A preconditioned MINRES method for the coupling of mixed-FEM and BEM in nonlinear problems*

Gabriel N. Gatica[†] Norbert Heuer[†]

Abstract

We provide an efficient solution procedure for the linearized Galerkin schemes arising from the combined use of mixed finite elements and boundary elements to solve nonlinear problems. As a model, we consider a nonlinear-linear transmission problem appearing in electromagnetism and steady heat conduction. Since the corresponding continuous and discrete variational formulations become nonlinear two-fold saddle point problems (also called dual-dual formulations), we propose to apply Newton's method to the Galerkin schemes, thus yielding linear systems with the same dual-dual structure. Hence, we follow previous works on this kind of operator equation and derive a preconditioned minimum residual method that guarantees a bounded number of iterations to solve these systems.

Keywords: Minimum residual iteration, mixed finite elements, boundary elements, transmission problems.

Mathematics Subject Classification (1991): 65N30, 65N38, 65N22, 65F10

1 Introduction

In some recent papers we have developed preconditioned minimum residual (MINRES) methods for symmetric indefinite linear systems with a two-fold saddle point structure (also called dual-dual structure) (see [6], [7]). These systems arise from the linearized Galerkin schemes for some nonlinear boundary value problems that are solved via a dual-mixed finite element method, and also for some linear-nonlinear transmission problems that are solved through the combined use of the same dual-mixed finite element method with Dirichlet-to-Neumann mappings (see, e.g. [1], [9], [10]). In these cases, the specific form of the stiffness matrix for the corresponding linear Galerkin systems is:

$$\begin{bmatrix} A_1 & B_1^T & 0 \\ B_1 & 0 & B^T \\ 0 & B & 0 \end{bmatrix}, \tag{1}$$

*This research was partially supported by Fondecyt-Chile through research project 1980122 and FONDAP Program in Applied Mathematics.

†Departamento de Ingeniería Matemática, Universidad de Concepción, Casilla 160-C, Concepción, Chile, {ggatica,norbert}@ing-mat.udec.cl

where A_1 is symmetric and positive definite. In [6], [7] we obtained precise estimates for the spectrum of matrices with this dual-dual structure and provided abstract convergence results for the preconditioned MINRES method. Moreover, we presented specific preconditioners that lead to bounded iteration numbers for this procedure.

The purpose of the present work is to extend the results from [6] and [7] to the case in which a linear-nonlinear transmission problem is solved via the coupling of mixed finite elements and boundary elements. As shown in [8], the stiffness matrix for the linear discrete scheme now becomes:

$$\begin{bmatrix} A_1 & B_1^T & 0 \\ B_1 & -S & B^T \\ 0 & B & 0 \end{bmatrix}, \tag{2}$$

where S is a positive semi-definite matrix. As a model, we consider a transmission problem arising in electromagnetism and steady heat conduction. However, our results should be easily extended to the hyperelastic transmission problem studied in [8].

An outline of the paper is as follows. In Section 2 we introduce the model problem, derive the associated variational formulation (in the form of a nonlinear two-fold saddle point problem), and state its unique solvability. We remark that, differently from our previous settings (see, e.g. [6], [10]), we do not require here any integral mean zero in the definition of the corresponding spaces. Indeed, we stabilize the hypersingular boundary integral operator with a least squares term and observe that the utilization of the spaces $H(\mathrm{div};\Omega)$ and $H^{1/2}(\Gamma)$ suffices to set a well posed two-fold saddle point formulation. Then, in Section 3 we define explicit finite element subspaces and briefly analyze the corresponding Galerkin scheme. In particular, a quasi-optimal error estimate is provided. The minimum residual method for the linearized Galerkin systems, which have the form given by (2), is investigated in Section 4. Here we first analyze the spectrum of matrices with the dual-dual structure (2), thus extending the corresponding results from [6]. Then, similarly as in [6] and [7], we show that without preconditioner (that is, with pure scaling of the basis functions) the number of iterations of the MINRES, which are required to solve the linearized discrete model problem, grows at most like $O(h^{-3/2})$ (where h is the mesh size for the finite element spaces), and that it remains bounded if a suitable preconditioner is applied. Several numerical results illustrating these facts are reported in Section 5.

2 The model problem

Let $\Omega \subset \mathbf{R}^2$ be a polygonal domain with boundary Γ and let $\Omega_c = \mathbf{R}^2 \setminus \bar{\Omega}$ be its complement. For given $f \in L^2(\Omega)$, $g_1 \in H^{1/2}(\Gamma)$, $g_2 \in H^{-1/2}(\Gamma)$ and a nonlinear function $\boldsymbol{\kappa} : \Omega \times [0,\infty) \to \mathbf{R}$, our model transmission problem reads as: *Find $u \in H^1(\Omega)$ and $u_c \in H^1_{\mathrm{loc}}(\Omega_c)$ such that*

$$\mathrm{div}\,(\boldsymbol{\kappa}(\cdot, |\nabla u|)\,\nabla u) = -f \quad \text{in } \Omega, \tag{3}$$

$$\Delta u_c = 0 \quad \text{in } \Omega_c, \tag{4}$$

$$u = u_c + g_1, \qquad \boldsymbol{\kappa}(\cdot, |\nabla u|)\,\frac{\partial u}{\partial \mathbf{n}} = \frac{\partial u_c}{\partial \mathbf{n}} + g_2 \quad \text{on } \Gamma, \tag{5}$$

subject to the decay condition

$$u_c(x) = a \log |x| + o(1) \quad as \quad |x| \to \infty. \tag{6}$$

for a constant a. Here $\mathbf{n}$ denotes the unit outward normal to Γ.

We assume that $\boldsymbol{\kappa} \in C^1(\Omega \times [0,\infty))$ and that there exist constants κ_0, $\kappa_1 > 0$, such that for all $(x,t) \in \Omega \times [0,\infty)$

$$\kappa_0 \leq \boldsymbol{\kappa}(x,t) \leq \kappa_1 \quad \text{and} \quad \kappa_0 \leq \boldsymbol{\kappa}(x,t) + t\,\frac{\partial \boldsymbol{\kappa}}{\partial t}\,(x,t) \leq \kappa_1\,. \tag{7}$$

In the following we derive a mixed variational formulation of (3)–(5). We need the space

$$H(\mathrm{div};\Omega) := \left\{\boldsymbol{\tau} \in [L^2(\Omega)]^2 : \ \mathrm{div}\,\boldsymbol{\tau} \in L^2(\Omega)\right\} \tag{8}$$

with inner product

$$\langle \boldsymbol{\sigma}, \boldsymbol{\tau} \rangle_{H(\mathrm{div};\Omega)} := \int_\Omega \boldsymbol{\sigma} \cdot \boldsymbol{\tau}\, dx + \int_\Omega \mathrm{div}\,\boldsymbol{\sigma}\ \mathrm{div}\,\boldsymbol{\tau}\, dx,$$

and induced norm

$$\|\boldsymbol{\tau}\|_{H(\mathrm{div};\Omega)} := \left(\|\boldsymbol{\tau}\|^2_{[L^2(\Omega)]^2} + \|\mathrm{div}\,\boldsymbol{\tau}\|^2_{L^2(\Omega)}\right)^{1/2}.$$

We first introduce the auxiliary unknowns

$$\boldsymbol{\sigma} := \boldsymbol{\kappa}(\cdot, |\nabla u|)\, \nabla u$$

and

$$\boldsymbol{\theta} := \nabla u \quad \text{in } \Omega, \tag{9}$$

so that

$$\boldsymbol{\sigma} = \boldsymbol{\kappa}(\cdot, |\boldsymbol{\theta}|)\, \boldsymbol{\theta} \quad \text{in } \Omega, \tag{10}$$

and (3) becomes

$$\mathrm{div}\,\boldsymbol{\sigma} = -f \quad \text{in } \Omega. \tag{11}$$

Then we multiply (9) by a test function $\boldsymbol{\tau} \in H(\mathrm{div};\Omega)$, integrate by parts on Ω and make use of the transmission condition $u = u_c + g_1$ on Γ. This gives

$$\int_\Omega \boldsymbol{\theta} \cdot \boldsymbol{\tau}\, dx + \int_\Omega u\, \mathrm{div}\,\boldsymbol{\tau}\, dx - \langle u_c + g_1, \boldsymbol{\tau} \cdot \mathbf{n} \rangle_\Gamma = 0 \tag{12}$$

for all $\boldsymbol{\tau} \in H(\mathrm{div};\Omega)$. Here, $\langle \cdot, \cdot \rangle_\Gamma$ stands for the duality pairing between $H^{1/2}(\Gamma)$ and $H^{-1/2}(\Gamma)$ with respect to the $L^2(\Gamma)$-inner product. Now, testing (10) with a function $\boldsymbol{\zeta} \in [L^2(\Omega)]^2$, we obtain

$$\int_\Omega \boldsymbol{\kappa}(\cdot, |\boldsymbol{\theta}|)\, \boldsymbol{\theta} \cdot \boldsymbol{\zeta}\, dx - \int_\Omega \boldsymbol{\sigma} \cdot \boldsymbol{\zeta}\, dx = 0 \quad \forall \boldsymbol{\zeta} \in [L^2(\Omega)]^2 \tag{13}$$

and, testing (11),

$$\int_\Omega v\, \mathrm{div}\,\boldsymbol{\sigma}\, dx = -\int_\Omega f v\, dx \quad \forall v \in L^2(\Omega). \tag{14}$$

The homogeneous linear problem in the exterior domain Ω_c is dealt with by boundary integral operators on Γ. Denoting by $\mathbf{V}$, $\mathbf{K}$, $\mathbf{K}'$ and $\mathbf{W}$ the integral operators of the single layer, double layer, adjoint of the double layer and hypersingular potentials, respectively, and using the well-known jump conditions (cf., e.g., [15]) and the decay condition (6) we obtain

$$u_c = (\frac{1}{2}\mathbf{I} + \mathbf{K})\phi - \mathbf{V}(\frac{\partial u_c}{\partial \mathbf{n}}) \quad \text{on } \Gamma,$$

and

$$\mathbf{W}\phi + (\frac{1}{2}\mathbf{I} + \mathbf{K}')(\frac{\partial u_c}{\partial \mathbf{n}}) = 0 \quad \text{on } \Gamma\,,$$

where

$$\phi := u_c|_\Gamma \in H^{1/2}(\Gamma)$$

and $\mathbf{I}$ is the identity operator on the corresponding spaces. Using the transmission condition (5), the integral equations read as

$$u_c = (\frac{1}{2}\mathbf{I} + \mathbf{K})\phi - \mathbf{V}(\boldsymbol{\sigma} \cdot \mathbf{n}) + \mathbf{V}(g_2) \quad \text{on } \Gamma \tag{15}$$

and

$$\mathbf{W}\phi + (\frac{1}{2}\mathbf{I} + \mathbf{K}')(\boldsymbol{\sigma} \cdot \mathbf{n}) = (\frac{1}{2}\mathbf{I} + \mathbf{K}')(g_2) \quad \text{on } \Gamma. \tag{16}$$

Eventually, we substitute (15) into (12), and we test (16) with functions $\psi \in H^{1/2}(\Gamma)$. Then, collecting (12), (13), (14) and (16) we arrive at the following variational formulation of the boundary value problem (3)–(5): *Find* $(\boldsymbol{\theta}, \boldsymbol{\sigma}, \phi, u) \in [L^2(\Omega)]^2 \times H(\mathrm{div};\Omega) \times H^{1/2}(\Gamma)/_{\mathbf{R}} \times L^2(\Omega)$ *such that*

$$\int_\Omega \boldsymbol{\kappa}(\cdot, |\boldsymbol{\theta}|)\, \boldsymbol{\theta} \cdot \boldsymbol{\zeta}\, dx - \int_\Omega \boldsymbol{\sigma} \cdot \boldsymbol{\zeta}\, dx = 0,$$

$$\begin{aligned} -\int_\Omega \boldsymbol{\theta} \cdot \boldsymbol{\tau}\, dx - \langle \mathbf{V}(\boldsymbol{\sigma} \cdot \mathbf{n}), \boldsymbol{\tau} \cdot \mathbf{n} \rangle_\Gamma + \langle (\frac{1}{2}\mathbf{I} + \mathbf{K})\phi, \boldsymbol{\tau} \cdot \mathbf{n} \rangle_\Gamma \\ - \int_\Omega u \,\mathrm{div}\, \boldsymbol{\tau}\, dx = -\langle \mathbf{V} g_2, \boldsymbol{\tau} \cdot \mathbf{n} \rangle_\Gamma - \langle g_1, \boldsymbol{\tau} \cdot \mathbf{n} \rangle_\Gamma, \end{aligned}$$

$$\langle \psi, \mathbf{W}\phi \rangle_\Gamma + \langle \psi, (\frac{1}{2}\mathbf{I} + \mathbf{K}')(\boldsymbol{\sigma} \cdot \mathbf{n}) \rangle_\Gamma = \langle \psi, (\frac{1}{2}\mathbf{I} + \mathbf{K}') g_2 \rangle_\Gamma,$$

$$-\int_\Omega v \,\mathrm{div}\, \boldsymbol{\sigma}\, dx = \int_\Omega f v\, dx,$$

for all $(\boldsymbol{\zeta}, \boldsymbol{\tau}, \psi, v) \in [L^2(\Omega)]^2 \times H(\mathrm{div};\Omega) \times H^{1/2}(\Gamma) \times L^2(\Omega)$.

Note that, since $\mathbf{K}(1) = -1/2$ and $\mathbf{W}(1) = 0$, the functions $(\boldsymbol{\zeta}, \boldsymbol{\tau}, \psi, v) = (0, 0, const., 0)$ are in the kernel of the bilinear part of the above formulation. Therefore, $\phi \in H^{1/2}(\Gamma)$ is unique only in the quotient space $H^{1/2}(\Gamma)/_{\mathbf{R}}$. However, given $\phi \in H^{1/2}(\Gamma)/_{\mathbf{R}}$, $\boldsymbol{\sigma} \cdot \mathbf{n}|_\Gamma \in H^{-1/2}(\Gamma)$ and $g_2 \in H^{-1/2}(\Gamma)$ then $u_c|_\Gamma$ is uniquely determined by (15). Here we need that $\mathbf{V}$ is invertible. In fact, by an appropriate scaling of the domain we may assume that the operator $\mathbf{V}:\ H^{-1/2}(\Gamma)\to H^{1/2}(\Gamma)$ is positive definite.

To fix the constant in $H^{1/2}(\Gamma)/_{\mathbf{R}}$ one may choose $\phi\in H_0^{1/2}(\Gamma)$ (the subspace of functions with integral mean zero) or one may introduce a Lagrangian multiplier for ϕ. Here, we add a least squares term to the hypersingular operator $\mathbf{W}$. We define

$$P:\ H^{1/2}(\Gamma)\to\mathbf{R},\qquad P\phi:=\int_\Gamma \phi\,ds$$

and consider, instead of (16),

$$\mathbf{W}\phi+P'P\phi+(\frac{1}{2}\mathbf{I}+\mathbf{K}')(\boldsymbol{\sigma}\cdot\mathbf{n})=(\frac{1}{2}\mathbf{I}+\mathbf{K}')(g_2)\quad\text{on }\Gamma. \tag{17}$$

Representing the solution ϕ of (17) like $\phi=\phi_0+c_\phi$ with $P\phi_0=0$ and $c_\phi\in\mathbf{R}$ we find that $\phi_0\in H_0^{1/2}(\Gamma)$ solves (16), and therefore ϕ solves (16) as well. Thus, $c_\phi\in\mathbf{R}$ satisfies $P'Pc_\phi=0$ such that $\phi\in H_0^{1/2}(\Gamma)$ is unique.

Let us define the positive definite operator operator

$$\tilde{\mathbf{W}}:=\mathbf{W}+P'P:\ H^{1/2}(\Gamma)\to H^{-1/2}(\Gamma).$$

Then the actual variational formulation of the model problem we deal with reads as follows: *Find* $(\boldsymbol{\theta},\boldsymbol{\sigma},\phi,u)\in[L^2(\Omega)]^2\times H(\mathrm{div};\Omega)\times H^{1/2}(\Gamma)\times L^2(\Omega)$ *such that*

$$\int_\Omega \boldsymbol{\kappa}(\cdot,|\boldsymbol{\theta}|)\,\boldsymbol{\theta}\cdot\boldsymbol{\zeta}\,dx-\int_\Omega\boldsymbol{\sigma}\cdot\boldsymbol{\zeta}\,dx=0, \tag{18}$$

$$-\int_\Omega\boldsymbol{\theta}\cdot\boldsymbol{\tau}\,dx-\langle\mathbf{V}(\boldsymbol{\sigma}\cdot\mathbf{n}),\boldsymbol{\tau}\cdot\mathbf{n}\rangle_\Gamma+\langle(\frac{1}{2}\mathbf{I}+\mathbf{K})\phi,\boldsymbol{\tau}\cdot\mathbf{n}\rangle_\Gamma$$

$$-\int_\Omega u\,\mathrm{div}\,\boldsymbol{\tau}\,dx=-\langle\mathbf{V}g_2,\boldsymbol{\tau}\cdot\mathbf{n}\rangle_\Gamma-\langle g_1,\boldsymbol{\tau}\cdot\mathbf{n}\rangle_\Gamma, \tag{19}$$

$$\langle\psi,\tilde{\mathbf{W}}\phi\rangle_\Gamma+\langle\psi,(\frac{1}{2}\mathbf{I}+\mathbf{K}')(\boldsymbol{\sigma}\cdot\mathbf{n})\rangle_\Gamma=\langle\psi,(\frac{1}{2}\mathbf{I}+\mathbf{K}')g_2\rangle_\Gamma, \tag{20}$$

$$-\int_\Omega v\,\mathrm{div}\,\boldsymbol{\sigma}\,dx=\int_\Omega fv\,dx, \tag{21}$$

for all $(\boldsymbol{\zeta},\boldsymbol{\tau},\psi,v)\in[L^2(\Omega)]^2\times H(\mathrm{div};\Omega)\times H^{1/2}(\Gamma)\times L^2(\Omega)$.

We rewrite the model problem in a more abstract form. Indeed, following the notation from [9], we introduce the spaces

$$X_1:=[L^2(\Omega)]^2\times H^{1/2}(\Gamma),\quad M_1:=H(\mathrm{div};\Omega),\quad M:=L^2(\Omega),$$

and the operators

$$\begin{aligned}\mathbf{A}_1:\ X_1\to X_1',\qquad&\mathbf{B}_1:\ X_1\to M_1',\\ \mathbf{B}:\ M_1\to M',\qquad&\mathbf{S}:\ M_1\to M_1'\end{aligned}$$

as follows:

$$
\begin{aligned}
[\mathbf{A}_1(\boldsymbol{\theta},\phi),(\boldsymbol{\zeta},\psi)] &:= \int_\Omega \boldsymbol{\kappa}(\cdot,|\boldsymbol{\theta}|)\,\boldsymbol{\theta}\cdot\boldsymbol{\zeta}\,dx + \langle\psi,\tilde{\mathbf{W}}\phi\rangle_\Gamma,\\
[\mathbf{B}_1(\boldsymbol{\theta},\phi),\boldsymbol{\tau}] &:= -\int_\Omega \boldsymbol{\theta}\cdot\boldsymbol{\tau}\,dx + \langle(\frac{1}{2}\mathbf{I}+\mathbf{K})\phi,\boldsymbol{\tau}\cdot\mathbf{n}\rangle_\Gamma,\\
[\mathbf{B}(\boldsymbol{\sigma}),v] &:= -\int_\Omega v\,\mathrm{div}\,\boldsymbol{\sigma}\,dx,\\
[\mathbf{S}(\boldsymbol{\sigma}),\boldsymbol{\tau}] &:= \langle\mathbf{V}(\boldsymbol{\sigma}\cdot\mathbf{n}),\boldsymbol{\tau}\cdot\mathbf{n}\rangle_\Gamma
\end{aligned}
$$

where, hereafter, the brackets $[\cdot,\cdot]$ denote the duality pairings induced by the corresponding operators. Let us also introduce a generic null operator/functional $\mathbf{O}$, and the functionals $\mathbf{G}_1 \in X_1'$, $\mathbf{G} \in M_1'$, $\mathbf{F} \in M'$ where

$$
\begin{aligned}
\mathbf{G}_1(\boldsymbol{\zeta},\psi) &:= \langle\psi,(\frac{1}{2}\mathbf{I}+\mathbf{K}')g_2\rangle_\Gamma,\\
\mathbf{G}(\boldsymbol{\tau}) &:= -\langle\mathbf{V}g_2,\boldsymbol{\tau}\cdot\mathbf{n}\rangle_\Gamma - \langle g_1,\boldsymbol{\tau}\cdot\mathbf{n}\rangle_\Gamma,\\
\mathbf{F}(v) &:= \int_\Omega f v\,dx.
\end{aligned}
$$

Using these operators and functionals the variational problem (18)–(21) can be equivalently formulated as: *Find* $((\boldsymbol{\theta},\phi),\boldsymbol{\sigma},u) \in \mathbf{X} := X_1 \times M_1 \times M$ *such that*

$$
\mathcal{A}\begin{bmatrix} (\boldsymbol{\theta},\phi)\\ \boldsymbol{\sigma}\\ u \end{bmatrix} := \begin{bmatrix} \mathbf{A}_1 & \mathbf{B}_1' & \mathbf{O}\\ \mathbf{B}_1 & -\mathbf{S} & \mathbf{B}'\\ \mathbf{O} & \mathbf{B} & \mathbf{O} \end{bmatrix}\begin{bmatrix} (\boldsymbol{\theta},\phi)\\ \boldsymbol{\sigma}\\ u \end{bmatrix} = \begin{bmatrix} \mathbf{G}_1\\ \mathbf{G}\\ \mathbf{F} \end{bmatrix}. \tag{22}
$$

Note that $\mathbf{A}_1$ is a nonlinear operator and that $\mathbf{B}_1$, $\mathbf{B}$ and $\mathbf{S}$ are linear and bounded. The operators $\mathbf{B}_1' : M_1 \to X_1'$ and $\mathbf{B}' : M \to M_1'$ are the adjoints of $\mathbf{B}_1$ and $\mathbf{B}$, respectively.

Thus, there exist positive constants γ_1 and γ such that

$$
|\,[\mathbf{B}_1(\boldsymbol{\zeta},\psi),\boldsymbol{\tau}]\,| \le \gamma_1\,\|(\boldsymbol{\zeta},\psi)\|_{X_1}\,\|\boldsymbol{\tau}\|_{M_1} \quad \text{for all } ((\boldsymbol{\zeta},\psi),\boldsymbol{\tau}) \in X_1 \times M_1 \tag{23}
$$

and

$$
|\,[\mathbf{B}(\boldsymbol{\tau},t),v]\,| \le \gamma\,\|(\boldsymbol{\tau},t)\|_{M_1}\,\|v\|_M \quad \text{for all } ((\boldsymbol{\tau},t),v) \in M_1 \times M\,. \tag{24}
$$

Also, using the assumptions on $\boldsymbol{\kappa}$ (cf. (7)) and the boundedness and positive definiteness of $\tilde{\mathbf{W}}$, one can prove (see, e.g. Theorem 3.8 in [12]) that $\mathbf{A}_1$ is strongly monotone and Lipschitz continuous, that is there exist α_0, $\alpha_1 > 0$ such that

$$
[\mathbf{A}_1(\mathbf{t}) - \mathbf{A}_1(\mathbf{s}),\mathbf{t}-\mathbf{s}] \ge \alpha_0\,\|\mathbf{t}-\mathbf{s}\|_{X_1}^2 \tag{25}
$$

and

$$
\|\mathbf{A}_1(\mathbf{t}) - \mathbf{A}_1(\mathbf{s})\|_{X_1'} \le \alpha_1\,\|\mathbf{t}-\mathbf{s}\|_{X_1} \tag{26}
$$

for all $\mathbf{t} := (\boldsymbol{\theta},\phi)$, $\mathbf{s} := (\boldsymbol{\zeta},\psi) \in X_1$.

Consequently, applying the theory developed in [9], one can prove the unique solvability of (22).

Theorem 2.1 *There exists a unique* $((\boldsymbol{\theta},\phi),\boldsymbol{\sigma},u)\in\mathbf{X}$ *solution of* (22)*. Moreover, there exists* $C>0$*, independent of the solution, such that*

$$\|((\boldsymbol{\theta},\phi),\boldsymbol{\sigma},u)\|_{\mathbf{X}} \le C\,\left(\|f\|_{L^2(\Omega)}+\|g_1\|_{H^{1/2}(\Gamma)}+\|g_2\|_{H^{-1/2}(\Gamma)}\right).$$

Proof. It suffices to observe that $\mathbf{S}$ is positive semi-definite and verify that $\mathbf{B}_1$ and $\mathbf{B}$ satisfy suitable inf-sup conditions. We refer to Theorem 2.3 in [11] for details. $\square$

3 Galerkin approximation for the model problem

Let $\mathcal{T}_h$ be a quasi-uniform regular triangulation of Ω by triangles T of diameter h_T with mesh size $h:=\sup_{T\in\mathcal{T}_h} h_T$. Next, we consider the canonical triangle with vertices $\hat{P}_1=(0,0)^T$, $\hat{P}_2=(1,0)^T$, and $\hat{P}_3=(0,1)^T$ as a reference triangle $\hat{T}$, and introduce a family of affine mappings $\{F_T\}_{T\in\mathcal{T}_h}$, such that $F_T(\hat{T})=T$.

We now consider the lowest order Raviart-Thomas spaces. To this end let

$$\mathcal{RT}_0(\hat{T}) := \text{span}\left\{\begin{pmatrix}1\\0\end{pmatrix},\begin{pmatrix}0\\1\end{pmatrix},\begin{pmatrix}\hat{x}_1\\\hat{x}_2\end{pmatrix}\right\}, \tag{27}$$

and for each triangle $T\in\mathcal{T}_h$ we define

$$\mathcal{RT}_0(T) := \{\boldsymbol{\tau}:\ \boldsymbol{\tau}=\hat{\boldsymbol{\tau}}\circ F_T^{-1},\hat{\boldsymbol{\tau}}\in\mathcal{RT}_0(\hat{T})\}.$$

Then, the finite element subspaces for the unknown $\boldsymbol{\sigma}$ are

$$M_{1,h} := \{\boldsymbol{\tau}_h\in H(\text{div};\Omega):\ \boldsymbol{\tau}_h|_T\in\mathcal{RT}_0(T)\quad\forall T\in\mathcal{T}_h\}\,.$$

Note that the conformity condition for $H(\text{div};\Omega)$ requires continuity of the normal components across the sides of the triangles. To deal with the unknown $\boldsymbol{\theta}$ it suffices to consider piecewise constant functions. We take

$$X_{1,h}^{\boldsymbol{\theta}} := \left\{\boldsymbol{\tau}_h\in[L^2(\Omega)]^2:\ \boldsymbol{\tau}_h|_T \text{ is a constant vector}\quad\forall T\in\mathcal{T}_h\right\}\,.$$

To approximate ϕ we consider the space of continuous piecewise linear functions on the mesh $\mathcal{G}_h:=\mathcal{T}_h|_\Gamma$:

$$X_{1,h}^{\phi} := \left\{\psi_h\in C^0(\Gamma):\ \psi_h|_\gamma\in\mathbf{P}_1(\gamma)\ \forall\gamma\in\mathcal{G}_h\right\}.$$

Here and in the following $\mathbf{P}_p(\mathcal{O})$ denotes the set of polynomials of degree at most p on the domain $\mathcal{O}$. Finally, let

$$X_{1,h} := X_{1,h}^{\boldsymbol{\theta}}\times X_{1,h}^{\phi}$$

and consider piecewise constant functions as the finite element subspace for the unknown u, i.e.

$$M_h := \{v_h\in L^2(\Omega):\ v_h|_T\in\mathbf{P}_0(T)\quad\forall T\in\mathcal{T}_h\}.$$

Then the Galerkin procedure for approximately solving our model problem (3)–(5) is: *Find* $(\boldsymbol{\theta}_h, \phi_h) \in X_{1,h}$, $\boldsymbol{\sigma}_h \in M_{1,h}$, $u_h \in M_h$ *such that*

$$\left[\begin{pmatrix} \mathbf{A}_1 & \mathbf{B}_1' & \mathbf{O} \\ \mathbf{B}_1 & -\mathbf{S} & \mathbf{B}' \\ \mathbf{O} & \mathbf{B} & \mathbf{O} \end{pmatrix} \begin{pmatrix} (\boldsymbol{\theta}_h, \phi_h) \\ \boldsymbol{\sigma}_h \\ u_h \end{pmatrix}, \begin{pmatrix} (\boldsymbol{\zeta}_h, \psi_h) \\ \boldsymbol{\tau}_h \\ v_h \end{pmatrix} \right] = \mathbf{G}_1(\boldsymbol{\zeta}_h, \psi_h) + \mathbf{G}(\boldsymbol{\tau}_h) + \mathbf{G}(v_h) \tag{28}$$

for all $(\boldsymbol{\zeta}_h, \psi_h) \in X_{1,h}$, $\boldsymbol{\tau}_h \in M_{1,h}$, $v_h \in M_h$. Here, $\mathbf{A}_1$, $\mathbf{B}_1$, $\mathbf{B}$, $\mathbf{S}$, $\mathbf{G}_1$, $\mathbf{G}$ and $\mathbf{F}$ are defined as in Section 2. The main properties ensuring unique solvability of the discrete Galerkin systems are discrete inf-sup conditions for $\mathbf{B}$ and $\mathbf{B}_1$. We define the discrete kernel of $\mathbf{B}$ by

$$\tilde{M}_{1,h} := \{ \boldsymbol{\tau}_h \in M_{1,h} : \; [\mathbf{B}(\boldsymbol{\tau}_h), v_h] = 0 \; \forall \, v_h \in M_h \} \,,$$

and recall that there hold the following discrete inf-sup conditions (see, e.g., [10]): There exist positive constants β and β_1, not depending on h, such that

$$\sup_{\boldsymbol{\tau}_h \in M_{1,h} \setminus \{0\}} \frac{[\mathbf{B}(\boldsymbol{\tau}_h), v_h]}{\|\boldsymbol{\tau}_h\|_{M_1}} \geq \beta \|v_h\|_M \quad \text{for all } v_h \in M_h, \tag{29}$$

and

$$\begin{aligned} \sup_{(\boldsymbol{\zeta}_h, \psi_h) \in X_{1,h} \setminus \{0\}} \frac{[\mathbf{B}_1(\boldsymbol{\zeta}_h, \psi_h), \boldsymbol{\tau}_h]}{\|(\boldsymbol{\zeta}_h, \psi_h)\|_{X_1}} \quad &\geq \quad \sup_{\boldsymbol{\zeta}_h \in X^{\boldsymbol{\theta}}_{1,h} \setminus \{0\}} \frac{[\mathbf{B}_1(\boldsymbol{\zeta}_h, 0), \boldsymbol{\tau}_h]}{\|\boldsymbol{\zeta}_h\|_{[L^2(\Omega)]^2}} \\ &\geq \quad \beta_1 \|\boldsymbol{\tau}_h\|_{M_1} \quad \text{for all } \boldsymbol{\tau}_h \in \tilde{M}_{1,h}. \end{aligned} \tag{30}$$

Therefore, we obtain the following quasi-optimal error estimate.

Theorem 3.1 *There exists a unique solution* $((\boldsymbol{\theta}_h, \phi_h), \boldsymbol{\sigma}_h, u_h) \in X_{1,h} \times M_{1,h} \times M_h$ *of* (28). *Let* $((\boldsymbol{\theta}, \phi), \boldsymbol{\sigma}, u)$ *be the unique solution of* (22) *and assume that* $\boldsymbol{\theta}|_T \in [H^1(T)]^2 \; \forall T \in \mathcal{T}_h$, $\phi \in H^{3/2}(\Gamma)$, $\boldsymbol{\sigma} \in [H^1(\Omega)]^2$, $\operatorname{div} \boldsymbol{\sigma} \in H^1(\Omega)$, *and* $u \in H^1(\Omega)$. *Then, there exists a positive constant* C*, independent of* h*, such that*

$$\begin{aligned} &\|((\boldsymbol{\theta}, \phi), \boldsymbol{\sigma}, u) - ((\boldsymbol{\theta}_h, \phi_h), \boldsymbol{\sigma}_h, u_h)\|_{\mathbf{X}} \\ &\leq C\, h \left\{ \sum_{T \in \mathcal{T}_h} \|\boldsymbol{\theta}\|^2_{[H^1(T)]^2} + \|\phi\|^2_{H^{3/2}(\Gamma)} + \|\boldsymbol{\sigma}\|^2_{[H^1(\Omega)]^2} + \|\operatorname{div} \boldsymbol{\sigma}\|^2_{H^1(\Omega)} + \|u\|^2_{H^1(\Omega)} \right\}^{1/2} . \end{aligned}$$

Proof. See Theorems 3.2 and 3.3 in [11] for details. $\square$

4 Iterative solution of the linear system

4.1 Preliminaries

In the following we use plain symbols for matrices which correspond to operators with the analogous bold symbol. For a coefficient vector which belongs to a certain function we use the same symbol as for the function and add an arrow for clarity.

We solve (28) by the Newton scheme. Hence, in each iteration we solve a problem of the form (28) but with a **linear** operator $\mathbf{A}_1 : X_1 \to X_1'$. Thus, our assumptions (25) and (26) now become

$$[\mathbf{A}_1(\mathbf{s}), \mathbf{s}] \;\geq\; \alpha_0 \, \|\mathbf{s}\|_{X_1}^2 \tag{31}$$

and

$$\| \, \mathbf{A}_1(\mathbf{s})\|_{X_1'} \;\leq\; \alpha_1 \, \|\mathbf{s}\|_{X_1} \tag{32}$$

for all $\mathbf{s} := (\boldsymbol{\zeta}, \psi) \in X_1$. Consequently, in each iteration we have to solve a linear system of the structure

$$\mathcal{A} \begin{pmatrix} \vec{\mathbf{t}}_h \\ \vec{\boldsymbol{\sigma}}_h \\ \vec{u}_h \end{pmatrix} := \begin{pmatrix} A_1 & B_1^T & 0 \\ B_1 & -S & B^T \\ 0 & B & 0 \end{pmatrix} \begin{pmatrix} \vec{\mathbf{t}}_h \\ \vec{\boldsymbol{\sigma}}_h \\ \vec{u}_h \end{pmatrix} = \begin{pmatrix} G_1 \\ G \\ F \end{pmatrix}, \tag{33}$$

where $A_1 \in \mathbf{R}^{L \times L}$ is symmetric and positive definite. Moreover, $S \in \mathbf{R}^{M \times M}$ is symmetric positive semi-definite and the matrices $B_1 \in \mathbf{R}^{M \times L}$ and $B \in \mathbf{R}^{N \times M}$ have maximum rank. As usual, B_1^T and B^T denote the transposes of B_1 and B, respectively. The solution vector $(\vec{\mathbf{t}}_h, \vec{\boldsymbol{\sigma}}_h, \vec{u}_h)$ corresponds to a solution $(\mathbf{t}_h, \boldsymbol{\sigma}_h, u_h) := ((\boldsymbol{\theta}_h, \phi_h), \boldsymbol{\sigma}_h, u_h)$ within the Hilbert space $X_{1,h} \times M_{1,h} \times M_h$.

The matrix $\mathcal{A}$ is symmetric and indefinite as can be seen by Sylvester's law of inertia. Analogously to [6] we define the non-singular matrices

$$T_1 := \begin{pmatrix} I & 0 \\ -B_1 A_1^{-1} & I \end{pmatrix} \quad \text{and} \quad T := \begin{pmatrix} I & 0 \\ -(0 \; B)(T_1 A T_1^T)^{-1} & I \end{pmatrix}$$

where

$$A := \begin{pmatrix} A_1 & B_1^T \\ B_1 & -S \end{pmatrix}.$$

We obtain

$$T \begin{pmatrix} T_1 & 0 \\ 0 & I \end{pmatrix} \mathcal{A} \begin{pmatrix} T_1^T & 0 \\ 0 & I \end{pmatrix} T^T = \begin{pmatrix} A_1 & 0 & 0 \\ 0 & -B_1 A_1^{-1} B_1^T & 0 \\ 0 & 0 & B(B_1 A_1^{-1} B_1^T + S)^{-1} B^T \end{pmatrix}.$$

Note that, since A_1 is positive definite and since B_1 has maximum rank, the matrix $B_1 A_1^{-1} B_1^T$ is positive definite and therefore, $B_1 A_1^{-1} B_1^T + S$ is positive definite as well. Moreover, since also B has maximum rank, $B(B_1 A_1^{-1} B_1^T + S)^{-1} B^T$ is positive definite. Therefore, by the above transformation, Sylvester's law of inertia yields the existence of $L + N$ positive and M negative eigenvalues for $\mathcal{A}$.

We use the minimum residual method (MINRES) for the iterative solution of the linearized symmetric and indefinite system (33). The MINRES belongs to the family of Krylov subspace methods. Stable formulations of this method for symmetric and indefinite systems are given in [17, 2]. If we denote the iterates by $\vec{x}_k$, $k = 0, 1, \dots$, with corresponding residual vectors $\vec{r}_k$, there holds

$$\|\vec{r}_k\|_{l^2} = \|\vec{b} - \mathcal{A}\vec{x}_k\|_{l^2} = \min_{\vec{x} \in \vec{x}_0 + \mathcal{K}_k(\mathcal{A}, \vec{r}_0)} \|\vec{b} - \mathcal{A}\vec{x}\|_{l^2}$$

where $\mathcal{K}_k(\mathcal{A}, \vec{r}_0) = \operatorname{span}\{\vec{r}_0, \mathcal{A}\vec{r}_0, \dots, \mathcal{A}^{k-1}\vec{r}_0\}$ denotes the Krylov subspace and $\vec{r}_0$ is the initial residual $\vec{r}_0 = \vec{b} - \mathcal{A}\vec{x}_0$. For a symmetric, positive definite preconditioner $\mathcal{P}$ this relation becomes

$$\|b - \mathcal{A}\vec{x}_k\|_{\mathcal{P}^{-1}} = \min_{\vec{x} \in \vec{x}_0 + \mathcal{K}_k(\mathcal{P}^{-1}\mathcal{A}, \mathcal{P}^{-1}\vec{r}_0)} \|\vec{b} - \mathcal{A}\vec{x}\|_{\mathcal{P}^{-1}}$$

where $\|\vec{z}\|^2_{\mathcal{P}^{-1}} = \vec{z}^T \mathcal{P}^{-1} \vec{z}$, see [25, 24].

The following theorem gives an error estimate for the MINRES which is based on bounds for the spectrum of the iteration matrix. Therefore, it is clear that, in order to study the convergence of the MINRES for our model problem, we need to investigate the spectrum of matrices with the special dual-dual structure. This will be done in this section below.

Theorem 4.1 [16, 24] *Let the set of eigenvalues Λ of $\mathcal{P}^{-1}\mathcal{A}$ be such that $\Lambda \subset [-a, -bh^{2\alpha}] \cup [ch^{\alpha}, d]$ for positive constants a, b, c, d, a small parameter $h > 0$ and a constant $\alpha > 0$. Then there holds*

$$\frac{\|\vec{r}_k\|_{l^2}}{\|\vec{r}_0\|_{l^2}} = C(1 - kO(h^{3\alpha/2}))$$

for all integer k and a constant $C > 0$. That is, the number of iterations of the preconditioned MINRES which are required to solve $\mathcal{A}\vec{x} = \vec{b}$ up to a given accuracy is bounded by $O(h^{-3\alpha/2})$.

Let us consider a symmetric positive definite preconditioner of the general block form

$$\mathcal{P} = \begin{pmatrix} P_1 & 0 & 0 \\ 0 & P_2 & 0 \\ 0 & 0 & P_3 \end{pmatrix} \tag{34}$$

where P_1 is a preconditioner for A_1 and P_2 and P_3 are equivalent to the inner products in $M_{1,h}$ and M_h, respectively:

$$c_0 \vec{\boldsymbol{\tau}}^T P_2 \vec{\boldsymbol{\tau}} \le \|\boldsymbol{\tau}\|^2_{M_1} \le c_1 \vec{\boldsymbol{\tau}}^T P_2 \vec{\boldsymbol{\tau}} \quad \text{for any } \boldsymbol{\tau} \in M_{1,h}, \tag{35}$$

$$C_0 \vec{v}^T P_3 \vec{v} \le \|v\|^2_M \le C_1 \vec{v}^T P_3 \vec{v} \quad \text{for any } v \in M_h. \tag{36}$$

Here, $\vec{\boldsymbol{\tau}}$ and $\vec{v}$ are the vectors of coefficients of $\boldsymbol{\tau}$ and v, respectively, for the chosen bases in $M_{1,h}$ and M_h. We do not require uniform equivalence, i.e., the positive numbers c_0, c_1, C_0, and C_1 may depend on h.

In the symmetric form the preconditioned stiffness matrix looks like

$$\begin{aligned}\tilde{\mathcal{A}} := \mathcal{P}^{-1/2}\mathcal{A}\mathcal{P}^{-1/2} &= \begin{pmatrix} P_1^{-1/2}A_1P_1^{-1/2} & P_1^{-1/2}B_1^TP_2^{-1/2} & 0 \\ P_2^{-1/2}B_1P_1^{-1/2} & -P_2^{-1/2}SP_2^{-1/2} & P_2^{-1/2}B^TP_3^{-1/2} \\ 0 & P_3^{-1/2}BP_2^{-1/2} & 0 \end{pmatrix} \\ &=: \begin{pmatrix} \tilde{A}_1 & \tilde{B}_1^T & 0 \\ \tilde{B}_1 & -\tilde{S} & \tilde{B}^T \\ 0 & \tilde{B} & 0 \end{pmatrix}.\end{aligned}$$

4.2 Spectrum estimate

By the previous calculations we know that the spectrum Λ of $\tilde{\mathcal{A}}$ is contained in the set

$$\Lambda \subset [-a, -b] \cup [c, d]$$

where a, b, c, and d are positive numbers which might depend on h. Moreover we can count and order the eigenvalues of $\tilde{\mathcal{A}}$ like

$$\mu_{-M} \leq \mu_{-M+1} \leq \cdots \leq \mu_{-1} < 0 < \mu_1 \leq \mu_2 \leq \cdots \leq \mu_{L+N}.$$

We also need the eigenvalues of $\tilde{A}_1$,

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_L,$$

the eigenvalues of $\tilde{S}$,

$$0 \leq \rho_1 \leq \rho_2 \leq \cdots \leq \rho_M,$$

the singular values of $\tilde{B}$,

$$0 < \eta_1 \leq \eta_2 \leq \cdots \leq \eta_N$$

and those of $\tilde{B}_1$,

$$0 < \sigma_1 \leq \sigma_2 \leq \cdots \leq \sigma_M.$$

Indeed, $\eta_1 > 0$ by the inf-sup condition (29) which also means that $\tilde{B}$ has maximum rank, see Lemma 4.1 below. There also holds $\sigma_1 > 0$. However, essential is the minimum singular value σ_m of the restriction $\tilde{B}_1^T|_{\ker(\tilde{B})}$ onto the kernel of $\tilde{B}$:

$$\sigma_m := \min_{\boldsymbol{\tau} \in \ker(\tilde{B})\setminus\{0\}} \max_{\mathbf{s} \in X_{1,h}\setminus\{0\}} \frac{\vec{\boldsymbol{\tau}}^T \tilde{B}_1 \vec{\mathbf{s}}}{\sqrt{\vec{\boldsymbol{\tau}}^T\vec{\boldsymbol{\tau}}}\sqrt{\vec{\mathbf{s}}^T\vec{\mathbf{s}}}}. \tag{37}$$

Here and in the following we identify matrices with their corresponding linear mappings.

Lemma 4.1 [6, Lemma 1] *There holds*

$$\beta_1\sqrt{\frac{c_0\lambda_1}{\alpha_1}} \leq \sigma_m \leq \sigma_M \leq \gamma_1\sqrt{\frac{c_1\lambda_L}{\alpha_0}}, \tag{38}$$

$$\beta\sqrt{c_0C_0} \leq \eta_1 \leq \eta_N \leq \gamma\sqrt{c_1C_1}. \tag{39}$$

The next lemma generalizes Lemma 2 of [6] to our more general situation. It presents abstract estimates for the spectrum of a matrix that has the structure of $\tilde{\mathcal{A}}$.

Lemma 4.2 *There holds*

$$\mu^*_{-M} \le \mu_{-M}, \tag{40}$$

$$\mu_{-1} \le \max\left\{ \frac{\lambda_L}{2} - \sqrt{\frac{\lambda_L^2}{4} + \sigma_m^2}, -\eta_1 \right\}, \tag{41}$$

$$\min\{\lambda_1, \mu_1^*\} \le \mu_1, \tag{42}$$

$$\mu_{L+N} \le \frac{\lambda_L}{2} + \sqrt{\frac{\lambda_L^2}{4} + \sigma_M^2 + \eta_N^2} \tag{43}$$

where $\mu^*_{-M} < 0$ is the smallest root of the cubic function

$$\mu^3 - \mu^2(\lambda_1 - \rho_M) - \mu(\sigma_M^2 + \lambda_1\rho_M + \eta_N^2) + \lambda_1\eta_N^2 \quad (\mu \le 0)$$

and

$$\mu_1^* := \begin{cases} \frac{\sigma_M^2+\lambda_1\rho_M+\eta_1^2}{2(\rho_M-\lambda_1)} - \sqrt{\left(\frac{\sigma_M^2+\lambda_1\rho_M+\eta_1^2}{2(\rho_M-\lambda_1)}\right)^2 - \frac{\lambda_1\eta_1^2}{\rho_M-\lambda_1}} & \textit{if real and } \lambda_1 < \rho_M, \\ \frac{\rho_M\eta_1^2}{\sigma_M^2+\rho_M^2+\eta_1^2} & \textit{if } \lambda_1 = \rho_M, \\ -\frac{\sigma_M^2+\lambda_1\rho_M+\eta_1^2}{2(\lambda_1-\rho_M)} + \sqrt{\left(\frac{\sigma_M^2+\lambda_1\rho_M+\eta_1^2}{2(\lambda_1-\rho_M)}\right)^2 + \frac{\lambda_1\eta_1^2}{\lambda_1-\rho_M}} & \textit{if } \lambda_1 > \rho_M, \\ \lambda_1 & \textit{else.} \end{cases}$$

Proof. Let us assume that μ is an eigenvalue of $\tilde{\mathcal{A}}$. Then there exists $(\vec{\mathbf{s}}, \vec{\boldsymbol{\tau}}, \vec{v}) \in \mathbf{R}^L \times \mathbf{R}^M \times \mathbf{R}^N$ which is not zero such that

$$\begin{array}{ccccccc} \tilde{A}_1\vec{\mathbf{s}} & + & \tilde{B}_1^T\vec{\boldsymbol{\tau}} & & & = & \mu\vec{\mathbf{s}}, \\ \tilde{B}_1\vec{\mathbf{s}} & - & \tilde{S}\vec{\boldsymbol{\tau}} & + & \tilde{B}^T\vec{v} & = & \mu\vec{\boldsymbol{\tau}}, \\ & & \tilde{B}\vec{\boldsymbol{\tau}} & & & = & \mu\vec{v}. \end{array} \tag{44}$$

First let us bound the negative eigenvalues of $\tilde{\mathcal{A}}$. Thus we assume $\mu < 0$ and then $\vec{\boldsymbol{\tau}} \neq 0$ by (44.a) and (44.c) since $\tilde{A}_1$ is positive definite. If $\mu < 0$ then $\tilde{A}_1 - \mu I$ is positive definite as well. By using (44.a) we obtain

$$\vec{\mathbf{s}} = -(\tilde{A}_1 - \mu I)^{-1}\tilde{B}_1^T\vec{\boldsymbol{\tau}}$$

and (44.b) then gives

$$\tilde{B}_1(\tilde{A}_1 - \mu I)^{-1}\tilde{B}_1^T\vec{\boldsymbol{\tau}} + \mu\vec{\boldsymbol{\tau}} + \tilde{S}\vec{\boldsymbol{\tau}} - \tilde{B}^T\vec{v} = 0.$$

Inserting $\vec{v} = \frac{1}{\mu}\tilde{B}\vec{\boldsymbol{\tau}}$ from (44.c) and multiplying by $\mu\vec{\boldsymbol{\tau}}^T$ this yields

$$\mu\vec{\boldsymbol{\tau}}^T\tilde{B}_1(\tilde{A}_1 - \mu I)^{-1}\tilde{B}_1^T\vec{\boldsymbol{\tau}} + \mu^2\vec{\boldsymbol{\tau}}^T\vec{\boldsymbol{\tau}} + \mu\vec{\boldsymbol{\tau}}^T\tilde{S}\vec{\boldsymbol{\tau}} - \vec{\boldsymbol{\tau}}^T\tilde{B}^T\tilde{B}\vec{\boldsymbol{\tau}} = 0. \tag{45}$$

Therefore, since $\mu < 0$ and $(\tilde{A}_1 - \mu I)^{-1} > 0$,

$$0 \geq \mu \sigma_M^2 (\lambda_1 - \mu)^{-1} + \mu^2 + \mu \rho_M - \eta_N^2.$$

We multiply this inequality by $-(\lambda_1 - \mu)$, which is negative, and obtain

$$0 \leq \mu^3 - \mu^2(\lambda_1 - \rho_M) - \mu(\sigma_M^2 + \lambda_1 \rho_M + \eta_N^2) + \lambda_1 \eta_N^2.$$

The cubic bound above eventually becomes negative for $\mu < 0$ small enough. Therefore, the minimum eigenvalue of $\tilde{\mathcal{A}}$ is bounded from below by the minimum root of that cubic expression. This proves (40).

Let us show the estimate (41). We consider the cases $\boldsymbol{\tau} \in \ker(\tilde{B})$ and $\boldsymbol{\tau} \notin \ker(\tilde{B})$. In the first case we conclude that $\vec{\boldsymbol{\tau}}^T \tilde{B}_1 \tilde{B}_1^T \vec{\boldsymbol{\tau}} \geq \sigma_m^2 \vec{\boldsymbol{\tau}}^T \vec{\boldsymbol{\tau}}$ by the definition of σ_m. Note that $\sigma_m > 0$ by Lemma 4.1. Therefore, by again using (45) and the positive semi-definiteness of $\tilde{S}$, and remembering that $\mu < 0$, we find that there holds

$$0 \leq \mu \sigma_m^2 (\lambda_L - \mu)^{-1} + \mu^2.$$

Analogously as before, and as in the proof of Lemma 2 in [6], this gives

$$\mu \leq \frac{\lambda_L}{2} - \sqrt{\frac{\lambda_L^2}{4} + \sigma_m^2}. \tag{46}$$

Now, for $\boldsymbol{\tau} \notin \ker(\tilde{B})$, there holds $\vec{\boldsymbol{\tau}}^T \tilde{B}^T \tilde{B} \vec{\boldsymbol{\tau}} \geq \eta_1^2 \vec{\boldsymbol{\tau}}^T \vec{\boldsymbol{\tau}}$ and from (45) we obtain

$$0 \leq \mu^2 - \eta_1^2,$$

i.e., for negative μ,

$$\mu \leq -\eta_1. \tag{47}$$

Combining (46) and (47) we obtain (41).

It remains to bound the positive eigenvalues of $\tilde{\mathcal{A}}$. First let us consider the case $\mu > 0$ and $\boldsymbol{\tau} \in \ker(\tilde{B})$. Then, due to (44.c), there holds $\vec{v} = 0$ and the eigenvalue problem (44) reduces to

$$\begin{aligned} \tilde{A}_1 \vec{\mathbf{s}} + \tilde{B}_1^T \vec{\boldsymbol{\tau}} &= \mu \vec{\mathbf{s}}, \\ \tilde{B}_1 \vec{\mathbf{s}} - \tilde{S} \vec{\boldsymbol{\tau}} &= \mu \vec{\boldsymbol{\tau}}. \end{aligned} \tag{48}$$

By combining (37) and Lemma 4.1 we know that $\tilde{B}_1$ satisfies an inf-sup condition on $\ker(\tilde{B})$. Therefore, we can use the results of Silvester and Wathen [18, Lemma 2.2] to estimate the positive eigenvalues μ of the eigenvalue problem (48). This yields

$$\lambda_1 \leq \mu \leq \frac{\lambda_L}{2} + \sqrt{\frac{\lambda_L^2}{4} + \sigma_M^2} \quad \left(\mu > 0, \boldsymbol{\tau} \in \ker(\tilde{B})\right). \tag{49}$$

Now let us assume that $\mu > 0$ and $\boldsymbol{\tau} \notin \ker(\tilde{B})$. Evidently we then also have $\vec{\boldsymbol{\tau}} \neq 0$. Moreover, we assume that $\mu \notin [\lambda_1, \lambda_L]$. Then $\tilde{A}_1 - \mu I$ is invertible and (45) holds in this case as well:

$$\mu\vec{\boldsymbol{\tau}}^T \tilde{B}_1(\tilde{A}_1 - \mu I)^{-1}\tilde{B}_1^T\vec{\boldsymbol{\tau}} + \mu^2\vec{\boldsymbol{\tau}}^T\vec{\boldsymbol{\tau}} + \mu\vec{\boldsymbol{\tau}}^T\tilde{S}\vec{\boldsymbol{\tau}} - \vec{\boldsymbol{\tau}}^T\tilde{B}^T\tilde{B}\vec{\boldsymbol{\tau}} = 0$$
$$\Big(\mu \notin [\lambda_1, \lambda_L], \boldsymbol{\tau} \notin \ker(\tilde{B})\Big)\,. \tag{50}$$

We consider the cases $0 < \mu < \lambda_1$ and $\mu > \lambda_L$ separately. If $0 < \mu < \lambda_1$ then $\tilde{A}_1 - \mu I$ is positive definite and (50) yields

$$0 \leq \mu\sigma_M^2(\lambda_1 - \mu)^{-1} + \mu^2 + \mu\rho_M - \eta_1^2.$$

Multiplying this by $-(\lambda_1 - \mu)$, which is negative, we obtain

$$\begin{aligned} 0 &\geq \mu^3 - \mu^2(\lambda_1 - \rho_M) - \mu(\sigma_M^2 + \lambda_1\rho_M + \eta_1^2) + \lambda_1\eta_1^2 \\ &\geq -\mu^2(\lambda_1 - \rho_M) - \mu(\sigma_M^2 + \lambda_1\rho_M + \eta_1^2) + \lambda_1\eta_1^2. \end{aligned} \tag{51}$$

If $\lambda_1 < \rho_M$ this means

$$\mu^2 - \mu\frac{\sigma_M^2 + \lambda_1\rho_M + \eta_1^2}{\rho_M - \lambda_1} + \frac{\lambda_1\eta_1^2}{\rho_M - \lambda_1} \leq 0,$$

i.e.,

$$\begin{aligned} \mu \ \geq \ & \frac{\sigma_M^2 + \lambda_1\rho_M + \eta_1^2}{2(\rho_M - \lambda_1)} - \sqrt{\left(\frac{\sigma_M^2 + \lambda_1\rho_M + \eta_1^2}{2(\rho_M - \lambda_1)}\right)^2 - \frac{\lambda_1\eta_1^2}{\rho_M - \lambda_1}} \\ & \Big(0 < \mu < \lambda_1 < \rho_M, \boldsymbol{\tau} \notin \ker(\tilde{B})\Big)\,. \end{aligned} \tag{52}$$

Of course, this makes only sense if the right hand side is a real number (which then is positive). If this is not the case then the assumptions made before lead to a contradiction and, thus, $\lambda_1 \geq \rho_M$ or $\mu \geq \lambda_1$. We continue to consider the situation $0 < \mu < \lambda_1$. If $\lambda_1 = \rho_M$ then (51) gives

$$-\mu(\sigma_M^2 + \lambda_1\rho_M + \eta_1^2) + \lambda_1\eta_1^2 \leq 0,$$

i.e.,

$$\mu \geq \frac{\lambda_1\eta_1^2}{\sigma_M^2 + \lambda_1\rho_M + \eta_1^2} = \frac{\rho_M\eta_1^2}{\sigma_M^2 + \rho_M^2 + \eta_1^2} \quad \Big(0 < \mu < \lambda_1 = \rho_M, \boldsymbol{\tau} \notin \ker(\tilde{B})\Big)\,. \tag{53}$$

If $\lambda_1 > \rho_M$, for $0 < \mu < \lambda_1$, then we obtain from (51)

$$\mu^2 + \mu\frac{\sigma_M^2 + \lambda_1\rho_M + \eta_1^2}{\lambda_1 - \rho_M} - \frac{\lambda_1\eta_1^2}{\lambda_1 - \rho_M} \geq 0,$$

i.e.,

$$\mu \geq -\frac{\sigma_M^2 + \lambda_1\rho_M + \eta_1^2}{2(\lambda_1 - \rho_M)} + \sqrt{\left(\frac{\sigma_M^2 + \lambda_1\rho_M + \eta_1^2}{2(\lambda_1 - \rho_M)}\right)^2 + \frac{\lambda_1\eta_1^2}{\lambda_1 - \rho_M}}$$

$$\Big(0 < \mu, \rho_M < \lambda_1, \boldsymbol{\tau} \notin \ker(\tilde{B})\Big)\,.$$

Together with (49), (52) and (53) this proves (42).

Now, for $\mu > \lambda_L$ and $\boldsymbol{\tau} \notin \ker(\tilde{B})$, $\tilde{A}_1 - \mu I$ is negative definite and since $\tilde{S}$ is positive semi-definite (50) gives

$$0 \geq \mu\sigma_M^2(\lambda_L - \mu)^{-1} + \mu^2 - \eta_N^2.$$

This relation already appeared in the proof of Lemma 2 in [6]. It yields, for $\mu > 0$,

$$\mu < \frac{\lambda_L}{2} + \sqrt{\frac{\lambda_L^2}{4} + \sigma_M^2 + \eta_N^2} \quad \left(\mu > \lambda_L, \boldsymbol{\tau} \notin \ker(\tilde{B})\right).$$

Together with (49) this proves (43). □

4.3 Application to the model problem

Now we apply the previous technical results to estimate the spectrum of the linearized stiffness matrix $\mathcal{A}$ of our problem (28). First we consider the case where no preconditioner is in use. Then, $P_1 = I \in \mathbf{R}^{L\times L}$, $P_2 = I \in \mathbf{R}^{M\times M}$, $P_3 = I \in \mathbf{R}^{N\times N}$ in (34), and from (35), (36) and the definition of $\lambda_1, \ldots, \lambda_L$ as the eigenvalues of $\tilde{A}_1 = P_1^{-1/2} A_1 P_1^{-1/2} = A_1$ it is clear that the scaling of basis functions influences the convergence of the MINRES. Here, $L = \dim X_{1,h}$, $M = \dim M_{1,h}$, $N = \dim M_h$. We choose the scalings such that the outer bounds of the spectrum of $\mathcal{A}$ will appear to be independent of h. Then, Theorem 4.1 is directly applicable giving a bound for the number of iterations of the MINRES.

- $X_{1,h}^{\boldsymbol{\theta}}$: the space of piecewise constant vectors. We scale such that the L^∞-norm of the basis functions is $O(h^{-1})$.
- $X_{1,h}^{\phi}$: this space consists of piecewise linear functions with respect to the partition $\mathcal{G}_h$ of Γ. We scale to $O(1)$.
- $M_{1,h}$: this is the Raviart-Thomas space of lowest order. We scale to $O(1)$.
- M_h: the space of piecewise constant functions on Ω. We scale to $O(h^{-1})$.

Before analyzing the spectrum of the linearized stiffness matrix $\mathcal{A}$ we give bounds for the eigenvalues and singular values of the individual contributions, for our model problem, under the chosen scalings of basis functions.

Lemma 4.3 *There holds*

$$\lambda_1 = O(h), \quad \lambda_L = O(1), \tag{54}$$
$$\rho_1 \geq 0, \quad \rho_M = O(h), \tag{55}$$
$$\sigma_m = O(h), \quad \sigma_M = O(h), \tag{56}$$
$$\eta_1 \geq O(h), \quad \eta_N = O(1). \tag{57}$$

Here, equality means that the given orders of h are sharp.

Proof. The numbers λ_1 and λ_L are the extremum eigenvalues of the matrix A_1 which is block-diagonal with two blocks consisting of a mass matrix for the space $X_{1,h}^{\boldsymbol{\theta}}$ and the discretization of the stabilized hypersingular operator $\tilde{\mathbf{W}}$ for the space $X_{1,h}^{\phi}$. Similar as in [6] for the hypersingular operator $\mathbf{W}$, one can estimate the spectrum of the matrix corresponding to $\tilde{\mathbf{W}} = \mathbf{W} + P'P$. Since $\langle \psi, \mathbf{W} + P'P\psi\rangle_\Gamma = \langle \psi, \mathbf{W}\psi\rangle_\Gamma + \langle P\psi, P\psi\rangle_\Gamma$ is equivalent to $\|\psi\|^2_{H^{1/2}(\Gamma)}$ one obtains by the inverse property

$$c\frac{\|\psi\|^2_{L^2(\Gamma)}}{\vec{\psi}^T\vec{\psi}} \le \frac{\langle \psi, \tilde{\mathbf{W}}\psi\rangle_\Gamma}{\vec{\psi}^T\vec{\psi}} \le Ch^{-1}\frac{\|\psi\|^2_{L^2(\Gamma)}}{\vec{\psi}^T\vec{\psi}} \quad \forall \psi \in X_{1,h}^{\phi}.$$

The constants c, C are independent of h. By the scaling $O(1)$ of the basis functions in $X_{1,h}^{\phi}$ we find that $\|\psi\|^2_{L^2(\Gamma)}$ is equivalent to h times the l_2-norm squared of the corresponding coefficient vector. Therefore

$$ch \le \frac{\langle \psi, \tilde{\mathbf{W}}\psi\rangle_\Gamma}{\vec{\psi}^T\vec{\psi}} \le C \quad \forall \psi \in X_{1,h}^{\phi}.$$

To show that the lower bound above is sharp we choose $\psi^* = 1$ on Γ. By the scaling $O(1)$ of the basis functions in $X_{1,h}^{\phi}$ we then have $\vec{\psi}^* = O(1)\,(1,\ldots,1)^T$, i.e., $(\vec{\psi}^*)^T\vec{\psi}^* \sim h^{-1}$. Therefore, the minimum eigenvalue of the matrix corresponding to $\tilde{\mathbf{W}}$ is bounded like

$$\min_{\psi\in X_{1,h}^{\phi}\setminus\{0\}} \frac{\langle \psi, \tilde{\mathbf{W}}\psi\rangle_\Gamma}{\vec{\psi}^T\vec{\psi}} \le \frac{\langle \psi^*, \tilde{\mathbf{W}}\psi^*\rangle_\Gamma}{(\vec{\psi}^*)^T\vec{\psi}^*} = \frac{\langle P(1), P(1)\rangle_\Gamma}{(\vec{\psi}^*)^T\vec{\psi}^*} \sim h.$$

Here, we also used the relation $\mathbf{W}(1) = 0$ such that $\tilde{\mathbf{W}}(\psi^*) = P'P(\psi^*)$. Now, since the mass matrix for the space $X_{1,h}^{\boldsymbol{\theta}}$ with scaling $O(h^{-1})$ is spectrally equivalent to $O(1)$, we obtain the sharp result (54).

To estimate the minimum eigenvalue of S from below it suffices to remark that the operator $\mathbf{V}:\ H^{-1/2}(\Gamma) \to H^{1/2}(\Gamma)$ is positive semi-definite. Then $\rho_1 \ge 0$ by the relation

$$[\mathbf{S}(\boldsymbol{\tau}), \boldsymbol{\tau}] = \langle \mathbf{V}(\boldsymbol{\tau}\cdot\mathbf{n}), \boldsymbol{\tau}\cdot\mathbf{n}\rangle_\Gamma \quad \text{for all } \boldsymbol{\tau} \in M_{1,h}$$

since $\boldsymbol{\tau}\cdot\mathbf{n} \in H^{-1/2}(\Gamma)$ by the trace theorem for $H(\mathrm{div};\Omega)$.

To estimate ρ_M we use the boundedness of $\mathbf{V}:\ H^{-1/2}(\Gamma) \to H^{1/2}(\Gamma)$ and write for $\boldsymbol{\tau} \in M_{1,h}$

$$\frac{[\mathbf{S}(\boldsymbol{\tau}), \boldsymbol{\tau}]}{\vec{\boldsymbol{\tau}}^T\vec{\boldsymbol{\tau}}} = \frac{\langle \mathbf{V}(\boldsymbol{\tau}\cdot\mathbf{n}), \boldsymbol{\tau}\cdot\mathbf{n}\rangle_\Gamma}{\vec{\boldsymbol{\tau}}^T\vec{\boldsymbol{\tau}}} \le c\frac{\|\boldsymbol{\tau}\cdot\mathbf{n}|_\Gamma\|^2_{H^{-1/2}(\Gamma)}}{\vec{\boldsymbol{\tau}}^T\vec{\boldsymbol{\tau}}} \le c\frac{\|\boldsymbol{\tau}\cdot\mathbf{n}|_\Gamma\|^2_{L^2(\Gamma)}}{\vec{\boldsymbol{\tau}}^T\vec{\boldsymbol{\tau}}}$$

which is bounded by $O(h)$ since we scale to $O(1)$ within $M_{1,h}$. Therefore, $\rho_M \le O(h)$. To show that this estimate is sharp we choose a function $\boldsymbol{\tau}^* \in M_{1,h}$ with zero normal components on interior sides and such that $\boldsymbol{\tau}^*\cdot\mathbf{n}|_\Gamma = 1$. Then we use that $\mathbf{V}$ is positive definite and obtain

$$\rho_M = \max_{\boldsymbol{\tau}\in M_{1,h}\setminus\{0\}} \frac{[\mathbf{S}(\boldsymbol{\tau}), \boldsymbol{\tau}]}{\vec{\boldsymbol{\tau}}^T\vec{\boldsymbol{\tau}}} \ge \frac{[\mathbf{S}(\boldsymbol{\tau}^*), \boldsymbol{\tau}^*]}{(\vec{\boldsymbol{\tau}}^*)^T\vec{\boldsymbol{\tau}}^*} \ge c\frac{\|\boldsymbol{\tau}^*\cdot\mathbf{n}|_\Gamma\|^2_{H^{-1/2}(\Gamma)}}{(\vec{\boldsymbol{\tau}}^*)^T\vec{\boldsymbol{\tau}}^*} \sim h.$$

Here, c is a generic positive constant that does not depend on h. Note that $\|\boldsymbol{\tau}^* \cdot \mathbf{n}|_\Gamma\|^2_{H^{-1/2}(\Gamma)}$ is a constant and that $(\vec{\boldsymbol{\tau}}^*)^T \vec{\boldsymbol{\tau}}^* = O(h^{-1})$ for the scaling $O(1)$ since only $O(h^{-1})$ coefficients are non-zero. Therefore, $\rho_M = O(h)$ which proves (55).

Now we consider the singular values of B_1 and B. The scalings used in this example correspond to those in [6] and from that reference we already know that $c_0 = O(h^2)$, $c_1 = O(1)$ (scaling constants in $M_{1,h}$, see (35)) and $C_0 = O(1)$, $C_1 = O(1)$ (scaling constants in M_h, see (36)).

For the singular values of B_1 we proceed as follows. Using Lemma 4.1 and the inf-sup condition (30) we obtain $O(h^{3/2}) \leq \sigma_m \leq \sigma_M \leq O(1)$. For our example, however, both estimates are not sharp and need to be improved.

Functions of the discrete kernel of the operator $\mathbf{B}$ are divergence free, see, e.g., [7]. Therefore, $\|\boldsymbol{\tau}\|_{H(\mathrm{div};\Omega)} = \|\boldsymbol{\tau}\|_{[L^2(\Omega)]^2}$ for all $\boldsymbol{\tau} \in \tilde{M}_{1,h}$, the discrete kernel of $\mathbf{B}$. This yields

$$
\begin{aligned}
\sigma_m &= \min_{\boldsymbol{\tau} \in \tilde{M}_{1,h} \setminus \{0\}} \max_{\mathbf{s} \in X_{1,h} \setminus \{0\}} \frac{\vec{\boldsymbol{\tau}}^T B_1 \vec{\mathbf{s}}}{\sqrt{\vec{\boldsymbol{\tau}}^T \vec{\boldsymbol{\tau}}} \sqrt{\vec{\mathbf{s}}^T \vec{\mathbf{s}}}} \geq \min_{\boldsymbol{\tau} \in \tilde{M}_{1,h} \setminus \{0\}} \max_{\boldsymbol{\zeta} \in X^{\boldsymbol{\theta}}_{1,h} \setminus \{0\}} \frac{-\int_\Omega \boldsymbol{\zeta} \cdot \boldsymbol{\tau}\, dx}{\sqrt{\vec{\boldsymbol{\tau}}^T \vec{\boldsymbol{\tau}}} \sqrt{\vec{\boldsymbol{\zeta}}^T \vec{\boldsymbol{\zeta}}}} \\
&= \min_{\boldsymbol{\tau} \in \tilde{M}_{1,h} \setminus \{0\}} \max_{\boldsymbol{\zeta} \in X^{\boldsymbol{\theta}}_{1,h} \setminus \{0\}} \frac{-\int_\Omega \boldsymbol{\zeta} \cdot \boldsymbol{\tau}\, dx}{\|\boldsymbol{\tau}\|_{H(\mathrm{div};\Omega)} \|\boldsymbol{\zeta}\|_{[L^2(\Omega)]^2}} \frac{\|\boldsymbol{\tau}\|_{[L^2(\Omega)]^2} \|\boldsymbol{\zeta}\|_{[L^2(\Omega)]^2}}{\sqrt{\vec{\boldsymbol{\tau}}^T \vec{\boldsymbol{\tau}}} \sqrt{\vec{\boldsymbol{\zeta}}^T \vec{\boldsymbol{\zeta}}}} \\
&\geq \beta_1 O(h).
\end{aligned}
$$

Here we used the discrete inf-sup property (30) of $\mathbf{B}_1$ and the equivalencies of norms $\|\boldsymbol{\tau}\|_{[L^2(\Omega)]^2} \simeq O(h)\sqrt{\vec{\boldsymbol{\tau}}^T \vec{\boldsymbol{\tau}}}$, $\|\boldsymbol{\zeta}\|_{[L^2(\Omega)]^2} \simeq \sqrt{\vec{\boldsymbol{\zeta}}^T \vec{\boldsymbol{\zeta}}}$ (by the chosen scalings $O(1)$ and $O(h^{-1})$ in $M_{1,h}$ and $X^{\boldsymbol{\theta}}_{1,h}$, respectively).

On the other hand, we use the continuity of $\mathbf{K} : \ L^2(\Gamma) \to L^2(\Gamma)$ and obtain for any $\boldsymbol{\tau} \in M_{1,h} \setminus \{0\}$ and $\mathbf{s} = (\boldsymbol{\zeta}, \psi) \in X_{1,h} \setminus (0,0)$ the estimate

$$
\begin{aligned}
\frac{\vec{\boldsymbol{\tau}}^T B_1 \vec{\mathbf{s}}}{\sqrt{\vec{\boldsymbol{\tau}}^T \vec{\boldsymbol{\tau}}} \sqrt{\vec{\mathbf{s}}^T \vec{\mathbf{s}}}} &= \frac{-\int_\Omega \boldsymbol{\zeta} \cdot \boldsymbol{\tau}\, dx + \langle (\frac{1}{2}\mathbf{I} + \mathbf{K})\psi, \boldsymbol{\tau} \cdot \mathbf{n} \rangle_\Gamma}{\sqrt{\vec{\boldsymbol{\tau}}^T \vec{\boldsymbol{\tau}}} \sqrt{\vec{\boldsymbol{\zeta}}^T \vec{\boldsymbol{\zeta}} + \vec{\psi}^T \vec{\psi}}} \\
&\leq c \frac{\|\boldsymbol{\zeta}\|_{[L^2(\Omega)]^2} \|\boldsymbol{\tau}\|_{[L^2(\Omega)]^2} + \|\psi\|_{L^2(\Gamma)} \|\boldsymbol{\tau} \cdot \mathbf{n}\|_{L^2(\Gamma)}}{\sqrt{\vec{\boldsymbol{\tau}}^T \vec{\boldsymbol{\tau}}} \sqrt{\vec{\boldsymbol{\zeta}}^T \vec{\boldsymbol{\zeta}} + \vec{\psi}^T \vec{\psi}}} \\
&\leq c \left(\frac{\|\boldsymbol{\zeta}\|_{[L^2(\Omega)]^2} \|\boldsymbol{\tau}\|_{[L^2(\Omega)]^2}}{\sqrt{\vec{\boldsymbol{\tau}}^T \vec{\boldsymbol{\tau}}} \sqrt{\vec{\boldsymbol{\zeta}}^T \vec{\boldsymbol{\zeta}}}} + \frac{\|\psi\|_{L^2(\Gamma)} \|\boldsymbol{\tau} \cdot \mathbf{n}\|_{L^2(\Gamma)}}{\sqrt{\vec{\boldsymbol{\tau}}^T \vec{\boldsymbol{\tau}}} \sqrt{\vec{\psi}^T \vec{\psi}}} \right) = O(h)
\end{aligned}
$$

due to the scalings $O(h^{-1})$, $O(1)$ and $O(1)$ for the bases in $X^{\boldsymbol{\theta}}_{1,h}$, $M_{1,h}$ and $X^{\phi}_{1,h}$, respectively. Here, c is a positive generic constant being independent of h. This proves both $\sigma_m \leq O(h)$ and $\sigma_M \leq O(h)$ such that, with $\sigma_m \geq O(h)$ from before, we obtain (56).

Finally, we consider the singular values of B. Using Lemma 4.1 and the inf-sup condition (29) we obtain $O(h) \leq \eta_1 \leq \eta_N \leq O(1)$.

To prove the sharpness of $\eta_N = O(1)$ let v^* be -1 on one element, e.g. on the first one T_1, and zero elsewhere. Then the corresponding vector is $\vec{v}^* = (-h\,O(1), 0, \ldots, 0)^T$ by the scaling $O(h^{-1})$. We obtain the estimate

$$\begin{aligned}
\eta_N &= \max_{v\in M_h\setminus\{0\}} \max_{\boldsymbol{\tau}\in M_{1,h}\setminus\{0\}} \frac{-\int_\Omega v \operatorname{div} \boldsymbol{\tau}\, dx}{\sqrt{\vec{v}^T\vec{v}}\sqrt{\vec{\boldsymbol{\tau}}^T\vec{\boldsymbol{\tau}}}} \geq \max_{\boldsymbol{\tau}\in M_{1,h}\setminus\{0\}} \frac{-\int_\Omega v^* \operatorname{div} \boldsymbol{\tau}\, dx}{\sqrt{(\vec{v}^*)^T\vec{v}^*}\sqrt{\vec{\boldsymbol{\tau}}^T\vec{\boldsymbol{\tau}}}} \\
&= O(h^{-1}) \max_{\boldsymbol{\tau}\in M_{1,h}\setminus\{0\}} \frac{\int_{T_1} \operatorname{div} \boldsymbol{\tau}\, dx}{\sqrt{\vec{\boldsymbol{\tau}}^T\vec{\boldsymbol{\tau}}}} = O(1).
\end{aligned}$$

For the latter estimate note that the integral over T_1 of a constant is $O(h^2)$, that $\|\boldsymbol{\tau}\|_{L^\infty(T_1)}$ is proportional to the corresponding coefficients of $\boldsymbol{\tau}|_{T_1}$ (due to the scaling $O(1)$ within $M_{1,h}$) and that the divergence gives another $O(h^{-1})$. This proves (57) and finishes the proof of the lemma. □

Now we are able to give an estimate of the spectrum of the linearized stiffness matrix $\mathcal{A}$ when no preconditioner is used.

Theorem 4.2 *Assume that the smallest eigenvalue λ_1 of A_1 and the largest eigenvalue ρ_M of S do not coincide for sufficiently small mesh size h. Then the eigenvalues of the linearized stiffness matrix $\mathcal{A}$ are covered by the set*

$$[-a, -bh^2] \cup [ch, d]$$

where the positive constants a, b, c, and d do not depend on h. The number of iterations of the minimum residual method which are required to solve the linearized system (33) *up to a given accuracy is bounded by $O(h^{-3/2})$.*

Proof. Abstract bounds for the eigenvalues of the linearized stiffness matrix $\mathcal{A}$ are given by Lemma 4.2. Lemma 4.3 provides estimates for the individual parameters (bounds for singular values and eigenvalues).

In order to estimate μ_{-M} we have to study the smallest root of the function in μ

$$\mu^3 - \mu^2(\lambda_1 - \rho_M) - \mu(\sigma_M^2 + \lambda_1\rho_M + \eta_N^2) + \lambda_1\eta_N^2.$$

Since the coefficients of this polynomial are confined to compact intervals the smallest root is bounded from below.

For the eigenvalue μ_{-1} Lemma 4.2 states the bound

$$\mu_{-1} \leq \max\left\{\frac{\lambda_L}{2} - \sqrt{\frac{\lambda_L^2}{4} + \sigma_m^2}, -\eta_1\right\}.$$

By the estimates of λ_L, σ_m and η_1 given above we obtain

$$\mu_{-1} \leq \max\left\{c - \sqrt{c^2 + O(h^2)}, -O(h)\right\} = -O(h^2).$$

To bound μ_1 from below we have to consider the expression μ_1^* defined in Lemma 4.2. In the (theoretical) case $\lambda_1 = \rho_M$,

$$\mu_1^* = \frac{\rho_M \eta_1^2}{\sigma_M^2 + \rho_M^2 + \eta_1^2} \geq c\rho_M \eta_1^2 = O(h^3).$$

But this case can be avoided by scaling the basis of $X_{1,h}$ or $M_{1,h}$ by a constant, which is assumed in this theorem. Now, with the condition that no complex numbers appear,

$$\begin{aligned}\mu_1^* &= \frac{\sigma_M^2 + \lambda_1\rho_M + \eta_1^2}{2(\rho_M - \lambda_1)} - \sqrt{\left(\frac{\sigma_M^2 + \lambda_1\rho_M + \eta_1^2}{2(\rho_M - \lambda_1)}\right)^2 - \frac{\lambda_1\eta_1^2}{\rho_M - \lambda_1}} \\ &= \frac{\sigma_M^2 + \lambda_1\rho_M + \eta_1^2}{2(\rho_M - \lambda_1)}\left(1 - \sqrt{1 - \frac{4(\rho_M - \lambda_1)\lambda_1\eta_1^2}{(\sigma_M^2 + \lambda_1\rho_M + \eta_1^2)^2}}\right) \geq O(h)\end{aligned}$$

since the quotient in the radical behaves like $O(1)$. Analogously, we obtain $\mu_1^* = O(h)$ in the case $\lambda_1 > \rho_M$. Therefore, in view of Lemma 4.2, $\mu_1 \geq O(h)$.

Finally, the largest eigenvalue of the linearized stiffness matrix $\mathcal{A}$ can be estimated with Lemma 4.2 by

$$\mu_{L+N} \leq \frac{\lambda_L}{2} + \sqrt{\frac{\lambda_L^2}{4} + \sigma_M^2 + \eta_N^2} = O(1)$$

which finishes the proof for the spectrum. The number of MINRES iterations which are required to solve the linearized system (33) up to a given accuracy is bounded by $O(h^{-3/2})$ due to Theorem 4.1. □

From the previous theorem we expect an increase of the number of MINRES iterations for the solution of the linearized system if the mesh size is reduced to improve the Galerkin approximation. From the proof of this theorem it is also clear that the ill-conditioned behaviour has at least two reasons. First, the matrix A_1 is ill-conditioned ($\lambda_L/\lambda_1 = O(h^{-1})$) which comes from the hypersingular operator. Second, the matrix S is ill-conditioned which is due to the weakly singular operator. However, also the ratio c_1/c_0 of the numbers describing the scaling (35) within $M_{1,h}$ deteriorates, and these numbers influence the bounds for the singular values of B_1 and B, see Lemma 4.1. The ill-conditioned behaviour of S and the non-scalability within $M_{1,h}$ can be dealt with by the same preconditioner, as can be seen below.

Let us assume that we have a spectrally equivalent preconditioner P_1^ϕ for the discretization of the hypersingular operator and that we have a preconditioner P_2 that is equivalent to the $H(\mathrm{div};\Omega)$-inner product in $M_{1,h}$. For P_2 we can simply take the discrete form of the $H(\mathrm{div};\Omega)$-inner product and for the hypersingular operator there are different possibilities to find good preconditioners, e.g. multigrid methods [22, 23], two and multilevel additive Schwarz methods [19], and wavelet-based techniques [3, 4, 20, 21]. Using these preconditioners we obtain the following result.

Theorem 4.3 *Let*

$$\mathcal{P} := \begin{pmatrix} I & 0 & 0 & 0 \\ 0 & P_1^{\phi} & 0 & 0 \\ 0 & 0 & P_2 & 0 \\ 0 & 0 & 0 & I \end{pmatrix}$$

be a preconditioner for the linearized stiffness matrix $\mathcal{A}$, yielding the preconditioned symmetric matrix $\tilde{\mathcal{A}}$, where $P_1^{\phi} \in \mathbf{R}^{l\times l}$ with $l = \dim X_{1,h}^{\phi}$, $P_2 \in \mathbf{R}^{M\times M}$ and the remaining blocks $I =$ $\mathrm{diag}\,(1,\ldots,1)$ are of the correct size. Further, assume that P_1^{ϕ} is uniformly spectrally equivalent to the block of $\mathcal{A}$ which belongs to the hypersingular operator $\tilde{\mathbf{W}} = \mathbf{W} + P'P$, and that P_2 is uniformly spectrally equivalent to the $H(\mathrm{div};\Omega)$-inner product in $M_{1,h}$. We choose the basis functions in $X_{1,h}^{\boldsymbol{\theta}}$ and M_h to be scaled to $O(h^{-1})$. Then, the spectrum of $\tilde{\mathcal{A}}$ is covered by the set

$$[-a,-b] \cup [c,d]$$

where the positive constants a, b, c, and d do not depend on h. Moreover, in that case, the number of iterations of the preconditioned minimum residual method for the solution of (33) *is bounded.*

Proof. The proof of Theorem 4.2 with some modifications holds in this case as well. We only have to mention that, due to the preconditioner for the block belonging to the hypersingular operator, there holds $\lambda_1 = O(1)$ and $\lambda_L = O(1)$ and, by the choice of the preconditioner P_2, we now have for generic positive constants c and C

$$c\vec{\boldsymbol{\tau}}^T P_2 \vec{\boldsymbol{\tau}} \le \|\boldsymbol{\tau}\|_{M_1}^2 \le C\vec{\boldsymbol{\tau}}^T P_2 \vec{\boldsymbol{\tau}} \quad \text{for any } \boldsymbol{\tau} \in M_{1,h}.$$

Therefore, the numbers c_0 and c_1 in (35) both are constant. The boundedness of the numbers C_0 and C_1 in (36) has already been checked in the proof of Theorem 4.2. Therefore, Lemma 4.1 proves the boundedness of the singular values σ_m, σ_M of B_1 and η_1, η_N of B. It remains to prove that the largest eigenvalue ρ_M of $\tilde{S}$ is bounded. This can be seen as follows.

$$\begin{aligned} \rho_M &= \max_{\boldsymbol{\tau}\in M_{1,h}\setminus\{0\}} \frac{[P_2^{-1/2}\mathbf{S}P_2^{-1/2}\boldsymbol{\tau},\boldsymbol{\tau}]}{\vec{\boldsymbol{\tau}}^T\vec{\boldsymbol{\tau}}} \\ &= \max_{\boldsymbol{\tau}\in M_{1,h}\setminus\{0\}} \frac{\langle \mathbf{V}(\boldsymbol{\tau}\cdot\mathbf{n}),\boldsymbol{\tau}\cdot\mathbf{n}\rangle_\Gamma}{[P_2\boldsymbol{\tau},\boldsymbol{\tau}]} \le c \max_{\boldsymbol{\tau}\in M_{1,h}\setminus\{0\}} \frac{\|\boldsymbol{\tau}\cdot\mathbf{n}\|_{H^{-1/2}(\Gamma)}^2}{\|\boldsymbol{\tau}\|_{H(\mathrm{div};\Omega)}^2}. \end{aligned}$$

Here we used the continuity of $\mathbf{V}: \ H^{-1/2}(\Gamma) \to H^{1/2}(\Gamma)$. Now by the trace theorem

$$\|\boldsymbol{\tau}\cdot\mathbf{n}\|_{H^{-1/2}(\Gamma)} \le c\|\boldsymbol{\tau}\|_{H(\mathrm{div};\Omega)} \quad \text{for any } \boldsymbol{\tau} \in H(\mathrm{div};\Omega)$$

and this proves that ρ_M is bounded by a constant that is independent of h. Therefore, all singular values and eigenvalues that appear in Lemma 4.2 remain bounded and the boundedness of the spectrum of $\mathcal{A}$ is proved by that lemma. The boundedness of the number of MINRES iterations follow from Theorem 4.1. $\square$

5 Numerical results

For our numerical example we choose $\Omega = (0,1) \times (0,1)$ and take the functions f, g_1 and g_2 in (3), (5) such that $u(x_1, x_2) = x_1^2 + x_2^2$ and $u_c(x_1, x_2) = \log |(x_1, x_2) - (1/2, 1/2)|$. For the non-linearity $\boldsymbol{\kappa}$ we choose $\boldsymbol{\kappa}(x,s) = \boldsymbol{\kappa}(s) = 2 + \frac{1}{1+s}$. Of course, then the regularity assumptions of Theorem 3.1 are satisfied. Considering a sequence of uniform triangular meshes we expect an approximation error which decays like $O(h)$. In Figure 1 the approximation errors for the individual unknowns $\boldsymbol{\theta}$, $\boldsymbol{\sigma}$, u are given. For completeness we also plot the norm $\|\phi_h - u_c|_\Gamma\|_{L^2(\Gamma)}$ which converges to a constant different from zero since ϕ_h is fixed by $P\phi_h = \int_\Gamma \phi_h \, ds = 0$ and $P(u_c|_\Gamma) \neq 0$ for our example. All other three approximations individually converge to the exact solutions like $O(h)$, and this confirms the theoretical error bound given by Theorem 3.1. For aspects concerning implementation we refer to [7] for the finite element part and to [5] for the integral operators.

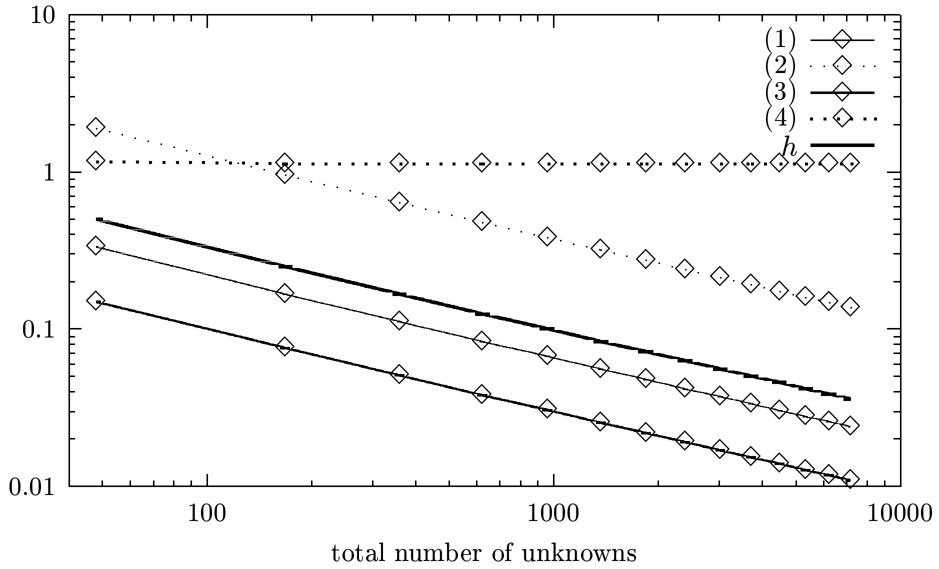

Figure 1: Approximation errors for the individual unknowns (1: $\|\boldsymbol{\theta} - \boldsymbol{\theta}_h\|_{[L^2(\Omega)]^2}$, 2: $\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{H(\mathrm{div}\,;\Omega)}$, 3: $\|u - u_h\|_{L^2(\Omega)}$, 4: $\|u_c - \phi_h\|_{L^2(\Gamma)}$).

Now we check the theoretical results of Section 4. We solve the non-linear system (28) by the Newton method with initial guess 0. The Newton iteration is stopped when the initial residual is reduced by a factor 10^{-7}. In each step of the Newton method we have to solve a linear system of the form (33) with matrix $\mathcal{A}$ or with preconditioned matrix $\tilde{\mathcal{A}}$. For the un-preconditioned MINRES we scale the basis functions in $X_{1,h}^{\boldsymbol{\theta}}$, $X_{1,h}^{\phi}$, $M_{1,h}$ and M_h to $O(h^{-1})$, $O(1)$, $O(1)$ and $O(h^{-1})$, respectively. Due to Theorem 4.2 we then expect that the eigenvalues of the stiffness

$M+L+N$	h	μ_{-M}	μ_{-1}	μ_1	μ_{L+N}
48	0.500	-2.674580	-0.054255	0.357457	2.667172
168	0.250	-2.773414	-0.013370	0.191241	2.772916
360	0.167	-2.800794	-0.005527	0.127842	2.800773
624	0.125	-2.811895	-0.002978	0.095660	2.811932
960	0.100	-2.817446	-0.001855	0.076324	2.817486
1368	0.083	-2.820610	-0.001265	0.063455	2.820644
1848	0.071	-2.822581	-0.000917	0.054283	2.822609
2400	0.063	-2.823891	-0.000695	0.047420	2.823914

Table 1: The extreme eigenvalues of the un-preconditioned matrix $\mathcal{A}$ in the first Newton step.

matrix $\mathcal{A}$ are contained in the set

$$[-a, -bh^2] \cup [ch, d]$$

where a, b, c, and d are positive numbers being independent of h. Table 1 presents the eigenvalues μ_{-M}, μ_{-1}, μ_1 and μ_{L+N} for our numerical example (taken from the first Newton step). Obviously, μ_{-M} and μ_{L+N} are bounded and μ_{-1} and μ_1 tend to zero if $h \to 0$. Indeed, $-\mu_{-1} = O(h^2)$ and $\mu_1 = O(h)$ as it can be seen from Figure 2 where a double logarithmic scale is used. Also from Theorem 4.2 we expect that the number of iterations of the MINRES is bounded by $O(h^{-3/2})$ in this case. In Figure 3 the maximum number of MINRES iterations over all Newton steps is given and the corresponding curve underlines the bound $O(h^{-3/2})$ without preconditioner. Here we stopped the MINRES iteration when the discrete l^2-norm of the initial residual was reduced by the factor 10^{-4}. It appears that the actual numbers of iterations are very large for this example. Therefore, one is forced to use a preconditioner.

For our example we simply take the $H(\mathrm{div}\,;\Omega)$-inner product and the matrix defined by the operator $\tilde{\mathbf{W}}$ itself to construct the preconditioners P_2 and P_1^{ϕ} in Theorem 4.3, respectively (and retain the scalings for $X_{1,h}^{\boldsymbol{\theta}}$ and M_h from above). Of course, this preconditioner is only of theoretical interest to confirm the results provided by Theorem 4.3. In practice other methods, e.g., as mentioned before that theorem are necessary. We expect a bounded spectrum for the preconditioned stiffness matrix $\tilde{\mathcal{A}}$ and also bounded iteration numbers of the preconditioned MINRES. Indeed, Table 2 confirms the boundedness of the extreme eigenvalues μ_{-M}, μ_{-1}, μ_1 and μ_{L+N}. Moreover, as presented by Figure 3, the iteration numbers are bounded when using the preconditioner and are very large when only using scaled basis functions. Finally, as it is also shown by Figure 3, for all mesh sizes h five Newton iterations are used to solve the non-linear system up to the relative accuracy 10^{-7}.

References

[1] M.A. Barrientos, G.N. Gatica and E.P. Stephan, *A mixed finite element method for nonlinear elasticity: two-fold saddle point approach and a-posteriori error estimate.* Tech-

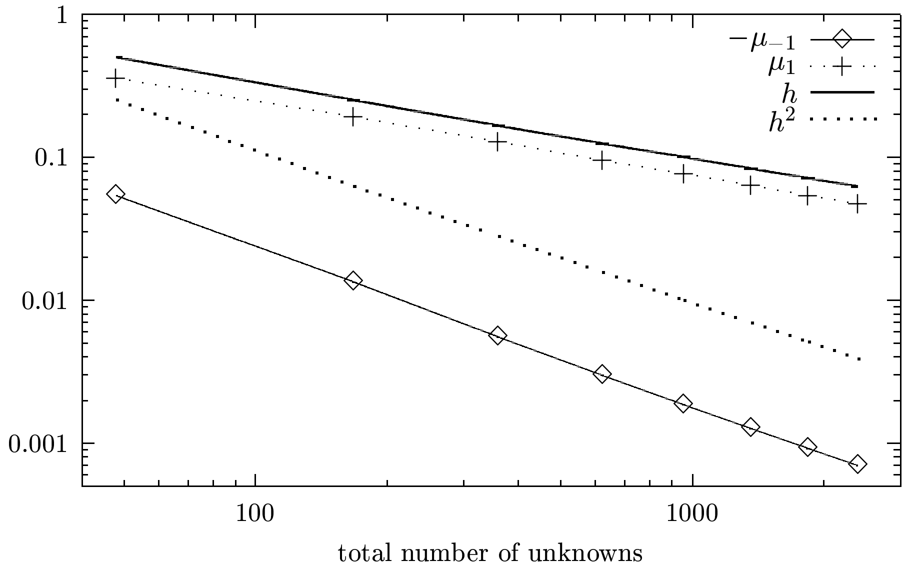

Figure 2: Asymptotic behavior of the extreme eigenvalues μ_{-1} and μ_1 of the un-preconditioned matrix $\mathcal{A}$ (first Newton step).

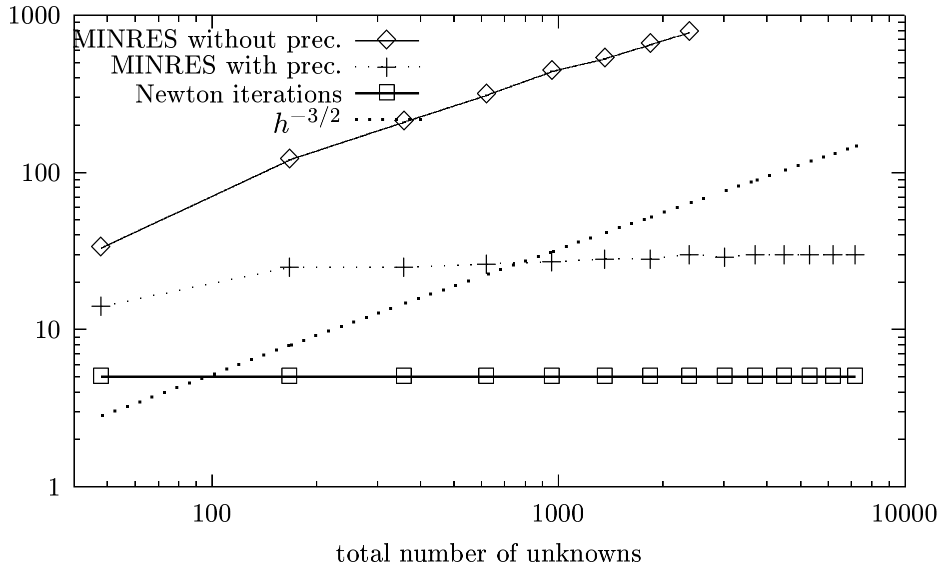

Figure 3: Number of MINRES iterations (maximum over all Newton steps) for reducing the initial residual by 10^{-4} and number of Newton iterations for reducing the initial residual by 10^{-7}.

$M+L+N$	h	μ_{-M}	μ_{-1}	μ_1	μ_{L+N}
48	0.500	-1.106929	-0.280776	0.641297	1.850382
168	0.250	-1.281106	-0.280776	0.638209	1.904767
360	0.167	-1.362907	-0.280775	0.637593	1.928042
624	0.125	-1.403690	-0.280773	0.637376	1.942155
960	0.100	-1.429430	-0.280771	0.637276	1.951966
1368	0.083	-1.447728	-0.280769	0.637221	1.959339
1848	0.071	-1.461666	-0.280765	0.637188	1.965163
2400	0.063	-1.472776	-0.280762	0.637167	1.969930

Table 2: The extreme eigenvalues of the preconditioned matrix $\tilde{\mathcal{A}}$ in the first Newton step.

nical Report 99-25, Departamento de Ingeniería Matemática, Universidad de Concepción, Chile, (1999).

[2] R. Chandra, S.C. Eisenstat, and M.H. Schultz, *The modified conjugate residual method for partial differential equations.* In Advances in Computer Methods for Partial Differential Equations II, R. Vichnevetsky, ed., IMACS, New Brunswick, 1977, pp. 13-19.

[3] W. Dahmen, S. Prössdorf, and R. Schneider, *Wavelet approximation methods for pseudodifferential equations II: Matrix compression and fast solution.* Advances in Computational Mathematics, 1 (1993), pp. 259-335.

[4] W. Dahmen, S. Prössdorf, and R. Schneider, *Wavelet approximation methods for pseudodifferential equations I: Stability and convergence.* Mathematische Zeitschrift, 215 (1994), pp. 583-620.

[5] V.J. Ervin, N. Heuer, and E.P. Stephan, *On the h-p version of the boundary element method for Symm's integral equation on polygons.* Computer Methods in Applied Mechanics and Engineering, 110 (1993), pp. 25–38.

[6] G.N. Gatica and N. Heuer, *Minimum residual iteration for a dual-dual mixed formulation of exterior transmission problems.* Numerical Linear Algebra and Applications, 8 (2001), pp. 147-164.

[7] G.N. Gatica and N. Heuer, *An expanded mixed finite element approach via a dual-dual formulation and the minimum residual method.* Journal of Computational and Applied Mathematics. To appear.

[8] G.N. Gatica and N. Heuer, *A dual-dual formulation for the coupling of mixed-FEM and BEM in hyperelasticity.* SIAM Journal on Numerical Analysis, 38 (2000), pp. 380-400.

[9] G.N. Gatica, N. Heuer and S. Meddahi, *On the numerical analysis of nonlinear two-fold saddle point problems.* Technical Report 00-03, Departamento de Ingeniería Matemática, Universidad de Concepción, Chile, (2000).

[10] G.N. Gatica and S. Meddahi, *A dual-dual mixed formulation for nonlinear exterior transmission problems.* Mathematics of Computation, posted on May 23, 2000, PII S 0025-5718(00)01267-9 (to appear in print).

[11] G.N. Gatica and S. Meddahi, *A fully discrete Galerkin scheme for a two-fold saddle point formulation of an exterior nonlinear problem.* Technical Report 00-05, Departamento de Ingeniería Matemática, Universidad de Concepción, Chile, (2000).

[12] G.N. Gatica and W.L. Wendland, *Coupling of mixed finite elements and boundary elements for linear and nonlinear elliptic problems.* Applicable Analysis, 63 (1996), pp. 39-75.

[13] M. Hahne, M. Maischak, E.P. Stephan, and A.J. Wathen, *Efficient preconditioners for coupled FEM–BEM equations.* Numerical Methods for Partial Differential Equations. To appear.

[14] N. Heuer, M. Maischak, and E.P. Stephan, *Preconditioned minimum residual iteration for the h-p version of the coupled FEM/BEM with quasi-uniform meshes.* Numerical Linear Algebra with Applications, 6 (1999), pp. 435–456.

[15] G.C. Hsiao, E.P. Stephan, and W.L. Wendland, *On the Dirichlet problem in elasticity for a domain exterior to an arc.* Journal of Computational and Applied Mathematics, 34 (1991), pp. 1-19.

[16] V.I. Lebedev, *Iterative methods for solving operator equations with a spectrum contained in several intervals.* USSR Comput. Math. and Math. Phys., 9 (1969), pp. 17-24.

[17] C.C. Paige and M.A. Saunders, *Solution of sparse indefinite systems of linear equations.* SIAM Journal on Numerical Analysis, 12 (1975), pp. 617-629.

[18] D.J. Silvester and A.J. Wathen, *Fast iterative solution of stabilised Stokes systems, Part II: Using general block preconditioners.* SIAM Journal on Numerical Analysis, 31 (1994), pp. 1352-1367.

[19] T. Tran and E.P. Stephan, *Additive Schwarz method for the h-version boundary element method.* Applicable Analysis, 60 (1996), pp. 63-84.

[20] T. Tran, E.P. Stephan, and S. Zaprianov, *Wavelet-based preconditioners for boundary integral equations.* Advances in Computational Mathematics, 9 (1998), pp. 233-249.

[21] T. von Petersdorff and C. Schwab, *Wavelet approximations for first kind boundary integral equations on polygons.* Numerische Mathematik, 74 (1996), pp. 479-516.

[22] T. von Petersdorff and E.P. Stephan, *On the convergence of the multigrid method for a hyper-singular integral equation of the first kind*. Numerische Mathematik, 57 (1990), pp. 379-391.

[23] T. von Petersdorff and E.P. Stephan, *Multigrid solvers and preconditioners for first kind integral equations*. Numerical Methods for Partial Differential Equations, 8 (1992), pp. 443-450.

[24] A.J. Wathen, B. Fischer, and D.J. Silvester, *The convergence rate of the minimum residual method for the Stokes problem*. Numerische Mathematik, 71 (1995), pp. 121-134.

[25] A.J. Wathen and D.J. Silvester, *Fast iterative solution of stabilised Stokes systems, Part I: Using simple diagonal preconditioners*. SIAM Journal on Numerical Analysis, 30 (1993), pp. 630-649.