ESAIM: M2AN 46 (2012) 39–57
DOI: 10.1051/m2an/2011017

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

A NUMERICAL MINIMIZATION SCHEME FOR THE COMPLEX HELMHOLTZ EQUATION

Russell B. Richins[1] and David C. Dobson[2]

Abstract. We use the work of Milton, Seppecher, and Bouchitté on variational principles for waves in lossy media to formulate a finite element method for solving the complex Helmholtz equation that is based entirely on minimization. In particular, this method results in a finite element matrix that is symmetric positive-definite and therefore simple iterative descent methods and preconditioning can be used to solve the resulting system of equations. We also derive an error bound for the method and illustrate the method with numerical experiments.

Mathematics Subject Classification. 65N30, 35A15.

Received July 28, 2010. Revised April 20, 2011.
Published online July 22, 2011.

1. Introduction

Many systems that result in steady-state oscillations can be modeled with the Helmholtz equation, but of particular interest are acoustic waves and transverse electric or transverse magnetic electromagnetic waves in inhomogeneous media. In each of these situations, the equation of interest can be expressed as

$$-\nabla\cdot\rho^{-1}\nabla P-\frac{\omega^2}{\kappa}P=0 \tag{1.1}$$

for appropriate choices of the complex-valued, spatially dependent material parameters ρ and κ, where $\omega>0$ is the frequency. The classical methods of deriving a weak form for this equation (with Dirichlet boundary conditions, for example) result in the variational equation

$$\int_\Gamma\left[\rho^{-1}\nabla P\cdot\nabla\overline{u}-\frac{\omega^2}{\kappa}P\overline{u}\right]\mathrm{d}x=0,\quad\forall u\in H_0^1(\Gamma), \tag{1.2}$$

which corresponds to a stationary principle, but not a minimization principle. In [9], Milton, Seppecher, and Bouchitté expand upon the work of Cherkaev and Gibiansky [3] for the conductivity equation to derive variational principles for (1.1) (as a special case of the more general equations of elasticity and electromagnetism) that are true minimization principles, provided the media are lossy. The minimization functional corresponds

Keywords and phrases. Variational methods, Helmholtz equation, finite element methods.

[1] Department of Mathematics, Michigan State University, East Lansing, 48824 Michigan, USA. richins@math.msu.edu
[2] Department of Mathematics, University of Utah, Salt Lake City, 84112 Utah, USA. dobson@math.utah.edu

Article published by EDP Sciences

© EDP Sciences, SMAI 2011

https://doi.org/10.1051/m2an/2011017 Published online by Cambridge University Press

physically to dissipated energy in the system, and is valid even for arbitrarily small coefficients of loss. While the framework presented in [9] results in nonstandard boundary conditions, Milton and Willis extend the principles to handle the classical Dirichlet and Neumann boundary conditions in [8].

In this paper we apply the finite element method to these minimization principles and thereby develop a numerical algorithm for solving (1.1) that can take advantage of the many efficient methods available for solving a symmetric, positive-definite system of linear equations. The outline of the paper is as follows. Sections 2 and 3 review the general variational formulation and boundary conditions introduced by Milton, Seppecher, Bouchitté, and Willis. In Section 4, we derive an error bound on certain finite element discretizations of the variational principle. In Sections 5, we describe a straightforward implementation of the finite element method on a square domain, with Dirichlet boundary conditions. In Section 6, we suggest a preconditioner for solving the resulting symmetric positive definite linear system *via* the preconditioned conjugate gradient method, and find conditions on the material coefficients under which we expect the best conditioning. Section 7 describes the results of some simple numerical experiments, and illustrates numerical convergence consistent with the error bounds from Section 4. Finally, in Section 8, we extend the method to handle Robin boundary conditions, and present some associated numerical examples.

2. Variational formulation

Our model problem is

$$\begin{cases} -\nabla \cdot \rho^{-1} \nabla P - \dfrac{\omega^2}{\kappa} P = 0 & \text{in } \Gamma \\ P = f & \text{on } \partial\Gamma \end{cases} \tag{2.1}$$

where Γ is an open, bounded subset of $\mathbb{R}^d$ ($d = 2$ or 3) with smooth boundary. For acoustic waves, ρ is the density, κ is the bulk modulus, ω is the frequency, and P is the pressure. Here ρ, κ, and P are all complex. In this section, we focus on Dirichlet boundary conditions for simplicity; Neumann conditions can be handled similarly. In [9], a general framework for generating variational principles is presented; here we will focus on the specific case of the model problem above.

First we define a dual variable v, called the complex velocity, by

$$\mathrm{i}\omega v = \rho^{-1} \nabla P,$$

and then we define two complex-valued fields and a matrix by

$$\mathcal{F} = \begin{pmatrix} \nabla P \\ P \end{pmatrix}, \ \mathcal{G} = \begin{pmatrix} -\mathrm{i}\omega v \\ -\mathrm{i}\omega \nabla \cdot v \end{pmatrix} \text{ and } Z = \begin{pmatrix} -\rho^{-1} & 0 \\ 0 & \frac{\omega^2}{\kappa} \end{pmatrix}.$$

Then we have

$$\begin{pmatrix} -\rho^{-1} & 0 \\ 0 & \frac{\omega^2}{\kappa} \end{pmatrix} \begin{pmatrix} \nabla P \\ P \end{pmatrix} = \begin{pmatrix} -\rho^{-1} \nabla P \\ \frac{\omega^2}{\kappa} P \end{pmatrix} = \begin{pmatrix} -\mathrm{i}\omega v \\ -\mathrm{i}\omega \nabla \cdot v \end{pmatrix},$$

which is equivalent to the constitutive relation

$$\mathcal{G} = Z\mathcal{F}.$$

We are still working with complex-valued quantities, and we need to move to real-valued quantities in order to talk about minimization. For this reason, we take the real and imaginary parts of the equation above to find (here $'$ denotes a real part and $''$ denotes an imaginary part)

$$\mathcal{G}' = Z'\mathcal{F}' - Z''\mathcal{F}'' \text{ and } \mathcal{G}'' = Z'\mathcal{F}'' + Z''\mathcal{F}',$$

https://doi.org/10.1051/m2an/2011017 Published online by Cambridge University Press

which in matrix form reads

$$\begin{pmatrix} \mathcal{G}'' \\ \mathcal{G}' \end{pmatrix} = \begin{pmatrix} Z'' & Z' \\ Z' & -Z'' \end{pmatrix} \begin{pmatrix} \mathcal{F}' \\ \mathcal{F}'' \end{pmatrix}.$$

The matrix in this constitutive relation is obviously not positive definite, but if $Z'' > \alpha I$, then we may use this relation to build saddle point variational principles, as detailed in [9].

In order to get a matrix that is positive definite, and therefore a minimization variational principle, we rearrange this equation, solving for the imaginary parts of $\mathcal{G}$ and $\mathcal{F}$. We find that

$$\mathcal{F}'' = (Z'')^{-1}Z'\mathcal{F}' - (Z'')^{-1}\mathcal{G}'$$

and

$$\mathcal{G}'' = Z'((Z'')^{-1}Z'\mathcal{F}' - (Z'')^{-1}G') + Z''\mathcal{F}' = (Z'' + Z'(Z'')^{-1}Z')\mathcal{F}' - Z'(Z'')^{-1}\mathcal{G}'.$$

In matrix form, the new constitutive relation is

$$\begin{pmatrix} \mathcal{G}'' \\ \mathcal{F}'' \end{pmatrix} = \mathcal{L} \begin{pmatrix} \mathcal{F}' \\ -\mathcal{G}' \end{pmatrix}, \tag{2.2}$$

where

$$\mathcal{L} = \begin{pmatrix} Z'' + Z'(Z'')^{-1}Z' & Z'(Z'')^{-1} \\ (Z'')^{-1}Z' & (Z'')^{-1} \end{pmatrix}.$$

The matrix $\mathcal{L}$ is positive definite as long as Z'' is. Indeed, given a vector $(\mathcal{F}', -\mathcal{G}')^T$, define $(\mathcal{G}'', \mathcal{F}'')^T$ by (2.2). Then

$$\begin{aligned} \begin{pmatrix} \mathcal{F}' \\ -\mathcal{G}' \end{pmatrix} \cdot \mathcal{L} \begin{pmatrix} \mathcal{F}' \\ -\mathcal{G}' \end{pmatrix} &= \begin{pmatrix} \mathcal{F}' \\ -\mathcal{G}' \end{pmatrix} \cdot \begin{pmatrix} \mathcal{G}'' \\ \mathcal{F}'' \end{pmatrix} = \mathcal{F}' \cdot \mathcal{G}'' - \mathcal{G}' \cdot \mathcal{F}'' \\ &= \mathcal{F}' \cdot (Z'\mathcal{F}'' + Z''\mathcal{F}') - (Z'\mathcal{F}' - Z''\mathcal{F}'') \cdot \mathcal{F}'' = \mathcal{F}' \cdot Z''\mathcal{F}' + \mathcal{F}'' \cdot Z''\mathcal{F}''. \end{aligned}$$

In the constitutive relation (2.2), it is convenient to separate the scalar variables from the vector variables. Let $r = -\rho^{-1}$ and $k = \kappa^{-1}$, and let

$$\mathcal{R} = \begin{pmatrix} r'' + r'(r'')^{-1}r' & r'(r'')^{-1} \\ (r'')^{-1}r' & (r'')^{-1} \end{pmatrix} \text{ and } \mathcal{K} = \begin{pmatrix} k'' + k'(k'')^{-1}k' & k'(k'')^{-1} \\ (k'')^{-1}k' & (k'')^{-1} \end{pmatrix}.$$

A calculation similar to the one done above shows that the matrices $\mathcal{R}$ and $\mathcal{K}$ are positive definite as long as there exist constants $\alpha > 0$ and $\beta < 0$ such that

$$\rho''(x) \geq \alpha I \text{ and } \kappa''(x) < \beta \ \forall \ x \in \Gamma. \tag{2.3}$$

In [10] it is explained that an imaginary ρ corresponds to inertial loss in the material, and imaginary κ corresponds to deformation losses. In [7] materials with complex effective density are constructed.

Assuming that condition (2.3) holds, we define the functional

$$\begin{aligned} Y(P', v'') &= \int_\Gamma \begin{pmatrix} \mathcal{F}' \\ -\mathcal{G}' \end{pmatrix} \cdot \mathcal{L} \begin{pmatrix} \mathcal{F}' \\ -\mathcal{G}' \end{pmatrix} \, \mathrm{d}x \\ &= \int_\Gamma \left[\begin{pmatrix} \nabla P' \\ -\omega v'' \end{pmatrix} \cdot \mathcal{R} \begin{pmatrix} \nabla P' \\ -\omega v'' \end{pmatrix} + \begin{pmatrix} \omega P' \\ -\nabla \cdot v'' \end{pmatrix} \cdot \mathcal{K} \begin{pmatrix} \omega P' \\ -\nabla \cdot v'' \end{pmatrix} \right] \mathrm{d}x. \end{aligned}$$

https://doi.org/10.1051/m2an/2011017 Published online by Cambridge University Press

Then if $P' \in H^1(\Gamma)$ and $v'' \in H(\mathrm{div}, \Gamma)$ are solutions to the Helmholtz equation, let $s \in H_0^1(\Gamma)$ and $T \in H_0(\mathrm{div}, \Gamma)$. We have

$$Y(P'+s, v''+T) = Y(P', v'') + 2\int_\Gamma \left[\begin{pmatrix} \nabla P' \\ -\omega v'' \end{pmatrix} \cdot \mathcal{R} \begin{pmatrix} \nabla s \\ -\omega T \end{pmatrix} + \begin{pmatrix} \omega P' \\ -\nabla \cdot v'' \end{pmatrix} \cdot \mathcal{K} \begin{pmatrix} \omega s \\ -\nabla \cdot T \end{pmatrix} \right] \mathrm{d}x + Y(s, T).$$

Since

$$\begin{aligned} &\int_\Gamma \left[\begin{pmatrix} \nabla P' \\ -\omega v'' \end{pmatrix} \cdot \mathcal{R} \begin{pmatrix} \nabla s \\ -\omega T \end{pmatrix} + \begin{pmatrix} \omega P' \\ -\nabla \cdot v'' \end{pmatrix} \cdot \mathcal{K} \begin{pmatrix} \omega s \\ -\nabla \cdot T \end{pmatrix} \right] \mathrm{d}x \\ &\quad = \int_\Gamma \left[\begin{pmatrix} \omega v' \\ \nabla P'' \end{pmatrix} \cdot \begin{pmatrix} \nabla s \\ -\omega T \end{pmatrix} + \begin{pmatrix} \nabla \cdot v' \\ \omega P'' \end{pmatrix} \cdot \begin{pmatrix} \omega s \\ -\nabla \cdot T \end{pmatrix} \right] \mathrm{d}x \\ &\quad = \int_\Gamma \left[\omega v' \cdot \nabla s - \omega \nabla P'' \cdot T + \omega \nabla \cdot v' s - \omega P'' \nabla \cdot T \right] \mathrm{d}x \\ &\quad = \int_\Gamma \left[\omega \nabla \cdot (s v') - \omega \nabla \cdot (P'' T) \right] \mathrm{d}x = \int_{\partial\Gamma} \left[\omega s v' \cdot n - \omega P'' T \cdot n \right] \mathrm{d}S = 0 \end{aligned}$$

because of the homogeneous boundary conditions satisfied by s and T, we have that

$$Y(P'+s, v''+T) = Y(P', v'') + Y(s, T) \geq Y(P', v''),$$

so (P', v'') minimizes Y over pairs $(s, T) \in H^1(\Gamma) \times H(\mathrm{div}, \Gamma)$ satisfying

$$s = P' \text{ and } T \cdot n = v'' \cdot n$$

on $\partial\Gamma$.

In [9] it is also shown that if P' and v'' minimize Y over all functions satisfying the same boundary conditions, then P' is the real part of the solution to the Helmholtz equation satisfying those boundary conditions, and v'' is the imaginary part of the corresponding complex velocity field.

3. Other boundary conditions

Instead of specifying boundary conditions on P' and v'', we can specify boundary conditions on their dual variables P'' and v' and solve the boundary value problem

$$\begin{cases} -\nabla \cdot \rho^{-1} \nabla P - \dfrac{\omega^2}{\kappa} P = 0 & \text{in } \Gamma \\ P'' = P_0'' & \text{on } \partial\Gamma \\ v' \cdot n = v_0' \cdot n & \text{on } \partial\Gamma. \end{cases} \tag{3.1}$$

https://doi.org/10.1051/m2an/2011017 Published online by Cambridge University Press

If (P'', v') are parts of a solution to the problem above, then for any choice of functions $s \in H^1(\Gamma)$ and $T \in H(\mathrm{div}, \Gamma)$, we have

$$\begin{aligned}
0 &= \int_\Gamma [-\omega(-\nabla \cdot v' + \nabla \cdot v')s - \omega(\nabla P'' - \nabla P'') \cdot T] \, \mathrm{d}x \\
&= \int_\Gamma [-\omega v' \cdot \nabla s - \omega \nabla \cdot v' s - \omega \nabla P'' \cdot T - \omega P'' \nabla \cdot T] \, \mathrm{d}x + \int_{\partial\Gamma} [\omega s v' \cdot n + \omega T \cdot n P''] \, \mathrm{d}S \\
&= \int_\Gamma \left[\begin{pmatrix} -\omega v' \\ \nabla P'' \end{pmatrix} \cdot \begin{pmatrix} \nabla s \\ -\omega T \end{pmatrix} + \begin{pmatrix} -\nabla \cdot v' \\ \omega P'' \end{pmatrix} \cdot \begin{pmatrix} \omega s \\ -\nabla \cdot T \end{pmatrix} \right] \mathrm{d}x \\
&\quad + \int_{\partial\Gamma} [\omega s v' \cdot n + \omega T \cdot n P''] \, \mathrm{d}S \\
&= \int_\Gamma \left[\begin{pmatrix} \nabla P' \\ -\omega v'' \end{pmatrix} \cdot \mathcal{R} \begin{pmatrix} \nabla s \\ -\omega T \end{pmatrix} + \begin{pmatrix} \omega P \\ -\nabla \cdot v'' \end{pmatrix} \cdot \mathcal{K} \begin{pmatrix} \omega s \\ -\nabla \cdot T \end{pmatrix} \right] \mathrm{d}x \\
&\quad + \omega \int_{\partial\Gamma} [s v' \cdot n + T \cdot n P''] \, \mathrm{d}S. \qquad (3.2)
\end{aligned}$$

This is just the weak form of the Euler-Lagrange equation for the variational problem

$$\inf_{(P', v'') \in H^1(\Gamma) \times H(\mathrm{div}, \Gamma)} \tilde{Y}(P', v''),$$

where

$$\begin{aligned}
\tilde{Y}(P', v'') = \int_\Gamma &\left[\begin{pmatrix} \nabla P' \\ -\omega v'' \end{pmatrix} \cdot \mathcal{R} \begin{pmatrix} \nabla P' \\ -\omega v'' \end{pmatrix} + \begin{pmatrix} \omega P \\ -\nabla \cdot v'' \end{pmatrix} \cdot \mathcal{K} \begin{pmatrix} \omega P' \\ -\nabla \cdot v'' \end{pmatrix} \right] \mathrm{d}x \\
&+ 2\omega \int_{\partial\Gamma} [P' v_0' \cdot n + v'' \cdot n P_0''] \, \mathrm{d}S. \qquad (3.3)
\end{aligned}$$

Neither of these combinations of boundary conditions is often encountered, but we can combine these boundary conditions to derive minimization functionals for the Dirichlet and Neumann boundary value problems. In order to solve the Dirichlet boundary value problem

$$\begin{cases} -\nabla \cdot \rho^{-1} \nabla P - \dfrac{\omega^2}{\kappa} P = 0 & \text{in } \Gamma \\ P = P_0 & \text{on } \partial\Gamma, \end{cases}$$

we solve the variational problem

$$\inf_{(P', v'') \in H_0^1(\Gamma) \times H(\mathrm{div}, \Gamma)} \tilde{Y}(P' + P_0', v'').$$

Since we are enforcing zero boundary values on the P', variable, we have that

$$\begin{aligned}
\tilde{Y}(P' + P_0', v'') = \int_\Gamma &\left[\begin{pmatrix} \nabla P_0' + \nabla P' \\ -\omega v'' \end{pmatrix} \cdot \mathcal{R} \begin{pmatrix} \nabla P_0' + \nabla P' \\ -\omega v'' \end{pmatrix} \right. \\
&\left. + \begin{pmatrix} \omega(P_0' + P) \\ -\nabla \cdot v'' \end{pmatrix} \cdot \mathcal{K} \begin{pmatrix} \omega(P_0' + P') \\ -\nabla \cdot v'' \end{pmatrix} \right] \mathrm{d}x + 2\omega \int_{\partial\Gamma} v'' \cdot n P_0'' \, \mathrm{d}S.
\end{aligned}$$

https://doi.org/10.1051/m2an/2011017 Published online by Cambridge University Press

The minimization functional for the Neumann problem is derived in an analogous manner. To solve the problem

$$\begin{cases} -\nabla \cdot \rho^{-1} \nabla P - \dfrac{\omega^2}{\kappa} P = 0 & \text{in } \Gamma \\ v \cdot n = v_0 \cdot n & \text{on } \partial\Gamma, \end{cases}$$

we solve the variational problem

$$\inf_{(P', v'') \in H^1(\Gamma) \times H_0(\mathrm{div}, \Gamma)} \tilde{Y}(P', v'' + v_0'').$$

In this case, since we are enforcing zero normal component on v'', we have that

$$\begin{aligned} \tilde{Y}(P', v'' + v_0'') = \int_\Gamma \Bigg[\begin{pmatrix} \nabla P' \\ -\omega(v'' + v_0'') \end{pmatrix} \cdot \mathcal{R} \begin{pmatrix} \nabla P' \\ -\omega(v'' + v_0'') \end{pmatrix} \\ + \begin{pmatrix} \omega P \\ -\nabla \cdot v'' - \nabla \cdot v_0'' \end{pmatrix} \cdot \mathcal{K} \begin{pmatrix} \omega P' \\ -\nabla \cdot v'' - \nabla \cdot v_0'' \end{pmatrix} \Bigg] \mathrm{d}x + 2\omega \int_{\partial\Gamma} P' v_0' \cdot n \, \mathrm{d}S. \end{aligned}$$

4. Error bound

In this section we give a bound on the error incurred by solving any of the minimization problems above over a finite dimensional subspace of the specified Sobolev spaces. We will give a more detailed account of exactly what the finite dimensional space looks like later on; in this section all that will matter is the highest degree of polynomials that the finite dimensional space contains. We will use the Bramble-Hilbert lemma to give a bound on the error.

In this chapter we will return to the notation of $\mathcal{F}$, $\mathcal{G}$, and $\mathcal{L}$ and drop the primes used to denote real and imaginary parts. Note that what follows applies to any of the boundary value problems discussed previously, since the bounds depend only on the corresponding bilinear form. Throughout this chapter, C is a constant independent of the solution (P, v) and the grid spacing h.

4.1. Bilinear form

Define the bilinear form B by

$$B(P, v; s, T) = \int_\Gamma \begin{pmatrix} \mathcal{F} \\ -\mathcal{G} \end{pmatrix} \cdot \mathcal{L} \begin{pmatrix} \mathcal{S} \\ -\mathcal{T} \end{pmatrix} \mathrm{d}x, \tag{4.1}$$

where as before, $\mathcal{F} = (\nabla P, P)^T$, $\mathcal{G} = (-\mathrm{i}\omega v, -\mathrm{i}\omega \nabla \cdot v)^T$, and $\mathcal{S}$ and $\mathcal{T}$ are generated from test function $s \in H^1(\Gamma)$ and $T \in H(\mathrm{div}, \Gamma)$ in the same fashion. Assume that there exist constants $\gamma_1, \gamma_2 > 0$ such that $\mathcal{L} > \gamma_2 I$ and that $[\mathcal{L}(x)]_{ij} \leq \gamma_1$ for a.e. $x \in \Gamma$. Let $V = H_0^1(\Gamma) \times H(\mathrm{div}, \Gamma)$, endowed with the norm

$$\|(u, G)\|_V = (\|u\|^2_{H^1(\Gamma)} + \|G\|^2_{H(\mathrm{div}, \Gamma)})^{\frac{1}{2}}.$$

Then it follows immediately from (4.1) that

$$B(P, v; s, T) \leq C\gamma_1 \|(P, v)\|_V \|(s, T)\|_V \tag{4.2}$$

and

$$B(P, v; P, v) \geq \gamma_2 \|(P, v)\|^2_V. \tag{4.3}$$

https://doi.org/10.1051/m2an/2011017 Published online by Cambridge University Press

4.2. **Minimization inequality**

Define the energy by

$$f(s,T) = \frac{1}{2}B(s,T;s,T) - F(s,T),$$

where $F : H^1(\Gamma)\times H(\mathrm{div},\Gamma) \to \mathbb{R}$ (in practice, F is usually composed of terms resulting from an inhomogeneous term and enforcement of the desired boundary conditions). If (u,G) is the minimizer of the energy, then this pair must satisfy the Euler-Lagrange equation

$$B(P,v;s,T) = F(s,T) \quad \forall s \in H_0^1(\Gamma),\ \forall\ T \in H(\mathrm{div},\Gamma),$$

so that

$$f(s,T) = f(P,v) + \frac{1}{2}B(P-s,v-T;P-s,v-T) \quad \forall\ s \in H_0^1(\Gamma)\ \forall\ T \in H(\mathrm{div},\Gamma).$$

Consider a finite dimensional subspace $V_N = V_{N1}\times V_{N2}$ of V, where V_{N1} is a finite dimensional subspace of $H^1(\Gamma)$ and V_{N2} is a finite dimensional subspace of $H(\mathrm{div},\Gamma)$. If (P_N, v_N) is such that

$$f(P_N,v_N) = \min_{(s,T)\in V_N} f(s,T),$$

then

$$[B(P-P_N,v-v_N;P-P_N,v-v_N)]^{\frac{1}{2}} = \min_{(s,T)\in V_N}[B(P-s,v-T;P-s,v-T)]^{\frac{1}{2}}.$$

Inequalities (4.2) and (4.3) imply that

$$\sqrt{\gamma_2}\|(s,T)\|_V \le \sqrt{B(s,T;s,T)} \le C\sqrt{\gamma_1}\|(s,T)\|_V \quad \forall\ (s,T)\in V,$$

so we have

$$\sqrt{\gamma_2}\|(P,v)-(P_N,v_N)\|_V \le \min_{(s,T)\in V_N} C\sqrt{\gamma_1}\|(P,v)-(s,T)\|_V. \tag{4.4}$$

Let F_1 be the orthogonal projection from $H^1(\Gamma)$ onto V_{N1}. Since F_1 is an orthogonal projection, it has $\|F_1\|_{B(H^1(\Gamma),H^1(\Gamma))} = 1$, where $B(H^1(\Gamma),H^1(\Gamma))$ is the set of bounded linear functions from $H^1(\Gamma)$ to $H^1(\Gamma)$. Also, define an operator $F_2 : H(\mathrm{div},\Gamma)\to V_{N2}$ by the solution of the variational inequality

$$\langle F_2G, Q-F_2G\rangle_{L^2(\Gamma,\mathbb{R}^d)} \ge \langle G, Q-F_2G\rangle_{L^2(\Gamma,\mathbb{R}^d)}\ \forall\ Q\in E_G,$$

over the set $E_G = \{v\in V_{N2} : \|\nabla\cdot v\|_{L^2(\Gamma)} \le \|\nabla\cdot G\|_{L^2(\Gamma)}\}$, which is a closed, convex subset of $L^2(\Gamma,\mathbb{R}^d)$. We then have

$$\|F_2G\|^2_{L^2(\Gamma,\mathbb{R}^d)} = \langle F_2G, F_2G\rangle_{L^2(\Gamma,\mathbb{R}^d)} \le \langle G, F_2G\rangle_{L^2(\Gamma,\mathbb{R}^d)} \le \|G\|_{L^2(\Gamma,\mathbb{R}^d)}\|F_2G\|_{L^2(\Gamma,\mathbb{R}^d)}.$$

If we take $s = F_1P$ and $T = F_2v$ in (4.4), then we have

$$\|(P,v)-(P_N,v_N)\|_V \le C\|(P-F_1P, v-F_2v)\|_V. \tag{4.5}$$

4.3. **Seminorm bounds**

We will discretize the domain Γ by by subdividing it into smaller regions, each of which can be seen as a suitable shifting and scaling of a "reference element". More precisely, if $\hat{e}$ is our reference element, there exist affine changes of variables $F_l(x) = Bx + x_l$ such that $F_l(\hat{e}) = e_l$, where e_l is the lth element (subdivision) in the finite element decomposition of Γ. In the case of rectangular elements in $\mathbb{R}^d$, for example, we can take $\hat{e} = (0,1)^d$, and then we have $B = hI_d$. In this section a hat will denote the corresponding function defined over the reference element.

https://doi.org/10.1051/m2an/2011017 Published online by Cambridge University Press

Let

$$[u,w]_s = \sum_{|\alpha|=s} \int_{\hat{e}} D^\alpha u \cdot D^\alpha w \, \mathrm{d}x \quad \text{and} \quad |w|_s^2 = [w,w]_s, \tag{4.6}$$

where for vector functions we define

$$D^\alpha w = \begin{pmatrix} D^\alpha w_1 \\ D^\alpha w_2 \\ \vdots \\ D^\alpha w_d \end{pmatrix}.$$

From [2] we get the inequalities

$$\begin{aligned} c^{-1}h^{s-\frac{d}{2}}|w|_{s,e_l} \le |\hat{w}|_s \le ch^{s-\frac{d}{2}}|w|_{s,e_l} \\ h^{s+\frac{d}{2}-1}|q|_{s,e_l} \le |\hat{q}|_s \le h^{s+\frac{d}{2}-1}|q|_{s,e_l} \\ h^{s+\frac{d}{2}}|\nabla\cdot q|_{s,e_l} \le |\nabla\cdot\hat{q}|_s \le h^{s+\frac{d}{2}}|\nabla\cdot q|_{s,e_l} \end{aligned} \tag{4.7}$$

for scalar functions w and vector functions q, where $w = \hat{w}\circ F^{-1}$ and $q = \hat{q}\circ F^{-1}$ and $|\cdot|_{s,e_l}$ denotes (4.6) with e_l in place of $\hat{e}$.

We now recall the following lemma from [1], which will be used in what follows.

Lemma 4.1 (Bramble-Hilbert lemma)**.** *For some region $\Omega \subset \mathbb{R}^2$ and some integer $k \ge -1$, let there be given a bounded linear functional*

$$f : H^{k+1}(\Omega) \to \mathbb{R},$$

satisfying $|f(u)| \le \delta\|u\|_{H^{k+1}(\Omega)}$ for all $u \in H^{k+1}(\Omega)$ for some δ independent of u. Suppose that $f(u) = 0$ for all $u \in P_k(\bar{\Omega})$. Then there exists a constant C, dependent only on Ω such that

$$|f(u)| \le C\delta|u|_{k+1}, \quad u \in H^{k+1}(\Omega).$$

Let us suppose that $\hat{P} \in H^{k+1}(\hat{e})$ and $\hat{v} \in H^j(\mathrm{div},\hat{e}) = \{q \in H^j(\hat{e},\mathbb{R}^d) : \nabla\cdot q \in H^j(\hat{e})\}$. For fixed elements $w \in H^s(\hat{e})$ and $Q \in H^s(\mathrm{div},\hat{e})$ define the functionals

$$f_1(u) = [u - F_1u, w]_s, \ f_2(G) = [G - F_2G, Q]_0, \ f_3(\nabla\cdot G) = [\nabla\cdot G - \nabla\cdot F_2G, \nabla\cdot Q]_0,$$

where $s = 0$ or $s = 1$. Then, since

$$\begin{aligned} |f_1(u)| &\le |u - F_1u|_s|w|_s \le (|u|_s + |F_1u|_s)|w|_s \le (\|u\|_{H^1(\Gamma)} + \|F_1u\|_{H^1(\Gamma)})|w|_s \\ &\le 2\|u\|_{H^1(\Gamma)}|w|_s \le 2\|u\|_{H^{k+1}(\Gamma)}|w|_s, \\ |f_2(G)| &\le |G - F_2G|_0|Q|_0 \le (|G|_0 + |F_2G|_0)|Q|_0 = (\|G\|_{L^2(\Gamma,\mathbb{R}^d)} + \|F_2G\|_{L^2(\Gamma,\mathbb{R}^d)})|Q|_0 \\ &\le 2\|G\|_{L^2(\Gamma),\mathbb{R}^d)}|Q|_0 \le \|G\|_{H^j(\Gamma,\mathbb{R}^d)}|Q|_0, \\ |f_3(\nabla\cdot\ G)| &\le |\nabla\cdot G - \nabla\cdot F_2G|_0|\nabla\cdot Q|_0 \le (|\nabla\cdot G|_0 + |\nabla\cdot F_2G|_0)|\nabla\cdot Q|_0 \\ &= (\|\nabla\cdot G\|_{L^2(\Gamma)} + \|\nabla\cdot F_2G\|_{L^2(\Gamma)})|\nabla\cdot Q|_0 \le 2\|\nabla\cdot G\|_{L^2(\Gamma)}|\nabla\cdot Q|_0 \le 2\|\nabla\cdot G\|_{H^j(\Gamma)}|\nabla\cdot Q|_0, \end{aligned}$$

and $F_1u = u$ for polynomials in V_{N1} and $F_2G = G$ for vectors of polynomials from V_{N2}, we can apply the Bramble-Hilbert lemma to find that there exists a constant such that

$$|f_1(\hat{P})| \le C|w|_s|\hat{P}|_{k+1}, \ |f_2(\hat{v})| \le C|Q|_0|\hat{v}|_j, \ |f_3(\nabla\cdot\hat{v})| \le C|\nabla\cdot Q|_0|\nabla\cdot\hat{v}|_j,$$

as long as k and j are small enough so that all polynomials of degree less than or equal to k are contained in the span of the basis functions representing $\hat{u}$ and all polynomials of degree less than or equal to j are contained in

https://doi.org/10.1051/m2an/2011017 Published online by Cambridge University Press

the span of the basis functions representing $\hat{G}$. For the elements used in the implementation to follow, we will have $j = k = 1$. By choosing $w = \hat{P} - F_1\hat{P}$ and $Q = \hat{v} - F_2\hat{v}$, we find that

$$|\hat{P} - F_1\hat{P}|_s \leq C|\hat{P}|_{k+1},\ |\hat{v} - F_2\hat{v}|_0 \leq C|\hat{v}|_j,\ |\nabla\cdot\hat{v} - \nabla\cdot F_2\hat{v}| \leq C|\nabla\cdot\hat{v}|_j.$$

Employing (4.7), we find that for $h \leq 1$,

$$\begin{aligned}
|P - F_1P|_{s,e_l} &\leq Ch^{\frac{d}{2}-s}|\hat{P} - F_1\hat{P}|_s \leq Ch^{\frac{d}{2}-s}|\hat{P}|_{k+1} \leq Ch^{k-s+1}|P|_{k+1,e_l},\\
|v - F_2v|_{0,e_l} &\leq h^{1-\frac{d}{2}}|\hat{v} - F_2\hat{v}|_0 \leq h^{1-\frac{d}{2}}C|\hat{v}|_j \leq Ch^j|v|_{j,e_l},\\
|\nabla\cdot v - \nabla\cdot F_2v|_{0,e_l} &\leq h^{-\frac{d}{2}}|\nabla\cdot\hat{v} - \nabla\cdot F_2\hat{v}|_0 \leq h^{-\frac{d}{2}}C|\nabla\cdot\hat{v}|_j \leq Ch^j|\nabla\cdot v|_{j,e_l}.
\end{aligned}$$

Returning to inequality (4.5), we find that

$$\begin{aligned}
&\|(P,v) - (P_N,v_N)\|_V^2 \leq C\|(P,v) - (F_1P,F_2v)\|_V^2\\
&= C\sum_{l=1}^{(N-1)^2}\left[|P - F_1P|_{0,e_l}^2 + |P - F_1P|_{1,e_l}^2 + |v - F_2v|_{0.e_l}^2 + |\nabla\cdot v - \nabla\cdot F_2v|_{0,e_l}^2\right]\\
&\leq C\sum_{l=1}^{(N-2)^2}\left[h^{2k+2}|P|_{k+1,e_l}^2 + h^{2k}|P|_{k+1,e_l}^2 + h^{2j}|v|_{j,e_l}^2 + h^{2j}|\nabla\cdot v|_{j.e_l}^2\right]\\
&\leq C(h^{2k}|P|_{k+1,\Gamma}^2 + h^{2j}(|v|_{j,\Gamma}^2 + |\nabla\cdot v|_{j,\Gamma}^2)).
\end{aligned}$$

Let $P_k(\bar{\Gamma})$ denote all polynomials of degree less than or equal to k on $\bar{\Gamma}$. We have now proved

Theorem 4.1. *If the solution $(P,v) \in H^{k+1}(\Gamma) \times H^j(div,\Gamma)$ and the finite element subspace used in the numerical method contains $P_k(\bar{\Gamma}) \times P_j(\bar{\Gamma}) \times P_j(\bar{\Gamma})$, then there is a constant C such that the error satisfies*

$$\|(P,v) - (P_N,v_N)\|_V^2 \leq C(h^{2k}|P|_{k+1,\Gamma}^2 + h^{2j}(|v|_{j,\Gamma}^2 + |\nabla\cdot v|_{j,\Gamma}^2)),$$

where $h \leq 1$ is the grid spacing.

4.4. **Regularity**

In order for the error bound to be meaningful, we must have $k, j \geq 1$ in Theorem 4.1, which means that at least

$$P \in H^2(\Gamma) \text{ and } v \in H^1(\text{div},\Gamma).$$

In the notation of the acoustic equation, if ρ^{-1} is positive definite, bounded, and C^1, then classical elliptic regularity theory such as in [5] guarantees that $P' \in H^2(\Gamma)$. Also since

$$v = -\frac{\mathrm{i}}{\omega}\rho^{-1}\nabla P,$$

we have that $v \in (H^1(\Gamma))^2$, and multiplying the acoustic equation through by $-1/\omega$ tells us that

$$\nabla\cdot v = \frac{\mathrm{i}\omega}{\kappa}P,$$

so $\nabla\cdot v \in H^1(\Gamma)$ as long as κ is at least C^1.

It would be more satisfying (and useful in other contexts) to have a regularity theory derived from the weak form of the equations presented herein, and this is a current topic of inquiry for the authors.

https://doi.org/10.1051/m2an/2011017 Published online by Cambridge University Press

5. Implementation of the finite element method

Our goal is to test the efficacy of this new variational principle, using a simple, explicit finite element implementation. Let us assume that $d = 2$ and $\Gamma = (0,1)^2$. In order to find a numerical solution for P', we introduce an $N \times N$ computational grid with equally spaced nodes (x_j, y_t) for $t, j = 1, 2, \ldots, N$ and grid spacing $h = 1/(N-1)$. We also introduce the piecewise bilinear finite element spaces

$$\Psi = \text{span}\left\{\left(1 - \frac{|x-x_j|}{h}\right)\left(1 - \frac{|y-y_t|}{h}\right)\chi_{tj} : 2 \le t, j \le N-1\right\}$$

$$\Phi_1 = \text{span}\left\{\begin{pmatrix}\left(1 - \frac{|x-x_j|}{h}\right)\left(1 - \frac{|y-y_t|}{h}\right)\\ 0\end{pmatrix}\chi_{tj} : 1 \le t, j \le N\right\}$$

$$\Phi_2 = \text{span}\left\{\begin{pmatrix}0\\ \left(1 - \frac{|x-x_j|}{h}\right)\left(1 - \frac{|y-y_t|}{h}\right)\end{pmatrix}\chi_{tj} : 1 \le t, j \le N\right\},$$

where

$$\chi_{tj}(x,y) = \begin{cases} 1 & \text{if } |x-x_j|, |y-y_t| \le h \\ 0 & \text{otherwise.}\end{cases}$$

We can re-index these elements with a single index by setting

$$\psi_k = \left(1 - \frac{|x-x_j|}{h}\right)\left(1 - \frac{|y-y_t|}{h}\right)\chi_{tj}, \text{ where } k = (t-2)(N-2) + j - 1,\ k = 1, \ldots, (N-2)^2,$$

$$\phi_{1k} = \begin{pmatrix}\left(1 - \frac{|x-x_j|}{h}\right)\left(1 - \frac{|y-y_t|}{h}\right)\\ 0\end{pmatrix}\chi_{tj} \text{ where } k = (t-1)N + j,\ k = 1, \ldots, N^2,$$

$$\phi_{2k} = \begin{pmatrix}0\\ \left(1 - \frac{|x-x_j|}{h}\right)\left(1 - \frac{|y-y_t|}{h}\right)\end{pmatrix}\chi_{tj} \text{ where } k = (t-1)N + j,\ k = 1, \ldots, N^2.$$

We assume that our finite element solution has the form

$$\begin{pmatrix} P' \\ v'' \end{pmatrix} = \begin{pmatrix} \psi_R + \sum_{k=1}^{(N-2)^2} \delta_k \psi_k \\ \sum_{k=1}^{N(N-1)} \beta_k \phi_{1k} + \sum_{k=1}^{N(N-1)} \gamma_k \phi_{2k} \end{pmatrix}.$$

Here ψ_R is any function that satisfies the desired Dirichlet boundary condition for P'. The weak form of the Euler-Lagrange equation for the variational principle is

$$0 = \int_\Gamma \left[\begin{pmatrix}\nabla P' \\ -\omega v''\end{pmatrix} \cdot \mathcal{R}\begin{pmatrix}\nabla s \\ -\omega T\end{pmatrix} + \begin{pmatrix}\omega P' \\ -\nabla \cdot v''\end{pmatrix} \cdot \mathcal{K}\begin{pmatrix}\omega s \\ -\nabla \cdot T\end{pmatrix}\right] \mathrm{d}x + \int_{\partial\Gamma} \omega T \cdot n P'' \, \mathrm{d}S \tag{5.1}$$

for any $s \in H_0^1(\Gamma)$ and any $T \in H(\text{div}, \Gamma)$. Substituting into (5.1), we get

$$\begin{aligned}\int_\Gamma \left[\begin{pmatrix}\sum \delta_k \nabla \psi_k \\ -\omega \sum \beta_k \phi_{1k} - \omega \sum \gamma_k \phi_{2k}\end{pmatrix} \cdot \mathcal{R}\begin{pmatrix}\nabla s \\ -\omega T\end{pmatrix} + \begin{pmatrix}\omega \sum \delta_k \psi_k \\ -\sum \beta_k \nabla \cdot \phi_{1k} - \sum \gamma_k \nabla \cdot \phi_{2k}\end{pmatrix} \cdot \mathcal{K}\begin{pmatrix}\omega s \\ -\nabla \cdot T\end{pmatrix}\right] \mathrm{d}x \\ = -\int_\Gamma \left[\begin{pmatrix}\nabla \psi_0 \\ 0\end{pmatrix} \cdot \mathcal{R}\begin{pmatrix}\nabla s \\ -\omega T\end{pmatrix} + \begin{pmatrix}\omega \psi_0 \\ 0\end{pmatrix} \cdot \mathcal{K}\begin{pmatrix}\omega s \\ -\nabla \cdot T\end{pmatrix}\right] \mathrm{d}x - \int_\Gamma \left[\omega \nabla \psi_I \cdot T + \omega \psi_I \nabla \cdot T\right] \mathrm{d}x,\end{aligned}$$

https://doi.org/10.1051/m2an/2011017 Published online by Cambridge University Press

where we have used the divergence theorem on the boundary integral, ψ_I is any function on Γ satisfying the desired Dirichlet boundary condition for P'', and $s \in H_0^1(\Gamma)$, $T \in H(\mathrm{div}, \Gamma)$ are arbitrary. In particular, this must hold when

$$s = \psi_k,\ T = 0 \text{ for } k = 1, \ldots, (N-2)^2$$

$$s = 0,\ T = \phi_{1k} \text{ for } k = 1, \ldots, N(N-1)$$

$$s = 0,\ T = \phi_{2k} \text{ for } k = 1, \ldots, N(N-1).$$

This gives rise to a system of equations of the form $A\alpha = b$, where A has the block form

$$A = \begin{pmatrix} A_1 & A_4 & A_6 \\ A_4 & A_2 & A_5 \\ A_6 & A_5 & A_3 \end{pmatrix} \tag{5.2}$$

and the blocks have entries

$$\begin{aligned}
(A_1)_{tj} &= \int_\Gamma \left[\begin{pmatrix} \nabla\psi_t \\ 0 \end{pmatrix} \cdot \mathcal{R} \begin{pmatrix} \nabla\psi_j \\ 0 \end{pmatrix} + \begin{pmatrix} \omega\psi_t \\ 0 \end{pmatrix} \cdot \mathcal{K} \begin{pmatrix} \omega\psi_j \\ 0 \end{pmatrix} \right] \mathrm{d}x \\
(A_2)_{tj} &= \int_\Gamma \left[\begin{pmatrix} 0 \\ -\omega\phi_{1t} \end{pmatrix} \cdot \mathcal{R} \begin{pmatrix} 0 \\ -\omega\phi_{ij} \end{pmatrix} + \begin{pmatrix} 0 \\ -\nabla\cdot\phi_{1t} \end{pmatrix} \cdot \mathcal{K} \begin{pmatrix} 0 \\ -\nabla\cdot\phi_{1j} \end{pmatrix} \right] \mathrm{d}x \\
(A_3)_{tj} &= \int_\Gamma \left[\begin{pmatrix} 0 \\ -\omega\phi_{2t} \end{pmatrix} \cdot \mathcal{R} \begin{pmatrix} 0 \\ -\omega\phi_{2j} \end{pmatrix} + \begin{pmatrix} 0 \\ -\nabla\cdot\phi_{2t} \end{pmatrix} \cdot \mathcal{K} \begin{pmatrix} 0 \\ -\nabla\cdot\phi_{2j} \end{pmatrix} \right] \mathrm{d}x \\
(A_4)_{tj} &= \int_\Gamma \left[\begin{pmatrix} 0 \\ -\omega\phi_{1t} \end{pmatrix} \cdot \mathcal{R} \begin{pmatrix} \nabla\psi_j \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ -\nabla\cdot\phi_{1t} \end{pmatrix} \cdot \mathcal{K} \begin{pmatrix} \omega\psi_j \\ 0 \end{pmatrix} \right] \mathrm{d}x. \\
(A_5)_{tj} &= \int_\Gamma \left[\begin{pmatrix} 0 \\ -\omega\phi_{2t} \end{pmatrix} \cdot \mathcal{R} \begin{pmatrix} 0 \\ -\omega\phi_{1j} \end{pmatrix} + \begin{pmatrix} 0 \\ -\nabla\cdot\phi_{2t} \end{pmatrix} \cdot \mathcal{K} \begin{pmatrix} 0 \\ -\nabla\cdot\phi_{1j} \end{pmatrix} \right] \mathrm{d}x \\
(A_6)_{tj} &= \int_\Gamma \left[\begin{pmatrix} 0 \\ -\omega\phi_{2t} \end{pmatrix} \cdot \mathcal{R} \begin{pmatrix} \nabla\psi_j \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ -\nabla\cdot\phi_{2t} \end{pmatrix} \cdot \mathcal{K} \begin{pmatrix} \omega\psi_j \\ 0 \end{pmatrix} \right] \mathrm{d}x.
\end{aligned} \tag{5.3}$$

The right-hand side vector b is partitioned as

$$b = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix},$$

where

$$\begin{aligned}
(b_1)_k = &- \int_\Gamma \left[\begin{pmatrix} \nabla\psi_R \\ 0 \end{pmatrix} \cdot \mathcal{R} \begin{pmatrix} \nabla\psi_k \\ 0 \end{pmatrix} + \begin{pmatrix} \omega\psi_R \\ 0 \end{pmatrix} \cdot \mathcal{K} \begin{pmatrix} \omega\psi_k \\ 0 \end{pmatrix} \right] \mathrm{d}x \\
(b_2)_k = &- \int_\Gamma \left[\begin{pmatrix} \nabla\psi_R \\ 0 \end{pmatrix} \cdot \mathcal{R} \begin{pmatrix} 0 \\ -\omega\phi_{1k} \end{pmatrix} + \begin{pmatrix} \omega\psi_R \\ 0 \end{pmatrix} \cdot \mathcal{K} \begin{pmatrix} 0 \\ -\nabla\cdot\phi_{1k} \end{pmatrix} \right] \mathrm{d}x \\
&- \int_\Gamma \left[\omega\nabla\psi_I \cdot \phi_{1k} + \omega\psi_I \nabla\cdot\phi_{1k} \right] \mathrm{d}x \\
(b_3)_k = &- \int_\Gamma \left[\begin{pmatrix} \nabla\psi_R \\ 0 \end{pmatrix} \cdot \mathcal{R} \begin{pmatrix} 0 \\ -\omega\phi_{2k} \end{pmatrix} + \begin{pmatrix} \omega\psi_R \\ 0 \end{pmatrix} \cdot \mathcal{K} \begin{pmatrix} 0 \\ -\nabla\cdot\phi_{2k} \end{pmatrix} \right] \mathrm{d}x \\
&- \int_\Gamma \left[\omega\nabla\psi_I \cdot \phi_{2k} + \omega\psi_I \nabla\cdot\phi_{2k} \right] \mathrm{d}x.
\end{aligned} \tag{5.4}$$

https://doi.org/10.1051/m2an/2011017 Published online by Cambridge University Press

The method for solving for P' and v'' can be easily modified to solve for P'' and v'. In this case the weak equation is

$$\int_\Gamma \left[\begin{pmatrix} \nabla P'' \\ \omega v' \end{pmatrix} \cdot \mathcal{R} \begin{pmatrix} \nabla s \\ \omega T \end{pmatrix} + \begin{pmatrix} -\omega P'' \\ -\nabla \cdot v' \end{pmatrix} \cdot \mathcal{K} \begin{pmatrix} -\omega s \\ -\nabla \cdot T \end{pmatrix} \right] \mathrm{d}x + \int_{\partial\Gamma} \omega T \cdot n P' \, \mathrm{d}S,$$

and all the methods above still apply. In fact, to obtain the new matrix for this formulation, we simply change the signs of the blocks A_4 and A_6, and the changes in b are mostly reversing signs and the roles of the two auxiliary functions ψ_R and ψ_I.

5.1. **Other discretizations**

Along with the discretization described in Section 5, we have experimented with two other implementations in which different basis functions are used to represent the variable v. The first of these uses the Raviart-Thomas $RT_{[0]}$ elements described in [2]. We found that the resulting finite element matrix is much more poorly scaled, with a condition number approximately twice as large as that obtained with the bilinear basis. The second method uses the $RT_{[1]}$ elements (also described in [2]). In this case, the higher-order basis functions obviously result in a somewhat less-sparse finite element matrix, and the condition number is approximately the same as that obtained with the all-bilinear discretization. In the numerical experiments described later, it was observed that the difference in error that resulted from these different discretizations was negligible.

6. Conditioning

As was mentioned, perhaps the greatest numerical advantage to having a minimization formulation for the Helmholtz equation is that the matrix produced by the finite element method is symmetric positive definite. This allows for the use of methods such as the conjugate gradient method to solve the system. Of course, the use of a preconditioning matrix in the conjugate gradient method can speed up the convergence considerably, which is especially important when solving the relatively large sparse systems generated by the finite element approach outlined above.

In our approach, there are three basic types of elements used: bilinear elements, first component bilinear vector elements, and second component bilinear vector elements. Each of these types of elements interacts with all of the other types, and these interactions are what give rise to the blocks in (5.2). Assuming that interactions among similar element types are most important, we choose the block Jacobi preconditioner

$$M = \begin{pmatrix} A_1 & 0 & 0 \\ 0 & A_2 & 0 \\ 0 & 0 & A_3 \end{pmatrix}.$$

Among all block diagonal preconditioners of this form, this choice of M minimizes the condition number of $M^{-\frac{1}{2}} A M^{-\frac{1}{2}}$ to within a factor of 3 of its minimum [4]. In our numerical experiments, we used an incomplete Choleski factorization of M as the preconditioner in the PCG iterations, which dramatically increases the efficiency over using M alone. Figure 1 shows the distribution of the eigenvalues of the matrix A before and after preconditioning for $N = 30$. In Figure 2, we see the how the number of PCG iterations grows with N in a specific example problem for several error tolerances.

A key component in ensuring that the system $A\alpha = b$ is well conditioned is for the matrix $\mathcal{L}$ (or equivalently $\mathcal{R}$ and $\mathcal{K}$) to have a coercivity constant that is as large as possible. For this reason, we expect better numerical results when the eigenvalues of $\mathcal{L}$ are bounded well away from zero. In the case of the Helmholtz equation, the matrix Z is diagonal, say $Z = \mathrm{diag}(c_1, \ldots, c_{d+1})$, which makes it possible to calculate the eigenvalues of $\mathcal{L}$. If D is an invertible matrix, then we may factor a block matrix

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

https://doi.org/10.1051/m2an/2011017 Published online by Cambridge University Press

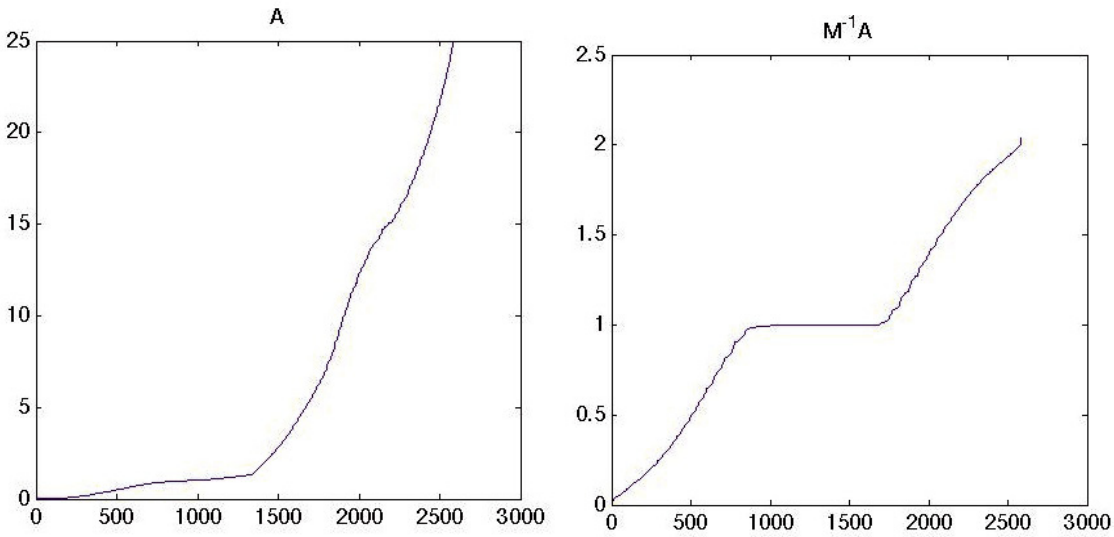


FIGURE 1. The distribution of the eigenvalues of A and the real parts of the eigenvalues of $M^{-1}A$ for $N = 30$.

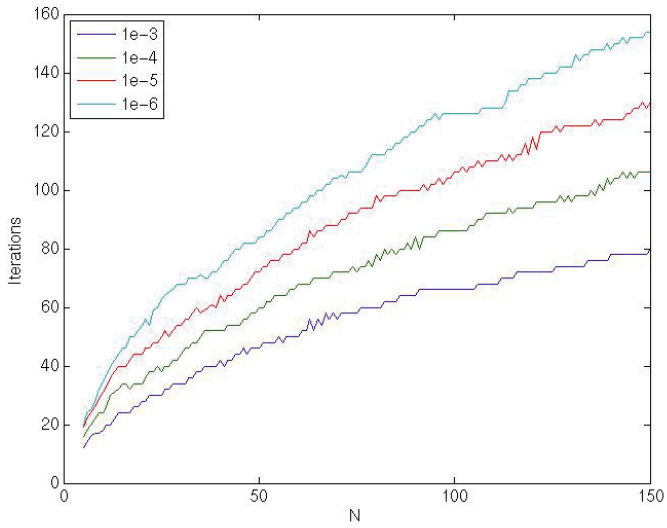


FIGURE 2. The growth of the number of PCG iterations required to solve a given problem with grid size for several error tolerances.

as

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I & B \\ 0 & D \end{pmatrix} \begin{pmatrix} A - BD^{-1}C & 0 \\ D^{-1}C & I \end{pmatrix},$$

which implies that

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(D)\det(A - BD^{-1}C).$$

https://doi.org/10.1051/m2an/2011017 Published online by Cambridge University Press

TABLE 1. The error in the finite element solution for various values of the grid size N.

N	h	$\|(P-P_N, v-v_N)\|_V$
30	0.0345	6.6162×10^{-4}
40	0.0256	3.6692×10^{-4}
50	0.0204	2.3252×10^{-4}
60	0.0169	1.6026×10^{-4}
70	0.0145	1.1722×10^{-4}
80	0.0127	8.9706×10^{-5}
90	0.0112	7.0686×10^{-5}
100	0.0101	5.7037×10^{-5}

Therefore,

$$\det(\mathcal{L}-\lambda I) = (-1)^{d+1}\det\begin{pmatrix} (Z'')^{-1}Z' & (Z'')^{-1}-\lambda I \\ Z''+Z'(Z'')^{-1}Z'-\lambda I & Z'(Z'')^{-1} \end{pmatrix}$$
$$= (-1)^{d+1}\det(Z'(Z'')^{-1})\det((Z'')^{-1}Z' + [-(Z')^{-1}+\lambda Z''(Z')^{-1}][Z''+Z'(Z'')^{-1}Z'-\lambda I])$$
$$= (-1)^{d+1}\det(Z'(Z'')^{-1})\det(\lambda^2[-Z''(Z')^{-1}] + \lambda[(Z')^{-1}+Z''(Z')^{-1}Z''+Z'] - (Z')^{-1}Z'').$$

In the case of diagonal Z, this implies that

$$\lambda = \frac{-a_j \pm \sqrt{a_j^2-b_j^2}}{-b_j} \quad j=1,\dots,d+1,$$

where

$$a_j = \frac{1}{c_j'} + \frac{(c_j'')^2}{c_j'} + c_j' \quad \text{and} \quad b_j = 2\frac{c_j''}{c_j'}.$$

If $Z'=0$, then $\mathcal{L}$ is diagonal, and its eigenvalues are those of Z'' and $(Z'')^{-1}$.

The above analysis tells us that the finite element problem will be better conditioned for those problems where the coefficients ρ and κ are such that Z is close to Ii, *i.e.* $\rho = \mathrm{i}I$ and $\kappa = -I\mathrm{i}$ (this would correspond to the limiting case where $a_j = b_j$). In many cases when we are presented with a problem where the coercivity constant for $\mathcal{L}$ is small, we can apply an appropriate rotation and scaling to the problem in order to get a finite element matrix that is better conditioned. By multiplying the problem (2.1) through by a complex constant $r\mathrm{e}^{\mathrm{i}\theta}$, we effectively replace Z with $r\mathrm{e}^{\mathrm{i}\theta}Z$, so we should choose r and θ so that $r\mathrm{e}^{\mathrm{i}\theta}Z$ is as close as possible to $\mathrm{i}I$. However, this may not always be possible, for example, when an isotropic $\rho(x)$ oscillates between values in the upper half of the complex plane that are close to 1 and -1.

7. NUMERICAL RESULTS

As an example, we demonstrate the error bound on the problem (1.1), with parameters $\rho = (-5+5\mathrm{i})I$, $\kappa = 4-4\mathrm{i}$ and $\omega = 2$. A solution is $P(x,y) = \mathrm{e}^{2\mathrm{i}x-3y}$. In this example we took

$$\psi_R = \mathrm{Re}(\mathrm{e}^{2\mathrm{i}x-3y}) + \sin(\pi x)\sin(\pi y),\ \ \psi_I = \mathrm{Im}(\mathrm{e}^{2\mathrm{i}x-3y}) + \sin(\pi x)\cdot 3\sin(\pi y))$$

and solved the problem on grids with $N = 3,\dots,100$. Table 1 shows the error in the finite element solution for various values of N. The errors were calculated using the trapezoidal rule with function evaluations on a grid with size $N = 1500$. Figure 3 demonstrates the method on a problem with non-constant coefficients, where the dissipation in the material is higher inside a disk (whose boundary is approximated by a staircase curve to fit the rectangular grid) centered in the unit square and with radius 0.25. The boundary conditions for the real part are oscillatory, while the boundary conditions for the imaginary part are simply an affine function.

https://doi.org/10.1051/m2an/2011017 Published online by Cambridge University Press

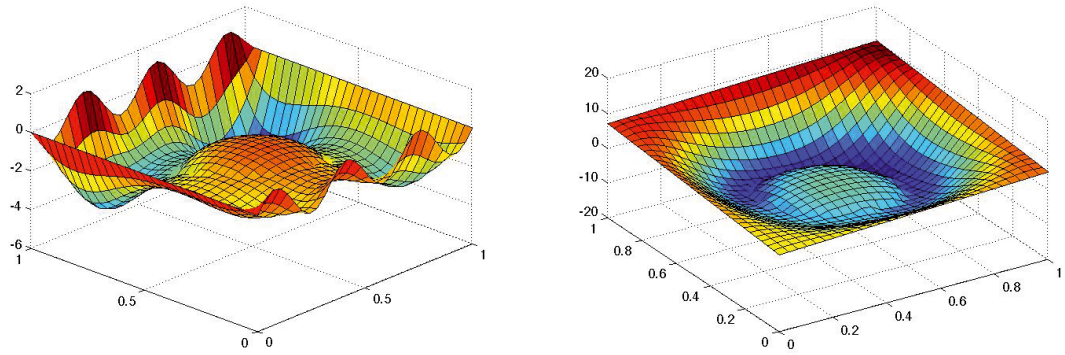


FIGURE 3. The solutions P' (left) and P'' (right), with $\omega = 10, \psi_R = \sin(6\pi x)\cos(3\pi y),\ \psi_I = 3x + 5y + 2,\ N = 30$. There is an approximately circular inclusion in the center of the domain with $\rho = .01 + .001\mathrm{i}$ and $\kappa = .01 - .003\mathrm{i}$ outside the inclusion and $\rho = -5 + 5\mathrm{i}$ and $\kappa = 4 - 4\mathrm{i}$ inside the inclusion.

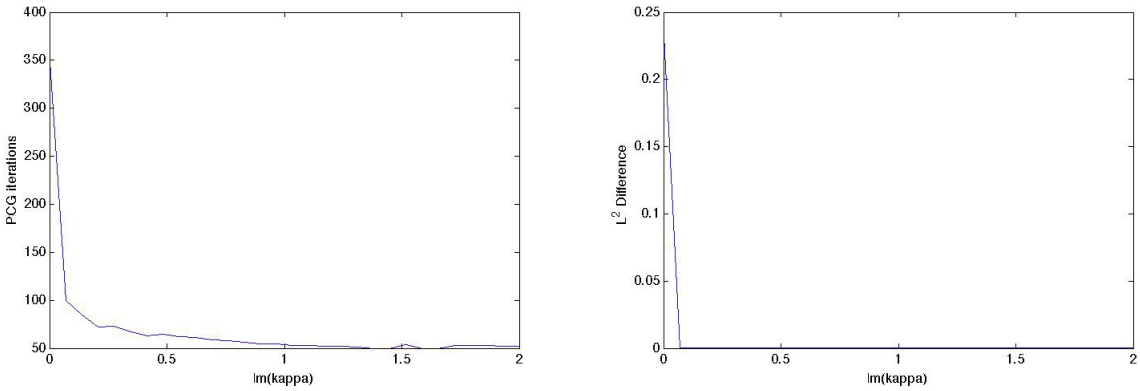


FIGURE 4. The graph on the left shows the increase in the number of PCG iterations as $\kappa'' \to 0$. On the right is the L^2 difference between solutions computed *via* minimization and stationary variational principles as κ'' vanishes.

In Figure 4, we see the effect of letting the imaginary part of κ approach zero while the rest of the problem parameters remain fixed. The parameters used to generate this figure were

$$N = 100,\ \omega = 2,\ \rho = 3 + \mathrm{i},\ \kappa' = 3.$$

On the left, we see that the loss of coercivity causes the number of PCG iterations to increase. On the right we see the L^2 difference between a solution using the minimization method and one computed from a finite element discretization of the stationary variational principle (1.2).

In Figure 5, a model problem is solved several times with increasing values of ω. As ω increases, the ratio ω/N is held approximately constant. On the left, we see the relative error in the $\|\cdot\|_V$-norm. On the right, we see that the number of PCG iterations required to solve the problems remains stable as ω increases.

https://doi.org/10.1051/m2an/2011017 Published online by Cambridge University Press

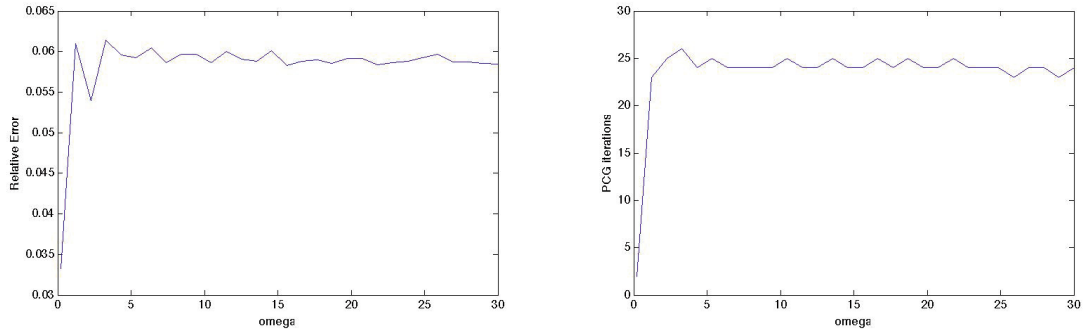


FIGURE 5. The graph on the left shows the relative error as ω increases. On the right is the number of PCG iterations to solve the problems for different ω.

8. Robin boundary conditions

8.1. Problem formulation

Another boundary condition that often appears is the Robin problem

$$\begin{cases} -\nabla \cdot \rho^{-1} \nabla P - \dfrac{\omega^2}{\kappa} P = 0 & \text{in } \Gamma, \\ P + av \cdot n = g & \text{on } \partial\Gamma, \end{cases}$$

where $a \in \mathbb{C}$. In order to deal with this boundary condition, which concerns both real and imaginary parts of the variables P and v simultaneously, we start with the minimization functional for the natural boundary conditions

$$Y(P', v'') + 2\omega \int_{\partial\Gamma} [P' v' \cdot n + P'' v'' \cdot n] \, \mathrm{d}S.$$

The Euler-Lagrange equation for the corresponding variational principle is

$$B(P', v'', s, T) = -\omega \int_{\partial\Gamma} [s v' \cdot n + P'' T \cdot n] \, \mathrm{d}S$$

where the bilinear form B is defined in (4.1). Notice that we can write the surface integral above as

$$-\omega \int_{\partial\Gamma} \begin{pmatrix} v' \cdot n \\ P'' \end{pmatrix} \cdot \begin{pmatrix} s \\ T \cdot n \end{pmatrix} \mathrm{d}S.$$

The vector on the right contains the primary variables for which we would like to solve, and the vector on the left contains the dual variables which we would like to eliminate using the Robin boundary condition. In terms of the vectors above, we can express the Robin condition as

$$M_1 \begin{pmatrix} P' \\ v'' \cdot n \end{pmatrix} + M_2 \begin{pmatrix} v' \cdot n \\ P'' \end{pmatrix} = \begin{pmatrix} g' \\ g'' \end{pmatrix},$$

where

$$M_1 = \begin{pmatrix} 1 & -a'' \\ 0 & a' \end{pmatrix} \text{ and } M_2 = \begin{pmatrix} a' & 0 \\ a'' & 1 \end{pmatrix}.$$

https://doi.org/10.1051/m2an/2011017 Published online by Cambridge University Press

Rearranging, we find that

$$\begin{pmatrix} v' \cdot n \\ P'' \end{pmatrix} = M_2^{-1} \begin{pmatrix} g' \\ g'' \end{pmatrix} - M_2^{-1} M_1 \begin{pmatrix} P' \\ v'' \cdot n \end{pmatrix},$$

so the surface integral term becomes

$$-\omega \int_{\partial\Gamma} \left[M_2^{-1} \begin{pmatrix} g' \\ g'' \end{pmatrix} - M_2^{-1} M_1 \begin{pmatrix} P' \\ v'' \cdot n \end{pmatrix} \right] \cdot \begin{pmatrix} S \\ T \cdot n \end{pmatrix} \mathrm{d}S$$

$$= -\omega \int_{\partial\Gamma} M_2^{-1} \begin{pmatrix} g' \\ g'' \end{pmatrix} \cdot \begin{pmatrix} P' \\ v'' \cdot n \end{pmatrix} \mathrm{d}S + \omega \int_{\partial\Gamma} M_2^{-1} M_1 \begin{pmatrix} P' \\ v'' \cdot n \end{pmatrix} \cdot \begin{pmatrix} P' \\ v'' \cdot n \end{pmatrix} \mathrm{d}S.$$

The new Euler-Lagrange equation for the Robin boundary condition is therefore

$$B(P', v''; s, T) - \omega \int_{\partial\Gamma} M_2^{-1} M_1 \begin{pmatrix} P' \\ v'' \cdot n \end{pmatrix} \cdot \begin{pmatrix} s \\ T \cdot n \end{pmatrix} \mathrm{d}S$$

$$= -\omega \int_{\partial\Gamma} M_2^{-1} \begin{pmatrix} g' \\ g'' \end{pmatrix} \cdot \begin{pmatrix} s \\ T \cdot n \end{pmatrix} \mathrm{d}S.$$

Since

$$M_2^{-1} = \frac{1}{a'} \begin{pmatrix} 1 & 0 \\ -a'' & a' \end{pmatrix},$$

we have

$$M_2^{-1} M_1 = \frac{1}{a'} \begin{pmatrix} 1 & -a'' \\ -a'' & |a|^2 \end{pmatrix},$$

which is positive definite as long as $a' > 0$. The new bilinear form above is guaranteed to be coercive as long as ρ and κ satisfy (2.3) and $a' < 0$.

To find a numerical solution for the Robin boundary value problem, we discretize using the finite element scheme presented in Section 5. Unfortunately, the surface integrals can no longer be converted to volume integrals by integration by parts and must be computed as they stand. In this case, the finite element matrix is written as the sum of two matrices $A - \omega B$, where A is of the form (5.2), and the blocks have entries (5.3), and another matrix B with the same block form and block entries

$$\begin{aligned}
(B_1)_{tj} &= \int_{\partial\Gamma} \begin{pmatrix} \psi_t \\ 0 \end{pmatrix} \cdot M_2^{-1} M_1 \begin{pmatrix} \psi_j \\ 0 \end{pmatrix} \mathrm{d}S \\
(B_2)_{tj} &= \int_{\partial\Gamma} \begin{pmatrix} 0 \\ \phi_{1t} \cdot n \end{pmatrix} \cdot M_2^{-1} M_1 \begin{pmatrix} 0 \\ \phi_{1j} \cdot n \end{pmatrix} \mathrm{d}S \\
(B_3)_{tj} &= \int_{\partial\Gamma} \begin{pmatrix} 0 \\ \phi_{2t} \cdot n \end{pmatrix} \cdot M_2^{-1} M_1 \begin{pmatrix} 0 \\ \phi_{2j} \cdot n \end{pmatrix} \mathrm{d}S \\
(B_4)_{tj} &= \int_{\partial\Gamma} \begin{pmatrix} 0 \\ \phi_{1t} \cdot n \end{pmatrix} \cdot M_2^{-1} M_1 \begin{pmatrix} \psi_j \\ 0 \end{pmatrix} \mathrm{d}S \\
(B_5)_{tj} &= \int_{\partial\Gamma} \begin{pmatrix} 0 \\ \phi_{2t} \cdot n \end{pmatrix} \cdot M_2^{-1} M_1 \begin{pmatrix} 0 \\ \phi_{1j} \cdot n \end{pmatrix} \mathrm{d}S \\
(B_6)_{tj} &= \int_{\partial\Gamma} \begin{pmatrix} 0 \\ \phi_{2t} \cdot n \end{pmatrix} \cdot M_2^{-1} M_1 \begin{pmatrix} \psi_j \\ 0 \end{pmatrix} \mathrm{d}S.
\end{aligned}$$

https://doi.org/10.1051/m2an/2011017 Published online by Cambridge University Press

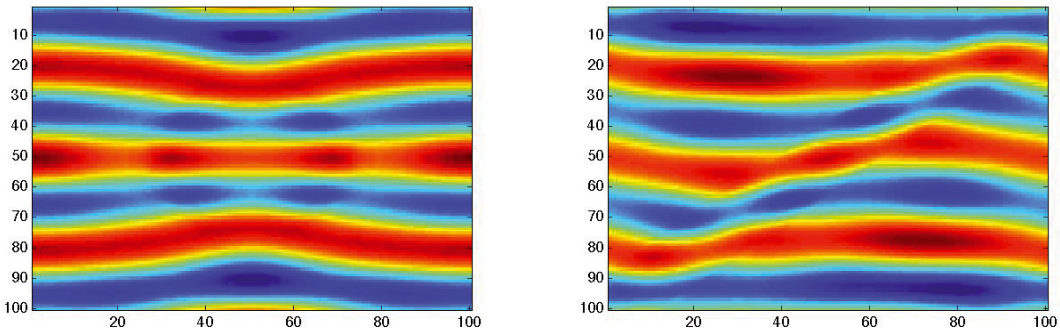


FIGURE 6. On the left is shown the solution to the Robin problem with a disc of high density material centered in the domain. On the right the disk is replaced with a bar of the same material angled from the lower left to the upper right of the domain.

The right-hand side vector b is also partitioned as $(b_1, b_2, b_3)^T$ with entries

$$\begin{aligned}
(b_1)_k &= -\omega \int_{\partial\Gamma} \begin{pmatrix} \psi_k \\ 0 \end{pmatrix} \cdot M_2^{-1} \begin{pmatrix} g' \\ g'' \end{pmatrix} \mathrm{d}S \\
(b_2)_k &= -\omega \int_{\partial\Gamma} \begin{pmatrix} 0 \\ \phi_{1k} \cdot n \end{pmatrix} \cdot M_2^{-1} \begin{pmatrix} g' \\ g'' \end{pmatrix} \mathrm{d}S \\
(b_3)_k &= -\omega \int_{\partial\Gamma} \begin{pmatrix} 0 \\ \phi_{2k} \cdot n \end{pmatrix} \cdot M_2^{-1} \begin{pmatrix} g' \\ g'' \end{pmatrix} \mathrm{d}S.
\end{aligned}$$

Assuming that the coercivity requirements (2.3) on ρ and κ are satisfied, and $a' < 0$, the system

$$(A - \omega B)\alpha = b$$

may be solved using the same preconditioned conjugate gradient approach as outlined previously.

8.2. **Numerical examples**

Here we present some numerical examples obtained by using the finite element method to solve problems with Robin boundary conditions. In these examples the Robin boundary conditions are imposed on $y = 0$ and $y = 1$, while on the other sides of the domain we have imposed periodic boundary conditions. On the left in Figure 6 is the solution with a circular scatterer with $\rho = 1 + .011\mathrm{i}$ outside the scatterer, $\rho = 2 + .011\mathrm{i}$ inside the scatterer, $\kappa = 1 + .011\mathrm{i}$ everywhere, $a = -1 + .333\mathrm{i}$ and $g = 3.33\mathrm{i}$. On the right, the circular scatterer is replaced by a bar angled across the domain, but the other parameters in the problem remain the same. These results were calculated using the $RT_{[0]}$ discretization for the v variable described in Section 5.1.

9. CONCLUSIONS

The variational principles of Milton, Seppecher, Bouchitté, and Willis make it possible to formulate the solution of the Helmholtz equation as a minimization, and this is reflected in the fact that the stiffness matrix for the finite element method is symmetric positive definite. This allows us to use classical iterative methods such as preconditioned conjugate gradient to solve the associated system, along with straightforward finite element error estimates. The primary advantage of this approach is that it allows the use of efficient iterative methods for the solution of the linear system. But there are also disadvantages in that the system has more unknowns, since we must solve for P and v simultaneously.

https://doi.org/10.1051/m2an/2011017 Published online by Cambridge University Press

More research is necessary to determine the circumstances under which this approach may be more effective than others currently in use. A particular point of interest is that even though the underlying minimization principles are valid for arbitrarily small loss coefficients, the conditioning of the associated finite element matrix deteriorates as the system becomes less dissipative. The general question of how solution efficiency depends on loss should be analyzed further.

We have only approached the scalar, two-dimensional Helmholtz equation in this paper, while the minimization principles of Milton, Seppecher, Bouchitté, and Willis apply to the full vector Maxwell equations, as well as the equations of linear elasticity. The general approach taken here should also apply in those cases. We note finally that in many applications for these models, one would like to apply nonlocal transmission or radiation boundary conditions in order to accurately handle unbounded domains. The problem of adapting these minimization methods to such boundary conditions remains open, although presumably the PML approach (see *e.g.* [6]) would apply.

Acknowledgements. Russell Richins is grateful for support from the National Science Foundation through grant DMS-0707978. Also, the authors would like to thank Graeme Milton for many helpful suggestions, and John Willis who, along with Graeme Milton, clarified an earlier formulation of the boundary value problem for us. The authors would also like to thank the referee for many valuable comments and suggestions.

References

[1] O. Axelsson and V.A. Barker, *Finite element solution of boundary value problems, theory and computation.* SIAM, Philidelphia, PA (2001).
[2] F. Brezzi and M. Fortin, *Mixed and hybrid finite element methods.* Springer-Verlag, New York, NY (1991).
[3] A.V. Cherkaev and L.V. Gibiansky, Variational principles for complex conductivity, viscoelasticity, and similar problems in media with complex moduli. *J. Math. Phys.* **35** (1994) 127–145.
[4] J. Demmel, The condition number of equivalence transformations that block diagonalize matrix pencils. *SIAM J. Num. Anal.* **20** (1983) 599–610.
[5] L.C. Evans, *Partial differential equations.* American Mathematical Society, Providence, RI (1998).
[6] I. Harari, M. Slavutin and E. Turkel, Analytical and numerical studies of a finite element PML for the Helmholtz equation. *J. Comp. Acoust.* **8** (2000) 121–137.
[7] G.W. Milton and J.R. Willis, On modifications of newton's second law and linear continuum elastodynamics. *Proc. R. Soc. A* **463** (2007) 855–880.
[8] G.W. Milton and J.R. Willis, Minimum variational principles for time-harmonic waves in a dissipative medium and associated variational principles of Hashin-Shtrikman type. *Proc. R. Soc. Lond.* **466** (2010) 3013–3032.
[9] G.W. Milton, P. Seppecher and G. Bouchitté, Minimization variational principles for acoustics, elastodynamics, and electromagnetism in lossy inhomogeneous bodies at fixed frequency. *Proc. R. Soc. A* **465** (2009) 367–396.
[10] V.V. Tyutekin and Y.V. Tyutekin, Sound absorbing media with two types of acoustic losses. *Acoust. Phys.* **56** (2010) 33–36.

https://doi.org/10.1051/m2an/2011017 Published online by Cambridge University Press