BALANCED ITERATIVE SOLVERS FOR LINEAR SYSTEMS ARISING FROM FINITE ELEMENT APPROXIMATION OF PDES

PRANJAL*

Abstract. This paper discusses the design and implementation of efficient solution algorithms for symmetric and nonsymmetric linear systems associated with finite element approximation of partial differential equations. The novel feature of our preconditioned MINRES solver (for symmetric systems) and preconditioned GMRES solver (for nonsymmetric systems) is the incorporation of error control in the natural norm (associated with the specific approximation space) in combination with a reliable and efficient a posteriori estimator for the PDE approximation error. This leads to a robust and balanced inbuilt stopping criterion: the iteration is terminated as soon as the algebraic error is insignificant compared to the approximation error.

Key words. Stochastic Bubnov–Galerkin approximation, parametric diffusion equation, Petrov-Galerkin approximation, convection-diffusion equation, a posteriori error analysis, iterative solvers, MINRES, GMRES, preconditioning.

1. Introduction. Mathematical models of many real-world phenomena often formulate as partial differential equations (PDEs) with boundary and/or initial conditions. Finite element methods are powerful tools for computing numerical solution of PDEs. Galerkin finite element approximation in space leads to a single linear(ized) system of equations whose coefficient matrix is ill-conditioned with respect to discretization parameters. Since the coefficient matrix has a well-defined sparse structure, iterative solution strategies can be effective nevertheless. The choice of iterative solver for solving the linear system depends on the nature of the coefficient matrix. Bubnov–Galerkin finite element (use of same trial and test functions) approximation of a diffusion equation leads to a symmetric positive definite linear system while Petrov–Galerkin finite element (use of different trial and test functions) approximation usually leads to a nonsymmetric linear system. Although the preconditioned conjugate gradient (CG) methods are commonly used to solve symmetric positive definite linear systems we will use the classic MINRES (minimal residual) algorithm of Paige & Saunders [12] to solve them. To solve nonsymmetric linear systems we will use the GMRES (generalized minimal residual) algorithm of Saad & Schultz [15].

Wathen [18] observed that finite element approximation of PDEs endows the problem with a 'natural' norm that is determined by the specific approximation space. Thus, in finite element setting, the PDE approximation error, the linear algebra error (algebraic error) and the total error at any iteration step are measured in this 'natural' norm. Henceforth, any reference to these errors will imply these errors measured in the 'natural' norm.

This paper investigates devising a 'balanced' inbuilt stopping criterion for iterative solvers of nonsymmetric and symmetric positive definite linear systems with PDE origins. This is an active research field (see Silvester & Simoncini [17], Arioli et al. [2]). The essence of the balanced stopping test is—the total error at any iteration step is approximately the sum of the approximation error and the algebraic error. For chosen discretization and problem parameters the approximation error is fixed but usually unknown. By balancing the algebraic error and the total error a 'balanced' inbuilt stopping test is obtained. In the finite element setting, the total error and the approximation error can be estimated using reliable and efficient a posteriori

*School of Mathematics, University of Manchester, Manchester, M13 9PL, United Kingdom, pranjal.pranjal@manchester.ac.uk

error estimators. Obtaining 'effective' upper and lower bounds on the algebraic error in terms of the readily computable and monotonically decreasing quantities of our chosen iterative solver is the novel feature of our stopping strategy.

The paper is structured as follows. We set up our symmetric positive definite linear algebra problem in section 2. This section also develops the rationale of our balanced stopping criterion. In section 3 using the S–IFISS toolbox [16] we present some computational results based on our stopping test for the target symmetric positive definite linear system. We set up our nonsymmetric linear system in section 4. In this section we also develop our balanced stopping test for the target nonsymmetric linear system. In section 5 we present a set of computational results that can be reproduced using the IFISS toolbox [7] and which confirm the effectiveness of our balanced stopping test. The section 6 contains the conclusions. Throughout the discussion $\mathbb{R}$ will denote the set of real numbers.

2. Parameter dependent linear systems. Parameterized linear systems are ubiquitous (see Butler et al. [4]). Here the entries of the matrix A and the solution vector u depend upon a finite number $m \in \{1, 2, \ldots\}$ of parameters $\mathbf{y} = [y_1, y_2, \ldots, y_m]$.

$$A(\mathbf{y})\, u(\mathbf{y}) = f.$$

Symmetric systems of this type arise in the solution of linear elliptic PDEs with random coefficients. An example might be a heat conduction problem in a region with m different materials: each having thermal conductivity coefficient that is not known precisely. This can be modelled as a stochastic steady-state diffusion PDE.[1]

2.1. Stochastic steady-state diffusion PDE. The stochastic steady diffusion equation is a model PDE which fits into the above framework. Let us suppose that the steady diffusion process is defined in a spatial domain $D \subset \mathbb{R}^d$, with an isotropic permeability tensor $K = \kappa I$ where $\kappa : D \times \Gamma \to \mathbb{R}$ is parameterized by m i.i.d. centered random variables, so that

$$\kappa(\vec{x}, y_1, \ldots, y_m) := \mu(\vec{x}) + \sigma \sum_{k=1}^{m} \psi_k(\vec{x})\, y_k. \tag{2.1}$$

Here $\mu(\vec{x})$ is the mean value of the permeability coefficient at the point $\vec{x} \in D, \sigma$ is the standard deviation of the parameter variation, $y_k \in \Gamma_k$ is the image of the kth random variable, $\Gamma := \Gamma_1 \times \cdots \times \Gamma_m$ and $\{\psi_k\}_{k=1}^m$ are given functions defined on D.

The associated solution $u(\vec{x}, \mathbf{y}) : D \times \Gamma \to \mathbb{R}$ satisfies almost surely

$$-\nabla \cdot K(\vec{x}, \mathbf{y})\, \nabla u(\vec{x}, \mathbf{y}) = f(\vec{x}), \qquad \vec{x} \in D \subset \mathbb{R}^d,\ (d = 2, 3),\ \mathbf{y} \in \Gamma, \tag{2.2a}$$

$$u(\vec{x}, \mathbf{y}) = g(\vec{x}), \qquad \vec{x} \in \partial D_D,\ \mathbf{y} \in \Gamma, \tag{2.2b}$$

$$K(\vec{x}, \mathbf{y})\, \nabla u(\vec{x}, \mathbf{y}) \cdot \vec{n} = 0, \qquad \vec{x} \in \partial D_N = \partial D \backslash \partial D_D,\ \mathbf{y} \in \Gamma, \tag{2.2c}$$

where ∂D_D, ∂D_N are the Dirichlet and the Neumann part of the spatial boundary ∂D. The vector $\vec{n}$ is the outward normal to ∂D. The source function f and the boundary data g are given deterministic functions.

[1] Stochastic Galerkin finite element (see Deb et al. [5]) approximation of a parameterized PDE results in a huge linear system. Since the existing storage requirements and computational flops increase with the size of a linear system, an inbuilt balanced stopping test will significantly reduce the computational work.

The weak formulation of (2.2) is to find u such that $u - \hat{g} \in W$ satisfies

$$\int_\Gamma \rho(\mathbf{y}) \int_D K(\vec{x},\mathbf{y})\, \nabla u(\vec{x},\mathbf{y}) \cdot \nabla w(\vec{x},\mathbf{y})\, \mathrm{d}\vec{x}\, \mathrm{d}\mathbf{y} = \int_\Gamma \rho(\mathbf{y}) \int_D f(\vec{x}) w(\vec{x},\mathbf{y})\, \mathrm{d}\vec{x}\, \mathrm{d}\mathbf{y}, \tag{2.3}$$

for all $w \in W$ (the space is defined below). Here $\hat{g}$ is a smooth extension of g into the domain and $\rho(\mathbf{y})$ denotes the joint probability density function defined on the product set Γ of a multivariate random variable defined on a probability space[2] $(\Gamma, \mathcal{B}(\Gamma), \pi)$.

Note that the left side of (2.3) characterizes the energy norm

$$\|w\|_E^2 := \int_\Gamma \rho(\mathbf{y}) \int_D K(\vec{x},\mathbf{y})\, |\nabla w(\vec{x},\mathbf{y})|^2\, \mathrm{d}\vec{x}\, \mathrm{d}\mathbf{y}, \tag{2.4}$$

so that the solution space is $W := \{u : \|u\|_E < \infty,\, u|_{\partial D_D \times \Gamma} = 0\} = H_0^1(D) \otimes L^2(\Gamma)$.

Stochastic Galerkin approximation of (2.3) is associated with choosing finite dimensional subspaces of the component spaces, that is $X_h \subset H_0^1(D)$, $S_p \subset L^2(\Gamma)$ and setting $X_h \otimes S_p =: W_{h,p} \subset W$ (see Lord et al. [11, section 9.5]). The parameter approximation space S_p consists of global (multivariate) polynomials of total polynomial degree $\leq p$ in the m parameters. The choice for X_h in a two-dimensional spatial domain is the usual piecewise bilinear ($\boldsymbol{Q}_1$) or biquadratic ($\boldsymbol{Q}_2$) finite element approximation.

This leads to the huge linear system with a Kronecker product ($\otimes$) structure

$$\mathcal{A}\mathbf{x} = \mathbf{f} \quad \Longleftrightarrow \quad \big(I \otimes A_0 + \sigma \textstyle\sum_{k=1}^m G_k \otimes A_k\big)\mathbf{x} = \mathbf{f}. \tag{2.5}$$

The matrices A_0 and A_k are essentially the sparse finite element stiffness matrices while G_k are the weighted gram matrices associated with the kth parameter. For the space S_p, it is sensible to choose a basis set $\{\xi_j\}_{j=1}^{n_\xi}$ that is orthonormal with respect to the probability measure π. This leads to sparse matrices G_k ($G_0 = I$, at most two nonzeros in any row otherwise) and means that matrix-vector products with the coefficient matrix $\mathcal{A}$ in (2.5) are cheap to compute—an essential ingredient for a computationally efficient iterative solution strategy. We note that if (2.3) is well posed, $\mathcal{A}$ is a symmetric positive-definite matrix. Also, since $\mathcal{A}$ is usually ill-conditioned with respect to stochastic and finite element discretization parameters, a preconditioner $\mathcal{M}$ is required. If σ is small relative to $\|A_0\|$, the positive-definite matrix $\mathcal{M}^{-1} := I \otimes A_0$ will be a close approximation to $\mathcal{A}$. This is known as mean based preconditioning see Powell & Elman [14].

2.2. A posteriori error estimation. A measure that is equally important is the energy norm of the solution when the permeability coefficient is given by the mean field, that is

$$\|w\|_{E_0}^2 := \int_\Gamma \rho(\mathbf{y}) \int_D \mu(\vec{x})\, |\nabla w(\vec{x},\mathbf{y})|^2\, \mathrm{d}\vec{x}\, \mathrm{d}\mathbf{y} \tag{2.6}$$

The key point here is that the two norms are equivalent whenever the formulation (2.3) is well posed (see Bespalov et al. [3, Proposition 4.2]); that is, there exist positive constants λ and Λ such that

$$\lambda \|w\|_{E_0}^2 \leq \|w\|_E^2 \leq \Lambda \|w\|_{E_0}^2 \qquad \forall w \in W. \tag{2.7}$$

[2]The triple $(\Gamma, \mathcal{B}(\Gamma), \pi)$ is assumed to define a probability space, see Lord et al. [11, section 4.1].

Let u, u_{hp}, $u_{hp}^{(k)}$ denote the true solution, exact numerical solution and numerical solution at the kth step of our iterative solver respectively. Then using the local problem error estimation strategy developed in [3, Lemma 4.1], one can approximate the mean energy error $\|u - u_{hp}^{(k)}\|_{E_0}$ a posteriori. In light of (2.7) one can compute an a posteriori error estimate $\eta^{(k)}$ for the total error at iteration k in the sense that [3, Theorem 4.1]

$$c\,\eta^{(k)} \;\leq\; \|u - u_{hp}^{(k)}\|_E \leq\; C\,\eta^{(k)}, \quad \text{with } \frac{C}{c} \sim O(1). \tag{2.8}$$

By Galerkin orthogonality at iteration k

$$\underbrace{\|u - u_{hp}^{(k)}\|_E^2}_{\text{total error}} \;=\; \underbrace{\|u - u_{hp}\|_E^2}_{\text{approximation error}} \;+\; \underbrace{\|u_{hp} - u_{hp}^{(k)}\|_E^2}_{\text{algebraic error}}. \tag{2.9}$$

Thus, assuming the a posteriori energy estimates $\eta^{(k)}$ and η are close (reliable and efficient)[3] estimates of the total error and the approximation error at the kth iteration step, (2.9) can be rewritten as

$$(\eta^{(k)})^2 \;\simeq\; \eta^2 \;+\; \|\mathbf{e}^{(k)}\|_{\mathcal{A}}^2, \tag{2.10}$$

where $\|u_{hp} - u_{hp}^{(k)}\|_E^2 \;=\; \|\mathbf{e}^{(k)}\|_{\mathcal{A}}^2 \;:=\; {\mathbf{e}^{(k)}}^T \mathcal{A}\, \mathbf{e}^{(k)} \;=\; {\mathbf{r}^{(k)}}^T \mathcal{A}^{-1}\, \mathbf{r}^{(k)}$,
and $\mathbf{e}^{(k)}$, $\mathbf{r}^{(k)}$ denote the iteration error and the residual at the kth iteration step.

The equivalence relation ($\simeq$) follows directly from (2.8). Note that the approximation error is fixed for chosen stochastic and spatial parameters. Thus, we are essentially constructing a sequence $\{\eta^{(k)}\}$ converging to η. A balanced stopping point is when the contribution of the algebraic error in the sum (2.10) becomes insignificant. So, we stop at iteration k^*, the smallest value of k such that

$$\|\mathbf{e}^{(k^*)}\|_{\mathcal{A}} \;\leq\; \eta^{(k^*)}. \tag{2.11}$$

Although preconditioned CG is used for solving symmetric positive definite linear systems, devising balanced stopping criterion using (2.11) is not easy [10]. The exact algebraic error is usually unknown. Computation of the algebraic error at any iteration step k using quantities available at iteration $k + d$ of preconditioned CG have been devised (see Arioli [1]).[4] But an optimal choice of the 'delay' parameter d for a generic problem is still an open question. Thus, we devise a balanced stopping test for preconditioned MINRES to solve symmetric positive definite linear systems.

2.3. A balanced stopping test for MINRES. In preconditioned MINRES, the quantity $\|\mathbf{r}^{(k)}\|_{\mathcal{M}} := \sqrt{{\mathbf{r}^{(k)}}^T \mathcal{M} \mathbf{r}^{(k)}}$[5] is readily computable and monotonically decreasing. We will obtain bounds for $\|\mathbf{e}^{(k)}\|_{\mathcal{A}}$ in terms of $\|\mathbf{r}^{(k)}\|_{\mathcal{M}}$. Since algebraic error $\|\mathbf{e}^{(k)}\|_{\mathcal{A}}^2 = {\mathbf{r}^{(k)}}^T \mathcal{A}^{-1} \mathbf{r}^{(k)}$, this involves calculating the Rayleigh quotient (see Golub & Van Loan [8, p. 453]) bounds for $\mathcal{A}^{-1}$ and $\mathcal{M}$. This is equivalent to calculating the largest (Θ) and the smallest (θ) eigenvalue of the preconditioned matrix $\mathcal{M}\mathcal{A}$.

$$\frac{1}{\Theta} \;\leq\; \frac{{(\mathbf{r}^{(0)})}^T \mathcal{A}^{-1}\, \mathbf{r}^{(0)}}{{(\mathbf{r}^{(0)})}^T \mathcal{M}\, \mathbf{r}^{(0)}}, \quad \frac{{(\mathbf{r}^{(k)})}^T \mathcal{A}^{-1}\, \mathbf{r}^{(k)}}{{(\mathbf{r}^{(k)})}^T \mathcal{M}\, \mathbf{r}^{(k)}} \;\leq\; \frac{1}{\theta}, \tag{2.12}$$

[3] This can be seen from the column for effectivity index η/(exact approximation error) in Tables 1 and 2 [3]. The effectivity is quite close to 1 thereby indicating that η is reliable and efficient.

[4] We have also devised a balanced stopping test for preconditioned CG using (2.11). However, we do not mention the results in this paper.

[5] Here $\|\cdot\|_{\mathcal{M}}$ indeed defines a norm since $\mathcal{M}$ is always a symmetric positive-definite matrix.

$$\frac{\|\mathbf{e}^{(k)}\|_{\mathcal{A}}}{\|\mathbf{e}^{(0)}\|_{\mathcal{A}}} \leq \sqrt{\frac{\Theta}{\theta}}\,\frac{\|\mathbf{r}^{(k)}\|_{\mathcal{M}}}{\|\mathbf{r}^{(0)}\|_{\mathcal{M}}} \Longleftrightarrow \|\mathbf{e}^{(k)}\|_{\mathcal{A}} \leq \frac{\sqrt{\Theta}}{\theta}\,\|\mathbf{r}^{(k)}\|_{\mathcal{M}}. \tag{2.13a}$$

$$\|\mathbf{e}^{(k)}\|_{\mathcal{A}} \leq \frac{1}{\sqrt{\theta}}\,\|\mathbf{r}^{(k)}\|_{\mathcal{M}}. \tag{2.13b}$$

The tighter bound (2.13b) will be used here. From (2.11) and (2.13b), we stop at iteration k^*, which is the smallest value of k such that

$$\frac{1}{\sqrt{\theta}}\,\|\mathbf{r}^{(k^*)}\|_{\mathcal{M}} \leq \eta^{(k^*)}. \tag{2.14}$$

3. Computational results. To provide a proof of concept, we present the results of computational experiments when stopping test (2.14) is applied to preconditioned symmetric positive definite linear systems arising from the model problem (2.2).

Following Deb et al. [5] the PDE problem (2.2) will be defined on a square domain $D = (-1,1) \times (-1,1)$ with zero Dirichlet boundary conditions everywhere on the boundary and $f(x_1, x_2) = \frac{1}{8}(2 - x_1^2 - x_2^2), \forall\,(x_1, x_2) \in D$. Rectangular $\boldsymbol{Q}_1$ (piecewise bilinear) finite elements are used on a uniform grid with mesh size $h = 2^{1-\ell}, \ell = 3, 4, 5, 6$. The diffusion coefficient κ in (2.1) is parameterized by uniform random variables defined on $\Gamma_k = [-1, 1]$, and the parameter approximation space S_p is spanned by complete Legendre polynomials of degree $p = 3$. The mean field in the expansion (2.1) is constant, $\mu(\mathbf{x}) = 1$. The correlation length is 2 in each coordinate direction and the spatial functions $\psi_k = \sqrt{3\lambda_k}\,\varphi_k$ in (2.1) are associated with eigenpairs $\{(\lambda_k, \varphi_k)\}_{k=1}^{m}$ of the (separable) covariance operator[6] $C(\vec{x}, \vec{x}') = \sigma^2 \exp\left(-\frac{1}{2}\|\vec{x} - \vec{x}'\|_{\ell_1}\right),\ \vec{x}, \vec{x}' \in D \subset \mathbb{R}^2$. We present results for different number of random variables m, standard deviation σ and mesh parameter h. A reference algebraic solution $\mathbf{x}$ can be computed in each case by solving the preconditioned discrete system with an absolute preconditioned residual ($\|\mathbf{r}^{(k)}\|_{\mathcal{M}}$) reduction tolerance of `1e-14`. Corresponding to this highly accurate solution $\mathbf{x}$, a reference a posteriori error estimate η can also be generated. The initial vector $\mathbf{x}^{(0)}$ for the solver is generated using MATLAB function `rand`.[7]

Representative results are presented in Figure 3.1. Each subplot shows the evolution of $\|\mathbf{r}^{(k)}\|_{\mathcal{M}}$, a posterior error estimator $\eta^{(k)}$ and $\frac{1}{\sqrt{\theta}}\,\|\mathbf{r}^{(k)}\|_{\mathcal{M}}$ at each iteration step k; with θ estimated[8] on the fly[9] as the smallest Ritz value $\theta^{(k)}$ of the tridiagonal Lanczos matrix in the Lanczos process of preconditioned MINRES; full details can be found in Greenbaum [9, section 2.5] The extremal Ritz values provide an accurate estimate of the extremal eigenvalues, even when iteration number k is small (see Parlett [13, chapter 13]). It can be seen that the sequence $\{\eta^{(k)}\}$ converges to the reference a posteriori error η on each plot. Note that we have taken 9 more extra iterations after convergence to check stopping at the correct place. Though we have computed $\eta^{(k)}$ at each step here to illustrate our method, in practice it should be

[6] The problem can be set up in S–IFISS by selecting example 2 in the driver `stoch_diff_testproblem`.

[7] The same initial vector is used for generating the reference solution and the algebraic solution based on our stopping test (2.14).

[8] Since the size of the linear system is huge, the matrix is neither computed/assembled in a practical implementation. So, MATLAB function `eig/eigs` cannot be used here to compute eigenvalues.

[9] Also we know that $\lambda \leq \theta$ and $\Theta \leq \Lambda$. This is not useful information in general, since a priori estimates of λ and Λ are pessimistic and/or hard to find.

computed periodically (for example, every 4-5 iterations) to have a minor impact on the overall algorithmic cost. Notice that the curves for $\|\mathbf{r}^{(k)}\|_{\mathcal{M}}$ and $\frac{1}{\sqrt{\theta^{(k)}}}\|\mathbf{r}^{(k)}\|_{\mathcal{M}}$ are not parallel initially but soon become parallel as $\theta^{(k)}$ converges to θ. The plots in Figure 3.2 further confirm this observation. In fact our computational experiments suggest no sign of the problematic 'ghost' (spurious) eigenvalues [8, p. 566] in any of these computations (see Figure 3.2).

To show the effectiveness of our stopping test (for various problem parameters), the iteration counts k^* needed to satisfy the stopping test (2.14) have been compared in Table 3.1 with iteration counts k_1 needed to satisfy a fixed absolute residual ($\|\mathbf{r}^{(k)}\|_{\mathcal{M}}$) reduction tolerance of `1e-5`. This tolerance value is a realistic user-input tolerance choice in the absence of inbuilt stopping test (2.14).[10] The table indicates that the number of iterations for convergence remains bounded even as the spatial grid is refined. This reconfirms that the mean based preconditioner $\mathcal{M}$ is spectrally equivalent to $\mathcal{A}$. Also when σ is increased, θ becomes increasingly smaller (close to zero).[11] The number of iterations for convergence based on (2.14) in the Table 3.1.1 is at least twice less as compared to those in Tables 3.1.2–3.1.3. Let $\eta^{(k^*)}$ be the corresponding a posteriori error estimate at the optimal stopping iteration k^* and $e_{\eta^*} := |\eta - \eta^{(k^*)}|$. The e_{η^*} columns in Table 3.1 show that $\eta^{(k^*)}$ has converged with a 'good' accuracy to the reference a posteriori error estimate η. Comparing the columns for iteration counts, we find that for the same approximation error we save a significant number of iterations by using our stopping test as compared to iteration counts k_1. This would result in significant savings in computational work of the solver (as compared to using fixed absolute residual ($\|\mathbf{r}^{(k)}\|_{\mathcal{M}}$) reduction tolerance of `1e-5` or tighter) if one were to solve the discrete preconditioned linear systems arising from adaptive finite element for the problem parameters. The number of degrees of freedom (#dof) of the resulting finite dimensional space which is equal to $\frac{(m+p)!}{m!p!}(2^l-1)^2$ is also tabulated. The savings in the computational work of the iterative solver becomes further significant in light of the huge size of these linear systems.

TABLE 3.1
Iteration counts for various problem parameters (Tables 3.1.1–3.1.3 (left–right))

ℓ	$\sigma = 0.3$, $m = 3$ and $p = 3$				$\sigma = 0.5$, $m = 3$ and $p = 3$				$\sigma = 0.5$, $m = 7$ and $p = 3$			
	k_1	k^*	e_{η^*}	#dof	k_1	k^*	e_{η^*}	#dof	k_1	k^*	e_{η^*}	#dof
3	12	6	7.2e-5	2744	26	11	2.2e-3	2744	33	13	7.8e-4	5880
4	13	7	2.1e-5	12600	30	14	5.5e-5	12600	43	18	8.9e-4	27000
5	13	8	1.5e-5	53816	31	16	2.5e-4	53816	48	22	1.2e-3	115320
6	14	9	7.2e-6	222264	33	17	7.3e-4	222264	52	26	2.5e-4	476280

4. A balanced stopping test for GMRES. In this section we will develop a balanced stopping test based on GMRES for solving nonsymmetric linear systems. The underlying PDE will be the deterministic convection-diffusion equation.[12]

[10] The user will not know in general the stopping point k^* a priori and is likely to provide a tighter tolerance than actually required. This would lead to wastage of computational work and time.

[11] Sharp bounds $[1-\tau, 1+\tau]$ for the Rayleigh quotient are established by Powell & Elman in [14, Theorem 3.8], where the factor τ is the sum of the norms $\|\psi_k\|_\infty$ of the functions in (2.1).

[12] The choice of *deterministic* convection-diffusion equation has been made because efficient and reliable a posteriori error estimators for parameteric version of this PDE have not been devised.

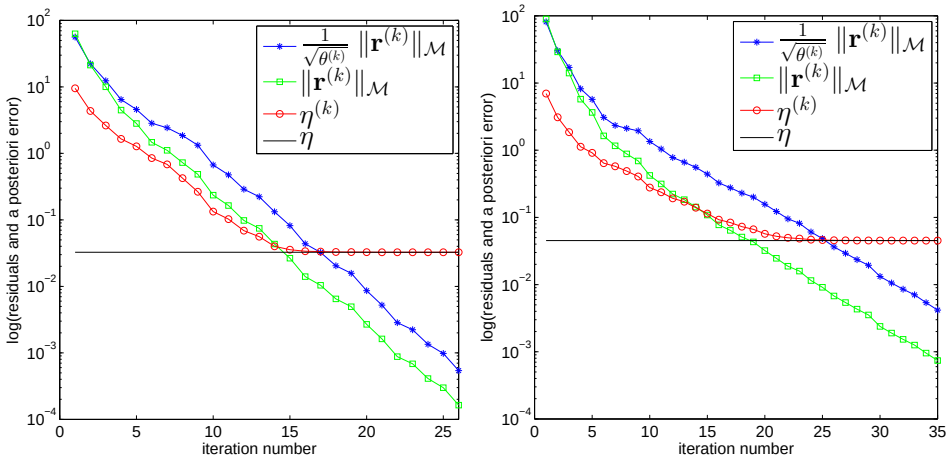

FIG. 3.1. *Convergence plots for preconditioned MINRES with* $h = 1/32$, $p = 3$ | $m = 5$, $\sigma = 0.5$ *(left),* $m = 7$, $\sigma = 0.5$ *(right).*

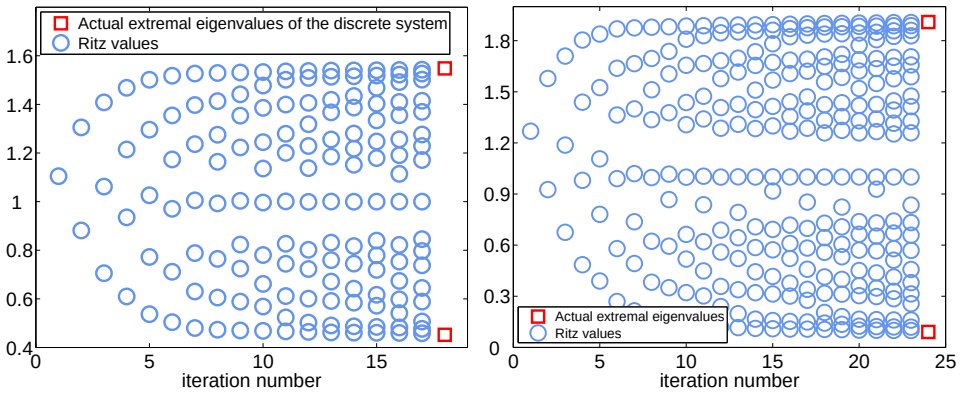

FIG. 3.2. *Ritz values plots with* $m = 5$, $p = 3$ | $h = 1/16$, $\sigma = 0.3$ *(left)* $h = 1/8$, $\sigma = 0.5$ *(right).*

4.1. Convection-diffusion PDE. Find the solution $u(\vec{x}) : D \to \mathbb{R}$ such that

$$\begin{aligned} -\nabla \cdot \epsilon \nabla u(\vec{x}) + \vec{w}(\vec{x}) \cdot \nabla u(\vec{x}) &= f(\vec{x}), \quad \vec{x} \in D \subset \mathbb{R}^d \, (d = 2, 3), \\ u(\vec{x}) &= g_D(\vec{x}), \quad \vec{x} \in \partial D_D, \\ \nabla u(\vec{x}) \cdot \vec{n} &= g_N(\vec{x}), \quad \vec{x} \in \partial D_N = \partial D \backslash \partial D_D. \end{aligned} \tag{4.1}$$

Here D is the spatial domain, $\vec{w}$ denotes the wind velocity and $\epsilon > 0$ is the diffusion coefficient. The source function f and the boundary data g_D, g_N are given functions. The boundary ∂D is the union of Dirichlet ∂D_D and Neumann ∂D_N components and $\vec{n}$ is the outward normal to the boundary.

The weak formulation of (4.1) is to find $u \in H^1_E$ such that

$$\epsilon \int_D \nabla u \cdot \nabla v \, \mathrm{d}\vec{x} + \int_D (w \cdot \nabla u) v \, \mathrm{d}\vec{x} = \int_D f v \, \mathrm{d}\vec{x} + \epsilon \int_{\partial D_N} g_N v \, \mathrm{d}\vec{x}, \quad \forall \, v \in H^1_{E_0}. \tag{4.2}$$

Here $H^1_E := \{u \in H^1(D) \,|\, u = g_D \text{ on } \partial D_D\}$, $H^1_{E_0} := \{u \in H^1(D) \,|\, u = 0 \text{ on } \partial D_D\}$.

Provided that (4.2) is well posed (see Elman et al. [6, p. 242]), the Petrov–Galerkin finite element method with streamline diffusion method [6, p. 251] results

in the nonsingular linear system

$$\mathcal{F}\mathbf{x} = \mathbf{f} \quad \text{or equivalently} \quad \mathcal{M}\mathcal{F}\mathbf{x} = \mathcal{M}\mathbf{f}. \tag{4.3}$$

Here $\mathcal{M}$ is a preconditioner. The linear system can be written as $\mathcal{F} = \epsilon\mathcal{A} + \mathcal{N} + \mathcal{S}$ [6, chapter 6]. Here $\mathcal{A}$ is a symmetric (diffusion matrix) positive definite matrix, N is a skew symmetric matrix (convection matrix) and S is the positive semi-definite stabilization matrix. Since $\mathcal{F}$ is nonsymmetric (if $\vec{w} \neq \vec{0}$) (preconditioned) GMRES is used to solve (4.3).

The 'natural' norm for Sobolev space $H^1(D)$ is the $L_2(D)$ norm of the gradient [18]. The a posteriori error estimators $\eta^{(k)}$, η are computed in this 'natural' norm which is inbuilt in IFISS [6, p. 264-265].

At iteration k, the triangle inequality gives

$$\underbrace{\|u - u_h^{(k)}\|_E^2}_{\text{total error}} \leq \underbrace{\|u - u_h\|_E^2}_{\text{approximation error}} + \underbrace{\|u_h - u_h^{(k)}\|_E^2}_{\text{algebraic error}}, \tag{4.4}$$

where u is the true solution, u_h is the exact numerical approximation, $u_h^{(k)}$ is the finite element approximation corresponding to iterate $\mathbf{x}^{(k)}$ and $\|\cdot\|_E$ denotes the $L_2(D)$ norm of the gradient.
Rewriting (4.4) in terms of the a posteriori error estimators and iteration error $\mathbf{e}^{(k)}$

$$(\eta^{(k)})^2 \underset{\sim}{\leq} \eta^2 + \|\mathbf{e}^{(k)}\|_{\mathcal{A}}^2, \tag{4.5}$$

where $\|u_h - u_h^{(k)}\|_E^2 = \|\mathbf{e}^{(k)}\|_{\mathcal{A}}^2 := {\mathbf{e}^{(k)}}^T \mathcal{A}\, \mathbf{e}^{(k)}$.

4.2. Balanced stopping test for preconditioned GMRES. We have

$$\mathbf{r}^{(k)} = \mathcal{F}\mathbf{e}^{(k)} \Longrightarrow \mathbf{e}^{(k)} = \mathcal{F}^{-1}\mathbf{r}^{(k)} \Longrightarrow \|\mathbf{e}^{(k)}\|_{\mathcal{A}}^2 = {\mathbf{r}^{(k)}}^T \mathcal{F}^{-T}\mathcal{A}\mathcal{F}^{-1}\mathbf{r}^{(k)}.$$

Assuming $\mathcal{F}$ is diagonalizable then in GMRES, $\|\mathbf{r}^{(k)}\|_2 := \sqrt{{\mathbf{r}^{(k)}}^T \mathbf{r}^{(k)}}$ is monotonically decreasing and readily computable. So, we will obtain bounds for $\|\mathbf{e}^{(k)}\|_{\mathcal{A}}$ in terms of $\|\mathbf{r}^{(k)}\|_2$. This involves computing the Rayleigh quotient bounds of $\mathcal{F}^{-T}\mathcal{A}\mathcal{F}^{-1}$ and the identity matrix $\mathcal{I}$—$\theta \leq \frac{{\mathbf{r}^{(k)}}^T \mathcal{F}^{-T}\mathcal{A}\mathcal{F}^{-1}\mathbf{r}^{(k)}}{{\mathbf{r}^{(k)}}^T \mathcal{I}\, \mathbf{r}^{(k)}} \leq \Theta$. Here θ, Θ are respectively the smallest and the largest eigenvalue of $\mathcal{F}^{-T}\mathcal{A}\mathcal{F}^{-1}$. This leads to[13]

$$\|\mathbf{e}^{(k)}\|_{\mathcal{A}} \leq \frac{\Theta}{\sqrt{\theta}}\, \|\mathbf{r}^{(k)}\|_2. \tag{4.6a}$$

$$\|\mathbf{e}^{(k)}\|_{\mathcal{A}} \leq \sqrt{\Theta}\, \|\mathbf{r}^{(k)}\|_2. \tag{4.6b}$$

Using the tighter bound (4.6b), we stop at iteration k^*, which is the smallest value of k such that

$$\sqrt{\Theta}\, \|\mathbf{r}^{(k^*)}\|_2 \leq \eta^{(k^*)} \tag{4.7}$$

Cheap, efficient and reliable[14] a posteriori error estimator $\eta^{(k)}$ is inbuilt in IFISS. As mentioned earlier, $\eta^{(k)}$ can be computed periodically to have a minor impact on the

[13] We have also devised a balanced stopping test in GMRES for linear system arising from every step of Picard and/or Newton step in Navier–Stokes equation but we do not mention results here.

[14] Reliability of the a posteriori estimator in this case not always guaranteed on coarser grids. This might result in an overestimation of the total error and hence premature stopping of our GMRES solver. In such situations the stopping test (4.6a) should be used.

TABLE 5.1
Iteration counts and errors for GMRES solver

ℓ	ILU			AMG			#dof
	k_1	k^*	e_{η^*}	k_1	k^*	e_{η^*}	
3	6	2	1.4e-2	5	2	1.5e-2	81
4	9	3	6.2e-3	5	3	2.6e-3	289
5	16	8	1.9e-3	5	3	8.9e-4	1089
6	37	19	5.0e-4	4	4	7.3e-5	4225

overall algorithmic cost. But we compute it here at every iteration step to show the effectiveness of our balanced stopping test. Computing the Rayleigh quotient bounds is equivalent to solving for the extremal eigenvalues of the generalized eigenvalue problem for $\mathcal{A}$ and $\mathcal{F}^T\mathcal{F}$. The matrices $\mathcal{F}^T\mathcal{F}$ and $\mathcal{A}$ are both symmetric positive definite and thus, the generalized eigenvalue problem can be converted to a symmetric positive definite algebraic eigenvalue problem through a Cholesky factorization of $\mathcal{F}^T\mathcal{F}$. Hence, all the eigenvalues of $\mathcal{F}^{-T}\mathcal{A}\mathcal{F}^{-1}$ are real and greater than zero. The finite element matrices obtained are quite sparse and relatively 'small' (see the rightmost column in Table 4.1), so the eigenvalues in the stopping test can be computed cheaply through MATLAB function `eigs`.[15]

5. Computational results. Again, to provide a proof of concept, we present the results of computational experiments when stopping test (4.7) is applied to preconditioned nonsymmetric linear systems arising from the model problem (4.1). Following [6, p. 240] the PDE problem (4.1) will be defined on a square domain $D = (-1,1) \times (-1,1)$. The source function is $f(x_1, x_2) = 0, \ \forall\, (x_1, x_2) \in D$. The wind velocity $\vec{w} = (2x_2(1 - x_1^2), -2x_1((1 - x_2^2))$ and zero Dirichlet boundary conditions are imposed everywhere on the boundary except at $x_1 = 1$ where $u = 1$. Rectangular $\boldsymbol{Q}_1$ (piecewise bilinear) finite elements are used on a uniform grid with mesh size $h = 2^{1-\ell}$, $\ell = 3, 4, 5, 6$. The diffusion coefficient $\epsilon = 1/64$ is fixed and the optimal inbuilt value of SUPG stabilization parameter [6, p. 253] is used.[16] The inbuilt preconditioners in IFISS–incomplete Cholesky factorization [6, p. 83] and algebraic multigrid (AMG) [6, p. 314] are used to solve (4.3) using preconditioned GMRES. In MATLAB notation incomplete LU (ILU) $[L, U] =$ `ilu`(F). The AMG preconditioner is used with its default inbuilt options in IFISS.

We present results for a hierarchy of finite element grids. For each grid, a reference solution $\mathbf{x}$ is computed using MATLAB backslash (Gaussian elimination) function along with the corresponding reference a posteriori error estimate η. The random initial vector is generated by MATLAB function `rand`. Let $\eta^{(k^*)}$ be the corresponding a posteriori error estimate at the balanced stopping iteration step k^* and let $e_{\eta^*} := |\eta - \eta^{(k^*)}|$. The e_{η^*} columns in Table 5.1 show that $\eta^{(k^*)}$ has converged with some accuracy to the reference a posteriori error estimate η. A comparison of the iteration counts k^* based on (4.7) with iteration counts k_1 (to satisfy a fixed absolute residual $\|\mathbf{r}^{(k)}\|_2$ reduction tolerance of `1e-5`) indicates significant savings in iteration counts especially if one is solving adaptively. Representative results are presented in Figure 5.1. Each subplot shows the evolution of $\eta^{(k)}$, $\|\mathbf{r}^{(k)}\|_2$ and $\sqrt{\Theta}\,\|\mathbf{r}^{(k)}\|_2$ with iteration index k. When $\|\mathbf{e}^{(k)}\|_{\mathcal{A}}$ is insignificant in (4.5), $\{\eta^{(k)}\}$ converges to η. In order to illustrate convergence we have taken nine more iterations after balanced stopping.

6. Conclusion. We have devised a balanced stopping test for symmetric positive definite and nonsymmetric linear systems with PDE origins. The viewpoint taken is that consideration of the PDE origins of such systems is essential to devise a balanced stopping test. It is shown that if a reliable and efficient a posteriori error estimation routine is available then a balanced inbuilt stopping criterion can be realized.

[15] For huge systems an alternative way of estimating the extremal eigenvalues needs to be devised.
[16] The problem can be setup by selecting problem 4 in the driver `cd_testproblem` of IFISS.

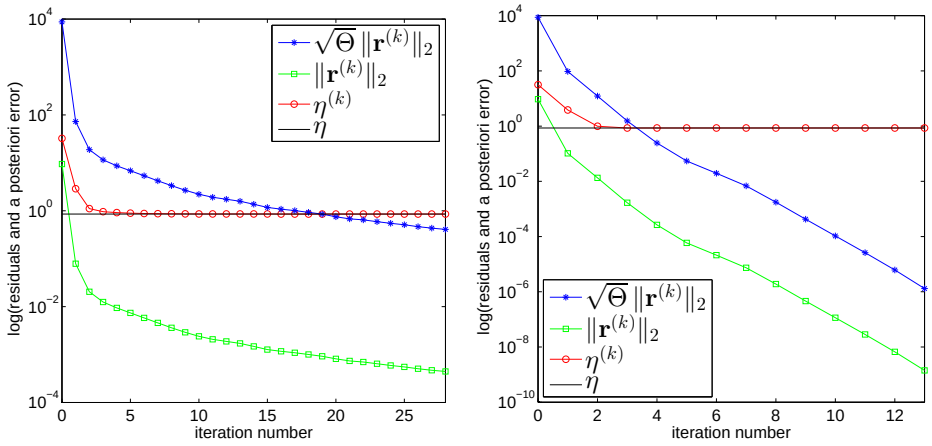

FIG. 5.1. *Convergence plots for ILU (left) and AMG (right) preconditioned GMRES for* $h = 1/32$.

REFERENCES

[1] M. ARIOLI, *A stopping criterion for the conjugate gradient algorithm in a finite element method framework*, Numer. Math., 97 (2004), pp. 1–24.

[2] M. ARIOLI, E. NOULHARD, AND A. RUSSO, *Stopping criteria for iterative methods: applications to pde's*, CALCOLO, Springer Verlag, 38 (2001), pp. 97–112.

[3] ALEX BESPALOV, CATHERINE POWELL, AND DAVID SILVESTER, *Energy norm a posteriori error estimation for parametric operator equations*, SIAM J. Sci. Comput., 36 (2014), pp. A339–A363.

[4] T. BUTLER, P. CONSTANTINE, AND T. WILDEY, *A posteriori error analysis of parametrized linear systems using spectral methods*, SIAM J. Matrix Anal. Appl., 33 (2012), pp. 195–209.

[5] M. K. DEB, I. M. BABUŠKA, AND J. T. ODEN, *Solution of stochastic partial differential equations using Galerkin finite element techniques*, Comput. Methods Appl. Mech. Engrg, 190 (2001), pp. 6359–6372.

[6] HOWARD ELMAN, DAVID SILVESTER, AND ANDY WATHEN, *Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics*, Oxford University Press, Oxford, UK, 2014. Second Edition.

[7] HOWARD C. ELMAN, ALISON RAMAGE, AND DAVID J. SILVESTER, *IFISS: A computational laboratory for investigating incompressible flow problems*, SIAM Review, 56 (2014), pp. 261–273.

[8] GENE GOLUB AND CHARLES VAN LOAN, *Matrix Computations*, The John Hopkins University Press, Baltimore, USA, 2013. Fourth Edition.

[9] A. GREENBAUM, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, PA, 1997.

[10] P. JIRÁNEK, Z. STRAKOŠ, AND M. VOHRALIK, *A posteriori error estimates including algebraic error and stopping criteria for iterative solvers*, SIAM J. Sci. Comput., 32 (2010), pp. 1567–1590.

[11] GABRIEL J. LORD, CATHERINE E. POWELL, AND TONY SHARDLOW, *An Introduction to Computational Stochastic PDEs*, Cambridge University Press, Cambridge, UK, 2014.

[12] C. C. PAIGE AND M. A. SAUNDERS, *Solution of sparse indefinite systems of linear equation*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.

[13] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, SIAM, Philadelphia, PA, 1998.

[14] CATHERINE E. POWELL AND HOWARD C. ELMAN, *Block-diagonal preconditioning for spectral stochastic finite-element systems*, IMA J. Numer. Anal., 29 (2009), pp. 350–375.

[15] Y. SAAD AND M. SCHULTZ, *A generalized minimal residual algorithm for solving non symmetric linear systems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869.

[16] DAVID J. SILVESTER, ALEX BESPALOV, AND CATHERINE E. POWELL, *S-IFISS version 1.02*, July 2015. available online at `http://www.manchester.ac.uk/ifiss/s-ifiss1.0.tar.gz`.

[17] DAVID J. SILVESTER AND VALERIA SIMONCINI, *An optimal iterative solver for symmetric indefinite systems stemming from mixed approximation*, ACM Trans. Math. Softw., 37 (2011).

[18] A. J. WATHEN, *Preconditioning and convergence in the right norm*, Int. J. Comput. Math., 84 (2007), pp. 1199–1209.