SIAM J. SCI. COMPUT.
Vol. 41, No. 1, pp. A402–A421

ROBUST PRECONDITIONING FOR STOCHASTIC GALERKIN FORMULATIONS OF PARAMETER-DEPENDENT NEARLY INCOMPRESSIBLE ELASTICITY EQUATIONS*

ARBAZ KHAN†, CATHERINE E. POWELL†, AND DAVID J. SILVESTER†

Abstract. We consider the nearly incompressible linear elasticity problem with an uncertain spatially varying Young's modulus. The uncertainty is modeled with a finite set of parameters with prescribed probability distribution. We introduce a novel three-field mixed variational formulation of the PDE model and discuss its approximation by stochastic Galerkin mixed finite element techniques. First, we establish the well-posedness of the proposed variational formulation and the associated finite-dimensional approximation. Second, we focus on the efficient solution of the associated large and indefinite linear system of equations. A new preconditioner is introduced for use with the minimal residual method. Eigenvalue bounds for the preconditioned system are established and shown to be independent of the discretization parameters and the Poisson ratio. The S-IFISS software used for computation is available online.

Key words. uncertain material parameters, linear elasticity, mixed approximation, stochastic Galerkin finite element method, preconditioning

AMS subject classifications. 65N30, 65F08, 35R60

DOI. 10.1137/18M117385X

1. Introduction.

The locking of finite element approximations when solving nearly incompressible elasticity problems is a significant issue in the computational engineering world. The standard way of preventing locking is to write the underlying equations as a system and introduce pressure as an additional unknown [9, 11]. Thus, the starting point for this work is the *Herrmann* formulation of linear elasticity

$$-\nabla \cdot \boldsymbol{\sigma} = \boldsymbol{f} \quad \text{in } D, \tag{1.1a}$$

$$\nabla \cdot \boldsymbol{u} + \frac{p}{\lambda} = 0 \quad \text{in } D, \tag{1.1b}$$

where D is a bounded Lipschitz polygon in $\mathbb{R}^2$ (polyhedral in $\mathbb{R}^3$). In this setting, the elastic deformation of the isotropic solid is defined in terms of the stress tensor $\boldsymbol{\sigma}$, the body force $\boldsymbol{f}$, the displacement field $\boldsymbol{u}$, and the Herrmann pressure p (auxiliary variable). The stress tensor is related to the strain tensor $\boldsymbol{\varepsilon}$ through the identities

$$\boldsymbol{\sigma} = 2\mu\boldsymbol{\varepsilon} - p\boldsymbol{I}, \qquad \boldsymbol{\varepsilon} = \frac{1}{2}\left(\nabla\boldsymbol{u} + (\nabla\boldsymbol{u})^\top\right).$$

The Lamé coefficients μ and λ satisfy $0 < \mu_1 < \mu < \mu_2 < \infty$ and $0 < \lambda < \infty$ and can be defined in terms of the Young's modulus E and the Poisson ratio ν via

$$\mu = \frac{E}{2(1+\nu)}, \quad \lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}.$$

*Submitted to the journal's Methods and Algorithms for Scientific Computing section March 5, 2018; accepted for publication (in revised form) October 30, 2018; published electronically February 5, 2019.

http://www.siam.org/journals/sisc/41-1/M117385.html

Funding: This work was supported by EPSRC grant EP/P013317/1. Part of this work was undertaken at the Isaac Newton Institute for Mathematical Sciences, Cambridge, during the Uncertainty Quantification programme, supported by EPSRC grant EP/K032208/1.

†School of Mathematics, University of Manchester, Manchester, M13 9PL, UK (arbaz.khan@manchester.ac.uk, c.powell@manchester.ac.uk, d.silvester@manchester.ac.uk).

Our focus is on uncertainty quantification. Specifically, we consider the case where the properties of the elastic material are varying spatially in an uncertain way. For example, this may be due to material imperfections or inaccurate measurements. To account for this uncertainty we model the Young's modulus E as a spatially varying random field. More precisely, we introduce a vector $\boldsymbol{y} = (y_1, \ldots, y_M)$ of parameters, with each $y_k \in \Gamma_k = [-1, 1]$, and represent E as an affine combination

$$E(\boldsymbol{x}, \boldsymbol{y}) := e_0(\boldsymbol{x}) + \sum_{k=1}^{M} e_k(\boldsymbol{x}) y_k, \quad \boldsymbol{x} \in D, \boldsymbol{y} \in \Gamma, \tag{1.2}$$

where $\Gamma = \Gamma_1 \times \cdots \times \Gamma_M \subset \mathbb{R}^M$ is our parameter domain. Such representations arise, for example, from truncated Karhunen–Loève expansions of second-order random fields. In (1.2), $e_0(\boldsymbol{x})$ typically represents the mean, and $e_k(\boldsymbol{x})y_k$ is a perturbation away from the mean. The resulting parameter-dependent problem is given by

$$-\nabla \cdot \boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{f}(\boldsymbol{x}), \quad \boldsymbol{x} \in D,\ \boldsymbol{y} \in \Gamma, \tag{1.3a}$$

$$\nabla \cdot \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{y}) + \frac{p(\boldsymbol{x}, \boldsymbol{y})}{\lambda(\boldsymbol{x}, \boldsymbol{y})} = 0, \qquad \boldsymbol{x} \in D,\ \boldsymbol{y} \in \Gamma, \tag{1.3b}$$

$$\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{g}(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial D_D,\ \boldsymbol{y} \in \Gamma, \tag{1.3c}$$

$$\boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{y})\boldsymbol{n} = \boldsymbol{0}, \qquad \boldsymbol{x} \in \partial D_N,\ \boldsymbol{y} \in \Gamma, \tag{1.3d}$$

where the boundary of the spatial domain is $\partial D = \partial D_D \cup \partial D_N$ with $\partial D_D \cap \partial D_N = \emptyset$ and $\partial D_D, \partial D_N \neq \emptyset$, the stress tensor is $\boldsymbol{\sigma} : D \times \Gamma \to \mathbb{R}^{d \times d}$ $(d = 2, 3)$, the strain tensor is $\boldsymbol{\varepsilon} : D \times \Gamma \to \mathbb{R}^{d \times d}$, the body force is $\boldsymbol{f} : D \to \mathbb{R}^d$, the displacement field is $\boldsymbol{u} : D \times \Gamma \to \mathbb{R}^d$, and the Herrmann pressure is $p : D \times \Gamma \to \mathbb{R}$. The Lamé coefficients are also parameter-dependent and spatially varying,

$$\mu(\boldsymbol{x}, \boldsymbol{y}) = \frac{E(\boldsymbol{x}, \boldsymbol{y})}{2(1+\nu)}, \quad \lambda(\boldsymbol{x}, \boldsymbol{y}) = \frac{E(\boldsymbol{x}, \boldsymbol{y})\nu}{(1+\nu)(1-2\nu)}.$$

Note that we assume that ν is a given fixed constant and that $0 < \mu_1 < \mu < \mu_2 < \infty$ and $0 < \lambda < \infty$ a.e. in $D \times \Gamma$.

Stochastic Galerkin finite element methods (SGFEMs) are a popular way of approximating solutions to parameter-dependent PDEs. Broadly speaking, we seek approximate solutions in tensor product spaces of the form $X_h \otimes S_\Lambda$, where X_h is an appropriate finite element space associated with a subdivision of D and S_Λ is, typically, a set of multivariate polynomials that are globally defined on the parameter domain Γ. This is a feasible strategy if the number of input parameters is modest, and the underlying solution is sufficiently smooth as a function of those parameters. We refer to Babuška, Tempone, and Zouraris [1] for a priori error estimates for SGFEM approximations of solutions to elliptic PDEs with parameter-dependent coefficients and Bespalov, Powell, and Silvester [2] for a priori error estimates for SGFEM approximations of solutions to mixed formulations of elliptic PDEs with parameter-dependent coefficients. A posteriori error analysis of linear elasticity with parameter-dependent coefficients is considered by Eigel et al. [4]. Crucial to the efficient implementation of SGFEMs is the need to separate the terms that depend on $\boldsymbol{x}$ from the terms that depend on $\boldsymbol{y}$ in the weak formulation of the problem. Here, since both μ and $1/\lambda$ appear in the PDE model (1.3), both E and E^{-1} appear in the formulation.

To address this difficulty, our idea here is to introduce a second auxiliary variable $\tilde{p} = p/E$ to give a distinctive three-field mixed formulation of (1.3): find $\boldsymbol{u} : D \times \Gamma \to$

$\mathbb{R}^d$ and $p, \tilde{p} : D \times \Gamma \to \mathbb{R}$ such that

$$-\nabla \cdot \boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{f}(\boldsymbol{x}), \quad \boldsymbol{x} \in D,\ \boldsymbol{y} \in \Gamma, \tag{1.4a}$$

$$\nabla \cdot \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{y}) + \tilde{\lambda}^{-1} \tilde{p}(\boldsymbol{x}, \boldsymbol{y}) = 0, \quad \boldsymbol{x} \in D,\ \boldsymbol{y} \in \Gamma, \tag{1.4b}$$

$$\tilde{\lambda}^{-1} p(\boldsymbol{x}, \boldsymbol{y}) - \tilde{\lambda}^{-1} E(\boldsymbol{x}, \boldsymbol{y}) \tilde{p}(\boldsymbol{x}, \boldsymbol{y}) = 0, \quad \boldsymbol{x} \in D,\ \boldsymbol{y} \in \Gamma, \tag{1.4c}$$

$$\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{g}(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial D_D,\ \boldsymbol{y} \in \Gamma, \tag{1.4d}$$

$$\boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{y}) \boldsymbol{n} = \boldsymbol{0}, \quad \boldsymbol{x} \in \partial D_N,\ \boldsymbol{y} \in \Gamma, \tag{1.4e}$$

where

$$\tilde{\lambda} = \frac{\lambda(\boldsymbol{x}, \boldsymbol{y})}{E(\boldsymbol{x}, \boldsymbol{y})} = \frac{\nu}{(1+\nu)(1-2\nu)},$$

is now a fixed constant. The advantage of (1.4) is that while E appears in the first and third equations, E^{-1} does not appear at all. As a result, the discrete problem associated with our SGFEM approximation has a structure that is relatively easy to exploit. This is a novel solution strategy and gives this work a distinctive edge.

The rest of the paper is organized as follows. Section 2 introduces a weak formulation of (1.4) and discusses well-posedness. In particular, a stability result is established with respect to a coefficient-dependent norm that is a generalization of the natural norm identified in our earlier work [11]. Section 3 introduces the finite-dimensional problem associated with SGFEM approximation and gives details of the associated linear algebra system that needs to be solved when computing the Galerkin solution. A novel preconditioner is introduced in section 4 and bounds for the eigenvalues of the preconditioned system are established. The preconditioning strategy is consistent with the philosophy of Mardal and Winther [15]: the diagonal blocks of the preconditioning matrix are associated with the norm for which the stability of the mixed approximation has been established. Finally, we present numerical results in section 5 to illustrate the efficiency and robustness when representative discrete problems are solved using the minimal residual method.

2. Weak formulation. First, we need to impose some conditions on the model inputs and define appropriate solution spaces. Recall that E is defined as in (1.2), where $\Gamma = \Gamma_1 \times \cdots \times \Gamma_M \subset \mathbb{R}^M$ is the parameter domain, and $\Gamma_k = [-1, 1]$.

Assumption 2.1. The random field $E \in L^\infty(D \times \Gamma)$ is uniformly bounded away from zero, i.e., there exist positive constants $E_{\min}$ and $E_{\max}$ such that

$$0 < E_{\min} \leq E(\boldsymbol{x}, \boldsymbol{y}) \leq E_{\max} < \infty \quad \text{a.e. in } D \times \Gamma. \tag{2.1}$$

To identify the lower bound, it will be convenient to further assume that

$$0 < e_0^{\min} \leq e_0(\boldsymbol{x}) \leq e_0^{\max} < \infty \quad \text{a.e. in } D \quad \text{and} \quad \frac{1}{e_0^{\min}} \sum_{k=1}^{M} ||e_k||_{L^\infty(D)} < 1. \tag{2.2}$$

Let $\pi(\boldsymbol{y})$ be a product measure with $\pi(\boldsymbol{y}) := \Pi_{k=1}^M \pi_k(y_k)$, where π_k denotes a measure on $(\Gamma_k, \mathcal{B}(\Gamma_k))$ and $\mathcal{B}(\Gamma_k)$ is the Borel σ-algebra on Γ_k. We will assume that the parameters y_k in (1.2) are images of independent mean zero uniform random variables on $[-1, 1]$ and choose π_k to be the associated probability measure. Now we can define the Bochner space

$$L^2_\pi(\Gamma, X(D)) := \left\{ v(\boldsymbol{x}, \boldsymbol{y}) : D \times \Gamma \to \mathbb{R}; ||v||_{L^2_\pi(\Gamma, X(D))} < \infty \right\},$$

where $X(D)$ is a normed vector space of real-valued functions on D with norm $||\cdot||_X$ and

$$||\cdot||_{L^2_\pi(\Gamma,X(D))} := \left(\int_\Gamma ||\cdot||_X^2 d\pi(\boldsymbol{y})\right)^{1/2}. \tag{2.3}$$

In our analysis, we will need the spaces

$$\boldsymbol{\mathcal{V}} := L^2_\pi(\Gamma, \boldsymbol{H}^1_{E_0}(D)), \quad \mathcal{W} := L^2_\pi(\Gamma, L^2(D)) \quad \text{and} \quad \boldsymbol{\mathcal{W}} := L^2_\pi(\Gamma, \boldsymbol{L}^2(D)),$$

where $\boldsymbol{H}^1_{E_0}(D) = \{\boldsymbol{v} \in \boldsymbol{H}^1(D), \boldsymbol{v}|_{\partial D_D} = \boldsymbol{0}\}$ and $\boldsymbol{H}^1(D) = \boldsymbol{H}^1(D;\mathbb{R}^d)$ is the usual vector-valued Sobolev space with associated norm $||\cdot||_1$. We assume that the load function satisfies $\boldsymbol{f} \in (L^2(D))^d$, and for simplicity, we choose $\boldsymbol{g} = \boldsymbol{0}$ on ∂D_D. In that case, the weak formulation of (1.4) is to find $(\boldsymbol{u}, p, \tilde{p}) \in \boldsymbol{\mathcal{V}} \times \mathcal{W} \times \mathcal{W}$ such that

$$a(\boldsymbol{u},\boldsymbol{v}) + b(\boldsymbol{v},p) = f(\boldsymbol{v}) \quad \forall \boldsymbol{v} \in \boldsymbol{\mathcal{V}}, \tag{2.4a}$$

$$b(\boldsymbol{u},q) - c(\tilde{p},q) = 0 \qquad \forall q \in \mathcal{W}, \tag{2.4b}$$

$$-c(p,\tilde{q}) + d(\tilde{p},\tilde{q}) = 0 \qquad \forall \tilde{q} \in \mathcal{W}. \tag{2.4c}$$

Here, we have

$$a(\boldsymbol{u},\boldsymbol{v}) := \alpha \int_\Gamma \int_D E(\boldsymbol{x},\boldsymbol{y})\boldsymbol{\varepsilon}(\boldsymbol{u}(\boldsymbol{x},\boldsymbol{y})) : \boldsymbol{\varepsilon}(\boldsymbol{v}(\boldsymbol{x},\boldsymbol{y}))d\boldsymbol{x}d\pi(\boldsymbol{y}), \tag{2.5}$$

$$b(\boldsymbol{v},p) := -\int_\Gamma \int_D p(\boldsymbol{x},\boldsymbol{y})\mathrm{div}\,\boldsymbol{v}(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}d\pi(\boldsymbol{y}), \tag{2.6}$$

$$c(p,q) := (\alpha\beta)^{-1} \int_\Gamma \int_D p(\boldsymbol{x},\boldsymbol{y})q(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}d\pi(\boldsymbol{y}), \tag{2.7}$$

$$d(p,q) := (\alpha\beta)^{-1} \int_\Gamma \int_D E(\boldsymbol{x},\boldsymbol{y})p(\boldsymbol{x},\boldsymbol{y})q(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}d\pi(\boldsymbol{y}), \tag{2.8}$$

$$f(\boldsymbol{v}) := \int_\Gamma \int_D f(\boldsymbol{x})\boldsymbol{v}(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}d\pi(\boldsymbol{y}) \tag{2.9}$$

with

$$\alpha := \frac{1}{1+\nu}, \qquad \beta := \frac{\nu}{(1-2\nu)}. \tag{2.10}$$

Note that α and β depend on the Poisson ratio ν but are fixed constants. Following convention, we will also define the bilinear form

$$\mathcal{B}(\boldsymbol{u},p,\tilde{p};\boldsymbol{v},q,\tilde{q}) = a(\boldsymbol{u},\boldsymbol{v}) + b(\boldsymbol{v},p) + b(\boldsymbol{u},q) - c(\tilde{p},q) - c(p,\tilde{q}) + d(\tilde{p},\tilde{q}), \tag{2.11}$$

so as to express (2.4) in the compact form: find $(\boldsymbol{u}, p, \tilde{p}) \in \boldsymbol{\mathcal{V}} \times \mathcal{W} \times \mathcal{W}$ such that

$$\mathcal{B}(\boldsymbol{u},p,\tilde{p};\boldsymbol{v},q,\tilde{q}) = f(\boldsymbol{v}) \quad \forall (\boldsymbol{v},q,\tilde{q}) \in \boldsymbol{\mathcal{V}} \times \mathcal{W} \times \mathcal{W}. \tag{2.12}$$

The next result establishes that the four bilinear forms appearing in (2.4) and hence the bilinear form $\mathcal{B}(\cdot,\cdot)$ in (2.12) are bounded.

LEMMA 2.1. *If E satisfies Assumption* 2.1*, then the following bounds hold:*

$$a(\boldsymbol{u},\boldsymbol{v}) \le \alpha E_{\max}||\nabla\boldsymbol{u}||_{\boldsymbol{\mathcal{W}}}||\nabla\boldsymbol{v}||_{\boldsymbol{\mathcal{W}}} \quad \forall \boldsymbol{u},\boldsymbol{v} \in \boldsymbol{\mathcal{V}}, \tag{2.13}$$

$$b(\boldsymbol{u},p) \le \sqrt{d}\,||\nabla\boldsymbol{u}||_{\boldsymbol{\mathcal{W}}}||p||_{\mathcal{W}} \qquad \forall \boldsymbol{u} \in \boldsymbol{\mathcal{V}},\ \forall p \in \mathcal{W}, \tag{2.14}$$

$$c(p,q) \le (\alpha\beta)^{-1}||p||_{\mathcal{W}}||q||_{\mathcal{W}} \qquad \forall p,q \in \mathcal{W}, \tag{2.15}$$

$$d(p,q) \le (\alpha\beta)^{-1}E_{\max}||p||_{\mathcal{W}}||q||_{\mathcal{W}} \quad \forall p,q \in \mathcal{W}. \tag{2.16}$$

Proof. All bounds follow from the Cauchy–Schwarz inequality and (2.1). ∎

The next result establishes that three of the bilinear forms appearing in (2.4) and (2.12) are coercive and that an inf-sup condition involving $b(\cdot,\cdot)$ is satisfied.

LEMMA 2.2. *If Assumption* 2.1 *is valid, then the following bounds hold:*

$$a(\boldsymbol{u},\boldsymbol{u}) \geq \alpha E_{\min} C_K ||\nabla \boldsymbol{u}||^2_{\boldsymbol{\mathcal{W}}} \quad \forall \boldsymbol{u} \in \boldsymbol{\mathcal{V}}, \tag{2.17}$$

$$c(p,p) \geq (\alpha\beta)^{-1} ||p||^2_{\mathcal{W}} \quad \forall p \in \mathcal{W}, \tag{2.18}$$

$$d(p,p) \geq (\alpha\beta)^{-1} E_{\min} ||p||^2_{\mathcal{W}} \quad \forall p \in \mathcal{W}, \tag{2.19}$$

where $0 < C_K \leq 1$ *is the Korn constant. In addition, there exists an inf-sup constant* $C_D > 0$ *such that*

$$\sup_{0 \neq \boldsymbol{v} \in \boldsymbol{\mathcal{V}}} \frac{b(\boldsymbol{v},q)}{||\nabla \boldsymbol{v}||_{\boldsymbol{\mathcal{W}}}} \geq C_D ||q||_{\mathcal{W}} \quad \forall q \in \mathcal{W}. \tag{2.20}$$

Proof. The first bound follows by combining (2.1) with Korn's inequality. The second and third bounds follow directly from the definition of the bilinear forms and (2.1). We can use arguments such as in [3, Lemma 11.2.3], [2, Lemma 7.2], (and references therein for nonconvex domains D) as well as [7, section 4.1.4] (for more general boundary conditions) to show that for any $q \in \mathcal{W}$ there exists a $\boldsymbol{w} \in \boldsymbol{\mathcal{V}}$ such that $\operatorname{div} \boldsymbol{w} = q$ and $C_D ||\nabla \boldsymbol{w}||_{\boldsymbol{\mathcal{W}}} \leq ||q||_{\mathcal{W}}$, where C_D is positive constant. Thus

$$\sup_{0 \neq \boldsymbol{v} \in \boldsymbol{\mathcal{V}}} \frac{b(\boldsymbol{v},q)}{||\nabla \boldsymbol{v}||_{\boldsymbol{\mathcal{W}}}} \geq \frac{-b(\boldsymbol{w},q)}{||\nabla \boldsymbol{w}||_{\boldsymbol{\mathcal{W}}}} = \frac{||q||^2_{\mathcal{W}}}{||\nabla \boldsymbol{w}||_{\boldsymbol{\mathcal{W}}}} \geq C_D ||q||_{\mathcal{W}}. \qquad \blacksquare$$

To establish that our problem formulation is well posed, we now introduce a coefficient-dependent norm $|||\cdot|||$ on $\boldsymbol{\mathcal{V}} \times \mathcal{W} \times \mathcal{W}$, defined by

$$|||(\boldsymbol{v},q,\tilde{q})|||^2 := \alpha ||\nabla \boldsymbol{v}||^2_{\boldsymbol{\mathcal{W}}} + (\alpha^{-1} + (\alpha\beta)^{-1}) ||q||^2_{\mathcal{W}} + (\alpha\beta)^{-1} ||\tilde{q}||^2_{\mathcal{W}}. \tag{2.21}$$

The well-posedness of (2.12) is addressed in the next two results.

LEMMA 2.3. *If Assumption* 2.1 *is valid, then for any* $(\boldsymbol{u},p,\tilde{p}) \in \boldsymbol{\mathcal{V}} \times \mathcal{W} \times \mathcal{W}$*, there exists* $(\boldsymbol{v},q,\tilde{q}) \in \boldsymbol{\mathcal{V}} \times \mathcal{W} \times \mathcal{W}$ *with* $|||(\boldsymbol{v},q,\tilde{q})||| \leq C_2 |||(\boldsymbol{u},p,\tilde{p})|||$*, satisfying*

$$\mathcal{B}(\boldsymbol{u},p,\tilde{p};\boldsymbol{v},q,\tilde{q}) \geq E_{\min} C_1 |||(\boldsymbol{u},p,\tilde{p})|||^2, \tag{2.22}$$

where C_1 *and* C_2 *depend on* $E_{\max}$*,* C_K*, and* C_D*.*

Proof. From (2.12) we have

$$\begin{aligned}\mathcal{B}(\boldsymbol{u},p,\tilde{p};\boldsymbol{u},-p,\tilde{p}) &= a(\boldsymbol{u},\boldsymbol{u}) + b(\boldsymbol{u},p) + b(\boldsymbol{u},-p) - c(\tilde{p},-p) - c(p,\tilde{p}) + d(\tilde{p},\tilde{p}) \\ &= a(\boldsymbol{u},\boldsymbol{u}) + d(\tilde{p},\tilde{p}) =: |\boldsymbol{u}|^2_a + |\tilde{p}|^2_d.\end{aligned}$$

Now, as a consequence of (2.20), since $p \in \mathcal{W}$, there exists a $\boldsymbol{w} \in \boldsymbol{\mathcal{V}}$ such that

$$-b(\boldsymbol{w},p) \geq C_D \alpha^{-1} ||p||^2_{\mathcal{W}}, \quad \alpha^{1/2} ||\nabla \boldsymbol{w}||_{\boldsymbol{\mathcal{W}}} \leq \alpha^{-1/2} ||p||_{\mathcal{W}}. \tag{2.23}$$

Using this particular $\boldsymbol{w}$ in (2.11) and using Lemma 2.1, it follows that

$$\begin{aligned}\mathcal{B}(\boldsymbol{u},p,\tilde{p};-\boldsymbol{w},0,0) &= -b(\boldsymbol{w},p) - a(\boldsymbol{u},\boldsymbol{w}) \\ &\geq C_D\, \alpha^{-1} ||p||^2_{\mathcal{W}} - |\boldsymbol{u}|_a |\boldsymbol{w}|_a \\ &\geq C_D\, \alpha^{-1} ||p||^2_{\mathcal{W}} - |\boldsymbol{u}|_a\, E^{1/2}_{\max} \alpha^{1/2} ||\nabla \boldsymbol{w}||_{\boldsymbol{\mathcal{W}}} \\ &\geq C_D\, \alpha^{-1} ||p||^2_{\mathcal{W}} - |\boldsymbol{u}|_a\, E^{1/2}_{\max} \alpha^{-1/2} ||p||_{\mathcal{W}} \\ &\geq C_D\, \alpha^{-1} ||p||^2_{\mathcal{W}} - \frac{\epsilon}{2} |\boldsymbol{u}|^2_a - \frac{\alpha^{-1} E_{\max}}{2\epsilon} ||p||^2_{\mathcal{W}}\end{aligned}$$

for any $\epsilon > 0$. From (2.11) and using (2.18) and (2.16) gives

$$\begin{aligned}\mathcal{B}(\boldsymbol{u}, p, \tilde{p}; 0, 0, -p) &= c(p,p) - d(\tilde{p}, p) \\ &\geq (\alpha\beta)^{-1}||p||^2_{\mathcal{W}} - |\tilde{p}|_d |p|_d \\ &\geq (\alpha\beta)^{-1}||p||^2_{\mathcal{W}} - |\tilde{p}|_d \, E_{\max}^{1/2}(\alpha\beta)^{-1/2}||p||_{\mathcal{W}} \\ &\geq (\alpha\beta)^{-1}||p||^2_{\mathcal{W}} - \frac{\epsilon_1}{2}|\tilde{p}|^2_d - \frac{(\alpha\beta)^{-1}E_{\max}}{2\epsilon_1}||p||^2_{\mathcal{W}}\end{aligned}$$

for any $\epsilon_1 > 0$. We now introduce two parameters $\delta > 0$ and $\delta' > 0$. Combining these two bounds gives

$$\begin{aligned}&\mathcal{B}(\boldsymbol{u}, p, \tilde{p}; \boldsymbol{u} - \delta\boldsymbol{w}, -p, \tilde{p} - \delta' p) \\ &\quad= \mathcal{B}(\boldsymbol{u}, p, \tilde{p}; \boldsymbol{u}, -p, \tilde{p}) + \delta\,\mathcal{B}(\boldsymbol{u}, p, \tilde{p}; -\boldsymbol{w}, 0, 0) + \delta'\,\mathcal{B}(\boldsymbol{u}, p, \tilde{p}; 0, 0, -p) \\ &\quad\geq |\boldsymbol{u}|^2_a + |\tilde{p}|^2_d + \delta\left(\frac{1}{\alpha}\left(C_D - \frac{E_{\max}}{2\epsilon}\right)||p||^2_{\mathcal{W}} - \frac{\epsilon}{2}|\boldsymbol{u}|^2_a\right) \\ &\qquad+ \delta'\left(\frac{1}{\alpha\beta}\left(1 - \frac{E_{\max}}{2\epsilon_1}\right)||p||^2_{\mathcal{W}} - \frac{\epsilon_1}{2}||\tilde{p}||^2_d\right) \\ &\quad= \left(1 - \frac{\delta\epsilon}{2}\right)|\boldsymbol{u}|^2_a + \left(\frac{\delta}{\alpha}\left(C_D - \frac{E_{\max}}{2\epsilon}\right) + \frac{\delta'}{\alpha\beta}\left(1 - \frac{E_{\max}}{2\epsilon_1}\right)\right)||p||^2_{\mathcal{W}} \\ &\qquad+ \left(1 - \frac{\delta'\epsilon_1}{2}\right)|\tilde{p}|^2_d.\end{aligned}$$

Next, making the specific choice

$$\epsilon = \frac{E_{\max}}{C_D}, \quad \delta = \frac{1}{\epsilon} = \frac{C_D}{E_{\max}}, \quad \epsilon_1 = E_{\max}, \quad \delta' = \frac{1}{\epsilon_1} = \frac{1}{E_{\max}}$$

and using (2.17) and (2.19) gives

$$\begin{aligned}&\mathcal{B}(\boldsymbol{u}, p, \tilde{p}; \boldsymbol{u} - \delta\boldsymbol{w}, -p, \tilde{p} - \delta' p) \\ &\quad\geq \frac{1}{2}|\boldsymbol{u}|^2_a + \frac{1}{2}\left(\frac{C_D^2}{\alpha E_{\max}} + \frac{1}{\alpha\beta E_{\max}}\right)||p||^2_{\mathcal{W}} + \frac{1}{2}|\tilde{p}|^2_d, \\ &\quad\geq \frac{1}{2}C_K E_{\min}\alpha\,||\nabla\boldsymbol{u}||^2_{\boldsymbol{\mathcal{W}}} + \frac{1}{2E_{\max}}\left(\frac{C_D^2}{\alpha} + \frac{1}{\alpha\beta}\right)||p||^2_{\mathcal{W}} + \frac{1}{2\alpha\beta}E_{\min}\,||\tilde{p}||^2_{\mathcal{W}}, \\ &\quad\geq C\left(\alpha||\nabla\boldsymbol{u}||^2_{\boldsymbol{\mathcal{W}}} + \left(\frac{1}{\alpha} + \frac{1}{\alpha\beta}\right)||p||^2_{\mathcal{W}} + \frac{1}{\alpha\beta}||\tilde{p}||^2_{\mathcal{W}}\right) =: C\,|||(\boldsymbol{u}, p, \tilde{p})|||^2,\end{aligned}$$

where $C = \frac{1}{2}\min\{E_{\min}C_K, \frac{C_D^2}{E_{\max}}, \frac{1}{E_{\max}}\}$. Since $E_{\min} \leq E_{\max}$ we have shown that (2.22) holds with $\boldsymbol{v} := \boldsymbol{u} - \delta\boldsymbol{w}$, $q := -p$, $\tilde{q} := \tilde{p} - \delta' p$ with $C \geq E_{\min}C_1$, where $C_1 = \frac{1}{2}\min\{C_K, \frac{C_D^2}{E_{\max}^2}, \frac{1}{E_{\max}^2}\}$. To complete the proof, we note that

$$\alpha||\nabla(\boldsymbol{u} - \delta\boldsymbol{w})||^2_{\boldsymbol{\mathcal{W}}} \leq 2\alpha||\nabla\boldsymbol{u}||^2_{\boldsymbol{\mathcal{W}}} + 2\delta^2\alpha||\nabla\boldsymbol{w}||^2_{\boldsymbol{\mathcal{W}}} \leq 2\alpha||\nabla\boldsymbol{u}||^2_{\boldsymbol{\mathcal{W}}} + 2\delta^2\alpha^{-1}||p||^2_{\mathcal{W}}.$$

Similarly,

$$(\alpha\beta)^{-1}||\tilde{p} - \delta' p||^2_{\mathcal{W}} \leq 2(\alpha\beta)^{-1}\|\tilde{p}\|^2_{\mathcal{W}} + 2\delta'^2(\alpha\beta)^{-1}\|p\|^2_{\mathcal{W}}.$$

Using the definition of the norm $|||\cdot|||$ then leads to the upper bound

$$\begin{aligned}
&|||(\boldsymbol{u}-\delta\boldsymbol{w},-p,\tilde{p}-\delta' p)|||^2 \\
&\qquad = \alpha||\nabla(\boldsymbol{u}-\delta\boldsymbol{w})||^2_{\boldsymbol{\mathcal{W}}} + \left(\frac{1}{\alpha}+\frac{1}{\alpha\beta}\right)||p||^2_{\mathcal{W}} + \frac{1}{\alpha\beta}||\tilde{p}-\delta' p||^2_{\mathcal{W}} \\
&\qquad \le (2+2\delta^2+2\delta'^2)\Big(\alpha||\nabla\boldsymbol{u}||^2_{\boldsymbol{\mathcal{W}}} + \Big(\frac{1}{\alpha}+\frac{1}{\alpha\beta}\Big)||p||^2_{\mathcal{W}} + \frac{1}{\alpha\beta}||\tilde{p}||^2_{\mathcal{W}}\Big) \\
&\qquad = C_2^2\,|||(\boldsymbol{u},p,\tilde{p})|||^2,
\end{aligned}$$

as required. ∎

The following theorem is an immediate consequence.

THEOREM 2.4. *Given that E satisfies condition* (2.1) *in Assumption* 2.1 *the three-field formulation* (2.12) *admits a unique solution $(\boldsymbol{u},p,\tilde{p})\in\boldsymbol{\mathcal{V}}\times\mathcal{W}\times\mathcal{W}$. Moreover,*

$$|||(\boldsymbol{u},p,\tilde{p})||| \le \frac{C_3}{E_{\min}}\alpha^{-1/2}||\boldsymbol{f}||_{L^2(D)}, \tag{2.24}$$

where C_3 depends on $E_{\max}$, C_K, and C_D.

Proof. Lemma 2.3 ensures that

$$C_1E_{\min}|||(\boldsymbol{u},p,\tilde{p})|||^2 \le \mathcal{B}(\boldsymbol{u},p,\tilde{p};\boldsymbol{v},q,\tilde{q}) = f(\boldsymbol{v}), \tag{2.25}$$

where $(\boldsymbol{v},q,\tilde{q})$ satisfies $|||(\boldsymbol{v},q,\tilde{q})||| \le C_2|||(\boldsymbol{u},p,\tilde{p})|||$. Applying Cauchy–Schwarz to the right-hand side then gives

$$\begin{aligned}
C_1E_{\min}|||(\boldsymbol{u},p,\tilde{p})|||^2 &\le \alpha^{-1/2}||\boldsymbol{f}||_{L^2(D)}\,\alpha^{1/2}||\boldsymbol{v}||_{L^2(\Gamma,L^2(D))} \\
&\le \alpha^{-1/2}||\boldsymbol{f}||_{L^2(D)}\,L|||(\boldsymbol{v},q,\tilde{q})||| \\
&\le \alpha^{-1/2}||\boldsymbol{f}||_{L^2(D)}\,LC_2\,|||(\boldsymbol{u},p,\tilde{p})|||,
\end{aligned}$$

where L is the Poincaré–Friedrichs constant associated with D. This implies (2.24) with $C_3 := LC_2/C_1$. ∎

3. Finite-dimensional formulation. To construct an SGFEM approximation of (2.4) we need to introduce a conforming finite element space

$$V_h = \operatorname{span}\{\phi_1(\boldsymbol{x}),\phi_2(\boldsymbol{x}),\ldots,\phi_{n_u}(\boldsymbol{x})\} \subset H^1_{E_0}(D)$$

and then define $\boldsymbol{V}_h$ to be the space of vector-valued functions whose components are in V_h. We will also require a compatible finite element space

$$W_h = \operatorname{span}\big\{\varphi_1(\boldsymbol{x}),\varphi_2(\boldsymbol{x}),\ldots,\varphi_{n_p}(\boldsymbol{x})\big\} \subset L^2(D),$$

in the sense that a *discrete* inf-sup condition

$$\sup_{0\neq\boldsymbol{v}\in\boldsymbol{V}_h}\frac{\int_D q\,\nabla\cdot\boldsymbol{v}}{||\nabla\boldsymbol{v}||_{\boldsymbol{L}^2(D)}} \ge \gamma||q||_{L^2(D)} \quad \forall q\in W_h \tag{3.1}$$

is satisfied with γ uniformly bounded away from zero (that is, independent of the mesh parameter h). Two specific inf-sup stable approximation pairs are included in our IFISS software [5] and thus have been extensively tested. These are $\boldsymbol{Q}_2$–Q_1 (continuous biquadratic approximation for the displacement and continuous bilinear

approximation for the pressure) and $\boldsymbol{Q}_2$–P_{-1} (continuous biquadratic approximation for the displacement and discontinuous linear approximation for the pressure) approximations for $\boldsymbol{V}_h$ and W_h defined on a rectangular element subdivision.[1]

Turning to the parametric discretization, let $\{\psi_i(y_j), i = 0, 1, \ldots\}$ denote the set of Legendre polynomials on Γ_j, where ψ_i has degree i. We fix $\psi_0 = 1$ and assume that the polynomials are normalized in the $L^2_{\pi_j}(\Gamma_j)$-sense, so that $\langle \psi_i, \psi_k \rangle_{\pi_j} = \delta_{i,k}$. Next, we choose a set of multi-indices $\Lambda \subset \mathbb{N}_0^M$ and define the set of multivariate polynomials

$$S_\Lambda := \operatorname{span}\left\{ \psi_{\boldsymbol{\alpha}}(\boldsymbol{y}) = \prod_{i=1}^{M} \psi_{\alpha_i}(y_i), \quad \boldsymbol{\alpha} \in \Lambda \right\} \subset L^2_\pi(\Gamma). \tag{3.2}$$

By construction, since π is a product measure, the basis functions for S_Λ are orthonormal with respect to the $L^2_\pi(\Gamma)$ inner product. We denote $\dim(S_\Lambda) = |\Lambda| = n_y$. For instance, if we choose $\Lambda = \{\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M), |\boldsymbol{\alpha}| \le p\}$, then $n_y = \frac{(M+p)!}{M!p!}$ and S_Λ contains multivariate polynomials of total degree p or less.

The finite-dimensional version of the three-field problem (2.4) is therefore to find $(\boldsymbol{u}_{h,\Lambda}, p_{h,\Lambda}, \tilde{p}_{h,\Lambda}) \in \boldsymbol{V}_{h,\Lambda} \times W_{h,\Lambda} \times W_{h,\Lambda}$ such that

$$a(\boldsymbol{u}_{h,\Lambda}, \boldsymbol{v}) + b(\boldsymbol{v}, p_{h,\Lambda}) = f(\boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \boldsymbol{V}_{h,\Lambda}, \tag{3.3a}$$

$$b(\boldsymbol{u}_{h,\Lambda}, q) - c(\tilde{p}_{h,\Lambda}, q) = 0 \qquad \forall q \in W_{h,\Lambda}, \tag{3.3b}$$

$$-c(p_{h,\Lambda}, \tilde{q}) + d(\tilde{p}_{h,\Lambda}, \tilde{q}) = 0 \qquad \forall \tilde{q} \in W_{h,\Lambda}, \tag{3.3c}$$

where we define $\boldsymbol{V}_{h,\Lambda} := \boldsymbol{V}_h \otimes S_\Lambda$ and $W_{h,\Lambda} := W_h \otimes S_\Lambda$.

The well-posedness of the discrete formulation follows from the stability estimate in Lemma 2.3 together with the discrete inf-sup condition (3.1).

LEMMA 3.1. *Assuming that E satisfies* (2.1) *and that the approximation pair $\boldsymbol{V}_h$, W_h is inf-sup stable, problem* (3.3) *admits a unique solution $(\boldsymbol{u}_{h,\Lambda}, p_{h,\Lambda}, \tilde{p}_{h,\Lambda}) \in \boldsymbol{V}_{h,\Lambda} \times W_{h,\Lambda}, \times W_{h,\Lambda}$ satisfying*

$$|||(\boldsymbol{u}_{h,\Lambda}, p_{h,\Lambda}, \tilde{p}_{h,\Lambda})||| \le \frac{C_5}{E_{\min}} \alpha^{-1/2} ||\boldsymbol{f}||_{L^2(D)}, \tag{3.4}$$

where C_5 depends on $E_{\max}$, C_K, and γ.

Remark 3.1. One could, in principle, approximate p and $\tilde{p}$ using different finite element spaces. In this case a second inf-sup condition relating the two pressure spaces would need to be satisfied to ensure a stable approximation overall.

3.1. Linear algebra aspects. We will restrict our attention to planar elasticity from this point onward.[2] To formulate the discrete linear system of equations associated with (3.3), a set of sparse matrices and vectors associated with the chosen basis functions for the approximation spaces V_h, W_h, and S_Λ will need to be assembled. To this end, we first define matrices $G_0, G_k \in \mathbb{R}^{n_y \times n_y}$ for $k = 1, \ldots, M$, by

$$[G_0]_{\boldsymbol{\alpha},\boldsymbol{\beta}} := \int_\Gamma \psi_{\boldsymbol{\alpha}}(\boldsymbol{y}) \psi_{\boldsymbol{\beta}}(\boldsymbol{y}) \, d\pi(\boldsymbol{y}), \qquad [G_k]_{\boldsymbol{\alpha},\boldsymbol{\beta}} := \int_\Gamma y_k \, \psi_{\boldsymbol{\alpha}}(\boldsymbol{y}) \psi_{\boldsymbol{\beta}}(\boldsymbol{y}) \, d\pi(\boldsymbol{y}),$$

where $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \Lambda$. In addition, we define the vector $\boldsymbol{g}_0 \in \mathbb{R}^{n_y}$ to be the first column of G_0. Since the basis functions for S_Λ have been chosen to be orthonormal, we have $G_0 = I$. In addition, due to the three-term recurrence of the underlying univariate

[1] Both of these mixed approximation strategies are inf-sup stable in a three-dimensional setting.

[2] The extension to three dimensions is completely straightforward.

families of Legendre polynomials, G_k has at most two nonzero entries per row, for each $k = 1, 2, \ldots, M$; see [16].

We will define the finite element matrix $A_{11}^k \in \mathbb{R}^{n_u \times n_u}$ associated with V_h by

$$[A_{11}^k]_{i,\ell} := \int_D e_k(\boldsymbol{x})\, \boldsymbol{\varepsilon} \begin{pmatrix} \phi_i(\boldsymbol{x}) \\ 0 \end{pmatrix} : \boldsymbol{\varepsilon} \begin{pmatrix} \phi_\ell(\boldsymbol{x}) \\ 0 \end{pmatrix} d\boldsymbol{x}, \qquad i, \ell = 1, \ldots, n_u,$$

for $k = 0, 1, \ldots, M$ and the matrix $A_{21}^k \in \mathbb{R}^{n_u \times n_u}$ by

$$[A_{21}^k]_{i,\ell} := \int_D e_k(\boldsymbol{x})\, \boldsymbol{\varepsilon} \begin{pmatrix} 0 \\ \phi_i(\boldsymbol{x}) \end{pmatrix} : \boldsymbol{\varepsilon} \begin{pmatrix} \phi_\ell(\boldsymbol{x}) \\ 0 \end{pmatrix} d\boldsymbol{x}, \qquad i, \ell = 1, \ldots, n_u.$$

The matrices $A_{12}^k, A_{22}^k \in \mathbb{R}^{n_u \times n_u}$ are defined analogously. We can also define matrices $B_1, B_2 \in \mathbb{R}^{n_p \times n_u}$ so that

$$[B_1]_{r,\ell} = -\int_D \varphi_r(\mathbf{x})\, \frac{\partial \phi_\ell(\mathbf{x})}{\partial x_1} d\boldsymbol{x}, \quad [B_2]_{r,\ell} = -\int_D \varphi_r(\mathbf{x})\, \frac{\partial \phi_\ell(\mathbf{x})}{\partial x_2} d\boldsymbol{x}$$

for $r = 1, \ldots, n_p$, $\ell = 1, \ldots, n_u$. The mass matrix $C \in \mathbb{R}^{n_p \times n_p}$ associated with W_h is defined by

$$[C]_{r,s} = \int_D \varphi_r(\mathbf{x})\, \varphi_s(\mathbf{x})\, d\mathbf{x}, \qquad r, s = 1, \ldots, n_p,$$

and the weighted mass matrices $D_k \in \mathbb{R}^{n_p \times n_p}$ are defined by

$$[D_k]_{r,s} = \int_D e_k(\mathbf{x}) \varphi_r(\mathbf{x})\, \varphi_s(\mathbf{x})\, d\mathbf{x}, \qquad r, s = 1, \ldots, n_p,$$

for $k = 0, 1, \ldots, M$. An important point is that if the coefficient $e_0(\boldsymbol{x})$ in the expansion of E is a constant, then $D_0 = e_0 C$. Moreover, if we choose $W_h = P_{-1}$ (discontinuous linear pressure approximation), then C is a diagonal matrix and so is D_k, for each $k = 0, 1, \ldots, M$. Finally, we define two vectors $\boldsymbol{f}_1, \boldsymbol{f}_2 \in \mathbb{R}^{n_u}$ associated with the body force $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}))^\top$, via

$$[\boldsymbol{f}_1]_\ell = \int_D f_1(\boldsymbol{x}) \phi_\ell(\boldsymbol{x})\, d\boldsymbol{x}, \qquad [\boldsymbol{f}_2]_\ell = \int_D f_2(\boldsymbol{x}) \phi_\ell(\boldsymbol{x})\, d\boldsymbol{x}, \qquad \ell = 1, \ldots, n_u.$$

Permuting the variables $p_{h,\Lambda}$ and $\tilde{p}_{h,\Lambda}$ in (3.3) and swapping the order of the second and third equations leads to a *saddle-point* system of $2(n_u + n_p)n_y$ equations

$$\begin{pmatrix} \mathcal{A} & \mathcal{B}^\top \\ \mathcal{B} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{v} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix}, \tag{3.5}$$

where $\mathbf{b}_1 = \mathbf{g}_0 \otimes \mathbf{f}_1$, $\mathbf{b}_2 = \mathbf{g}_0 \otimes \mathbf{f}_2$ with vectors

$$\mathbf{v} = \begin{pmatrix} \mathbf{u} \\ \tilde{\mathbf{p}} \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix},$$

defined so that $\mathbf{u}$, $\tilde{\mathbf{p}}$, and $\mathbf{p}$ are the coefficient vectors representing $u_{h,\Lambda}$, $\tilde{p}_{h,\Lambda}$, and $p_{h,\Lambda}$, respectively, in the chosen bases. The coefficient matrix in (3.5) is symmetric with

$$\mathcal{A} := \left(\begin{array}{cc|c} \alpha \sum_{k=0}^M G_k \otimes A_{11}^k & \alpha \sum_{k=0}^M G_k \otimes A_{21}^k & \mathbf{0} \\ \alpha \sum_{k=0}^M G_k \otimes A_{12}^k & \alpha \sum_{k=0}^M G_k \otimes A_{22}^k & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & (\alpha\beta)^{-1} \sum_{k=0}^M G_k \otimes D_k \end{array}\right) \tag{3.6}$$

and

$$\mathcal{B} := \left(\begin{array}{cc|c} G_0 \otimes B_1 & G_0 \otimes B_2 & -(\alpha\beta)^{-1} G_0 \otimes C \end{array} \right). \tag{3.7}$$

Note that due to its very large size, we do not assemble the full coefficient matrix. Operations are only performed via the actions of G_0, G_k, A_{11}^k, A_{12}^k, A_{21}^k, A_{22}^k, B_1, B_2, C, and D_k.

The best way to solve a symmetric saddle-point system iteratively is to use the minimal residual method; see [6, Chapter 4]. Since our system is ill-conditioned, preconditioning is a critical component of our solution strategy.

4. Preconditioning. Following [8], [12], [15], and [19] the most natural preconditioner for the saddle-point system (3.5) is a block preconditioning matrix

$$P_{\text{approx}} = \begin{pmatrix} \mathcal{A}_{\text{approx}} & 0 \\ 0 & \mathcal{S}_{\text{approx}} \end{pmatrix},$$

where $\mathcal{A}_{\text{approx}}$ and $\mathcal{S}_{\text{approx}}$ are matrices that are chosen to represent the matrix $\mathcal{A}$ and the Schur complement $\mathcal{S} = \mathcal{B}\mathcal{A}^{-1}\mathcal{B}^\top$. An important requirement is that the work needed to apply the action of $\mathcal{A}_{\text{approx}}^{-1}$ and $\mathcal{S}_{\text{approx}}^{-1}$ is proportional to the dimension of the associated approximation space.

4.1. Approximation of $\mathcal{A}$. The obvious way to approximate (3.6) is given by

$$\mathcal{A}_{\text{approx}} = \left(\begin{array}{cc|c} \alpha\, G_0 \otimes A_{11}^0 & \alpha\, G_0 \otimes A_{21}^0 & \mathbf{0} \\ \alpha\, G_0 \otimes A_{12}^0 & \alpha\, G_0 \otimes A_{22}^0 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & (\alpha\beta)^{-1}\,(G_0 \otimes D_0) \end{array} \right), \tag{4.1}$$

where we denote the diagonal blocks in (4.1) by

$$\mathcal{A}_{\text{approx},1} := \alpha \begin{pmatrix} G_0 \otimes A_{11}^0 & G_0 \otimes A_{21}^0 \\ G_0 \otimes A_{12}^0 & G_0 \otimes A_{22}^0 \end{pmatrix}, \quad \mathcal{A}_{\text{approx},2} := \frac{1}{\alpha\beta}\,(G_0 \otimes D_0)\,. \tag{4.2}$$

The fact that $G_0 = I$ means that the nonzero terms in (4.1) are all block diagonal. Since the finite element matrices A_{11}^0, A_{12}^0, A_{21}^0, A_{22}^0, and D_0 all involve the mean coefficient e_0, we will refer to this strategy as a *mean-based* approximation. The following lemma quantifies the effectiveness of this approximation.

LEMMA 4.1. *Let $\mathcal{A}$ and $\mathcal{A}_{approx}$ be defined in* (3.6) *and* (4.1). *If Assumption* 2.1 *holds, the eigenvalues of $\mathcal{A}_{approx}^{-1}\mathcal{A}$ lie in the bounded interval $\left[E_{\min}/e_0^{\max}, E_{\max}/e_0^{\min}\right]$.*

Proof. The eigenvalues of $\mathcal{A}_{\text{approx}}^{-1}\mathcal{A}$ can be separated into two distinct sets, each associated with one of the diagonal blocks of $\mathcal{A}$, that is,

$$\mathcal{A}_1 := \alpha \begin{pmatrix} \sum_{k=0}^M G_k \otimes A_{11}^k & \sum_{k=0}^M G_k \otimes A_{21}^k \\ \sum_{k=0}^M G_k \otimes A_{12}^k & \sum_{k=0}^M G_k \otimes A_{22}^k \end{pmatrix}, \quad \mathcal{A}_2 := \frac{1}{\alpha\beta} \sum_{k=0}^M G_k \otimes D_k.$$

Let us consider the first block. For any $\boldsymbol{v} \in \mathbb{R}^{2n_u n_y}$ there is an associated function

$\boldsymbol{r} \in \boldsymbol{V}_{h,\Lambda}$ and using (2.1) and (2.2) gives

$$\begin{aligned}\boldsymbol{v}^\top \mathcal{A}_1 \boldsymbol{v} = a(\boldsymbol{r},\boldsymbol{r}) &= \alpha \int_\Gamma \int_D E(\boldsymbol{x},\boldsymbol{y})\, \boldsymbol{\varepsilon}(\boldsymbol{r}) : \boldsymbol{\varepsilon}(\boldsymbol{r}) d\boldsymbol{x} d\pi(\boldsymbol{y}) \\ &\leq \frac{E_{\max}}{e_0^{\min}}\, \alpha \int_\Gamma \int_D e_0(\boldsymbol{x})\, \boldsymbol{\varepsilon}(\boldsymbol{r}) : \boldsymbol{\varepsilon}(\boldsymbol{r}) d\boldsymbol{x} d\pi(\boldsymbol{y}), \\ &= \frac{E_{\max}}{e_0^{\min}}\, \boldsymbol{v}^\top \mathcal{A}_{\text{approx},1} \boldsymbol{v}. \end{aligned} \tag{4.3}$$

Similarly,

$$\boldsymbol{v}^\top \mathcal{A}_1 \boldsymbol{v} \geq \frac{E_{\min}}{e_0^{\max}}\, \boldsymbol{v}^\top \mathcal{A}_{\text{approx},1} \boldsymbol{v}. \tag{4.4}$$

Combining (4.3) and (4.4) gives, for any $\boldsymbol{v} \neq \boldsymbol{0}$,

$$\frac{E_{\min}}{e_0^{\max}} \leq \frac{\boldsymbol{v}^\top \mathcal{A}_1 \boldsymbol{v}}{\boldsymbol{v}^\top \mathcal{A}_{\text{approx},1} \boldsymbol{v}} \leq \frac{E_{\max}}{e_0^{\min}}. \tag{4.5}$$

Let us consider the second block. For any $\boldsymbol{w} \in \mathbb{R}^{n_p n_y}$ we can define a function $s \in W_{h,\Lambda}$ such that

$$\begin{aligned}\boldsymbol{w}^\top \mathcal{A}_2 \boldsymbol{w} = d(s,s) &= (\alpha\beta)^{-1} \int_\Gamma \int_D E(\boldsymbol{x},\boldsymbol{y}) s(\boldsymbol{x},\boldsymbol{y})\, s(\boldsymbol{x},\boldsymbol{y})\, d\boldsymbol{x} d\pi(\boldsymbol{y}), \\ \boldsymbol{w}^\top \mathcal{A}_{\text{approx},2} \boldsymbol{w} &= (\alpha\beta)^{-1} \int_\Gamma \int_D e_0(\boldsymbol{x}) s(\boldsymbol{x},\boldsymbol{y})\, s(\boldsymbol{x},\boldsymbol{y})\, d\boldsymbol{x} d\pi(\boldsymbol{y}).\end{aligned}$$

Making use of (2.1) and (2.2) again gives

$$\frac{E_{\min}}{e_0^{\max}} \leq \frac{\boldsymbol{w}^\top \mathcal{A}_2 \boldsymbol{w}}{\boldsymbol{w}^\top \mathcal{A}_{\text{approx},2} \boldsymbol{w}} \leq \frac{E_{\max}}{e_0^{\min}}. \tag{4.6}$$

Combining the bounds for the two Rayleigh quotients completes the proof. ∎

4.2. Refined approximations of $\boldsymbol{\mathcal{A}}$. Inverting the $\mathcal{A}_{\text{approx},1}$ block of (4.1) is computationally expensive. To address this we will look for block diagonal alternatives of the form

$$\tilde{\mathcal{A}}_{\text{approx},1} := \alpha \begin{pmatrix} G_0 \otimes \mathbb{A}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & G_0 \otimes \mathbb{A}_{22} \end{pmatrix}. \tag{4.7}$$

Herein, we will consider two different choices of $\mathbb{A}_{11}$, and $\mathbb{A}_{22}$. The first option is to take $\mathbb{A}_{11} = \mathbb{A}_{22} = \mathbb{A} := 2(A_{11}^0 + A_{22}^0)/3$. Note that if $e_0 = 1$, then for any $\boldsymbol{v} \in \mathbb{R}^{n_u}$ we have $\boldsymbol{v}^\top \mathbb{A} \boldsymbol{v} = \|\nabla v_h\|^2_{L^2(D)}$, where v_h is the finite element function in V_h represented by $\boldsymbol{v}$. That is, $\mathbb{A}$ gives a discrete representation of the scalar Laplacian operator.

LEMMA 4.2. *Let $\mathbb{A}_{11} = \mathbb{A}_{22} = 2(A_{11}^0 + A_{22}^0)/3$. If Assumption 2.1 holds, then all eigenvalues $\sigma_{\mathcal{A}}$ of $\mathcal{A}_{approx}^{-1}\mathcal{A}$, where $\mathcal{A}_{approx}$ has leading diagonal block (4.7), satisfy*

$$\sigma_{\mathcal{A}} \in \left[C_K \frac{E_{\min}}{e_0^{\max}}, \frac{E_{\max}}{e_0^{\min}} \right], \tag{4.8}$$

where $0 < C_K \leq 1$ is the Korn constant.

Proof. For any $\boldsymbol{v} \in \mathbb{R}^{2n_u n_y}$, we can define a function $\boldsymbol{r} \in \boldsymbol{V}_{h,\Lambda}$ such that

$$\begin{aligned}
\boldsymbol{v}^\top \mathcal{A}_1 \boldsymbol{v} &= \alpha \int_\Gamma \int_D E(\boldsymbol{x},\boldsymbol{y}) \boldsymbol{\varepsilon}(\boldsymbol{r}) : \boldsymbol{\varepsilon}(\boldsymbol{r}) d\boldsymbol{x} d\pi(\boldsymbol{y}) \\
&\leq E_{\max}\, \alpha \int_\Gamma \int_D \nabla \boldsymbol{r} : \nabla \boldsymbol{r} d\boldsymbol{x} d\pi(\boldsymbol{y}) \\
&\leq \frac{E_{\max}}{e_0^{\min}}\, \alpha \int_\Gamma \int_D e_0(\boldsymbol{x}) \nabla \boldsymbol{r} : \nabla \boldsymbol{r} d\boldsymbol{x} d\pi(\boldsymbol{y}) \\
&= \frac{E_{\max}}{e_0^{\min}} \boldsymbol{v}^\top \tilde{\mathcal{A}}_{\text{approx},1} \boldsymbol{v}. \qquad (4.9)
\end{aligned}$$

Analagously,

$$\begin{aligned}
\boldsymbol{v}^\top \mathcal{A}_1 \boldsymbol{v} &\geq E_{\min}\, \alpha \int_\Gamma \int_D \boldsymbol{\varepsilon}(\boldsymbol{r}) : \boldsymbol{\varepsilon}(\boldsymbol{r}) d\boldsymbol{x} d\pi(\boldsymbol{y}) \\
&\geq E_{\min}\, \alpha C_K \int_\Gamma \int_D \nabla \boldsymbol{r} : \nabla \boldsymbol{r} d\boldsymbol{x} d\pi(\boldsymbol{y}) \\
&\geq \frac{E_{\min}}{e_0^{\max}}\, \alpha C_K \int_\Gamma \int_D e_0(\boldsymbol{x}) \nabla \boldsymbol{r} : \nabla \boldsymbol{r} d\boldsymbol{x} d\pi(\boldsymbol{y}) \\
&= \frac{E_{\min}}{e_0^{\max}} C_K \boldsymbol{v}^\top \tilde{\mathcal{A}}_{\text{approx},1} \boldsymbol{v}. \qquad (4.10)
\end{aligned}$$

Combining (4.9) and (4.10) leads to bounds for the Rayleigh quotient

$$\frac{E_{\min}}{e_0^{\max}} C_K \leq \frac{\boldsymbol{v}^\top \mathcal{A}_1 \boldsymbol{v}}{\boldsymbol{v}^\top \tilde{\mathcal{A}}_{\text{approx},1} \boldsymbol{v}} \leq \frac{E_{\max}}{e_0^{\min}}$$

and hence for the eigenvalues of $\tilde{\mathcal{A}}_{\text{approx},1}^{-1} \mathcal{A}_1$. The bound (4.6) provides a bound for the eigenvalues of $\mathcal{A}_{\text{approx},2}^{-1} \mathcal{A}_2$. Combining these two bounds gives the stated result. ∎

For our second choice of $\mathbb{A}_{11}$ and $\mathbb{A}_{22}$, we simply discard the off-diagonal blocks of the mean-based approximation $\mathcal{A}_{\text{approx},1}$. The strategy will not be pursued here since it results in an inferior eigenvalue bound.

LEMMA 4.3. *Let* $\mathbb{A}_{11} = A_{11}^0$ *and* $\mathbb{A}_{22} = A_{22}^0$. *If Assumption* 2.1 *holds, then all eigenvalues* $\sigma_{\mathcal{A}}$ *of* $\mathcal{A}_{approx}^{-1} \mathcal{A}$, *where* $\mathcal{A}_{approx}$ *has leading diagonal block* (4.7), *satisfy*

$$\sigma_{\mathcal{A}} \in \left[C_K \frac{E_{\min}}{e_0^{\max}}, 2 \frac{E_{\max}}{e_0^{\min}} \right], \qquad (4.11)$$

where $0 < C_K \leq 1$ *is the Korn constant.*

Proof. The proof is a minor variation of that of Lemma 4.2. By obtaining bounds for both Rayleigh quotients separately, we find that the eigenvalues lie in

$$\left[C_K \frac{E_{\min}}{e_0^{\max}}, 2 \frac{E_{\max}}{e_0^{\min}} \right] \cup \left[\frac{E_{\min}}{e_0^{\max}}, \frac{E_{\max}}{e_0^{\min}} \right],$$

which yields the stated result. ∎

Remark 4.1. The bounds (4.8) and (4.11) depend on the Young's modulus E and on the Korn constant C_K but are independent of all discretization parameters.

4.3. Approximation of $\mathcal{S}$. Given a block diagonal approximation to $\mathcal{A}_1$ of the form (4.7), an approximation to the Schur complement matrix $\mathcal{S}$ can be constructed so that $\tilde{\mathcal{S}}_{\text{approx}} := \mathcal{B}\tilde{\mathcal{A}}^{-1}_{\text{approx}}\mathcal{B}^\top$. Since this is a dense matrix it is not a practical preconditioner. The next result introduces a sparse block diagonal matrix $P_{\mathcal{S}}$ and establishes that it is spectrally equivalent to $\tilde{\mathcal{S}}_{\text{approx}}$.

LEMMA 4.4. *Suppose that $\tilde{\mathcal{S}}_{approx} := \mathcal{B}\tilde{\mathcal{A}}^{-1}_{approx}\mathcal{B}^\top$, where in the definition* (4.7) *we make the choice $\mathbb{A}_{11} = \mathbb{A}_{22} = 2(A^0_{11} + A^0_{22})/3$. Defining*

$$P_S := (\alpha^{-1} + (\alpha\beta)^{-1})\, I \otimes C, \tag{4.12}$$

where C is the pressure mass matrix, we have

$$\theta^2 \leq \frac{\boldsymbol{w}^\top \tilde{\mathcal{S}}_{approx}\boldsymbol{w}}{\boldsymbol{w}^\top P_{\mathcal{S}}\boldsymbol{w}} \leq \Theta^2 \qquad \forall \boldsymbol{w} \in \mathbb{R}^{n_p n_y} \tag{4.13}$$

with $\theta^2 = \gamma^2/e_0^{\max}$, $\Theta^2 = 2/e_0^{\min}$, where γ is the discrete inf-sup constant in (3.1) *associated with the finite element spaces $\boldsymbol{V}_h$ and W_h.*

Proof. Using the definitions of $\mathcal{B}$ and $\tilde{\mathcal{A}}_{\text{approx}}$ and the fact that $G_0 = I$ gives

$$\begin{aligned}
\tilde{\mathcal{S}}_{\text{approx}} &= (I \otimes B_1)\left(\alpha\left(I \otimes \mathbb{A}_{11}\right)\right)^{-1}(I \otimes B_1)^\top + (I \otimes B_2)\left(\alpha\left(I \otimes \mathbb{A}_{22}\right)\right)^{-1}(I \otimes B_2)^\top \\
&\quad + \left(-(\alpha\beta)^{-1} I \otimes C\right)\left((\alpha\beta)^{-1}(I \otimes D_0)\right)^{-1}\left(-(\alpha\beta)^{-1} I \otimes C\right)^\top \\
&= \alpha^{-1}\left(I \otimes \left(B_1 \mathbb{A}_{11}^{-1} B_1^\top\right) + I \otimes \left(B_2 \mathbb{A}_{22}^{-1} B_2^\top\right)\right) + (\alpha\beta)^{-1}\left(I \otimes C D_0^{-1} C^\top\right) \\
&= \alpha^{-1}(I \otimes X) + (\alpha\beta)^{-1}\left(I \otimes C D_0^{-1} C^\top\right), \qquad (4.14)
\end{aligned}$$

where $X := \left(B_1 \mathbb{A}_{11}^{-1} B_1^\top + B_2 \mathbb{A}_{22}^{-1} B_2^\top\right)$.

The fact that the matrices $\mathbb{A}_{11}$ and $\mathbb{A}_{22}$ represent discrete Laplacian operators weighted by the mean field e_0 gives the matrix X a structure that can be exploited. Specifically we can combine the bounds in [6, Proposition 3.24] with the bounds on e_0 in (2.2) to give a two-sided bound

$$\frac{\gamma^2}{e_0^{\max}} \leq \frac{\boldsymbol{w}^\top (I \otimes X)\, \boldsymbol{w}}{\boldsymbol{w}^\top (I \otimes C)\, \boldsymbol{w}} \leq \frac{2}{e_0^{\min}} \qquad \forall \boldsymbol{w} \in \mathbb{R}^{n_p n_y},$$

where γ is the inf-sup constant (as defined in (3.1)). We also have a two-sided bound for the two component mass matrices

$$e_0^{\min} C \leq D_0 \leq e_0^{\max} C, \tag{4.15}$$

where the inequalities hold entrywise. Combining these results with (4.14) gives

$$\begin{aligned}
\boldsymbol{w}^\top \tilde{\mathcal{S}}_{\text{approx}} \boldsymbol{w} &= \alpha^{-1} \boldsymbol{w}^\top (I \otimes X)\, \boldsymbol{w} + (\alpha\beta)^{-1} \boldsymbol{w}^\top \left(I \otimes C D_0^{-1} C^\top\right) \boldsymbol{w} \\
&\leq 2(\alpha e_0^{\min})^{-1} \boldsymbol{w}^\top (I \otimes C)\, \boldsymbol{w} + (\alpha\beta e_0^{\min})^{-1} \boldsymbol{w}^\top (I \otimes C)\, \boldsymbol{w} \\
&\leq 2(e_0^{\min})^{-1} \boldsymbol{w}^\top P_{\mathcal{S}} \boldsymbol{w}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\boldsymbol{w}^\top \tilde{\mathcal{S}}_{\text{approx}} \boldsymbol{w} &\geq (\alpha e_0^{\max})^{-1} \gamma^2 \boldsymbol{w}^\top (I \otimes C)\, \boldsymbol{w} + (\alpha\beta e_0^{\max})^{-1} \boldsymbol{w}^\top I \otimes C \boldsymbol{w} \\
&= \gamma^2 (e_0^{\max})^{-1} \boldsymbol{w}^\top P_{\mathcal{S}} \boldsymbol{w}. \qquad (4.16)
\end{aligned}$$

Combining the upper and lower bounds leads to the stated result. ∎

Remark 4.2. In a practical setting, the mean field e_0 in (1.2) is often taken to be constant. In this case we could define $P_{\mathcal{S}} := (\alpha^{-1} + (\alpha\beta)^{-1})\, e_0^{-1} I \otimes C$ and get a refined estimate

$$\theta^2 := \gamma^2 \leq \frac{\boldsymbol{w}^\top \tilde{\mathcal{S}}_{approx} \boldsymbol{w}}{\boldsymbol{w}^\top P_{\mathcal{S}} \boldsymbol{w}} \leq 2 := \Theta^2 \qquad \forall \boldsymbol{w} \in \mathbb{R}^{n_p n_y}. \tag{4.17}$$

Notice that P_S is block diagonal but if we choose $W_h = P_{-1}$, then C is diagonal and hence so is P_S.

We will summarize our preferred methodology at this point: the preconditioner of choice is a block diagonal matrix

$$\mathcal{P} := \begin{pmatrix} \tilde{\mathcal{A}}_{\text{approx},1} & 0 & 0 \\ 0 & \mathcal{A}_{\text{approx},2} & 0 \\ 0 & 0 & P_{\mathcal{S}} \end{pmatrix}, \tag{4.18}$$

where $\tilde{\mathcal{A}}_{\text{approx},1}$ is as defined in (4.7) with $\mathbb{A}_{11} = \mathbb{A}_{22} = 2(A^0_{11} + A^0_{22})/3$, $\mathcal{A}_{\text{approx},2}$ is defined in (4.2), and P_S is defined in (4.12) (or else as in (4.17) if e_0 is constant).

We note that each of the three diagonal blocks of the preconditioner, $\tilde{\mathcal{A}}_{\text{approx},1}$, $\mathcal{A}_{\text{approx},2}$, and $P_{\mathcal{S}}$, provides a discrete representation of a norm that is equivalent to one of the terms in the norm (2.21). This strategy is consistent with the preconditioning philosophy of Mardal and Winther [15] and ensures that the eigenvalues of the preconditioned system can be bounded independently of the discretization parameters. This is formally expressed in the following concluding result.

THEOREM 4.5. *Let $\mu_{\min}$ and $\mu_{\max}$ be the extremal eigenvalues of $\mathcal{A}^{-1}_{approx}\mathcal{A}$, where $\mathcal{A}_{approx}$ has leading diagonal block* (4.7) *with $\mathbb{A}_{11} = \mathbb{A}_{22} = 2(A^0_{11} + A^0_{22})/3$. Then the eigenvalues of*

$$\mathcal{P}^{-1/2} \begin{pmatrix} \mathcal{A} & \mathcal{B}^\top \\ \mathcal{B} & 0 \end{pmatrix} \mathcal{P}^{-1/2} \tag{4.19}$$

lie in the union of the intervals

$$\begin{aligned} &\left[\frac{1}{2}\left(\mu_{\min} - \sqrt{\mu^2_{\min} + 4\Theta^2}\right), \frac{1}{2}\left(\mu_{\max} - \sqrt{\mu^2_{\max} + 4\theta^2}\right)\right] \\ &\qquad \cup \left[\mu_{\min}, \frac{1}{2}\left(\mu^2_{\max} + \sqrt{\mu_{\max} + 4\Theta^2}\right)\right], \end{aligned} \tag{4.20}$$

where the constants θ and Θ are given in Lemma 4.4 *if $P_{\mathcal{S}}$ is as defined in* (4.12) *or else are given in* (4.17) *if $P_{\mathcal{S}}$ has the alternative definition given in Remark* 4.2.

Proof. The proof follows from Lemma 2.1 of [18] and Corollary 3.4 of [17]. ∎

Recall that bounds for the eigenvalues of $\tilde{\mathcal{A}}^{-1}_{approx}\mathcal{A}$ are given in (4.8). Hence, the bounds for the eigenvalues for the preconditioned system depend *only* on the discrete inf-sup constant γ in (3.1), the Korn constant C_K, and the ratios $E_{\min}/e_0^{\max}$ and $E_{\max}/e_0^{\min}$. Note that the eigenvalue bounds are robust in the incompressible limit. A direct consequence of our eigenvalue bound is that the number of MINRES iterations needed to converge to a fixed tolerance when solving the Galerkin system is guaranteed to be bounded by a constant that is independent of all discretization parameters as well as the Poisson ratio. This will be illustrated by numerical results in the final section.

5. Numerical results. In this section we consider a representative test problem taken from the S-IFISS toolbox [20] and we study the practical performance of the block diagonal preconditioning strategy that was analyzed above. The spatial domain is $D = (-1,1) \times (-1,1)$. We impose a homogeneous Neumann boundary condition on the right edge $\partial D_N = \{1\} \times (-1,1)$ and a zero essential boundary condition for the displacement on $\partial D_D = \partial D \setminus \partial D_N$. The body force is chosen to be $\boldsymbol{f} = (1,1)^\top$. The Young's modulus has constant mean value one and takes the form

$$E(\boldsymbol{x},\boldsymbol{y}) = 1 + \sigma\sqrt{3}\sum_{m=1}^{M}\sqrt{\lambda_m}\varphi_m(\boldsymbol{x})y_m, \tag{5.1}$$

where σ is the standard deviation and $\{(\lambda_m,\varphi_m)\}$ are the eigenpairs of the integral operator associated with $(1/\sigma^2)C(\boldsymbol{x},\boldsymbol{x}')$, where

$$C(\boldsymbol{x},\boldsymbol{x}') = \sigma^2 \exp\left(-\frac{1}{2}||\boldsymbol{x}-\boldsymbol{x}'||_1\right), \quad \boldsymbol{x},\boldsymbol{x}' \in D. \tag{5.2}$$

For the spatial approximation, we use $\boldsymbol{Q}_2 - P_{-1} - P_{-1}$ mixed finite elements, that is, continuous biquadratic approximation for the displacement and discontinuous linear approximation for both of the Lagrange multipliers. In this case, the approximation $P_{\mathcal{S}}$ to the Schur complement appearing in the preconditioner (4.18) is a diagonal matrix, and so is $\mathcal{A}_{\text{approx},2}$. For the parametric approximation, we choose S_Λ to be the set of polynomials of total degree p or less in $y_1,\ldots,y_M$ on $\Gamma = [-1,1]^M$. In Table 5.1 we record the number of spatial degrees of freedom associated with the finite element discretization (as the spatial refinement level ℓ is varied) and in Table 5.2 we record the dimension of the space S_Λ (when M and p are varied). Recall that the number of equations to be solved is $2(n_u+n_p)n_y$. For example, when we have $M = 10$ input parameters, the grid level is set to $\ell = 6$, and the polynomial degree is $p = 4$, we have over 14 million equations to solve.

TABLE 5.1
Number of deterministic degrees of freedom associated with $\boldsymbol{Q}_2 - P_{-1} - P_{-1}$ approximation.

Deterministic degrees of freedom			
Refinement level (l)	n_u	n_p	$2(n_u+n_p)$
4	240	192	864
5	992	768	3,520
6	4,032	3,072	14,208

TABLE 5.2
Number of parametric degrees of freedom associated with the chosen multi-index set Λ.

	n_y		
p	$M=5$	$M=8$	$M=10$
3	56	165	286
4	126	495	1,001

We examine the eigenvalues of the preconditioned SGFEM system first. The `est_minres` code that is built into S-IFISS exploits the connection with the Lanczos algorithm (see [6, section 2.4]) and generates accurate harmonic Ritz value estimates of the underlying eigenvalue spectrum as the preconditioned system is being solved. Details are given in Silvester and Simoncini [19]. The extremal eigenvalue estimates

are computed on the fly and are reproduced in Tables 5.3 and 5.4.[3] We consider two values of the Poisson ratio ν and two values for the standard deviation σ (values which guarantee that all realizations of E are positive) and vary M and ℓ. The polynomial degree $p = 3$ is fixed. We observe that the widths of the intervals containing the estimated eigenvalues are independent of the spatial discretization parameter as well as the number M of parameters. While the intervals are slightly wider for $\nu = 0.49999$ than for $\nu = 0.4$, they are bounded as $\nu \to 1/2$. The interior eigenvalue bounds are closer to the origin, however, for the larger value of σ. The price we pay for using a *mean-based* preconditioner (which is by definition block diagonal) is that the resulting eigenvalue bounds depend on the ratios $E_{\min}/e_0^{\max}$ and $E_{\max}/e_0^{\min}$. In this example, these quantities depend on σ. However, we stress that σ cannot be chosen arbitrarily large. Assumption 2.1 must be satisfied; otherwise the problem is not well posed. Preconditioning schemes that are not mean-based may lead to more tightly clustered eigenvalues but in general are not as computationally efficient.

TABLE 5.3
Bound for eigenvalues of preconditioned SGFEM system, $\sigma = 0.085$, $p = 3$.

Computed eigenvalue		
$l = 5$		
M	$\nu = .4$	$\nu = .49999$
5	$[-0.8287, -0.3369] \cup [0.2737, 1.8332]$	$[-0.9347, -0.1892] \cup [0.2878, 1.8886]$
8	$[-0.8305, -0.3368] \cup [0.2722, 1.8408]$	$[-0.9058, -0.1891] \cup [0.2859, 1.8934]$
10	$[-0.8311, -0.3367] \cup [0.2720, 1.8427]$	$[-0.9064, -0.1891] \cup [0.2857, 1.8949]$
$l = 6$		
5	$[-0.8291, -0.3368] \cup [0.2731, 1.8358]$	$[-0.9047, -0.1890] \cup [0.2866, 1.8910]$
8	$[-0.8323, -0.3366] \cup [0.2715, 1.8448]$	$[-0.9084, -0.1890] \cup [0.2849, 1.8986]$
10	$[-0.8334, -0.3366] \cup [0.2713, 1.8469]$	$[-0.9094, -0.1890] \cup [0.2848, 1.9006]$

TABLE 5.4
Bound for eigenvalues of preconditioned SGFEM system, $\sigma = 0.17$, $p = 3$.

Computed eigenvalue		
$l = 5$		
M	$\nu = .4$	$\nu = .49999$
5	$[-0.9291, -0.3178] \cup [0.2318, 1.9435]$	$[-0.9491, -0.1789] \cup [0.2428, 1.9935]$
8	$[-0.8797, -0.3171] \cup [0.2268, 1.9566]$	$[-0.9538, -0.1789] \cup [0.2358, 2.0052]$
10	$[-0.8817, -0.3169] \cup [0.2264, 1.9604]$	$[-0.9555, -0.1788] \cup [0.2352, 2.0086]$
$l = 6$		
5	$[-0.9206, -0.3176] \cup [0.2307, 1.9454]$	$[-0.9507, -0.1787] \cup [0.2413, 1.9964]$
8	$[-0.8836, -0.3167] \cup [0.2254, 1.9623]$	$[-0.9581, -0.1787] \cup [0.2346, 2.0126]$
10	$[-0.8857, -0.3166] \cup [0.2251, 1.9663]$	$[-0.9600, -0.1785] \cup [0.2336, 2.0167]$

Next, we consider the implementation and cost of the preconditioning scheme. To apply the preconditioner in each iteration, we need to invert two diagonal matrices of size $n_p n_y$ and apply the action of the inverse of $\tilde{\mathcal{A}}_{\text{approx},1}$. The latter is block diagonal, with $2 \cdot n_y$ copies of $\alpha \mathbb{A}_{11}$ on the main diagonal. Hence, we simply need to apply the action of $\mathbb{A}_{11}^{-1}$ on $2 \cdot n_y$ vectors. Since $\mathbb{A}_{11}$ represents the scalar Laplacian operator, there are many efficient ways to approximate the action of $\mathbb{A}_{11}^{-1}$ (e.g., using multigrid) with $\mathcal{O}(n_u)$ cost. We use a direct solver based on the MATLAB in-built `lu` factorization routine. Note that the factorization of $\mathbb{A}_{11}$ needs to be performed only *once*.

[3]The associated MINRES relative residual tolerance is set to 10^{-6}. Bounds are unchanged if we rerun the experiments with a tighter tolerance.

In Table 5.5 we record the number of MINRES iterations required to reduce the preconditioned residual error to 10^{-6} for the case $\sigma = 0.085$, with $p = 3$ fixed and varying M and ℓ. In Tables 5.6 and 5.7 we record the number of iterations required when $\sigma = 0.17$ with $p = 3$ and $p = 4$ fixed, respectively. The timings were recorded running S-IFISS on a MacBook Pro laptop with 16GB of memory and a 2.3GHz Intel Core i5 processor. We observe that for a fixed value of σ, the iteration counts remain stable as the discretization parameters ℓ and p are varied. Moreover, the iteration counts stay bounded when working with values of ν arbitrarily close to 1/2.

TABLE 5.5
MINRES iteration counts for stopping tolerance 10^{-6}*, and timings in seconds (in parentheses),* $\sigma = 0.085$*,* $p = 3$*.*

M	$\nu = .4$	$\nu = .49$	$\nu = .499$	$\nu = .4999$	$\nu = .49999$
			$l = 5$		
5	56(3.9)	74(5.3)	78(5.7)	78(5.7)	78(5.8)
8	56(10)	75(12.8)	78(13.7)	79(13.7)	79(13.4)
10	56(16.5)	75(22.5)	79(23.6)	79(23.4)	79(23.5)
			$l = 6$		
5	56(14.6)	75(19.7)	79(20.9)	79(20.5)	79(20.9)
8	56(45.2)	75(60.5)	79(64.2)	79(63.8)	79(64)
10	56(86)	75(114.3)	79(120.5)	79(118.1)	79(117.3)

TABLE 5.6
MINRES iteration counts for stopping tolerance 10^{-6}*, and timings in seconds (in parentheses),* $\sigma = 0.17$*,* $p = 3$*.*

M	$\nu = .4$	$\nu = .49$	$\nu = .499$	$\nu = .4999$	$\nu = .49999$
			$l = 5$		
5	66(5.4)	86(6.3)	90(6.8)	92(6.9)	92(6.9)
8	67(11.5)	88(15.4)	92(15.8)	93(16.4)	93(15.9)
10	67(19.9)	88(27)	93(28.4)	93(28.1)	93(28.6)
			$l = 6$		
5	66(18.3)	88(24.4)	92(25.5)	92(25.5)	92(25.5)
8	67(55.4)	88(70.7)	93(75.8)	93(75.4)	93(76.6)
10	67(102.4)	89(134.9)	93(140.4)	95(145)	95(142)

TABLE 5.7
MINRES iteration counts for stopping tolerance 10^{-6}*, and timings in seconds (in parentheses),* $\sigma = 0.17$*,* $p = 4$*.*

M	$\nu = .4$	$\nu = .49$	$\nu = .499$	$\nu = .4999$	$\nu = .49999$
			$l = 5$		
5	67(8.1)	90(11.6)	95(12)	95(12)	95(12)
8	70(34.4)	93(44.4)	97(45.6)	98(48.9)	98(48.1)
10	70(69.4)	93(94.1)	98(96.7)	98(96.5)	98(94.2)
			$l = 6$		
5	69(39.9)	91(50.6)	95(54.7)	96(54.6)	96(53.6)
8	70(176.8)	94(233.3)	98(249.4)	98(249.2)	98(250.6)
10	70(378.2)	94(513.9)	98(538.1)	98(538.7)	98(534.3)

We address an important question before concluding, namely, given the availability of robust nonintrusive methods for solving PDEs with parametric uncertainty, why is our approach even worth considering? Our preconditioning strategy requires $2 \cdot n_y \cdot n_{\text{iter}}$ decoupled solves with a symmetric and positive definite matrix of size n_u representing the scalar Laplacian operator, where n_{iter} is the required number of

iterations. It therefore can be implemented with standard off-the-shelf technology in a nonintrusive way. However, to test against a conventional sampling method, we also solved our three-field test problem using an off-the-shelf implementation of sparse grid interpolation. Specifically, we sampled the Young's modulus (5.1) at the sets of Clenshaw–Curtis sparse grid collocation points provided by the MATLAB toolbox SPINTERP [13, 14] and then solved the associated sequence of n_c (deterministic) saddle-point systems of dimension $n = 2(n_u + n_p)$ using the default sparse direct solver (\) that is built into MATLAB. Results are presented in Tables 5.8 and 5.9 for collocation levels 3 and 4. Here, setting the collocation "level" to be equal to p means that the resulting interpolation scheme is at least exact for polynomials of total degree p or less. The number of points, and hence the number of systems to solve, increases with the level number and the number of parameters M. The timings (using the same laptop) listed should be compared with those given in Tables 5.6 and 5.7, which are for the same value of σ, with $p = 3$ and 4, respectively.

TABLE 5.8

Number n_c of collocation points and timings (assembly time + solve time) in seconds for the case $\sigma = 0.17$ and collocation level 3.

Collocation level = 3		
M	$\nu = .4$	$\nu = .49999$
$l = 5$		
5	241 (14.5 + 8.8)	241 (16.1 + 9.4)
8	849 (56.3 + 30.6)	849 (54.1 + 31.0)
10	1,581 (106.3 + 54.6)	1,581 (107.1 + 57.2)
$l = 6$		
5	241 (32.3 + 41.8)	241 (33.5 + 42.4)
8	849 (140.0 + 149.3)	849 (140.8 + 151.4)
10	1,581 (297.3 + 269.4)	1,581 (298.0 + 273.1)

TABLE 5.9

Number n_c of collocation points and timings (assembly time + solve time) in seconds for the case $\sigma = 0.17$ and collocation level 4.

Collocation level = 4		
M	$\nu = .4$	$\nu = .49999$
$l = 5$		
5	801 (41.0 + 28.2)	801 (44.0 + 30.1)
8	3,937 (213.1 + 143.4)	3,937 (213.2 + 145.1)
10	8,801 (521.7 + 318.7)	8,801 (525.1 + 320.8)
$l = 6$		
5	801 (101.4 + 140)	801 (101.6 + 145.1)
8	3,937 (632.6 + 682.9)	3,937 (632.7 + 683.4)
10	8,801 (1,642.2 + 1,546.8)	8,801 (1,642.0 + 1,545.6)

The timings reported in Tables 5.8 and 5.9 are the times taken to assemble and solve all n_c saddle-point systems. Thanks to the stochastically linear nature of the three-field problem, it is not necessary to recompute the system matrix from scratch for each collocation point (which would be the case for the two-field problem), and we exploited this to give a fair comparison. The MATLAB sparse solver is highly tuned and incredibly efficient. For spatial refinement level $l = 6$ the discrete three-field system associated with one collocation point has dimension $n = 14{,}208$ (see Table 5.1). A preconditioned Krylov solver for such a saddle-point system is not able to compete with \ on such a small problem. An attractive feature of the sparse direct solver is that the timings are essentially unaffected when σ is varied. This is

in contrast to our mean-based preconditioned iterative solver which is needed for the much larger SGFEM saddle-point systems.

What is also evident is that the timings for the off-the-shelf nonintrusive approach (with the collocation level chosen to be equal to p) are considerably slower than the bespoke preconditioned solver times for the SGFEM systems. A direct comparison of the accuracy of the two approximation approaches is beyond the scope of this paper. However, a major advantage of using stochastic Galerkin approximation is that it enables one to derive efficient and reliable a posteriori error estimates (with two-sided error bounds). Moreover, one can design fully adaptive solution algorithms with adaptivity driven by estimates for the potential error reduction associated with proposed refinement schemes. This has been achieved for linear elasticity problems in the recent work [10], which exploits the preconditioning methodology developed herein. A rigorous underpinning for adaptive algorithms that combine spatial refinement with a posteriori parametric refinement is significantly more challenging to achieve when working outside of a Galerkin projection framework.

6. Conclusions. This work analyzes parameter-robust discretizations and the construction of preconditioners for linear elasticity problems with uncertain material parameters. Having introduced a new three-field formulation of the problem, it is rigorously shown that preconditioners that are based on mapping properties associated with a specific parameter-dependent norm are robust with respect to variations of the Poisson ratio, the choice of finite element spaces, as well as the discretization parameters. The theoretical results are confirmed by a systematically designed set of numerical experiments. There are several generalizable aspects of our work. The idea of introducing an additional auxiliary variable to avoid working with parameter-dependent coefficients that exhibit rational nonlinearities is applicable to other PDE problems. The ideas underpinning the construction of the block diagonal preconditioner apply to other PDEs with parameter-dependent coefficients where stochastic Galerkin approximation leads to a discrete problem with the structure (3.5). Finally, the preconditioning methodology gives a starting point for designing efficient solution algorithms for more challenging (e.g., nonlinear) elasticity models with uncertain material coefficients.

Acknowledgments. We are grateful to Eman Almoalim for her collocation software and we thank three anonymous referees for very constructive comments.

REFERENCES

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

[2] A. Bespalov, C. E. Powell, and D. J. Silvester, *A priori error analysis of stochastic Galerkin mixed approximations of elliptic PDEs with random data*, SIAM J. Numer. Anal., 50 (2012), pp. 2039–2063.

[3] S. Brenner and R. Scott, *The Mathematical Theory of Finite Element Methods*, Texts in Appl. Math. 15, Springer, New York, 2007.

[4] M. Eigel, C. J. Gittelson, C. Schwab, and E. Zander, *Adaptive stochastic Galerkin FEM*, Comput. Methods Appl. Mech. Engrg., 270 (2014), pp. 247–269.

[5] H. Elman, A. Ramage, and D. Silvester, *IFISS: A computational laboratory for investigating incompressible flow problems*, SIAM Rev., 56 (2014), pp. 261–273, https://doi.org/10.1137/120891393.

[6] H. Elman, D. Silvester, and A. Wathen, *Finite Elements and Fast Iterative Solvers: With Applications in Incompressible Fluid Dynamics*, 2nd ed., Oxford University Press, Oxford, UK, 2014.

[7] A. Ern and J.-L. Guermond, *Theory and Practice of Finite Elements*, Springer, New York, 2004.

[8] 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.

[9] L. R. Herrmann, *Elasticity equations for incompressible and nearly incompressible materials by a variational theorem*, AIAA J., 3 (1965), pp. 1896–1900.

[10] A. Khan, A. Bespalov, C. E. Powell, and D. J. Silvester, *Robust A Posteriori Error Estimation for Stochastic Galerkin Formulations of Parameter-Dependent Linear Elasticity Equations*, https://arxiv.org/abs/1810.07440, 2018.

[11] A. Khan, C. E. Powell, and D. J. Silvester, *Robust A Posteriori Error Estimators for Mixed Approximation of Nearly Incompressible Elasticity*, https://arxiv.org/abs/1710.03328, 2017.

[12] A. Klawonn, *An optimal preconditioner for a class of saddle point problems with a penalty term*, SIAM J. Sci. Comput., 19 (1998), pp. 540–552.

[13] A. Klimke, *Sparse Grid Interpolation Toolbox–User's Guide*, Technical report IANS report 2007/017, University of Stuttgart, 2007.

[14] A. Klimke and B. Wohlmuth, *Algorithm* 847: *SPINTERP: Piecewise multilinear hierarchical sparse grid interpolation in MATLAB*, ACM Trans. Math. Software, 31 (2005).

[15] K.-A. Mardal and R. Winther, *Preconditioning discretizations of systems of partial differential equations*, Numer. Linear Algebra Appl., 18 (2011), pp. 1–40.

[16] 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.

[17] C. E. Powell and D. J. Silvester, *Optimal preconditioning for Raviart–Thomas mixed formulation of second-order elliptic problems*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 718–738.

[18] T. Rusten and R. Winther, *A preconditioned iterative method for saddlepoint problems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 887–904.

[19] D. Silvester and V. Simoncini, *An optimal iterative solver for symmetric indefinite systems stemming from mixed approximation*, ACM Trans. Math. Software, 37 (2011). https://doi.org/10.1145/1916461.1916466.

[20] D. J. Silvester, A. Bespalov, and C. E. Powell, *Stochastic IFISS (S-IFISS) Version* 1.04, http://www.manchester.ac.uk/ifiss/sifiss.html (2017).