e+1	4.7	9.1e-1	7.6e-2	8.8e-4	4.0e-6	7.9e-8	1.4e-9
2^{-4}	3.2e+1	9.5	1.7	2.4e-1	3.5e-3	1.1e-5	1.0e-7	2.7e-9
2^{-5}	6.4e+1	2.4e+1	3.6	9.6e-1	1.5e-2	4.7e-5	1.5e-6	6.6e-9

present example the observed convergence for $\lambda - \lambda_h$ is better than that Theorem 5.1 predicts.

A comparison of the exact Newton and inexact Newton solvers is shown in Table 2 for mesh size $h = 2^{-4}$. The convergence for other mesh sizes is similar. These tests indicate that the inexact Newton solver is nearly as efficient as the exact Newton solver. In Table 3, the convergence of the inexact Newton solver with different mesh sizes is shown. It shows the mesh independence property of the preconditioned iterative solver.

7.2. A harmonic map with singularity. As it is well known, the solution of the harmonic map problem is generally not unique and may have singularities even with smooth data. In order to show the applicability of our algorithms for these problems, we test a problem with a singular solution, i.e., $\mathbf{u} = (x/r, y/r)$, with $r = k\sqrt{x^2 + y^2}$ and $\lambda = -|\nabla \mathbf{u}|^2$ on $\Omega = [-0.5, 0.5] \times [0.5, 0.5]$. The pair (\mathbf{u}, λ) corresponds to a classical solution of the saddle point system away from the origin, but $\|\mathbf{u}\|_1 = \infty$. Therefore, this example is not covered by our theoretical results, but we include the example to illustrate additional effects. The Dirichlet boundary conditions are obtained from the analytical solution, while the initial value for λ is $\lambda_0 = 0$ everywhere except in $(0, 0)$, where $\lambda = 1$. The initial value for \mathbf{u} is shown in Figure 1(a). The computed solution is shown in Figure 1(b). The numerical errors are given in Table 4. The errors indicate that both \mathbf{u}_h and λ_h converge linearly to the solution when measured in L^2 . It is interesting to observe that we get convergence for $\|u - \mathbf{u}_h\|_0$ and $\|\lambda - \lambda_u\|_0$ even without mesh refinement around the singularity.

For this example, the Newton solvers are unstable and do not always converge. Thus, we have used the following iteration to produce the initial value for the Newton solvers:

$$(7.5) \quad \begin{pmatrix} -\mathbf{L}_h & \text{diag}(\mathbf{u}_n) \\ \text{diag}(\mathbf{u}_n)^t & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{u}_{n+1} - \mathbf{u}_n \\ \lambda_{n+1} - \lambda_n \end{pmatrix} = \begin{pmatrix} \mathbf{L}_h \mathbf{u}_n - \lambda_n \mathbf{u}_n \\ (1 - |\mathbf{u}_n|^2)/2 \end{pmatrix}.$$

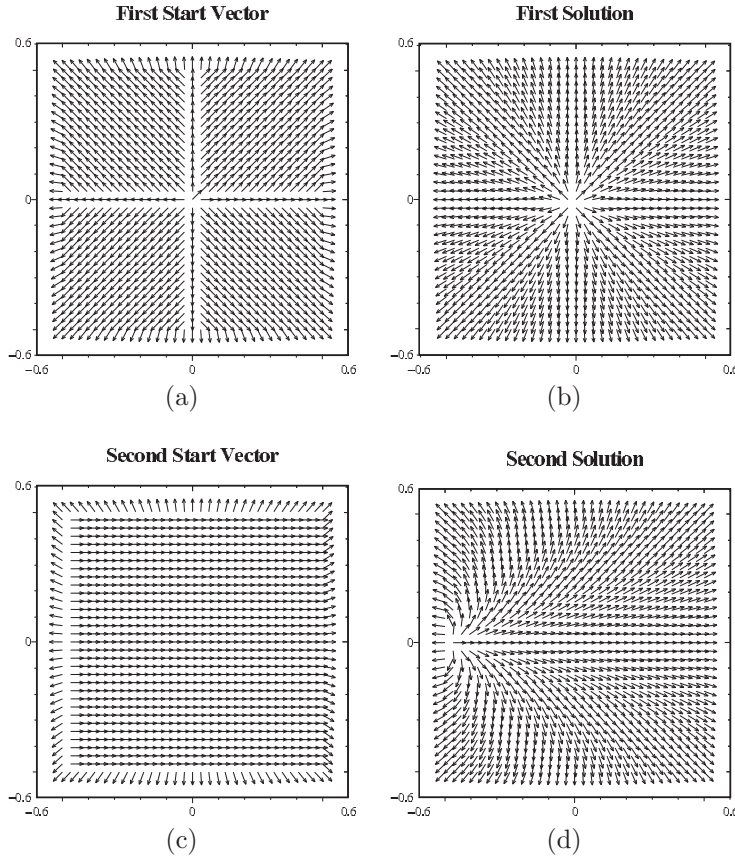


FIG. 1. Plot of the initial solutions and the computed solutions. (a) The first initial solution. (b) The solution for (a). (c) The second initial solution. (d) The solution for (c).

TABLE 4
Errors with respect to h for the singular problem.

h	2^{-3}	2^{-4}	2^{-5}	2^{-6}
$\ \mathbf{u} - \mathbf{u}_h\ _0$	2.2e-1	1.3e-1	7.4e-2	4.0e-2
$\ \lambda - \lambda_h\ _0$	8.3e-1	4.1e-1	2.1e-1	1.0e-1

TABLE 5
Convergence for the inexact Newton solver for the singular problem.

e_1	e_5	e_{10}	e_{11}	e_{12}	e_{13}	e_{14}
1.1e+1	6.4e-1	1.1e-1	8.1e-2	9.7e-4	2.4e-7	1.2e-8

Compared with (7.3), the matrix Λ_n has been dropped. This iterative scheme is globally convergent and is normally slower than the Newton solvers. Its convergence properties will be analyzed and discussed elsewhere. We do ten iterations of (7.5), and the inexact Newton solver is then turned on. The results are shown in Table 5 for $h = 2^{-4}$, where it is clear that we have quadratic convergence in the last iterations.

For the smooth problem tested in section 7.1, it seems that the iterative solution always converges to the same solution no matter what kind of initial solution we use. For the problem here, we have noticed that the saddle point problem may have

multiple solutions. With another initial solution, as shown in Figure 1(c), we obtain another solution, which is shown in Figure 1(d).

Acknowledgment. The authors are grateful to Kent Mardal who has supplied the numerical experiments for this work.

REFERENCES

- [1] F. ALOUGES, *A new algorithm for computing liquid crystal stable configurations: The harmonic mapping case*, SIAM J. Numer. Anal., 34 (1997), pp. 1708–1726.
- [2] D. N. ARNOLD, R. S. FALK, AND R. WINTHER, *Preconditioning discrete approximations of the Reissner–Mindlin plate model*, M2AN Math. Model. Numer. Anal., 31 (1997), pp. 517–557.
- [3] D. N. ARNOLD, R. S. FALK, AND R. WINTHER, *Preconditioning in $H(\text{div})$ and applications*, Math. Comp., 66 (1997), pp. 957–984.
- [4] J. BARRETT, S. BARTELS, X. FENG, AND A. PROHL, *A convergent and constraint-preserving finite element method for the p -harmonic flow into spheres*, SIAM J. Numer. Anal., 45 (2007), pp. 905–927.
- [5] S. BARTELS, *Stability and convergence of finite-element approximation schemes for harmonic maps*, SIAM J. Numer. Anal., 43 (2005), pp. 220–238.
- [6] D. BERTSEKAS, *Constrained Minimization and Lagrange Multiplier Methods*, Athena Scientific, Belmont, MA, 1996.
- [7] H. BREZIS, *The interplay between analysis and topology in some nonlinear PDE problems*, Bull. Amer. Math. Soc., 40 (2003), pp. 179–201.
- [8] F. BREZZI, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers*, RAIRO Anal. Numér., 8 (1974), pp. 129–151.
- [9] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer, New York, 1991.
- [10] F. BREZZI, J. RAPPAZ, AND P. RAVIART, *Finite dimensional approximation of nonlinear problems Part I: Branches of nonsingular solution*, Numer. Math., 36 (1980), pp. 1–25.
- [11] Y. CHEN AND M. STRUWE, *Existence and partial regularity results for the heat flow for harmonic maps*, Math. Z., 201 (1989), pp. 83–103.
- [12] Y. CHEN, *The weak solutions to the evolution of harmonic maps*, Math. Z., 201 (1989), pp. 69–74.
- [13] R. COHEN, R. HARDT, D. KINDERLEHRER, S. LIN, AND M. LUSKIN, *Minimum energy configurations for liquid crystals: Computational results*, in Theory and Applications of Liquid Crystals, IMA Vol. Math. Appl. 5, Springer, New York, 1987, pp. 99–121.
- [14] Q. DU, B. GUO, AND J. SHEN, *Fourier spectral approximation to a dissipative system modeling the flow of liquid crystals*, SIAM J. Numer. Anal., 39 (2001), pp. 735–762.
- [15] Q. DU, B. GUO, AND J. SHEN, *Corrigendum: Fourier spectral approximation to a dissipative system modeling the flow of liquid crystals*, SIAM J. Numer. Anal., 41 (2003), pp. 796–798.
- [16] W. E AND X. WANG, *Numerical Methods for the Landau–Lifshitz equation*, SIAM J. Numer. Anal., 38 (2000), pp. 1647–1665.
- [17] I. EKELAND AND R. TEMAM, *Convex Analysis and Variational Problems*, Classics Appl. Math. 28, SIAM, Philadelphia, PA, 1999.
- [18] H. ELMAN, D. SILVESTER, AND A. J. WATHEN, *Finite Elements and Fast Iterative Solvers with Applications in Incompressible Fluid Dynamics*, Oxford University Press, London, 2005.
- [19] R. GLOWINSKI AND P. LE TALLEC, *Augmented Lagrangian and Operator Splitting Methods in Nonlinear Mechanics*, SIAM, Philadelphia, PA, 1989.
- [20] R. GLOWINSKI, P. LIN, AND X. PAN, *An operator-splitting method for a liquid crystal model*, Comput. Phys. Comm., 152 (2003), pp. 242–252.
- [21] R. HARDT, D. KINDERLEHRER, AND M. LUSKIN, *Remarks about the mathematical theory of liquid crystals*, in Calculus of Variations and Partial Differential Equations, Lecture Notes in Math. 1340, Springer, Berlin, 1988, pp. 123–138.
- [22] F. HÉLEIN, *Régularité des applications faiblement harmoniques une surface et une variété riemannienne*, C. R. Acad. Sci. Paris, 312 (1991), pp. 591–596.
- [23] W. JÄGER AND H. KAUL, *Uniqueness and stability of harmonic maps and their Jacobi fields*, Manuscripta Math., 28 (1979), pp. 269–291.
- [24] J. JOST, *Riemannian Geometry and Geometric Analysis*, 4th ed., Springer, Heidelberg, 2005.
- [25] S. LIN AND M. LUSKIN, *Relaxation methods for liquid crystal problems*, SIAM J. Numer. Anal., 26 (1989), pp. 1310–1324.

- [26] M. LYSAKER, S. OSHER, AND X.-C. TAI, *Noise removal using smoothed normals and surface fitting*, IEEE Trans. Image Process., 13 (2004), pp. 1345–1357.
- [27] A. QUARTERONI, R. SACCO, AND F. SALERI, *Numerical Mathematics*, Springer, New York, 2000.
- [28] T. RUSTEN AND R. WINTHER, *A preconditioned iterative method for saddlepoint problems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 887–904.
- [29] R. SOCHEN AND S. T. YAU, *Lectures on Harmonic Maps*, International Press, Somerville, MA, 1997.
- [30] M. STRUWE, *Variational Methods*, 3rd ed., Springer, New York, 2000.
- [31] X. C. TAI AND J. C. XU, *Global and uniform convergence of subspace correction methods for some convex optimization problems*, Math. Comp., 71 (2001), pp. 105–124.
- [32] L. VESE AND S. OSHER, *Numerical methods for p -harmonic flows and applications to image processing*, SIAM J. Numer. Anal., 40 (2002), pp. 2085–2104.
- [33] J. XU, *Iterative methods by space decomposition and subspace correction*, SIAM Rev., 34 (1992), pp. 581–613.