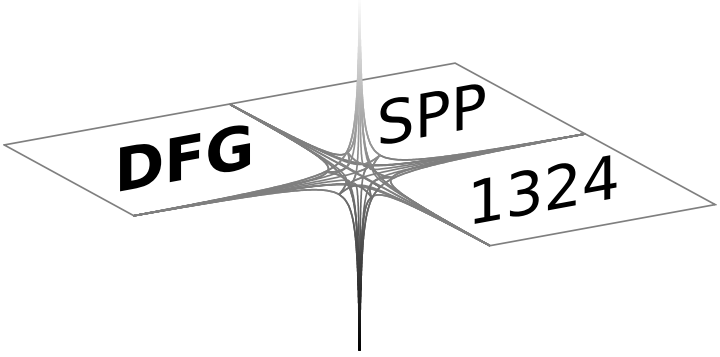
DFG-Schwerpunktprogramm 1324

„Extraktion quantifizierbarer Information aus komplexen Systemen”

Preconditioning Stochastic Galerkin Saddle Point Problems

C. E. Powell, E. Ullmann

Preprint 31



Edited by

AG Numerik/Optimierung
Fachbereich 12 - Mathematik und Informatik
Philipps-Universität Marburg
Hans-Meerwein-Str.
35032 Marburg

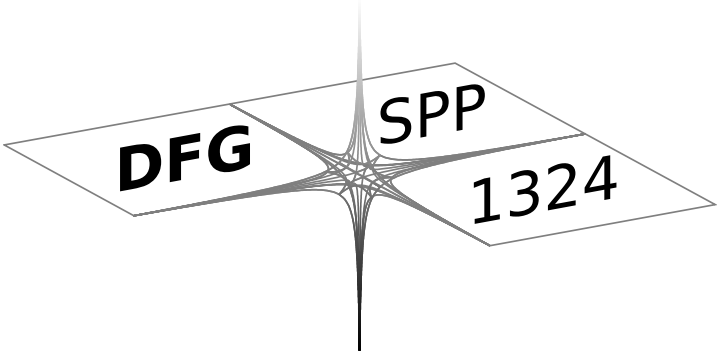
DFG-Schwerpunktprogramm 1324

„Extraktion quantifizierbarer Information aus komplexen Systemen”

Preconditioning Stochastic Galerkin Saddle Point Problems

C. E. Powell, E. Ullmann

Preprint 31



The consecutive numbering of the publications is determined by their chronological order.

The aim of this preprint series is to make new research rapidly available for scientific discussion. Therefore, the responsibility for the contents is solely due to the authors. The publications will be distributed by the authors.

PRECONDITIONING STOCHASTIC GALERKIN SADDLE POINT SYSTEMS

CATHERINE E. POWELL † AND ELISABETH ULLMANN ‡

Abstract. Mixed finite element discretizations of deterministic second-order elliptic partial differential equations (PDEs) lead to saddle point systems for which the study of iterative solvers and preconditioners is mature. Galerkin approximation of solutions of stochastic second-order elliptic PDEs, which couple standard mixed finite element discretizations in physical space with global polynomial approximation on a probability space, also give rise to linear systems with familiar saddle point structure. For stochastically nonlinear problems, the solution of such systems presents a serious computational challenge. The blocks are sums of Kronecker products of pairs of matrices associated with two distinct discretizations and the systems are large, reflecting the curse of dimensionality inherent in most stochastic approximation schemes. Moreover, for the problems considered herein, the leading blocks of the saddle point matrices are block-dense and the cost of a matrix vector product is non-trivial.

We implement a stochastic Galerkin discretization for the steady-state diffusion problem written as a mixed first-order system. The diffusion coefficient is assumed to be a lognormal random field, approximated via a nonlinear function of a finite number of unbounded random parameters. We study the resulting saddle point systems and investigate the efficiency of block-diagonal preconditioners of Schur complement and augmented type, for use with MINRES. By introducing so-called Kronecker product preconditioners we improve the robustness of cheap, mean-based preconditioners with respect to the statistical properties of the stochastically nonlinear diffusion coefficients.

Key words. saddle point matrices, preconditioning, Kronecker product, multigrid, stochastic Galerkin finite element method, lognormal random field, $H(div)$ approximation

AMS subject classifications. 35R60, 65C20, 65F10, 65N22, 65N30

1. Introduction.

We are interested in the design of efficient and robust preconditioners for a class of linear systems of equations with symmetric and indefinite coefficient matrices of the form

$$\widehat{C} := \begin{bmatrix} \widehat{A} & \widehat{B}^\top \\ \widehat{B} & 0 \end{bmatrix} \tag{1.1}$$

where $\widehat{A}$ is symmetric and positive definite and $\widehat{B}$ has full row rank. Such systems arise, notably, in the solution of PDEs via mixed finite element methods (e.g. see [10], [26], [16]). Today there is a large community of researchers dedicated to the task of solving $\widehat{C}\boldsymbol{x} = \boldsymbol{b}$ and the spectral properties of $\widehat{C}$, appropriate iterative solvers and preconditioners have been well studied, [6]. The appearance of the zero matrix in the (2,2) block and the fact that $\widehat{A}$ is positive definite mean that $\widehat{C}$ falls into a relatively easy class of saddle point matrices, for which the minimal residual method (MINRES, [25]) is an optimal iterative solver. Convergence can be accelerated using symmetric and positive definite preconditioners, of which there are two well-known types.

Writing $\widehat{S} = \widehat{B}\widehat{A}^{-1}\widehat{B}^\top$, the classical Schur complement preconditioner is

$$\widehat{P}_S := \begin{bmatrix} \widehat{A} & 0 \\ 0 & \widehat{S} \end{bmatrix}. \tag{1.2}$$

$\widehat{P}_S^{-1}\widehat{C}$ has only three distinct eigenvalues [24, Remark 1] and so (in exact arithmetic) preconditioned MINRES converges in at most three iterations. Obtaining a practical implementation of this preconditioner depends on our ability to approximate the actions of $\widehat{A}^{-1}$ and $\widehat{S}^{-1}$ on vectors, cheaply.

Alternatively, so-called augmented preconditioners of the form,

†School of Mathematics, University of Manchester, Oxford Road, Manchester, M13 9PL, United Kingdom (c.powell@manchester.ac.uk)

‡Institut für Numerische Mathematik und Optimierung, Technische Universität Bergakademie Freiberg, D-09596 Freiberg, Germany (ullmann@math.tu-freiberg.de). This author's research was partially supported by the DFG-Priority Program 1324 and by the U. S. Department of Energy under grant DEFG0204ER25619 during a visit at the University of Maryland, College Park.

$$\widehat{P}_A := \begin{bmatrix} \widehat{A} + \gamma^{-1}\widehat{B}^\top \widehat{W}^{-1}\widehat{B} & 0 \\ 0 & \gamma\widehat{W} \end{bmatrix} \tag{1.3}$$

which have their roots in the augmented Lagrangian method, are being adapted with success in many applications (e.g. see [16], [34]). Choosing the parameter γ and the symmetric positive definite weight matrix $\widehat{W}$ appropriately, is key. A smart choice of γ can force the eigenvalues of $\widehat{P}_A^{-1}\widehat{C}$ to cluster at ± 1, forcing MINRES to converge rapidly. For ease of solution, $\widehat{W}$ is typically chosen as an identity or mass matrix. For PDE problems, however, there is often an underlying bilinear form from the weak formulation that drives the choice of $\widehat{W}$ as $\widehat{A} + \widehat{B}^\top \widehat{W}^{-1}\widehat{B}$ is the natural matrix representation of a particular PDE operator. Obtaining a practical version of $\widehat{P}_A$ depends on the availability of cheap algorithms to approximate the action of $(\widehat{A} + \gamma^{-1}\widehat{B}^\top \widehat{W}^{-1}\widehat{B})^{-1}$ for the chosen $\widehat{W}$.

In this work, we are concerned specifically with saddle point matrices of the form

$$\widehat{C} := \begin{bmatrix} G_0 \otimes A_0 + \sum_{n=1}^{N} G_n \otimes A_n & G_0 \otimes B^\top \\ G_0 \otimes B & 0 \end{bmatrix} \tag{1.4}$$

where $\otimes$ denotes the Kronecker product. Matrices with this structure arise from stochastic Galerkin (SG) mixed finite element formulations of two-field PDE problems with random coefficients. Examples include the Darcy flow problem with random permeability coefficients and the Stokes problem with random viscosity. $A_0, A_1, \ldots, A_N$ and B are finite element matrices associated with the physical domain. They are sparse and usually ill-conditioned with respect to the finite element mesh size and, here, the statistical properties of the PDE coefficients. The matrices $G_0, G_1, \ldots, G_N$ represent multiplication operators on a probability space associated with the random PDE coefficients. Their structural and spectral properties (see [12], [27], [30]) are governed by our choice of discretization on the probability space. We assume that the (1,1) block in (1.4) is positive definite and so linear systems with this $\widehat{C}$ can be solved via preconditioned MINRES with the block-diagonal preconditioners described above. However, there are additional computational challenges. Due to the Kronecker product structure, the dimension of the system can be huge even if the physical domain is only two-dimensional. If the matrices $G_0, G_1, \ldots, G_N$ are not sparse and/or if N is large, then the cost of a matrix vector product is non-trivial.

In [11] and [15] an SG mixed formulation of the steady-state diffusion problem is studied and block-diagonal preconditioners for the resulting saddle point matrices are proposed. In those works, however, the diffusion coefficient is a *linear* function of M *bounded* random parameters, yielding $N = M$ and well-conditioned, sparse matrices $G_0, G_1, \ldots, G_M$ in (1.4). Here, we extend our earlier work to a new, more challenging model problem. We consider again the steady-state diffusion problem but now the diffusion coefficient is a *nonlinear* function of M *unbounded* random parameters. This has serious consequences for the linear algebra and new preconditioners are required.

1.1. Outline. In Section 2 we describe the model problem and an appropriate SG mixed finite element discretization. Properties of the resulting saddle point matrices are discussed in Section 3. In Section 4 we study a Schur complement preconditioner from [11] and in Section 5 we revisit a preconditioner from [15], which is equivalent to $\widehat{P}_A$ in (1.3) with $\gamma = 1$ and a certain $\widehat{W}$. We make essential improvements to both preconditioners for the new model problem by combining them with best Kronecker product approximation (see [33]). Practical preconditioners are derived by exploiting appropriate multigrid methods, and numerical results are presented in Section 6.

2. Stochastic steady-state diffusion problem. In many applications, there is a growing need to solve PDEs with inputs that are subject to uncertainty. Specifically, we mean problems

where the uncertainty stems from lack of knowledge about the data and one or more inputs to the PDE(s) of interest cannot be stated as functions of $\boldsymbol{x} \in D$ where $D \subset \mathbb{R}^d$ is the physical domain. Suppose T is a coefficient function that is not known at every $\boldsymbol{x} \in D$ but whose values at two distinct spatial locations are connected via a prescribed covariance function. We can model T as a random field, a real-valued function $T(\boldsymbol{x}, \omega) : D \times \Omega \to \mathbb{R}$ where $\omega \in \Omega$ is an abstract label for a realization of T. In deterministic models we prescribe $T = T(\boldsymbol{x})$ for every $\boldsymbol{x} \in D$ but here, $T = T(\boldsymbol{x}, \omega)$ and we prescribe a statistical distribution of $T(\boldsymbol{x}, \cdot)$ for every $\boldsymbol{x} \in D$ and a covariance function $C_T(\boldsymbol{x}, \boldsymbol{y})$.

The steady-state diffusion problem with random diffusion coefficient T can be stated as: find a random field $u(\boldsymbol{x}, \omega)$ that solves, P-almost surely,

$$\begin{aligned} -\nabla\cdot T(\boldsymbol{x},\omega)\nabla u(\boldsymbol{x},\omega) &= f(\boldsymbol{x}) && \text{in } D\times\Omega,\\ u(\boldsymbol{x},\omega) &= g(\boldsymbol{x}) && \text{on } \partial D_D\times\Omega,\\ \boldsymbol{n}\cdot T\nabla u(\boldsymbol{x},\omega) &= 0 && \text{on } \partial D_N\times\Omega. \end{aligned} \tag{2.1}$$

Formally, P is the probability measure associated with a probability space $(\Omega, \mathscr{F}, P)$. To solve (2.1), we first approximate $T(\boldsymbol{x}, \omega)$ by a function $T_M(\boldsymbol{x}, \boldsymbol{\xi})$ of an appropriate set of M independent random parameters $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_M]^\top$ taking values in $\Gamma \subseteq \mathbb{R}^M$ (e.g. see [9]). Crucially, this converts the stochastic PDE (2.1) to an $(M + d)$-dimensional deterministic one and conventional discretization schemes can be applied. We then seek a random field $u(\boldsymbol{x}, \boldsymbol{\xi})$ that solves, with probability one,

$$\begin{aligned} -\nabla\cdot T_M(\boldsymbol{x},\boldsymbol{\xi})\nabla u(\boldsymbol{x},\boldsymbol{\xi}) &= f(\boldsymbol{x}) && \text{in } D\times\Gamma,\\ u(\boldsymbol{x},\boldsymbol{\xi}) &= g(\boldsymbol{x}) && \text{on } \partial D_D\times\Gamma,\\ \boldsymbol{n}\cdot T\nabla u(\boldsymbol{x},\boldsymbol{\xi}) &= 0 && \text{on } \partial D_N\times\Gamma. \end{aligned} \tag{2.2}$$

The well-posedness of (2.1)–(2.2), primal variational formulations, and approximation schemes based on finite element spatial discretizations have been widely studied (see [9], [4], [5], [3], [13], [23], [35]). Stochastic Galerkin approximation, specifically, has been studied in [4] and [9] and solvers for the resulting symmetric positive definite linear systems have been studied in [27], [32] and [30].

2.1. Mixed formulation. We are concerned with the more challenging problem of finding a *pair* of random fields $\boldsymbol{q} = \boldsymbol{q}(\boldsymbol{x}, \omega)$ and $u = u(\boldsymbol{x}, \omega)$ such that, P-almost surely,

$$\begin{aligned} T^{-1}(\boldsymbol{x},\omega)\boldsymbol{q}(\boldsymbol{x},\omega) + \nabla u(\boldsymbol{x},\omega) &= 0\\ \nabla\cdot\boldsymbol{q}(\boldsymbol{x},\omega) &= f(\boldsymbol{x}) && \text{in } D\times\Omega,\\ u(\boldsymbol{x},\omega) &= g(\boldsymbol{x}) && \text{on } \partial D_D\times\Omega,\\ \boldsymbol{n}\cdot\boldsymbol{q}(\boldsymbol{x},\omega) &= 0 && \text{on } \partial D_N\times\Omega. \end{aligned} \tag{2.3}$$

We assume that $D \subset \mathbb{R}^2$ is a convex bounded open set and $0 < T_{min} \le T(\boldsymbol{x}, \omega) \le T_{max}$ a.e. in $D \times \Omega$. Given an approximation $T_M^{-1}(\boldsymbol{x}, \boldsymbol{\xi}) : D \times \Gamma \to \mathbb{R}$ to T^{-1} in terms of a finite set of M independent random parameters we solve a corresponding $(M + 2)$-dimensional boundary value problem

$$\begin{aligned} T_M^{-1}(\boldsymbol{x},\boldsymbol{\xi})\boldsymbol{q}(\boldsymbol{x},\boldsymbol{\xi}) + \nabla u(\boldsymbol{x},\boldsymbol{\xi}) &= 0\\ \nabla\cdot\boldsymbol{q}(\boldsymbol{x},\boldsymbol{\xi}) &= f(\boldsymbol{x}) && \text{in } D\times\Gamma,\\ u(\boldsymbol{x},\boldsymbol{\xi}) &= g(\boldsymbol{x}) && \text{on } \partial D_D\times\Gamma,\\ \boldsymbol{n}\cdot\boldsymbol{q}(\boldsymbol{x},\boldsymbol{\xi}) &= 0 && \text{on } \partial D_N\times\Gamma. \end{aligned} \tag{2.4}$$

The associated weak problem of finding $\boldsymbol{q}(\boldsymbol{x}, \boldsymbol{\xi}) \in V$ and $u(\boldsymbol{x}, \boldsymbol{\xi}) \in W$ satisfying

$$\langle a(\boldsymbol{q}, \boldsymbol{r})\rangle + \langle b(\boldsymbol{r}, u)\rangle = \langle -(g, \boldsymbol{n}\cdot\boldsymbol{r})_{\partial D_D}\rangle \qquad \forall\, \boldsymbol{r} \in V, \tag{2.5}$$

$$\langle b(\boldsymbol{q}, v)\rangle = \langle -(f, v)\rangle \qquad \forall\, v \in W, \tag{2.6}$$

where $a(\cdot,\cdot) : H(div; D) \times H(div; D) \to \mathbb{R}$ and $b(\cdot,\cdot) : H(div; D) \times L^2(D) \to \mathbb{R}$ are defined via

$$a(\boldsymbol{u}, \boldsymbol{v}) = \int_D T_M^{-1}\boldsymbol{u}\cdot\boldsymbol{v}\, d\boldsymbol{x}, \qquad b(\boldsymbol{v}, p) = -\int_D p\,\nabla\cdot\boldsymbol{v}\, d\boldsymbol{x},$$

is well-posed in $V = L^2_\rho(\Gamma, H_0(div; D))$ and $W = L^2_\rho(\Gamma, L^2(D))$. Here $\langle\cdot\rangle = \int_\Gamma \rho(\boldsymbol{\xi}) \cdot d\boldsymbol{\xi}$ denotes the expectation operator and $\rho(\boldsymbol{\xi})$ is the joint probability density function of $\boldsymbol{\xi}$, which is known once a distribution has been chosen for the independent parameters ξ_m. For any Hilbert space X with norm $\|\cdot\|_X$, $L^2_\rho(\Gamma, X) = \left\{v : \Gamma \to X \mid \left\langle \| v \|^2_X \right\rangle < \infty\right\}$.

Introducing finite-dimensional spaces $V_h \subset H_0(div; D), W_h \subset L^2(D), S_d \subset L^2_\rho(\Gamma)$ leads to the stochastic Galerkin saddle point problem: find $\boldsymbol{q}_{hd}(\boldsymbol{x}, \boldsymbol{\xi}) \in V_h \otimes S_d$ and $u_{hd}(\boldsymbol{x}, \boldsymbol{\xi}) \in W_h \otimes S_d$ satisfying

$$\langle a(\boldsymbol{q}_{hd}, \boldsymbol{r}_{hd})\rangle + \langle b(\boldsymbol{r}_{hd}, u_{hd})\rangle = \langle -(g, \boldsymbol{n} \cdot \boldsymbol{r}_{hd})_{\partial D_D}\rangle \qquad \forall\, \boldsymbol{r}_{hd} \in V_h \otimes S_d \tag{2.7}$$

$$\langle b(\boldsymbol{q}_{hd}, v_{hd})\rangle = \langle -(f, v_{hd})\rangle \qquad \forall\, v_{hd} \in W_h \otimes S_d. \tag{2.8}$$

If we introduce bases $V_h = \text{span}\{\boldsymbol{\varphi}_j(\boldsymbol{x})\}_{j=1}^{N_q}$, $W_h = \text{span}\{\phi_j(\boldsymbol{x})\}_{j=1}^{N_u}$, and $S_d = \text{span}\{\psi_\ell(\boldsymbol{\xi})\}_{\ell=1}^{N_{\boldsymbol{\xi}}}$ then

$$\boldsymbol{q}_{hd}(\boldsymbol{x}, \boldsymbol{\xi}) = \sum_{\ell=1}^{N_{\boldsymbol{\xi}}} \sum_{j=1}^{N_q} q_{j,\ell}\, \boldsymbol{\varphi}_j(\boldsymbol{x})\psi_\ell(\boldsymbol{\xi}), \qquad u_{hd}(\boldsymbol{x}, \boldsymbol{\xi}) = \sum_{\ell=1}^{N_{\boldsymbol{\xi}}} \sum_{j=1}^{N_u} u_{j,\ell}\, \phi_j(\boldsymbol{x})\psi_\ell(\boldsymbol{\xi}). \tag{2.9}$$

We will use $V_h = RT_0(D)$ and $W_h = P_0(D)$, the lowest-order Raviart-Thomas approximation [29] based on a partition of the physical domain D into triangles. For S_d we will use M-variate polynomials of total degree d in $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_M]^\top$. For these specific choices, the well-posedness of (2.7)–(2.8) was established in [15] under the assumption that

$$0 < T_{M,min} \le T_M(\boldsymbol{x}, \omega) \le T_{M,max} \qquad \text{a.e. in } D \times \Gamma. \tag{2.10}$$

However, *any* deterministic inf-sup stable pair V_h-W_h used in conjunction with *any* finite-dimensional subspace of $L^2_\rho(\Gamma)$ leads to an inf-sup stable approximation in a certain pair of norms; see Section 5.

2.2. Lognormal diffusion coefficient. The question remains as to what is a suitable approximation $T_M^{-1}(\boldsymbol{x}, \boldsymbol{\xi})$? The boundary value problem (2.3) provides a model for groundwater flow. In that scenario, u and $\boldsymbol{q}$ denote the pressure and velocity field, respectively, and T is the permeability coefficient. In [11] and [15], T^{-1} is approximated directly by an M-term Karhunen-Loève (KL) expansion [21] which is a linear function of M uncorrelated random variables ξ_m. In flow models, however, T often follows a lognormal distribution (e.g. see [14]) and cannot be approximated well using a linear combination of random parameters.

Here, we assume $T = \exp(G)$ where $G = G(\boldsymbol{x}, \omega)$ is a correlated Gaussian random field, with given mean $\langle G(\boldsymbol{x}, \omega)\rangle = \mu_G(\boldsymbol{x})$ and covariance function

$$C_G(\boldsymbol{x}, \boldsymbol{y}) = \langle (G(\boldsymbol{x}, \omega) - \mu_G(\boldsymbol{x}))\ (G(\boldsymbol{y}, \omega) - \mu_G(\boldsymbol{y}))\rangle = \sigma_G^2 V_G(\boldsymbol{x}, \boldsymbol{y}). \tag{2.11}$$

Our starting point is a Karhunen-Loève expansion [21] for G,

$$G(\boldsymbol{x}, \omega) = \mu_G + \sigma_G \sum_{m=1}^{\infty} \sqrt{\lambda_m} k_m(\boldsymbol{x}) \xi_m(\omega), \tag{2.12}$$

where μ_G and σ_G are the (spatially) constant mean and standard deviation of G, $(\lambda_m, k_m)_{m=1}^\infty$ are the eigenpairs of the integral operator associated with $V_G(\boldsymbol{x}, \boldsymbol{y})$ in (2.11) and $(\xi_1, \xi_2, \ldots, \xi_m, \ldots)$ are uncorrelated, independent standard Gaussian random variables. The eigenvalues are assumed to be listed in descending order so that $(\lambda_m, k_m)_{m=1}^M$ denote the eigenpairs corresponding to the M largest eigenvalues. Truncating after M terms gives $G_M(\boldsymbol{x}, \boldsymbol{\xi}) : D \times \Gamma \to \mathbb{R}$ where $\boldsymbol{\xi} := [\xi_1, \xi_2, \ldots, \xi_M]^\top$ and $\Gamma = \mathbb{R}^M$. Finally, we choose $T_M^{-1}(\boldsymbol{x}, \boldsymbol{\xi})$ as the truncated Wiener polynomial chaos expansion [19] of T^{-1} containing only polynomials in $\xi_1, \xi_2, \ldots, \xi_M$ which leads to (2.4).

DEFINITION 2.1 (M-dimensional multi-indices). *An M-dimensional multi-index $\boldsymbol{\alpha} \in \mathbb{N}_0^M$ is a sequence of non-negative integers $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_M)$. We define $|\boldsymbol{\alpha}| := \sum_{m=1}^M \alpha_m$, $\boldsymbol{\alpha}! := \prod_{m=1}^M \alpha_m!$, and write $\mathscr{I} = \mathbb{N}_0^M$. For a specified $M \in \mathbb{N}$ and $d \in \mathbb{N}_0$, denote*

$$\mathscr{I}_d := \{\boldsymbol{\alpha} \in \mathscr{I},\ |\boldsymbol{\alpha}| \le d\}, \qquad \mathscr{I}_d^+ := \{\boldsymbol{\alpha} \in \mathscr{I}_d,\ |\boldsymbol{\alpha}| > 0\}. \tag{2.13}$$

There is a bijection $\boldsymbol{\iota} : \{1, \ldots, J\} \mapsto \mathscr{I}_d$, $J = |\mathscr{I}_d| = \binom{M+d}{M}$, *that assigns a unique integer* $j \in \{1, \ldots, J\}$ *to each multi-index* $\boldsymbol{\iota}(j) \in \mathscr{I}_d$ *and vice versa.*

Consider, now, the set of multivariate polynomials

$$\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) := \prod_{m=1}^{M} \psi_{\alpha_m}^{(m)}(\xi_m), \qquad \boldsymbol{\alpha} \in \mathscr{I}, \tag{2.14}$$

where $\psi_k^{(m)}$ denotes the univariate Hermite polynomial of exact degree $k \in \mathbb{N}_0$. Such polynomials are orthonormal with respect to the Gaussian probability density function (pdf)

$$\rho_m(\xi_m) = (\sqrt{2\pi})^{-1} \exp(-\xi_m^2/2), \qquad m \in \mathbb{N}. \tag{2.15}$$

Hence $\langle \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) \psi_{\boldsymbol{\beta}}(\boldsymbol{\xi}) \rangle = \delta_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ and the polynomials in (2.14) are also orthonormal. Collectively, they form an M-dimensional polynomial chaos and provide a useful basis for $L^2_\rho(\Gamma)$. Assuming $T_M^{-1} \in L^2_\rho(\Gamma)$ for any $\boldsymbol{x} \in D$, we can then write

$$T_M^{-1}(\boldsymbol{x}, \boldsymbol{\xi}) = \sum_{\boldsymbol{\alpha} \in \mathscr{I}} t_{\boldsymbol{\alpha}}(\boldsymbol{x}) \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}), \qquad t_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \left\langle T_M^{-1}(\boldsymbol{x}, \boldsymbol{\xi}) \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) \right\rangle. \tag{2.16}$$

We can also use a subset of the polynomials in (2.14) to provide a basis for S_d. We choose

$$S_d := \operatorname{span}\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) : \boldsymbol{\alpha} \in \mathscr{I}_d\},$$

which has dimension $N_{\boldsymbol{\xi}} = \binom{M+d}{d} = \frac{(M+d)!}{M!d!}$.

Noting that $T^{-1}(\boldsymbol{x}, \boldsymbol{\xi}) = \exp(-G(\boldsymbol{x}, \boldsymbol{\xi}))$ and exploiting the orthogonality properties of the polynomial chaos functions yields explicit formulae for the spatial coefficient functions in (2.16) in terms of the known KL expansion functions for G from (2.12). From [22, Chapter I, Theorem 3.1]) it follows

$$t_0(\boldsymbol{x}) = \left\langle T^{-1} \right\rangle = \exp(-\mu_G + \sigma_G^2/2), \qquad \boldsymbol{\alpha} \in \mathscr{I},\ |\boldsymbol{\alpha}| = 0, \tag{2.17}$$

$$t_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \left\langle T^{-1} \right\rangle \frac{(-1)^{|\boldsymbol{\alpha}|} \sigma_G^{|\boldsymbol{\alpha}|}}{\sqrt{\alpha!}} \prod_{m=1}^{M} \left(\sqrt{\lambda_m} k_m(\boldsymbol{x}) \right)^{\alpha_m}, \quad \boldsymbol{\alpha} \in \mathscr{I},\ |\boldsymbol{\alpha}| > 0. \tag{2.18}$$

3. Stochastic Galerkin equations. Using the expansion (2.16), the SG mixed finite element approximation (2.7)–(2.8) leads to the set of Galerkin equations,

$$\begin{bmatrix} \widehat{A} & \widehat{B}^\top \\ \widehat{B} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{q} \\ \boldsymbol{u} \end{bmatrix} = \begin{bmatrix} \boldsymbol{g} \\ \boldsymbol{f} \end{bmatrix} \tag{3.1}$$

where $\boldsymbol{q}$ and $\boldsymbol{u}$ contain the coefficients in the expansions of u_{hd} and $\boldsymbol{q}_{hd}$ in (2.9) and

$$\widehat{A} = G_0 \otimes A_0 + \sum_{\boldsymbol{\alpha} \in \mathscr{I}_{2d}^+} G_{\boldsymbol{\alpha}} \otimes A_{\boldsymbol{\alpha}}, \qquad \widehat{B} = G_0 \otimes B. \tag{3.2}$$

We have separated out the symmetric, positive definite term $G_0 \otimes A_0$ which, together with $\widehat{B}$, represents the discretized *mean problem*, i.e., the deterministic problem obtained by replacing the random field T^{-1} with $\left\langle T^{-1} \right\rangle$. The right Kronecker product factors in (3.2) are finite element matrices that can be produced using standard finite element code. We have

$$A_0 \in \mathbb{R}^{N_q \times N_q}, \qquad [A_0]_{i,k} = (t_0 \boldsymbol{\varphi}_k, \boldsymbol{\varphi}_i), \qquad i,k = 1, \ldots, N_{\boldsymbol{q}}, \tag{3.3a}$$

$$A_{\boldsymbol{\alpha}} \in \mathbb{R}^{N_q \times N_q}, \qquad [A_{\boldsymbol{\alpha}}]_{i,k} = (t_{\boldsymbol{\alpha}} \boldsymbol{\varphi}_k, \boldsymbol{\varphi}_i), \qquad i,k = 1, \ldots, N_{\boldsymbol{q}},\ \boldsymbol{\alpha} \in \mathscr{I}_{2d}^+, \tag{3.3b}$$

$$B \in \mathbb{R}^{N_u \times N_q}, \qquad [B]_{i,k} = -(\nabla \cdot \boldsymbol{\varphi}_i, \phi_k), \qquad i = 1, \ldots, N_u,\ k = 1, \ldots, N_{\boldsymbol{q}}, \tag{3.3c}$$

where $(\cdot,\cdot)$ denotes the $L^2(D)$ inner-product. A_0 and $A_{\boldsymbol{\alpha}}$ have the structure of mass matrices, each one weighted by a different coefficient function from (2.18), and B is the matrix representation of the divergence operator. Although A_0 is positive definite (and so is $\widehat{A}$ if (2.10) holds), the coefficient functions $t_{\boldsymbol{\alpha}}$ are not strictly positive functions and so the matrices $A_{\boldsymbol{\alpha}}$ are in general indefinite. The left Kronecker product factors in (3.4) are defined in terms of the basis functions for S_d,

$$G_0 \in \mathbb{R}^{N_{\boldsymbol{\xi}}\times N_{\boldsymbol{\xi}}}, \qquad [G_0]_{j,\ell} = \left\langle \psi_{\boldsymbol{\iota}(\ell)}\psi_{\boldsymbol{\iota}(j)} \right\rangle, \qquad j,\ell = 1,\ldots,N_{\boldsymbol{\xi}}, \tag{3.4a}$$

$$G_{\boldsymbol{\alpha}} \in \mathbb{R}^{N_{\boldsymbol{\xi}}\times N_{\boldsymbol{\xi}}}, \qquad [G_{\boldsymbol{\alpha}}]_{j,\ell} = \left\langle \psi_{\boldsymbol{\alpha}}\psi_{\boldsymbol{\iota}(\ell)}\psi_{\boldsymbol{\iota}(j)} \right\rangle, \qquad j,\ell = 1,\ldots,N_{\boldsymbol{\xi}},\ \boldsymbol{\alpha} \in \mathscr{I}_{2d}^{+}. \tag{3.4b}$$

There is actually a matrix $G_{\boldsymbol{\alpha}}$ in (3.2) for each polynomial $\psi_{\boldsymbol{\alpha}}$ in the expansion (2.16). However, it can be shown (see [23]) that $G_{\boldsymbol{\alpha}}$ is the zero matrix for all multi-indices $\boldsymbol{\alpha} \in \mathscr{I}\backslash\mathscr{I}_{2d}$. The Galerkin projection onto S_d effectively truncates the infinite expansion of T_M^{-1} in (2.16) after a finite number of terms and the sum in (3.2) involves only multi-indices $\boldsymbol{\alpha} \in \mathscr{I}_{2d}$.

The spectral properties of the SG matrices $G_{\boldsymbol{\alpha}}$ in (3.4b) are key to determining an efficient solution strategy for our model problem. Unfortunately, explicit formulae for their eigenvalues remain elusive (except for low values of d, see [12]). It can be shown, however, that they are very ill-conditioned with respect to the discretization parameter d. In Corollary 4.6 we will show the spectral radius of each one can be bounded above by a quantity that is $O(\exp(Md)\exp(|\boldsymbol{\alpha}|/2))$.

3.1. Computational aspects. When (3.1) is solved iteratively, matrix-vector products with the Galerkin matrix dominate the cost of an iteration. Since we use orthonormal basis functions for S_d, G_0 in (3.4a) is the identity matrix and $\widehat{B}$ is block-diagonal. B is sparse and so matrix-vector products with $\widehat{B}$ can be performed in $O(N_{\boldsymbol{\xi}}N_q)$ operations. Performing multiplications with $\widehat{A}$, however, is challenging. A linear combination of the matrices $G_{\boldsymbol{\alpha}}$, with $\boldsymbol{\alpha} \in \mathscr{I}_{2d}$, is fully populated, since for every $j,\ell = 1,\ldots,N_{\boldsymbol{\xi}}$ there exists an $\boldsymbol{\alpha} \in \mathscr{I}_{2d}$, such that $[G_{\boldsymbol{\alpha}}]_{\boldsymbol{\iota}(j),\boldsymbol{\iota}(\ell)} \neq 0$, (cf. [23, Theorem 18]). Consequently, $\widehat{A}$ is block-dense. Each $A_{\boldsymbol{\alpha}}$ is sparse so matrix-vector products with the completely assembled $\widehat{A}$ can be performed in $O(N_{\boldsymbol{\xi}}^2 N_q)$ operations. However, storing $\widehat{A}$ rapidly consumes memory on desktop computers. The alternative is to only store the $|\mathscr{I}_{2d}|$ Kronecker factors $A_{\boldsymbol{\alpha}}$ and $G_{\boldsymbol{\alpha}}$ in (3.2). Writing $|\mathscr{I}_{2d}| = \binom{M+2d}{M} = N_{\boldsymbol{\xi}}N_p$, where $N_p := \prod_{k=d+1}^{2d}\frac{M+k}{k} \ll N_{\boldsymbol{\xi}}$, matrix-vector products with $\widehat{A}$ can then be performed in $O((N_{\boldsymbol{\xi}}N_q + N_{\boldsymbol{\xi}}^2 N_q)|\mathscr{I}_{2d}|) = O((N_{\boldsymbol{\xi}}^2 + N_{\boldsymbol{\xi}}^3)N_q N_p)$ operations.

In short, the saddle point matrix in (3.1) has the structure displayed in (1.4); the (1,2) block is block-diagonal and the (1,1)-block is block-dense with $N+1 = |\mathscr{I}_{2d}|$ terms. Fewer terms could be retained but this corresponds to a premature truncation of T_M^{-1} in (2.16). Such an approximation to T_M^{-1} does not necessarily satisfy a bound like (2.10).

3.2. Stochastically linear diffusion coefficient. The differences between the saddle point systems in (3.1) and those encountered in previous work stem from the choice of approximation in (2.16). Suppose, as was assumed in [11] and [15], that we had started from a *linear* KL expansion

$$T_M^{-1}(\boldsymbol{x},\boldsymbol{\xi}) = t_0 + \sigma\sum_{n=1}^{M}\sqrt{\lambda_n}k_n(\boldsymbol{x})\xi_n(\omega). \tag{3.5}$$

Then, instead of (3.2) we obtain,

$$\widehat{A} = G_0 \otimes A_0 + \sum_{n=1}^{M} G_n \otimes A_n, \qquad \widehat{B} = G_0 \otimes B, \tag{3.6}$$

where, with $t_n(\boldsymbol{x}) = \sigma\sqrt{\lambda_n}k_n(\boldsymbol{x})$,

$$G_n \in \mathbb{R}^{N_{\boldsymbol{\xi}}\times N_{\boldsymbol{\xi}}}, \qquad [G_n]_{j,\ell} = \left\langle \xi_n\psi_{\boldsymbol{\iota}(\ell)}\psi_{\boldsymbol{\iota}(j)} \right\rangle, \qquad j,\ell = 1,\ldots,N_{\boldsymbol{\xi}},\ n = 1,\ldots M, \tag{3.7a}$$

$$A_n \in \mathbb{R}^{N_{\boldsymbol{q}}\times N_{\boldsymbol{q}}}, \qquad [A_n]_{i,k} = (t_n(\boldsymbol{x})\boldsymbol{\varphi}_k, \boldsymbol{\varphi}_i), \qquad i,k = 1,\ldots,N_{\boldsymbol{q}},\ n = 1,\ldots M. \tag{3.7b}$$

The eigenvalues of the matrices G_n in (3.7a) are known explicitly (see [27], [12]). For Gaussian random variables, the condition number of each G_n grows, at worst, like $O(\sqrt{d})$. Each G_n is also *sparse* with only two non-zero entries per row. $\widehat{A}$ has at most $2M+1$ non-zero blocks per row and if $M \ll N_{\boldsymbol{\xi}}$, matrix-vector products with $\widehat{A}$ can be performed in only $O(N_{\boldsymbol{\xi}} N_q)$ operations.

For (3.5) to satisfy (2.10), σ must be small relative to $t_0 = \langle T^{-1} \rangle$ and the term $G_0 \otimes A_0$ dominates in (3.6). The approximation $\widehat{A} \approx G_0 \otimes A_0$ was exploited in [11] to obtain a preconditioner of type (1.2). For stochastically linear problems, such mean-based approximations are effective within the regime of statistical parameters where the problem is well-posed. For our model problem, however, there are no restrictions on σ_G relative to t_0, and $\widehat{A}$ is very ill-conditioned when σ_G and d are large. Mean-based preconditioners are not effective. In [15], an augmented preconditioner (1.3) is proposed with $\widehat{W} = \widehat{N}$ where $\widehat{N}$ represents the natural norm on $W_h \otimes S_d$ (see Section 5). It is not based on a mean-based approximation and so is more robust. However, it is practical only if the G_n matrices are sparse and $N_{\boldsymbol{\xi}}$ and $|\mathscr{I}_{2d}^+|$ are small. For the stochastically nonlinear problem considered here, the number of Kronecker product pairs $|\mathscr{I}_{2d}^+|$ can be very large (see Table 3.1).

TABLE 3.1
Dimension of S_d and number of Kronecker product pairs in $\widehat{A}$ for varying M and d.

	$M=5$			$M=10$			$M=20$		
	$d=1$	$d=2$	$d=3$	$d=1$	$d=2$	$d=3$	$d=1$	$d=2$	$d=3$
$N_{\boldsymbol{\xi}} = \lvert\mathscr{I}_d\rvert$	6	21	56	11	66	286	21	231	1,771
$N+1 = \lvert\mathscr{I}_{2d}\rvert$	21	126	462	66	1,001	8,008	231	10,626	230,231

3.3. Stochastic Galerkin versus sampling methods. Approximations based on sampling (e.g. Monte Carlo, stochastic collocation methods [3]) lead to decoupled deterministic problems, the number of which usually exceeds $N_{\boldsymbol{\xi}}$. For stochastically linear problems, where optimal solvers exist, SG methods are preferred. If $T_M(\boldsymbol{x}, \boldsymbol{\xi})$ is a *nonlinear* function of $\boldsymbol{\xi}$ it is less clear whether SG methods are competitive, as robust solvers are lacking. $G_0 \otimes A_0$ is not a good approximation to $\widehat{A}$ due, in part, to a dramatic increase in the ill-conditioning of the G_n matrices. For SG finite element discretizations of (2.2), preconditioners have been suggested in [30] and [32]. SG methods can only be competitive for challenging PDEs, however, if robust preconditioners are found for the coupled linear systems of equations they yield. SG systems therefore warrant serious investigation before conclusions about the efficacy of an approximation scheme for a specific PDE can be made.

4. Schur complement preconditioners. Applying $\widehat{P}_S^{-1}$ in (1.2) requires expensive solves with $\widehat{A}$ and working with the exact Schur complement is infeasible. An obvious first step towards a practical preconditioner for (3.1) is to replace $\widehat{A}$ by a symmetric, positive definite and sparse (e.g. diagonal) matrix $\widehat{X}$, leading to the preconditioner

$$\widehat{P}_X := \begin{bmatrix} \widehat{X} & 0 \\ 0 & \widehat{B}\widehat{X}^{-1}\widehat{B}^\top \end{bmatrix}, \tag{4.1}$$

whose performance, according to the following result, depends only on the choice of $\widehat{X}$.

LEMMA 4.1. *Let $0 < \nu_1 \le \nu_2 \le \cdots \le \nu_n$, where $n = N_{\boldsymbol{q}} N_{\boldsymbol{\xi}}$, denote the eigenvalues of $\widehat{X}^{-1}\widehat{A}$. The eigenvalues of $\widehat{P}_X^{-1}\widehat{C}$ lie in the union of the intervals*

$$\left[\frac{1}{2}(\nu_1 - \sqrt{\nu_1^2+4}), \frac{1}{2}(\nu_n - \sqrt{\nu_n^2+4})\right] \cup \left[\nu_1, \frac{1}{2}(\nu_n + \sqrt{\nu_n^2+4})\right]. \tag{4.2}$$

Proof. See, for example, [28, Corollary 3.3]. □

Lemma 4.1 hints that the eigenvalues of $\widehat{X}^{-1}\widehat{A}$ should be tightly clustered. On the other hand, linear systems with coefficient matrix $\widehat{X}$ must be solvable with much less effort than those with $\widehat{A}$. Satisfying both requirements is a tall order. We focus first on the latter, and consider approximations to $\widehat{A}$ of the form $\widehat{X} = L \otimes R$ where $L \in \mathbb{R}^{N_\xi \times N_\xi}$ and $R \in \mathbb{R}^{N_q \times N_q}$ are symmetric and positive definite. This respects the structure of $\widehat{A}$ in (3.1) and allows for efficient implementations as $\widehat{X}^{-1} = L^{-1} \otimes R^{-1}$. Motivated by [11] (see also Section 4.2) we investigate $\widehat{X} = L \otimes D_0$ where $R = D_0 := \operatorname{diag}(A_0)$.

4.1. Spectral bounds for $\widehat{X}^{-1}\widehat{A}$. According to Lemma 4.1, the number of preconditioned MINRES iterations required to solve (3.1) with $\widehat{X} = L \otimes D_0$ in (4.1), depends on the eigenvalues of

$$\widehat{X}^{-1}\widehat{A} = L^{-1} \otimes D_0^{-1}A_0 + \sum_{\boldsymbol{\alpha} \in \mathscr{I}_{2d}^+} L^{-1}G_{\boldsymbol{\alpha}} \otimes D_0^{-1}A_{\boldsymbol{\alpha}}, \tag{4.3}$$

which we now investigate.

LEMMA 4.2. *Let $D_0 = \operatorname{diag}(A_0)$ and define $A_{\boldsymbol{\alpha}}$, $\boldsymbol{\alpha} \in \mathscr{I}_{2d}^+$, as in* (3.3b)*. If $V_h = RT_0(D)$ is based on uniform meshes of right-angled triangles (for example), then (a) the eigenvalues of $D_0^{-1}A_0$ lie in the interval $\left[\frac{1}{2}, \frac{3}{2}\right]$ and (b) the eigenvalues of $D_0^{-1}A_{\boldsymbol{\alpha}}$ lie in the bounded interval $[-\frac{3}{2}c_{\boldsymbol{\alpha}}, \frac{3}{2}c_{\boldsymbol{\alpha}}]$, where*

$$c_{\boldsymbol{\alpha}} := \max_{\boldsymbol{x} \in D} \left| t_{\boldsymbol{\alpha}}(\boldsymbol{x}) t_0^{-1} \right| = \frac{\sigma_G^{|\boldsymbol{\alpha}|}}{\sqrt{\boldsymbol{\alpha}!}} \prod_{m=1}^{M} \left(\sqrt{\lambda_m} \|k_m\|_{L^\infty(D)} \right)^{\alpha_m}, \ \boldsymbol{\alpha} \in \mathscr{I}_{2d}^+. \tag{4.4}$$

Proof. Assertion (a) is shown the proof of [11, Lemma 4.3]. For any $\boldsymbol{q} \in \mathbb{R}^{N_q} \backslash \{\boldsymbol{0}\}$, we may define an $\boldsymbol{r} \in V_h$ by $\boldsymbol{r}(\boldsymbol{x}) = \sum q_i \boldsymbol{\varphi}_i(\boldsymbol{x})$, where $V_h = \operatorname{span}\{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_{N_q}\}$. Then, using (2.18), we obtain

$$\left| \boldsymbol{q}^\top A_{\boldsymbol{\alpha}} \boldsymbol{q} \right| = \left| \frac{\sigma_G^{|\boldsymbol{\alpha}|}}{\sqrt{\boldsymbol{\alpha}!}} \int_D \prod_{m=1}^{M} \left(\sqrt{\lambda_m} k_m(\boldsymbol{x}) \right)^{\alpha_m} \left\langle T^{-1} \right\rangle \boldsymbol{r} \cdot \boldsymbol{r} \, d\boldsymbol{x} \right| \leq c_{\boldsymbol{\alpha}} \boldsymbol{q}^\top A_0 \boldsymbol{q}.$$

For any $\boldsymbol{q} \in \mathbb{R}^{N_q} \backslash \{\boldsymbol{0}\}$, $-c_{\boldsymbol{\alpha}} \leq \frac{\boldsymbol{q}^\top A_{\boldsymbol{\alpha}} \boldsymbol{q}}{\boldsymbol{q}^\top A_0 \boldsymbol{q}} \leq c_{\boldsymbol{\alpha}}$ which, in combination with (a), gives the result. ∎

Decay rates of the KL eigenvalues λ_m and pointwise bounds on the eigenfunctions k_m, are derived in [31]. Using these results, we may obtain an upper bound for the constant $c_{\boldsymbol{\alpha}}$ in (4.4).

COROLLARY 4.3. *For a bounded domain $D \subset \mathbb{R}^2$, let the covariance function $C_G \in L^2(D \times D)$ in* (2.11) *be piecewise analytic in the sense of [31, Definition 2.15]. Then, we have the upper bound*

$$c_{\boldsymbol{\alpha}} \leq \kappa_1^{|\boldsymbol{\alpha}|} e^{-\kappa_2 |\boldsymbol{\alpha}|} \frac{\sigma_G^{|\boldsymbol{\alpha}|}}{\sqrt{\boldsymbol{\alpha}!}}, \qquad \boldsymbol{\alpha} \in \mathscr{I}_{2d}^+, \tag{4.5}$$

where $\kappa_1, \kappa_2 > 0$ are constants independent of M, d and $\boldsymbol{\alpha}$. If $C_G \in L^2(D \times D)$ is piecewise smooth in the sense of [31, Definition 2.15], there exists a constant $\kappa > 0$ independent of M, d and $\boldsymbol{\alpha}$, with

$$c_{\boldsymbol{\alpha}} \leq \kappa^{|\boldsymbol{\alpha}|} \frac{\sigma_G^{|\boldsymbol{\alpha}|}}{\sqrt{\boldsymbol{\alpha}!}}, \qquad \boldsymbol{\alpha} \in \mathscr{I}_{2d}^+. \tag{4.6}$$

Proof. If C_G is piecewise analytic, combining an estimate for λ_m in [31, Proposition 2.18] and an upper bound for $\|k_m\|_{L^\infty(D)}$ in terms of $|\lambda_m|$ from [31, Theorem 2.24], it can be shown that

$$\sqrt{\lambda_m} \|k_m\|_{L^\infty(D)} \leq \kappa_1 \exp(-\kappa_2 \sqrt{m}),$$

for all $m \geq 1$ with constants $\kappa_1, \kappa_2 > 0$ independent of m. Thus,

$$\prod_{m=1}^{M} (\sqrt{\lambda_m} \|k_m\|_{L^\infty(D)})^{\alpha_m} \leq \prod_{m=1}^{M} (\kappa_1 \exp(-\kappa_2 \sqrt{m}))^{\alpha_m} \leq \prod_{m=1}^{M} (\kappa_1 \exp(-\kappa_2))^{\alpha_m} = \kappa_1^{|\boldsymbol{\alpha}|} \exp\left(-\kappa_2 |\boldsymbol{\alpha}|\right),$$

yielding the upper bound for $c_{\boldsymbol{\alpha}}$ in (4.5). Analogously, for piecewise smooth covariance kernels, from [31, Corollary 2.22] and [31, Theorem 2.24] it follows that there exists a constant $\kappa > 0$ independent of m, such that, for all $m \geq 1$, $\sqrt{\lambda_m}\|k_m\|_{L^\infty(D)} \leq \kappa$, and this completes the proof. ∎

Lemma 4.2 tells us that the eigenvalues of $D_0^{-1}A_{\boldsymbol{\alpha}}$ are bounded independently of the discretization parameter h. However, since $|\boldsymbol{\alpha}| \leq 2d$, Corollary 4.3 suggests that those eigenvalues depend on d. To analyze the eigenvalues of $\widehat{X}^{-1}\widehat{A}$, we also need to study the matrices $G_{\boldsymbol{\alpha}}$, $\mathrm{diag}(G_{\boldsymbol{\alpha}})$ and $L^{-1}G_{\boldsymbol{\alpha}}$. To this end, we recall, first, an eigenvalue bound from [12]. Denote by $(\eta_{m,i}, w_{m,i})_{i=1}^{\delta_m}$ the nodes and weights, respectively, of the δ_m-point Gaussian quadrature rule associated with ρ_m in (2.15). Then,

$$\left\langle \psi^{(m)} \right\rangle = \int_{\Gamma_m} \psi^{(m)}(\xi_m)\rho_m(\xi_m)\, d\xi_m \approx \sum_{i=1}^{\delta_m} \psi^{(m)}(\eta_{m,i})w_{m,i}, \qquad m = 1, \ldots, M.$$

Each quadrature rule on Γ_m is exact for univariate polynomials $\psi^{(m)} \in \mathrm{span}\{1, \xi_m, \ldots, \xi_m^{2\delta_m-1}\}$. We can then define a tensor product quadrature rule on $\Gamma = \Gamma_1 \times \Gamma_2 \cdots \times \Gamma_M$ using the grid,

$$\Xi_{\boldsymbol{\delta}} := \bigtimes_{m=1}^{M} \{\eta_{m,1}, \eta_{m,2}, \ldots, \eta_{m,\delta_m}\}, \tag{4.7}$$

which can be used to establish the following theoretical results.

Lemma 4.4. *([12, Corollary 13]) The eigenvalues of $G_{\boldsymbol{\alpha}}$ in* (3.4b) *lie in $[\theta_{\boldsymbol{\alpha}}, \Theta_{\boldsymbol{\alpha}}]$, where*

$$\theta_{\boldsymbol{\alpha}} := \min\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{\eta}) : \boldsymbol{\eta} \in \Xi_{\boldsymbol{\delta}}\}, \qquad \Theta_{\boldsymbol{\alpha}} := \max\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{\eta}) : \boldsymbol{\eta} \in \Xi_{\boldsymbol{\delta}}\}, \qquad \boldsymbol{\alpha} \in \mathscr{I}_{2d}^{+}, \tag{4.8}$$

$\boldsymbol{\delta} \in \mathbb{N}_0^M$ is a multi-index with $\delta_m := d + \left\lceil \frac{\alpha_m+1}{2} \right\rceil$, $m = 1, \ldots, M$ and $\Xi_{\boldsymbol{\delta}}$ is defined as in (4.7).

Lemma 4.5. *Let $G_{\boldsymbol{\alpha}}$, $\boldsymbol{\alpha} \in \mathscr{I}_{2d}^{+}$, be defined as in* (3.4b). *The eigenvalues of* $\mathrm{diag}(G_{\boldsymbol{\alpha}})$ *also lie in the interval $[\theta_{\boldsymbol{\alpha}}, \Theta_{\boldsymbol{\alpha}}]$, with $\theta_{\boldsymbol{\alpha}}$ and $\Theta_{\boldsymbol{\alpha}}$ defined in* (4.8).

Proof. Define $\boldsymbol{\delta} \in \mathbb{N}_0^M$ as in Lemma 4.4. The largest eigenvalue of $\mathrm{diag}(G_{\boldsymbol{\alpha}})$ satisfies:

$$\begin{aligned}\lambda_{max}(\mathrm{diag}(G_{\boldsymbol{\alpha}})) &= \max_{j=1,\ldots,N_{\boldsymbol{\xi}}} \left\langle \psi_{\boldsymbol{\alpha}}\psi_{\boldsymbol{\iota}(j)}^2 \right\rangle = \max_{j=1,\ldots,N_{\boldsymbol{\xi}}} \textstyle\prod_{m=1}^{M} \sum_{i=1}^{\delta_m} \psi_{\alpha_m}^{(m)}(\eta_{m,i})[\psi_{\boldsymbol{\iota}(j)_m}(\eta_{m,i})]^2 w_{m,i} \\ &\leq \max_{j=1,\ldots,N_{\boldsymbol{\xi}}} \Theta_{\boldsymbol{\alpha}} \textstyle\prod_{m=1}^{M} \sum_{i=1}^{\delta_m} [\psi_{\boldsymbol{\iota}(j)_m}(\eta_{m,i})]^2 w_{m,i} = \max_{j=1,\ldots,N_{\boldsymbol{\xi}}} \Theta_{\boldsymbol{\alpha}} \left\langle \psi_{\boldsymbol{\iota}(j)}^2 \right\rangle = \Theta_{\boldsymbol{\alpha}}.\end{aligned}$$

The lower bound for the smallest eigenvalue of $\mathrm{diag}(G_{\boldsymbol{\alpha}})$ follows analogously. ∎

We can now investigate the constants $\theta_{\boldsymbol{\alpha}}$ and $\Theta_{\boldsymbol{\alpha}}$ in (4.8) in more detail.

Corollary 4.6. *The eigenvalues of $G_{\boldsymbol{\alpha}}$ in* (3.4b) *lie in the interval $[-b_{\boldsymbol{\alpha}}, b_{\boldsymbol{\alpha}}]$ where*

$$b_{\boldsymbol{\alpha}} := \exp(M(2d+1)/2)\exp(|\boldsymbol{\alpha}|/2), \qquad \boldsymbol{\alpha} \in \mathscr{I}_{2d}^{+}. \tag{4.9}$$

Proof. Apply Lemma 4.4 with the polynomials $\psi_k^{(m)}(\xi_m) = H_k(\xi_m/\sqrt{2})/\sqrt{2^k k!}$ generated by the Gaussian pdf (2.15), where H_k is the Hermite polynomial of degree k. For $k > 1$, all roots of H_k lie in $(-\sqrt{2k-2}, \sqrt{2k-2})$, see [20, Theorem 4] and so all roots of $\psi_k^{(m)}$ are contained in $(-2\sqrt{k-1}, 2\sqrt{k-1})$. The upper bound $|\psi_k^{(m)}(\xi_m)| \leq \exp(\xi_m^2/4)$ follows from the fact that $|H_k(\xi)| \leq \sqrt{2^k k!}\exp(\xi^2/2)$, cf. [18]. Hence, with δ_m defined in Lemma 4.4, and $\Xi_{\boldsymbol{\delta}}$ as in (4.7),

$$\begin{aligned}|\lambda(G_{\boldsymbol{\alpha}})| &\leq \max\{|\psi_{\boldsymbol{\alpha}}(\boldsymbol{\eta})|, \boldsymbol{\eta} \in \Xi_{\boldsymbol{\delta}}\} \leq \prod_{m=1}^{M} |\psi_{\alpha_m}^{(m)}(2\sqrt{\delta_m - 1})| \leq \prod_{m=1}^{M} \exp(\delta_m - 1) \\ &\leq \prod_{m=1}^{M} \exp(d + (\alpha_m + 1)/2) = \exp(M(2d+1)/2)\exp(|\boldsymbol{\alpha}|/2).\end{aligned}$$

∎

Using these spectral inclusion bounds for the $G_{\boldsymbol{\alpha}}$ matrices we can now prove the following result.

THEOREM 4.7. *Let $\widehat{X} = L \otimes D_0$, where $D_0 = \mathrm{diag}(A_0)$, and $L \in \mathbb{R}^{N_{\boldsymbol{\xi}} \times N_{\boldsymbol{\xi}}}$ is any symmetric positive definite matrix whose eigenvalues lie in $[\ell_{min}, \ell_{max}]$ with $0 < \ell_{min} \leq \ell_{max}$. For each $\boldsymbol{\alpha} \in \mathscr{I}_{2d}^+$, define $c_{\boldsymbol{\alpha}}$ as in (4.4) and $b_{\boldsymbol{\alpha}}$ as in (4.9). Then, the eigenvalues of $\widehat{X}^{-1}\widehat{A}$ lie in the interval*

$$\left[\frac{1}{2\ell_{max}} - \frac{3\tau}{2\ell_{min}}, \frac{3(1+\tau)}{2\ell_{min}}\right], \qquad \tau := \sum_{\boldsymbol{\alpha} \in \mathscr{I}_{2d}^+} b_{\boldsymbol{\alpha}} c_{\boldsymbol{\alpha}}. \tag{4.10}$$

Proof. We can estimate the largest eigenvalue of $\widehat{X}^{-1}\widehat{A}$ as follows,

$$\begin{aligned}
\lambda_{max}(\widehat{X}^{-1}\widehat{A}) &= \max_{\boldsymbol{v} \in \mathbb{R}^{N_{\boldsymbol{\xi}} N_q} \setminus \{0\}} \frac{\boldsymbol{v}^\top \widehat{A} \boldsymbol{v}}{\boldsymbol{v}^\top \widehat{X} \boldsymbol{v}} = \max_{\boldsymbol{v} \in \mathbb{R}^{N_{\boldsymbol{\xi}} N_q} \setminus \{0\}} \frac{\boldsymbol{v}^\top \left(\sum_{\boldsymbol{\alpha} \in \mathscr{I}_{2d}} G_{\boldsymbol{\alpha}} \otimes A_{\boldsymbol{\alpha}}\right) \boldsymbol{v}}{\boldsymbol{v}^\top (L \otimes D_0) \boldsymbol{v}} \\
&\leq \sum_{\boldsymbol{\alpha} \in \mathscr{I}_{2d}} \max_{\boldsymbol{v} \in \mathbb{R}^{N_{\boldsymbol{\xi}} N_q} \setminus \{0\}} \frac{\boldsymbol{v}^\top (G_{\boldsymbol{\alpha}} \otimes A_{\boldsymbol{\alpha}}) \boldsymbol{v}}{\boldsymbol{v}^\top (L \otimes D_0) \boldsymbol{v}} = \sum_{\boldsymbol{\alpha} \in \mathscr{I}_{2d}} \lambda_{max}(L^{-1} G_{\boldsymbol{\alpha}} \otimes D_0^{-1} A_{\boldsymbol{\alpha}}) \\
&= \sum_{\boldsymbol{\alpha} \in \mathscr{I}_{2d}} \max_{i,j} \{\gamma_{\boldsymbol{\alpha}}^{(i)} \cdot \nu_{\boldsymbol{\alpha}}^{(j)}, \text{ where } \gamma_{\boldsymbol{\alpha}}^{(i)} \in \lambda(L^{-1} G_{\boldsymbol{\alpha}}) \text{ and } \nu_{\boldsymbol{\alpha}}^{(j)} \in \lambda(D_0^{-1} A_{\boldsymbol{\alpha}})\}.
\end{aligned}$$

The eigenvalues of L^{-1} lie in $[\ell_{max}^{-1}, \ell_{min}^{-1}]$ so part (a) of Lemma 4.2 tells us that the eigenvalues of $L^{-1}G_0 \otimes D_0^{-1}A_0$ lie in the interval $[\frac{1}{2}\ell_{max}^{-1}, \frac{3}{2}\ell_{min}^{-1}]$. Combining part (b) of Lemma 4.2 with the fact that the eigenvalues of $L^{-1}G_{\boldsymbol{\alpha}}$ lie in $[-\ell_{max}^{-1} b_{\boldsymbol{\alpha}}, \ell_{min}^{-1} b_{\boldsymbol{\alpha}}]$ results in spectral bounds for $L^{-1}G_{\boldsymbol{\alpha}} \otimes D_0^{-1}A_{\boldsymbol{\alpha}}$, for each $\boldsymbol{\alpha} \in \mathscr{I}_{2d}^+$. In particular, $\lambda_{max}(L^{-1}G_{\boldsymbol{\alpha}} \otimes D_0^{-1}A_{\boldsymbol{\alpha}}) \leq \frac{3}{2}\ell_{min}^{-1} b_{\boldsymbol{\alpha}} c_{\boldsymbol{\alpha}}$. Summing over $\boldsymbol{\alpha}$ yields an upper bound on $\lambda_{max}(\widehat{X}^{-1}\widehat{A})$. The lower bound for $\lambda_{min}(\widehat{X}^{-1}\widehat{A})$ follows analogously. ∎

Note that in (4.10), τ depends on d, σ_G and M. If C_G is a piecewise analytic covariance function we can combine the upper bound for $c_{\boldsymbol{\alpha}}$ in (4.5) with the definition of $b_{\boldsymbol{\alpha}}$ in (4.9) to obtain

$$\tau \leq \sum_{\boldsymbol{\alpha} \in \mathscr{I}_{2d}^+} e^{M(2d+1)/2} e^{|\boldsymbol{\alpha}|/2} \kappa_1^{|\boldsymbol{\alpha}|} e^{-\kappa_2 |\boldsymbol{\alpha}|} \frac{\sigma^{|\boldsymbol{\alpha}|}}{\sqrt{\boldsymbol{\alpha}!}} = e^{M(2d+1)/2} \sum_{\boldsymbol{\alpha} \in \mathscr{I}_{2d}^+} \frac{(\kappa_1 \sigma_G e^{1/2-\kappa_2})^{|\boldsymbol{\alpha}|}}{\sqrt{\boldsymbol{\alpha}!}}.$$

Similarly, if C_G is piecewise smooth, we obtain

$$\tau = \sum_{\boldsymbol{\alpha} \in \mathscr{I}_{2d}^+} b_{\boldsymbol{\alpha}} c_{\boldsymbol{\alpha}} \leq \sum_{\boldsymbol{\alpha} \in \mathscr{I}_{2d}^+} e^{M(2d+1)/2} e^{|\boldsymbol{\alpha}|/2} \kappa^{|\boldsymbol{\alpha}|} \frac{\sigma^{|\boldsymbol{\alpha}|}}{\sqrt{\boldsymbol{\alpha}!}} = e^{M(2d+1)/2} \sum_{\boldsymbol{\alpha} \in \mathscr{I}_{2d}^+} \frac{(\kappa \sigma_G \sqrt{e})^{|\boldsymbol{\alpha}|}}{\sqrt{\boldsymbol{\alpha}!}}.$$

Hence, the message from (4.10) is that a good choice of L is one that damps out ill-conditioning in $\widehat{X}^{-1}\widehat{A}$ with respect to d, σ_G and M. We now consider three suggestions for the matrix L.

4.2. Mean-based preconditioners. The preconditioning scheme from [11] corresponds to $L = I$ (the $N_{\boldsymbol{\xi}} \times N_{\boldsymbol{\xi}}$ identity matrix). Then, $\widehat{X} = I \otimes D_0 = \mathrm{diag}(G_0 \otimes A_0)$ and (4.1) is the sparse matrix

$$\widehat{P}_0 = \begin{bmatrix} I \otimes D_0 & 0 \\ 0 & I \otimes S_0 \end{bmatrix}, \tag{4.11}$$

where we have introduced $S_0 := B D_0^{-1} B^\top \in \mathbb{R}^{N_u \times N_u}$. Lemma 4.1 says that the efficiency of $\widehat{P}_0$ as a preconditioner, depends solely on the spectrum of

$$\widehat{X}^{-1}\widehat{A} = I \otimes D_0^{-1} A_0 + \sum_{\boldsymbol{\alpha} \in \mathscr{I}_{2d}^+} G_{\boldsymbol{\alpha}} \otimes D_0^{-1} A_{\boldsymbol{\alpha}}. \tag{4.12}$$

COROLLARY 4.8. *The eigenvalues of $\widehat{P}_0^{-1}\widehat{C}$ are contained in the union of the intervals in (4.2) with $\nu_1 \geq \frac{1}{2} - \frac{3}{2}\tau$ and $\nu_n \leq \frac{3}{2} + \frac{3}{2}\tau$, where τ is defined as in (4.10).*

Proof. The result follows by combining Lemma 4.1 and Theorem 4.7 with $L = I$ and $\ell_{min} = \ell_{max} = 1$. ∎

For our model problem, we shall see that $\widehat{X} = I \otimes D_0$ fails to be a robust approximation to $\widehat{A}$ with respect to d and σ_G, leading to unacceptably high MINRES iteration counts.

4.3. Kronecker product preconditioners. We can improve on (4.11) using a Kronecker product approximation, introduced for the primal problem (2.1) in [32]. Instead of approximating $\widehat{A}$ by the diagonal of a single term in (3.2), the idea is to choose $\widehat{X} = G \otimes D_0$, where

$$G = \operatorname{argmin}\{H \in \mathbb{R}^{N_{\boldsymbol{\xi}} \times N_{\boldsymbol{\xi}}} : ||\widehat{A} - H \otimes D_0||_F\}$$

and $\|\cdot\|_F$ denotes the Frobenius norm. The solution is available in closed form [33, Theorem 3],

$$G = I + \sum_{\boldsymbol{\alpha} \in \mathscr{I}_{2d}^+} \frac{\operatorname{tr}(A_{\boldsymbol{\alpha}}^\top D_0)}{\operatorname{tr}(D_0^\top D_0)} G_{\boldsymbol{\alpha}}. \tag{4.13}$$

Note that $\operatorname{tr}(A_{\boldsymbol{\alpha}}^\top D_0) = \sum_{i=1}^{N_q} [A_{\boldsymbol{\alpha}}]_{i,i} [D_0]_{i,i}$ and so the coefficients in (4.13) can be computed cheaply. Moreover, since $\widehat{A}$ and D_0 are symmetric and positive definite, so are G [33, Theorem 10] and $\widehat{X} = G \otimes D_0$. Since $\widehat{B}\widehat{X}^{-1}\widehat{B}^\top = G^{-1} \otimes S_0$, we arrive at the preconditioner

$$\widehat{P}_1 = \begin{bmatrix} G \otimes D_0 & 0 \\ 0 & G^{-1} \otimes S_0 \end{bmatrix}. \tag{4.14}$$

Applying Lemma 4.1, the efficiency of $\widehat{P}_1$ depends on the spectrum of

$$\widehat{X}^{-1}\widehat{A} = G^{-1} \otimes D_0^{-1} A_0 + \sum_{\boldsymbol{\alpha} \in \mathscr{I}_{2d}^+} G^{-1} G_{\boldsymbol{\alpha}} \otimes D_0^{-1} A_{\boldsymbol{\alpha}}. \tag{4.15}$$

Comparing (4.15) with (4.12), we see that some of the ill-conditioning in $G_{\boldsymbol{\alpha}}$ and $D_0^{-1}A_{\boldsymbol{\alpha}}$ can potentially be damped out by G in (4.14); the preconditioner (4.11) offers no such possibility. Since G in (4.13) is not diagonal in general, approximating the action of $\widehat{P}_1^{-1}$ is more costly than $\widehat{P}_0^{-1}$. Since $\operatorname{diag}(G)$ is positive definite, we can also consider the cheaper preconditioner

$$\widehat{P}_2 = \begin{bmatrix} \operatorname{diag}(G) \otimes D_0 & 0 \\ 0 & \operatorname{diag}(G)^{-1} \otimes S_0 \end{bmatrix}. \tag{4.16}$$

The (2,2) block of $\widehat{P}_2$ is then block-diagonal with $N_{\boldsymbol{\xi}}$ sparse blocks of size $N_u \times N_u$ as in (4.11).

COROLLARY 4.9. *Let G be defined as in* (4.13). *Let τ be defined as in* (4.10) *and suppose $\tau < 1$. If $\widehat{X} = G \otimes D_0$ or $\widehat{X} = \operatorname{diag}(G) \otimes D_0$ then the eigenvalues of $\widehat{P}_X^{-1}\widehat{C}$ are bounded and lie in the union of the intervals in* (4.2) *with $\nu_1 \geq \frac{1}{2(1+\tau)} - \frac{3\tau}{2(1-\tau)}$ and $\nu_n \leq \frac{3(1+\tau)}{2(1-\tau)}$.*

Proof. Apply Theorem 4.7 with $L = G$ and $L = \operatorname{diag}(G)$. If $\tau < 1$, we show that we can choose $\ell_{min} = 1 - \tau$ and $\ell_{max} = 1 + \tau$ in (4.10), which, in combination with Lemma 4.1, yields the assertion. Define $D_{\boldsymbol{\alpha}} := \operatorname{diag}(A_{\boldsymbol{\alpha}})$, $\boldsymbol{\alpha} \in \mathscr{I}_{2d}^+$. Then, in (4.13), we obtain

$$\frac{|\operatorname{tr}(A_{\boldsymbol{\alpha}}^\top D_0)|}{\operatorname{tr}(D_0^\top D_0)} = \frac{|\operatorname{tr}(D_{\boldsymbol{\alpha}}^\top D_0)|}{\operatorname{tr}(D_0^\top D_0)} \leq \frac{\|D_{\boldsymbol{\alpha}}\|_F \|D_0\|_F}{\|D_0\|_F^2} = \frac{\|D_{\boldsymbol{\alpha}}\|_F}{\|D_0\|_F}.$$

Furthermore, we estimate

$$\begin{aligned} \|D_{\boldsymbol{\alpha}}\|_F^2 &= \sum_{i=1}^{N_u} \left(\int_D \langle T^{-1} \rangle \frac{\sigma_G^{|\boldsymbol{\alpha}|}}{\sqrt{\boldsymbol{\alpha}!}} \prod_{m=1}^{M} (\sqrt{\lambda_m} k_m(\boldsymbol{x}))^{\alpha_m} \boldsymbol{\varphi}_i \cdot \boldsymbol{\varphi}_i \, d\boldsymbol{x} \right)^2 \\ &\leq \frac{\sigma_G^{2|\boldsymbol{\alpha}|}}{\boldsymbol{\alpha}!} \prod_{m=1}^{M} (\sqrt{\lambda_m} \|k_m\|_{L^\infty(D)})^{2\alpha_m} \sum_{i=1}^{N_u} \left(\int_D \langle T^{-1} \rangle \boldsymbol{\varphi}_i \cdot \boldsymbol{\varphi}_i \, d\boldsymbol{x} \right)^2 = c_{\boldsymbol{\alpha}}^2 \|D_0\|_F^2. \end{aligned}$$

Thus, $-c_{\boldsymbol{\alpha}} \leq \frac{\mathrm{tr}(A_{\boldsymbol{\alpha}}^\top D_0)}{\mathrm{tr}(D_0^\top D_0)} \leq c_{\boldsymbol{\alpha}}$, $\boldsymbol{\alpha} \in \mathscr{I}_{2d}^+$, and consequently,

$$\lambda_{max}(G) \leq 1 + \sum_{\boldsymbol{\alpha}\in\mathscr{I}_{2d}^+} c_{\boldsymbol{\alpha}} \lambda_{max}(G_{\boldsymbol{\alpha}}) \leq 1 + \sum_{\boldsymbol{\alpha}\in\mathscr{I}_{2d}^+} c_{\boldsymbol{\alpha}} b_{\boldsymbol{\alpha}} = 1 + \tau, \tag{4.17}$$

where $b_{\boldsymbol{\alpha}}$ is defined in (4.9). The bound $\lambda_{min}(G) \geq 1 - \tau$ follows analogously, giving the desired result for $L = G$. In Lemma 4.5 we established that for each $\boldsymbol{\alpha} \in \mathscr{I}_{2d}^+$ the spectral bounds for $G_{\boldsymbol{\alpha}}$ also hold for $\mathrm{diag}(G_{\boldsymbol{\alpha}})$. Hence, replacing $G_{\boldsymbol{\alpha}}$ by $\mathrm{diag}(G_{\boldsymbol{\alpha}})$ in (4.17) also yields the upper bound

$$\lambda_{max}(\mathrm{diag}(G)) \leq 1 + \sum_{\boldsymbol{\alpha}\in\mathscr{I}_{2d}^+} c_{\boldsymbol{\alpha}} \lambda_{max}(\mathrm{diag}(G_{\boldsymbol{\alpha}})) \leq 1 + \sum_{\boldsymbol{\alpha}\in\mathscr{I}_{2d}^+} c_{\boldsymbol{\alpha}} b_{\boldsymbol{\alpha}} \leq 1 + \tau,$$

and similarly, $\lambda_{min}(\mathrm{diag}(G)) \geq 1 - \tau$, which completes the proof for the case $L = \mathrm{diag}(G)$. ∎

The spectral inclusion bounds in Corollaries 4.8–4.9 are of course of limited value in practise. They do not provide information on the clustering of the eigenvalues and cannot be used to predict a priori, which preconditioner will perform best in terms of iteration counts. Indeed, we have derived the same bounds for both $\widehat{P}_1$ and $\widehat{P}_2$ in Corollary 4.9 but we shall see in Section 6 that the performance of these preconditioners, in terms of MINRES iterations, is by no means the same. The bounds do tell us, however, that iteration counts, for all the preconditioners, are likely to be affected by σ_G, d and M since those parameters influence τ. When $\tau < 1$, we can see that the bound in Corollary 4.8 for $\widehat{P}_0$ is better than the bound in Corollary 4.9 for the preconditioners $\widehat{P}_1$ and $\widehat{P}_2$. This fits with our intuition since for $\tau < 1$, σ_G and d have to be small and we'd expect the mean-based preconditioner $\widehat{P}_0$ to perform adequately in that case.

4.4. Practical Schur complement preconditioners. Computing the actions of $\widehat{P}_0^{-1}$, $\widehat{P}_1^{-1}$, and $\widehat{P}_2^{-1}$ involves solving $N_{\boldsymbol{\xi}}$ linear systems with the sparse coefficient matrix $S_0 = BD_0^{-1}B^\top$. Since S_0 is a discrete representation of the elliptic differential operator $\nabla \cdot \langle T \rangle \nabla$, those systems can be solved approximately in $O(N_u)$ operations, using a wide variety of standard multigrid methods. In Section 6 we apply, specifically, one V-cycle of a black-box algebraic multigrid method (AMG, see [7]). To analyze the impact of this extra approximation, let V_0 be an approximation to S_0 that satisfies

$$0 < \theta^2 \leq \frac{\boldsymbol{w}^\top S_0 \boldsymbol{w}}{\boldsymbol{w}^\top V_0 \boldsymbol{w}} \leq \Theta^2 \qquad \forall \boldsymbol{w} \in \mathbb{R}^{N_u} \backslash \{\boldsymbol{0}\}, \qquad \text{for some } \theta, \Theta \in \mathbb{R}^+. \tag{4.18}$$

LEMMA 4.10. *Let $\widehat{X} = L \otimes D_0$, where $L \in \mathbb{R}^{N_{\boldsymbol{\xi}} \times N_{\boldsymbol{\xi}}}$ is symmetric and positive definite. Let $0 < \nu_1 \leq \cdots \leq \nu_n$, $n = N_{\boldsymbol{q}} N_{\boldsymbol{\xi}}$, denote the eigenvalues of $\widehat{X}^{-1}\widehat{A}$. The eigenvalues of $\widehat{P}_{amg}^{-1}\widehat{C}$, where*

$$\widehat{P}_{amg} = \begin{bmatrix} L \otimes D_0 & 0 \\ 0 & L^{-1} \otimes V_0 \end{bmatrix}, \tag{4.19}$$

and $\widehat{C}$ denotes the Galerkin matrix in (3.1)*, lie in the union of the intervals*

$$\left[\frac{1}{2}(\nu_1 - \sqrt{\nu_1^2 + 4\Theta^2}), \frac{1}{2}(\nu_n - \sqrt{\nu_n^2 + 4\theta^2})\right] \cup \left[\nu_1, \frac{1}{2}(\nu_n + \sqrt{\nu_n^2 + 4\Theta^2})\right].$$

Proof. (4.19) is a preconditioner of the form (1.2) with $\widehat{X} = L \otimes D_0$ where $\widehat{B}\widehat{X}^{-1}\widehat{B}^T = L^{-1} \otimes S_0$ has been approximated by $L^{-1} \otimes V_0$. The result follows using [28, Corollary 3.4] by noting that the efficiency of that approximation only depends on the constants in (4.18) since

$$\frac{\boldsymbol{v}^T(\widehat{B}\widehat{X}^{-1}\widehat{B})\boldsymbol{v}}{\boldsymbol{v}^T\left(L^{-1} \otimes V_0\right)\boldsymbol{v}} = \frac{\boldsymbol{v}^T\left(L^{-1} \otimes S_0\right)\boldsymbol{v}}{\boldsymbol{v}^T\left(L^{-1} \otimes V_0\right)\boldsymbol{v}} = \frac{\boldsymbol{w}^T S_0 \boldsymbol{w}}{\boldsymbol{w}^T V_0 \boldsymbol{w}}$$

for any $\boldsymbol{v} = \boldsymbol{u} \otimes \boldsymbol{w} \in \mathbb{R}^{N_{\boldsymbol{\xi}} N_u} \backslash \{\boldsymbol{0}\}$ with $\boldsymbol{u} \in \mathbb{R}^{N_{\boldsymbol{\xi}}}$ and $\boldsymbol{w} \in \mathbb{R}^{Nu}$. ∎

Applying the (1,1) block of $\widehat{P}_{amg}$ in each MINRES iteration requires $N_{\boldsymbol{\xi}}$ solves with the diagonal matrix D_0 and $N_{\boldsymbol{q}}$ solves with L. If L is fully populated, this requires $O(N_{\boldsymbol{q}}(N_{\boldsymbol{\xi}}+N_{\boldsymbol{\xi}}^2))$ operations, assuming a Cholesky decomposition of L is given. If L is diagonal, applying the (1,1) block of (4.19) costs only $O(N_{\boldsymbol{q}}N_{\boldsymbol{\xi}})$ operations. Applying the (2,2) block of the preconditioner requires $N_{\boldsymbol{\xi}}$ single AMG V-cycles on linear systems with coefficient matrix S_0 and N_u multiplications with L. This requires $O(N_u(N_{\boldsymbol{\xi}}+N_{\boldsymbol{\xi}}^2))$ operations in general, or $O(N_uN_{\boldsymbol{\xi}})$ operations if L is diagonal. In summary, applying the preconditioner costs less than one matrix-vector product with the saddle point matrix, if L is diagonal. Moreover, if we store the Kronecker product factors of $\widehat{C}$ instead of assembling it, applying $\widehat{P}_{amg}$ is cheaper than one matrix-vector product with $\widehat{C}$ even for a fully populated matrix L (cf. the discussion at the end of Section 3.1).

5. Augmented preconditioners. We now focus on preconditioners of the form (1.3). To motivate a certain choice of weight matrix $\widehat{W}$ we review, first, the discrete inf-sup condition.

The natural norms on the spaces V and W are, respectively,

$$\| \boldsymbol{v} \|_V^2 = \left\langle \| \boldsymbol{v} \|_{H(div;D)}^2 \right\rangle, \qquad \| w \|_W^2 = \left\langle \| w \|_{L^2(D)}^2 \right\rangle. \tag{5.1}$$

Define the finite element matrices $A_I \in \mathbb{R}^{N_q \times N_q}$, $D \in \mathbb{R}^{N_q \times N_q}$ and $N \in \mathbb{R}^{N_u \times N_u}$ via

$$[A_I]_{ij} = \left(\boldsymbol{\varphi}_i, \boldsymbol{\varphi}_j\right), \qquad [D]_{ij} = \left(\nabla\cdot\boldsymbol{\varphi}_i, \nabla\cdot\boldsymbol{\varphi}_j\right), \qquad [N]_{rs} = \left(\phi_r, \phi_s\right). \tag{5.2}$$

Recalling that $G_0 = I$, for $\boldsymbol{v}_{hd} \in V_h \otimes S_d$ and $w_{hd} \in W_h \otimes S_d$, we have

$$\| \boldsymbol{v}_{hd} \|_V^2 = \boldsymbol{v}^T(\widehat{A}_I + \widehat{D})\boldsymbol{v}, \qquad \| w_{hd} \|_W^2 = \boldsymbol{w}^T\widehat{N}\boldsymbol{w}, \tag{5.3}$$

where $\widehat{D} = I \otimes D$, $\widehat{A}_I = I \otimes A_I$, $\widehat{N} = I \otimes N$ and $\boldsymbol{v} \in \mathbb{R}^{N_q}$ and $\boldsymbol{w} \in \mathbb{R}^{N_u}$ are the coefficient vectors associated with $\boldsymbol{v}_{hd}$ and w_{hd} respectively. If discontinuous pressure approximation is used, note that N is a diagonal matrix.

For our specific V_h, W_h and S_d, it was shown in [15] that $\exists\, \widehat{\beta} > 0$ depending only on the physical domain and the Raviart-Thomas interpolation operator $\Pi_h : H^1(D) \to V_h$ [8, Ch.3] such that:

$$\sup_{\boldsymbol{v}_{hd} \in V_h \otimes S_d \setminus \{0\}} \frac{\langle b\left(\boldsymbol{v}_{hd}, w_{hd}\right)\rangle}{\| \boldsymbol{v}_{hd} \|_V} \geq \widehat{\beta} \| w_{hd} \|_W \qquad \forall\, w_{hd} \in W_h \otimes S_d. \tag{5.4}$$

Equivalently, we have

$$\widehat{\beta}^2 \leq \frac{\boldsymbol{v}^T\widehat{B}(\widehat{A}_I + \widehat{D})^{-1}\widehat{B}^T\boldsymbol{v}}{\boldsymbol{v}^T\widehat{N}\boldsymbol{v}} \qquad \forall \boldsymbol{v} \in \mathbb{R}^{N_{\boldsymbol{\xi}}N_u} \setminus \{\boldsymbol{0}\}$$

and writing $\boldsymbol{v} = \boldsymbol{u} \otimes \boldsymbol{w}$ where $\boldsymbol{u} \in \mathbb{R}^{N_{\boldsymbol{\xi}}}$ and $\boldsymbol{w} \in \mathbb{R}^{N_u}$ gives

$$\widehat{\beta}^2 \leq \frac{\boldsymbol{v}^T\left(I \otimes B(A_I + D)^{-1}B^T\right)\boldsymbol{v}}{\boldsymbol{v}^T\left(I \otimes N\right)\boldsymbol{v}} = \frac{\boldsymbol{w}^T B(A_I + D)^{-1}B^T\boldsymbol{w}}{\boldsymbol{w}^T N\boldsymbol{w}}. \tag{5.5}$$

Now, for *any* deterministic finite element spaces V_h, W_h that satisfy the usual inf-sup condition

$$\sup_{\boldsymbol{v}_h \in V_h \setminus \{0\}} \frac{b\left(\boldsymbol{v}_h, w_h\right)}{\| \boldsymbol{v}_h \|_{H(div;D)}} \geq \beta \| w_h \|_{L^2(D)} \qquad \forall\, w_h \in W_h, \tag{5.6}$$

there exists a constant $\beta > 0$ independent of the characteristic mesh size h such that

$$\beta^2 \leq \frac{\boldsymbol{w}^T B\left(A_I + D\right)^{-1} B^T\boldsymbol{w}}{\boldsymbol{w}^T N\boldsymbol{w}} \qquad \forall\, \boldsymbol{w} \in \mathbb{R}^{N_u} \setminus \{\boldsymbol{0}\}.$$

From (5.5) we see that $\widehat{\beta}$ coincides with the deterministic inf-sup constant. This is not a coincidence. Starting from any deterministic pair V_h and W_h satisfying (5.6), the commutativity diagram (see [8, pp.131]) that connects $H(div;D)$, $L^2(D)$, V_h and W_h can be easily re-drawn for the tensor product spaces $H(div;D) \otimes S_d$, $L^2(D) \otimes S_d$, $V_h \otimes S_d$ and $W_h \otimes S_d$ for any $S_d \subset L^2_\rho(\Gamma)$.

5.1. H(div) preconditioning. In [15], the so-called $H(div)$ preconditioner

$$\widehat{P}_{div} := \begin{bmatrix} \widehat{A}+\widehat{D} & 0 \\ 0 & \widehat{N} \end{bmatrix}, \tag{5.7}$$

is studied, where $\widehat{N} = I \otimes N$, $\widehat{D} = \widehat{B}^T \widehat{N}^{-1} \widehat{B} = I \otimes B^T N^{-1} B$ and N is the deterministic mass matrix. $\widehat{P}_{div}$ is equivalent to (1.3) with $\gamma = 1$ and weight matrix $\widehat{W} = \widehat{N}$. Thanks to (5.3), $\widehat{N}$ provides a discrete representation of the norm $\|\cdot\|_W$ on $W_h \otimes S_d$. We also have

$$\boldsymbol{v}^T \left(\widehat{A}+\widehat{D}\right) \boldsymbol{v} = \left\langle \left(T_M^{-1} \boldsymbol{v}_{hd}, \boldsymbol{v}_{hd}\right)\right\rangle + \left\langle (\nabla \cdot \boldsymbol{v}_{hd}, \nabla \cdot \boldsymbol{v}_{hd})\right\rangle = \left\langle \| \boldsymbol{v}_{hd} \|^2_{div,T_M} \right\rangle. \tag{5.8}$$

Choosing $\widehat{W} = \widehat{N}$ allows us to express the eigenvalues of $\widehat{P}_{div}^{-1}\widehat{C}$ in terms of the inf-sup constant $\widehat{\beta}$ in (5.4). This leads to spectral inclusion bounds that are independent of the parameters h, d and M. To obtain bounds that are also independent of $T_{min,M}$ and $T_{max,M}$ in (2.10) we consider a parameterized version of the same preconditioner.

THEOREM 5.1. *Let $\gamma > 0$. The generalized eigenvalue problem,*

$$\begin{pmatrix} \widehat{A} & \widehat{B}^T \\ \widehat{B} & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{q} \\ \boldsymbol{u} \end{pmatrix} = \lambda \begin{pmatrix} \widehat{A}+\gamma^{-1}\widehat{B}^T\widehat{N}^{-1}\widehat{B} & 0 \\ 0 & \gamma\widehat{N} \end{pmatrix} \begin{pmatrix} \boldsymbol{q} \\ \boldsymbol{u} \end{pmatrix} \tag{5.9}$$

has $N_{\boldsymbol{\xi}} N_q$ eigenvalues at $+1$. The remaining $N_{\boldsymbol{\xi}} N_u$ values are negative and lie in the bounded interval

$$\left(-1, -\frac{\widehat{\beta}^2\ T_{min,M}}{\gamma + \widehat{\beta}^2\ T_{min,M}}\right], \tag{5.10}$$

where $T_{min,M}$ is defined in (2.10) *and $\widehat{\beta}$ is the inf-sup constant defined in* (5.4).

Proof. It is easy to see that all positive eigenvalues lie in a cluster at $+1$. This is a major benefit of augmented preconditioning. The remaining eigenvalues are negative and satisfy

$$\widehat{B}\left(\widehat{A}+\gamma^{-1}\widehat{B}^T\widehat{N}^{-1}\widehat{B}\right)^{-1}\widehat{B}^T\boldsymbol{u} = -\lambda\gamma\widehat{N}\boldsymbol{u} \tag{5.11}$$

(see [34], [28]). The positive values $-\lambda$ coincide with the values $\frac{\sigma_i}{1+\sigma_i}$ where $\sigma_i > 0$ is an eigenvalue of $(\gamma\widehat{N})^{-1}\widehat{B}\widehat{A}^{-1}\widehat{B}^T$. From (5.4) we have, for any $w_{hd} \in W_h \otimes S_d$

$$T_{min,M}^{\frac{1}{2}}\widehat{\beta} \parallel w_{hd} \parallel_W \leq \sup_{\boldsymbol{v} \in V_h \otimes S_d \setminus \{\boldsymbol{0}\}} \frac{\langle (\nabla \cdot \boldsymbol{v}_{hd}, w_{hd})\rangle}{\left\langle \| T_M^{-\frac{1}{2}} \boldsymbol{v}_{hd} \|_{L^2(D)} \right\rangle} = \max_{v \in \mathbb{R}^{N_q}} \frac{\boldsymbol{w}^T\widehat{B}\boldsymbol{v}}{(\boldsymbol{v}^T\widehat{A}\boldsymbol{v})^{\frac{1}{2}}} = \left(\boldsymbol{w}^T\widehat{B}\widehat{A}^{-1}\widehat{B}^T\boldsymbol{w}\right)^{\frac{1}{2}}.$$

Since $\| w_{hd} \|^2_W = \boldsymbol{w}^T\widehat{N}\boldsymbol{w}$, we have $\min_i \{\sigma_i\} \geq \gamma^{-1}\widehat{\beta}^2 T_{min,M}$ and the result follows. □

Theorem 5.1 indicates that as $\gamma \to 0$ the negative eigenvalues cluster at -1. Choosing $\gamma = O(T_{min,M})$ leads to a bound that is independent of the PDE coefficients. However, it is not desirable to choose γ too small, as this can cause numerical difficulties for preconditioned MINRES.

5.2. Alternative weight matrix. Now consider the alternative preconditioner

$$\widehat{P}_{L,div} := \begin{bmatrix} \widehat{A}+\gamma^{-1}\widehat{B}^T\widehat{W}_L^{-1}\widehat{B} & 0 \\ 0 & \gamma\widehat{W}_L \end{bmatrix}, \tag{5.12}$$

where $\widehat{W}_L = L^{-1} \otimes N$ and $L \in \mathbb{R}^{N_{\boldsymbol{\xi}} \times N_{\boldsymbol{\xi}}}$ is *any* symmetric positive definite matrix. The following result says that if γ and L are chosen appropriately, preconditioned MINRES will converge in a few iterations, independently of all the problem parameters.

LEMMA 5.2. *Let L be a symmetric positive definite matrix. $\widehat{P}_{L,div}^{-1}\widehat{C}$ has $N_{\boldsymbol{\xi}}N_q$ eigenvalues at $+1$. The remaining $N_{\boldsymbol{\xi}}N_u$ eigenvalues are negative and contained in the interval $(-1,-c]$, where*

$$c := \frac{\widehat{\beta}^2\,\ell_{min}\,T_{min,M}}{\gamma+\widehat{\beta}^2\,\ell_{min}\,T_{min,M}} > 0, \tag{5.13}$$

$T_{min,M}$ is defined in (2.10), $\widehat{\beta}$ is defined in (5.4) and $\ell_{min}>0$ is the minimum eigenvalue of L.

Proof. Follow the proof of Theorem 5.1 and substitute $\widehat{W}_L = L^{-1}\otimes N$ for $\widehat{N}$. This time, $\sigma_i>0$ is an eigenvalue of $(\gamma\widehat{W}_L)^{-1}\left(\widehat{B}\widehat{A}^{-1}\widehat{B}^T\right)$ and from (5.4) we have

$$\widehat{\beta}^2 \le \frac{1}{T_{min,M}}\frac{\boldsymbol{w}^T\widehat{B}\widehat{A}^{-1}\widehat{B}^T\boldsymbol{w}}{\boldsymbol{w}^T\widehat{W}_L\boldsymbol{w}}\cdot\frac{\boldsymbol{w}^T\widehat{W}_L\boldsymbol{w}}{\boldsymbol{w}^T\widehat{N}\boldsymbol{w}} \le \frac{1}{\ell_{min}T_{min,M}}\frac{\boldsymbol{w}^T\widehat{B}\widehat{A}^{-1}\widehat{B}^T\boldsymbol{w}}{\boldsymbol{w}^T\widehat{W}_L\boldsymbol{w}} \quad \forall \boldsymbol{w}\in\mathbb{R}^{N_u}\setminus\{\boldsymbol{0}\}.$$

□

To apply (5.12) we need to be able to approximate the action of $(\widehat{A}+\gamma^{-1}\widehat{B}^T\widehat{W}_L^{-1}\widehat{B})^{-1}$. A multigrid method that exploits the bilinear form (5.8) was introduced in [15] to solve linear systems of equations with coefficient matrix $\widehat{A}+\widehat{B}^T\widehat{N}^{-1}\widehat{B}$. If $L=I$, $\widehat{W}_L = I\otimes N$ and

$$\boldsymbol{v}^T\left(\widehat{A}+\gamma^{-1}\widehat{B}^T\widehat{W}_L^{-1}\widehat{B}\right)\boldsymbol{v} = \left\langle\left(T_M^{-1}\boldsymbol{v}_{hd},\boldsymbol{v}_{hd}\right)\right\rangle + \gamma^{-1}\left\langle(\nabla\cdot\boldsymbol{v}_{hd},\nabla\cdot\boldsymbol{v}_{hd})\right\rangle. \tag{5.14}$$

The multigrid method from [15] can be applied, even if $\gamma\neq 1$. It becomes excessively expensive, however, as M and d increase as it requires exact solves with matrices of dimension $O(N_{\boldsymbol{\xi}})\times O(N_{\boldsymbol{\xi}})$ at each smoothing step, at each level. For an arbitrary matrix L there is no obvious bilinear form on $V_h\otimes S_d$ to which $\widehat{A}+\gamma^{-1}\widehat{B}^T\widehat{W}_L^{-1}\widehat{B}$ corresponds that can be used to develop a practical solution algorithm. Our motivation for allowing $L\neq I$ is as follows.

5.3. Cheaper preconditioner. For brevity, let $\widehat{D}_L = \widehat{B}^T\widehat{W}_L^{-1}\widehat{B}$. To develop a cheaper preconditioner than the one studied in [15] we want to replace $\widehat{A}$ in (5.12) with an approximation of the form $\widehat{X}=L\otimes A_0$. If $\widehat{W}_L = L^{-1}\otimes N$ then

$$\widehat{X}+\gamma^{-1}\widehat{D}_L = L\otimes A_0+\gamma^{-1}L\otimes B^TN^{-1}B = L\otimes\left(A_0+\gamma^{-1}B^TN^{-1}B\right). \tag{5.15}$$

The right Kronecker factor in (5.15) is associated with a weighted *deterministic* $H(div)$ bilinear form on V_h. Specifically, for any $\boldsymbol{v}_h\in V_h$, with associated coefficient vector $\boldsymbol{v}\in\mathbb{R}^{N_q}$, we have

$$\boldsymbol{v}^T\left(A_0+\gamma^{-1}B^TN^{-1}B\right)\boldsymbol{v} = (t_0\,\boldsymbol{v}_h,\boldsymbol{v}_h)+\gamma^{-1}\left(\nabla\cdot\boldsymbol{v}_h,\nabla\cdot\boldsymbol{v}_h\right). \tag{5.16}$$

Crucially, this means that existing deterministic solvers can be exploited (e.g. see [1], [17]) since

$$(\widehat{X}+\gamma^{-1}\widehat{D}_L)^{-1} = L^{-1}\otimes\left(A_0+\gamma^{-1}B^TN^{-1}B\right)^{-1}.$$

To analyze the efficiency of the resulting preconditioner

$$\widehat{P}_{X,div} := \begin{bmatrix}\widehat{X}+\gamma^{-1}\widehat{D}_L & 0\\ 0 & \gamma\widehat{W}_L\end{bmatrix} \tag{5.17}$$

compared to (5.12), we need the following result.

LEMMA 5.3. *Let $\widehat{X}$ be symmetric and positive definite and let $0<\nu_1\le\ldots\le\nu_n$ where $n=N_{\boldsymbol{\xi}}N_q$, be the eigenvalues of $\widehat{X}^{-1}\widehat{A}$. The generalised eigenvalue problem*

$$(\widehat{A}+\gamma^{-1}\widehat{D}_L)\boldsymbol{v} = \lambda(\widehat{X}+\gamma^{-1}\widehat{D}_L)\boldsymbol{v} \tag{5.18}$$

has $N_{\boldsymbol{\xi}}(N_q-N_u)$ eigenvalues independent of γ which are contained in the interval $[\nu_1,\nu_n]$. All $N_{\boldsymbol{\xi}}N_q$ eigenvalues are contained in the interval $[\alpha_1,\alpha_2]$ where $\alpha_1=\min(1,\nu_1)$ and $\alpha_2=\max(1,\nu_n)$.

Proof. $\widehat{X}$ and $\widehat{A}$ are positive definite and $\widehat{D}_L$ is positive semi-definite so $\lambda > 0$. If $\boldsymbol{v} \in \text{null}(\widehat{B})$ then $\widehat{A}\boldsymbol{v} = \lambda\widehat{X}\boldsymbol{v}$. $\widehat{B}$ has a nullspace of dimension $N_{\boldsymbol{\xi}}(N_q - N_u)$, so at least $N_{\boldsymbol{\xi}}(N_q - N_u)$ eigenvalues are contained in the interval $[\nu_1, \nu_n]$. Rearranging (5.18) gives

$$\gamma(\widehat{A} - \lambda\widehat{X})\boldsymbol{v} = (\lambda - 1)\widehat{D}_L\boldsymbol{v}. \tag{5.19}$$

If $(\lambda - 1) \leq 0$ then $\widehat{A} - \lambda\widehat{X}$ is negative semi-definite and $\nu_1 \leq \lambda \leq 1$. If $\lambda > 1$ then $\widehat{A} - \lambda\widehat{X}$ is positive semi-definite and $1 < \lambda \leq \nu_n$. Hence $\lambda \in [\nu_1, 1] \cup (1, \nu_n]$. □

Note that if, for the chosen $\widehat{X}$, we have $\nu_1 \leq 1 \leq \nu_n$ then all eigenvalues in (5.18) lie in $[\nu_1, \nu_n]$.

REMARK 5.4. *Lemma 5.3 says that a subset of eigenvalues in* (5.18) *are insensitive to* γ *and depend only on* $\widehat{X}$. *Observe also that if* $\boldsymbol{v} \notin$ *null*$(\widehat{B})$, $\widehat{D}_L\boldsymbol{v} \neq \boldsymbol{0}$ *and so if* $\gamma \to 0$ *in* (5.19) *then* $\lambda \to 1$. *Hence the remaining eigenvalues can be forced to cluster at* $+1$ *by choosing* γ *small enough.*

LEMMA 5.5. *Let* $\widehat{X}$ *satisfy the conditions of Lemma 5.3 and let* $\widehat{W}_L = L^{-1} \otimes N$ *where* L *is symmetric and positive definite. Then, the eigenvalues of*

$$\begin{pmatrix} \widehat{A} & \widehat{B}^T \\ \widehat{B} & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{q} \\ \boldsymbol{u} \end{pmatrix} = \lambda \begin{pmatrix} \widehat{X} + \gamma^{-1}\widehat{D}_L & 0 \\ 0 & \gamma\widehat{W}_L \end{pmatrix} \begin{pmatrix} \boldsymbol{q} \\ \boldsymbol{u} \end{pmatrix} \tag{5.20}$$

are contained in the union of the intervals

$$\left[-\sqrt{\alpha_2},\ \frac{1}{2}\left(\alpha_1(1-c) - \sqrt{\alpha_1^2(c-1)^2 + 4\,c\,\alpha_1}\right)\right] \cup [\alpha_1, \alpha_2], \tag{5.21}$$

where c *is defined in* (5.13) *and* $\alpha_1 = \min(1, \nu_1)$ *and* $\alpha_2 = \max(1, \nu_n)$.

Proof. This result is similar to [15, Theorem 6] which is for the deterministic problem with $\gamma = 1$, $L = I$ and $\alpha_2 = 1$. Eliminating $\boldsymbol{u}$ in (5.20) and rearranging gives

$$\lambda(\widehat{A} + \gamma^{-1}\widehat{D}_L)\boldsymbol{q} + \gamma^{-1}(1 - \lambda)\widehat{D}_L\boldsymbol{q} = \lambda^2(\widehat{X} + \gamma^{-1}\widehat{D}_L)\boldsymbol{q}.$$

Let $\lambda > 0$. Since $\text{null}(\widehat{D}_L) = \text{null}(\widehat{B})$ at least $N_{\boldsymbol{\xi}}(N_q - N_u)$ eigenvalues are contained in $[\nu_1, \nu_n]$ and are independent of γ. The remaining $N_{\boldsymbol{\xi}}N_u$ positive eigenvalues must satisfy $\lambda \to 1$ as $\gamma \to 0$. Lemma 5.3 says that $\gamma^{-1}(1 - \lambda)\boldsymbol{q}^T\widehat{D}_L\boldsymbol{q} \leq (\lambda^2\alpha_1^{-1} - \lambda)\boldsymbol{q}^T(\widehat{A} + \gamma^{-1}\widehat{D}_L)\boldsymbol{q}$. If $\lambda \leq 1$ then since $\boldsymbol{q}^T\widehat{D}_L\boldsymbol{q} \geq 0$ we must have $\lambda^2\alpha_1^{-1} - \lambda \geq 0$ and so $\lambda \in [\alpha_1, 1]$. Similarly, if $\lambda > 1$ then $\lambda^2\alpha_2^{-1} - \lambda \leq 0$ and so $\lambda \in (1, \alpha_2]$. Hence the positive eigenvalues belong to $[\alpha_1, 1] \cup (1, \alpha_2] = [\alpha_1, \alpha_2]$.

Now let $\lambda < 0$. Eliminating $\boldsymbol{q}$ gives

$$\widehat{B}(\lambda(\widehat{X} + \gamma^{-1}\widehat{D}_L) - \widehat{A})^{-1}\widehat{B}^T\boldsymbol{u} = \lambda\gamma\widehat{W}_L\boldsymbol{u}.$$

These $N_{\boldsymbol{\xi}}N_u$ eigenvalues coincide with the values $\frac{\sigma_i}{1+\sigma_i}$ where $-1 < \sigma_i < 0$ is an eigenvalue of

$$\widehat{B}\left(\lambda\left(\widehat{X} + \gamma^{-1}\widehat{D}_L\right) - \left(\widehat{A} + \gamma^{-1}\widehat{D}_L\right)\right)^{-1}\widehat{B}^T\boldsymbol{u} = \sigma\gamma\widehat{W}_L\boldsymbol{u}.$$

The eigenvalues of $(\lambda(\widehat{X} + \gamma^{-1}\widehat{D}_L) - (\widehat{A} + \gamma^{-1}\widehat{D}_L))^{-1}\boldsymbol{x} = \mu(\widehat{A} + \gamma^{-1}\widehat{D}_L))^{-1}\boldsymbol{x}$ satisfy

$$\frac{\lambda\mu}{1+\mu} = \frac{\boldsymbol{x}^T(\widehat{A} + \gamma^{-1}\widehat{D}_L)\boldsymbol{x}}{\boldsymbol{x}^T(\widehat{X} + \gamma^{-1}\widehat{D}_L)\boldsymbol{x}}. \tag{5.22}$$

Setting $\boldsymbol{x} = \widehat{B}^T\boldsymbol{u}$ and using Lemma 5.3 it can then be shown that

$$\frac{\alpha_2}{\lambda - \alpha_2} \leq \frac{\boldsymbol{u}^T\widehat{B}(\lambda(\widehat{X} + \gamma^{-1}\widehat{D}_L) - (\widehat{A} + \gamma^{-1}\widehat{D}_L))^{-1}\widehat{B}^T\boldsymbol{u}}{\boldsymbol{u}^T\widehat{B}(\widehat{A} + \gamma^{-1}\widehat{D}_L)^{-1}\widehat{B}^T\boldsymbol{u}} \leq \frac{\alpha_1}{\lambda - \alpha_1}$$

and using Lemma 5.2 that

$$\frac{\alpha_2}{\lambda-\alpha_2} \le \frac{\boldsymbol{u}^T\widehat{B}(\lambda(\widehat{X}+\gamma^{-1}\widehat{D}_L)-(\widehat{A}+\gamma^{-1}\widehat{D}_L))^{-1}\widehat{B}^T\boldsymbol{u}}{\boldsymbol{u}^T(\gamma\widehat{W}_L)\boldsymbol{u}} \le \frac{c\,\alpha_1}{\lambda-\alpha_1}.$$

Hence $\frac{\alpha_2}{\lambda} \le \frac{\sigma}{1+\sigma} \le \frac{c\,\alpha_1}{\lambda+\alpha_1(c-1)}$. This gives $\alpha_2 \ge \lambda^2$ and $\lambda^2+\lambda\alpha_1\,(c-1)-c\alpha_2 \ge 0$ and solving for λ gives the result. □

REMARK 5.6. *If $\widehat{X}=\widehat{A}$, $\alpha_1=1=\alpha_2$ and* (5.21) *reduces to the bound in Lemma 5.2.*

REMARK 5.7. *If $\gamma\to 0$ then, in* (5.13)*, $c\to 1$, leading to spectral inclusion bounds* (5.21) *that depend only on α_1 and α_2 i.e. on the choice of $\widehat{X}$. However, in the limit $\gamma\to 0$,* (5.21) *is pessimistic. In* (5.22)*, $\boldsymbol{x}=\widehat{B}^T\boldsymbol{u}\notin null(\widehat{B})$. By Remark 5.4, the bound $[\alpha_1,\alpha_2]$ is pessimistic for that subset of eigenvalues of* (5.18) *and, in fact, for the given $\boldsymbol{x}$, $\frac{\lambda\mu}{1+\mu}\to 1$ as $\gamma\to 0$. Asymptotically, then* (5.21) *reduces to $[-1]\cup[\alpha_1,\alpha_2]$. A subset of the positive eigenvalues of the preconditioned system in* (5.20) *are also forced to $+1$ as $\gamma\to 0$. There remain, however, $N_{\boldsymbol{\xi}}(N_q-N_u)$ positive eigenvalues lying in $[\nu_1,\nu_n]$ that are completely insensitive to γ and that can only be controlled via the choice of $\widehat{X}$.*

In view of (5.15) we'd like to choose $\widehat{X}=L\otimes A_0$ so that known deterministic solvers can be exploited. In that case, we have

$$\widehat{X}^{-1}\widehat{A} = L^{-1}\otimes I + \sum_{\boldsymbol{\alpha}\in\mathscr{I}_{2d}^+} L^{-1}G_{\boldsymbol{\alpha}}\otimes A_0^{-1}A_{\boldsymbol{\alpha}}. \tag{5.23}$$

From our study of Schur-complement preconditioners, it is clear that we cannot achieve spectral inclusion bounds for $\widehat{X}^{-1}\widehat{A}$ that are independent of *all* the problem parameters. However, by Remark 5.7, there is hope that by choosing γ appropriately, the weakness of the approximation $\widehat{X}$ will only have an impact on the positive eigenvalues of the preconditioned system.

LEMMA 5.8. *Let $\widehat{X}=L\otimes A_0$ where $L\in\mathbb{R}^{N_{\boldsymbol{\xi}}\times N_{\boldsymbol{\xi}}}$ is any symmetric positive definite matrix whose eigenvalues lie in the interval $[\ell_{min},\ell_{max}]$, where $0<\ell_{min}\le\ell_{max}$. For each $\boldsymbol{\alpha}\in\mathscr{I}_{2d}^+$, define $c_{\boldsymbol{\alpha}}$ as in* (4.4)*, and $b_{\boldsymbol{\alpha}}$ as in* (4.9)*. Then, the eigenvalues of $\widehat{X}^{-1}\widehat{A}$ lie in the interval*

$$\left[\frac{1}{\ell_{max}}-\frac{\tau}{\ell_{min}},\frac{(1+\tau)}{\ell_{min}}\right], \qquad \tau := \sum_{\boldsymbol{\alpha}\in\mathscr{I}_{2d}^+} b_{\boldsymbol{\alpha}}c_{\boldsymbol{\alpha}} > 0. \tag{5.24}$$

Proof. Follow the proof of Theorem 4.7 replacing $D_0=\mathrm{diag}(A_0)$ with A_0. □

Since $\widehat{A}$ is positive definite, $\nu_1=\lambda_{min}(\widehat{X}^{-1}\widehat{A})>0$ but the lower bound in (5.24) is not necessarily positive. The message is clear, however. Since τ increases with d,σ_G and M a good choice of L is a matrix that damps out ill-conditioning caused by those parameters.

5.4. Mean-based preconditioners. For comparison, consider first $L=I$. Then $\widehat{X}=I\otimes A_0$ and (5.17) is the mean-based preconditioner

$$\widehat{P}_{0,div} := \begin{bmatrix} I\otimes(A_0+\gamma^{-1}B^TN^{-1}B) & 0 \\ 0 & I\otimes\gamma N \end{bmatrix}. \tag{5.25}$$

COROLLARY 5.9. *The eigenvalues of $\widehat{P}_{0,div}^{-1}\widehat{C}$ are bounded and contained in the union of the intervals in* (5.21) *where c is defined in* (5.13) *with $l_{min}=1$, $\alpha_1=\nu_1\ge 1-\tau$ and $\alpha_2=\nu_n\le 1+\tau$.*

Proof. From (5.23) we have $\widehat{X}^{-1}\widehat{A}=\widehat{I}+\widehat{S}$ where $\widehat{I}$ is the $N_{\boldsymbol{\xi}}N_q\times N_{\boldsymbol{\xi}}N_q$ identity matrix and $\widehat{S}$ is indefinite. Using Lemma 5.8 with $L=I$ gives $1-\tau\le\nu_1$ and $\nu_n\le 1+\tau$. We also have $\nu_1\le 1\le\nu_n$ and so the result follows by Lemma 5.5. □

When σ_G and d are large we can expect this preconditioner, like the Schur-complement preconditioner $\widehat{P}_0$, to lose efficiency.

5.5. Kronecker product preconditioners. Alternatively, we can again adopt the best-fit approach introduced in [32]. If we choose $\widehat{X} = Q \otimes A_0$ with

$$Q = \text{argmin}\{H \in \mathbb{R}^{N_{\boldsymbol{\xi}} \times N_{\boldsymbol{\xi}}} : ||\widehat{A} - H \otimes A_0||_F\}$$

we arrive at the preconditioner

$$\widehat{P}_{1,div} := \begin{bmatrix} Q \otimes (A_0 + \gamma^{-1} B^T N^{-1} B) & 0 \\ 0 & Q^{-1} \otimes \gamma N \end{bmatrix} \tag{5.26}$$

where Q is the symmetric and positive definite matrix

$$Q = I + \sum_{\boldsymbol{\alpha} \in \mathscr{I}_{2d}^+} \frac{\text{tr}(A_{\boldsymbol{\alpha}}^\top A_0)}{\text{tr}(A_0^\top A_0)} G_{\boldsymbol{\alpha}} \tag{5.27}$$

(cf. [33, Theorem 3]). Since $L = Q \in \mathbb{R}^{N_{\boldsymbol{\xi}} \times N_{\boldsymbol{\xi}}}$ is not diagonal we also consider

$$\widehat{P}_{2,div} := \begin{bmatrix} \text{diag}(Q) \otimes (A_0 + \gamma^{-1} B^T N^{-1} B) & 0 \\ 0 & \text{diag}(Q)^{-1} \otimes \gamma N \end{bmatrix}. \tag{5.28}$$

COROLLARY 5.10. *The eigenvalues of $\widehat{P}_{1,div}^{-1}\widehat{C}$ and $\widehat{P}_{2,div}^{-1}\widehat{C}$ are bounded and contained in the union of the intervals in* (5.21) *where c is defined in* (5.13) *with $l_{min} \geq 1 - \tau$.*

Proof. Apply Lemma 5.5. The lower bound for l_{min} can be established as in Corollary 4.9. □

5.6. Practical augmented preconditioners. In [1]–[2], a geometric multigrid method for solving linear systems of equations arising from discretizations of the bilinear form (5.16) is analyzed. Specifically, [2] shows that when t_0 is constant, the matrix V_0^{div} whose inverse corresponds to the application of one multigrid V-cycle to a system with coefficient matrix $A_0 + \gamma^{-1} B^T N^{-1} B$, satisfies

$$1 - \delta \leq \frac{\boldsymbol{v}^T (A_0 + \gamma^{-1} B^T N^{-1} B) \boldsymbol{v}}{\boldsymbol{v}^T V_0^{div} \boldsymbol{v}} \leq 1, \qquad \forall \boldsymbol{v} \in \mathbb{R}^{N_q} \setminus \{\boldsymbol{0}\} \tag{5.29}$$

where $\delta > 0$ is a constant depending only on the number of smoothing steps performed. In Section 6 we implement the method from [1]–[2]. However, *any* V_0^{div} which satisfies (5.29) and whose inverse can be applied in $O(N_q)$ work, can be used as a building block to obtain practical versions of the preconditioners (5.25), (5.26) and (5.28). Consider, then

$$\widehat{P}_{L,div,mg} := \begin{bmatrix} L \otimes V_0^{div} & 0 \\ 0 & L^{-1} \otimes \gamma N \end{bmatrix}. \tag{5.30}$$

COROLLARY 5.11. *Let $V_0^{div} \in \mathbb{R}^{N_q \times N_q}$ be any matrix that satisfies* (5.29). *The eigenvalues of $\widehat{P}_{L,div,mg}^{-1}\widehat{C}$ are contained in the union of the intervals* (5.21) *where α_1 is replaced by $\widehat{\alpha}_1 = (1-\delta)\alpha_1$.*

Proof. Follow the proof of Lemma 5.5 replacing $\widehat{X}_L + \gamma^{-1}\widehat{D}_L$ in the (1,1) block of the preconditioner with $L \otimes V_0^{div}$. For $\lambda > 0$ we obtain

$$\lambda(\widehat{A} + \gamma^{-1}\widehat{D}_L)\boldsymbol{q} + \gamma^{-1}(1-\lambda)\widehat{D}_L\boldsymbol{q} = \lambda^2 (L \otimes V_0^{div})\boldsymbol{q}. \tag{5.31}$$

Observe that for any $\boldsymbol{q} \in \mathbb{R}^{N_{\boldsymbol{\xi}} N_q}$ with $\boldsymbol{q} = \boldsymbol{u} \otimes \boldsymbol{v}$,

$$\begin{aligned} \frac{\boldsymbol{q}^T(\widehat{A} + \gamma^{-1}\widehat{D}_L)\boldsymbol{q}}{\boldsymbol{q}^T(L \otimes V_0^{div})\boldsymbol{q}} &= \frac{\boldsymbol{q}^T(\widehat{A} + \gamma^{-1}\widehat{D}_L)\boldsymbol{q}}{\boldsymbol{q}^T(L \otimes (A_0 + \gamma^{-1}B^T N^{-1}B)\boldsymbol{q}} \, \frac{\boldsymbol{q}^T(L \otimes (A_0 + \gamma^{-1}B^T N^{-1}B))\boldsymbol{q}}{\boldsymbol{q}^T(L \otimes V_0^{div})\boldsymbol{q}} \\ &= \frac{\boldsymbol{q}^T(\widehat{A} + \gamma^{-1}\widehat{D}_L)\boldsymbol{q}}{\boldsymbol{q}^T(\widehat{X}_L + \gamma^{-1}\widehat{D}_L)\boldsymbol{q}} \, \frac{\boldsymbol{v}^T(A_0 + \gamma^{-1}B^T N^{-1}B)\boldsymbol{v}}{\boldsymbol{v}^T V_0^{div}\boldsymbol{v}}. \end{aligned}$$

Hence, if (5.29) holds, the eigenvalues of $(\widehat{A} + \gamma^{-1}\widehat{D}_L)\boldsymbol{v} = \mu(L \otimes V_0^{div})\boldsymbol{v}$ belong to the interval $[\widehat{\alpha}_1, \alpha_2]$ where $\widehat{\alpha}_1 = (1-\delta)\alpha_1$ and (from Lemma 5.3) $\alpha_1 = \min(1, \nu_1)$ and $\alpha_2 = \max(1, \nu_n)$ with $\nu_1 = \lambda_{min}(\widehat{X}_L^{-1}\widehat{A})$ and $\nu_n = \lambda_{max}(\widehat{X}_L^{-1}\widehat{A})$. In (5.31) we then have

$$\gamma^{-1}(1-\lambda)\boldsymbol{q}^T\widehat{D}_L\boldsymbol{q} \leq (\lambda^2\widehat{\alpha}_1^{-1} - \lambda)\boldsymbol{q}^T(\widehat{A} + \gamma^{-1}\widehat{D})\boldsymbol{q}.$$

If $\lambda \leq 1$ then $\lambda^2\widehat{\alpha}_1^{-1} - \lambda \leq 0$ so $\lambda \in [\widehat{\alpha}_1, 1]$. If $\lambda > 1$ then $\lambda^2\alpha_2^{-1} - \lambda \leq 0$ and so $\lambda \in (1, \alpha_2]$. Hence the positive eigenvalues belong to $[\widehat{\alpha}_1, 1] \cup (1, \alpha_2] = [\widehat{\alpha}_1, \alpha_2]$. The bound for the negative eigenvalues follows from (5.22) with $\widehat{X}_L + \gamma^{-1}\widehat{D}_L$ replaced by $L \otimes V_0^{div}$ and α_1 replaced by $\widehat{\alpha}_1$. □

REMARK 5.12. *The performance of $\widehat{P}_{L,div,mg}$ in* (5.30) *depends on δ in* (5.29). *For the multigrid method from [1]–[2], δ depends on the number of smoothing steps and potentially on $t_0(\boldsymbol{x})$ (if it is not spatially constant) but is independent of the augmentation parameter γ.*

Applying the (1,1) block of $\widehat{P}_{L,div,mg}$, in each MINRES iteration, requires $N_{\boldsymbol{\xi}}$ single V-cycles of a multigrid method on systems with coefficient matrix $A_0 + \gamma^{-1}B^TN^{-1}B$ and N_q solves with L. Assuming the multigrid method is optimal, the cost per iteration is $O((N_{\boldsymbol{\xi}} + N_{\boldsymbol{\xi}}^2)N_q)$ if $L = Q$ (and a Cholesky decomposition is given) or $O(N_{\boldsymbol{\xi}}N_q)$ if L is diagonal. Inverting the (2,2) block requires N_u multiplications with L and $N_{\boldsymbol{\xi}}$ solves with the diagonal matrix N. The associated cost is $O((N_{\boldsymbol{\xi}} + N_{\boldsymbol{\xi}}^2)N_u)$ if $L = Q$ or $O(N_{\boldsymbol{\xi}}N_u)$ if L is diagonal. Once again, if $\widehat{C}$ is not assembled, the cost of applying each preconditioner is less than the cost of a matrix-vector product with $\widehat{C}$. Note that increasing the number of smoothing steps ν in a multigrid method improves the constant δ in (5.29), so MINRES iterations, and hence matrix-vector products, can be saved for a fixed L, by increasing ν.

6. Numerical examples. We discretize (2.3) on the unit square with $f(\boldsymbol{x}) = 1$ and $g(\boldsymbol{x}) = 0$. For the spatial discretization we employ $RT_0 - P_0$ triangular elements. The finite element mesh consists of 32^2 squares, each divided into two triangles, resulting in $N_{\boldsymbol{x}} = N_q + N_u = 5,184$ spatial degrees of freedom. The coefficient T is modelled as a lognormal random field as discussed in Section 2.2. The Gaussian field G is a truncated KL expansion, with $\mu_G = 1$ and standard deviation $\sigma_G \geq 0$. The covariance function is $\mathrm{Cov}_G(\boldsymbol{x}, \boldsymbol{y}) = \sigma_G^2 r K_1(r)$ where $r = \|\boldsymbol{x} - \boldsymbol{y}\|_2$ and K_1 is the modified Bessel function of the second kind with order one. We use $M = 5$ random variables in (2.12) to capture 97% of the Gaussian field's total variance and S_d consists of polynomials of total degree d in those 5 variables. The dimension of the resulting Galerkin matrix is given in Table 6.1.

TABLE 6.1
Values of $N_{\boldsymbol{\xi}}$, $N_{\boldsymbol{x}}N_{\boldsymbol{\xi}}$ and number of terms $N+1$ in the Kronecker product representation of $\widehat{A}$.

	d=1	d=2	d=3	d=4	d=5
$N_{\boldsymbol{\xi}}$	6	21	56	126	252
$N+1$	21	126	462	1,287	3,003
$N_{\boldsymbol{x}}N_{\boldsymbol{\xi}}$	31,104	108,864	290,304	653,184	1,306,368

Below, we report preconditioned MINRES iteration counts for the model problem and investigate the robustness of all the preconditioners discussed, with respect to d and σ_G. All experiments were performed in MATLAB 7.5 and the stopping criterion for the iteration was a reduction of the Euclidean norm of the preconditioned relative residual error to $tol = 10^{-8}$.

6.1. Schur complement preconditioning. First, we apply the preconditioners $\widehat{P}_0$, $\widehat{P}_1$, and $\widehat{P}_2$ from (4.11), (4.14) and (4.16), and the cheaper multigrid versions (4.19). The multigrid experiments were performed with a MATLAB version of the black-box AMG code `HSL_MI20` [7] using one pre and post smoothing step. Timings are reported in parentheses (in seconds) and include set-up.

As expected, $\widehat{P}_0$ is not robust with respect to variations in d and σ_G. Replacing $L = I$ with $L = G$ and $\mathrm{diag}(G)$, saves a significant number of iterations when d and σ_G are large. For $d = 4$ and

TABLE 6.2
Iteration counts (and timings) for exact and multigrid versions of Schur-complement preconditioners

		exact version				multigrid version			
	σ_G	d=1	d=2	d=3	d=4	d=1	d=2	d=3	d=4
$\widehat{P}_0$	0.2	45	53	61	69	47 (2)	57 (22)	65 (245)	74 (2979)
	0.4	57	83	114	148	61 (3)	89 (35)	121 (534)	148 (6321)
	0.6	72	129	210	320	77 (3)	139 (53)	225 (1369)	345 (13794)
	0.8	93	204	397	698	99 (4)	219 (84)	425 (2604)	747 (29808)
	1.0	118	316	730	1489	126 (5)	339 (129)	785 (4719)	1597 (63911)
$\widehat{P}_1$	0.2	37	40	43	46	40 (2)	42 (13)	45 (169)	48 (1291)
	0.4	43	50	58	65	45 (2)	53 (15)	60 (229)	68 (1688)
	0.6	47	63	78	95	51 (2)	66 (18)	81 (307)	99 (2553)
	0.8	54	80	108	141	56 (2)	82 (24)	111 (419)	145 (3651)
	1.0	61	98	146	203	64 (3)	102 (28)	152 (575)	210 (5177)
$\widehat{P}_2$	0.2	45	53	60	67	47 (2)	55 (15)	64 (235)	72 (1799)
	0.4	56	78	103	131	58 (2)	83 (22)	111 (411)	140 (3477)
	0.6	70	114	175	252	73 (3)	123 (32)	185 (689)	266 (6645)
	0.8	84	165	292	474	90 (3)	175 (44)	307 (1142)	496 (12435)
	1.0	100	234	476	865	107 (4)	248 (62)	495 (1843)	892 (17356)

$\sigma_G = 1.0$, using $\widehat{P}_1$ in place of $\widehat{P}_0$ saves 16 hours of computation time! Although the cost per iteration is higher for $\widehat{P}_1$ than $\widehat{P}_2$, the number of iterations saved with $\widehat{P}_1$ is substantial and increases with d. The deficiency of *all* the Schur-complement preconditioners is that they rely on cheap approximations to both $\widehat{A}$ *and* $\widehat{A}^{-1}$. Here, approximating $\widehat{A}$ with $L \otimes \mathrm{diag}(A_0)$ provides a cheap preconditioner that can be implemented with known deterministic algorithms. The approximation is optimal with respect to h, μ_G and M but the approximation is just too weak with respect to d and σ_G.

6.2. Augmented preconditioning. Next, we apply the preconditioners from Section 5. First we apply the ideal preconditioner $\widehat{P}_{div}$ with $\gamma = 10^{-3}$. The results in Table 6.3 confirm the result from Theorem 5.1, namely that the eigenvalues of $\widehat{P}_{div}^{-1}\widehat{C}$ are clustered at ± 1 when γ is small enough.

TABLE 6.3
MINRES *iteration counts for exact version of the full preconditioner* $\widehat{P}_{div}$.

d	$\sigma_G = 0.2$	$\sigma_G = 0.4$	$\sigma_G = 0.6$	$\sigma_G = 0.8$	$\sigma_G = 1.0$
1	3	3	3	3	3
2	3	3	3	3	3
3	3	3	3	3	3

Fixing $\gamma = 10^{-3}$ we apply the preconditioners $\widehat{P}_{0,div}$, $\widehat{P}_{1,div}$, and $\widehat{P}_{2,div}$ defined in (5.25), (5.26) and (5.28) and the corresponding multigrid versions (5.30). Choosing a smaller value of γ leads to smaller iteration counts but can skew the norm in which the preconditioned MINRES iteration is converging. Timings are reported in parentheses (in seconds) and include set-up time. The $H(div)$ multigrid method we have applied is from [1] and is best implemented in parallel. Our experiments were performed in serial. To minimize computing times, the multigrid preconditioners are applied with d pre and post smoothing steps per V-cycle.

We observe, as it typical with augmented preconditioners, that iteration counts are lower than for the Schur-complement preconditioners. We don't need to approximate A_0 by a diagonal matrix and

TABLE 6.4
Iteration counts (and timings) for exact and multigrid versions of augmented preconditioners

		exact version				multigrid version			
	σ_G	d=1	d=2	d=3	d=4	$d=1$	d=2	d=3	d=4
$\widehat{P}_{0,div}$	0.2	6	8	9	10	24 (4)	19 (11)	16 (45)	16 (398)
	0.4	7	12	15	18	28 (4)	24 (12)	24 (67)	27 (670)
	0.6	8	15	21	27	33 (4)	32 (16)	35 (97)	43 (712)
	0.8	10	17	28	44	39 (5)	42 (25)	51 (165)	67 (1521)
	1.0	12	20	41	69	46 (6)	54 (31)	72 (288)	100 (1636)
$\widehat{P}_{1,div}$	0.2	6	7	8	8	21 (3)	15 (10)	13 (54)	13 (220)
	0.4	8	10	11	12	24 (4)	19 (12)	18 (74)	17 (430)
	0.6	9	13	15	17	26 (4)	21 (13)	21 (86)	23 (579)
	0.8	10	16	19	22	28 (4)	24 (14)	26 (107)	29 (737)
	1.0	12	18	23	28	31 (4)	29 (17)	33 (134)	31 (973)
$\widehat{P}_{2,div}$	0.2	7	8	9	10	24 (4)	19 (12)	16 (65)	16 (402)
	0.4	8	12	15	17	28 (4)	24 (14)	24 (98)	26 (652)
	0.6	9	16	21	26	32 (5)	31 (18)	34 (137)	41 (675)
	0.8	11	19	32	43	37 (6)	40 (23)	49 (197)	61 (993)
	1.0	13	25	43	66	43 (6)	52 (31)	68 (273)	94 (1536)

results can always be tuned by changing the augmentation parameter γ. Although the underlying approximations to $\widehat{A}$ are still weak, the impact is reduced. Ultimately, no choice of L we have found yields an optimal approximation to $\widehat{A}$ with respect to d and σ_G. However, using $\widehat{P}_{1,div}$ leads to significant computational savings compared to the mean-based preconditioner $\widehat{P}_{0,div}$ when σ_G and d are large. The savings are more moderate than for $\widehat{P}_1$ however since $\widehat{P}_{0,div}$ leads to far lower iteration counts than $\widehat{P}_0$. This time, no savings are achieved with $\widehat{P}_{2,div}$.

7. Conclusions. We have analyzed preconditioners of Schur complement and augmented type for saddle point systems arising from mixed finite element Galerkin discretizations of second-order elliptic PDEs with random, lognormally distributed coefficients. We suggested improvements to mean-based preconditioners based on best Kronecker product approximation. Spectral inclusion bounds for the preconditioned Galerkin matrix reveal that none of the preconditioners are optimal with respect to d, the degree of polynomials used to construct S_d or the standard deviation σ_G of the underlying Gaussian random field. However, the Kronecker product preconditioners $\widehat{P}_1$ and $\widehat{P}_{1,div}$ are far more robust with respect to those parameters than $\widehat{P}_0$ and $\widehat{P}_{0,div}$. The augmented preconditioners yield lower MINRES iteration counts than the Schur complement preconditioners. On the other hand, the Schur complement preconditioners are parameter-free, and require only a fast solver for deterministic elliptic problems. The augmented preconditioners contain a parameter which needs tuning and rely on more specialised multigrid techniques.

Uncertainty quantification is becoming an increasingly important aspect of mathematical modelling. However, the study of efficient linear algebra techniques for the systems of equations that arise from stochastic mixed finite element methods is in its infancy. Motivated by deterministic saddle point preconditioners, this work highlights a need for more sophisticated solution strategies.

REFERENCES

[1] DOUGLAS N. ARNOLD, RICHARD S. FALK, AND R. WINTHER, *Preconditioning in H*(div) *and applications*, Math. Comp., 66 (1997), pp. 957–984.

[2] ———, *Multigrid in H(div) and H(curl)*, Numer. Math., 85 (2000), pp. 197–217.
[3] I. Babuška, F. Nobile, and R. Tempone, *A stochastic collocation method for elliptic partial differential equations with random input data*, SIAM J. Numer. Anal., 45 (2007), pp. 1005–1034.
[4] I. Babuška, R. Tempone, and G. E. Zouraris, *Galerkin finite element approximations of stochastic elliptic partial differential equations*, SIAM J. Numer. Math., 42 (2004), pp. 800–825.
[5] ———, *Solving elliptic boundary value problems with uncertain coefficients by the finite element method: the stochastic formulation*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 1251–1294.
[6] M. Benzi, G. H. Golub, and J. Liesen, *Numerical solution of saddle point problems*, Acta Numer., 14 (2005), pp. 1–137.
[7] J. Boyle, M. D. Mihajlović, and J. A. Scott, *HSL_MI20: an efficient AMG preconditioner*, Tech. Report RAL-TR-2007-021, SFTC Rutherford Appleton Laboratory, Didcot, December 2007.
[8] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
[9] 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.
[10] H. Elman, D. Silvester, and A. Wathen, *Finite Elements and Fast Iterative Solvers*, Oxford University Press, Oxford, 2005.
[11] O. G. Ernst, C. E. Powell, D. J. Silvester, and E. Ullmann, *Efficient solvers for a linear stochastic Galerkin mixed formulation of diffusion problems with random data*, SIAM J. Sci. Comput., 31 (2009), pp. 1424–1447.
[12] O. G. Ernst and E. Ullmann, *Stochastic Galerkin matrices*, To appear in SIAM J. Matrix Anal. Appl., 2009.
[13] P. Frauenfelder, C. Schwab, and R. A. Todor, *Finite elements for elliptic problems with stochastic coefficients*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 205–228.
[14] A. Freeze, *A stochastic-conceptual analysis of one-dimensional groundwater flow in nonuniform homogeneous media*, Water Resour. Res., 11 (1975), pp. 725–740.
[15] D. G. Furnival, H. C. Elman, and C. E. Powell, *H(div) preconditioning for a mixed finite element formulation of the stochastic diffusion problem*, Tech. Report CS-TR-4918, University of Maryland Department of Computer Science, 2008. To appear in Math. Comp., 2009.
[16] C. Greif and D. Schötzau, *Preconditioners for the discretized time-harmonic maxwell equations in mixed form*, Numer. Linear. Algebra Appl, 14 (2007), pp. 281–297.
[17] R. Hiptmair, *Multigrid method for H(div) in three dimensions*, Electron. Trans. Numer. Anal., 6 (1997), pp. 133–152.
[18] J. Indritz, *An inequality for Hermite polynomials*, Proc. Amer. Math. Soc., 12 (1961), pp. 981–983.
[19] S. Janson, *Gaussian Hilbert spaces*, Cambridge University Press, Cambridge, 1997.
[20] I. Krasikov, *Nonnegative quadratic forms and bounds on orthogonal polynomials*, J. Approx. Theory, 111 (2001), pp. 31–49.
[21] M. Loève, *Probability Theory*, vol. II, Springer-Verlag, New York - Heidelberg - Berlin, 4th ed., 1978.
[22] P. Malliavin, *Stochastic Analysis*, Springer-Verlag, Berlin, 1997.
[23] H. G. Matthies and A. Keese, *Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 1295–1331.
[24] M. F. Murphy, G. H. Golub, and A. J. Wathen, *A note on preconditioning for indefinite linear systems*, SIAM J. Sci. Comput., 21 (2000), pp. 1969–1972.
[25] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.
[26] I. Perugia and V. Simoncini, *Block-diagonal and indefinite symmetric preconditioners for mixed finite element formulations*, Numer. Linear. Algebra Appl, 7 (2000), pp. 585–616.
[27] C. E. Powell and H. C. Elman, *Block-diagonal preconditioning for spectral stochastic finite-element systems*, IMA J. Numer. Anal., 29 (2009), pp. 350–375.
[28] C. E. Powell and D. Silvester, *Optimal preconditioning for Raviart-Thomas mixed formulation of second-order elliptic problems*, SIAM J. Matrix Anal. Appl., 25 (2004), pp. 718–738.
[29] P.-A. Raviart and J. M. Thomas, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of the Finite Element Method, I. Galligani and E. Magenes, eds., no. 606 in Lect. Notes in Math., Springer-Verlag, New York, 1977, pp. 292–315.
[30] E. Rosseel and S. Vandewalle, *Iterative solvers for the stochastic finite element method*, To appear in SIAM J. Sci. Comput., 2009.
[31] C. Schwab and R. A. Todor, *Karhunen-Loève approximation of random fields by generalized fast multipole methods*, J. Comput. Phys., 127 (2006), pp. 100–122.
[32] E. Ullmann, *A Kronecker product preconditioner for stochastic Galerkin finite element discretizations*, Tech. Report 04-2008, Fakultät für Mathematik und Informatik, Technische Universität Bergakademie Freiberg, 2008.
[33] C. F. Van Loan and N. Pitsianis, *Approximation with Kronecker products*, in Linear algebra for large scale and real- time applications. Proceedings of the NATO Advanced Study Institute, Leuven, Belgium, August 3 - 14, 1992., Marc S. Moonen, Gene H. Golub, and Bart L. R. de Moor, eds., Kluwer Academic Publishers, Dordrecht, 1993, pp. 293–314.

[34] P. S. Vassilevski and R. Lazarov, *Preconditioning mixed finite element saddle-point elliptic problems*, Numer. Linear. Algebra Appl, 3 (1996), pp. 1–20.
[35] D. Xiu and G. E. Karniadakis, *The Wiener-Askey polynomial chaos for stochastic differential equations*, SIAM J. Sci. Comput., 24 (2002), pp. 619–644.

Preprint Series DFG-SPP 1324

http://www.dfg-spp1324.de

Reports

[1] R. Ramlau, G. Teschke, and M. Zhariy. A Compressive Landweber Iteration for Solving Ill-Posed Inverse Problems. Preprint 1, DFG-SPP 1324, September 2008.

[2] G. Plonka. The Easy Path Wavelet Transform: A New Adaptive Wavelet Transform for Sparse Representation of Two-dimensional Data. Preprint 2, DFG-SPP 1324, September 2008.

[3] E. Novak and H. Woźniakowski. Optimal Order of Convergence and (In-) Tractability of Multivariate Approximation of Smooth Functions. Preprint 3, DFG-SPP 1324, October 2008.

[4] M. Espig, L. Grasedyck, and W. Hackbusch. Black Box Low Tensor Rank Approximation Using Fibre-Crosses. Preprint 4, DFG-SPP 1324, October 2008.

[5] T. Bonesky, S. Dahlke, P. Maass, and T. Raasch. Adaptive Wavelet Methods and Sparsity Reconstruction for Inverse Heat Conduction Problems. Preprint 5, DFG-SPP 1324, January 2009.

[6] E. Novak and H. Woźniakowski. Approximation of Infinitely Differentiable Multivariate Functions Is Intractable. Preprint 6, DFG-SPP 1324, January 2009.

[7] J. Ma and G. Plonka. A Review of Curvelets and Recent Applications. Preprint 7, DFG-SPP 1324, February 2009.

[8] L. Denis, D. A. Lorenz, and D. Trede. Greedy Solution of Ill-Posed Problems: Error Bounds and Exact Inversion. Preprint 8, DFG-SPP 1324, April 2009.

[9] U. Friedrich. A Two Parameter Generalization of Lions' Nonoverlapping Domain Decomposition Method for Linear Elliptic PDEs. Preprint 9, DFG-SPP 1324, April 2009.

[10] K. Bredies and D. A. Lorenz. Minimization of Non-smooth, Non-convex Functionals by Iterative Thresholding. Preprint 10, DFG-SPP 1324, April 2009.

[11] K. Bredies and D. A. Lorenz. Regularization with Non-convex Separable Constraints. Preprint 11, DFG-SPP 1324, April 2009.

[12] M. Döhler, S. Kunis, and D. Potts. Nonequispaced Hyperbolic Cross Fast Fourier Transform. Preprint 12, DFG-SPP 1324, April 2009.

[13] C. Bender. Dual Pricing of Multi-Exercise Options under Volume Constraints. Preprint 13, DFG-SPP 1324, April 2009.

[14] T. Müller-Gronbach and K. Ritter. Variable Subspace Sampling and Multi-level Algorithms. Preprint 14, DFG-SPP 1324, May 2009.

[15] G. Plonka, S. Tenorth, and A. Iske. Optimally Sparse Image Representation by the Easy Path Wavelet Transform. Preprint 15, DFG-SPP 1324, May 2009.

[16] S. Dahlke, E. Novak, and W. Sickel. Optimal Approximation of Elliptic Problems by Linear and Nonlinear Mappings IV: Errors in L_2 and Other Norms. Preprint 16, DFG-SPP 1324, June 2009.

[17] B. Jin, T. Khan, P. Maass, and M. Pidcock. Function Spaces and Optimal Currents in Impedance Tomography. Preprint 17, DFG-SPP 1324, June 2009.

[18] G. Plonka and J. Ma. Curvelet-Wavelet Regularized Split Bregman Iteration for Compressed Sensing. Preprint 18, DFG-SPP 1324, June 2009.

[19] G. Teschke and C. Borries. Accelerated Projected Steepest Descent Method for Nonlinear Inverse Problems with Sparsity Constraints. Preprint 19, DFG-SPP 1324, July 2009.

[20] L. Grasedyck. Hierarchical Singular Value Decomposition of Tensors. Preprint 20, DFG-SPP 1324, July 2009.

[21] D. Rudolf. Error Bounds for Computing the Expectation by Markov Chain Monte Carlo. Preprint 21, DFG-SPP 1324, July 2009.

[22] M. Hansen and W. Sickel. Best m-term Approximation and Lizorkin-Triebel Spaces. Preprint 22, DFG-SPP 1324, August 2009.

[23] F.J. Hickernell, T. Müller-Gronbach, B. Niu, and K. Ritter. Multi-level Monte Carlo Algorithms for Infinite-dimensional Integration on $\mathbb{R}^{\mathbb{N}}$. Preprint 23, DFG-SPP 1324, August 2009.

[24] S. Dereich and F. Heidenreich. A Multilevel Monte Carlo Algorithm for Lévy Driven Stochastic Differential Equations. Preprint 24, DFG-SPP 1324, August 2009.

[25] S. Dahlke, M. Fornasier, and T. Raasch. Multilevel Preconditioning for Adaptive Sparse Optimization. Preprint 25, DFG-SPP 1324, August 2009.

[26] S. Dereich. Multilevel Monte Carlo Algorithms for Lévy-driven SDEs with Gaussian Correction. Preprint 26, DFG-SPP 1324, August 2009.

[27] G. Plonka, S. Tenorth, and D. Roşca. A New Hybrid Method for Image Approximation using the Easy Path Wavelet Transform. Preprint 27, DFG-SPP 1324, October 2009.

[28] O. Koch and C. Lubich. Dynamical Low-rank Approximation of Tensors. Preprint 28, DFG-SPP 1324, November 2009.

[29] E. Faou, V. Gradinaru, and C. Lubich. Computing Semi-classical Quantum Dynamics with Hagedorn Wavepackets. Preprint 29, DFG-SPP 1324, November 2009.

[30] D. Conte and C. Lubich. An Error Analysis of the Multi-configuration Time-dependent Hartree Method of Quantum Dynamics. Preprint 30, DFG-SPP 1324, November 2009.

[31] C. E. Powell and E. Ullmann. Preconditioning Stochastic Galerkin Saddle Point Problems. Preprint 31, DFG-SPP 1324, November 2009.