Optimal iterative solvers for linear nonsymmetric systems and nonlinear systems with PDE origins: Balanced black-box stopping tests

Pranjal, Prasad and Silvester, David J.

2018

MIMS EPrint: **2018.13**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/
And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

Optimal iterative solvers for linear nonsymmetric systems and nonlinear systems with PDE origins: Balanced black-box stopping tests

Pranjal* and David Silvester†

Abstract. This paper discusses the design of efficient algorithms for solving linear nonsymmetric systems and nonlinear systems associated with FEM approximation of elliptic PDEs. The novel feature of the designed linear solvers like GMRES, BICGSTAB(ℓ), TFQMR, and nonlinear solvers like Newton and Picard, is the incorporation of error control in the 'natural norm' in combination with an effective a posteriori estimator for the PDE approximation error. This leads to robust and optimal black-box stopping criteria: the iteration is terminated as soon as the algebraic error is insignificant compared to the approximation error.

Key words. FEM approximation of PDEs, a posteriori FEM error estimators, nonsymmetric linear systems, iterative solvers, GMRES, BICGSTAB(ℓ), TFQMR, Newton solvers, preconditioning.

AMS subject classifications. 65C20, 65N30, 65N15

1. Introduction and problem description.

Consider the following boundary value problem: find the solution $u(\vec{x}) : \Omega \to \mathbb{R}$ such that

$$\mathcal{L}(\vec{x})u(\vec{x}) = f(\vec{x}), \qquad \forall\, \vec{x} \in \Omega, \tag{1.1a}$$

$$\mathcal{B}(\vec{x})u(\vec{x}) = g(\vec{x}), \qquad \forall\, \vec{x} \in \partial\Omega. \tag{1.1b}$$

where $\Omega \subset \mathbb{R}^d\, (d = 1, 2, 3, \ldots)$ denotes the spatial domain and $\partial\Omega$ is the spatial boundary. Here $\mathcal{L}$ is the (possibly nonlinear) elliptic PDE operator, $\mathcal{B}$ denotes the boundary operator, f is the given (scalar) source term, g denotes boundary value, and u the true solution. Equation (1.1) is discretized here using finite element method (FEM) [2] and the corresponding linear or nonlinear discrete system is solved by a linear or a nonlinear iterative solver respectively. Linear iterative process is considered next.

1.1. Linear systems.

For chosen spatial discretization parameter h, FEM for solving (1.1) results in solving for a linear system, that is,

$$\mathcal{F}_h \mathbf{x}_h = \mathbf{b}_h \iff \mathcal{M}_h^{-1}\mathcal{F}_h \mathbf{x}_h = \mathcal{M}_h^{-1}\mathbf{b}_h, \tag{1.2}$$

where the matrix $\mathcal{M}_h$ is a preconditioner. The matrix $\mathcal{F}_h$ and the vector $\mathbf{b}_h$ are known quantities arising from the FEM process while the unknown algebraic solution $\mathbf{x}_h$ denotes the coordinate vector of the FEM solution u_h in the chosen FEM basis.

The numerical solution of (1.1) essentially involves two types of errors which are: approximation error ($\|u - u_h\|_{\mathcal{E}}$) and algebraic error ($\|u_h - u_h^{(k)}\|_{\mathcal{E}}$). Here $u_h^{(k)}$ denotes the FEM solution formed from the solution $\mathbf{x}_h^{(k)}$ at the kth step of the chosen iterative solver. Also, $\|\cdot\|_{\mathcal{E}}$ denotes the natural norm. Wathen [24] has observed that FEM approximation of a PDE endows the problem with a natural norm, which is determined by the specific approximation space. Typically, in FEM setting, the PDE approximation error and the algebraic error are measured in this natural norm.

1.1.1. Research objective.

The approximation error is fixed for chosen spatial discretization parameter. Solving iteratively the corresponding discrete linear(ized)

*M.E. department, University of Wisconsin-Madison, USA (pranjalprasad21@gmail.com).
†School of Mathematics, University of Manchester, UK (d.silvester@manchester.ac.uk).

system(s) to a very high accuracy is not desirable. This is because a highly accurate iterative solution may require too many iterations and simply waste computational resources without decreasing the approximation error. On the other hand, if the iterations are stopped too early the iterative solution will not be a good approximation to the exact solution. This paper attempts to address these issues by presenting optimal balanced black-box stopping tests in Krylov solvers [8, sections 11.3–11.4] for solving linear systems with PDE origins. This is an active research field; see [17, 20, 11, 14, 15].

1.1.2. Motivation. Typically, using an optimal balanced black-box stopping methodology would usually lead to huge computational savings and in any case would definitely rule out premature stopping of the chosen iterative solver.

1.2. Nonlinear systems. When FEM approximation to solve (1.1) results in a nonlinear discrete system, nonlinear solvers like Newton solvers etc., are used to solve them. Starting with a given initial guess $u_h^{(0)}$, nonlinear solvers typically construct a sequence of iterates $\{u_h^{(l+1)}\}$, $l = 0, 1, \cdots$ satisfying

$$u_h^{(l+1)} = u_h^{(l)} + \delta u_h^{(l)}. \tag{1.3}$$

The problem of optimal balanced black-box stopping of the nonlinear iterative solver is of interest here in the same manner as that for a linear solver. Also, the 'correction' term $\delta u_h^{(l)}$ at each nonlinear iterative step l requires solving a linear system for the basis coefficients of $\delta u_h^{(l)}$. So, optimal balanced black-box stopping criteria developed in linear solvers can be applied in solving the linear system arising at each nonlinear iterative step too.

Paper organization. The structure of this paper is as follows. The problem statement has already been discussed in section 1. The general solution methodology for linear and nonlinear systems is presented in sections 2 and 3 respectively. The optimal balanced black-box stopping tests are derived therein. In section 4, the main contribution of this work is highlighted while in section 5 optimal balanced black-box stopping tests are presented in GMRES [18] and suboptimal Krylov solvers of non-symmetric linear systems with PDE origins. In section 6, using the IFISS [4] toolbox in MATLAB, computational results illustrating the devised optimal balanced black-box methodology are presented and discussed for convection-diffusion equations and Navier–Stokes equations. An optimal balanced black-box stopping test in (nonlinear) Newton/Picard solver for solving the Navier–Stokes PDE is also presented therein. Conclusions are presented in section 7. Note that C and c denote generic constants throughout. Also, l denotes a nonlinear iterative step and k denotes a linear iterative step throughout this paper.

2. Solution methodology for linear solvers. For a given approximation, the approximation error is fixed. The triangle inequality at (linear solver) iteration k ($k = 0, 1, 2, \ldots$) gives the following decomposition of the total error,

$$\underbrace{\|u - u_h^{(k)}\|_{\mathcal{E}}}_{\text{Total error}} \leq \underbrace{\|u - u_h\|_{\mathcal{E}}}_{\text{Approximation error}} + \underbrace{\|u_h - u_h^{(k)}\|_{\mathcal{E}}}_{\text{Algebraic error}}. \tag{2.1}$$

The total error at iteration step k is nothing but the approximation error obtained from kth FEM iterate $u_h^{(k)}$ (which in turn is obtained from $\mathbf{x}_h^{(k)}$ whose components are

the coefficients in the FEM basis representation of $u_h^{(k)}$). Estimation of total errors (approximation error) and algebraic error is discussed next.

In FEM setting, the approximation error (and hence total errors) can be measured a priori or a posteriori [23]. A priori approximation error estimation usually requires the PDE solution to satisfy some regularity conditions which may not hold or/and may not be easily verifiable a priori. On the other hand, robust a posteriori approximation error estimation techniques are popular for driving the FEM procedure adaptively and are generally readily available in the sense that

$$c\,\eta_h \;\le\; \|u - u_h\|_{\mathcal{E}} \;\le\; C\eta_h, \quad \text{with } \frac{C}{c} \sim O(1). \tag{2.2}$$

where η_h denotes the a posteriori error estimate of the approximation error. Also, the algebraic error $\|u_h - u_h^{(k)}\|_{\mathcal{E}}$ can usually be expressed in terms of the iteration error $\mathbf{e}_h^{(k)} := \mathbf{x}_h - \mathbf{x}_h^{(k)}$ norm $\|\mathbf{e}_h^{(k)}\|_{E_h} := \sqrt{(\mathbf{e}_h^{(k)})^T E_h \mathbf{e}_h^{(k)}}$, that is,

$$c_1\|\mathbf{e}_h^{(k)}\|_{E_h} \;\le\; \|u_h - u_h^{(k)}\|_{\mathcal{E}} \;\le\; C_1\|\mathbf{e}_h^{(k)}\|_{E_h}, \quad \text{with } \frac{C_1}{c_1} \sim O(1). \tag{2.3}$$

Here E_h is a 'suitable' (see section 4 for more details) symmetric positive-definite matrix such that $\|\cdot\|_{E_h}$ indeed defines a norm.

Thus, if $\eta_h^{(k)}$, η_h, $\|\mathbf{e}_h^{(k)}\|_{E_h}$ are tight estimates of the total error (at iteration k), the approximation error, and the algebraic error (at iteration k) respectively, then in light of the above discussions (2.1) becomes

$$\eta_h^{(k)} \;\simeq\; \eta_h + \|\mathbf{e}_h^{(k)}\|_{E_h}, \qquad k = 0, 1, 2, \ldots. \tag{2.4}$$

The equivalence $\simeq$ in (2.4) follows from (2.2) and (2.3).

Remark 2.1. Notice from (2.4) that when the contribution of $\|\mathbf{e}_h^{(k)}\|_{E_h}$ to the sum is 'small' then $\{\eta_h^{(k)}\}$ would converge with some accuracy to (unknown but fixed) η_h and one stops 'optimally', that is, the total error a posteriori error estimate cannot be significantly reduced further. Thus, the iterative strategy here can be looked upon as constructing a sequence $\{\eta_h^{(k)}\}$ converging to η_h.

Thus, one would stop optimally when $\|\mathbf{e}_h^{(k)}\|_{E_h}$ 'balances' $\eta_h^{(k)}$ in the sum (2.4), that is, stop at the first iteration k^* such that

$$\|\mathbf{e}_h^{(k^*)}\|_{E_h} \;\le\; \eta_h. \tag{2.5}$$

At this specific iteration k^*, it follows from (2.1) that the total error estimate $\eta_h^{(k)}$ will be bounded by twice the unknown approximation error (up to the constants in (2.2)). Under the assumption that the equivalence (2.4) represents an equality, a practical stopping test is obtained as in (2.5) except that the right-hand-side is given by

$$\|\mathbf{e}_h^{(k^*)}\|_{E_h} \;\le\; \theta\eta_h^{(k^*)}, \tag{2.6}$$

where $0 < \theta \le 1$. Observe that $\theta = 1/2$ corresponds to equality in (2.4). Note that the numerical results presented later have been produced with $\theta = 1$.

Generally, the iteration error $\mathbf{e}_h^{(k)}$ is unknown (and hence $\|\mathbf{e}_h^{(k)}\|_{E_h}$ is unknown too) since the exact algebraic solution $\mathbf{x}_h$ is not usually available. Usually some norm

$\|\mathbf{r}_h^{(k)}\|_{S_h} := \sqrt{(\mathbf{r}_h^{(k)})^T S_h \mathbf{r}_h^{(k)}}$ (where S_h is a symmetric positive-definite matrix) of the iteration residual $\mathbf{r}_h^{(k)} := \mathbf{b}_h - \mathcal{F}_h \mathbf{x}_h^{(k)}$ is readily computable and monotonically decreasing with respect to the iteration count k of the chosen solver. Obtaining tractable upper and lower bounds on the $\|\mathbf{e}_h^{(k)}\|_{E_h}$ norm in terms of the surrogate norm $\|\mathbf{r}_h^{(k)}\|_{S_h}$ is the novel feature of the optimal balanced black-box stopping strategy that is presented here. Moreover, this work states in these bounds, the exact positive constants $\lambda_h \in \mathbb{R}, \Lambda_h \in \mathbb{R}$ such that

$$\lambda_h \|\mathbf{r}_h^{(k)}\|_{S_h}^2 \leq \|\mathbf{e}_h^{(k)}\|_{E_h}^2 \leq \Lambda_h \|\mathbf{r}_h^{(k)}\|_{S_h}^2, \qquad k = 0, 1, 2, \ldots. \tag{2.7}$$

Equation (2.7) leads to the following two bounds on $\|\mathbf{e}_h^{(k)}\|_{E_h}$, that is,

$$\frac{\|\mathbf{e}_h^{(k)}\|_{E_h}}{\|\mathbf{e}_h^{(0)}\|_{E_h}} \leq \sqrt{\frac{\Lambda_h}{\lambda_h}} \frac{\|\mathbf{r}_h^{(k)}\|_{S_h}}{\|\mathbf{r}_h^{(0)}\|_{S_h}} \Longleftrightarrow \|\mathbf{e}_h^{(k)}\|_{E_h} \leq \frac{\Lambda_h}{\sqrt{\lambda_h}} \|\mathbf{r}_h^{(k)}\|_{S_h}, \tag{2.8a}$$

$$\|\mathbf{e}_h^{(k)}\|_{E_h} \leq \sqrt{\Lambda_h}\, \|\mathbf{r}_h^{(k)}\|_{S_h}. \tag{2.8b}$$

Using $\mathbf{r}_h^{(k)} = \mathcal{F}_h \mathbf{e}_h^{(k)} \Longleftrightarrow \|\mathbf{e}_h^{(k)}\|_{E_h}^2 = (\mathbf{r}_h^{(k)})^T \mathcal{F}_h^{-T} E_h \mathcal{F}_h^{-1} \mathbf{r}_h^{(k)}$. Thus, using (2.7) this implies that obtaining a lower (λ_h) bound and an upper (Λ_h) bound on the quantity $\frac{\|\mathbf{e}_h^{(k)}\|_{E_h}^2}{\|\mathbf{r}_h^{(k)}\|_{S_h}^2} = \frac{(\mathbf{r}_h^{(k)})^T \mathcal{F}_h^{-T} E_h \mathcal{F}_h^{-1} \mathbf{r}_h^{(k)}}{(\mathbf{r}_h^{(k)})^T S_h \mathbf{r}_h^{(k)}}$ requires calculating the extremal Rayleigh quotients [8, p. 453] of $\mathcal{F}_h^{-T} E_h \mathcal{F}_h^{-1}$ and S_h. This is equivalent to computing the extremal (outer most) eigenvalues, that is, the smallest eigenvalue and the largest eigenvalue of the symmetric positive-definite generalized eigenvalue problem for $\mathcal{F}_h^{-T} E_h \mathcal{F}_h^{-1}$ and S_h (or equivalently of E_h and $\mathcal{F}_h^T S_h \mathcal{F}_h$). This generalized eigenvalue problem can be transformed (theoretically through a Cholesky factorization of S_h) into a symmetric positive-definite algebraic eigenvalue problem (keeping the eigenvalues same) and hence λ_h and Λ_h will both be positive.

In light of (2.6), the bounds in (2.8) lead to the following strong and weak optimal balanced black-box stopping tests: stop at the first iteration k^* such that either holds,

$$\frac{\Lambda_h}{\sqrt{\lambda_h}} \|\mathbf{r}_h^{(k^*)}\|_{S_h} \leq \eta_h^{(k^*)}; \quad \sqrt{\Lambda_h}\, \|\mathbf{r}_h^{(k^*)}\|_{S_h} \leq \eta_h^{(k^*)},$$

or equivalently,

$$\|\mathbf{r}_h^{(k^*)}\|_{S_h} \leq \frac{\sqrt{\lambda_h}}{\Lambda_h} \eta_h^{(k^*)} \text{ (\textbf{Strong stop})}; \quad \|\mathbf{r}_h^{(k^*)}\|_{S_h} \leq \frac{1}{\sqrt{\Lambda_h}} \eta_h^{(k^*)} \text{ (\textbf{Weak stop})}. \tag{2.9}$$

In terms of the number of iterations (and hence computational work and time) for convergence, the strong stopping test cannot perform better than the weak stopping test since $\frac{\sqrt{\lambda_h}}{\Lambda_h} = \frac{\sqrt{\lambda_h}}{\sqrt{\Lambda_h}\sqrt{\Lambda_h}} \leq \frac{1}{\sqrt{\Lambda_h}}$. Moreover, the strong stopping test involves an additional overhead of computing the smallest eigenvalue λ_h. Thus, it would be prudent to employ the weak stopping test whenever possible.

Remark 2.2. A crucial point to note here is that if the employed a posteriori error estimator overestimates the approximation error (total errors), it will be better to employ the strong stopping test for otherwise the weak stopping test might lead to premature stopping.

3. Solution methodology for nonlinear solvers. It follows from (1.3) that at any nonlinear iterative step $l+1$,

$$\|u_h^{(l+1)} - u_h^{(l)}\|_{\mathcal{E}} = \|\delta u_h^{(l)}\|_{\mathcal{E}}. \tag{3.1}$$

Note that since norm of difference is greater than or equal to the difference of norms, $\|u_h^{(l+1)} - u_h^{(l)}\|_{\mathcal{E}} \geq \|u_h^{(l+1)} - u\|_{\mathcal{E}} - \|u_h^{(l)} - u\|_{\mathcal{E}}$, therefore (3.1) becomes

$$\|u_h^{(l+1)} - u\|_{\mathcal{E}} \leq \|u_h^{(l)} - u\|_{\mathcal{E}} + \|\delta u_h^{(l)}\|_{\mathcal{E}}. \tag{3.2}$$

If $\eta_h^{(l)}$ are 'tight' a posteriori approximation error estimators at l nonlinear iterative step such that $c\,\eta_h^{(l)} \leq \|u - u_h^{(l)}\|_{\mathcal{E}} \leq C\eta_h^{(l)}$, with $\frac{C}{c} \sim O(1)$, then (3.2) can be rewritten as

$$\eta_h^{(l+1)} \simeq \eta_h^{(l)} + \|\delta u_h^{(l)}\|_{\mathcal{E}}. \tag{3.3}$$

Note that $\{\eta_h^{(l)}\}$ ultimately converges to true a posteriori approximation error estimate η_h. So $\forall\, l \geq \hat{l}$ (say), $\eta_h^{(l)}$ are $\eta_h^{(l+1)}$ are essentially the same. Using this idea, one can optimally stop the nonlinear iteration when the contribution from $\|\delta u_h^{(l)}\|_{\mathcal{E}}$ in (3.2) and (3.3) is insignificant. Thus, in light of the discussion in section 2, stop the nonlinear iteration at the smallest value of $(l+1)$ of l^* such that

$$\|\delta u_h^{(l^*)}\|_{\mathcal{E}} \leq \eta_h^{(l^*+1)}. \tag{3.4}$$

4. Main contribution. An iterative (linear or nonlinear) solver employing the optimal balanced stopping methodology that is presented here will be a black-box solver. In the various Krylov solvers used next for solving $\mathcal{M}_h^{-1}\mathcal{F}_h\mathbf{x}_h = \mathcal{M}_h^{-1}\mathbf{b}_h$, the eigenvalue problem for finding λ_h, Λ_h will be specified explicitly by identifying suitable E_h and S_h. For the remainder of this paper $\mathcal{F}_h$ will be assumed to be nonsymmetric. Note that this work has appeared as chapters 4 and 5 in first author's PhD thesis [17].

5. Optimal balanced black-box stopping tests for nonsymmetric linear solvers. Generalized minimal residual (GMRES) method [18] with preconditioning is the popular method for solving nonsymmetric linear systems. The iteration error norm $\|\mathbf{e}_h^{(k)}\|_{E_h} = \sqrt{(\mathbf{r}_h^{(k)})^T\mathcal{F}_h^{-T}E_h\mathcal{F}_h^{-1}\mathbf{r}_h^{(k)}}$ can be bounded here by GMRES solver's readily available and monotonically decreasing (with iteration count k) residual norm $\|\mathbf{r}_h^{(k)}\|_2 := \sqrt{(\mathbf{r}_h^{(k)})^T\mathbf{r}_h^{(k)}}$ (S_h is the identity matrix here). This involves computing the smallest eigenvalue λ_h and the largest eigenvalue Λ_h of the generalized symmetric positive-definite eigenvalue problem for E_h and $\mathcal{F}_h^T\mathcal{F}_h$. A popular choice of E_h for nonsymmetric linear systems is the symmetric part of the coefficient matrix $\mathcal{F}_h$, that is, $\frac{\mathcal{F}_h^T+\mathcal{F}_h}{2}$ provided it is also positive-definite. Thus, in this case, λ_h and Λ_h are the smallest and the largest eigenvalue respectively of the generalized eigenvalue problem for $\frac{\mathcal{F}_h^T+\mathcal{F}_h}{2}$ and $\mathcal{F}_h^T\mathcal{F}_h$. In light of (2.9), the eigenvalue problem above leads to the following strong and weak stopping criteria in preconditioned GMRES for solving nonsymmetric linear systems: stop at the first iteration k^* such that either holds,

$$\|\mathbf{r}_h^{(k^*)}\|_2 \leq \frac{\sqrt{\lambda_h}}{\Lambda_h}\eta_h^{(k^*)}\ (\textbf{Strong stop}); \quad \|\mathbf{r}_h^{(k^*)}\|_2 \leq \frac{1}{\sqrt{\Lambda_h}}\eta_h^{(k^*)}\ (\textbf{Weak stop}). \tag{5.1}$$

Note that (5.1) is a black-box stopping test as opposed to the devised stopping test in [25, chapter 5] and involves a posteriori error bounds as opposed to [1].

It is further proposed here that that (5.1) can be used in suboptimal solvers like BICGSTAB(ℓ) [21], TFQMR [6] etc., which unlike GMRES require fixed and less storage requirements per iteration like and for which little convergence theory exists currently. This is provided that breakdowns in such solvers are handled adequately (see [7]) and these solvers converge at least to the order of the PDE approximation error.

5.1. Computational logistics of optimal balanced black-box stopping tests. Optimal balanced black box tests in preconditioned MINRES [13] solver for solving symmetric positive-definite and symmetric indefinite linear systems can be found in [17, chapters 2–3]. There $E_h = \mathcal{M}_h^{-1}$ and hence preconditioner was involved in the optimal stopping test's eigenvalue problem of interest. So, λ_h, Λ_h could be estimated cheaply there by exploiting Ritz (eigenvalues of the MINRES Lanczos matrix) and harmonic Ritz value (eigenvalues of 'modified' MINRES Lanczos matrix) relations with the required eigenvalues of the preconditioned coefficient matrix of interest. But unlike MINRES, a cheap method for computing λ_h, Λ_h in (5.1) is ongoing research since here the eigenvalue problem of interest requires computing the eigenvalues of a matrix different from the preconditioned coefficient matrix. So, Ritz/harmonic Ritz value relations are difficult to exploit. Using MATLAB `eigs` is an alternative but it might be more expensive if the matrix has 'many' nonzero entries. However, ($\mathcal{F}_h$) coefficient matrix arising from FEM approximation of PDEs are sparse in general and hence employing `eigs` to obtain largest and smallest eigenvalues of generalized eigenvalue problem for $\frac{\mathcal{F}_h^T + \mathcal{F}_h}{2}$ and $\mathcal{F}_h^T \mathcal{F}_h$ might not be too expensive. Also, note that $\eta_h^{(k)}$ should be computed periodically (say every 4–5 iterations) to minimize its impact on the overall algorithmic cost.

6. Model test problems. The storage requirements and computational flops increase with the size and the number of linear systems (which are usually huge for stochastic PDEs). An optimal balanced black-box stopping test might save significant computational work of an iterative solver (especially for stochastic PDEs) and in any case it would rule out premature stopping. Note that the optimal balanced black-box stopping methodology developed here remains applicable for solving the discrete systems arising from numerical approximation of the corresponding stochastic PDE too.

Observe that availability of a tight posteriori approximation error estimator is crucial for devising an optimal balanced black-box stopping methodology. Tight a posteriori PDE approximation error estimators for parametric convection-diffusion equations and parametric Navier–Stokes equations is still an ongoing research and hence their deterministic counterparts are chosen here to illustrate the developed optimal balanced black-box stopping methodology.

6.1. Convection-diffusion equations. Convection-diffusion equations are used for modelling various phenomena such as the temperature of a fluid moving along a heated wall, the transfer and diffusion of pollutants, etc.; see [5, chapter 18].

Following the notation in [3, p. 234], the steady-state scalar convection-diffusion solution $u(\vec{x}) : D \to \mathbb{R}$ satisfies

$$-\nabla \cdot \epsilon(\vec{x}) \nabla u(\vec{x}) + \vec{w}(\vec{x}) \cdot \nabla u(\vec{x}) = f(\vec{x}), \quad \forall\, \vec{x} \in \Omega \subset \mathbb{R}^d \, (d = 2, 3), \tag{6.1a}$$

$$u(\vec{x}) = g_D(\vec{x}), \quad \forall\, \vec{x} \in \partial\Omega_D, \tag{6.1b}$$

$$\nabla u(\vec{x}) \cdot \vec{n} = g_N(\vec{x}), \quad \forall\, \vec{x} \in \partial\Omega_N = \partial\Omega \setminus \partial\Omega_D. \tag{6.1c}$$

Here Ω is the spatial domain, $\vec{w}$ denotes the wind, and $\epsilon := \kappa I$ is the isotropic permeability tensor, $\kappa : \Omega \to \mathbb{R}$. The quantities f, g_D, g_N are given functions and $\vec{n}$ denotes the normal to boundary $\partial\Omega$, which is the union of the Dirichlet ($\partial\Omega_D$) and the Neumann ($\partial\Omega_N$) spatial boundary.

For the simplicity of exposition, the diffusion coefficient $\epsilon > 0$ will be assumed to be independent of the spatial coordinates. Also, it will be assumed that $\nabla \cdot \vec{w} = 0$.

6.1.1. Galerkin FEM formulation. The Galerkin FEM formulation of (6.1) is to find $u_h \in S_E^h$ such that

$$\epsilon \int_\Omega (\nabla u_h \cdot \nabla v_h) + \int_\Omega (\vec{w} \cdot \nabla u_h) v_h = \int_\Omega f v_h + \epsilon \int_{\partial\Omega_N} g_N v_h, \qquad \forall\, v_h \in S_0^h, \tag{6.2}$$

where S_E^h and S_0^h are finite dimensional subspaces of H_E^1 and $H_{E_0}^1$ respectively. Here

$$\begin{aligned} H_E^1(\Omega) &:= \{v \in H^1(\Omega) \,|\, v = g_D \text{ on } \partial\Omega_D\}, \\ H_{E_0}^1(\Omega) &:= \{v \in H^1(\Omega) \,|\, v = 0 \text{ on } \partial\Omega_D\}, \\ H^1(\Omega) &:= \{u \in L^2(\Omega) \,|\, D^{\boldsymbol{\alpha}} u \in L^2(\Omega), \forall\, |\boldsymbol{\alpha}| \le 1\}, \end{aligned}$$

where $D^{\boldsymbol{\alpha}}$ is distributional derivative of u, $|\boldsymbol{\alpha}| := \sum_{i=1}^d \alpha_i$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$ is a multiindex, see [12, p. 434].

Wathen [24] advocates that a natural norm for a function u in the Sobolev space $H_{E_0}^1$ is the L^2 norm of its gradient, that is, $\|\nabla u\|_{L^2(\Omega)} := \sqrt{\int_\Omega (\nabla u_h)^2}$. However, this need not be the only meaningful norm for measuring errors associated with (6.2). An alternative norm known as the streamline diffusion norm is also discussed in [3, p. 252]. This norm arises when the streamline diffusion method introduced by [10] is used for overcoming the drawbacks associated with the Galerkin discretization.[1] This leads to a slightly different FEM formulation to (6.2) known as the streamline diffusion FEM formulation [3, p. 252].

The corresponding streamline diffusion norm is

$$\|u_h\|_{\text{sd}} := (\epsilon \, \|\nabla u_h\|_{L^2(\Omega)}^2 + \delta \, \|\vec{w} \cdot \nabla u_h\|_{L^2(\Omega)}^2)^{1/2}, \tag{6.3}$$

where δ denotes the stabilization parameter [3, p. 253]. For convection dominated problems, that is, for large Peclet numbers (small ϵ), the solution u_h is dominated by its behaviour along the streamlines, and hence $\|u_h\|_{\text{sd}}$ which involves the streamline derivative $\|\vec{w} \cdot \nabla u_h\|_{L^2(\Omega)}$ is a more meaningful measure than $\|\nabla u_h\|_{L^2(\Omega)}$.

The IFISS toolbox employs streamline diffusion stabilization for solving (6.1), but measures errors in the L^2 norm of the gradient. The balanced stopping test will be based on this norm. However, the stopping methodology can easily be modified to cater to the streamline diffusion norm in (6.3).

Having formulated the streamline diffusion FEM formulation, the target linear system is set up in the next subsection.

6.1.2. Matrix formulation. The Galerkin FEM formulation (6.2) gives rise to the following system of linear equations with given coefficient matrix $\mathcal{F}_h$, given

[1] Galerkin approximation for (6.1) is inaccurate if the mesh is not fine enough to resolve the layers in the solution and these inaccuracies may also propagate and pollute the approximated solution in regions where the exact solution is well behaved. An alternative way to handle boundary layers is by using Shishkin grids; see [19].

right-hand-side vector $\mathbf{b}_h$, and unknown vector $\mathbf{x}_h$.

$$\mathcal{F}_h \mathbf{x}_h = \mathbf{b}_h \quad \Longleftrightarrow \quad \mathcal{M}_h^{-1}\mathcal{F}_h \mathbf{x}_h = \mathcal{M}_h^{-1}\mathbf{b}_h, \tag{6.4}$$

where $\mathcal{M}_h$ is a preconditioner. When lower order (piecewise linear or bilinear) finite elements along with stabilization are employed, the coefficient matrix $\mathcal{F}_h$ in (6.4) has the following form,

$$\mathcal{F}_h \;=\; \epsilon\, A_h \;+\; N_h \;+\; S_h,$$

see [3, p. 272] for more details. For FEM discretization without streamline diffusion stabilization, $\mathcal{F}_h = \epsilon A_h + N_h$ for finite elements of any order. The matrices under consideration are quite structured. The matrix A_h is symmetric and positive-definite provided Dirichlet boundary conditions exist over an interval ($\int_{\partial\Omega_D} \neq 0$), however small. The stabilization matrix S_h is symmetric and positive-semidefinite. The matrix N_h is a skew-symmetric matrix [3, p. 241, pp. 271–272]. Thus, $\mathcal{F}_h$ is a nonsymmetric matrix that will be assumed to be invertible throughout this paper. Iterative solvers like GMRES, BICGSTAB(ℓ), and TFQMR are popular for solving nonsymmetric linear systems.

6.1.3. A posteriori approximation error estimation. The a posteriori approximation error estimator that will be employed here for the deterministic convection-diffusion equations is reliable but need not always be efficient. It is reliable in the sense that the global upper bound on the true error does not depend on the mesh parameter h and the diffusion parameter ϵ. However, it might not always be possible that the a posteriori error estimate is a lower bound on the local (elemental) approximation error [3, theorem 6.9, proposition 6.11, pp. 264–265]. According to [3, p. 265], efficiency issue is generic for any local error estimator whenever boundary layers are not resolved by the FEM approximation. Hence, streamline diffusion stabilization is necessary for dealing adequately (but not completely!) with such situations.

To demonstrate that the employed a posteriori error estimator is a 'close' estimate of the approximation error, some computational results are presented for the test problem described in (next) subsection 6.1.4. The a posteriori error η_h and the actual approximation error $\|\nabla(u_{ref} - u_h)\|_{L^2(\Omega)} := \sqrt{\int_\Omega \nabla(u_{ref} - u_h)^2}$ using a reference solution are tabulated in Table 1 for a sequence of uniform grids. Since the exact

TABLE 1
Approximation errors, a posteriori errors, and effectivity indices for convection-diffusion test problem on uniform grids.

h	η_h	$\|\nabla(u_{ref} - u_h)\|_{L^2(\Omega)}$	β_{eff}	$\mathcal{P}_h^{r_{\max}}$
1/16	1.0562	4.1162	0.25	3.87
1/32	0.8556	2.6216	0.33	1.97
1/64	0.8018	1.5380	0.52	0.99
1/128	0.7855	0.7571	1.04	0.50

solution to the model problem is not available, a reference solution u_{ref} is computed on a fine ($h = 1/256$) spatial 512×512 uniform grid. This reference solution is then compared with the computed FEM solution u_h (which is linearly interpolated using MATLAB `interp2` function for compatible comparison with the reference solution) for grids with $h = 1/16, 1/32, 1/64, 1/128$. The corresponding effectivity index, that is, $\beta_{\text{eff}} = \frac{\eta_h}{\|\nabla(u_{ref} - u_h)\|_{L^2(\Omega)}}$ is also presented. The column for β_{eff} in Table 1 indicates that

the a posteriori error estimator is an 'acceptably close' estimate of the approximation error. In fact as the mesh is refined and the layers in the solution are resolved, (that is, maximum mesh Peclet number [3, p. 253] $\mathcal{P}_h^{r_{\max}}$ approaches ≤ 1) $\beta_{\text{eff}} \to 1$. Note that the computation of a posteriori error estimator employed here is quite cheap since it requires solving for a local 5×5 linear system on each element.

6.1.4. Experimental results. From Table 1, observe that the computed a posteriori approximation errors do not overestimate the corresponding approximation error. In light of Remark 2.2 this implies that the weak stopping test in (5.1) can be used as the optimal balanced black-box stopping test in GMRES, BICGSTAB(ℓ), and TFQMR for solving (6.4). Thus, one needs to compute the largest eigenvalue Λ_h of the generalized eigenvalue problem for $\frac{\mathcal{F}_h^T+\mathcal{F}_h}{2\epsilon}$ and $\mathcal{F}_h^T\mathcal{F}_h$. (Note that here the iteration errors are measured in $E_h = A_h$ norm, and the symmetric positive-definite part of $\mathcal{F}_h$ is $A_h = \frac{\mathcal{F}_h^T+\mathcal{F}_h}{2\epsilon}$ and not just $\frac{\mathcal{F}_h^T+\mathcal{F}_h}{2}$ as mentioned in section 4). To reiterate, the employed iterative solver will stop at the first iteration k^* such that $\sqrt{\Lambda_h}\ \|\mathbf{r}_h^{(k^*)}\|_2 \leq \eta_h^{(k^*)}$ (**Weak stop**). Also, note that Λ_h is computed here using MATLAB `eigs`. Some alternative approaches (still under further research) towards cheaper estimation of Λ_h are discussed in [17, chapter 4].

The results here are presented for GMRES, BICGSTAB(2), and TFQMR. The choice $\ell = 2$ for BICGSTAB(ℓ) is quite popular and widespread among practitioners; see [3, p. 296]. Roundoff errors might pollute the residual norm computed from short-term recurrences for suboptimal Krylov solvers. In order to avoid these inaccuracies, $\|\mathbf{r}_h^{(k)}\|_2$ is computed here after forming the residual explicitly, that is, $\mathbf{r}_h^{(k)} = \mathbf{b}_h - \mathcal{F}_h\mathbf{x}_h^{(k)}$. It is claimed here that in presence of tight a posteriori approximation error estimators, the balanced stopping test can be employed optimally for suboptimal iterative methods too provided breakdowns are handled adequately and these algorithms 'converge' at least to the accuracy of the true approximation error.

Four test problems based on (6.1) are present in IFISS software in MATLAB. Computational results are presented here for the fourth test problem. This problem is characterized by a recirculating wind $\vec{w}$ and has discontinuous Dirichlet boundary conditions leading to the formation of boundary layers near the corners of the domain [3, p. 240].

The convection-diffusion problem (6.1) is defined on $\Omega = (-1, 1) \times (-1, 1)$ with the source function $f(x_1, x_2) = 0$, $\forall\,(x_1, x_2) \in \Omega$. Rectangular piecewise bilinear ($\boldsymbol{Q}_1$) finite elements are used on a sequence of uniform grids. The viscosity parameter $\epsilon = 1/64$ is fixed and the optimal inbuilt value of the stabilization parameter δ is used; see [3, p. 253]. This problem can be set up by choosing test problem `4` after running the driver `cd_testproblem` in IFISS.

There are four preconditioners built in IFISS for the discrete convection-diffusion problem. They are: diagonal (DIAG) preconditioner, that is, the diagonal matrix formed from the diagonal elements of $\mathcal{F}_h$, incomplete LU (ILU), geometric multigrid (GMG), and algebraic multigrid (AMG) preconditioners; see [3, chapter 7]. Results are presented here only for ILU and AMG preconditioners for each of the iterative methods. Let $\mathbf{x}_h$ denote the MATLAB backslash (Gaussian elimination) solution on each grid. Henceforth, this will be regarded as the reference (true) algebraic solution. This will be used for comparison with the result $\mathbf{x}_h^{(k^*)}$ computed using the balanced stopping test. From $\mathbf{x}_h$, the reference (true) a posteriori error estimate η_h is computed. The starting vector $\mathbf{x}_h^{(0)}$ is generated using the MATLAB function `rand`. The balanced stopping test that is used in preconditioned GMRES and BICGSTAB(ℓ) is

TABLE 2
GMRES iteration counts and errors for ILU (left), AMG (right) preconditioning on uniform grids for discrete convection-diffusion system.

h	k_{tol1}	k_{tol2}	k^*	$e^*_{\eta_h}$
1/16	19	24	7	1.9e-3
1/32	43	54	19	4.4e-4
1/64	113	144	54	1.4e-4
1/128	288	374	148	2.9e-5

h	k_{tol1}	k_{tol2}	k^*	$e^*_{\eta_h}$
1/16	6	10	3	1.4e-3
1/32	7	11	4	7.7e-5
1/64	8	14	4	4.4e-5
1/128	7	14	5	1.8e-6

implemented in `gmres_r` and `bicgstab_ell` in IFISS respectively, while the balanced stopping test in preconditioned TFQMR is incorporated in the existing MATLAB function for this solver. Also, let $\eta_h^{(k^*)}$ denote the a posteriori error estimate at the optimal stopping iteration k^* and $e^*_{\eta_h} := |\eta_h - \eta_h^{(k^*)}|$. These values are tabulated in the Tables 2 to 4 for each preconditioner on every grid level for both uniform and stretched grids. The insights from these numbers are quite generic, which are summarised in the following paragraphs. The $e^*_{\eta_h}$ columns show that $\{\eta_h^{(k)}\}$ has converged

TABLE 3
BICGSTAB(2) iteration counts and errors for ILU (left), AMG (right) preconditioning on uniform grids for discrete convection-diffusion system.

h	k_{tol1}	k_{tol2}	k^*	$e^*_{\eta_h}$
1/16	12	16	6	4.1e-5
1/32	30	38	15	1.7e-4
1/64	86	114	48	2.8e-5
1/128	236	290	124	3.5e-5

h	k_{tol1}	k_{tol2}	k^*	$e^*_{\eta_h}$
1/16	4	6	2	2.0e-4
1/32	4	6	3	3.7e-6
1/64	4	8	4	8.0e-6
1/128	4	8	4	3.4e-6

with a good accuracy to the reference a posteriori error estimate η_h at the balanced stopping iteration. To show the effectiveness of the balanced stopping test, the iteration counts k^* needed to satisfy the balanced stopping test have been compared with iteration counts $k_{\text{tol1}}, k_{\text{tol2}}$ needed to satisfy a fixed relative residual $\dfrac{\|\mathbf{r}_h^{(k)}\|_2}{\|\mathbf{r}_h^{(0)}\|_2}$ reduction tolerance of `1e-6` (which is the default tolerance in MATLAB solvers) and `1e-9` respectively. These tolerance values are a realistic user-input tolerance choices in the absence of a balanced stopping test. The user will not know in general the stopping point k^* a priori and is more likely to provide a tighter/coarser tolerance than actually required. This would lead to unnecessary computations/premature stopping.

A comparison of the corresponding columns for ILU iteration counts shows that for the same approximation error, a significant number of iterations is saved by using the balanced stopping test. This would result in significant savings in computational work of the solver (as compared to using fixed relative residual $\dfrac{\|\mathbf{r}_h^{(k)}\|_2}{\|\mathbf{r}_h^{(0)}\|_2}$ reduction tolerance `1e-6` or tighter) if one were to solve the (preconditioned) linear systems arising from adaptive finite element for the chosen problem parameters. The linear systems that are solved are of size: 1089×1089, 4225×4225, 16641×16641, and 66049×66049. These computational savings are even more striking in light of the huge size of some of these systems. Also, notice that in the case of AMG iterations, not much savings (in terms of iteration counts) is achieved by using the optimal balanced black-box stopping test. However, using the optimal balanced black-box stopping test does ensure that the employed solver has not stopped prematurely.

TABLE 4
TFQMR iteration counts and errors for ILU (left), AMG (right) preconditioning on uniform grids for discrete convection-diffusion system.

h	k_{tol1}	k_{tol2}	k^*	$e^*_{\eta_h}$
1/16	32	37	15	7.5e-4
1/32	73	84	36	8.3e-5
1/64	193	234	105	4.4e-5
1/128	534	684	345	4.8e-5

h	k_{tol1}	k_{tol2}	k^*	$e^*_{\eta_h}$
1/16	9	12	4	4.1e-5
1/32	7	14	4	4.4e-4
1/64	9	19	4	1.9e-5
1/128	8	17	5	4.3e-5

Among the employed iterative methods here, BICGSTAB(2) performs the best with each preconditioner. Between GMRES and TFQMR, GMRES converges slightly faster. However, using GMRES over TFQMR could be memory extensive in terms of storage. In any case, the optimal balanced black-box stopping test provides an optimal stopping point for suboptimal Krylov solvers like TFQMR etc. Indeed this is crucially dependent on the fact that these suboptimal solvers do not break down prematurely. Note that the main aim of these computational results is not to compare the convergence rates of GMRES and various suboptimal Krylov solvers but to illustrate that an optimal balanced black-box stopping test (5.1) can be employed for suboptimal solvers as well.

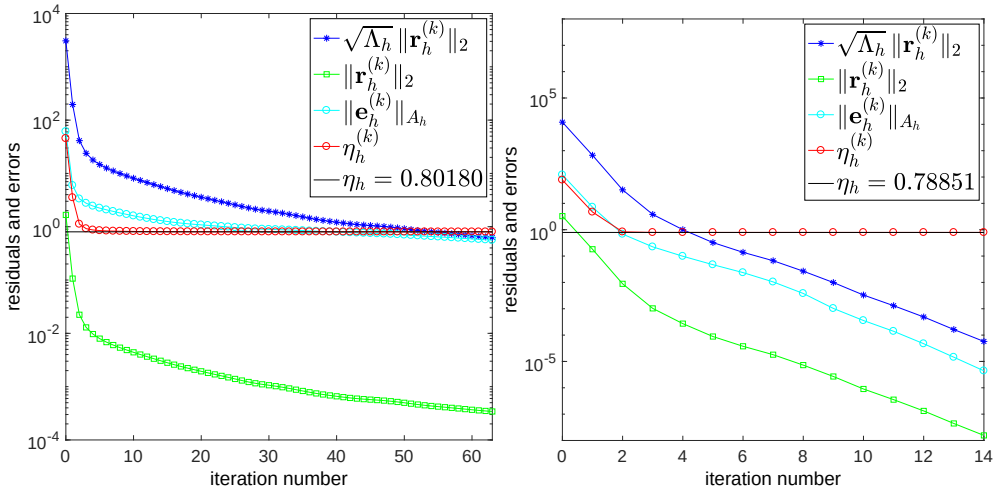

FIG. 1. *Errors vs iteration number for convection-diffusion test problem for ILU preconditioned GMRES with $h = 1/64$ (left) and AMG preconditioned GMRES with $h = 1/128$ (right).*

In order to gain further insight from the numerical experiments, the evolution of the following quantities—$\eta_h^{(k)}$, $\|\mathbf{e}_h^{(k)}\|_{A_h}$, $\|\mathbf{r}_h^{(k)}\|_2$, and the (weak) algebraic error bound $\sqrt{\Lambda_h}\|\mathbf{r}_h^{(k)}\|_2$ is also plotted. The optimal balanced black-box (weak) stopping test stops optimally when the $\sqrt{\Lambda_h}\|\mathbf{r}_h^{(k)}\|_2$ curve is below the $\eta_h^{(k)}$ curve. From the plots it follows that when the contribution of $\|\mathbf{e}_h^{(k)}\|_{A_h}$ to the sum $\eta_h + \|\mathbf{e}_h^{(k)}\|_{A_h}$ is insignificant,[2] $\{\eta_h^{(k)}\}$ converges to η_h.

Indeed this is the case in all plots of Figures 1 to 3. In order to illustrate this convergence, iterations have been continued for nine more steps after optimal stopping

[2]From visual inspection this seems to occur soon after $\|\mathbf{e}_h^{(k)}\|_{A_h} \leq \eta_h$. Generally, both these quantities are unknown. So, a priori knowledge of optimal stopping step is generally difficult.

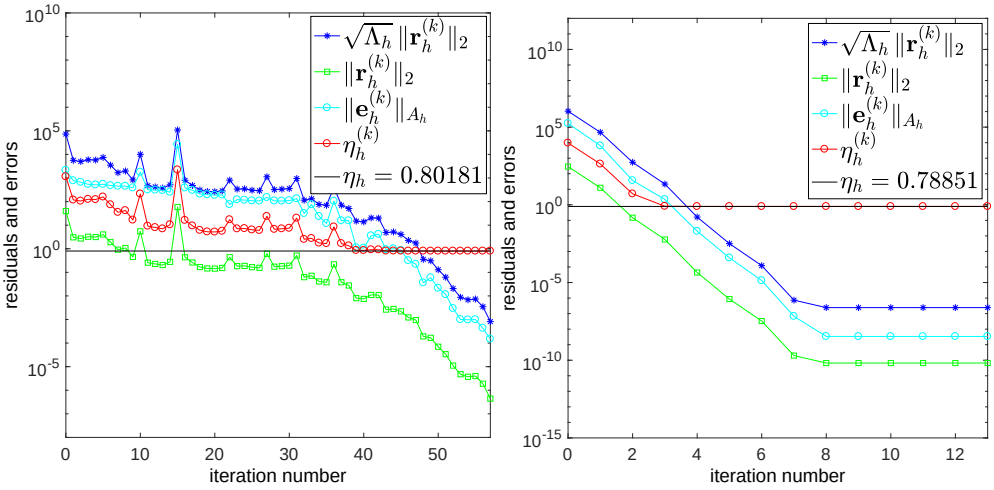

FIG. 2. *Errors vs iteration number for convection-diffusion test problem for ILU preconditioned BICGSTAB(2) with* $h = 1/64$ *(left) and AMG preconditioned BICGSTAB(2) with* $h = 1/128$ *(right).*

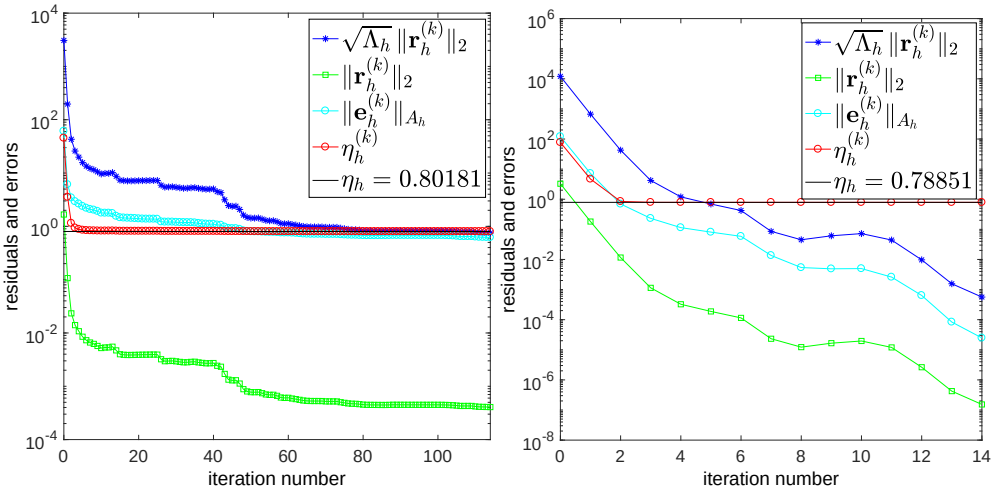

FIG. 3. *Errors vs iteration number for convection-diffusion test problem for ILU preconditioned TFQMR with* $h = 1/64$ *(left) and AMG preconditioned TFQMR with* $h = 1/128$ *(right).*

in each plot. This also illustrates optimal stopping at the correct iteration, that is $\{\eta_h^{(k)}\}$ converges to η_h on each plot. In each plot of Figure 1, it is noticed that after a initial burn in period, the rate of convergence of $\|\mathbf{r}_h^{(k)}\|_2$ is constant and is the (famous) asymptotic convergence factor [3, p. 290] of GMRES. Also, in Figure 1, Euclidean norm of the residual $\|\mathbf{r}_h^{(k)}\|_2$ is monotonically decreasing in GMRES while it exhibits irregular behaviour for BICGSTAB(2) and TFQMR; see ILU plots in Figures 2 and 3. However, a 'good' preconditioner smoothes out the irregular behaviour to a large extent; see Figures 2 and 3 for AMG preconditioned BICGSTAB(2) and TFQMR respectively.

6.2. Navier–Stokes equations. Navier–Stokes equations form the fundamental model of an incompressible Newtonian fluid such as air etc [3, p. 333 ff.]. Similar

to the convection-diffusion equations, the steady-state Navier–Stokes solution $(\vec{u}, p)$ is defined on a spatial domain $\Omega \subset \mathbb{R}^d$, $(d = 2, 3)$, where the vector valued velocity function $\vec{u}(\vec{x}) : \Omega \to \mathbb{R}^d$ and the scalar valued pressure function $p(\vec{x}) : \Omega \to \mathbb{R}$ satisfy

$$-\nu\nabla\cdot\nabla\vec{u}(\vec{x}) + \vec{u}(\vec{x})\cdot\nabla\vec{u}(\vec{x}) + \nabla p(\vec{x}) = \vec{f}(\vec{x}), \qquad \forall\,\vec{x}\in\Omega, \tag{6.5a}$$

$$\nabla\cdot\vec{u}(\vec{x}) = 0, \qquad \forall\,\vec{x}\in\Omega, \tag{6.5b}$$

$$\vec{u}(\vec{x}) = \vec{w}(\vec{x}), \qquad \forall\,\vec{x}\in\partial\Omega_D, \tag{6.5c}$$

$$\nu\nabla\vec{u}(\vec{x})\cdot\vec{n} - \vec{n}p(\vec{x}) = \vec{0}, \qquad \forall\,\vec{x}\in\partial\Omega_N. \tag{6.5d}$$

The functions $\vec{f}$, $\vec{w}$ are given and $\partial\Omega_D$, $\partial\Omega_N$ are the Dirichlet and Neumann parts respectively of the spatial boundary $\partial\Omega$. Kinematic velocity $\nu > 0$ is given and $\vec{n}$ denotes the outward normal to $\partial\Omega$. Note that the presence of the convective term gives the Navier–Stokes equations a nonlinear behaviour.

6.2.1. Mixed FEM formulation. The mixed FEM formulation of (6.5) is to find $\vec{u}_h \in \mathbf{X}^1_E \subset \mathbf{H}^1_E(\Omega)$ and $p_h \in M^h \subset L^2(\Omega)$ such that

$$\begin{aligned}\nu\int_\Omega \nabla\vec{u}_h : \nabla\vec{v}_h + \int_\Omega (\vec{u}_h\cdot\nabla\vec{u}_h)\cdot\vec{v}_h - \int_\Omega p_h\,(\nabla\cdot\vec{v}_h) &= \int_\Omega \vec{f}\cdot\vec{v}_h, \ \forall\,\vec{v}_h\in\mathbf{X}^h_{E_0}\subset\mathbf{H}^1_{E_0}(\Omega),\\ \int_\Omega q_h\,(\nabla\cdot\vec{u}_h) &= 0, \qquad \forall\,q_h\in M^h,\end{aligned} \tag{6.6}$$

where the spaces $\mathbf{H}^1_E(\Omega), \mathbf{H}^1_{E_0}(\Omega)$ are the vector versions of spaces $H^1_E(\Omega), H^1_{E_0}(\Omega)$ respectively, defined in (6.2) and $\nabla\vec{u} : \nabla\vec{v}$ denotes componentwise dot product. The solution of (6.6) involves nonlinear iterations that requires solving a linearized problem at each iterative step.

Starting with a given initial guess $(\vec{u}^{(0)}_h, p^{(0)}_h) \in \mathbf{X}^1_E \times M^h$, a sequence $\{(\vec{u}^{(l+1)}_h, p^{(l+1)}_h)\}$, $l = 0, 1, \cdots$ of iterates is constructed satisfying (6.6) such that [3, pp. 344, 341],

$$\vec{u}^{(l+1)}_h = \vec{u}^{(l)}_h + \delta\vec{u}^{(l)}_h, \qquad p^{(l+1)}_h = p^{(l)}_h + \delta p^{(l)}_h. \tag{6.7}$$

Plugging (6.7) in (6.6) gives

$$\begin{aligned}D(\vec{u}^{(l)}_h, \delta\vec{u}^{(l)}_h, \vec{v}_h) + \nu\int_\Omega \nabla\delta\vec{u}^{(l)}_h : \nabla\vec{v}_h - \int_\Omega \delta p^{(l)}_h\,(\nabla\cdot\vec{v}_h) &= R^{(l)}(\vec{v}_h), \qquad \forall\,\vec{v}_h\in\mathbf{X}^h_{E_0},\\ \int_\Omega q_h\,(\nabla\cdot\vec{u}^{(l)}_h) &= r^{(l)}(\vec{q}_h), \qquad \forall\,q_h\in M^h,\end{aligned} \tag{6.8}$$

where

$$\begin{aligned}R^{(l)}(\vec{v}_h) &= \int_\Omega \vec{f}\cdot\vec{v}_h - \int_\Omega (\vec{u}^{(l)}_h\cdot\nabla\vec{u}^{(l)}_h)\cdot\vec{v}_h - \nu\int_\Omega \nabla\vec{u}^{(l)}_h : \nabla\vec{v}_h + \int_\Omega p^{(l)}_h\,(\nabla\cdot\vec{v}_h),\\ r^{(l)}(\vec{q}_h) &= -\int_\Omega q_h\,(\nabla\cdot\vec{u}^{(l)}_h),\end{aligned}$$

$$D(\vec{u}^{(l)}_h, \delta\vec{u}^{(l)}_h, \vec{v}_h) = \int_\Omega (\delta\vec{u}^{(l)}_h\cdot\nabla\delta\vec{u}^{(l)}_h)\cdot\vec{v}_h + \int_\Omega (\delta\vec{u}^{(l)}_h\cdot\nabla\vec{u}^{(l)}_h)\cdot\vec{v}_h + \int_\Omega (\vec{u}^{(l)}_h\cdot\nabla\delta\vec{u}^{(l)}_h)\cdot\vec{v}_h.$$

6.2.2. Newton iteration. Dropping the quadratic term $\int_\Omega (\delta\vec{u}_h^{(l)} \cdot \nabla \delta\vec{u}_h^{(l)}) \cdot \vec{v}_h$ of D and substituting in (6.8) leads to solving a linear problem for the Newton correction $(\delta\vec{u}^{(l)}, \delta p^{(l)})$ at the lth iterative step. That is, $\forall\, (\vec{v}_h, q_h) \in \mathbf{X}_{E_0}^h \times M^h$, find $(\delta\vec{u}_h^{(l)}, \delta p_h^{(l)}) \in \mathbf{X}_{E_0}^h \times M^h$ such that

$$\int_\Omega (\delta\vec{u}_h^{(l)} \cdot \nabla \vec{u}_h^{(l)}) \cdot \vec{v}_h \;+\; \int_\Omega (\vec{u}_h^{(l)} \cdot \nabla \delta\vec{u}_h^{(l)}) \cdot \vec{v}_h \;+\; \nu \int_\Omega \nabla \delta\vec{u}_h^{(l)} : \nabla \vec{v}_h \\ - \int_\Omega \delta p_h^{(l)}\, (\nabla \cdot \vec{v}_h) \;=\; R^{(l)}(\vec{v}_h) \tag{6.9a}$$

$$\int_\Omega q_h\, (\nabla \cdot \vec{u}_h^{(l)}) \;=\; r^{(l)}(\vec{q}_h). \tag{6.9b}$$

6.2.3. Picard iteration. Further linearization is achieved by dropping the linear term $\int_\Omega (\delta\vec{u}_h^{(l)} \cdot \nabla \vec{u}_h^{(l)}) \cdot \vec{v}_h$ in (6.9). This leads to solving a linear problem for the Picard correction $(\delta\vec{u}^{(l)}, \delta p^{(l)})$ at the lth iterative step.

6.2.4. Matrix formulation. Let $\{\vec{\phi}_j\}_{j=1}^{n_u}$ be a basis for $\mathbf{X}_{E_0}^h$. Then any arbitrary $\delta\vec{u}_h^{(l)} \in \mathbf{X}_{E_0}^h$ can be expressed as $\delta\vec{u}_h^{(l)} = \sum_{j=1}^{n_u} \Delta u_j^{(l)} \vec{\phi}_j$, $\Delta u_j^{(l)} \in \mathbb{R}$. Also, $\{\vec{\phi}_j\}_{j=1}^{n_u}$ can be extended (loosely speaking) [3] to form a basis for $\mathbf{X}_E^h$, so that any $\vec{u}_h^{(l)} \in \mathbf{X}_E^h$ can be expanded as $\vec{u}_h^{(l)} = \sum_{j=1}^{n_u + n_\partial} u_j^{(l)} \vec{\phi}_j$, $u_j^{(l)} \in \mathbb{R}$, where the term $\sum_{j=n_u+1}^{n_u+n_\partial} u_j^{(l)} \vec{\phi}_j$ interpolates the boundary data on $\partial\Omega_D$.

Similarly, if $\{\psi_k\}_{k=1}^{n_p}$ be a basis for M^h, then any $p_h^{(l)}, \delta p_h^{(l)} \in M^h$ has an expression $p_h^{(l)} = \sum_{k=1}^{n_p} p_k^{(l)} \psi_k$, $\delta p_h^{(l)} = \sum_{k=1}^{n_p} \Delta p_k^{(l)} \psi_k$, $p_k^{(l)}, \Delta p_k^{(l)} \in \mathbb{R}$. Since $\vec{u}_h^{(l)}, p_h^{(l)}$ are known from the previous iterative step, their basis coefficients are known too.

Using these basis exapansions in (6.9) leads to the following discrete (Newton) system of linear equations at the lth nonlinear iterative step,

$$\begin{bmatrix} \nu\mathbf{A}_h + \mathbf{N}_h^{(l)} + \mathbf{W}_h^{(l)} & B_h^T \\ B_h & O \end{bmatrix} \begin{bmatrix} \Delta\mathbf{u}_h^{(l)} \\ \Delta\mathbf{p}_h^{(l)} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_h^{(l)} \\ \mathbf{g}_h^{(l)} \end{bmatrix}. \tag{6.10}$$

The symmetric positive-definite matrix $\mathbf{A}_h$ (vector-Laplacian matrix) is the block diagonal matrix with the usual FEM stiffness matrix on its diagonals and $\mathbf{N}_h^{(l)}$ is the vector convection matrix, (the scalar versions of both were introduced in (6.4)). Solution vectors $\Delta\mathbf{u}_h^{(l)} = [\Delta u_1^{(l)}, \ldots, \Delta u_{n_u}^{(l)}]^T \in \mathbb{R}^{n_u}$, $\Delta\mathbf{p}_h^{(l)} = [\Delta p_1^{(l)}, \ldots, \Delta p_{n_p}^{(l)}]^T \in \mathbb{R}^{n_p}$ and for the entries of $\mathbf{A}_h$, B_h, $\mathbf{N}_h^{(l)}$, $\mathbf{W}_h^{(l)}$, $\mathbf{f}_h^{(l)}$, and $\mathbf{g}_h^{(l)}$, see [3, p. 348]. Note the dependence of vector convection matrix $\mathbf{N}_h$, Newton derivative matrix $\mathbf{W}_h$, and right-hand-side vectors $\mathbf{f}_h, \mathbf{g}_h$ on the nonlinear iterative step.

Dropping the Newton derivative matrix in (6.10) results in the linear system arising from Picard iteration, which is,

$$\begin{bmatrix} \nu\mathbf{A}_h + \mathbf{N}_h^{(l)} & B_h^T \\ B_h & O \end{bmatrix} \begin{bmatrix} \Delta\mathbf{u}_h^{(l)} \\ \Delta\mathbf{p}_h^{(l)} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_h^{(l)} \\ \mathbf{g}_h^{(l)} \end{bmatrix}. \tag{6.11}$$

In any case, the (Newton or Picard) coefficient matrix in (6.10) or (6.11) respectively

[3] $\mathbf{X}_E^h$ is not a vector space unless its elements (which are functions) are zero on the boundary.

is nonsymmetric.[4] Thus, Krylov solvers like GMRES, BIGSTAB(ℓ) etc., will be used for solving the associated linear systems (6.10) or (6.11).

6.2.5. An optimal balanced black-box stopping test for linear solver. A natural norm for measuring errors arising from mixed FEM approximation (6.6) is $\|(\vec{u},p)\|_{\mathcal{E}} := \|\nabla\vec{u}\|_{L^2(\Omega)} + \|p\|_{L^2(\Omega)}, \forall\,(\vec{u},p) \in \mathbf{H}^1_{E_0}(\Omega) \times L^2(\Omega)$. The associated vector norm $\|\cdot\|_{E_h}$ is defined as

$$\|\mathbf{e}_h\|_{E_h} := \sqrt{\mathbf{e}_h^T E_h \mathbf{e}_h} = \sqrt{\mathbf{e}_1^T \mathbf{A}_h \mathbf{e}_1 + \mathbf{e}_2^T Q_h \mathbf{e}_2}, \qquad \forall\, \mathbf{e}_h = [\mathbf{e}_1^T, \mathbf{e}_2^T]^T \in \mathbb{R}^{n_u+n_p},$$

where $E_h := \begin{bmatrix} \mathbf{A}_h & O \\ O & Q_h \end{bmatrix}$. Here E_h is a symmetric positive-definite matrix and therefore $\|\cdot\|_{E_h}$ is a norm on $\mathbb{R}^{n_u+n_p}$. Here $Q_h = [q_{kj}]$, $q_{kj} := \int_\Omega \psi_k\psi_j\ \forall\, k,\ j = 1,\ldots,n_p$ is the pressure mass matrix. Note that by construction, for a given approximation, E_h is independent of nonlinear iterative step l. Also, observe that unlike the convection-diffusion case E_h is not simply the symmetric positive-definite part of the Navier–Stokes coefficient matrix in (6.10) or (6.11).

For any two nonnegative real numbers a and b [3, p. 213]

$$\sqrt{a+b} \le \sqrt{a} + \sqrt{b} \le \sqrt{2}\sqrt{a+b}. \tag{6.12}$$

Using (6.12), for any $(\vec{v}_h, q_h) \in \mathbf{X}^h_{E_0} \times M^h$, $\|\cdot\|_{\mathcal{E}}$ is equivalent to $\|\cdot\|_{E_h}$ in the sense that

$$\sqrt{\mathbf{v}_h^T \mathbf{A}_h \mathbf{v}_h + \mathbf{q}_h^T Q_h \mathbf{q}_h} \ \le\ \|(\vec{v}_h, q_h)\|_{\mathcal{E}} \ \le\ \sqrt{2}\sqrt{\mathbf{v}_h^T \mathbf{A}_h \mathbf{v}_h + \mathbf{q}_h^T Q_h \mathbf{q}_h}, \tag{6.13}$$

where $\mathbf{v}_h$, $\mathbf{q}_h$ are the coordinates of $\vec{v}_h$, q_h with respect to velocity and pressure basis respectively. Since, the coefficient matrix in (6.10) or (6.11) is nonsymmetric and GMRES, BICGSTAB(ℓ) etc., will be used to solve them. Hence the stopping methodology developed in section 4 can be applied in solving (6.10) or (6.11). At the lth nonlinear iteration, the iteration residual at kth step of the linear solver is

$$\mathbf{r}_h^{(l_k)} = [(\mathbf{f}_h^{(l)})^T, (\mathbf{g}_h^{(l)})^T]^T - \mathcal{F}_h^{(l)}[(\Delta\mathbf{u}_h^{(l_k)})^T, (\Delta\mathbf{p}_h^{(l_k)})^T]^T.$$

Here, $\mathcal{F}_h^{(l)}$ denotes the (Newton or Picard) nonsymmetric coefficient matrix of the (linearized) discrete Navier–Stokes system in (6.10) or (6.11) at the lth (nonlinear) iterative step. Proceeding as in section 4, at the lth nonlinear iteration, linear solvers GMRES, BICGSTAB(ℓ) etc., solving linear systems (6.10) or (6.11) are stopped at the first iteration l_{k^*} such that either holds,

$$\|\mathbf{r}_h^{(l_{k^*})}\|_2 \le \frac{\sqrt{\lambda_h^{(l)}}}{\Lambda_h^{(l)}}\eta_h^{(l_{k^*})}\ (\textbf{Strong stop}); \quad \|\mathbf{r}_h^{(l_{k^*})}\|_2 \le \frac{1}{\sqrt{\Lambda_h^{(l)}}}\eta_h^{(l_{k^*})}\ (\textbf{Weak stop}). \tag{6.14}$$

Here $\lambda_h^{(l)}$ and $\Lambda_h^{(l)}$ are the smallest and the largest eigenvalues respectively of the generalized eigenvalue problem for E_h and $(\mathcal{F}_h^{(l)})^T\mathcal{F}_h^{(l)}$.

[4] A stabilization matrix similar to the Stokes equations [3, chapter 3] is employed (for lower order finite elements) in place of the zero block of the coefficient matrix [3, p. 349].

The a posteriori error estimator $\eta_h^{(l_k)}$ that is employed in (6.14) is equivalent to the total error (approximation error at the kth iteration) in the sense that

(6.15)
$$c\,\eta_h^{(l_k)} \le \|\nabla(\delta\vec{u}^{(l)} - \delta\vec{u}_h^{(l_k)})\|_{L^2(\Omega)} + \|\delta p^{(l)} - \delta p_h^{(l_k)}\|_{L^2(\Omega)} \le C\,\eta_h^{(l_k)}, \quad \text{with } \frac{C}{c} \sim O(1).$$

At the lth iterative step $\vec{u}_h^{(l)}, p_h^{(l)}$ is known. It follows from (6.7) that $\delta\vec{u}_h^{(l)} = \vec{u}_h^{(l+1)} - \vec{u}_h^{(l)}$ and $\delta p_h^{(l)} = p_h^{(l+1)} - p_h^{(l)}$. This implies that (6.10) or (6.11) essentially solves for the basis coefficients of $(\vec{u}_h^{(l+1)}, p_h^{(l+1)})$. Thus, essentially one can use the same a posteriori approximation error estimators to estimate approximation errors a posteriori for $(\delta\vec{u}_h^{(l)}, \delta p_h^{(l)})$ as those for $(\vec{u}_h^{(l+1)}, p_h^{(l+1)})$.[5]

6.2.6. An optimal balanced black-box stopping test for nonlinear solver. Using $\|(\vec{u}_h, p_h)\|_{\mathcal{E}} := \|\nabla\vec{u}_h\|_{L^2(\Omega)} + \|p_h\|_{L^2(\Omega)}$, if $(\vec{u}, p)$ denotes the true solution, then following section 3 leads to

(6.16)
$$\|\nabla(\vec{u}_h^{(l+1)} - \vec{u})\|_{L^2(\Omega)} + \|(p_h^{(l+1)} - p)\|_{L^2(\Omega)} \le \left(\|\nabla(\vec{u}_h^{(l)} - \vec{u})\|_{L^2(\Omega)} + \|p_h^{(l)} - p\|_{L^2(\Omega)}\right) + \left(\|\nabla\delta\vec{u}_h^{(l)}\|_{L^2(\Omega)} + \|\delta p_h^{(l)}\|_{L^2(\Omega)}\right).$$

From (6.13) it follows that

$$\|\nabla\delta\vec{u}_h^{(l)}\|_{L^2(\Omega)} + \|\delta p_h^{(l)}\|_{L^2(\Omega)} \simeq \sqrt{(\Delta\mathbf{u}_h^{(l)})^T\mathbf{A}_h\Delta\mathbf{u}_h^{(l)} + (\Delta\mathbf{p}_h^{(l)})^T Q_h\Delta\mathbf{p}_h^{(l)}}. \tag{6.17}$$

In presence of 'tight' a posteriori error estimator $\eta_{h_{\text{sol}}}^{(l)}$, which is equivalent to the approximation error at the lth nonlinear iteration in the sense that

(6.18)
$$c\,\eta_{h_{\text{sol}}}^{(l)} \le \|\nabla(\vec{u}_h^{(l)} - \vec{u})\|_{L^2(\Omega)} + \|p_h^{(l)} - p\|_{L^2(\Omega)} \le C\,\eta_{h_{\text{sol}}}^{(l)}, \quad \text{with } \frac{C}{c} \sim O(1),$$

using (6.17) and (6.18) in (6.16) leads to

$$\eta_{h_{\text{sol}}}^{(l+1)} \simeq \eta_{h_{\text{sol}}}^{(l)} + \sqrt{(\Delta\mathbf{u}_h^{(l)})^T\mathbf{A}_h\Delta\mathbf{u}_h^{(l)} + (\Delta\mathbf{p}_h^{(l)})^T Q_h\Delta\mathbf{p}_h^{(l)}}. \tag{6.19}$$

Using the strong or weak stopping criterion (6.14) or any other stopping criterion for linear iteration, $(\Delta\mathbf{u}_h^{(l)}, \Delta\mathbf{p}_h^{(l)})$ is replaced by $(\Delta\mathbf{u}_h^{(l_{k^*})}, \Delta\mathbf{p}_h^{(l_{k^*})})$[6] in (6.19) which becomes

$$\eta_{h_{\text{sol}}}^{(l+1)} \simeq \eta_{h_{\text{sol}}}^{(l)} + \sqrt{(\Delta\mathbf{u}_h^{(l_{k^*})})^T\mathbf{A}_h\Delta\mathbf{u}_h^{(l_{k^*})} + (\Delta\mathbf{p}_h^{(l_{k^*})})^T Q_h\Delta\mathbf{p}_h^{(l_{k^*})}}. \tag{6.20}$$

Thus, in spirit of section 3 stop the nonlinear iteration at l^* which is the smallest value of $(l+1)$ such that

$$\sqrt{(\Delta\mathbf{u}_h^{(l^*_{k^*})})^T\mathbf{A}_h\Delta\mathbf{u}_h^{(l^*_{k^*})} + (\Delta\mathbf{p}_h^{(l^*_{k^*})})^T Q_h\Delta\mathbf{p}_h^{(l^*_{k^*})}} \le \eta_{h_{\text{sol}}}^{(l^*+1)}. \tag{6.21}$$

Note that alternative nonlinear iteration stopping strategies do exist, see [22] for more details. However, these are neither optimal nor black-box in the sense presented in (6.21).

[5] This is not a rigorous mathematical statement. A proof for this statement is an ongoing research.
[6] This k^* will in general be different for different l.

6.2.7. A posteriori error estimation. Computation of a posteriori error estimates for the Navier–Stokes mixed FEM formulation entails solving local Poisson problems for each component of velocity [3, p. 352 ff.]. In fact it has been stated in [3, proposition 8.9, p. 354] that a posteriori error estimators for stabilized $\boldsymbol{Q}_1$-$\boldsymbol{P}_0$ rectangular finite elements are reliable in the sense that the global upper bound on the approximation error does not depend on the parameters of the continuous problem. Thus, results presented in the next section are thus based on stabilized $\boldsymbol{Q}_1$-$\boldsymbol{P}_0$ rectangular finite elements.

6.2.8. Computational logistics. At the lth nonlinear iteration, $\|\mathbf{r}_h^{(l_k)}\|_2$ is readily available as a by-product of GMRES iteration. The eigenvalues $\Lambda_h^{(l)}$ and $\lambda_h^{(l)}$ involved in the (linear) stopping test (6.14) are computed using MATLAB `eigs`. Also, a cheap but an additional cost arises in computing the matrix-vector products in the left-hand-side of the nonlinear balanced stopping test (6.21).

The resulting algorithm `NAVIER_NEWTON_GMRES` is presented in Figure 4. Note that the coefficient matrix $\mathcal{F}_h^{(l)}$ is never assembled for `GMRES_Navier_balanced`. Instead intelligent matrix-vector products are carried out using the structure of $\mathcal{F}_h^{(l)}$ (see the coefficient matrix structure in (6.10) and (6.11). The same is true for any choice of a preconditioner $\mathcal{M}_h^{(l)}$. Also, a random initial guess can be used for each call of `GMRES_Navier_balanced`. Note that in practice, the a posteriori error estimate $\eta_{h_{\text{sol}}}^{(l+1)}$ should be computed (and hence the nonlinear stopping test (6.21) be tested) periodically. The algorithm in Figure 4 can easily be modified to cater to this situation. The same holds true for the (linearized) optimal balanced black-box stopping inside `GMRES_Navier_balanced`.

6.2.9. Experimental results. Results of some computational experiments in IFISS are presented in this section as a proof-of-concept. The test problem for this purpose is the flow over a backward-facing step problem; see [9], [16]. In order to illustrate the robustness of the linear and the nonlinear balanced stopping test (6.14) and (6.21) respectively, results are presented here for various values of viscosity (hence varying Reynolds number). The grid level ($h = 6$) is fixed and $\boldsymbol{Q}_1$-$\boldsymbol{P}_0$ rectangular finite elements are employed on $2^h \times (2^h \times 3)$ grid.

Since no stabilization for the convection term is inbuilt in IFISS for the Navier–Stokes equations, the a posteriori error estimator is expected to overestimate the true error. Thus, employing the weak stopping test in (6.14) for linear iterations might lead to premature stopping. Hence, the strong stopping test in (6.14) will be used here. The modified pressure convection-diffusion preconditioner [3, chapter 9] is employed as a preconditioner for all cases in the GMRES solver for solving the linear(ized) system arising at each nonlinear iterative step. Moreover, results are presented here only for the Newton iterations. However, the optimal balanced black-box stopping criterion for both linear and nonlinear iterations is applicable to Picard iterations as well. Also, note that the initial guess for the Newton iteration in each case is the (inbuilt) solution of the corresponding Stokes problem.

At each grid level and for various values of viscosity, a reference 'true' solution is computed. This is done by solving the test problem using Newton iteration to a tight nonlinear relative residual tolerance of `1e-12`. From this true solution, 'true' a posteriori error estimate $\eta_{h_{\text{sol}}}$ is computed. Also, let the difference between the true a posteriori error estimate and the computed a posteriori error estimate at the nonlinear iteration l be denoted by $e_{\eta_{h_{\text{sol}}}}^{(l_{k^*})} := |\eta_{h_{\text{sol}}}^{(l_{k^*})} - \eta_{h_{\text{sol}}}|$.

Algorithm: `NAVIER_NEWTON_GMRES`
given functions `GMRES_Navier_balanced`, `matvecA`, `matvecQ`, `Navier_error_est`

solve the corresponding Stokes problem to obtain starting guess: $(\vec{u}_h^{(0)}, p_h^{(0)})$

`for` $l = 0, 1, 2, \ldots$ `until` convergence `do`

Inner iteration (GMRES solver)
% `GMRES_Navier_balanced`: solves (6.10) or (6.11) using preconditioned GMRES with (or without) stopping test (6.14)

% Coefficient matrix $\mathcal{F}_h^{(l)}$, right-hand-side $[\mathbf{f}_h^{(l)^T}, \mathbf{g}_h^{(l)^T}]^T$, preconditioner $\mathcal{M}_h^{(l)}$

compute the vector of basis coefficients for $\delta\vec{u}_h^{(l)}$ and $\delta p_h^{(l)}$:
$[\Delta\mathbf{u}_h^{(l)^T}, \Delta\mathbf{p}_h^{(l)^T}]^T$ = `GMRES_Navier_balanced`$(\mathcal{F}_h^{(l)}, [\mathbf{f}_h^{(l)^T}, \mathbf{g}_h^{(l)^T}]^T, \mathcal{M}_h^{(l)})$

Outer iteration (Nonlinear solver)
update solution: $\vec{u}_h^{(l+1)} = \vec{u}_h^{(l)} + \delta\vec{u}_h^{(l)}, \qquad p_h^{(l+1)} = p_h^{(l)} + \delta p_h^{(l)}$

% `Navier_error_est` computes the a posteriori error estimate
compute a posteriori error estimate: $\eta_{h_{\text{sol}}}^{(l+1)}$ = `Navier_error_est` $(\vec{u}_h^{(l+1)}, p_h^{(l+1)})$

% `matvecA`$(\cdot)$, `matvecQ`$(\cdot)$ compute the action of $\mathbf{A}_h$, Q_h on a vector respectively.
stopping test:
if $\sqrt{(\Delta\mathbf{u}_h^{(l)})^T \texttt{matvecA}(\Delta\mathbf{u}_h^{(l)}) + (\Delta\mathbf{p}_h^{(l)})^T \texttt{matvecQ}(\Delta\mathbf{p}_h^{(l)})} \leq \eta_{h_{\text{sol}}}^{(l+1)}$
convergence, break l loop
endif

`enddo`

FIG. 4. *The* `NAVIER_NEWTON_GMRES` *algorithm expressed in pseudo-code.*

Similarly, on each grid level, a 'true' MATLAB backslash solution is computed for linear system arising at each step of the nonlinear iteration. From this true solution, 'true' a posteriori error estimate $\eta_h^{(l)}$ is also computed. Also, let the difference between the true a posteriori error estimate and the computed a posteriori error estimate[7] at linear stopping iteration k be denoted by $e_{\eta_h}^{(l_{k^*})} := |\eta_h^{(l_{k^*})} - \eta_h^{(l)}|$. Each linear system was also solved using GMRES to a (iteration) relative residual $\frac{\|\mathbf{r}_h^{(l_k)}\|_2}{\|\mathbf{r}_h^{(l_0)}\|_2}$ tolerance of `1e-6` and `1e-9` for a comparison with balanced stopping GMRES solver. The same preconditioner and the same initial random vector is used in all these solvers for solving any particular linear system. Also, let $l_{k_{\text{tol1}}}$, $l_{k_{\text{tol2}}}$ denote the number of iterations needed to satisfy GMRES relative residual tolerance of `1e-6` and `1e-9` respectively.

The Navier–Stokes PDE (6.5) is defined on a L-shaped (flow over a backward-facing step) domain $\Omega = (-1, 5) \times (-1, 1) \setminus (-1, 0] \times (-1, 0]$. Poiseuille flow profile is imposed on the inflow boundary ($x_1 = -1, 0 \leq x_2 \leq 1$), $\vec{x} = (x_1, x_2) \in \Omega$ and zero velocity condition is imposed on the walls. Neumann boundary conditions are defined everywhere on the outflow boundary ($x_1 = 5, -1 < x_2 < 1$). The forcing term $\vec{f}$ is zero. This problem can be generated in IFISS by choosing example `2` when running

[7] This a posteriori approximation error estimate is for the linearized part $(\delta\vec{u}_h^{(l_k)}, \delta p_h^{(l_k)})$.

the driver `navier_testproblem` [3, p. 335]. The balanced stopping test in GMRES is implemented in IFISS function `gmres_r` while the nonlinear balanced stopping test is incorporated in the function `solve_step_navier` in IFISS.

TABLE 5
Navier–Stokes test problem with Newton iteration on a $2^h \times (2^h \times 3)$ ($h = 6$) grid with $\nu = 1/50$.

l	$l_{k_{\text{tol1}}}$	$l_{k_{\text{tol2}}}$	l_{k^*}	$e_{\eta_h}^{(l_{k^*})}$	$e_{\eta_{h_{\text{sol}}}}^{(l_{k^*})}$	$\Lambda_h^{(l)}$	$\lambda_h^{(l)}$
1	29	39	23	3.6e-05	3.1e-02	2.6e+05	2.2e-01
2	38	48	33	7.1e-08	3.3e-04	4.4e+05	2.2e-01
3	42	53	36	1.3e-09	5.2e-07	4.0e+05	2.1e-01

In Tables 5 and 6, a comparison of $l_{k_{\text{tol1}}}$, $l_{k_{\text{tol2}}}$ numbers with the corresponding l_{k^*} values shows that employing the (strong) linear stopping test (6.14) leads to savings in iteration counts. In each table at the lth Newton iteration and at the linear optimal balanced black-box stopping iteration l_{k^*}, $e_{\eta_h}^{(l_{k^*})}$ columns show that the preconditioned GMRES solution of the linearized part has converged with some accuracy to the true linearized solution. In other words, $\{\eta_h^{(l_k)}\}$ has converged with some accuracy to true $\eta_h^{(l)}$. At the nonlinear balanced stopping iteration l^*, $e_{\eta_{h_{\text{sol}}}}^{(l_{k^*})}$ columns exhibit convergence with some accuracy of $\{\eta_{h_{\text{sol}}}^{(l_{k^*})}\}$ to the true a posteriori approximation error estimate $\eta_{h_{\text{sol}}}$.

The eigenvalues $\Lambda_h^{(l)}$, $\lambda_h^{(l)}$ used in the linear stopping criterion are also tabulated in these tables. These numbers exhibit some structure thereby suggesting that there might be an expression for these quantities in terms of the parameters of the problem. However, this aspect has not been investigated in this work.

TABLE 6
Navier–Stokes test problem with Newton iteration on a $2^h \times (2^h \times 3)$ ($h = 6$) grid with $\nu = 1/100$.

l	$l_{k_{\text{tol1}}}$	$l_{k_{\text{tol2}}}$	l_{k^*}	$e_{\eta_h}^{(l_{k^*})}$	$e_{\eta_{h_{\text{sol}}}}^{(l_{k^*})}$	$\Lambda_h^{(l)}$	$\lambda_h^{(l)}$
1	35	46	46	2.0e-05	2.9e-01	9.0e+05	8.5e-02
2	53	66	52	2.0e-08	7.1e-03	4.9e+06	8.5e-02
3	55	68	52	1.8e-08	2.2e-04	3.0e+06	8.5e-02
4	59	73	55	8.2e-10	8.4e-07	2.6e+06	8.5e-02

Evolution of errors with iteration number are plotted in Figure 5 at 4th Newton iteration on 64×192 grid for $\nu = 1/100$. On the plot for linear iteration observe that at the optimal balanced black-box linear stopping iteration l_{k^*}, the curve for $\eta_h^{(l_k)}$ converges with some accuracy to the line for $\eta_h^{(l)}$. This convergence is further illustrated by continuing for 9 more iterations after balanced linear stopping. Note that $\{\eta_h^{(l_k)}\}$ converges to $\eta_h^{(l)}$ when $\|\mathbf{e}_h^{(l_k)}\|_{E_h}$ curve goes below the (black) line for $\eta_h^{(l)}$. However, as mentioned earlier, for a given approximation, the iteration error $\mathbf{e}_h^{(l_k)} = [(\Delta\mathbf{u}_h^{(l)})^T, (\Delta\mathbf{p}_h^{(l)})^T]^T - [(\Delta\mathbf{u}_h^{(l_k)})^T, (\Delta\mathbf{p}_h^{(l_k)})^T]^T$ is rarely known a priori. Also, on the plot for Newton iteration (right) notice that at the optimal balanced black-box stopping nonlinear iteration number four, the curve for $\eta_{h_{\text{sol}}}^{(l_{k^*})}$ converges with some accuracy to the line for the true a posteriori approximation error estimate $\eta_{h_{\text{sol}}}$. This convergence is further illustrated by continuing for 2 more iterations after optimal balanced black-box nonlinear stopping.

As mentioned earlier, currently, the eigenvalues for optimal stopping of linear iterative solvers for solving nonsymmetric linear system (at each nonlinear iterative

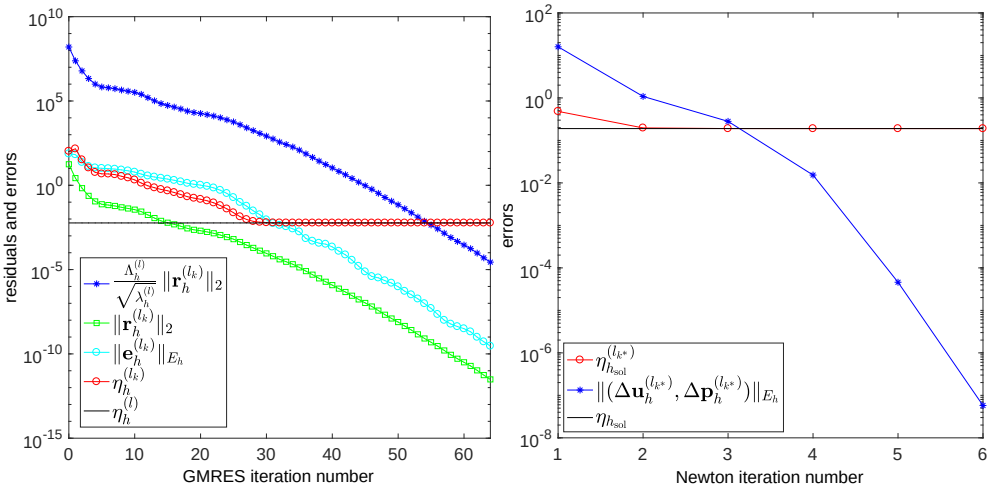

FIG. 5. *Errors vs iteration number for Navier–Stokes test problem on a* 64×192 *grid with* $\nu = 1/100$ *for Newton iteration (right) and GMRES iteration (left) at* $l = 4$*th Newton iteration.*

step) cannot be estimated cheaply on-the-fly as compared to those for symmetric linear systems. So, the cost of computing the eigenvalues for the linear stopping test may offset the computational savings (in terms of number of linear solver iterations for convergence) if large number of nonlinear iterations are required. Hence, it is prudent here to use only the optimal balanced black-box nonlinear stopping test (6.21).

7. Conclusions. In this paper, optimal balanced black-box stopping criteria have been devised in linear (GMRES, suboptimal solvers like BICGSTAB(ℓ), TFQMR etc.) solvers for nonsymmetric linear(ized) systems arising from FEM approximation of convection-diffusion equations and Navier–Stokes equations. Moreover, an optimal balanced black-box stopping criterion for nonlinear (Newton or Picard) iterations for solving Navier–Stokes equations has also been derived. Using optimal balanced black-box stopping tests may not only save unnecessary computational but also rules out premature stopping of the employed linear and/or nonlinear iterative solvers.

The optimal balanced black-box stopping strategies presented here are quite generic. They can be suitably modified to cater for varied linear and nonlinear iterative procedures for solving nonsymmetric linear(ized) systems arising from numerical approximation of a PDE. This is provided cheap and tight a posteriori (or a priori) approximation error estimators are available along with cheap tractable bounds on the relevant errors that are generally difficult to compute.

REFERENCES

[1] M. Arioli, D. Loghin, and A. J. Wathen, *Stopping criteria for iterations in finite element methods*, Numer. Math., 99 (2005), pp. 381–410. https://doi.org/10.1007/s00211-004-0568-z.

[2] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer, USA, 2008. Third Edition.

[3] H. Elman, D. Silvester, and A. Wathen, *Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics*, Oxford University Press, UK, 2014. Second Edition.

[4] H. C. Elman, A. Ramage, and D. J. Silvester, *IFISS: A computational laboratory for investigating incompressible flow problems*, SIAM Review, 56 (2014), pp. 261–273. https:

//doi.org/10.1137/120891393.

[5] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Computational Differential Equations*, Cambridge University Press, USA, 1996. First Edition.

[6] R. W. Freund, *A transpose-free quasi-minimal residual algorithm for non-hermitian linear systems*, SIAM J. Sci. Comput., 14 (1993), pp. 470–482. https://doi.org/10.1137/0914029.

[7] R. W. Freund, M. H. Gutknecht, and N. M. Nachtigal, *An implementation of the look-ahead Lanczos algorithm for non-hermitian matrices*, SIAM J. Sci. Comput., 14 (1993), pp. 137–158. https://doi.org/10.1137/0914009.

[8] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The John Hopkins University Press, USA, 2013. Fourth Edition.

[9] P. M. Gresho, D. K. Gartling, J. R. Torczynski, K. A. Cliffe, K. H. Winters, T. J. Garratt, A. Spence, and J. W. Goodrich, *Is the steady viscous incompressible two-dimensional flow over a backward-facing step at re = 800 stable?*, Int. J. Numer. Methods Fluids, 17 (1993), pp. 501–541. https://doi.org/10.1002/fld.1650170605.

[10] T. J. R. Hughes and A. Brooks, *A multi-dimensional upwind scheme with no crosswind diffusion*, In: Finite Element Methods for Convection Dominated Flows, ASME Winter Annual Meeting, T. Hughes (Ed.), New York, USA, 34 (1979), pp. 19–35. https://www.researchgate.net/publication/297681092.

[11] P. Jiránek, Z. Strakos, and M. Vohralík, *A posteriori error estimates including algebraic error and stopping criteria for iterative solvers*, SIAM J. Sci. Comput., 32 (2010), pp. 1567–1590. https://doi.org/10.1137/08073706X.

[12] J. T. Oden and L. F. Demkowicz, *Applied Functional Analysis*, CRC Press, USA, 1996. First Edition.

[13] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629. https://doi.org/10.1137/0712047.

[14] D. A. D. Pietro, E. Flauraud, M. Vohralík, and S. Yousef, *A posteriori error estimates, stopping criteria, and adaptivity for multiphase compositional refinement for thermal multiphase compositional flows in porous media*, Journal of Comp. Phy., 276 (2014), pp. 163–187. https://doi.org/10.1016/j.jcp.2014.06.061.

[15] D. A. D. Pietro, M. Vohralík, and S. Yousef, *An a posteriori-based, fully adaptive algorithm with adaptive stopping criteria and mesh refinement for thermal multiphase compositional flows in porous media*, Comput. Math. Appl., 68 (2014), pp. 2331–2347. https://doi.org/10.1016/j.camwa.2014.08.008.

[16] C. E. Powell and D. J. Silvester, *Preconditioning steady-state Navier–Stokes equations with random data*, SIAM J. Sci. Comput., 34 (2012), pp. A2482–A2506. https://doi.org/10.1137/120870578.

[17] Pranjal, *Optimal iterative solvers for linear systems with stochastic PDE origins: Balanced black-box stopping tests, PhD thesis*, University of Manchester, UK, 2017. http://eprints.maths.manchester.ac.uk/2596/.

[18] Y. Saad and M. Schultz, *A generalized minimal residual algorithm for solving nonsymmetric linear systems.*, SIAM J. Sci. Comput., 7 (1986), pp. 856–869. https://doi.org/10.1137/0907058.

[19] G. I. Shishkin, *Methods of constructing grid approximations for singularly perturbed boundary-value problems. Condensing grid methods*, Russian J. Numer. Anal. Math. Modelling, 7 (1992), pp. 537–562. https://doi.org/10.1515/rnam.1992.7.6.537.

[20] D. Silvester and Pranjal, *An optimal solver for linear systems arising from stochastic FEM approximation of diffusion equations with random coefficients*, SIAM/ASA J. Uncertainty Quantification, 4 (2016), pp. 298–311. https://doi.org/10.1137/15M1017740.

[21] G. L. G. Sleijpen and D. R. Fokkema, *BICGSTAB(L) for linear equations involving unsymmetric matrices with complex spectrum*, Elec. Trans. Numer. Anal., 1 (1993), pp. 11–32. https://etna.mcs.kent.edu/vol.1.1993/pp11-32.dir/pp11-32.pdf.

[22] Syamsudhuha and D. J. Silvester, *Efficient solution of the steady-state Navier–Stokes equations using a multigrid preconditioned Newton–Krylov method*, Int. J. Numer. Methods Fluids, 43 (2003), pp. 1407–1427. https://doi.org/10.1002/fld.627.

[23] R. Verfürth, *A Posteriori Error Estimation Techniques for Finite Element Methods*, Oxford University Press, UK, 2013. First Edition.

[24] A. Wathen, *Preconditioning and convergence in the right norm*, Int. J. Comput. Math., 84 (2007), pp. 1199–1209. https://doi.org/10.1080/00207160701355961.

[25] C. T. Wu, *On the implementation of an accurate and efficient solver for convection-diffusion equations, PhD thesis*, University of Maryland, USA, 2003. https://drum.lib.umd.edu/handle/1903/32.