An optimal iterative solver for symmetric indefinite linear systems with PDE origins: Balanced black-box stopping tests

Pranjal, Prasad

2018

ISSN 1749-9097

An optimal iterative solver for symmetric indefinite linear systems with PDE origins: Balanced black-box stopping tests

Pranjal

Abstract This work discusses the design of efficient algorithms for solving symmetric indefinite linear systems arising from FEM approximation of PDEs. The distinctive feature of the preconditioned MINRES solver that is used here 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 a robust and optimal black-box stopping criterion: the iteration is terminated as soon as the algebraic error is insignificant compared to the approximation error.

Keywords FEM approximation of PDEs, a posteriori FEM error estimators, symmetric indefinite linear systems, iterative solvers, MINRES, preconditioning.

Mathematics Subject Classification (2000) 65C20 · 65N30 · 65N15

1 Introduction

1.1 Problem

Numerical solution of a partial differential equation (PDE) together with initial and/or boundary conditions essentially involves two types of errors—(PDE) approximation error and algebraic error (which arises from solving the usually huge discrete linear system iteratively). For chosen discretization parameters, the approximation error is fixed . Solving iteratively the corresponding discrete linear(ized) system(s) to a very high accuracy is not desirable. This is because a highly accurate iterative solution may require too many iterations and would simply waste computational resources without any decrease in 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 work attempts to handle these issues by presenting *optimal balanced black-box* stopping tests in iterative solvers, specifically Minimal Residual

Pranjal
University of Wisconsin-Madison, Madison, Wisconsin 53706, USA
E-mail: pranjalprasad21@gmail.com

(MINRES) solver [17], (the popular method of choice) for solving symmetric indefinite linear systems with PDE origins. This is an active research field in general; see [21,22,12,19,20].

1.2 Solution methodology

In order to stop *optimally*, that is, by avoiding premature stopping and unnecessary computations, it is important to use the fundamental relation between the algebraic error and the approximation error. For a given approximation (that is, for a fixed approximation error), at any iteration step the total error (which can be regarded as the approximation error obtained from the solution computed at that iteration step) is essentially the sum of the approximation error and the algebraic error; all the errors are measured in some ‘natural’ norm (this issue is addressed later). By balancing the algebraic error and the total error, a *balanced* stopping test is obtained.

Generally, the algebraic error is unknown since the exact algebraic solution is not usually available. Obtaining tractable upper and lower bounds on the algebraic error in terms of a readily computable and monotonically decreasing quantity (if any) of the chosen iterative solver is the distinctive feature of the devised stopping strategy. Moreover, there are no user-defined constants in the optimal balanced stopping tests presented in this paper. Thus, iterative solvers incorporating such optimal balanced stopping strategies will be *black-box* solvers.

Wathen [25] has observed that finite element (FEM) approximation (see [5]) of a PDE endows the FEM problem with a natural norm for measuring errors, which is determined by the approximation space chosen. Typically, in FEM setting, the PDE approximation error and the algebraic error are measured in this natural norm. Also, note that the approximation error can be measured a priori or/and a posteriori (see [24]). A priori approximation error estimation usually requires the 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 generally readily available. Moreover, a posteriori error estimation can be used for driving the FEM procedure adaptively. Hence, a posteriori approximation error estimation approach is used in this paper.

This paper has 9 sections. In section 2, a review is presented of the work done towards developing an optimal balanced black-box stopping test in MINRES solver for solving symmetric indefinite linear systems (in particular saddle point systems) arising from mixed FEM approximation of PDEs. The main contribution of this paper towards existing research is also summarized therein. The weak form and the mixed FEM set up of the underlying PDE (here Stokes equations) is done in section 3 and the target linear system is formulated in section 4. An overview of MINRES is presented in section 5 and a discussion about block preconditioning for accelerating MINRES convergence in solving the target linear system is presented in section 6. The optimal balanced black-box stopping tests in MINRES are derived in section 7. Computational results that are produced using the IFISS [8] toolbox are discussed in section 8. A summary of the paper is presented in section 9. For the sake of brevity, the term balanced stopping test will usually be used in place of optimal balanced black-box stopping test throughout this paper. Note that this work has appeared as chapter 3 in author's PhD thesis [21].

2 Saddle point linear systems

Large linear systems in saddle point form are ubiquitous. They frequently arise in optimization and in mixed finite element approximation of problems arising in fluid and solid mechanics. In matrix form such systems usually have a 2×2 block form

$$\begin{bmatrix} A & B^T \\ B & -C \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix}, \tag{2.1}$$

where $A \in \mathbb{R}^{n \times n}$ is symmetric positive-definite, $C \in \mathbb{R}^{m \times m}$ is symmetric positive semi-definite, $B \in \mathbb{R}^{m \times n}$, $\mathbf{u}, \mathbf{f} \in \mathbb{R}^n$ and $\mathbf{p}, \mathbf{g} \in \mathbb{R}^m$ with $n \geq m$. The coefficient matrix in (2.1) is always symmetric indefinite and so (preconditioned) MINRES is used for solving (2.1). An introduction about discrete saddle point systems and a detailed discussion on numerical methods for solving them can be found in [4].

An optimal balanced black-box stopping test in preconditioned MINRES for solving (2.1) arising from mixed FEM approximation of PDEs has been devised in [23]. In their analysis the matrix C is taken to be the zero matrix. An extension of their algorithm `EST_MINRES` henceforth called `SADDLE_MINRES` is presented in this paper. The solver `SADDLE_MINRES` has essentially the same ingredients as the `EST_MINRES` solver: it employs a block preconditioner to accelerate MINRES convergence with a rate that is independent of problem parameters and incorporates a balanced stopping strategy to maximize efficiency. The balanced stopping test is obtained by balancing the a posteriori approximation error estimate with the iteration error in the natural norm associated with the underlying PDE. Similar to [23], tractable bounds on the usually unobservable (natural) norm of the iteration error are obtained in terms of the monotonically decreasing preconditioner norm of MINRES iteration residual.

Balanced stopping criterion for symmetric indefinite linear systems arising from mixed FEM approximation of PDEs have also been studied in detail in [2]. Their stopping criterion is based on a priori approximation error bounds and the constants involved in the balanced stopping test are also estimated a priori, which may or may not be straightforward to estimate a priori in general. This is in contrast to the material presented here and in [23] where the approximation error is estimated a posteriori and the constants involved in the balanced stopping test are estimated *on-the-fly*.

2.1 Main contribution

Unlike `EST_MINRES`, the optimal balanced black-box stopping strategy presented here provides not only for solving saddle point systems with a nonzero matrix C but a general framework (by presenting the precise eigenvalue problem to solve for the constants required for balanced stopping test) in (preconditioned) MINRES for solving symmetric indefinite linear systems with PDE origins. Moreover, the constant in the balanced stopping test of [23] has been ‘improved’ in this paper in the sense that one now stops optimally a ‘bit’ earlier than using the balanced stopping test of [23].

3 Deterministic steady-state Stokes equations

Stochastic Galerkin FEM (which results in a huge linear system) and stochastic collocation methods (which result in solving for many smaller linear systems) are the popular choices for solving parametric PDEs [11]. Since the existing storage requirements and computational flops increase with the size and the number of linear systems, an optimal balanced black-box stopping test might save significant computational work of an iterative solver and in any case it would rule out premature stopping. Note that the optimal balanced black-box stopping methodology presented here is applicable for solving both the parametric and the corresponding deterministic PDE.

Stokes equations are archetypal PDEs, which on mixed FEM approximation give rise to symmetric indefinite linear systems. A posteriori error estimators play an important role in devising a balanced stopping test (see section 7), however, ‘tight’ a posteriori approximation error estimators for stochastic Stokes equations have not yet been developed. Thus, it is sufficient to focus on devising a balanced stopping test in MINRES for solving the symmetric indefinite linear system arising from mixed FEM approximation of deterministic Stokes equations.

Stokes equations are used for modelling flows at ‘low speed’. Examples include highly viscous and confined flows such as flow of blood etc.; see [7, p. 119]. Following the notation in [7, p. 119], the steady-state Stokes solution $(\overrightarrow{u}, p)$ is defined on a spatial domain $D \subset \mathbb{R}^d$, $(d = 1, 2, 3)$, where the vector valued velocity function $\overrightarrow{u}(\overrightarrow{x}) : D \to \mathbb{R}^d$ and the scalar valued pressure $p(\overrightarrow{x}) : D \to \mathbb{R}$ satisfy

$$\begin{aligned} -\nabla \cdot \nabla \overrightarrow{u}(\overrightarrow{x}) + \nabla p(\overrightarrow{x}) &= \overrightarrow{0}, & \forall \overrightarrow{x} &\in D, & \text{(3.1a)}\\ \nabla \cdot \overrightarrow{u}(\overrightarrow{x}) &= 0, & \forall \overrightarrow{x} &\in D, & \text{(3.1b)}\\ \overrightarrow{u}(\overrightarrow{x}) &= \overrightarrow{w}(\overrightarrow{x}), & \forall \overrightarrow{x} &\in \partial D_D, & \text{(3.1c)}\\ \nabla \overrightarrow{u}(\overrightarrow{x}) \cdot \overrightarrow{n} - \overrightarrow{n} p(\overrightarrow{x}) &= \overrightarrow{s}(\overrightarrow{x}), & \forall \overrightarrow{x} &\in \partial D_N. & \text{(3.1d)} \end{aligned}$$

Here ∂D_D and ∂D_N are the Dirichlet and Neumann parts respectively of the spatial boundary ∂D. The functions $\overrightarrow{w}$, $\overrightarrow{s}$ are given and $\overrightarrow{n}$ denotes the outward normal to ∂D. The set of real numbers is denoted by $\mathbb{R}$.

3.1 Weak formulation

The weak formulation of (3.1) is to find $\overrightarrow{u} \in \mathbf{H}^1_E(D)$ and $p \in L^2(D)$ such that

$$\begin{aligned} a(\overrightarrow{u}, \overrightarrow{v}) + b(\overrightarrow{v}, p) &= f(\overrightarrow{v}), & \forall \overrightarrow{v} &\in \mathbf{H}^1_{E_0}(D), \\ b(\overrightarrow{u}, q) &= 0, & \forall q &\in L^2(D), \end{aligned} \tag{3.2}$$

where

$$\begin{aligned}
a(\overrightarrow{u}, \overrightarrow{v}) &:= \int_D \nabla\overrightarrow{u} : \nabla\overrightarrow{v} - \int_D p\,(\nabla \cdot \overrightarrow{v}),\\
&\nabla\overrightarrow{u} : \nabla\overrightarrow{v} \text{ denotes componentwise dot product},\\
b(\overrightarrow{u}, q) &:= \int_D q\,(\nabla \cdot \overrightarrow{u}), \qquad f(\overrightarrow{v}) := \int_{\partial D_N} \overrightarrow{s} \cdot \overrightarrow{v},\\
L^2(D) &:= \{p : D \to \mathbb{R} \,|\, \int_D p^2 < \infty\},\\
H^1(D) &:= \{u \in L^2(D) \,|\, \Omega^{\boldsymbol{\alpha}} u \in L^2(D), \forall\, |\boldsymbol{\alpha}| \leq 1\},
\end{aligned}$$

$\Omega^{\boldsymbol{\alpha}}$ is distributional derivative of u, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$ is a multiindex, $|\boldsymbol{\alpha}| := \sum_{i=1}^{d} \alpha_i$,

$$\begin{aligned}
\mathbf{H}^1_E(D) &:= \{\overrightarrow{v} \in H^1(D)^d \,|\, \overrightarrow{v} = \overrightarrow{w} \text{ on } \partial D_D\},\\
\mathbf{H}^1_{E_0}(D) &:= \{\overrightarrow{v} \in H^1(D)^d \,|\, \overrightarrow{v} = \overrightarrow{0} \text{ on } \partial D_D\}.
\end{aligned}$$

Here $H^1(D)^d$ is the d-fold Cartesian product of the $H^1(D)$ space. For definition of distributional derivative, see [16, p. 434].

3.2 Mixed FEM formulation

Choosing finite dimensional subspaces in (3.2), $\mathbf{X}^h_E \subset \mathbf{H}^1_E(D)$, $\mathbf{X}^h_{E_0} \subset \mathbf{H}^1_{E_0}(D)$, $M^h \subset L^2(D)$ leads to a mixed FEM formulation: find $\overrightarrow{u_h} \in \mathbf{X}^h_E$, $p_h \in M^h$ such that

$$\begin{aligned}
a(\overrightarrow{u_h}, \overrightarrow{v_h}) + b(\overrightarrow{v_h}, p_h) &= f(\overrightarrow{v_h}), \qquad \forall\, \overrightarrow{v_h} \in \mathbf{X}^h_{E_0},\\
b(\overrightarrow{u_h}, q_h) &= 0, \qquad \forall\, q_h \in M^h.
\end{aligned} \tag{3.3}$$

Let $\{\overrightarrow{\phi_j}\}_{j=1}^{n_u}$ be a basis for the finite dimensional space $\mathbf{X}^h_{E_0}$. It can be extended (loosely speaking)[1] to form a basis $\{\overrightarrow{\phi_j}\}_{j=1}^{n_u+n_\partial}$ for $\mathbf{X}^h_E$, so that any $\overrightarrow{u_h} \in \mathbf{X}^h_E$ can be written as

$$\overrightarrow{u_h} = \sum_{j=1}^{n_u + n_\partial} u_j \overrightarrow{\phi_j}, \qquad u_j \in \mathbb{R}, \tag{3.4}$$

where the known term $\sum_{j=n_u+1}^{n_u+n_\partial} u_j \overrightarrow{\phi_j}$ interpolates the boundary data on ∂D_D.

Similarly, if $\{\psi_k\}_{k=1}^{n_p}$ be a basis for M^h, then any $p_h \in M^h$ has an expansion

$$p_h = \sum_{k=1}^{n_p} p_k \psi_k, \qquad p_k \in \mathbb{R}. \tag{3.5}$$

[1] The space $\mathbf{X}^1_E$ is not a vector space unless $\overrightarrow{w} = \overrightarrow{0}$.

4 Block matrix form

Plugging the basis expansions from equations (3.4) and (3.5) in (3.3) results in the following block matrix formulation[2]

$$\begin{bmatrix} \mathbf{A} & B^T \\ B & O \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix}. \tag{4.1}$$

The symmetric positive-definite matrix $\mathbf{A}$ (henceforth called *vector-Laplacian matrix*) is a block diagonal matrix with the usual FEM stiffness matrix on its diagonals and the matrix B is called the *divergence matrix*. Also, $\mathbf{u} = [u_1, \ldots, u_{n_u}]^T \in \mathbb{R}^{n_u}$, $\mathbf{p} = [p_1, \ldots, p_{n_p}]^T \in \mathbb{R}^{n_p}$, and the entries of $\mathbf{A},\ B,\ \mathbf{f}$, and $\mathbf{g}$ are [7, p. 130]

$$\begin{aligned}
\mathbf{A} &= [a_{ij}] \in \mathbb{R}^{n_u \times n_u}, & a_{ij} &:= \int_D \nabla \overrightarrow{\phi_i} : \nabla \overrightarrow{\phi_j}, \\
B &= [b_{kj}] \in \mathbb{R}^{n_p \times n_u}, & b_{kj} &:= -\int_D \psi_k \, (\nabla \cdot \overrightarrow{\phi_j}), \\
\mathbf{f} &= [f_i] \in \mathbb{R}^{n_u}, & f_i &:= \int_{\partial D_N} \overrightarrow{s} \cdot \overrightarrow{\phi_i} - \sum_{j = n_u + 1}^{n_u + n_\partial} u_j \int_D \nabla \overrightarrow{\phi_i} : \nabla \overrightarrow{\phi_j}, \\
\mathbf{g} &= [g_k] \in \mathbb{R}^{n_p}, & g_k &:= \sum_{j = n_u + 1}^{n_u + n_\partial} u_j \int_D \psi_k \, (\nabla \cdot \overrightarrow{\phi_j}).
\end{aligned} \tag{4.2}$$

For the Stokes equations (continuous and discrete) to be well-posed, a compatibility condition needs to be satisfied at the inflow and outflow boundaries (if any). Moreover, if the discrete system (4.1)–(4.2) is to be a faithful representation of the continuous problem (3.1), then the mixed FEM velocity and pressure spaces need to be chosen carefully such that they satisfy an *inf-sup* (or correspondingly a (discrete) uniform inf-sup) stability condition; see [7, p. 133 ff.] for more details. Typically, choosing more pressure basis functions than velocity basis functions necessarily results in a singular linear system.

Using the popular (piecewise quadratic) $\boldsymbol{Q}_2$–$\boldsymbol{P}_{-1}$ (piecewise linear, discontinuous across elemental boundaries) finite elements or the (Taylor–Hood) $\boldsymbol{Q}_2$–$\boldsymbol{Q}_1$ (piecewise bilinear pressure) finite elements for velocity and pressure space combination leads to inf-sup stable approximations on a rectangular grid. However, the use of higher order finite elements might not always provide more accurate FEM solutions, especially if the true solution is not very regular. Because of this reason and from the ease of programming and computational efficiency, $\boldsymbol{Q}_1$–$\boldsymbol{P}_0$ (piecewise constant pressure) finite elements or $\boldsymbol{Q}_1$–$\boldsymbol{Q}_1$ finite elements are attractive choices for velocity-pressure FEM basis. But these approximations are not inf-sup stable on a rectangular grid. In order to make these finite element methods stable, a symmetric positive semi-definite *stabilization matrix* C is introduced in place of the zero block of the coefficient matrix in (4.1). A detailed discussion about the stabilization rationale and strategy for the discrete Stokes system can be found in [7, pp. 139–149].

[2] Some discretizations of the Stokes equations can lead to nonsymmetric linear systems. But such discretizations are not considered here.

The symmetric coefficient matrix $K := \begin{bmatrix} \mathbf{A} & B^T \\ B & -C \end{bmatrix}$ of stabilized discrete Stokes system is always indefinite; this follows by applying Sylvester's law of inertia on the matrix K [7, p. 189]. Moreover, it will be assumed that K is nonsingular, that is, it has no zero eigenvalue. Since K is symmetric indefinite, MINRES is the popular and robust iterative method of choice for solving discrete linear systems with coefficient matrix K.

5 An overview of MINRES

Iteratively solving $K\mathbf{x} = \mathbf{b}$ using MINRES [7, chapter 4] involves constructing a sequence of iterates $\mathbf{x}^{(k)}$ $(k = 1, 2, \ldots)$ from the shifted Krylov space

$$\mathbf{x}^{(0)} + \text{span}\,\{\mathbf{r}^{(0)}, K\mathbf{r}^{(0)}, \ldots, K^{k-1}\mathbf{r}^{(0)}\}, \tag{5.1}$$

where $\mathbf{x}^{(0)}$ is the initial solution vector, $\mathbf{r}^{(0)} = \mathbf{b} - K\mathbf{x}^{(0)}$ is the initial residual and the spanning space $\mathrm{K}_k(K, \mathbf{r}^{(0)}) := \text{span}\,\{\mathbf{r}^{(0)}, K\mathbf{r}^{(0)}, \ldots, K^{k-1}\mathbf{r}^{(0)}\}$ is the Krylov subspace of order k generated by the matrix K and the vector $\mathbf{r}^{(0)}$. In the context here $\mathbf{x} := \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix}$, $\mathbf{b} := \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix}$. The residual $\mathbf{r}^{(k)}$ at the kth iterative step is

$$\mathbf{r}^{(k)} = \mathbf{b} - K\mathbf{x}^{(k)} = \mathbf{r}^{(0)} + \text{span}\,\{K\mathbf{r}^{(0)}, K^2\mathbf{r}^{(0)}, \ldots, K^k\mathbf{r}^{(0)}\}. \tag{5.2}$$

The MINRES method chooses the iterate $\mathbf{x}^{(k)}$ from the space (5.1) such that it minimizes the Euclidean norm $\|\mathbf{r}^{(k)}\|_2 := \sqrt{(\mathbf{r}^{(k)})^T\mathbf{r}^{(k)}}$ of the corresponding residual $\mathbf{r}^{(k)}$ over the shifted space in the right-hand-side of (5.2).

A basis of orthonormal vectors $\{\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(k)}\}$ is constructed for the Krylov space (5.1), where $\mathbf{w}^{(1)} := \mathbf{b}/\|\mathbf{b}\|$. This construction process of basis is known as the Lanczos method [14] where the basis vectors are generated iteratively using the recurrence

$$KW_k = W_kT_k + t_{k+1,k}\,\mathbf{w}^{(k+1)}\mathbf{e}_k^T =: W_{k+1}\,\underline{T}_k, \tag{5.3}$$

where $W_k := [\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(k)}]$ and $\mathbf{e}_k$ is the kth vector of the canonical basis. The tridiagonal symmetric matrix T_k contains the orthogonalization coefficients and $\underline{T}_k$ is the tridiagonal matrix T_k with an additional final row $[0, \ldots, 0, t_{k+1,k}]$; for complete details see [10, section 2.5]. The constant $t_{k+1,k}$ is chosen such that $\|\mathbf{w}^{(k+1)}\| = 1$. The Lanczos step (5.3) provides the following characterization of the iterate $\mathbf{x}^{(k)}$ and the residual $\mathbf{r}^{(k)}$

$$\mathbf{x}^{(k)} = \mathbf{x}^{(0)} + W_k\mathbf{y}^{(k)}, \tag{5.4a}$$

$$\mathbf{r}^{(k)} = \mathbf{b} - K\mathbf{x}^{(k)} = W_{k+1}\left(\mathbf{e}_1\|\mathbf{r}^{(0)}\| - \underline{T}_k\mathbf{y}^{(k)}\right). \tag{5.4b}$$

By solving the least squares problem $\min_{\mathbf{y}}(\mathbf{e}_1\|\mathbf{r}^{(0)}\| - \underline{T}_k\mathbf{y})$, the minimizing solution $\mathbf{x}^{(k)}$ is computed. Here $\mathbf{e}_1$ is the first canonical basis vector in $(k + 1)$ dimensions. In order to solve the least squares problem, a QR factorization (see [9, p. 246]) of $\underline{T}_k$ is performed using k Givens rotations.

The eigenvalues of $T_k = W_k^T K W_k$ are known as the *Ritz values* (see [9, p. 551]). These can be computed cheaply and readily in the Lanczos method at each iterative step of the MINRES solver. As the iteration progresses, the extremal Ritz values provide an increasingly better approximation to the corresponding extremal eigenvalues of K or of $M^{-1}K$ if the matrix is preconditioned with matrix M. This point will be discussed further in section 7.6.

From the minimal residual criterion, the following MINRES convergence estimate is obtained.

$$\|\mathbf{r}^{(k)}\|_2 \leq \min_{p_k \in \Pi_k,\, p_k(0)=1} \max_j |p_k(\lambda_j)| \; \|\mathbf{r}^{(0)}\|_2, \tag{5.5}$$

where Π_k denotes the set of real polynomials of degree less than or equal to k and λ_j's are the eigenvalues of K. In case of preconditioned linear system with (symmetric positive-definite) preconditioner M, (5.5) becomes [7, p. 192]

$$\frac{\|\mathbf{r}^{(k)}\|_{M^{-1}}}{\|\mathbf{r}^{(0)}\|_{M^{-1}}} \leq \min_{p_k \in \Pi_k,\, p_k(0)=1} \max_j |p_k(\lambda_j)|, \tag{5.6}$$

where $\|\mathbf{r}^{(k)}\|_{M^{-1}} := \sqrt{(\mathbf{r}^{(k)})^T M^{-1} \mathbf{r}^{(k)}}$ is monotonically decreasing with iteration count k in preconditioned MINRES.

6 Block preconditioning

Typically, matrices arising from FEM approximation are ill-conditioned with respect to discretization parameters. Thus, preconditioning is required to accelerate convergence. It is advocated in [15] that block diagonal preconditioners are intrinsic choices for symmetric linear systems (in saddle point problems), which arise from numerical approximation of PDEs. Proceeding in this flavour, it has been argued in [7, p. 194 ff.] that the symmetric matrix $\begin{bmatrix} \mathbf{A} & O \\ O & B\mathbf{A}^{-1}B^T + C \end{bmatrix}$ is the 'desired' but an impractical preconditioner (since $B\mathbf{A}^{-1}B^T + C$ is a dense matrix and hence computing its inverse and also of vector-Laplacian matrix is not cheap) for solving the preconditioned linear system

$$M^{-1}K \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} = M^{-1} \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix}, \tag{6.1}$$

where $M := \begin{bmatrix} \mathbf{P} & O \\ O & S \end{bmatrix}$ is a preconditioner. A practical choice of $\mathbf{P}$ is a block diagonal matrix with each block a preconditioner for the scalar Laplacian matrix (which is on the diagonal of $\mathbf{A}$). It would be ideal to have $\mathbf{P}$ to be spectrally equivalent to $\mathbf{A}$, that is, there exist positive constants δ_1 and Δ_1 that are independent of discretization parameters such that

$$\delta_1 \leq \frac{\mathbf{u}^T\mathbf{A}\mathbf{u}}{\mathbf{u}^T\mathbf{P}\mathbf{u}} \leq \Delta_1, \qquad \forall\, \mathbf{u} \in \mathbb{R}^{n_u}. \tag{6.2}$$

Indeed this is the case when a Laplacian multigrid preconditioner is used; see [7, lemma 4.2, p. 197]. For the block S of the preconditioner, a good choice is the pressure mass matrix $Q = [q_{kl}],\ \ q_{kl} := \int_D \psi_k \psi_l,\ \forall k,\, l = 1, \ldots, n_p$ [7, p. 172]. The matrix Q is spectrally equivalent to the matrix $B\mathbf{A}^{-1}B^T + C$, that is, there exist positive constants γ and $\mathit{\Gamma}$ that are independent of discretization parameters such that the following holds [7, p. 193–194].

$$\gamma^2 \le \frac{\mathbf{q}^T(B\mathbf{A}^{-1}B^T + C)\mathbf{q}}{\mathbf{q}^T Q\mathbf{q}} \le \mathit{\Gamma}^2 \le d, \qquad \forall\, \mathbf{q} \in \mathbb{R}^{n_p} \text{ and } \mathbf{q} \ne \mathbf{1}, \tag{6.3}$$

where d is the dimension of the domain D. In fact the particular choice of $S = \mathtt{diag}(Q)$ for continuous ($\boldsymbol{P}_1$ or $\boldsymbol{Q}_1$)[3] makes S spectrally equivalent to Q, that is, there exist positive constants δ_2 and $\mathit{\Delta}_2$ that are independent of discretization parameters [7, pp. 198–199] such that

$$\delta_2^2 \le \frac{\mathbf{q}^T Q\mathbf{q}}{\mathbf{q}^T S\mathbf{q}} \le \mathit{\Delta}_2^2, \qquad \forall\, \mathbf{q} \in \mathbb{R}^{n_p}. \tag{6.4}$$

Note that the constant γ in (6.3) is the uniform inf-sup constant when $C = 0$ and $\delta^2 = 2\gamma^2$, where δ is the uniform inf-sup constant for the case when $C \ne 0$.

Having formulated a mixed FEM matrix formulation of (3.1) and discussed briefly about MINRES preconditioners to be used for solving the corresponding discrete linear system, the balanced stopping strategy is presented in the next section.

7 A balanced stopping test

According to [25], a natural norm for a function in the space of square integrable functions is its L^2 norm while the L^2 norm of the gradient of the function is a natural choice if the function is in $H^1_{E_0}$. Thus, a natural choice of norm ($\|\cdot\|_{\mathcal{E}}$) for any $(\overrightarrow{u}, p) \in \mathbf{H}^1_{E_0}(D) \times L^2(D)$ is[4]

$$\|(\overrightarrow{u}, p)\|_{\mathcal{E}} := \|\nabla \overrightarrow{u}\|_2 + \|p\|_2, \tag{7.1}$$

where the L^2 norm $\|\cdot\|_2$ is defined as $\|p\|_2 := (\int_D p^2)^{1/2}$. In terms of vectors, $\|\cdot\|_{\mathcal{E}}$ translates into the norm $\|\cdot\|_E$

$$\|\mathbf{e}\|_E := \sqrt{\mathbf{e}^T E\mathbf{e}} = \sqrt{\mathbf{e}_1^T \mathbf{A}\mathbf{e}_1 + \mathbf{e}_2^T Q\mathbf{e}_2}, \qquad \forall\, \mathbf{e} = [\mathbf{e}_1^T,\, \mathbf{e}_2^T]^T \in \mathbb{R}^{n_u + n_p}, \tag{7.2}$$

where $\mathbf{e}_1 \in \mathbb{R}^{n_u}, \mathbf{e}_2 \in \mathbb{R}^{n_p}$, and $E := \begin{bmatrix} \mathbf{A} & O \\ O & Q \end{bmatrix}$. Since the vector-Laplacian matrix $\mathbf{A}$ and the pressure mass matrix Q are both symmetric positive-definite, the matrix E is also symmetric positive-definite and hence $\|\cdot\|_E$ is indeed a norm on $\mathbb{R}^{n_u + n_p}$.

[3] For $\boldsymbol{P}_0$ pressure approximation, Q is in fact diagonal.

[4] Note that $\nabla \overrightarrow{u}$ is to be interpreted componentwise.

7.1 Error equation

For a given approximation, by the triangle inequality at iteration k

$$\underbrace{\|(\overrightarrow{u} - \overrightarrow{u_h^{(k)}}, p - p_h^{(k)})\|_{\mathcal{E}}}_{\text{total error}} \leq \underbrace{\|(\overrightarrow{u} - \overrightarrow{u_h}, p - p_h)\|_{\mathcal{E}}}_{\text{approximation error}} + \underbrace{\|(\overrightarrow{u_h} - \overrightarrow{u_h^{(k)}}, p_h - p_h^{(k)})\|_{\mathcal{E}}}_{\text{algebraic error}}, \tag{7.3}$$

where $(\overrightarrow{u}, p)$ is the true solution, $(\overrightarrow{u_h}, p_h)$ is the true mixed FEM solution, and $(\overrightarrow{u_h^{(k)}}, p_h^{(k)})$ is the FEM solution formed from the kth iterate of the chosen iterative solver. From (7.1), it follows by the definition of $\| \cdot \|_{\mathcal{E}}$ that

$$\|(\overrightarrow{u_h} - \overrightarrow{u_h^{(k)}}, p_h - p_h^{(k)})\|_{\mathcal{E}} = \|\nabla(\overrightarrow{u_h} - \overrightarrow{u_h^{(k)}})\|_2 + \|p_h - p_h^{(k)}\|_2. \tag{7.4}$$

Note that

$$\begin{aligned} \|\nabla(\overrightarrow{u_h} - \overrightarrow{u_h^{(k)}})\|_2 &= \sqrt{(\mathbf{e}_1^{(k)})^T \mathbf{A} \mathbf{e}_1^{(k)}}, \quad \mathbf{e}_1^{(k)} = [u_{1_h} - u_{1_h}^{(k)}, \ldots, u_{n_{u_h}} - u_{n_{u_h}}^{(k)}]^T, \\ \|p_h - p_h^{(k)}\|_2 &= \sqrt{(\mathbf{e}_2^{(k)})^T Q \mathbf{e}_2^{(k)}}, \quad \mathbf{e}_2^{(k)} = [p_{1_h} - p_{1_h}^{(k)}, \ldots, p_{n_{p_h}} - p_{n_{p_h}}^{(k)}]^T, \end{aligned} \tag{7.5}$$

where $\overrightarrow{u_h} - \overrightarrow{u_h^{(k)}} = \sum_{i=1}^{n_u}(u_{i_h} - u_{i_h}^{(k)})\overrightarrow{\phi_i}$, $p_h - p_h^{(k)} = \sum_{j=1}^{n_p}(p_{j_h} - p_{j_h}^{(k)})\psi_j$. Also, for any two nonnegative real numbers a and b [7, p. 213]

$$\sqrt{a+b} \leq \sqrt{a} + \sqrt{b} \leq \sqrt{2}\sqrt{a+b}. \tag{7.6}$$

Putting $a = \|\nabla(\overrightarrow{u_h} - \overrightarrow{u_h^{(k)}})\|_2^2$, $b = \|p_h - p_h^{(k)}\|_2^2$ in (7.6) and using (7.5), (7.2) gives

$$\|\mathbf{e}^{(k)}\|_E \leq \|\nabla(\overrightarrow{u_h} - \overrightarrow{u_h^{(k)}})\|_2 + \|p_h - p_h^{(k)}\|_2 \leq \sqrt{2}\|\mathbf{e}^{(k)}\|_E. \tag{7.7}$$

For enclosed flow problems, a slight variant of the L^2 norm known as the 'quotient space norm' $\| \cdot \|_{0,D}$ is used for measuring pressure.

Here $\|q_h\|_{0,D} = \|q_h - \frac{1}{|D|}\int_D q_h\|_2$, $|D| = \int_D$ for any $q_h \in M^h$ [7, p. 128]. Note that

$$\begin{aligned} \|q_h - \frac{1}{|D|}\int_D q_h\|_2^2 &= \int_D \left(q_h - \frac{1}{|D|}\int_D q_h\right)^2 \\ &= \int_D q_h\, q_h + \int_D \left(\frac{1}{|D|}\int_D q_h\right)^2 - \int_D 2q_h \left(\frac{1}{|D|}\int_D q_h\right) \\ &= \|q_h\|_2^2 + \frac{1}{|D|}\left(\int_D q_h\right)^2 - 2\frac{1}{|D|}\left(\int_D q_h\right)^2 \\ &= \|q_h\|_2^2 - \frac{1}{|D|}\left(\int_D q_h\right)^2 \leq \|q_h\|_2^2, \end{aligned} \tag{7.8}$$

since $\frac{1}{|D|}\left(\int_D q_h\right)^2 \geq 0$. So, $\|q_h\|_{0,D} \leq \|q_h\|_2,\ \forall\, q_h \in M^h$. Thus, the algebraic error $\|(\overrightarrow{u_h} - \overrightarrow{u_h^{(k)}}, p_h - p_h^{(k)})\|_{\mathcal{E}} = \|\nabla(\overrightarrow{u_h} - \overrightarrow{u_h^{(k)}})\|_2 + \|p_h - p_h^{(k)}\|_{0,D}$ (in quotient space norm) can be bounded from above by the usual L^2 norm of the algebraic error, that is

$$\|\nabla(\overrightarrow{u_h} - \overrightarrow{u_h^{(k)}})\|_2 + \|p_h - p_h^{(k)}\|_{0,D} \;\leq\; \|\nabla(\overrightarrow{u_h} - \overrightarrow{u_h^{(k)}})\|_2 + \|p_h - p_h^{(k)}\|_2. \quad (7.9)$$

Using (7.9) one can obtain the same bound (7.7) for the enclosed flow algebraic error at kth iterative step in terms of $\|\mathbf{e}^{(k)}\|_E$ norm of the kth iteration error.

A handle on the approximation error and the total error (approximation error at the kth iteration) is obtained with a posteriori error estimators η and $\eta^{(k)}$ respectively. The a posteriori error estimator $\eta^{(k)}$ is equivalent to the total error in the sense that

$$c_1\,\eta^{(k)} \;\leq\; \|\nabla(\overrightarrow{u} - \overrightarrow{u_h^{(k)}})\|_2 + \|p - p_h^{(k)}\|_2 \;\leq\; C_1\,\eta^{(k)}, \quad \text{with } \frac{C_1}{c_1} \sim O(1), \quad (7.10)$$

If the a posteriori error estimators η and $\eta^{(k)}$ are assumed to be 'close' estimates of the approximation error and total error (at kth iteration step) respectively, then the error equation (7.3) can be rewritten as

$$\eta^{(k)} \;\simeq\; \eta + \|\mathbf{e}^{(k)}\|_E, \qquad k = 0, 1, 2, \ldots. \quad (7.11)$$

The relation $\simeq$ is a result of (7.10) and (7.7). In fact it follows from (7.11) that when the norm $\|\mathbf{e}^{(k)}\|_E$ of the iteration error $\mathbf{e}^{(k)}$ is 'small', then $\{\eta^{(k)}\}$ converges to η. Thus, one would stop optimally when $\|\mathbf{e}^{(k)}\|_E$ and the a posteriori error estimate $\eta^{(k)}$ of the total error are balanced, that is, stop at the first iteration k^* such that

$$\|\mathbf{e}^{(k^*)}\|_E \;\leq\; \eta^{(k^*)}. \quad (7.12)$$

Remark 7.1 Notice from (7.11) and (7.12) that at the optimal stopping iteration k^*, $\{\eta^{(k)}\}$ would converge with some accuracy to η. Thus, the iterative strategy here can be looked upon as constructing a sequence $\{\eta^{(k)}\}$ converging to η.

In the subsequent subsections, a brief discussion on the a posteriori error estimation for the Stokes equations is done and tractable bounds on difficult to compute $\|\mathbf{e}^{(k)}\|_E$ are derived.

7.2 Tractable bounds on algebraic error

In preconditioned MINRES with symmetric positive-definite preconditioner M, the norm $\|\mathbf{r}^{(k)}\|_{M^{-1}} := \sqrt{\mathbf{r}^{(k)^T} M^{-1}\mathbf{r}^{(k)}}$ is monotonically decreasing with iteration count k and hence a suitable surrogate norm for computations in place of $\|\mathbf{e}^{(k)}\|_E$. Here $\mathbf{r}^{(k)} := K\mathbf{e}^{(k)}$ is the residual at iteration k. Thus, one obtains an expression for the algebraic error at kth iterative step in terms of the iteration residual $\mathbf{r}^{(k)}$, that is

$$\|\mathbf{e}^{(k)}\|_E^2 \;=\; (\mathbf{e}^{(k)})^T E\mathbf{e}^{(k)} \;=\; (\mathbf{r}^{(k)})^T K^{-T} E K^{-1}\mathbf{r}^{(k)}. \quad (7.13)$$

It follows from (7.13) that bounding $\|\mathbf{e}^{(k)}\|_E$ by $\|\mathbf{r}^{(k)}\|_{M^{-1}}$ requires computing constants c_2 and C_2 such that

$$c_2 \le \frac{\left(\mathbf{r}^{(k)}\right)^T K^{-T} E K^{-1} \mathbf{r}^{(k)}}{(\mathbf{r}^{(k)})^T M^{-1} \mathbf{r}^{(k)}} \le C_2, \tag{7.14}$$

This leads to computing extremal Rayleigh quotient [9, p. 453] bounds of $K^{-T}EK^{-1}$ and M^{-1}, that is, find $\lambda_{\text{min}}, \lambda_{\text{max}} \in \mathbb{R}$ such that

$$\lambda_{\text{min}} \le \frac{\mathbf{v}^T K^{-T} E K^{-1} \mathbf{v}}{\mathbf{v}^T M^{-1} \mathbf{v}} \le \lambda_{\text{max}}, \qquad \forall\, \mathbf{v} \in \mathbb{R}^{n_u + n_p}. \tag{7.15}$$

Equation (7.15) implies that one needs to compute generalized extremal eigenvalues for $K^{-T}EK^{-1}$ and M^{-1}, that is, find the extremal eigenvalues λ such that

$$K^{-T}EK^{-1}\mathbf{y} = \lambda M^{-1}\mathbf{y}, \qquad \mathbf{y} \in \mathbb{R}^{n_u + n_p} \text{ is an eigenvector.} \tag{7.16}$$

Note that the matrices K, E are symmetric so the matrix $K^{-T}EK^{-1}$ is also symmetric. Also, since M is symmetric positive-definite, its inverse M^{-1} is also symmetric positive-definite. So, the generalized eigenvalue problem (7.16) can be converted (theoretically) into a symmetric algebraic eigenvalue problem through a Cholesky factorization of M^{-1}. Hence all λ's in (7.16) are real. Let $\mathbf{z} = K^{-1}\mathbf{y}$, then (7.16) becomes

$$K^{-T}E\mathbf{z} = \lambda M^{-1}K\mathbf{z}, \qquad \mathbf{z} \in \mathbb{R}^{n_u + n_p}. \tag{7.17}$$

It is clear from the discussions in section that an ideal but an impractical choice for the preconditioner M is the matrix E. A more practical choice is where the matrices $\mathbf{P}$ and S satisfy (6.2) and (6.4) respectively and hence M is spectrally equivalent to E. Thus, for 'good' choices of $\mathbf{P}$ and S, M will 'behave like' E after a 'few' iterations. Substituting M for E in (7.17), and using that K is symmetric gives

$$\left(M^{-1}K\right)^{-1}\mathbf{z} = \lambda M^{-1}K\mathbf{z}, \qquad \mathbf{z} \in \mathbb{R}^{n_u + n_p}. \tag{7.18}$$

Let $W := M^{-1}K$, then (7.18) can be rearranged as the following eigenvalue problem

$$W^2\mathbf{z} = \mu\mathbf{z}, \qquad \mathbf{z} \in \mathbb{R}^{n_u + n_p}, \tag{7.19}$$

where $\mu = 1/\lambda$. Note that since $W = M^{-1}K$ is symmetric and nonsingular, all its eigenvalues are real and nonzero. So, the eigenvalues μ's of W^2 (which are the squares of eigenvalues of W) are all real and greater than zero. So, any λ cannot be zero; in fact all λ's are greater than zero.

In light of (7.17), (7.18), and (7.19) the eigenvalue problem (7.16) is transformed into finding the largest (μ_{max}) and smallest (μ_{min}) eigenvalues of W^2 such that

$$W^2\mathbf{z} = \mu\mathbf{z}, \qquad \mathbf{z} \in \mathbb{R}^{n_u + n_p} \text{ is an eigenvector.} \tag{7.20}$$

Since the eigenvalues of W^2 are just the square of the eigenvalues of W, it is sufficient to compute the eigenvalues of W. In fact, one obtains

$$\mu_{\text{max}} = \max\{|\theta^{+}_{\text{max}}|^2, |\theta^{-}_{\text{min}}|^2\}, \tag{7.21a}$$

$$\mu_{\text{min}} = \min\{|\theta^{+}_{\text{min}}|^2, |\theta^{-}_{\text{max}}|^2\}, \tag{7.21b}$$

where θ's are eigenvalues of W such that

θ^+_{max} – maximum positive eigenvalue, θ^+_{min} – minimum positive eigenvalue,
θ^-_{max} – maximum negative eigenvalue, θ^-_{min} – minimum negative eigenvalue.

7.3 Stopping criteria

Using $\lambda_{\text{min}} = \frac{1}{\mu_{\text{max}}}, \lambda_{\text{max}} = \frac{1}{\mu_{\text{min}}}$; (7.14), (7.15), and (7.21) can be combined into

$$\frac{1}{\max\{|\theta^+_{\text{max}}|^2, |\theta^-_{\text{min}}|^2\}} \leq \frac{(\mathbf{r}^{(k)})^T K^{-T} E K^{-1} \mathbf{r}^{(k)}}{(\mathbf{r}^{(k)})^T M^{-1} \mathbf{r}^{(k)}} \leq \frac{1}{\min\{|\theta^+_{\text{min}}|^2, |\theta^-_{\text{max}}|^2\}}. \tag{7.22}$$

It follows from (7.22) that

$$\frac{1}{\sqrt{\max\{|\theta^+_{\text{max}}|^2, |\theta^-_{\text{min}}|^2\}}} \leq \frac{\|\mathbf{e}^{(0)}\|_E}{\|\mathbf{r}^{(0)}\|_{M^{-1}}}; \frac{\|\mathbf{e}^{(k)}\|_E}{\|\mathbf{r}^{(k)}\|_{M^{-1}}} \leq \frac{1}{\sqrt{\min\{|\theta^+_{\text{min}}|^2, |\theta^-_{\text{max}}|^2\}}}. \tag{7.23}$$

Equation (7.23) leads to the following upper bounds on $\|\mathbf{e}^{(k)}\|_E$, that is

$$\|\mathbf{e}^{(k)}\|_E \leq \frac{1}{\sqrt{\min\{|\theta^-_{\text{max}}|^2, |\theta^+_{\text{min}}|^2\}}} \|\mathbf{r}^{(k)}\|_{M^{-1}},$$

$$\begin{aligned} \frac{\|\mathbf{e}^{(k)}\|_E}{\|\mathbf{e}^{(0)}\|_E} &\leq \sqrt{\frac{\max\{|\theta^+_{\text{max}}|^2, |\theta^-_{\text{min}}|^2\}}{\min\{|\theta^-_{\text{max}}|^2, |\theta^+_{\text{min}}|^2\}}} \frac{\|\mathbf{r}^{(k)}\|_{M^{-1}}}{\|\mathbf{r}^{(0)}\|_{M^{-1}}} \qquad (7.24) \\ \Longleftrightarrow \|\mathbf{e}^{(k)}\|_E &\leq \frac{\sqrt{\max\{|\theta^+_{\text{max}}|^2, |\theta^-_{\text{min}}|^2\}}}{\min\{|\theta^-_{\text{max}}|^2, |\theta^+_{\text{min}}|^2\}} \|\mathbf{r}^{(k)}\|_{M^{-1}}. \end{aligned}$$

Thus, from (7.12) it follows that an optimal stopping point is the first iteration k^* at which one of the following tests is satisfied

$$\frac{\sqrt{\max\{|\theta^+_{\text{max}}|^2, |\theta^-_{\text{min}}|^2\}}}{\min\{|\theta^-_{\text{max}}|^2, |\theta^+_{\text{min}}|^2\}} \|\mathbf{r}^{(k^*)}\|_{M^{-1}} \leq \eta^{(k^*)}. \tag{7.25}$$

$$\frac{1}{\sqrt{\min\{|\theta^-_{\text{max}}|^2, |\theta^+_{\text{min}}|^2\}}} \|\mathbf{r}^{(k^*)}\|_{M^{-1}} \leq \eta^{(k^*)}. \tag{7.26}$$

Henceforth, the stopping test (7.25) will be called the stronger stopping test while the stopping test (7.26) will be called the weaker stopping test.

7.4 A posteriori error estimation

The a posteriori error estimation technique used in the software IFISS for Stokes equations is due to [1] and it essentially involves solving a local Poisson problem for each velocity component; see [7, section 3.4.2]. The a poseriori error estimator based on this strategy provides 'acceptable' close estimates of the true total (approximation) errors. In fact for $\boldsymbol{Q}_1$–$\boldsymbol{P}_0$ rectangular finite elements, this a posteriori error estimator is both a global upper bound, (that is, it is reliable) and a local elementwise bound (that is, it is efficient) on the actual error; see [13] for full details. A comparison of η [7, table 3.4, p. 169] and 'actual' approximation error $\|\nabla(\overrightarrow{u} - \overrightarrow{u_h})\|_2 + \|p - p_h\|_{0,2}$ [7, table 3.3, p. 166] are tabulated in Table 7.1. The results presented therein are for the Stokes test problem 1 [7, p. 126] in section 8.1 with $\boldsymbol{Q}_1$–$\boldsymbol{P}_0$ rectangular finite elements on a uniform grid and mesh step size h.

Table 7.1 Actual approximation errors, a posteriori errors, and effectivity indices for $\boldsymbol{Q}_1$–$\boldsymbol{P}_0$ rectangular finite elements on uniform grids for Stokes test problem 1.

h	η	$\|\nabla(\overrightarrow{u} - \overrightarrow{u_h})\|_2 + \|p - p_h\|_{0,2}$	β_{eff}
1/4	9.501	18.729	0.51
1/8	5.307	8.853	0.59
1/16	2.761	4.290	0.64
1/32	1.399	2.116	0.66

The entries for corresponding effectivity index $\beta_{\text{eff}} = \dfrac{\eta}{\|\nabla(\overrightarrow{u} - \overrightarrow{u_h})\|_2 + \|p - p_h\|_{0,2}}$ in Table 7.1 show that a posteriori approximation error estimator employed here is an 'acceptable close' estimate of the true error.

7.5 Computational logistics

The M^{-1} norm of the iteration residual, that is, $\|\mathbf{r}^{(k)}\|_{M^{-1}}$ is readily available in preconditioned MINRES. Also, it is advisable in general to compute $\eta^{(k)}$ periodically (say every 4–5 iterations) to minimize the overall algorithmic cost. The eigenvalues involved in the stopping test (7.25) and (7.26) can be estimated cheaply on-the-fly, the strategy for which is described in the next subsection.

7.6 Cheap estimation of eigenvalues in stopping test

Note that the extremal eigenvalues θ^{+}_{max}, θ^{-}_{min} of the preconditioned matrix can cheaply be estimated by the corresponding extremal Ritz values θ^{k+}_{max}, θ^{k-}_{min} (the maximum positive Ritz value and the minimum negative Ritz value respectively) of the Lanczos matrix T_k (see section 5) in preconditioned MINRES.[5] As the iteration progresses,

[5] This relationship was also exploited in [22].

the extremal Ritz values provide an increasingly better approximation to the corresponding extremal eigenvalues of $M^{-1}K$. *This holds true for even small iteration index k*, and has been discussed extensively in [18, chapter 13].

But for the interior most eigenvalues θ^{+}_{min} and θ^{-}_{max}, the Ritz values usually provide a poor estimation. So, the interior most eigenvalues are estimated here by computing the corresponding interior most eigenvalues θ^{k+}_{min}, θ^{k-}_{max} of the following generalized eigenvalue problem

$$\underline{T}_k^T \underline{T}_k \mathbf{y} \;=\; \theta_{\text{har}} T_k \mathbf{y}, \qquad \mathbf{y} \text{ is an eigenvector.} \tag{7.27}$$

where $\underline{T}_k$ is the $\mathbb{R}^{(k+1)\times k}$ Lanczos matrix. The eigenvalues θ_{har} in (7.27) are known as *harmonic Ritz values*; see [3, section 3.2, p. 41–43]. Here θ^{k+}_{min} and θ^{k-}_{max} denote the minimum positive harmonic Ritz value and the maximum negative harmonic Ritz value respectively. Unlike the Ritz values, which approximate first the extremal eigenvalues of the preconditioned matrix, the harmonic Ritz values approximate first the interior most eigenvalues of the preconditioned matrix.[6] This is better than using Ritz values to estimate the actual interior most eigenvalues since the interior most Ritz values might take a long time to provide a good approximation (if at all) to the interior most eigenvalues.

Further insight into the eigenvalues of the preconditioned matrix is obtained from the following result in [7, theorem 4.7, p. 201].

Theorem 7.1 *The eigenvalues of $M^{-1}K$ satisfy*

$$\begin{aligned} -\varDelta_2^2\left(\varGamma^2 + \varUpsilon\right) \;\le\; \theta^{-}_{\text{min}} \;\le\; \theta^{-}_{\text{max}} \;\le\; \frac{1}{2}\left(\delta_1 \,-\, \sqrt{\delta_1^2 \,+\, 4\delta_1\gamma^2\delta_2^2}\right), \\ \delta_1 \;\le\; \theta^{+}_{\text{min}} \;\le\; \theta^{+}_{\text{max}} \;\le\; \varDelta_1 \,+\, \varGamma^2\varDelta_2^2, \end{aligned} \tag{7.28}$$

where $\delta_1, \varDelta_1, \delta_2, \varDelta_2, \gamma,$ and $\varGamma$ are the same as in (6.2), (6.4), and (6.3) respectively. The constant $\varUpsilon$ satisfies

$$\frac{\mathbf{q}^T C\mathbf{q}}{\mathbf{q}^T Q\mathbf{q}} \le \varUpsilon, \qquad \forall\, \mathbf{q} \in \mathbb{R}^{n_p}. \tag{7.29}$$

Proceeding along the lines of [23] note that for $M = E$, $\delta_1 = \varDelta_1 = 1$. Also, for $\boldsymbol{P}_0$ pressure approximation $\delta_2 = \varDelta_2 = 1$. In any case if preconditioner blocks $\mathbf{P}$ and S 'closely' approximate $\mathbf{A}$ and Q respectively, then

$$\delta_1 \simeq 1, \varDelta_1 \simeq 1, \delta_2 \simeq 1, \text{ and } \varDelta_2 \simeq 1. \tag{7.30}$$

Also, the asymptotic simplification $(1+x)^{\frac{1}{2}} = 1 + \frac{1}{2}x$ gives

$$\frac{1}{2}\left(\delta_1 \,-\, \sqrt{\delta_1^2 \,+\, 4\delta_1\gamma^2\delta_2^2}\right) \simeq \frac{1}{2}\left(1 \,-\, \sqrt{1 \,+\, 4\gamma^2}\right) \simeq -\gamma^2. \tag{7.31}$$

Combining (7.28), (7.30), and (7.31) leads to

$$\theta^{-}_{\text{max}} \;\simeq\; -\gamma^2 \;\le\; 1 \;\simeq\; \theta^{+}_{\text{min}}. \tag{7.32}$$

[6] This approach was also adopted in [23].

The validity of the equivalence $\theta^{-}_{\text{max}} \simeq -\gamma^2$ for $C = 0$ case can be further confirmed from the discussions in [7, pp. 196–197]. If $\gamma^2 \leq 1$, which is usually the case[7] then from (7.32) it follows $\dfrac{1}{\sqrt{\min\{|\theta^{-}_{\text{max}}|^2, |\theta^{+}_{\text{min}}|^2\}}} = \dfrac{1}{\sqrt{|\theta^{-}_{\text{max}}|^2}} \simeq \dfrac{1}{\sqrt{\gamma^4}} = \dfrac{1}{\gamma^2}$. In light of this analysis, the weaker stopping test (7.26) can be transformed into

$$\frac{1}{\gamma^2}\, \|\mathbf{r}^{(k^*)}\|_{M^{-1}} \leq \eta^{(k^*)} \quad \Longleftrightarrow \quad \|\mathbf{r}^{(k^*)}\|_{M^{-1}} \leq \gamma^2\, \eta^{(k^*)}. \tag{7.33}$$

An equivalence similar to (7.32) holds for maximum positive eigenvalue and minimum negative eigenvalue

$$\theta^{-}_{\text{min}} \simeq -(\varGamma^2 + \varUpsilon) \leq (1 + \varGamma^2) \simeq \theta^{+}_{\text{max}}. \tag{7.34}$$

Since $0 \leq \varUpsilon \leq 1$ [7, p. 200], from (7.34) $\sqrt{\max\{|\theta^{+}_{\text{max}}|^2, |\theta^{-}_{\text{min}}|^2\}} = \sqrt{|\theta^{+}_{\text{max}}|^2} \simeq (1 + \varGamma^2) \leq 1 + d$. Combining this with $\min\{|\theta^{-}_{\text{max}}|^2, |\theta^{+}_{\text{min}}|^2\} = |\theta^{-}_{\text{max}}|^2 \simeq \gamma^4$, the stronger stopping test (7.25) becomes

$$\frac{1+d}{\gamma^4}\, \|\mathbf{r}^{(k^*)}\|_{M^{-1}} \leq \eta^{(k^*)} \quad \Longleftrightarrow \quad \|\mathbf{r}^{(k^*)}\|_{M^{-1}} \leq \frac{\gamma^4}{1+d}\, \eta^{(k^*)}. \tag{7.35}$$

In presence of 'tight' a posteriori error estimators and 'good' preconditioner blocks, the stopping test (7.33) or (7.35) can be used and they hold for both $C = 0$ and $C \neq 0$.

Table 7.2 Comparison of literature and improved stopping tests for $\boldsymbol{Q}_2$–$\boldsymbol{P}_1$ finite elements on rectangular uniform grids for Stokes test problem 1.

h	k^*_{lit}	e^*_{lit}	k^*_{imp}	e^*_{imp}
1/8	10	6.0e-2	8	5.4e-2
1/16	17	1.0e-5	15	5.6e-3
1/32	21	3.1e-4	19	1.7e-3
1/64	24	1.1e-4	24	1.1e-4

In fact for the case $C = 0$, using (7.33) one stops optimally a 'bit' earlier than using the stopping test $\|\mathbf{r}^{(k^*)}\|_{M^{-1}} \leq \dfrac{\gamma^2}{\sqrt{2}}\, \eta^{(k^*)}$ of [23]. This happens because $\dfrac{\gamma^2}{\sqrt{2}} < \gamma^2$ and hence an 'improvement' of constants (over those in the existing literature) involved in the stopping test for the case $C = 0$ has been obtained here. Note that this improvement is only a theoretical result. Since $\sqrt{2} \approx 1.41$, in practice a gain of only 'very few' (if any) iterations is obtained by using γ^2 over $\dfrac{\gamma^2}{\sqrt{2}}$ in balanced stopping (7.33); see Table 7.2. The stopping iteration k^*_{lit} and k^*_{imp} corresponding to

[7] In fact $\gamma^2 \leq \varGamma^2 \leq d$, where d (equal to 2 or 3 here) denotes the dimensionality of the domain D. But for $C = 0$ with Dirichlet boundary conditions in $\mathbb{R}^2, \gamma^2 \leq 1$; see [7, theorem 3.22, p. 174].

the stopping test in [23] and (7.33) respectively are tabulated in Table 7.2. Also, at each grid level tabulated are $e^*_{\text{lit}} := |\eta - \eta^{(k^*_{\text{lit}})}|$ and $e^*_{\text{imp}} := |\eta - \eta^{(k^*_{\text{imp}})}|$. These denote the corresponding absolute differences in a posteriori error estimates from the actual a posteriori estimate η obtained using the 'true' solution (MATLAB backslash solution). It follows from Table 7.2 that savings of only a few iterations is obtained on using the stopping test (7.33) over that in [23]. Also, at the stopping iteration for both these stopping tests, the sequence $\{\eta^{(k)}\}$ has converged with some accuracy to the true η; see columns for e^*_{lit} and e^*_{imp}. These numbers have been obtained by running `itsolve_stokes` with default options in IFISS toolbox of MATLAB after setting up the Stokes test problem 1 that is described in section 8.1. The constants involved in the stopping test can be modified suitably in the function `param_est` in IFISS.

7.7 Choice of stopping test

A drawback of using the stopping test (7.25) or (7.26) is that they might lead to premature stopping because one or more of the computed extremal Ritz values or the interior most harmonic Ritz values would have not yet converged to their corresponding (discrete system) actual eigenvalue respectively. Although this convergence is usually quite fast, it is generally difficult to determine beforehand the iteration count at which they will converge. Hence, it is proposed here to store the required Ritz and harmonic Ritz values of previous 4–5 consecutive iterations and apply the stopping test (7.25) or (7.26) only when the absolute successive differences of these values for each of the required quantities is below a prescribed tolerance of 10^{-2} (say).[8]

Substituting (7.33) for (7.26) and (7.35) for (7.25) overcomes this drawback. This is because the constants in the stopping tests now depend on d, which is trivially known and the discrete inf-sup constant γ, which in many practical applications is known beforehand and depends only on the topology of the spatial domain; see [6]. However, the stopping tests (7.33) and (7.35) were derived using many equivalences ($\simeq$) which may not be tight in general. Hence, in presence of a preconditioner M which is spectrally equivalent to E, it will be better to employ the weaker or the stronger stopping test based on interior most harmonic Ritz values and extremal Ritz values.

The resulting algorithm known as `SADDLE_MINRES` in the software IFISS is given in the form of pseudo-code in Figure 7.1. The external functions `matvecK`, `precM` compute the action of the matrices K and M^{-1} on a vector respectively while the function `Stokes_error_est` computes the a posteriori error estimate. Also, $\mathbf{b} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix}$ denotes the right-hand-side vector in Figure 7.1. This algorithm can easily be modified for the weaker stopping test. A practical implementation of this algorithm should incorporate periodic computations of the a posteriori error estimate. Also, it should involve storage of previous 4–5 values from consecutive iterations for each of the Ritz and the harmonic Ritz values involved in the balanced stopping test.

[8] For the weaker stopping test this procedure has to be done for only the harmonic Ritz values.

```
Algorithm: SADDLE_MINRES
given vectors $\mathbf{b}$, $\mathbf{x}^{(0)}$ and functions matvecK, precM, param_intest, param_extest
Stokes_error_est
..........................................................................................
set $\mathbf{r}^{(0)} = \mathbf{b} - \mathtt{matvecK}\,(\mathbf{x}^{(0)})$, $\hat{\mathbf{r}}^{(0)} = \mathtt{precM}\,(\mathbf{r}^{(0)})$, $\rho_0 = \sqrt{(\mathbf{r}^{(0)})^T \hat{\mathbf{r}}^{(0)}}$
initialize basis vectors: $\mathbf{w} = \hat{\mathbf{r}}^{(0)}/\rho_0$, $\mathbf{p}^{(-1)} = \mathbf{0}$, $\mathbf{p}^{(0)} = \mathbf{r}^{(0)}/\rho_0$
initialize auxiliary vectors: $\mathbf{d}^{(-1)} = \mathbf{0}$, $\mathbf{d}^{(0)} = \mathbf{0}$
initialize projected right-hand side: $f = \rho_0$
..........................................................................................
for $k = 1, 2, \ldots$ until convergence do
    generate new basis and auxiliary vectors: $\mathbf{p}^{(k)} = \mathtt{matvecK}\,(\mathbf{w}),\ \mathbf{d}^{(k)} = \mathbf{w}$
        if $k>1$, $t_{k-1,k} = t_{k,k-1},\ \mathbf{p}^{(k)} = \mathbf{p}^{(k)} - \mathbf{p}^{(k-1)} t_{k-1,k}$
    $t_{k,k} = \mathbf{w}^T \mathbf{p}^{(k)},\ \mathbf{p}^{(k)} = \mathbf{p}^{(k)} - \mathbf{p}^{(k-1)} t_{k,k}$
    compute preconditioned basis vector: $\mathbf{w} = \mathtt{precM}\,(\mathbf{p}^{(k)})$
    $t_{k+1,k} = \sqrt{\mathbf{w}^T \mathbf{p}^{(k)}},\ \mathbf{p}^{(k)} = \mathbf{p}^{(k)}/t_{k+1,k}$ , $\mathbf{w} = \mathbf{w}/t_{k+1,k}$
    compute parameters for stopping test:
    coefext = param_extest $(T_k)$
    coefint = param_intest $(T_k, t_{k+1,k})$
    coef = coefext/(coefint)$^2$
    apply previous rotations:
        if $k>2$, $\rho_{1:2} = S_{k-2} t_{k-2:k-1,k},\ \rho_{2:3} = S_{k-1}[\rho_2; t_{k,k}]$
        elseif $k=2$, $\rho_{2:3} = S_{k-1} t_{1:2,2}$
        elseif $k = 1$, $\rho_3 = t_{1,1}$
    compute new rotations:
      $\hat{\delta} = \sqrt{\rho_3^2 + t_{k+1,k}^2},\ c = |\rho_3|/\hat{\delta},\ s = \mathrm{sign}(\rho_3) t_{k+1,k}/\hat{\delta}$
    apply new rotations: $\rho_3 = c\rho_3 + s t_{k+1,k},\ \hat{f} = -sf,\ f = cf,\ S_k = [c\ s; -s\ c]$
    update auxiliary vector: $\mathbf{d}^{(k)} = (\mathbf{d}^{(k)} - \mathbf{d}^{(k-1)}\rho_1 - \mathbf{d}^{(k-2)}\rho_2)/\rho_3$
    update solution: $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \mathbf{d}^{(k)} \hat{f}$
    compute discretization error estimate : $\eta^{(k)}$ = Stokes_error_est $(\mathbf{x}^{(k)})$
    stopping test: if coef$\cdot|\hat{f}| \le \eta^{(k)}$, convergence
    update residual norm: $f = \hat{f}$
enddo
```

```
function coefext = param_extest $(T_k)$
    compute the smallest negative eigenvalue $\theta_{\min}^{k-}$ and the largest positive eigenvalue
    $\theta_{\max}^{k+}$ of $T_k$
        if $|\theta_{\min}^{k-}|^2 \le |\theta_{\max}^{k+}|^2$ set coefext = $\theta_{\max}^{k+}$
        else set coefext = $|\theta_{\min}^{k-}|$
endfunction
```

```
function coefint = param_intest $(T_k, t_{k+1,k})$
    compute the smallest positive eigenvalue $\theta_{\min}^{k+}$ and the largest negative eigenvalue
    $\theta_{\max}^{k-}$ of generalized eigenvalue problem $\underline{T}_k^T \underline{T}_k$ and $T_k$
        if $|\theta_{\max}^{k-}|^2 \le |\theta_{\min}^{k+}|^2$ set coefint = $|\theta_{\max}^{k-}|$
        else set coefint = $|\theta_{\min}^{k+}|$
endfunction
```

Fig. 7.1 The SADDLE_MINRES algorithm expressed in pseudo-code.

8 Computational results

To provide a proof-of-concept, some computational results are presented in this section for two test problems in IFISS. The stronger stopping test (7.25) is employed for both the test problems in order to exhibit the nuances associated with using a stopping test based on both interior most and exterior most eigenvalues of the preconditioned matrix. It has also been observed from computations for the test problems considered here that the relevant extremal Ritz values and the interior most harmonic Ritz values have converged with some accuracy before optimal stopping has been reached. So, one does not need to store previous 4-5 values from consecutive iterations for these quantities. Also, instead of computing the a posteriori error estimator periodically, it is computed here at each iteration to illustrate the balanced stopping methodology.

There are four preconditioners built in IFISS for the discrete Stokes problem. They are: diagonal (DIAG) preconditioner—the diagonal matrix which is formed from the diagonal elements of $\mathbf{A}$ and the diagonal entries of Q—the block ideal preconditioner E, block geometric multigrid (GMG), and block algebraic multigrid (AMG) [7, chapter 4] preconditioners. Results are presented here for block ideal and block AMG preconditioners for both the test problems. Note that the block AMG preconditioner is employed with its specified default settings in IFISS.

Piecewise bilinear ($\boldsymbol{Q}_1$) finite elements are used for FEM velocity space and $\boldsymbol{P}_0$ finite elements are employed for FEM pressure space on rectangular grids. The uniform mesh step size h is used for the test problem 1 while $2^l \times (2^l \times 3)$ grids are employed for the test problem 2. The built-in *stabilization* parameter value in IFISS is used for setting up the matrix block C in K for both the test problems.

8.1 Test Problem 1

The Stokes PDE (3.1) is defined on a square domain $D = (-1, 1) \times (-1, 1)$ with Dirichlet boundary condition specified everywhere on the boundary. This (enclosed flow) problem [7, p. 126] can be generated by choosing example `4` when running the driver `stokes_testproblem` in IFISS. On a given grid, the 'true' algebraic solution $\mathbf{x}$ is obtained from (block ideal/block AMG) preconditioned MINRES with a tight relative residual $\frac{\|\mathbf{r}^{(k)}\|_{M^{-1}}}{\|\mathbf{r}^{(0)}\|_{M^{-1}}}$ reduction tolerance of `1e-14`. From $\mathbf{x}$, the 'exact' a posteriori error estimate η is computed. The starting vector $\mathbf{x}^{(0)}$ is generated using the MATLAB function `rand`. Also, let $\eta^{(k^*)}$ denote the a posteriori error estimate at the optimal stopping iteration k^* and $e^*_\eta := |\eta - \eta^{(k^*)}|$. These values are tabulated

Table 8.1 MINRES iteration counts and errors along with extremal Ritz values and interior most harmonic Ritz values for block ideal preconditioning on uniform grids for Stokes test problem 1.

h	k_{tol1}	k_{tol2}	k^*	e^*_η	$\theta^{k^*_-}_{\text{min}}$	$\theta^{k^*_-}_{\text{max}}$	$\theta^{k^*_+}_{\text{min}}$	$\theta^{k^*_+}_{\text{max}}$	#dof
1/16	33	48	15	1.3e-2	-1.2994	-0.2911	1.000	1.6152	3202
1/32	33	48	24	5.3e-4	-1.3173	-0.1949	1.000	1.6170	12546
1/64	33	50	27	1.2e-4	-1.3184	-0.1841	1.000	1.6175	49666
1/128	33	50	30	2.8e-5	-1.3192	-0.1781	1.000	1.6177	197634

in Tables 8.1 and 8.2 for block ideal and block AMG preconditioner respectively on various grids.

Table 8.2 MINRES iteration counts and errors along with extremal Ritz values and interior most harmonic Ritz values for block AMG preconditioning on uniform grids for Stokes test problem 1.

h	k_{tol1}	k_{tol2}	k^*	e^*_η	$\theta_{\text{min}}^{k^*_-}$	$\theta_{\text{max}}^{k^*_-}$	$\theta_{\text{min}}^{k^*_+}$	$\theta_{\text{max}}^{k^*_+}$	$\#$dof
1/16	37	54	18	1.1e-5	-1.3010	-0.2815	0.8676	1.5989	3202
1/32	39	55	27	5.2e-6	-1.3069	-0.2017	0.8375	1.6093	12546
1/64	41	58	31	8.4e-7	-1.3088	-0.1816	0.8159	1.6119	49666
1/128	41	58	35	1.7e-6	-1.3095	-0.1756	0.8070	1.6134	197634

The e^*_η columns show that $\{\eta^{(k)}\}$ has converged with a good accuracy to the true a posteriori error estimate η at the balanced stopping iteration, see Remark 7.1. The effectiveness of the balanced stopping test can be gauged by comparing the iteration counts k^* needed to satisfy the balanced stopping test with the iteration counts $k_{\text{tol1}}, k_{\text{tol2}}$ needed to satisfy a fixed relative residual $\dfrac{\|\mathbf{r}^{(k)}\|_{M^{-1}}}{\|\mathbf{r}^{(0)}\|_{M^{-1}}}$ reduction tolerance of `1e-6` (which is the default tolerance in MATLAB solvers) and `1e-9` respectively. In the absence of a balanced stopping test, these are realistic choices for algebraic error tolerance. It is unlikely that the user will know the stopping point k^* a priori and is likely to provide a tighter tolerance than actually required. This would result in needless computations. A quick glance at the columns for optimal iteration counts k^* and those of k_{tol2} shows that a significant number of iterations is wasted (without decreasing the approximation error) by not using the balanced stopping test. Typically, employing the balanced stopping test (7.25) or (7.26) would result in significant savings in computational work of the solver, especially if one were to solve the underlying PDE adaptively using FEM. These computational savings are further significant in light of huge size of some of these linear systems; see the last ($\#$dof) column in Tables 8.1 and 8.2.

The stopping tests (7.35) and (7.33) suggest that the relevant eigenvalues involved in the stopping tests (7.25) and (7.26) are independent of the discretization parameters. Indeed this is the case, which can be seen from the column entries at balanced stopping iteration for extremal Ritz values $\theta_{\text{min}}^{k^*_-}, \theta_{\text{max}}^{k^*_+}$ and interior most harmonic Ritz values $\theta_{\text{max}}^{k^*_-}, \theta_{\text{min}}^{k^*_+}$ estimates of the corresponding eigenvalues of the discrete system. Also, a comparison of the corresponding eigenvalue (Ritz and harmonic Ritz) estimates for block AMG and block ideal preconditioners shows that block AMG approximates the block ideal preconditioner quite closely.

Further insight into the intricacies associated with applying the stronger stopping test (7.25) is provided by Figures 8.1, 8.2, and 8.3 for both block ideal and block AMG preconditioning on a uniform grid with $h = \dfrac{1}{128}$. On both plots of Figure 8.1, note that at the optimal stopping iteration k^* (the iteration where the red curve for $\eta^{(k)}$ is first above the blue curve for $\|\mathbf{r}^{(k)}\|_{M^{-1}}$) $\{\eta^{(k)}\}$ has converged with some accuracy to the exact a posteriori error estimate η. The convergence is

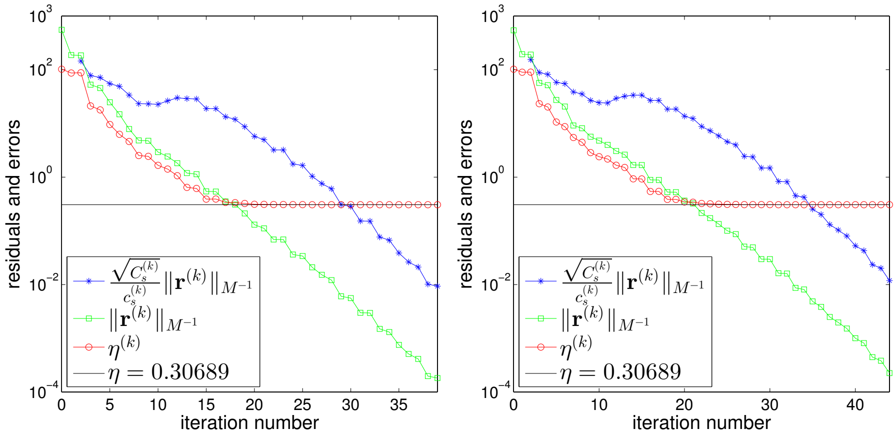

Fig. 8.1 Errors vs iteration number for block ideal (left) and block AMG (right) preconditioned MINRES on a uniform grid $h = 1/128$ for Stokes test problem 1.

further illustrated by continuing for 9 more iterations after balanced stopping where the red curve for $\eta^{(k)}$ always 'stays' on the black line for η. Note that on these plots $C_s^{(k)} := \max\{|\theta_{\text{max}}^{k+}|^2, |\theta_{\text{min}}^{k-}|^2\}$ and $c_s^{(k)} := \min\{|\theta_{\text{max}}^{k-}|^2, |\theta_{\text{min}}^{k+}|^2\}$.

The convergence of extremal Ritz values and interior most harmonic Ritz values at the balanced stopping iteration to the corresponding eigenvalues of the discrete problem can be seen from Figures 8.2 and 8.3 respectively. The actual extremal and interior most eigenvalues of the preconditioned (block ideal and block AMG) matrix on these plots are estimated as the corresponding Ritz and harmonic Ritz values respectively. Preconditioned MINRES is run 'long enough' here to ensure that these estimates have 'converged' (this was ascertained by looking at the values of these estimates). Note that the data plotted in Figures 8.2 and 8.3 corresponds to the entries in the last row for block ideal and block AMG preconditioner respectively in Tables 8.1 and 8.2. Also, the plots continue for 9 more iterations after balanced stopping to illustrate that the converged extremal Ritz values and interior most harmonic Ritz values stay convergent to the corresponding discrete system eigenvalues.[9]

The Ritz value plots in Figure 8.2 further suggest that there are no *ghost (spurious copies)* of extremal Ritz values. The same is suggested for interior most harmonic Ritz values in Figure 8.3. In contrast there are ghost Ritz values for interior most Ritz values r_{max}^{k-} (the maximum negative Ritz value at the kth step) and r_{min}^{k+} (the minimum positive Ritz value at the kth step); see Figure 8.2. This is also the case for the extremal harmonic Ritz values h_{max}^{k+} (the maximum positive harmonic Ritz value at the kth step) and h_{min}^{k-} (the minimum negative harmonic Ritz value at the kth step); see Figure 8.3. Thus, θ_{max}^{+} and θ_{min}^{-} should be estimated by the corresponding extremal Ritz values while θ_{min}^{+} and θ_{max}^{-} should be estimated by the corresponding interior most harmonic Ritz values. This is consistent with the discussion in section 7.6.

[9] Lanczos method can lose orthogonalization after convergence of Ritz vectors and hence these estimates might not remain converged. But implementing the Lanczos procedure in MINRES with reorthogonalization solves this issue.

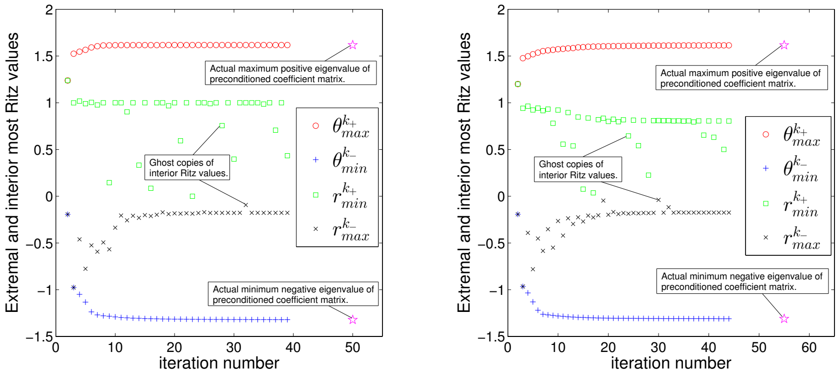

Fig. 8.2 Computed Ritz values for block ideal (left) and block AMG (right) MINRES on a uniform grid $h = 1/128$ for Stokes test problem 1.

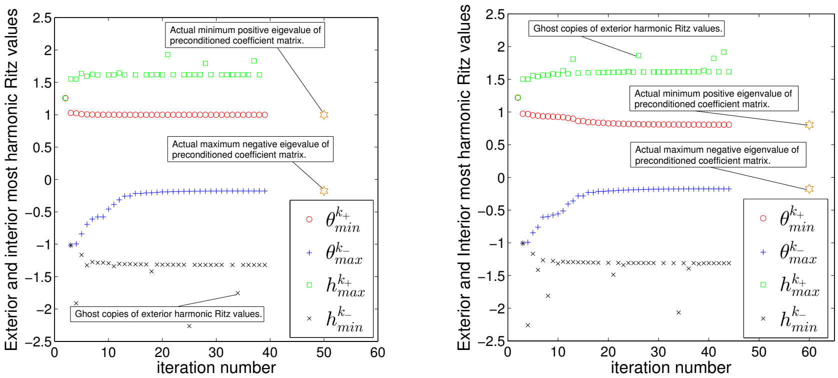

Fig. 8.3 Computed harmonic Ritz values for block ideal (left) and block AMG (right) MINRES on a uniform grid $h = 1/128$ for Stokes test problem 1.

The discrete inf-sup constant can also be estimated on-the-fly as suggested in the work of [23]. It follows from Theorem 7.1 that if the bounds in (7.28) are 'tight' then

$$\gamma^2 = \frac{(\theta^-_{\text{max}})^2 - \theta^-_{\text{max}}\theta^+_{\text{min}}}{\theta^+_{\text{min}}}. \tag{8.1}$$

In light of the Lanczos estimates for the extremal and interior most eigenvalues of the preconditioned matrix, (8.1) can be rewritten as

$$(\gamma^{(k)})^2 = \frac{(\theta^{k_-}_{\text{max}})^2 - \theta^{k_-}_{\text{max}}\theta^{k_+}_{\text{min}}}{\theta^{k_+}_{\text{min}}}. \tag{8.2}$$

Thus, the balanced stopping strategy also provides a cheap estimate for γ on-the-fly; see Figure 8.4. The 'true' γ in Figure 8.4 is computed by running (block ideal and

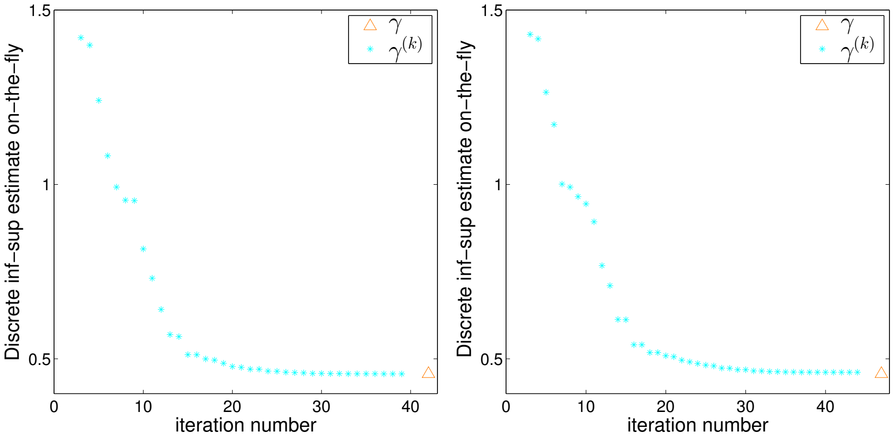

Fig. 8.4 Computed discrete inf-sup constant for block ideal (left) and block AMG (right) preconditioned MINRES on a uniform grid $h = 1/128$ for Stokes test problem 1.

block AMG) preconditioned MINRES 'long enough' to ensure convergence (from inspection of the estimate values).

A closer examination of Figure 8.1 shows that $\{\eta^{(k)}\}$ has converged to η much before balanced stopping iteration on each plot. In fact if one were to apply the weaker stopping test (7.26) then this is the iteration at which one would stop optimally. However, there is always the pitfall of premature stopping due to nonconvergence of the interior most harmonic Ritz values. A way to overcome this issue of premature stopping has been discussed in section 7.7. To reiterate, the results are presented here for the stronger stopping test (7.25) only to illustrate the nuances associated with optimal stopping for symmetric indefinite systems. In general, the weaker stopping test should be used in practice.

Note that the plots in Figure 8.1 merit a further investigation in devising an optimal balanced black-box stopping test that is independent of the extremal and interior most eigenvalues of the preconditioned matrix. If $\sqrt{\min\{|\theta^-_{\text{max}}|^2, |\theta^+_{\text{min}}|^2\}} \geq 1$, then

$$\frac{1}{\sqrt{\min\{|\theta^-_{\text{max}}|^2, |\theta^+_{\text{min}}|^2\}}}\ \|\mathbf{r}^{(k)}\|_{M^{-1}} \leq \|\mathbf{r}^{(k)}\|_{M^{-1}}. \tag{8.3}$$

The weaker stopping test (7.26) in light of (8.3) can be transformed into the following. Stop at the first iteration k^* such that

$$\|\mathbf{r}^{(k^*)}\|_{M^{-1}} \leq \eta^{(k^*)}. \tag{8.4}$$

However, application of the stopping test (8.4) depends on the implicit assumption that $\sqrt{\min\{|\theta^-_{\text{max}}|^2, |\theta^+_{\text{min}}|^2\}} \geq 1$, which is not always true; see the corresponding entries for θ^+_{min} and θ^-_{max} in Tables 8.1 and 8.2.

8.2 Test Problem 2

The Stokes PDE (3.1) is defined on a L-shaped ('flow over a backward-facing step') domain $D = (-1,5) \times (-1,1) \setminus (-1,0] \times (-1,0]$. Poiseuille flow profile is imposed on the inflow boundary ($x_1 = -1, 0 \leq x_2 \leq 1$) for $\overrightarrow{x} = (x_1, x_2) \in D$, and zero velocity condition is imposed on the walls. Neumann boundary conditions are defined everywhere on the outflow boundary ($x_1 = 5, -1 < x_2 < 1$) [7, p. 124]. This problem can be generated IFISS by choosing example `2` when running the driver `stokes_testproblem`.

Table 8.3 MINRES iteration counts and errors along with extremal Ritz values and interior most harmonic Ritz values for block ideal preconditioning on $2^l \times (2^l \times 3)$ grids for Stokes test problem 2.

l	k_{tol1}	k_{tol2}	k^*	e^*_η	$\theta^{k^*_-}_{\text{min}}$	$\theta^{k^*_-}_{\text{max}}$	$\theta^{k^*_+}_{\text{min}}$	$\theta^{k^*_+}_{\text{max}}$	#dof
4	53	73	51	6.2e-6	-1.3632	-0.0242	1.000	1.7909	2242
5	55	73	54	3.5e-6	-1.3638	-0.0242	1.000	1.8109	8706
6	53	76	58	1.4e-6	-1.3669	-0.0242	1.000	1.8184	34306
7	53	77	61	3.1e-7	-1.3671	-0.0241	1.000	1.8214	136194

Table 8.4 MINRES iteration counts and errors along with extremal Ritz values and interior most harmonic Ritz values for block AMG preconditioning on $2^l \times (2^l \times 3)$ grids for Stokes test problem 2.

l	k_{tol1}	k_{tol2}	k^*	e^*_η	$\theta^{k^*_-}_{\text{min}}$	$\theta^{k^*_-}_{\text{max}}$	$\theta^{k^*_+}_{\text{min}}$	$\theta^{k^*_+}_{\text{max}}$	#dof
4	59	80	55	1.1e-5	-1.3540	-0.0241	0.8025	1.7191	2242
5	63	84	61	5.2e-6	-1.3571	-0.0241	0.7865	1.7294	8706
6	63	86	65	8.4e-7	-1.3576	-0.0240	0.7606	1.7334	34306
7	63	88	69	1.7e-6	-1.3577	-0.0241	0.7290	1.7361	136194

Results are tabulated for block ideal and block AMG preconditioned MINRES in Tables 8.3 and 8.4 for this test problem on various $2^l \times (2^l \times 3)$ grids. The quantities in these tables are defined exactly in the same way as for the test problem 1.[10]

The insights from the results here is essentially similar to those for test problem 1. As compared to the test problem 1, the slower convergence is due to a singularity in the problem near the 'step' which is reflected in the largest negative eigenvalue estimate (see the $\theta^{k^*_-}_{\text{max}}$ column in Tables 8.3 and 8.4) of the preconditioned matrix, which is more closer to zero than $\theta^{k^*_-}_{\text{max}}$ of test problem 1 (where there was no singularity in the problem).

From Figure 8.5 note that it is possible that the curve of $\eta^{(k)}$ may fall below the line of true a posteriori estimate η. However, as the iteration proceeds, ultimately the sequence $\{\eta^{(k)}\}$ converges with some accuracy to η.

[10] However, here the 'true' algebraic solution $\mathbf{x}$ is obtained from preconditioned MINRES with a tight relative residual reduction tolerance of `1e-12` instead of `1e-14` since (preconditioned) MINRES gives a warning that latter 'input tolerance may not be achievable by MINRES' on some grids.

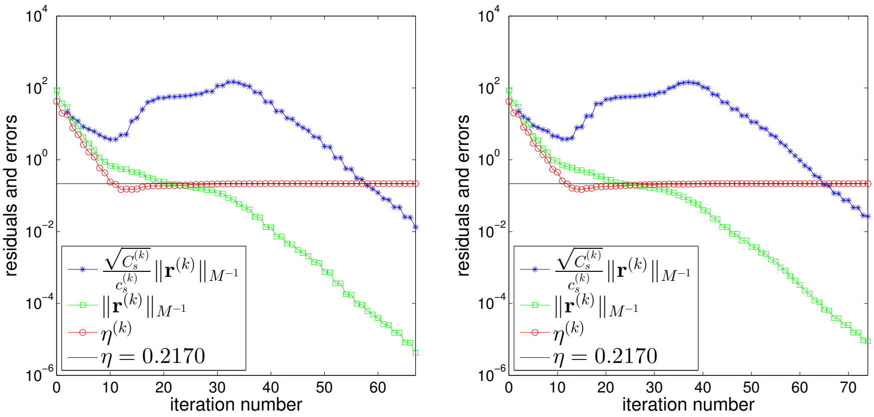

Fig. 8.5 Errors vs iteration number for block ideal (left) and block AMG (right) preconditioned MINRES on a 128×384 grid for Stokes test problem 2.

The results from both the test problems illustrate that employing an optimal balanced black-box stopping strategy not only avoids unnecessary computations but also rules out premature stopping of the preconditioned MINRES solver.

9 Conclusions

An optimal balanced black-box stopping test is devised (in a generic sense) in this paper in MINRES with preconditioning for solving (saddle point) symmetric indefinite linear systems arising from FEM discretization of an underlying PDE (Stokes equations in particular). The constants in the balanced stopping test are estimated cheaply on-the-fly. This is achieved by exploiting the relationship between Ritz, harmonic Ritz values (obtained from the Lanczos process in preconditioned MINRES) and the relevant eigenvalues of the preconditioned matrix involved in the balanced stopping test. Typically, employing such a balanced stopping strategy would avoid premature stopping and generally lead to significant computational savings. The stopping strategy presented here has extended the work done in this direction by [23]. In particular, the methodology presented here for deriving the constants involved in the balanced stopping test is quite generic as compared from that in [23]. Also, the constant involved the stopping test of [23] has been ‘improved’ in the sense that one can now stop optimally a few iterations earlier than using their stopping test.

The optimal balanced black-box stopping methodology presented in this thesis can be generalized for any iterative solver of a linear(ized) symmetric indefinite system arising from numerical approximation of a PDE. The only prerequisites for this purpose are the existence of a cheap and tight a posteriori error estimator for the approximation error along with cheap and tractable bounds on the algebraic error.

Acknowledgements The author would like to thank his PhD supervisor Professor (Dr.) David J. Silvester for introducing him to this field of research.

References

1. Ainsworth, M., Oden, J.: A posteriori error estimators for Stokes and Oseen equations. SIAM J. Numer. Anal. **34**(1), 228–245 (1997). https://doi.org/10.1137/S0036142994264092
2. Arioli, M., Loghin, D.: Stopping criteria for mixed finite element problems. Elec. Trans. on Numer. Anal. **29**, 178–192 (2008). https://etna.ricam.oeaw.ac.at/vol.29.2007-2008/pp178-192.dir/pp178-192.pdf
3. Bai, Z., Demmel, J., Dongarra, J., Ruhe, A., van der Vorst, H.: Templates for the solution of Algebraic Eigenvalue Problems: A Practical guide. SIAM, USA (2000)
4. Benzi, M., Golub, G.H., Liesen, J.: Numerical solution of saddle point problems. Acta Numerica **14**, 1–137 (2005). https://doi.org/10.1017/S0962492904000212
5. Brenner, S.C., Scott, L.R.: The Mathematical Theory of Finite Element Methods. Springer, USA (2008). Third Edition
6. Chizhonkov, E.V., Olshanskii, M.A.: On the domain geometry dependence of the LBB condition. ESAIM: M2AN **34**(5), 935–951 (2000). https://doi.org/10.1051/m2an:2000110
7. Elman, H., Silvester, D., Wathen, A.: Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics. Oxford University Press, UK (2014). Second Edition
8. Elman, H.C., Ramage, A., Silvester, D.J.: IFISS: A computational laboratory for investigating incompressible flow problems. SIAM Review **56**(2), 261–273 (2014). https://doi.org/10.1137/120891393
9. Golub, G.H., Van Loan, C.F.: Matrix Computations. The John Hopkins University Press, USA (2013). Fourth Edition
10. Greenbaum, A.: Iterative Methods for Solving Linear Systems. SIAM, USA (1997). First Edition
11. Gunzburger, M.D., Webster, C.G., Zhang, G.: Stochastic finite element methods for partial differential equations with random input data. Acta Numerica **23**, 521–650 (2014). https://doi.org/10.1017/S0962492914000075
12. Jiránek, P., Strakos, Z., Vohralík, M.: A posteriori error estimates including algebraic error and stopping criteria for iterative solvers. SIAM J. Sci. Comput. **32**(3), 1567–1590 (2010). https://doi.org/10.1137/08073706X
13. Kay, D., Silvester, D.: A posteriori error estimation for stabilized mixed approximations of the Stokes equations. SIAM J. Sci. Comput. **24**(1), 1321–1336 (1999). https://doi.org/10.1137/S1064827598333715
14. Lanczos, C.: An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. J. Research Nat. Bur. Standards **45**(4), 255–282 (1950). https://doi.org/10.6028/jres.045.026
15. Mardal, K., Winther, R.: Preconditioning discretizations of systems of partial differential equations. Numerical Linear Algebra with Applications **18**(1), 1–40 (2011). https://doi.org/10.1002/nla.716
16. Oden, J.T., Demkowicz, L.F.: Applied Functional Analysis. CRC Press, USA (1996). First Edition
17. Paige, C.C., Saunders, M.A.: Solution of sparse indefinite systems of linear equations. SIAM J. Numer. Anal. **12**(4), 617–629 (1975). https://doi.org/10.1137/0712047
18. Parlett, B.N.: The Symmetric Eigenvalue Problem. SIAM, USA (1998)
19. Pietro, D.A.D., Flauraud, E., Vohralík, M., Yousef, S.: 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**, 163–187 (2014). https://doi.org/10.1016/j.jcp.2014.06.061
20. Pietro, D.A.D., Vohralík, M., Yousef, S.: 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**(12 B), 2331–2347 (2014). https://doi.org/10.1016/j.camwa.2014.08.008
21. Pranjal: Optimal iterative solvers for linear systems with stochastic PDE origins: Balanced black-box stopping tests, PhD thesis. University of Manchester, UK (2017). PhD Thesis
22. Silvester, D., Pranjal: An optimal solver for linear systems arising from stochastic FEM approximation of diffusion equations with random coefficients. SIAM/ASA J. Uncertainty Quantification **4**(1), 298–311 (2016). https://doi.org/10.1137/15M1017740
23. Silvester, D.J., Simoncini, V.: An optimal iterative solver for symmetric indefinite systems stemming from mixed approximation. ACM Trans. Math. Softw. **37**(4) (2011). https://doi.org/10.1145/1916461.1916466
24. Verfürth, R.: A Posteriori Error Estimation Techniques for Finite Element Methods. Oxford University Press, UK (2013). First Edition
25. Wathen, A.: Preconditioning and convergence in the right norm. Int. J. Comput. Math. **84**(8), 1199–1209 (2007). https://doi.org/10.1080/00207160701355961