Robustness of an elementwise parallel finite element method for convection-diffusion problems

Citation for published version (APA):

DOI:
10.1137/S1064827595293545

Document status and date:
Published: 01/01/1998

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
ROBUSTNESS OF AN ELEMENTWISE PARALLEL FINITE ELEMENT METHOD FOR CONVECTION-DIFFUSION PROBLEMS

W. J. LAYTON†, J. M. MAUBACH†, AND P. J. RABIER†

Abstract. We consider an elementwise data-parallel finite element procedure, recently proposed by Layton and Rabier [Appl. Math. Lett., 5 (1992), pp. 67–70], [J. Numer. Linear Algebra Appl., 2 (1995), pp. 363–394], applied to singularly perturbed convection-diffusion equations with possibly highly anisotropic diffusion. It is shown that the number of iterations required for the solution of the linear algebraic system is proportional to the inverse of the smallest grid element diameter, uniformly in the diffusion parameter and the degree of anisotropy. This is optimal, since the method can in some cases use only element matrices and load vectors and the algorithm requires only local communication on the physical mesh between adjacent elemental subdomains. Our analysis includes both the usual Galerkin formulation and the streamline upwind finite element formulations. The convergence result holds with conforming elements of any order or elements of arbitrary order from a new family of nonconforming elements.

Key words. parallel, iterative method, robustness

AMS subject classifications. 65N, 65F

P11. S1064827595293545

1. Introduction. This paper addresses the issue of robustness of iterative methods for solving the linear algebraic systems arising from discretizations of singularly perturbed convection-diffusion equations. Robustness, like efficiency, is a central and highly arguable issue through all of numerical analysis and has the general meaning that the algorithm under consideration converges uniformly in the problem’s most important parameters. The performance of a “robust” algorithm should not degrade as one moves further from “model” problems.

We consider the linear system arising from finite element discretizations of singularly perturbed convection-diffusion problems. For \( \Omega \subset \mathbb{R}^2 \), a polygonal domain with boundary \( \Gamma \) and outward unit normal \( n \), the problem is to find \( u : \bar{\Omega} \rightarrow \mathbb{R} \) satisfying

\[
\begin{align*}
-\nabla \cdot (\epsilon(x,y)\nabla u) + b \cdot \nabla u + gu &= f \quad \text{in } \Omega, \\
u &= 0 \quad \text{on } \Gamma_0, \\
\nabla u \cdot n &= 0 \quad \text{on } \Gamma \setminus \Gamma_0,
\end{align*}
\]

where \( \Gamma_0 \subset \Gamma \) is nonempty and contains all boundary points where \( b \cdot n \neq 0 \). We specifically study the case where the eigenvalues \( \lambda \) of the \( 2 \times 2 \) (symmetric) diffusivity tensor \( \epsilon(x,y) \) satisfy

\[
0 \leq \epsilon_{\min} := \inf_{(x,y)} \lambda_{\min}(\epsilon(x,y)) \leq \sup_{(x,y)} \lambda_{\max}(\epsilon(x,y)) := \epsilon_{\max} \leq 1 \quad \text{in } \bar{\Omega},
\]

which includes anisotropic problems and problems convection-dominated in some regions and diffusion-dominated in others. The iterative method studied is a variant of the Peaceman–Rachford operator splitting method using a new splitting, introduced in [21, 22] and studied further in [7, 10, 16, 17, 18, 19, 20, 26, 31, 32] based upon...
a decomposition of the physical elements into two groups. There are several options regarding the grouping of the elements. If conforming elements are used the groups can correspond to a two-coloring of the domain into topological strips. If a member of a new family of nonconforming elements is used, the groups can even correspond to a collection of noninteracting individual elements. To fix ideas let us suppose for the moment that the nonconforming linear element (or one of the new higher order versions; see section 2) is chosen. Then each iteration costs one multiply and one solve with each element matrix, and a local communication. If conforming elements are used, then each iteration costs instead one multiply and one solve with the stiffness matrix corresponding to each topological strip. In either case, the execution speed of each iteration is quite fast, and the resulting algorithm is massively data-parallel.

Let $T$ denote the iteration operator of the iterative solution procedure of Algorithm 2.1 (given precisely in section 2). The principal result of this report 1) states that the Algorithm 2.1 is robust, and 2) gives a worst case estimate of the number of iterations for convergence of the algorithm.

In [22] an abstract convergence theorem was proven for methods including Algorithm 2.1 as a special case. The conditions of this convergence theory are all satisfied, so convergence of Algorithm 2.1 follows immediately. Rates of convergence do not follow from either the results or the approach of [22]. Thus, although we use herein several technical results of [22], we must take an approach tied more directly to the specific model convection-diffusion problem studied herein to obtain Theorem 2.1’s rates of convergence and their uniformity in $ϵ$. In [19] a preliminary (and weaker) version of a robustness result was announced without proof.

Let $|·|$ be the euclidian vector and matrix norms and $δ$ a piecewise constant discretization parameter. The case $δ = 0$ corresponds to the usual Galerkin formulation, and $δ = O(h)$ corresponds to the streamline upwind method [17, 24, 33]. Provided that $g - \frac{1}{2} \nabla \cdot b > 0$, the basic result of this paper, Theorem 2.1, is that the Algorithm 2.1 converges in $O((\min h_e)^{-1})$ iterations (in the sense of requiring $O((\min h_e)^{-1}$ iterations per significant digit of accuracy sought) uniformly in all problem parameters, where $h_e$ denotes the diameter of the element $e$. Specifically, there is a constant $C > 0$, depending only upon $b, g$, and the smallest angle of the triangulation, such that

\[
\sup \left\{ |T| : 0 \leq \epsilon_{\text{min}} \leq \epsilon_{\text{max}} \leq 1, 0 \leq h_e \leq O(h_e), 0 \leq \epsilon_{\text{max}} \leq \epsilon_{\text{min}} \leq \infty \right\} \leq 1 - C \min h_e,
\]  

(1.3)

where $|·|$ is the norm on $T$ induced by the euclidian norm. This result implies uniformity in the diffusivity tensor $\epsilon(x, y)$, the method chosen ($δ = 0$ centered or $δ = O(h)$ streamline upwind), and even in the variability and anisotropy in the diffusivity, without any assumption on the mesh other than the usual angle condition. This result holds in both the conforming and nonconforming case for all values of $p$, the element’s polynomial degree.

The uniformity in the diffusion parameter in (1.3) is a remarkable property of Algorithm 2.1. The right-hand side “$1 - O(h)$” in (1.3) for quasi-uniform meshes is optimal for an algorithm (such as the one considered herein) which requires only nearest neighbor communications on the physical mesh and which uses only subdomain stiffness matrices and load vectors corresponding to subdomains of width $O(h)$.

With regard to the discretization chosen, the problem (1.1) with $g - \frac{1}{2} \nabla \cdot b > 0$ has solution operator uniformly bounded as $\epsilon_{\text{min}} \to 0$. Thus, a robust iterative method
should be uniform in $\epsilon$ for both the usual finite element method (FEM) and the streamline upwind FEM. The former leads to linear systems which are very sensitive and, for small $\epsilon$, give many other solution procedures serious problems in addition to producing severely oscillating solutions. The robustness of our algorithm shows, for example, that convergence is not affected by choosing $\delta = 0$ in regions where $\epsilon_{\text{min}} = O(1)$ (or where $u$ is smooth) and $\delta_{\epsilon} = O(h_{\epsilon})$ otherwise.

As a point of comparison, consider the usual model Poisson problem on the unit square using a quasi-uniform mesh. The only parameter here is the meshwidth $h$, and the linear system has uniformly bounded inverse as $h \to 0$. Thus, “robust” for symmetric problems means optimally “efficient” or “optimal order” in $h$. Optimal order methods for symmetric boundary value problems, such as multilevel methods, have been extensively developed in recent years. In the nonsymmetric case (1.1) the most important problem parameters are contained in the diffusion tensor $\epsilon$. Thus, an obvious distinction between “robustness,” “efficiency,” and “optimal order” arises here. For example, the computational effort of one optimal order method of Chen and Temam [6] for (1.1) is estimated there to be (for a uniformly refined mesh and $\epsilon = \epsilon I$) $C(\epsilon h^{-2} |\epsilon n(h)|)$ floating point operations where, rescaling their estimates appropriately, $C(\epsilon) = O(\epsilon^{-2})$ a quite good result. In contrast, in this setting, the complexity of Algorithm 2.1 is $Ch^{-3}$ floating point operations with $C$ uniform in $\epsilon$. The sequential superiority of “optimal order” methods for symmetric or nearly symmetric problems is clear when $\epsilon_{\text{min}} = O(1)$, as is the advantage of Algorithm 2.1 for $\epsilon$ small with respect to $h$.

The search for “good” iterative methods for highly nonsymmetric problems continues to be an extremely important and active area of numerical analysis. We mention in particular the methods of Strikwerda [29], Han et al. [12], Kellogg [17], Chen and Temam [6], Sonneveld (CGS method) [30], and Axelsson (GCGLS) method [2] as especially promising in connection with good block-type preconditioners [5]. Reusken [28] considers a novel two-level method for solving the linear system arising from the first-order upwind discretization of constant coefficient convection-diffusion equations with $\epsilon = \epsilon I$ subject to periodic boundary conditions. He shows that for such problems, the “local modes” of the error converge uniformly in $\epsilon$. Such local mode analysis does not take into account the presence of boundary and interior layers for realistic boundary conditions, the nonnormality of the coefficient matrices for other boundary conditions, and the resulting mathematical gap between normal mode convergence and convergence in various norms.

The first-order upwind discretization, studied in [28], also artificially increases the diffusion coefficient to $O(\epsilon + h)$. (Streamline diffusion discretizations of through-flow–type problems can have similar effects.) Nevertheless, the work of Reusken [28], verifying the necessary condition of robustness in local modes, is an important first step for two-level methods. To our knowledge, the present algorithm is still the only one for which a uniform convergence result of the form (1.3) is both true and accessible to a fully rigorous mathematical proof. We do not have any experience attempting to increase the efficiency of our method by using it as a smoother in a two-level method; its use as a preconditioner is mentioned in [19].

Concurrently, there has been a very active research effort directed at the development of elementwise data-parallel finite element procedures under various names like “element-by-element preconditioning” and “finite elements without assembly.” See, for example, Hughes, Levit, and Winget [13], Nour-Omid and Parlett [27], and Wathen [34]. The (usual) motivations for these procedures are that the “assembly” of the
global stiffness matrix typically entails significant global communication, calculation of element matrices is embarrassingly parallel, and, in nonlinear problems, calculation of element matrices and assembly of global matrices can often dominate the entire calculation. Since the solution procedures of [21, 22] give a particularly simple elementwise data-parallel solution procedure which requires only local communications on the physical mesh when nonconforming elements are used, the use of our solution procedure with these elements is a natural choice for an adaptive algorithm to be used in a parallel environment. The basic algorithm shall therefore be presented for the case of nonconforming elements (in Algorithm 2.1). (The algorithm is presented in the conforming case in [22].) The potential of Algorithm 2.1 on various parallel architectures has been demonstrated in, for example, Sunmonu [30, 31], Layton et al. [20], and Crosbie and Ervin [7].

2. Parallel finite element methods and convergence estimate. We focus on the “model” problem (1.1) of finding $u(x, y)$ satisfying

\[
\begin{cases}
-\nabla \cdot (\epsilon(x, y) \nabla u) + b(x, y) \cdot \nabla u + g(x, y)u = f & \text{in } \Omega, \\
u = 0 & \text{on } \Gamma_0, \quad \nabla u \cdot n = 0 & \text{on } \Gamma \setminus \Gamma_0,
\end{cases}
\]

where the coefficients are sufficiently smooth; $\Omega$ is a planar, polygonal domain with outward unit normal $n$ and boundary $\Gamma$; and $\Gamma_0 \subset \Gamma$ is nonempty, closed, and contains all points where $b \cdot n \neq 0$. We shall assume

\[
|b(x, y)| \leq b_{\text{max}} < \infty, \quad 0 \leq \epsilon_{\text{min}} \leq \epsilon_{\text{max}} \leq 1 \text{ in } \Omega,
\]

\[
g - \frac{1}{2} \nabla \cdot b > g_{\text{min}} > 0 \text{ in } \bar{\Omega}.
\]

Remark 2.1. The assumption (2.2b) ensures that $\|u\|_{L^2(\Omega)} \leq C\|f\|_{L^2(\Omega)}$ uniformly in $\epsilon$. Thus it is reasonable to ask a method to be uniform in $\epsilon$ under (2.1). If (2.2b) fails, e.g., $g - \frac{1}{2} \nabla \cdot b \equiv 0$, then (for constant $\epsilon$ say)

\[
\|u\|_{L^2(\Omega)} \leq C\|\nabla u\|_{L^2(\Omega)} \leq C\epsilon^{-1}\|f\|_{H^{-1}(\Omega)}.
\]

Thus, robustness might be defined when (2.2b) fails so as to allow an $O(\epsilon^{-1})$ growth in the complexity as $\epsilon \to 0$. The case $g - \frac{1}{2} \nabla \cdot b \equiv 0$ is beyond the scope of the present report.

Let $X^{h,1}$ denote the space of nonconforming linear Crouzeix–Raviart elements (introduced in [8]). These elements are associated with midedges of the given edge-to-edge triangulation of $\Omega$, $H^1(\Omega)$. $X^{h,p}$ will denote the nonconforming finite element, recently introduced in [26], of degree $p$. For odd $p \geq 1$ these are quite simple and involve node placements as depicted in Figure 2.1.

The case of even $p$ is also fully resolved in [26]. Its (more technical) presentation is given there in detail.

Remark 2.2. The original paper of Crouzeix and Raviart [8] presents the linear element and one higher order example—an element containing full cubics and some quartics. The $p$ odd elements of [26] are systematic and simple; they contain only full $p$-degree polynomials. The case of even $p$ is notorious in that the natural attempts fail. For example, taking the two Gauss points on each edge for $p = 2$ does not result in a $P_2$ unisolvent set of degrees of freedom. The case of even $p$ is also fully resolved in [26]. There are interesting connections between nonconforming elements and mixed-type discretizations using conforming elements (see Arbogast and Chen [1]), which are yet to be explored for either this new family or Algorithm 2.1.
The subspace $X_0^{h,p}(\Omega)$ of $X^{h,p}(\Omega)$ will consist of those functions vanishing at the nodes lying on $\Gamma_0$. Without the loss of generality, we suppose that no element $e$ has more than one edge on $\Gamma_0$. Let $\delta \in L^\infty(\Omega)$ be an $O(h)$ elementwise constant parameter. The usual streamline diffusion ($|\delta|_\infty, \Omega = O(h)$) to the usual Galerkin ($\delta = 0$) family of discretizations of (2.1) is given by: find $u^h \in X_0^{h,p}(\Omega)$ satisfying (with $\delta_e := |\delta|_e$)

$$\sum_{e \in \Pi^h(\Omega)} \left[ a_e(u^h, v) - \int_e f(v + \delta_e b \cdot \nabla v) \right] = 0 \quad \forall \ v \in X_0^{h,p}(\Omega), \tag{2.3}$$

with $a_e(\cdot, \cdot)$ $\delta$-dependent and given by

$$a_e(u, v) := \sigma_e(u, v) + \tilde{\sigma}_e(u, v) + \beta_e(u, v), \tag{2.4}$$

$$\sigma_e(u, v) := \int_e \left[ (\epsilon \nabla u) \cdot \nabla v + \delta_e (b \cdot \nabla u)(b \cdot \nabla v) \right. \right. \left. \left. + \int_e \left( g - \frac{1}{2} \nabla \cdot (b + \delta_e g b) \right) uv \right], \tag{2.5}$$

$$\tilde{\sigma}_e(u, v) := \frac{1}{2} \int_e (1 - \delta_e g)[(b \cdot \nabla u)v - (b \cdot \nabla v)u], \tag{2.6}$$

$$\beta_e(u, v) := \int_e \delta_e \nabla \cdot (\epsilon(x, y) \nabla u)(b \cdot \nabla v).$$

Note that the “conforming” part of $a_e(\cdot, \cdot)$ has been explicitly divided into its symmetric part $\sigma_e(\cdot, \cdot)$ and skew-symmetric part $\tilde{\sigma}_e(\cdot, \cdot)$. One crucial point here is that the bilinear form $a_e$ differs from the more natural choice

$$\tilde{a}_e(u, v) := \int_e \left[ -\delta_e \nabla \cdot (\epsilon(x, y) \nabla u)(b \cdot \nabla v) + (\epsilon(x, y) \nabla u) \cdot \nabla v + (b \cdot \nabla u)v \right.$$

$$\left. + \delta_e (b \cdot \nabla u)(b \cdot \nabla v) + gu(v + \delta_e b \cdot \nabla v) \right],$$

insofar as the “conforming” part of $a_e(\cdot, \cdot)$ has been explicitly skew-symmetrized. The difference between $a_e(\cdot, \cdot)$ and $\tilde{a}_e(\cdot, \cdot)$ involves discarding boundary integrals (which vanish when $u$ is the true solution of (2.1) due to either the boundary condition $u = 0$
on $\Gamma_0$ or $b \cdot n = 0$ on $\Gamma \setminus \Gamma_0$) which appear in the splitting of the above expression into symmetric and skew-symmetric parts. On the one hand, the contribution of these boundary integrals can be incorporated into the “consistency error.” On the other hand, the choice of the element bilinear forms $a_e(\cdot, \cdot)$ in (2.4)–(2.6) instead of $\tilde{a}_e(\cdot, \cdot)$ is essential, as the key property required by [13, 22] $a_e(u, u) \geq 0$ (i.e., $\sigma_e(u, u) + \beta_e(u, u) \geq 0$) does not hold with $a_e$ replaced by $\tilde{a}_e$.

The case $\delta = 0$ is precisely the usual Galerkin method with explicitly skew-symmetrized convection term. The case $\delta > 0$ corresponds to the streamline upwind method in which $\delta = O(h)$ when $\epsilon_{\text{max}} \leq h$ and $\delta = 0$ when $\epsilon_{\text{min}} \geq h$. The proof of stability and convergence of this method needs an extra smallness condition on $\delta$ which takes the following form. On each element $e$, $\delta_e = \delta_{0,e} h_e$, where $\delta_{0,e}$ is small enough that

\begin{equation}
\tag{2.7a}
\tilde{g}_e := g_e - \frac{1}{2} \nabla \cdot (b + \delta_e g b) \geq \frac{1}{2} \left( g - \frac{1}{2} \nabla \cdot b \right)_{\text{min}} > 0
\end{equation}

and, when $p > 1$, $\delta_{0,e}$ is also small enough that

\begin{equation}
\tag{2.7b}
0 \leq \delta_{0,e} \leq \frac{1}{2} \frac{h_e \epsilon_{\text{min},e}}{C_p E [\epsilon e_{\text{max}}^2 + h_e^2 |\nabla \epsilon|^2_{\text{min}}]},
\end{equation}

where $\epsilon_{\text{max},e} = \sup_{(x,y) \in \epsilon} \lambda_{\text{max}}(\epsilon(x,y))$ and $\epsilon_{\text{min},e} = \inf_{(x,y) \in \epsilon} \lambda_{\text{min}}(\epsilon(x,y))$, and the infinity norms of the tensors $\epsilon$ and $\nabla \epsilon$ are defined precisely in the Appendix. These conditions (2.7a,b) ensure that $\sigma_e(u, u) > 0$ and (when $p > 1$) $\sigma_e(u, u) + \beta_e(u, u) > 0$; see Lemma A.1. The condition (2.7b) is the variable $\epsilon(x, y)$ version of the usual restriction on $\delta_{0,e}$, in the constant $\epsilon$ case (compare with [15, p. 299, Remark 1.6]). We note that if conforming or nonconforming linear elements are used, this $\beta_e(u, v)$ perturbation term is identical zero. There have been many studies of the discretization error of these methods for various model problems [17, 26, 33]. In particular, the combination of nonconforming elements with this variational formulation has been systematically validated in [26, 33]. Thus, we take (2.3) as representative of a family of accepted discretization methods for (2.1).

We shall present the solution algorithm in the case of nonconforming elements. This requires some preliminary formalism. Let $e \in \Pi_1(\Omega)$ be a generic element which contains $\alpha$ nodes. Let $N_1, N_2, \ldots, N_\alpha$ (resp., $N_1, N_2, \ldots, N_{\alpha-p}$) be the $\alpha$ (resp., $\alpha-p$) nodes on the element $e$ if $e$ has no (resp., one) edge on $\Gamma$. Let $\phi_{N_i}, 1 \leq i \leq \alpha$ (resp., $\alpha-p$), be the associated nodal basis functions. Recall that $A_e$ is the $\alpha \times \alpha$ matrix (resp., $(\alpha-p) \times (\alpha-p)$)

\begin{equation}
\tag{2.8}
[A_e]_{ij} := a_e(\phi_{N_i}, \phi_{N_j}) \quad \text{for all nodes }\quad N_i, N_j \quad \text{on the element } e.
\end{equation}

Also, recall that $f_e$ is the $\alpha$-vector (resp., $(\alpha-p)$ vector) given by $[f_e] := \int_e f \phi_N dx$ for all $N_i$ nodes on $e$.

The solution algorithm exploits additivity of the integrals with respect to subdomains in the variational formulation. To this end, let $\Pi_1(\Omega)$ be two-colored, each element being colored “red” or “black.” If conforming elements are used then each connected red or black subdomain is a topological strip. If nonconforming elements are used the elements are easily two-colored using, for example, the algorithm of [14]. The connected red or black subdomains are then either individual elements or two adjacent elements. The most straightforward case is when each interior vertex is shared by an even number of triangles. In this case (illustrated in Figure 2.2) the connected, same-color subdomains are individual elements (and the mesh is said to be “two-colorable”).
Fig. 2.2. An example of a two-colored mesh and the associated node orderings.

(To reduce the number of cases considered at each step we shall suppose the vertex degree of interior vertices of $\Pi^h(\Omega)$ is even so that $\Pi^h(\Omega)$ is two-colorable, as illustrated in Figure 2.2.) It is convenient to consider three global node orderings: fixed, red, and black. The fixed ordering ($F$) is the original one determined by the mesh construction. The red ordering ($R$) enumerates first nodes belonging to the red elements, consecutively if they lie in the same red triangle and then the few remaining (boundary) nodes not on any red triangle. The black ordering ($B$) is defined, mutatis mutandis, by replacing “red” with “black” in the above description. Associated with these are permutation matrices $P^{B-R}_R$, $P^{R-B}_R$ ($= (P^{B-R}_R)^*$), $P^{F-R}_R$, $P^{R-F}_R$ ($= (P^{F-R}_R)^*$), $P^{F-B}_B$, $P^{B-F}_B$ ($= (P^{F-B}_B)^*$) which reorder in the indicated manner. For example, if $x$ represents nodal data in the red ordering then $P^{B-R}_R x$ is the same data in the black ordering. An example of a colored mesh with the red and black ordering is given in Figure 2.2. The elementwise parallel
Algorithm 2.1. Elementwise Parallel Finite Element Method.

1. Assemble in parallel: all element matrices $A_e, e \in \Pi^h(\Omega)$.
2. Choose: $\mu > 0$ and initial guess $u^n_h$.
3. For color = red, black, red, black, ..., until convergence, do:
   3a. For all $e \in \Pi^h(\Omega)$ with specified color compute in parallel:
       $$r_e := f_e + (\mu I - A_e)u^n_e.$$ 
   3b. For all $e \in \Pi^h(\Omega)$ with specified color solve in parallel:
       $$(A_e + \mu I)u^{n+1}_e = r_e.$$ 
   3c. Reorder to other color ordering and switch to other color:
       $$u^{n+1} \leftarrow P_{\text{other color}} u^{n+1}.$$ 
4. Upon convergence: Set
       $$u := P_{\text{fixed color}} u^{n+1} + P_{\text{other color}} u^n.$$ 

Remark 2.3. Selection of the Parameter $\mu$. For $\epsilon$ small (and constant), theory and computation all point to choosing $\mu = \mu_0 h$. Thus, a good value of $\mu_0$ can be found on a very coarse mesh and then used thereafter. (In all our experiments in section 4 we have used $\mu = \mu_0 h$.) For larger $\epsilon$, $\mu_0$ should be increased; see Table 4.2. In fact, a heuristic calculation suggests formulas like $\mu \sim \sqrt{h}(\epsilon + h)$. The parameter $\mu$ can also be chosen node by node as a function of the local meshwidth, or element by element. Although these (minor) modifications do not alter the theoretical result in Theorem 2.1, we have found that they improve the efficiency of Algorithm 2.1 in practical calculations substantially.

Remark 2.4. Comparison with Element-by-Element Preconditioning. The early and important element-by-element preconditioning method of Hughes, Levit, and Winget [13] can be described (in our formalism) as follows. With conforming elements, color the domain, using a “greedy” algorithm, so that no triangle is adjacent across a vertex with a same-color triangle. This requires $k = \text{vertex degree} (\Pi^h(\Omega)) + 1$ colors. Assemble the associated stiffness matrices $A_\ell, \ell = 1, \ldots, k$ which (analogous to the discussion which follows this remark) are each block diagonal with color $\ell$ element matrices for blocks in an appropriate $\ell$th node ordering. The method of [13] then uses $\Pi_{\ell=1}^k (I + A_\ell)^{-1}$ as a preconditioner for the global system. Available theory and experiments all suggest that the preconditioned system’s condition number is still $O(h^{-2})$ (with a smaller constant). On the other hand, if our method instead is used as a preconditioner, the original system’s condition number is reduced from $O(h^{-2})$ to $O(h^{-1})$ uniformly in the problem parameters. This comparative advantage is due to two factors: first, by restricting ourselves to two colors (at the cost of somewhat larger blocks) we can access the powerful convergence theory of methods involving splittings into only two addends; second, because of these two additive terms in our splittings we obtain an error equation for which a good scaling of the acceleration parameter $\mu$ is obtainable.

As presented above it is obvious that the algorithm is embarassingly data-parallel with elemental data as the logical units. Algorithm 2.1 arises from the classical
Dyakonov splitting method (see Marchuk [23]) using the elementwise parallel splitting of [21, 22]. More precisely, order the “red” (resp., “black”) elements \(e^R\) (resp., \(e^B\)) consistently with the “red” (resp., “black”) ordering of the nodes, say \(e^R_1, \ldots, e^R_{L_{Red}}\) (resp., \(e^B_1, \ldots, e^B_{L_{Black}}\)), and let \(A_R\) and \(A_B\) be the \(N^h \times N^h\) (\(N^h = \dim X^h_0(\Omega)\)) block diagonal matrices

\[
A_R := \begin{pmatrix}
A^R_{e_1} & & \\
& \ddots & \\
& & A^R_{e_{L_{Red}}} \\
0 & & \\
& \ddots & 0 \\
0 & & \\
\end{pmatrix},
\]

\[
A_B := \begin{pmatrix}
A^B_{e_1} & & \\
& \ddots & \\
& & A^B_{e_{L_{Black}}} \\
0 & & \\
& \ddots & 0 \\
0 & & \\
\end{pmatrix}.
\]

If \(A := (a(\phi_j, \phi_i))\) is the matrix representing the bilinear form \(a := \sum_{e \in \Pi^h(\Omega)} a_e\) in the nodal basis \(\Phi = \{\phi_1, \ldots, \phi_{N^h}\}\) of \(X^h_0(\Omega)\), then

\[
A = P^F_R A_R P^R_F + P^F_B A_B P^B_F,
\]

and the iteration operator induced by Algorithm 2.1 is given by

\[
u^{n+2} = P^B_F T_\mu(A_B) P^B_R T_\mu(A_R) P^R_F u^n,
\]

where

\[
T_\mu(z) = (\mu + z)^{-1}(\mu - z).
\]

The main result of this paper is Theorem 2.1 below, which states that Algorithm 2.1 converges robustly.

**Theorem 2.1.** Let \(X^h,p_0(\Omega)\) denote the space of \(p\)th order nonconforming elements of [26], \(p \geq 1\). Let \(\Pi^h(\Omega)\) be two-colorable and suppose the minimum angle in \(\Pi^h(\Omega)\) is bounded away from zero. Suppose \(\mu = \mu_0 \min_e h_e\), where \(\mu_0\) is a constant and \(h_e\) denotes the diameter of the element \(e\). Suppose (2.7a) holds; (2.7b) holds when \(p > 1\) and also

\[
0 \leq \epsilon_{\min} \leq \epsilon_{\max} \leq 1, \quad 0 \leq |\delta|_{\infty, \Omega} \leq O(h).
\]

Then, there is a constant \(C > 0\) depending only upon \(g, b, \mu_0\) and the smallest angle of the triangulation, such that, for \(h > 0\) small enough,

\[
|P^F_B T_\mu(A_B) P^B_R T_\mu(A_R) P^R_F| \leq 1 - C \min_e h_e.
\]

**Remark 2.5.** The proof of Theorem 2.1 uses only general features of piecewise polynomial finite element spaces such as \(X^h_{0,p}(\Omega)\). As a result, it remains valid, with only minor and local modifications to its proof, when \(X^h(\Omega)\) consists of the conforming elements of any degree \(p\).
3. Proof of Theorem 2.1. The proof of Theorem 2.1, given in this section, involves three ingredients. The first ingredient is the crucial inequality

\[(3.1) \quad |P_{T_\mu}(A_B)P_{T_\mu}(A_R)P_{T_\mu}(A_F)| \leq \max\{|T_\mu(A_e)| : e \in \Pi^h(\Omega)\},\]

proved in Theorem 3.1 later on and based upon results of [22] and a variant of a technical estimate by Douglas and Pearcey [9]. This is a nontrivial result since \(|T_\mu(A_R)| = |T_\mu(A_B)| = 1\) while \(|T_\mu(A_e)| < 1\) for each element \(e\). The second is the following estimates for the eigenvalues of the symmetric part \(H_e\) and skew-symmetric part \(S_e\) of the element matrix \(A_e:\)

\[\begin{align*}
(3.2a) \quad & \lambda_{\min}(H_e) \geq C_0^{-1} h_e^2, \\
(3.2b) \quad & \lambda_{\max}(H_e) \leq C_0(\epsilon_{\max,e} + \kappa_e + h_e + h_e^2), \\
(3.2c) \quad & |S_e| \leq C_0 h_e,
\end{align*}\]

where \(C_0 > 0\) depends only upon \(g, b, \\mu, \) and the minimum angle of the triangulation, and \(h_e\) denotes the diameter of \(e\). For completeness, (3.2) is proven in the Appendix. The third ingredient involves manipulation of the quotient representing \(|T_\mu(A_e)|\) and applying (3.1) and (3.2) in the manner which now follows.

Let \(x \in \mathbb{R}^\alpha \setminus \{0\}\) (or, resp., \(\mathbb{R}^{\alpha-p} \setminus \{0\}\) if \(e\) has an edge on \(\Gamma_0\)) be fixed and define \(v = (\mu I + A_e)^{-1} x\). Then, with \(\langle \cdot, \cdot \rangle\) denoting the canonical inner product of \(\mathbb{R}^\alpha\) (resp., \(\mathbb{R}^{\alpha-p}\)),

\[|T_\mu(A_e)x|^2 = \langle (\mu I - A_e)(\mu I + A_e)^{-1} x, (\mu I - A_e)(\mu I + A_e)^{-1} x \rangle \]

\[= \langle (\mu I - A_e)v, (\mu I - A_e)v \rangle = \mu^2|v|^2 - 2\mu \langle v, A_e v \rangle + |A_e v|^2 \]

\[= 1 - 4\mu\gamma_e(v), \]

where

\[\gamma_e(v) := \frac{\langle v, A_e v \rangle}{\mu^2|v|^2 + 2\mu \langle v, A_e v \rangle + |A_e v|^2}. \tag{3.4}\]

As \(\mu = \mu_0 \min_h h_e\), Theorem 2.1 follows from (3.1), (3.3) and \((1 - 4C_\mu \min_h h_e)^{1/2} \leq 1 - 2C_\mu \min_h h_e\) (\(C_0\) as defined in (3.2c)) provided that

\[\inf_{v \neq 0} \gamma_e(v) \geq C > 0, \tag{3.5}\]

where \(C\) depends only upon \(g, b, \mu_0,\) and the smallest angle of the triangulation.

Consider therefore \(\gamma_e(v)\). Note that \(\langle v, A_e v \rangle = \langle v, H_e v \rangle\) and

\[|A_e v|^2 \leq (|H_e v| + |S_e v|)^2 \leq 2|H_e v|^2 + 2|S_e v|^2 \leq 2|H_e v|^2 + 2C_0^2 h_e^2 |v|^2,\]

where (3.2c) was used to bound \(|S_e v|\). Inserting these into (3.4) gives

\[\gamma_e(v) \geq \frac{\langle v, H_e v \rangle}{\mu^2|v|^2 + 2\mu \langle v, H_e v \rangle + 2|H_e v|^2 + 2C_0^2 h_e^2 |v|^2}. \]
Therefore, (3.5) is verified provided that
\[ w \quad \text{hence, by (3.2a), we have for} \quad h > 0 \text{ small enough (and since } 0 \leq h \leq h_0 \text{)} \]
\[ \text{we may take any suitable } w \text{ such that} \quad \langle w, w \rangle = \min \{ C, e^{-h_0^2} \} \text{.} \]
\[ \text{In particular, the function} \quad \lambda \mapsto \lambda / (\mu_0 h_0 + \lambda)^2 \text{ is increasing in } [0, \lambda_{\text{min}}] \text{, and hence, by (3.2a), we have for } h > 0 \text{ small enough (and since } 0 \leq h \leq h_0 \text{)} \]
\[ \lambda_{\text{min}} (\mu_0 h_0 + \lambda_{\text{min}})^2 \geq \frac{C_0^{-1} h_0^2}{(\mu_0 h_0 + C_0^{-1} h_0^2)^2} = \frac{1}{C_0(\mu_0 + C_0^{-1} h_0^2)} \geq \frac{1}{4C_0^2} \mu_0^2. \]

Suppose now that the minimum in (3.7) is achieved for \( \lambda = \lambda_{\text{max}} \). As observed above, this happens only if \( \lambda_{\text{max}} \geq \mu_0 h_0 \). In particular, the function \( \lambda \mapsto \lambda / (\mu_0 h_0 + \lambda)^2 \) is decreasing in \( [\lambda_{\text{max}}, \infty) \). By (3.2b) we have \( \lambda_{\text{max}} \leq C_0(\epsilon_{\text{max}}, e + \delta + h_0^2) \leq C_0(1 + \delta + h_0^2) \) since \( 0 \leq e \leq 1 \), whence \( \lambda_{\text{max}} \leq 2C_0 \) for \( h > 0 \) small enough since \( |\delta|_{\infty, 1} = O(h) \) and \( 0 \leq h \leq h_0 \). Thus,
\[ \frac{\lambda_{\text{max}} (\mu_0 h_0 + \lambda_{\text{max}})^2}{(\mu_0 h_0 + 2C_0 h_0^2)^2} \geq \frac{1}{8C_0}, \]

since \( \mu_0 h_0 \leq 2C_0 \) for \( h > 0 \) small enough. By combining (3.8) and (3.9), we see that \( C \) in (3.7) may be chosen as
\[ C = \min \left\{ \frac{1}{4C_0^2 h_0^2}, \frac{1}{8C_0} \right\} \]

To complete the proof of Theorem 2.1, it remains to establish the validity of inequality (3.1). In this aim, we first recall some concepts and results from [22].
Let $V$ denote a finite element space and $\Phi = \{\phi_1, \ldots, \phi_N\}$ a basis of $V$. A bilinear form $b : V \times V \to \mathbb{R}$ is said to be $\Phi$-coercive if $b(v, v) \geq 0 \forall v \in V$, and the following condition holds: if $v = \sum_{i=1}^{n} x_i \phi_i$ and $b(v, v) = 0$, then for each index $1 \leq i \leq N$, one has either $x_i = 0$ or $b(\phi_i, \cdot) = 0$.

**Lemma 3.1.** Let $\Phi^h = \{\phi_1, \ldots, \phi_N\}$ denote the nodal basis of the space $X^h_0(\Omega)$. Then, each bilinear form $a_e$ (see (2.4)) is $\Phi^h$-coercive.

**Proof.** For $v \in X^h_0(\Omega)$, we have $a_e(v, v) = \sigma(v, v) + \beta_e(v, v) \geq 0$ by (2.4), (2.6), (2.7) and Lemma A.2. Now, let $v = \sum_{i=1}^{N^h} x_i \phi_i$ and suppose $a_e(v, v) = 0$. Assume first that $e$ has no edge lying on $\Gamma_0$ and let $N_1, N_2, \ldots, N_\alpha$ be the nodes associated with the element $e$. For $i \notin \{N_1, N_2, \ldots, N_\alpha\}$, we have $\phi_{i_e} = 0$ and hence $a_e(\phi_i, \cdot) = a_e(\cdot, \phi_i) = 0$ from (2.4)–(2.6). Thus, the hypothesis $a_e(v, v) = 0$ also reads $\sum_{k,t=1}^{N_h} x_{N_h} x_{N_h} a_e(\phi_{N_1}, \phi_{N_1}) = 0$; i.e., with $x = (x_{N_1}, x_{N_2}, \ldots, x_{N_\alpha}) \in \mathbb{R}^\alpha$ and (2.8),

$$\langle A_e x, x \rangle = 0.$$

But $\langle A_e x, x \rangle = \langle H_e x, x \rangle$ and $H_e$ is positive definite (e.g., by (3.2a)), whence $x = 0$. In summary, we have $a_e(\phi_i, \cdot) = 0$ for $i \notin \{N_1, N_2, \ldots, N_\alpha\}$ and $x_i = 0$ if $i \in \{N_1, N_2, \ldots, N_\alpha\}$, as required for $\Phi^h$-coercivity. If $e$ has one edge lying on $\Gamma_0$, there are $\alpha - p$ nodes $N_1, \ldots, N_{\alpha - p}$ associated with $e$, and the above arguments can be repeated with $\{N_1, N_2, \ldots, N_{\alpha - p}\}$ replacing $\{N_1, N_2, \ldots, N_\alpha\}$. \(\Box\)

As a result of $\Phi^h$-coercivity of the element bilinear forms $a_e$ together with the relation

$$\sum_{e \in \Pi^h(\Omega)} a_e(v, v) > 0 \quad \forall \ v \in X^h_0(\Omega) \setminus \{0\}$$

(obvious from (2.4)–(2.6)), Lemma 3.2 follows directly from [22, Theorem 3.1].

**Lemma 3.2.** The block diagonal matrices $A_R$ and $A_B$ have the following properties:

(i) $\mathbb{R}^{N_h} = \ker A_{R/B} \oplus \text{rgy} A_{R/B}$ (orthogonal direct sum),

(ii) $\text{rgy} A_{R/B} P_{R/B} \subset \ker P_{B/A} B_{R/B} P_{F/B}$.

**Remark 3.1.** Part (i) of Lemma 3.2 also follows from the definition of $A_R$ and $A_B$ and the remark that the nonzero diagonal blocks $A_e$ in $A_R$ or $A_B$ are invertible. Note also that (i) also holds with $A_{R/B}$ replaced by $P_{R/B} A_{R/B} P_{R/B}$ since $P_{R/B}$ is orthogonal with inverse $P_{R/B}$. In contrast, (ii) is not true with $P_{R/B} A_{R/B} P_{R/B}$ replaced by $A_{R/B}$. \(\Box\)

As a final preliminary step, we need the following variant of a result by Douglas and Peacry [9, Lemma 3.2].

**Lemma 3.3.** For $i = 1, 2$, let $T_i \in \mathcal{L}(\mathbb{R}^N)$. Suppose that there are spaces $M_i, i = 1, 2$, such that

(i) $T_{i|M_i} = I$, $T_i(M_i) \subset M_i$,

(ii) $M_1 \perp M_2$.

Set

$$\nu_i := \|T_{i|M_i}\| \quad \text{(euclidian norm)}.$$

Then,

$$|T_1 T_2| \leq \max\{\nu_1, \nu_2, \nu_1 \nu_2\}.$$
Proof. Let \( x \in \mathbb{R}^N \) be written as \( x = x_2 + x_2^\perp \) according to the orthogonal splitting \( \mathbb{R}^N = M_2 \oplus M_2^\perp \). Since \( T_2|_{M_2} = I \), we have \( T_2x = x_2 + y_2 \) with \( y_2 := T_2x_2^\perp \in M_2^\perp \) (because \( T_2(M_2^\perp) \subset M_2^\perp \)).

From the above, \( T_1T_2x = T_1x_2 + T_1y_2 \). Write \( y_2 = y_{21} + y_{21}^\perp \) according to the orthogonal splitting \( \mathbb{R}^N = M_1 \oplus M_1^\perp \). Note that the same decomposition for \( x_2 \) yields \( x_2 = x_2^\perp_2 \) because \( x_2 \in M_2 \) and \( M_2 \perp M_1 \) by hypothesis. Thus, \( T_1T_2x = T_1y_{21} + T_1(x_2^\perp + y_{21}^\perp) = y_{21} + T_1(x_2^\perp + y_{21}^\perp) \) since \( T_1(M_1) = I \). As a result,

\[
|T_1T_2x|^2 = |y_{21}|^2 + |T_1(x_2^\perp + y_{21}^\perp)|^2,
\]

because \( T_1(M_1^\perp) \subset M_1^\perp \), and hence

\[
|T_1T_2x|^2 \leq |y_{21}|^2 + \nu_1^2|x_2^\perp + y_{21}^\perp|^2 = (1 - \nu_1^2)|y_{21}|^2 + \nu_1^2(|x_2^\perp + y_{21}^\perp|^2 + |y_{21}|^2).
\]

Since \( x_2^\perp = x_2 \) and \( y_{21} = y_{21} + y_{21}^\perp \), this also reads

\[
|T_1T_2x|^2 \leq (1 - \nu_1^2)|y_{21}|^2 + \nu_1^2|x_2 + y_{21}|^2 = (1 - \nu_1^2)|y_{21}|^2 + \nu_1^2(|x_2|^2 + |y_{21}|^2),
\]

because \( x_2 \) and \( y_{21} \) are orthogonal, as noticed earlier. By using \( |y_{21}|^2 = |y_{21}|^2 + |y_{21}^\perp|^2 \), this yields

\[
|T_1T_2x|^2 \leq |y_{21}|^2 + \nu_1^2|x_2|^2 + \nu_1^2|y_{21}^\perp|^2 \leq \nu_1^2|x_2|^2 + \max\{1, \nu_1^2\}|y_{21}|^2.
\]

Finally, from \( |y_{21}| = |T_2x_2^\perp| \leq \nu_2|x_2^\perp| \), we find

\[
|T_1T_2x|^2 \leq \nu_1^2|x_2|^2 + \nu_2^2\max\{1, \nu_1^2\}|y_{21}|^2 \leq \max\{\nu_1, \nu_2, \nu_1\nu_2\}|x|^2,
\]

and the conclusion follows. \( \Box \)

Remark 3.2. In [9], Douglas and Peaczy assume that \( T_1 \) and \( T_2 \) of Lemma 3.3 are contractions for the euclidian norm. On the other hand, they do not assume that \( M_1 \perp M_2 \) and give an estimate for \( |T_1T_2| \) depending upon the angle between \( M_1 \) and \( M_2 \). By using Lemma 3.2 instead of Douglas and Peaczy’s estimate in the proof of Theorem 3.1 below, we shall avoid redundancies by not having to check the contraction hypothesis. \( \Box \)

Theorem 3.1. Inequality (3.1) holds; i.e.,

\[
|P^F_{R}T_\mu(A_R)P^R_{\mu}(A_R)P^R_F| \leq \max\{|T_\mu(A_e)|: e \in \Pi^h(\Omega)\}.
\]

Proof. For notational convenience, set

\[
A_R = P^F_{\mu}A_RP^F_F, \quad A_B = P^F_BA_BP^F_F.
\]

Since \( P^F_F = (P^F_{\mu})^{-1} \) and \( P^F_B = (P^F_F)^{-1} \), we have

\[
P^F_{R}T_\mu(A_R)P^R_F = T_\mu(A_R), \quad P^F_B T_\mu(A_B)P^R_F = T_\mu(A_B),
\]

and by using \( P^F_R = P^F_B P^F_F \) (meaning that to pass from the “red” to the “black” ordering, we may first pass from “red” to “fixed” and next from “fixed” to “black”) it follows that

\[
P^F_{R}T_\mu(A_B)P^R_{R}T_\mu(A_R)P^R_F = T_\mu(A_B)T_\mu(A_R).
\]
Thus, the inequality (3.1) to be proved here also reads

\[(3.11) \quad |T_\mu(\tilde{A}_B)T_\mu(\tilde{A}_R)| \leq \max\{|T_\mu(A_e)| : e \in \Pi^h(\Omega)\}.
\]

In Lemma 3.3, choose \(T_1 := T_\mu(\tilde{A}_B), T_2 := T_\mu(\tilde{A}_R),\) and \(M_1 := \ker \tilde{A}_B, M_2 := \ker \tilde{A}_R.\) Then, \(T_\mu(\tilde{A}_{R/B})_{|\ker \tilde{A}_{R/B}} = I\) and \(T_\mu(\tilde{A}_{R/B})_{|\text{rge} \tilde{A}_{R/B}} \subset \text{rge} \tilde{A}_{R/B}.\) Thus, \(T_1\) and \(T_2\) satisfy condition (i) of Lemma 3.3 by Lemma 3.2 (i) and Remark 3.1. That condition (ii) of Lemma 3.3 also holds follows from Lemma 3.2 (ii). Therefore, we have

\[(3.12) \quad |T_\mu(\tilde{A}_B)T_\mu(\tilde{A}_R)| \leq \max\{\nu_B, \nu_R, \nu_B\nu_R\},
\]

where

\[\nu_R/B := |T_\mu(\tilde{A}_{R/B})_{|\text{rge} \tilde{A}_{R/B}}|.
\]

By the orthogonality of the permutation matrices \(P_{B/R}^{F} \) and \(P_{B/R}^{B} = (P_{B/R}^{F})^{-1}, \nu_B/R\)

above is also given by

\[(3.13) \quad \nu_R/B = |T_\mu(\tilde{A}_{B/R})_{|\text{rge} \tilde{A}_{R/B}}|.
\]

The matrix \(A_R\) is block diagonal with nonzero blocks \(A_e\) where \(e \in \Pi^h(\Omega)\) ranges over all the “red” elements. Since these blocks are invertible (e.g., by (3.2a)), \(\text{rge} A_R\) is exactly the direct product of the spaces \(\text{rge} A_e \simeq \mathbb{R}^a\) or \(\mathbb{R}^{a-p}\) as \(e\) runs over all the “red” elements. Together with the block diagonal structure of \(A_R\), this makes it obvious that \(\nu_R\) in (3.13) is the maximum of \(|T_\mu(A_e)|\) as \(e\) runs over all the “red” elements. From either the estimate (3.3) (clearly independent of (3.1) being valid or not) or “Kellogg’s Lemma” in [16], we infer that \(\nu_R < 1.\) Replacing “red” by “black” in the foregoing arguments, and since each element \(e \in \Pi^h(\Omega)\) is either “red” or “black,” we infer that

\[\max\{\nu_R, \nu_B\} = \max\{|T_\mu(A_e)| : e \in \Pi^h(\Omega)\} < 1,
\]

whence \(\max\{\nu_R, \nu_B, \nu_B\nu_R\} = \max\{\nu_R, \nu_B\} = \max\{|T_\mu(A_e)| : e \in \Pi^h(\Omega)\},\) and the validity of (3.1) now follows from (3.13).

Remark 3.3. The only place in the proof of convergence in which the particular subdivision of \(\Omega\) is used is in the eigenvalue estimates (3.2a,b,c). However, if \(D\) is any subdomain, completely analogous estimates hold for the stiffness matrices on \(D\) (replace \(A_e, H_e, S_e\) with \(A_D, H_D, S_D\) in (3.2)), for conforming and nonconforming elements. In fact, the proof of (3.2) in all these cases is hardly modified from that given in the appendix in the case \(D = \epsilon.\) Thus, it follows easily and directly that the appropriately modified variant of Algorithm 2.1 converges robustly with conforming or nonconforming elements and larger subdomains.

4. Numerical experiments. This section illustrates the robustness of the algorithm on several test problems with \(\epsilon = \epsilon I\) and a quasi-uniform mesh. Accordingly, we give a test of the method varying \(\epsilon\) and \(h\) separately, thereby illustrating both the uniformity in \(\epsilon\) and the \(O(h^{-1})\) estimate on the number of iterations.

We have selected the “rotating pulse” problem for our tests because it captures some of the interesting features of internal flow problems, such as stagnation points,
where \( \varphi \) parameter \( \delta \) slope \( \pm /h \) mesh of 1 poor for the iterations required for convergence. At first sight the method appears to be \( \epsilon \) choice for small \( \epsilon \) as \([26]\) concerning the \( \mu \) the relaxation parameter \( \epsilon \) for a uniform mesh approximation. The final approximate solution obtained is depicted in Figure 4.1. Note that solution quality is good highly locally refined meshes \([24, 25]\) is illustrated at length in experiments in \([18, 19, 20, 21, 22, 26]\).

The following table for the case \( p = 1 \) further illustrates that a good choice of \( \mu_0 \) for \( \epsilon \) small may not be a good one for the slightly nonsymmetric case of \( \epsilon = O(1) \). This is consistent with the possibility advanced in Remark 2.1 that \( \mu_0 \) should be increased as \( \epsilon \) increases.

The case \( p = 1 \) of nonconforming linears also illustrates a subtle point developed in \([26]\) concerning the consistency error in nonconforming approximations. Specifically, the consistency error for small \( \epsilon \) for \( p = 1 \) suffers from a global deterioration to \( O(1) \)

<table>
<thead>
<tr>
<th>( \epsilon \backslash h )</th>
<th>( h = \frac{1}{8} )</th>
<th>( h = \frac{1}{16} )</th>
<th>( h = \frac{1}{32} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 10^{-6} )</td>
<td>1480</td>
<td>2160</td>
<td>2640</td>
</tr>
<tr>
<td>( 10^{-4} )</td>
<td>140</td>
<td>520</td>
<td>270</td>
</tr>
<tr>
<td>( 10^{-2} )</td>
<td>30</td>
<td>40</td>
<td>50</td>
</tr>
<tr>
<td>( 10^{-4} )</td>
<td>20</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>( 10^{-6} )</td>
<td>20</td>
<td>20</td>
<td>40</td>
</tr>
</tbody>
</table>
Table 4.2
Iteration counts for two different $\mu_0$'s ($\mu_0 = 32.5/\mu_0 = .25$) for the case $p = 1$ of nonconforming linears.

<table>
<thead>
<tr>
<th>$\epsilon \setminus h$</th>
<th>$h = \frac{1}{8}$</th>
<th>$h = \frac{1}{16}$</th>
<th>$h = \frac{1}{32}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon = 10^{-2}$</td>
<td>20/110</td>
<td>40/190</td>
<td>80/320</td>
</tr>
<tr>
<td>$\epsilon = 10^{-3}$</td>
<td>60/20</td>
<td>100/30</td>
<td>180/60</td>
</tr>
<tr>
<td>$\epsilon = 10^{-4}$</td>
<td>70/10</td>
<td>120/20</td>
<td>220/30</td>
</tr>
<tr>
<td>$\epsilon = 10^{-5}$</td>
<td>70/10</td>
<td>130/20</td>
<td>220/30</td>
</tr>
<tr>
<td>$\epsilon = 10^{-6}$</td>
<td>70/10</td>
<td>130/20</td>
<td>220/30</td>
</tr>
</tbody>
</table>

Fig. 4.1. Approximate solution to rotating pulse problem with nonconforming linears ($p = 1$).

which is not experienced for $p > 1$; compare the solution quality of Figure 4.2 ($p = 1$) to Figure 4.1 ($p = 3$).

As a last remark, we note that the very low iteration counts needed for the parallel finite element method are not an artifact of the regularizing effect of the streamline diffusion discretization. Indeed, when the linear system arising from the usual centered Galerkin method ($\delta \equiv 0$) is solved to the same tolerance $Tol = 10^{-4}$ only 30 iterations are required (when $\epsilon = 10^{-4}, h = \frac{1}{32}$, and $p = 3$)—the same number as for the streamline diffusion method. (Solution quality is naturally not expected to be satisfactory for the centered method.)

Appendix A. In this Appendix, the three estimates (3.2) of $\lambda_{\min}(H_e), \lambda_{\max}(H_e)$, and $|S_e|$ are proven. The method of proof follows well-worn paths in finite element theory; see, for example, Axelsson and Barker [4], Axelsson [3], or Fried [11] for more information about these techniques.

For $e \in \Pi_h(\Omega)$ a generic triangle let $X_0^{h,p}(e)$ denote $X_0^{h,p}(\Omega)_e$ and let $h_e$ be the diameter of $e$. Since $\Pi_h(\Omega)$ is assumed to satisfy the minimum angle condition, the
usual inverse and norm-equivalence estimates hold for $X_0^{k,p}(e)$. We record them in Proposition A.1 below for use in this section.

**Proposition A.1.** Let the minimum angle in $\Pi^h(\Omega)$ be bounded away from zero. Given an element $e \in \Pi^h(\Omega)$, let $\phi_{N_i}, 1 \leq i \leq \alpha$ (or $1 \leq i \leq \alpha - p$ if $e$ has an edge on $\Gamma_0$), be the nodal basis functions associated with the nodes in $e \setminus \Gamma_0$. There is a constant $C_p \geq 1$ depending only upon the smallest angle of the triangulation such that, for any $\psi = \sum_{i=1}^\alpha a_i \phi_i \in X_0^{k,p}(e)$ we have

\[
(A.1) \quad |\psi|_{1,e} \leq C_p h_e^{-1} |\psi|_{0,e}, \quad |\Delta \psi|_{0,e} \leq C_p h_e^{-1} |\psi|_{1,e},
\]

and

\[
(A.2) \quad \frac{1}{C_p} h_e \left( \sum_{i=1}^\alpha a_i^2 \right)^{\frac{1}{2}} \leq |\psi|_{0,e} \leq C_p h_e \left( \sum_{i=1}^\alpha a_i^2 \right)^{\frac{1}{2}}.
\]

Recall that $\epsilon_{\text{max},e}$ and $\epsilon_{\text{min},e}$ are given by

\[
\epsilon_{\text{max},e} = \frac{\epsilon_{\text{max},e}}{\epsilon_{\text{min},e}} = \sup_{(x,y) \in e} \frac{1}{\inf(\lambda_{\text{max},e}/\epsilon(x,y))}
\]

and that $\epsilon_{\text{max}} = \max_e \epsilon_{\text{max},e}, \epsilon_{\text{min}} = \min_e \epsilon_{\text{min},e}$. Further, recall $|\epsilon|_{\infty,e} = \epsilon_{\text{ess}} \sup\{|\epsilon|_{\infty}(x) : x \in e\}$, where $|\cdot|_{\infty}$ is the $\ell_\infty$ matrix norm. We will also need an $\infty$-like norm on the 3-tensor $\nabla \epsilon$. Let $\gamma = \gamma_{ijk}$ be a 3-tensor and define the norm $|\gamma|_{\gamma} := \max\{\sum_{i,j=1}^2 \gamma_{ij} w_i : (w_1, w_2) \in \mathbb{R}^2, w_1^2 + w_2^2 = 1\}$. We shall denote by $|\epsilon|_{\infty,e} := \epsilon_{\text{ess}} \sup\{|\nabla \epsilon|_{\gamma}(x) : x \in e\}$ and $|\nabla \epsilon|_{\gamma} := \max_e |\epsilon|_{\infty,e}$.

**Lemma A.1.** Let $\delta \geq 0$ with $|\delta|_{\infty,\Omega}$ small enough that $0 < \frac{1}{2} (g - \frac{1}{2} \nabla \cdot b)_{\text{min}} \leq \tilde{g} \leq 2 (g - \frac{1}{2} \nabla \cdot b)_{\text{max}}$ (see (2.7) for the definition of $\tilde{g}$) and $\frac{1}{2} \leq 1 - |\delta g|_{\infty,\Omega} \leq 2$. Then, for every $e \in \Pi^h(\Omega)$ and $u, v \in X_0^{k,p}(e)$, we have

\[
(A.3a) \quad \sigma_e(u, u) \geq \frac{1}{2} \left( g - \frac{1}{2} \nabla \cdot b \right)_{\text{min}} |u|_{0,e}^2 + \epsilon_{\text{min},e} |u|_{1,e}^2 + \delta |b \cdot \nabla u|_{0,e}^2.
\]
(A.3b) \[ |\sigma_e(u, v)| \leq \epsilon_{\text{max}, e} |u|_{1, e} |v|_{1, e} + \delta_e |b \cdot \nabla u|_{0, e} |b \cdot \nabla v|_{0, e} + 2 \left( g - \frac{1}{2} \nabla \cdot b \right)_{\text{max}, e} |u|_{0, e} |v|_{0, e} \]

(A.3c) \[ |\vec{\sigma}_e(u, v)| \leq |b|_{\infty, e}(|u|_{1, e} |v|_{0, e} + |u|_{0, e} |v|_{1, e}) \]

and, provided additionally that (A.1) holds,

(A.3d) \[ |\beta_e(u, v)| \leq \delta_e (|\epsilon|_{\infty, e} C_p h_e^{-1} + |\nabla \epsilon|_{\infty, e}) |u|_{1, e} |b \cdot \nabla v|_{0, e}. \]

**Proof.** The proof is trivial. \( \Box \)

**Lemma A.2.** Let \( \Pi^h(\Omega) \) satisfy an angle condition and suppose that, on each element \( e \), \( \delta_e = \delta_{0, e} h_e \) where \( 0 \leq \delta_{0, e} \) if \( p = 1 \) and if \( p \geq 2 \), \( \delta_{0, e} \) is chosen small enough that

(A.3e) \[ 0 \leq \delta_{0, e} \leq \frac{h_e \epsilon_{\text{min}, e}}{2 C_p^2 |\epsilon|_{\text{loc}, e}^2 + h_e^2 |\nabla \epsilon|_{\text{loc}, e}^2}. \]

Then, for all \( u \in X_0^{h, p}(e) \),

\[ \sigma_e(u, u) + \beta_e(u, u) \geq \frac{1}{2} \epsilon_{\text{min}, e} |u|_{1, e}^2 + \frac{1}{2} \delta_e |b \cdot \nabla u|_{0, e}^2 + \frac{1}{2} \left( g - \frac{1}{2} \nabla \cdot b \right)_{\text{min}, e} |u|_{0, e}^2. \]

Using (A.3a) and the above then gives for \( u \in X_0^{h, p}(e) \)

\[ |\beta_e(u, v)| \leq \frac{\epsilon_{\text{min}, e}}{2} |u|_{1, e}^2 + \frac{2}{\epsilon_{\text{min}, e}} (\delta_e^2 |\epsilon|_{\text{loc}, e}^2 C_p^2 h_e^{-2} + \delta_e^2 |\nabla \epsilon|_{\text{loc}, e}^2)|b \cdot \nabla v|_{0, e}^2. \]

The result now follows from (A.3f) by requiring that

\[ 1 - \delta_e(|\epsilon|_{\text{loc}, e}^2 h_e^{-2} \epsilon_{\text{min}, e}^{-1} C_p^2 + |\nabla \epsilon|_{\text{loc}, e}^2 \epsilon_{\text{min}, e}^{-1}) \delta_e > \frac{1}{2}. \] \( \Box \)

**Lemma A.3.** Suppose the hypotheses of Lemmas A.1 and A.2 hold. Then,

(A.4) \[ \lambda_{\text{max}}(H_e) \leq C_p^4 (\epsilon_{\text{max}, e} + \epsilon_{\text{min}, e} \delta_e h_e^{-1} C_p |b|_{\text{loc}, e} + \delta_e |\nabla \epsilon|_{\text{loc}, e} |b|_{\text{loc}, e} + \delta_e |b|_{\text{loc}, e}) + 2 C_p^2 h_e^2 \left( g - \frac{1}{2} \nabla \cdot b \right)_{\text{max}, e} \]

(A.5) \[ \lambda_{\text{min}}(H_e) \geq h_e^2 \beta \left( g - \frac{1}{2} \nabla \cdot b \right)_{\text{min}, e}, \]
where \( C_p \geq 1 \) is the constant obtained in Proposition A.1.

Proof. For the sake of argument, suppose first that no edge of \( e \) lies on \( \Gamma_0 \). Let \( \lambda \in \text{Sp}(\mathbf{H}_0) \) be associated with the eigenvector \( \mathbf{a} = (a_1, a_2, \ldots, a_\alpha) \in \mathbb{R}^3 \) and set \( \psi = \sum_{i=1}^\alpha a_i \phi_i \in X_0^{h,p}(e) \), where \( \phi_i \), \( i = 1, 2, \ldots, \alpha \), are the nodal basis functions associated with the nodes in \( e \). We have \( \lambda = \mathbf{a}^T \mathbf{H}_0 \mathbf{a} / |\mathbf{a}|^2 = [\sigma_\epsilon(\psi, \psi) + \beta_\epsilon(\psi, \psi)] / |\mathbf{a}|^2 \) and hence, from (A.3b,d) and the second inequality in (A.2),

\[
\lambda \leq \frac{C^2 h^2 \sigma_\epsilon + \beta_\epsilon(\psi, \psi)}{|\psi|^2_{0,e}}
\]

\[
\leq \frac{C^2 h^2}{|\psi|^2_{0,e}} \left\{ \epsilon_{\text{max},e} |\nabla \psi|_{0,e}^2 + \delta_\epsilon |\mathbf{b}|_{\infty}^2 \right\}
\]

\[
+ \frac{C^2 h^2}{|\psi|^2_{0,e}} \left\{ \frac{\delta_\epsilon}{|\mathbf{b}|_{\infty}^2} \right\}
\]

\[
\leq C^4_p (\epsilon_{\text{max},e} + \delta_\epsilon |\mathbf{b}|_{\infty}^2) + C^2 h^2 \cdot 2 \left\{ |\nabla \psi|_{\infty}^2 + |\mathbf{b}|_{\infty}^2 \right\}
\]

where \( \epsilon_{\text{min},e} \leq h_e \) is assumed. Thus, (A.4) follows by substituting (A.1) into the above inequality. For the proof of (A.5), use (A.3g) and the first inequality in (A.2).

If one edge of \( e \) lies in \( \Gamma_0 \), the above arguments can be repeated verbatim with \( \mathbf{a} = (a_1, a_2, \ldots, a_{\alpha-1}) \in \mathbb{R}^{\alpha-1} \).

Lemma A.4. Suppose the hypotheses of Lemma A.2 hold. Then

\[
(A.6) \quad |\mathbf{S}_e| \leq h_e C^\prime, \quad C^\prime = |\mathbf{b}|_{\infty,e}(2 - \delta_\epsilon g|\infty, C^3_p + C^4_p),
\]

where \( C_p \geq 1 \) is the constant obtained in Proposition A.1.

Proof. Once again, assume first that \( e \) has no edge lying on \( \Gamma_0 \). We have \( |\mathbf{S}_e| = \sup_{\mathbf{a}, \mathbf{z} \in \mathbb{R}^\alpha \setminus \{0\}} \langle \mathbf{S}_e, \mathbf{a}, \mathbf{z} \rangle / |\mathbf{a}| |\mathbf{z}| \) where \( \langle \cdot, \cdot \rangle \) is the canonical inner product of \( \mathbb{R}^\alpha \). Let \( \mathbf{a} = (a_1, a_2, \ldots, a_\alpha) \), \( \mathbf{z} = (z_1, z_2, \ldots, z_\alpha) \) and (using the notation of the proof of Lemma A.2 above) set \( \psi_\mathbf{a} = \sum_{i=1}^\alpha a_i \phi_i \in X_0^{h,p}(e), \psi_\mathbf{z} = \sum_{i=1}^\alpha z_i \phi_i \in X_0^{h,p}(e). \) Then, using (A.2), (A.3d), and (A.3c)

\[
|\mathbf{S}_e| = \sup_{\mathbf{a}, \mathbf{z} \in \mathbb{R}^\alpha \setminus \{0\}} \frac{\delta_\epsilon \psi_\mathbf{a}, \psi_\mathbf{z}}{\mathbf{a}} + \frac{1}{2} (\beta_\epsilon \psi_\mathbf{a}, \psi_\mathbf{z} - \beta_\epsilon \psi_\mathbf{z}, \psi_\mathbf{a})
\]

\[
\leq \frac{C^2 h^2}{|\psi_\mathbf{a}|_{0,e}|\psi_\mathbf{z}|_{0,e}} \left\{ |\nabla \psi_\mathbf{a}|_{1,e}^2 \right\}
\]

\[
+ \delta_\epsilon |\mathbf{b}|_{\infty,e}(\epsilon_{\text{max},e} C_p h^{-1} + |\nabla \epsilon|_{\infty,e}) |\psi_\mathbf{a}|_{1,e} |\psi_\mathbf{z}|_{1,e}.
\]

The inverse assumption (A.1) now gives

\[
|\mathbf{S}_e| \leq 2 |\mathbf{b}|_{\infty,e}(1 - \delta_\epsilon g|\infty, C^3_p h^{-1} + \delta_\epsilon |\mathbf{b}|_{\infty,e} C^4_p (\epsilon_{\text{max},e} C_p h^{-1} + |\nabla \epsilon|_{\infty}).
\]

(The last term on the right-hand side is absent if \( p = 1 \) since the \( \beta_\epsilon(\cdot, \cdot) \) term is dropped from the variational formulation in this case.)
If $\epsilon_{\min,e} \geq h_e$ then $\delta_{e,0,e}$ is chosen to be zero on $e$, and the result holds trivially. Now suppose, on the contrary, that $\epsilon_{\min,e} < h_e$. Then,

$$
\delta_e |b|_\infty C_p^4 (\epsilon_{\text{max},e} C_p h_e^{-1} + |\nabla \epsilon|_\infty)
\leq \delta_{0,e,h_e} |b|_\infty C_p^4 (\epsilon_{\text{max},e} C_p h_e^{-1} + |\nabla \epsilon|_\infty)
\leq h_e |b|_\infty h_e^2 (2 C_p^4 \epsilon_{\text{max},e} C_p h_e^{-1} + h_e^2 |\nabla \epsilon|_\infty^2)
\leq h_e |b|_\infty C_p^4 \epsilon_{\min,e} h_e / h_e^2
\leq |b|_\infty C_p^4 h_e.
$$

Thus, for $\epsilon_{\text{max},e} \leq h_e$ in this case, it follows that

$$
|S_e| \leq h_e \{ 2 |b|_\infty |1 - \delta_e g|_\infty C_p^3 + |b|_\infty C_p^4 \},
$$

which is the desired result. If $e$ has one edge lying in $\Gamma_0$, repeat the above arguments with $a, z \in \mathbb{R}^{n-2} \setminus \{0\}$. □

**PROPOSITION A.2.** The estimates (3.2) hold, i.e.,

$$
\lambda_{\min}(H_e) \geq C_0^{-1} h_e^2,
\lambda_{\max}(H_e) \leq C_0 (\epsilon_{\text{max},e} + \delta_e + h_e + h_e^2),
|S_e| \leq C_0 h_e,
$$

where $C_0 > 0$ is a constant depending only upon $g, b$ and the smallest angle of the triangulation.

**Note.** From our standing assumption about the triangulation and Lemmas A.1 and A.2, we may indeed choose

$$
C_0 = \max \left\{ C_p^4, (|\epsilon|_\infty + \delta_e |b|_\infty^2)^2, C_2^2 \left( g - \frac{1}{2} \nabla \cdot b \right)_\text{max}, 2C_p^2 |b|_\infty \frac{2C_p^2}{(g - \frac{1}{2} \nabla \cdot b)_{\text{min}}}, 2 |b|_\infty |1 - \delta_e g|_\infty \cdot C_2^2 \right\},
$$

where $C_p > 0$ is the constant obtained in Proposition A.1, hence depending only upon the smallest angle of the triangulation. □

**Remark.** The Case of Larger Subdomains, Conforming Elements, Etc. Since the above bounds upon the subdomain stiffness matrix follow from simple properties of the piecewise polynomial finite element space which hold on arbitrary groups of elements, the spectral bounds of Proposition A.2 hold much more generally. For example, let $D$ be a subdomain consisting of a union of elements, $A_D, H_D, S_D$ be the stiffness matrix of $D$, and its hermitian and skew-hermitian parts. Then the following bounds hold by essentially the same argument:
\[ \lambda_{\min}(H_D) \geq C_0^{-1}(\min_{e \subset D} h_e)^2, \]
\[ \lambda_{\max}(H_D) \leq C_0 \left( e_{\max, D} + \max_{e \subset D} \delta_e + \max_{e \subset D} h_e + \left( \max_{e \subset D} h_e \right)^2 \right), \]
\[ |S_D| \leq C_0 \max_{e \subset D} h_e, \]

where the max's and min's in the expression for \( C_0 \) are taken over \( D \).

Acknowledgment. We thank the (unknown) referees for their careful reading and constructive comments which have improved the paper.

REFERENCES


