An optimal iterative solver for linear systems arising from SFEM approximation of diffusion equations with random coefficients

Silvester, David and Pranjal,

2015

MIMS EPrint: **2015.20**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

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

ISSN 1749-9097

AN OPTIMAL ITERATIVE SOLVER FOR LINEAR SYSTEMS ARISING FROM SFEM APPROXIMATION OF DIFFUSION EQUATIONS WITH RANDOM COEFFICIENTS

DAVID SILVESTER† AND PRANJAL‡

Abstract. This paper discusses the design and implementation of efficient solution algorithms for symmetric linear systems associated with stochastic Galerkin approximation of elliptic PDE problems with correlated random data. The novel feature of our iterative solver is the incorporation of error control in the natural "energy" norm in combination with an effective a posteriori estimator for the PDE approximation error. This leads to a robust and optimally efficient stopping criterion: the iteration is terminated as soon as the algebraic error is insignificant compared to the approximation error.

Key words. Stochastic Galerkin approximation, PDEs with random data, parametric operator equations, a posteriori error analysis, iterative solvers, MINRES, optimal preconditioning.

AMS subject classifications. 35R60, 65C20, 65N30, 65N15.

1. Introduction. Stochastic spectral finite element methods are becoming popular for the numerical solution of steady-state diffusion problems whose permeability coefficients are random fields. Galerkin approximation of the probabilistic dimension of the PDE coupled with the finite element discretization in space gives rise to a single linear(ized) system of equations whose coefficient matrix is symmetric positive definite but is ill-conditioned with respect to the discretization parameters. Since the coefficient matrix has a well-defined sparse structure, iterative solution methods can still be extremely effective. While preconditioned conjugate gradient (CG) methods are commonly used for this purpose, we will focus on the classic MINRES algorithm of Paige & Saunders [11] in this work. Specifically we will develop a new solver for discretized diffusion problems with uncertain coefficients by extending the EST_MINRES algorithm developed for discrete saddle-point systems by Silvester & Simoncini in [14]. The extended algorithm has two important features. First, the use of a block preconditioner which will guarantee convergence independent of the problem parameters, and second, the use of a stopping criterion for the iterative solver which balances the approximation error with the algebraic error. The latter requires a reliable a posteriori error estimation technique for computing the approximation error. The specific implementation discussed in this paper builds on the energy error estimation framework developed by Bespalov et al. in [3].

Wathen [15] observed that numerical finite element approximation of PDEs endows the problem with a 'natural norm' that is determined by the specific approximation space. If u, u_h, $u_h^{(k)}$ are the true solution, the Galerkin solution and algebraic solution at the kth step ($k = 0, 1, \ldots$) of the iterative solver respectively, then it is shown that the 'natural norm' on the total error ($u - u_h^{(k)}$) at the kth step of iteration consists of two parts—the natural norm of the true approximation error ($u - u_h$) and the natural norm of the algebraic error ($u_h - u_h^{(k)}$). Note that the total error is actually the approximation error at the kth step of the iteration. For symmet-

‡School of Mathematics, University of Manchester, Manchester, M13 9PL, United Kingdom, pranjal.chess@gmail.com

†School of Mathematics, University of Manchester, Manchester, M13 9PL, United Kingdom, d.silvester@manchester.ac.uk

ric positive-definite linear systems the 'natural norm' is the underlying *energy* norm. Unfortunately, computing errors in the energy norm is anything but straightforward.

The preconditioned MINRES algorithm minimizes the norm of the residual of the linear algebra system in terms of a norm that involves the preconditioner. This 'surrogate norm' is monotonically decreasing and is readily computable. Thus, the choice of preconditioner is important for two different reasons. First for accelerating convergence, and second, for ensuring that the convergence criterion matches the 'natural norm' of the discrete problem. In order to determine the stopping criterion of the iterative solver, we must balance this energy norm of the algebraic error with the energy norm of the approximation error at that step of iteration. Our stopping criterion provides an inbuilt tolerance for the solver by balancing the approximation error and the algebraic error. In this sense our solver is optimal.

The target linear algebra problem is set up in section 2 and the natural (energy) norm is identified. This section also contains an overview of preconditioned MINRES and develops the rationale for our stopping methodology. In section 3 we present a set of computational results that can be reproduced using the S-IFISS toolbox [2] and which confirm the effectiveness of our optimal stopping strategy. Throughout the discussion, $\|\cdot\|$, $(\cdot,\cdot)$, $\mathbb{R}$ will denote the ℓ_2 norm, the ℓ_2 inner product and the set of real numbers respectively.

2. Parameter dependent linear systems. The goal of this work is to develop effective methods for solving parameterized symmetric linear systems of equations of the form

$$A(y)\,u(y) = f, \tag{2.1}$$

where the entries in the matrix $A \in \mathbb{R}^{n\times n}$ (and hence the solution vector u) depend on a set of m parameters $y = [y_1, y_2, \ldots, y_m]^T$. Working in a statistical setting the parameters might represent independent observations of uniform random variables taking values in a bounded interval $[a, b]$. In practice the parameter dependence is often taken to be *linear*, in which case we can decompose the coefficient matrix so that

$$A(y) = A_0 + \sigma \sum_{k=1}^{m} y_k A_k, \tag{2.2}$$

where $A_0 \in \mathbb{R}^{n\times n}$ might be associated with the *mean* of the coefficients and the parameter $\sigma > 0$ might represent the standard deviation of the fluctuations. In typical applications, the matrix A_0 will be positive definite, but the matrices A_k may be indefinite. In such cases, $A(y)$ can only be guaranteed to be invertible when σ is small enough. We will return to this issue later.

Symmetric systems of this type arise in the solution of linear elliptic partial differential equations with *random coefficients*. An example might be a heat conduction problem in a region containing m different materials: each having thermal conductivity coefficient that is not known precisely. Discretization of such a PDE problem (for example, using finite element approximation in space) typically leads to a linear algebra system with a coefficient matrix of the form (2.2).

Classical Monte-Carlo (MC) methods are traditionally used to produce sample solutions from independent realisations of parameter inputs. Statistics of the stochastic solution may then be generated by postprocessing the sample solutions. Although they are robust and easily parallelizable, MC sampling can be an incredibly inefficient

use of computational resources, especially when solving large scale models where a single deterministic solve of the original system is computationally expensive. An efficient alternative to MC sampling is to use a stochastic Galerkin method. These first appeared in the engineering literature in the 1990s (see Ghanam & Spanos [6]) and have been extensively studied over the last decade (see, for example, Deb et al. [4], Babuska et al. [1]). Using this approach, the parameters are approximated by multivariate polynomials (Legendre polynomials in the case of uniform random variables) of total degree p. Each solution coefficient is then written as a linear combination of the polynomial basis functions

$$u_i = u_i^1 \xi_1 + u_i^2 \xi_2 + \ldots + u_i^{n_\xi} \xi_{n_\xi} \tag{2.3}$$

and the system is projected (in a least-squares sense) to give the best approximation to the solution u from a finite-dimensional subspace $S_p = \text{span}\{\xi_j\}_{j=1}^{n_\xi}$. The Galerkin projection process leads to the linear algebra system

$$A_0 X + \sigma \sum_{k=1}^{m} A_k X G_k = F \tag{2.4}$$

where X is the $n \times n_\xi$ matrix of the unknown coefficients u_i^j and G_k is the weighted Gram matrix associated with the kth parameter. Note that writing $\mathbf{x} = \texttt{vec}(X)$ leads to an equivalent high-dimensional system (of dimension $n \cdot n_\xi$) with a characteristic Kronecker product structure

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

2.1. A model PDE problem. The simplest example of a problem that fits into the framework above is the model of a diffusion process 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. centred random variables, so that

$$\kappa(\mathbf{x}, y_1, \ldots, y_m) = \mu(\mathbf{x}) + \sum_{k=1}^{m} \psi_k(\mathbf{x})\, y_k. \tag{2.6}$$

Here $\mu(\mathbf{x})$ is the mean value of the permeability coefficient at the point $\mathbf{x} \in D$, $y_k \in \Gamma_k$ 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 $u(\mathbf{x}, \mathbf{y}) : D \times \Gamma \to \mathbb{R}$ satisfies

$$-\nabla \cdot K(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}), \qquad \mathbf{x} \in D \subset \mathbb{R}^d,\ (d = 2, 3),\ \mathbf{y} \in \Gamma, \tag{2.7a}$$

$$u(\mathbf{x}, \mathbf{y}) = g(\mathbf{x}), \qquad \mathbf{x} \in \partial D_D,\ \mathbf{y} \in \Gamma, \tag{2.7b}$$

$$K(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y}) \cdot \vec{n} = 0, \qquad \mathbf{x} \in \partial D_N = \partial D \backslash \partial D_D,\ \mathbf{y} \in \Gamma, \tag{2.7c}$$

almost surely, where ∂D_D, ∂D_N are the Dirichlet and the Neumann part of the boundary of D, and f and g are given deterministic functions.

The variational formulation of (2.7) is associated with the space W of functions that are zero for all realizations on the Dirichlet boundary, and whose 'stochastic energy' (defined later) is finite. The goal is to find u such that $u - g \in W$ satisfies

$$\left\langle \int_D K(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y}) \cdot \nabla w(\mathbf{x}, \mathbf{y})\, \mathrm{d}\mathbf{x} \right\rangle = \left\langle \int_D f(\mathbf{x}) w(\mathbf{x}, \mathbf{y})\, \mathrm{d}\mathbf{x} \right\rangle \tag{2.8}$$

for all $w \in W$, where $\langle\cdot\rangle$ denotes the expected value of a multivariate random variable defined on the space[1] $(\Gamma, \mathcal{B}(\Gamma), \pi)$ with joint probability density function $\rho(\mathbf{y})$ defined on the product set Γ. It can thus be restated as: find $u - g \in W$ such that

$$\int_\Gamma \rho(\mathbf{y}) \int_D K(\mathbf{x},\mathbf{y})\, \nabla u(\mathbf{x},\mathbf{y}) \cdot \nabla w(\mathbf{x},\mathbf{y})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} = \int_\Gamma \rho(\mathbf{y}) \int_D f(\mathbf{x}) w(\mathbf{x},\mathbf{y})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y}, \tag{2.9}$$

for all $w \in W$. Note that the left side of (2.9) characterises the energy norm

$$\|w\|_W^2 = \int_\Gamma \rho(\mathbf{y}) \int_D K(\mathbf{x},\mathbf{y})\, |\nabla w(\mathbf{x},\mathbf{y})|^2\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y}, \tag{2.10}$$

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

A measure that will be equally important later is the energy norm of the solution when the permeability coefficient is given by the mean field, that is

$$\|w\|_E^2 = \int_\Gamma \rho(\mathbf{y}) \int_D \mu(\mathbf{x})\, |\nabla w(\mathbf{x},\mathbf{y})|^2\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} = \left\langle \int_D \mu(\mathbf{x})\, |\nabla w(\mathbf{x},\mathbf{y})|^2\, \mathrm{d}\mathbf{x} \right\rangle. \tag{2.11}$$

Identifying the form of the ('natural') energy norms of the finite-dimensional version of (2.8) is the first step towards expressing the energy norm of the algebraic error in terms of a computable norm of the algebraic residual. The key point here is that the two norms are equivalent whenever the formulation (2.8) is well posed (see [3]); that is, there exist positive constants λ and Λ such that

$$\lambda \|w\|_W^2 \le \|w\|_E^2 \le \Lambda \|w\|_W^2 \qquad \forall w \in W. \tag{2.12}$$

Galerkin approximation of (2.9) is associated with choosing finite dimensional subspaces of the component spaces, that is $X_h \subset H_0^1(D)$ and $S_p \subset L^2(\Gamma)$ so that $W_{h,p} = H_0^1(D) \otimes L^2(\Gamma)$. Full details can be found in [3] or [10, section 9.5]. When generating test problems using the S-IFISS toolbox [2], the spatial domain D is two-dimensional and the approximation is either piecewise bilinear ($\boldsymbol{Q}_1$) or biquadratic ($\boldsymbol{Q}_2$) on a rectangular grid. This leads to sparse (stiffness) matrices A_0 and A_k in (2.4) and (2.5). In contrast, the parameter approximation space S_p is composed of (multivariate) gobal polynomials, as in (2.3). For this space, it is sensible to choose a basis set $\{\xi_j\}_{j=1}^{n_\xi}$ that is orthogonal 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 an effective iterative solver.

2.2. A fast iterative solver. Looking at the structure of our target system (2.5) it is clear that the positive-definite matrix $\mathcal{M}^{-1} = I \otimes A_0$ will give an effective approximation of $\mathcal{A}$ whenever σ is small relative to $\|A_0\|$. The use of $\mathcal{M}$ as a *preconditioner* for the system (2.5) will be a key component of our iterative solution strategy. (In the stochastic Galerkin literature, this is sometimes referred to as *mean-based* preconditioning.) Since $I \otimes A_0$ is a block-diagonal matrix, the action of its inverse can be effected by a single sparse factorisation (of A_0) followed by a n_ξ forward and backward substitutions. Characterising a precise stopping criteria for our solver will require accurate estimates of Rayleigh quotient bounds θ, Θ satisfying

$$\theta \le \frac{\mathbf{x}^T \mathcal{A}\, \mathbf{x}}{\mathbf{x}^T \mathcal{M}^{-1}\, \mathbf{x}} \le \Theta, \quad \forall \mathbf{x} \in \mathbb{R}^{n \cdot n_\xi}. \tag{2.13}$$

[1] The triple $(\Gamma, \mathcal{B}(\Gamma), \pi)$ is assumed to define a probability space, see Lord et al. [10, section 4.1].

An analysis of the spectral equivalence of $\mathcal{M}^{-1}$ and $\mathcal{A}$ for our model PDE problem can be found in Powell & Elman [13]. Note that since (2.12) is the infinite dimensional analogue of (2.13), 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. To address this, an effective way of computing estimates of θ and Θ on-the-fly will be presented in section 2.4.

Solving our target system (2.5) using MINRES requires an initial vector $\mathbf{x}^{(0)}$, and computes a sequence of iterates $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}, \ldots$ from the shifted Krylov space

$$\mathbf{x}^{(0)} + \text{span}\{\mathbf{r}^{(0)}, \mathcal{A}\mathbf{r}^{(0)}, \ldots, \mathcal{A}^{(k-1)}\mathbf{r}^{(0)}\}, \tag{2.14}$$

where $\mathbf{r}^{(0)} = \mathbf{f} - \mathcal{A}\mathbf{x}^{(0)}$ is the initial residual and k is the iteration number . The characteristic feature of MINRES (see Liesen & Strakos [9, chap. 2]) is that the iterate $\mathbf{x}^{(k)}$ minimizes the Euclidean (ℓ_2) norm $\|\cdot\|$ of the corresponding residual $\mathbf{r}^{(k)}$ over the shifted Krylov space $\mathbf{r}^{(0)} + \text{span}\{\mathcal{A}\mathbf{r}^{(0)}, \mathcal{A}^2\mathbf{r}^{(0)}, \ldots, \mathcal{A}^k\mathbf{r}^{(0)}\}$.

The convergence estimate for unpreconditioned MINRES resulting from the minimum residual criteria is the *optimal* polynomial estimate

$$\|\mathbf{r}^{(k)}\| \leq \min_{p_k \in \Pi_k,\, p_k(0)=1} \max_j |p_k(\lambda_j)|\, \|\mathbf{r}^{(0)}\|, \tag{2.15}$$

where Π_k is the set of real polynomials of degree less than or equal to k and $\{\lambda_j\}$ are the eigenvalues of $\mathcal{A}$ (see Elman et al. [5, p. 191]). Let $\mathcal{M}$ be the mean-based preconditioner for the system (2.5). Since $\mathcal{M}$ is a (symmetric) positive definite matrix, we can write $\mathcal{M}^{-1} = HH^T$ and consider the preconditioned system

$$H^{-1}\,\mathcal{A}\,H^{-T}\mathbf{y} = H^{-1}\,\mathbf{f}, \qquad \mathbf{y} = H^T\mathbf{x}. \tag{2.16}$$

Clearly, any solution to (2.5) is also a solution to (2.16) and vice-versa. If MINRES is applied to the symmetric system (2.16) then $\|H^{-1}\mathbf{r}^{(k)}\|$ will be minimized over the space $H^{-1}(\mathbf{r}^{(0)} + \text{span}\{\mathcal{A}\mathcal{M}\mathbf{r}^{(0)}, (\mathcal{A}\mathcal{M})^2\mathbf{r}^{(0)}, \ldots, (\mathcal{A}\mathcal{M})^k\mathbf{r}^{(0)}\})$ at the kth iteration. In fact, we have

$$\|H^{-1}\mathbf{r}^{(k)}\|^2 = \|\mathbf{r}^{(k)}\|_{\mathcal{M}}^2 = (\mathbf{r}^{(k)})^T\mathcal{M}\,\mathbf{r}^{(k)}. \tag{2.17}$$

So, for preconditioned MINRES the convergence estimate analogous to (2.15) is

$$\frac{\|\mathbf{r}^{(k)}\|_{\mathcal{M}}}{\|\mathbf{r}^{(0)}\|_{\mathcal{M}}} \leq \min_{p_k \in \Pi_k,\, p_k(0)=1} \max_j |p_k(\lambda_j)|, \tag{2.18}$$

where λ_j are the eigenvalues of the coefficient matrix $H^{-1}\,\mathcal{A}\,H^{-T}$ in (2.16). Note that, from the similarity transformation, $\mathcal{M}\mathcal{A} = H^{-T}\,(H^{-1}\,\mathcal{A}\,H^{-T})\,H^T$ it follows that λ_j are also the eigenvalues of $\mathcal{M}\mathcal{A}$. (Note that H is not needed in actual computation—one only requires the action of $\mathcal{M}$ on a vector.) The optimal bound in (2.18) can be weakened to a bound over the fixed interval $[\lambda, \Lambda]$ by using the Rayleigh quotient bounds on the eigenvalues of the preconditioned matrix in (2.13),

$$\frac{\|\mathbf{r}^{(k)}\|_{\mathcal{M}}}{\|\mathbf{r}^{(0)}\|_{\mathcal{M}}} \leq \min_{p_k \in \Pi_k,\, p_k(0)=1} \max_{z\in[\theta,\Theta]} |p_k(z)| \tag{2.19a}$$

$$\leq \min_{p_k \in \Pi_k,\, p_k(0)=1} \max_{z\in[\lambda,\Lambda]} |p_k(z)| =: \rho_k. \tag{2.19b}$$

This implies that, when applied to our model problem, the preconditioned MINRES iteration is guaranteed to satisfy a prescibed residual error tolerance in a fixed number of iterations. The convergence rate of the solver will thus be bounded *independently* of the stochastic Galerkin discretization parameters h and p.

2.3. An optimal stopping criterion. The real goal is to bound the norm of the energy error in terms of the preconditioned residual norm that is minimised at every step of MINRES. To this end, if $\mathbf{r}^{(0)}$ is the initial residual and $\mathbf{r}^{(k)}$ is the residual at the kth iteration step, then inverting the eigenvalue bounds in (2.13) gives

$$\frac{1}{\Theta} \;\le\; \frac{(\mathbf{r}^{(0)})^T \mathcal{A}^{-1}\,\mathbf{r}^{(0)}}{(\mathbf{r}^{(0)})^T \mathcal{M}\,\mathbf{r}^{(0)}}, \quad \frac{(\mathbf{r}^{(k)})^T \mathcal{A}^{-1}\,\mathbf{r}^{(k)}}{(\mathbf{r}^{(k)})^T \mathcal{M}\,\mathbf{r}^{(k)}} \;\le\; \frac{1}{\theta}, \tag{2.20}$$

The algebraic error at the kth step is $\mathbf{e}^{(k)} = \mathbf{x} - \mathbf{x}^{(k)} = \mathcal{A}^{-1}\mathbf{r}^{(k)}$, thus

$$\|\mathbf{e}^{(k)}\|_{\mathcal{A}}^2 = (\mathbf{e}^{(k)})^T \mathcal{A}\,\mathbf{e}^{(k)} = (\mathbf{r}^{(k)})^T \mathcal{A}^{-1}\,\mathbf{r}^{(k)} = \|\mathbf{r}^{(k)}\|_{\mathcal{A}^{-1}}^2. \tag{2.21}$$

Combining (2.21) with (2.20) and (2.19b) gives the bounds

$$\frac{\|\mathbf{e}^{(k)}\|_{\mathcal{A}}}{\|\mathbf{e}^{(0)}\|_{\mathcal{A}}} \;\le\; \sqrt{\frac{\Theta}{\theta}}\;\frac{\|\mathbf{r}^{(k)}\|_{\mathcal{M}}}{\|\mathbf{r}^{(0)}\|_{\mathcal{M}}} \le \sqrt{\frac{\Lambda}{\lambda}}\,\rho_k \tag{2.22a}$$

$$\|\mathbf{e}^{(k)}\|_{\mathcal{A}} \;\le\; \frac{1}{\sqrt{\theta}}\;\|\mathbf{r}^{(k)}\|_{\mathcal{M}}. \tag{2.22b}$$

The quantity $\frac{1}{\sqrt{\theta}}\;\|\mathbf{r}^{(k)}\|_{\mathcal{M}}$ will be called the algebraic error *bound* in the rest of the paper.

To devise the optimal stopping criterion, we make the premise that the *algebraic* error at a given step of the iteration cannot be worse than the *approximation* error at that step. To see what this means in the context of our model problem let $u_h - g \in W_{h,p}$ be the approximation to the PDE solution u. Then the vector $\mathbf{x}$ solving (2.5) will be the coordinate vector of u_h with respect to a chosen ordered basis. Next, let $\mathbf{x}^{(k)}$ be the coordinate vector at the kth iteration of the algebraic solver. Corresponding to this kth iterate one can form the kth approximation $u_h^{(k)}$ and estimate the (mean) energy error $\|u - u_h^{(k)}\|_W$ a posteriori (for example, by using the energy estimator developed in [3]). That is, one can compute $\eta^{(k)}$ satisfying

$$c\,\eta^{(k)} \;\le\; \|u - u_h^{(k)}\|_E \le\; C\,\eta^{(k)}, \qquad \frac{C}{c} \sim O(1). \tag{2.23}$$

Combining the triangle inequality with Galerkin orthogonality gives the decomposition into the Galerkin solution approximation error and the algebraic error,

$$\|u - u_h^{(k)}\|_E^2 = \|u - u_h\|_E^2 + \|u_h - u_h^{(k)}\|_E^2. \tag{2.24}$$

Thus, assuming the a posteriori error estimates η and η^k are close estimates of the true approximation error and the approximation error at the kth iteration step, respectively, (2.24) can be rewritten as

$$\eta^{(k)} \;\simeq\; \eta \;+\; \underbrace{\|u_h - u_h^{(k)}\|_E}_{\|\mathbf{e}^{(k)}\|_{\mathcal{A}}}, \qquad k \;=\; 0, 1, 2, \ldots \tag{2.25}$$

Our iteration procedure can thus be looked upon as that of constructing a sequence, $\{\eta^{(k)}\}$, which converges to η. An efficient stopping point is the point when the contribution of the energy norm of the algebraic error to the sum in (2.25) becomes insignificant. In the light of (2.22b) our preconditioned MINRES iteration will be stopped at iteration k^*, the smallest value of k for which

$$\frac{1}{\sqrt{\theta}}\;\|\mathbf{r}^{(k)}\|_{\mathcal{M}} \;\le\; \eta^{(k)}. \tag{2.26}$$

A clever way of estimating the constant θ on the fly will be discussed next.

2.4. A posteriori estimation of the norm equivalence bounds. As already mentioned, MINRES generates a sequence of approximations $\mathbf{x}^{(k)}$, $k = 1, 2, \ldots$ from a shifted Krylov space such that the residual $\mathbf{r}^{(k)} = \mathbf{f} - \mathcal{A}\mathbf{x}^{(k)}$ is minimized. Let us suppose that $\{\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(k)}\}$ is a set of orthonormal vectors spanning the k-dimensional Krylov space, with $\mathbf{w}^{(1)} = \mathbf{f}/\|\mathbf{f}\|$, and let $W_k = [\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(k)}]$. These basis vectors can be generated iteratively using the recurrence:

$$\mathcal{A}W_k = W_k T_k + t_{k+1,k}\, \mathbf{w}^{(k+1)} \mathbf{e}_k^T =: W_{k+1}\, \underline{T}_k, \tag{2.27}$$

where $\mathbf{e}_k$ is the kth vector of the canonical basis and T_k is a tridiagonal symmetric matrix containing the orthogonalization coefficients; full details can be found in Greenbaum [8, sect. 2.5]. Using (2.27) leads to the following characterisation of the iterate, $\mathbf{x}^{(k)} = W_k \mathbf{y}^{(k)}$,

$$\mathbf{r}^{(k)} = \mathbf{f} - \mathcal{A}\mathbf{x}^{(k)} = V_{k+1} \left(\mathbf{e}_1 \|\mathbf{f}\| - \underline{T}_k \mathbf{y}^{(k)} \right). \tag{2.28}$$

The minimizing solution $\mathbf{x}^{(k)}$ may then be found by solving the least squares problem $\min_{\mathbf{y}} \|\mathbf{e}_1\|\mathbf{f}\| - \underline{T}_k \mathbf{y}\|$. The Lanczos relation in (2.27) can also be used to show that $T_k = W_k^T \mathcal{A} W_k$ so that the eigenvalues of T_k, also known as the *Ritz values*,[2] provide approximations to the eigenvalues of $\mathcal{A}$ (or of $\mathcal{M}\mathcal{A}$ if the matrix is preconditioned).

In our setting we would like to estimate the extremal eigenvalues θ, Θ of the preconditioned matrix associated with (2.13) on the fly. What works in our favour is the fact that extremal Ritz values can be readily computed at every step of the MINRES iteration and that they provide accurate estimates of the extremal eigenvalues, even when k (the number of iterations) is relatively small. This key aspect is discussed in Parlett [12, chap. 13]. The efficiency of this eigenvalue estimation strategy will be confirmed by the computational experiments presented in the next section.

3. Computational results. To give a proof of concept, the results of computational experiments when the stopping test (2.26) is applied to systems of the form (2.1) derived from the model PDE problem (2.7) will be presented in this section. Representative systems (2.5) can be generated for this purpose using the S-IFISS toolbox [2]. Our test problem is defined on a square domain $D = (-1, 1) \times (-1, 1)$ with source function $f(\mathbf{x}) = \frac{1}{8}(2 - x_1^2 - x_2^2)$ and zero Dirichlet condition everywhere on the boundary. The spatial approximation space X_h is a piecewise bilinear finite element space on a uniform of grid of square elements (here $h > 0$ denotes the length of each element edge). We will present results for $h = 2^{1-l}, (l = 3, 4, 5, 6)$.

The S-IFISS software generates a diffusion coefficient κ in (2.6) with uniform random variables defined on $\Gamma_k = [-1, 1]$, and the parameter approximation space S_p is spanned by complete polynomials of degree p. The mean field in the expansion (2.6) is constant, $\mu(\mathbf{x}) = 1$, and the spatial functions $\psi_k = \sqrt{3\lambda_k}\varphi_k$ in (2.6) are associated with eigenpairs $\{(\lambda_k, \varphi_k)\}_{k=1}^m$ of the (separable) covariance operator

$$C(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{1}{2} \|\mathbf{x} - \mathbf{x}'\|_{\ell_1} \right), \quad \mathbf{x}, \mathbf{x}' \in D \subset \mathbb{R}^2, \tag{3.1}$$

where σ denotes the standard deviation, and the correlation length is 2. We note that the resulting model problem is essentially the same as that considered in [4] and [13].

[2] The idea of exploiting the Lanczos connection was introduced by Silvester & Simoncini [14] in the context of saddle-point problems. The main difference is that harmonic Ritz values are used in [14] in place of Ritz values—estimates of *interior* eigenvalues are needed in the saddle-point case.

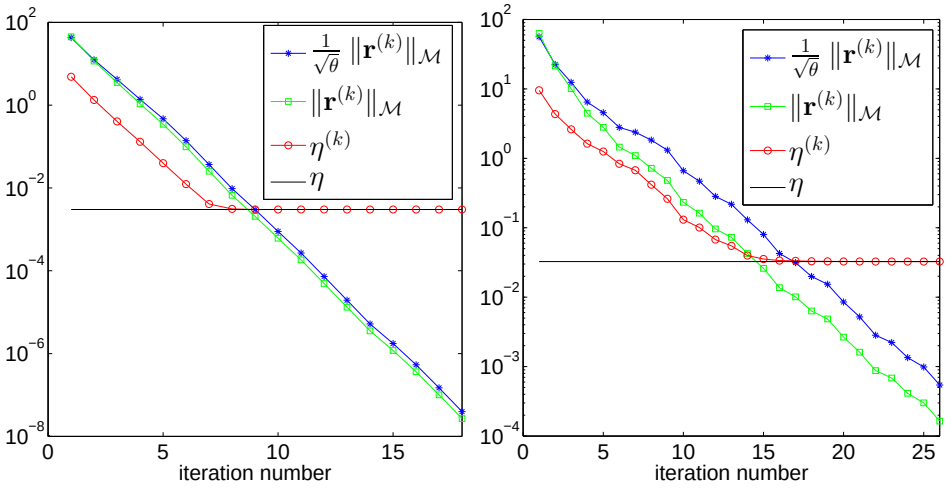

FIG. 1. *Errors vs iteration number for optimally preconditioned MINRES for the model PDE problem with* $h = 1/32$, $m = 5$, $p = 3$ | $\sigma = 0.3$ *(left),* $\sigma = 0.5$ *(right).*

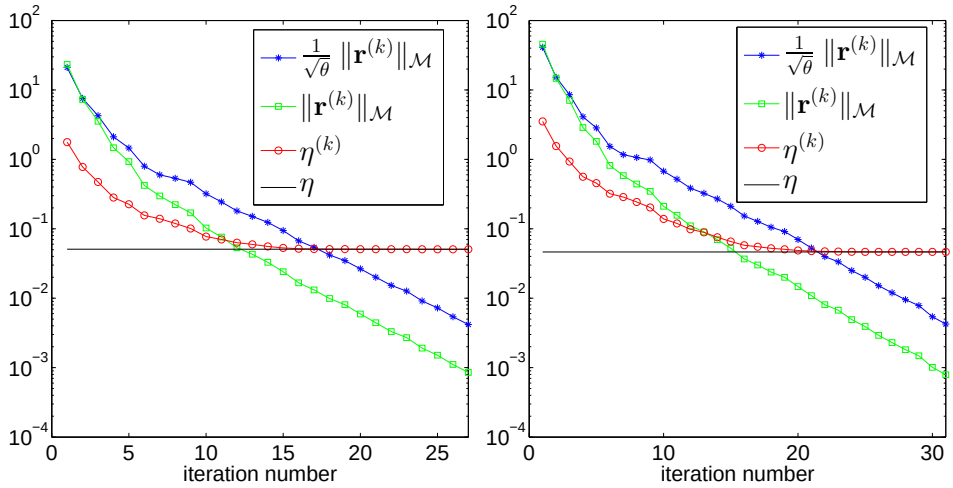

FIG. 2. *Errors vs iteration number for optimally preconditioned MINRES for the model PDE problem with* $m = 7$, $p = 3$ *for* $\sigma = 0.5$ | $h = 1/8$ *(left),* $h = 1/16$ *(right).*

The discretised problems were set up by running `stoch_diff_testproblem`. The resulting algebraic system was solved by calling a new function `stoch_est_minres`. A reference solution was computed in each case by turning off our optimal stopping criterion in `stoch_est_minres`, and solving the discrete system with a preconditioned residual reduction tolerance of `1e-14`. We will compare the reference solution $\mathbf{x}$ obtained in this way with the result $\mathbf{x}^{(k^*)}$ computed using the optimal stopping test. The starting vector $\mathbf{x}^{(\mathbf{0})}$ is always generated using the MATLAB function `rand`.

Some representative results are presented in Figures 1 and 2. These figures show the evolution of the residual error $\|\mathbf{r}\|_{\mathcal{M}}$ together with the algebraic error bound; with θ estimated at each step k using the strategy outlined in section 2.4. The approxima-

tion error $\eta^{(k)}$ is also plotted at each step: it can be seen to converge to the estimate η associated with the reference solution. Note that we take 9 additional iterations after convergence to ensure that we have stopped at the right place. The results in Figure 1 show that the solver takes longer to converge when the standard deviation is increased (all other parameters being kept constant). This is to be expected since λ in (2.12) (and θ in (2.13) become increasingly negative when σ is increased.[3] (Indeed our PDE problem is not well posed for $\sigma > \sigma^* \approx 0.55$). The iteration automatically stops after 9 iterations when $\sigma = 0.3$, but takes about twice as may iterations when $\sigma = 0.5$. The results in Figure 2 illustrate that the convergence of the solver is essentially independent of the spatial discretization (assuming that m and σ are kept fixed). This is what we mean by fast convergence!

The number of iterations needed to satisfy the optimal stopping test (2.26) is compared in Tables 1–3 with the number (k_{tol}) needed to satisfy a fixed (absolute) tolerance of `1e-3`. These tables provide additional evidence that, for fixed stochastic parameters, the number of iterations stays bounded as the spatial grid is increasingly refined. This boundedness is also evident in the tabulated extremal Ritz values computed at iteration k^*.

TABLE 1
Iteration counts and Rayleigh quotients estimates for the case $\sigma = 0.3$ $m = 5$ and $p = 3$.

l	k_{tol}	k^*	θ^*	Θ^*
3	8	6	0.5276	1.5044
4	9	7	0.4833	1.5257
5	10	8	0.4734	1.5283
6	10	9	0.4708	1.5311

TABLE 2
Iteration counts and Rayleigh quotients estimates for the case $\sigma = 0.5$ $m = 5$ and $p = 3$.

l	k_{tol}	k^*	θ^*	Θ^*
3	17	11	0.1358	1.8789
4	20	14	0.1110	1.8941
5	21	16	0.1042	1.9032
6	22	17	0.1029	1.9045

TABLE 3
Iteration counts and Rayleigh quotients estimates for the case $\sigma = 0.5$ $m = 7$ and $p = 3$.

l	k_{tol}	k^*	θ^*	Θ^*
3	21	13	0.0908	1.9315
4	27	18	0.0558	1.9590
5	30	22	0.0413	1.9649

The stopping point Ritz estimates turn out to be extremely good approximations of the actual extremal eigenvalues. This is evident in Figures 1 and 2 —the algebraic bound curve is too close to the curve of the norm of the preconditioned residual for

[3]Sharp bounds $[1-\tau, 1+\tau]$ for the Rayleigh quotient (2.13) are established by Powell & Elman in [13, Theorem 3.8], where the factor τ is the sum of the norms $\|\psi_k\|_\infty$ of the functions in (2.6). These bounds suggest that convergence will also be affected if m is increased with σ kept fixed.

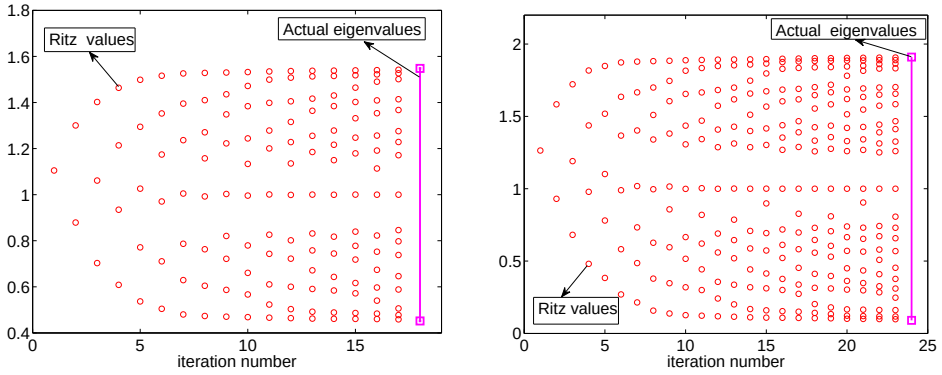

FIG. 3. *Computed Ritz values for the model PDE problem with* $m = 5$, $p = 3 \mid h = 1/16$ *and* $\sigma = 0.3$ *(left)* $h = 1/8$ *and* $\sigma = 0.5$ *(right).*

the first few iterations but the two curves rapidly become parallel as $\theta^{(k)}$ converges to θ. This is also illustrated by the plots in Figure 3 showing the convergence of the Ritz values. A final observation is that there is no sign of any "ghost" eigenvalues (see Golub & van Loan [7, p. 566, sect. 10.3.5]) in any of these computations.

REFERENCES

[1] I. M. Babuška, R. Tempone, and G. E. Zouraris, *Galerkin finite element approximations of stochastic elliptic partial differential equations*, SIAM J. Numer. Anal., 42 (2004), pp. 800–825.

[2] Alex Bespalov, Catherine Powell, and David Silvester, *S-IFISS ver. 1.0*, September 2013. available online at `http://www.manchester.ac.uk/ifiss/s-ifiss1.0.tar.gz`.

[3] ———, *Energy norm a posteriori error estimation for parametric operator equations*, SIAM J. Sci. Comput., 36 (2014), pp. A339–A363. `http://dx.doi.org/10.1137/130916849`.

[4] 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.

[5] 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.

[6] Roger G. Ghanem and Pol D. Spanos, *Stochastic finite elements: a spectral approach*, Springer-Verlag, New York, 1991.

[7] Gene Golub and Charles Van Loan, *Matrix Computations*, The John Hopkins University Press, Baltimore, USA, 2013. Fourth Edition.

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

[9] Jörg Liesen and Zdenek Strakos, *Krylov Subspace Methods, Principles and Analysis*, Oxford University Press, Oxford, UK, 2012.

[10] Gabriel J. Lord, Catherine E. Powell, and Tony Shardlow, *An Introduction to Computational Stochastic PDEs*, Cambridge University Press, Cambridge, UK, 2014.

[11] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equation*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.

[12] B. N. Parlett, *The Symmetric Eigenvalue Problem*, SIAM, Philadelphia, PA, 1998.

[13] 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.

[14] David J. Silvester and Valeria Simoncini, *An optimal iterative solver for symmetric indefinite systems stemming from mixed approximation*, ACM Trans. Math. Softw., 37 (2011). `http://dx.doi.org/10.1145/1916461.1916466`.

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