Available online at www.sciencedirect.com
ScienceDirect
Journal of Computational and Applied Mathematics 204 (2007) 493–504
JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS
www.elsevier.com/locate/cam

Error analysis and preconditioning for an enhanced DtN-FE algorithm for exterior scattering problems

Leonid Chindelevitch[a], David P. Nicholls[b], Nilima Nigam[a,*]

[a] *Department of Mathematics and Statistics, McGill University, 805 Sherbrooke West, Montréal, Que., Canada H3A 2K6*
[b] *Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, 851 South Morgan Street, Chicago, IL 60607, USA*

Received 16 October 2005; received in revised form 26 January 2006

Abstract

In this paper we present an error analysis for a high-order accurate combined Dirichlet-to-Neumann (DtN) map/finite element (FE) algorithm for solving two-dimensional exterior scattering problems. We advocate the use of an *exact* DtN (or Steklov–Poincaré) map at an artificial boundary exterior to the scatterer to truncate the unbounded computational region. The advantage of using an exact DtN map is that it provides a transparent condition which does not reflect scattered waves unphysically. Our algorithm allows for the specification of quite general artificial boundaries which are perturbations of a circle. To compute the DtN map on such a geometry we utilize a boundary perturbation method based upon recent theoretical work concerning the analyticity of the DtN map. We also present some preliminary work concerning the preconditioning of the resulting system of linear equations, including numerical experiments.

Keywords: Bounded obstacle scattering; Finite elements; Transparent boundary conditions; Boundary perturbations; Acoustics; Electromagnetics

1. Introduction

We consider the scattering of time-harmonic acoustic (electromagnetic) radiation by a bounded, sound-hard (perfectly conducting) obstacle D in $\mathbb{R}^2$. Our treatment remains the same for other boundary conditions with obvious modifications. The boundary of D, denoted by ∂D, is only required to be Lipschitz continuous. It is well known [6] that the scattered field $v(x)$ satisfies

$$\Delta v + k^2 v = 0, \quad x \in \mathbb{R}^2 \backslash (\bar{D}), \tag{1a}$$

$$\partial_n v = g, \quad x \in \partial D, \tag{1b}$$

$$\lim_{r\to\infty} \sqrt{r}(\partial_r v - \mathrm{i}kv) = 0. \tag{1c}$$

Here k is specified by the wavenumber of the incident radiation and n is the unit normal pointing exterior to D. The Neumann data g are given in terms of the incident wave. The Sommerfeld radiation condition (1c) is prescribed to

* Corresponding author.
E-mail address: nigam@math.mcgill.ca (N. Nigam).

0377-0427/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2006.02.051

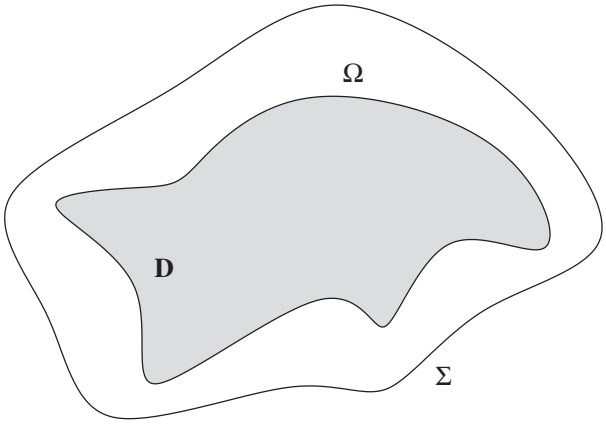

Fig. 1. D is the obstacle and Σ is the enclosing artificial boundary.

ensure uniqueness of solutions [6]. A considerable challenge to numerical simulation of solutions of (1) is the infinite nature of the domain, coupled with the desire to faithfully enforce the Sommerfeld radiation condition.

A common approach to the numerical solution of (1) is to introduce an artificial boundary Σ properly enclosing $\bar{D}$, and then to discretize the annular domain between them, Ω (see Fig. 1). This introduces a natural domain decomposition which leads to a system of equations coupled across Σ and equivalent to (1):

$$\Delta u + k^2 u = 0, \quad x \in \Omega, \tag{2a}$$

$$\partial_n u = g, \quad x \in \partial D, \tag{2b}$$

$$\partial_N u = \partial_N w, \quad x \in \Sigma, \tag{2c}$$

$$u = w, \quad x \in \Sigma, \tag{2d}$$

$$\Delta w + k^2 w = 0, \quad x \in \mathrm{Ext}(\Sigma), \tag{2e}$$

$$\lim_{r\to\infty} \sqrt{r}(\partial_r w - \mathrm{i}kw) = 0, \tag{2f}$$

where N is an inward-pointing normal to $\mathrm{Int}(\Sigma)$. Gathering (2d)–(2f), we note that the resulting problem

$$\Delta w + k^2 w = 0, \quad x \in \mathrm{Ext}(\Sigma), \tag{3a}$$

$$w = u, \quad x \in \Sigma, \tag{3b}$$

$$\lim_{r\to\infty} \sqrt{r}(\partial_r w - \mathrm{i}kw) = 0 \tag{3c}$$

has a unique solution w for a given Dirichlet trace $u|_\Sigma$. We can thus use (3) to define a Dirichlet-to-Neumann (DtN) map $\mathcal{G}$, where $\mathcal{G} : H^s(\Sigma) \to H^{s-1}(\Sigma)$ is specified as $\mathcal{G}[u|_\Sigma] := \partial_N w|_\Sigma$. This map is often called the Steklov–Poincaré map, and includes information regarding the outgoing nature of the wave. We can use $\mathcal{G}$ to rewrite the original exterior problem (1) equivalently as

$$\Delta u + k^2 u = 0, \quad x \in \Omega, \tag{4a}$$

$$\partial_n u = g, \quad x \in \Gamma, \tag{4b}$$

$$\partial_N u - \mathcal{G}[u] = 0, \quad x \in \Sigma, \tag{4c}$$

where we will now write $\mathcal{G}[u|_\Sigma]$ simply as $\mathcal{G}[u]$.

If we can accurately compute $\mathscr{G}[u]$, solving problem (4) provides the solution of the original problem (1) in the near field in the sense that $v|_\Omega = u$. We can therefore solve the reduced problem (4) for u. To compute the solution of (1), v, in the infinite region $\mathbb{R}^2\backslash\overline{\mathrm{Int}(\Sigma)}$, we use the traces $u|_\Sigma$ and $\partial_n u|_\Sigma$ and the representation formula,

$$v(x) = \int_\Sigma u(y)\partial_{n_y} g(x, y) - g(x, y)\partial_{n_y} u(y)\,\mathrm{d}y, \quad x \in \mathbb{R}^2\backslash(\bar{D}).$$

Here $g(x, y) := (\mathrm{i}/4)H_0^{(1)}(k|x - y|)$ and $H_0^{(1)}$ is the zeroth Hankel function of the first kind.

Accurate and efficient computation of the DtN map is a non-trivial task, but several algorithms exist, see e.g., [15,10,8]. Other work includes reducing the exterior problem to the form (4) to ensure u is outgoing, but without actually computing $\mathscr{G}[u]$, e.g., [4,3,9]. Clearly, if the artificial boundary Σ is such that $\mathbb{R}^2\backslash\mathrm{Int}(\Sigma)$ is separable, we can compute $\mathscr{G}[u]$ using separation of variables. This idea was exploited, for example, in [13,16,12], where Σ is chosen to be a circle. In recent work, we extended this to quite general Σ which are *perturbations* of a circle [17]. This, of course, allows irregularly shaped obstacles to be enclosed more tightly by the artificial boundary. In fact, we showed that if $\mathscr{G}_0$ denotes the DtN map on a circle, and if $\Sigma := \{(r, \theta)|r = a + \delta f(\theta), \theta \in [0, 2\pi]\}$, then $\mathscr{G}$ is an analytic perturbation of $\mathscr{G}_0$. By this we mean that, for integer $s \geqslant 0$ and data $\mu \in H^{s+1/2}$,

$$\mathscr{G}[\mu] = \sum_{n=0}^{\infty} \mathscr{G}_n[\mu]\delta^n, \tag{5}$$

where the series converges strongly in the operator norm from $H^{s+3/2}([0, 2\pi])$ to $H^{s+1/2}([0, 2\pi])$, provided $f \in C^{s+2}([0, 2\pi])$. This result was recently extended in [18] to admit the strong convergence of (5) from $H^{1/2}([0, 2\pi])$ to $H^{-1/2}([0, 2\pi])$ which is of crucial importance in establishing the theorems of this paper. Of course, for the application at hand μ will be the Dirichlet trace of the field, ξ. In the Appendix we provide some details for one approach to the *numerical* simulation of these $\mathscr{G}_n$.

In what follows, we denote by $\mathscr{G}^N$ the operator

$$\mathscr{G}^N[\mu] := \sum_{n=0}^{N} \mathscr{G}_n[\mu]\delta^n,$$

obtained by truncating the series (5) after N terms. We shall show, in Section 3, that the weak form of problem (4) is well-posed when $\mathscr{G}$ is replaced by $\mathscr{G}^N$, and examine the error introduced due to this substitution. This is the key theoretical contribution of this paper.

Once one has an efficient and accurate boundary condition at the artificial boundary Σ, the annular region between the scatterer D and Σ can be discretized using, for example, a finite element method. Typically, the resulting system of linear equations will be solved iteratively, and needs to be preconditioned. In fact, since the exact DtN map is a non-local operator, we may expect the matrices involved in the linear system to have dense sub-matrices. At the discrete level this illustrates one of the major difficulties in prescribing these artificial boundary conditions: accurate boundary conditions are non-local and lead to dense systems which lose accuracy when sparsity is artificially enforced. Fortunately, a natural preconditioner exists for the linear system resulting from our method. Its performance is compared with that of some other possible choices in Section 4.

We note that the error analysis in the first part of the paper does not influence the preconditioning strategies presented in the second part. Indeed, we regard these two aspects of our algorithm as deeply important, complementing features: it is our goal to present an algorithm which is provably robust, as well as amenable to optimization in terms of implementation. We also note that our preconditioner of choice will be dependent on our specific choice of artificial boundary condition, and is not hence a generic "blackbox" preconditioner.

2. Variational formulation

Let $V := H^1(\Omega)$ with the associated norm denoted by $\|\cdot\|_V$, and denote by $(\cdot, \cdot)$ the inner product on $L^2(\Omega)$. The variational formulation of (4) is:

Find $u \in V$ such that for all $v \in V$,

$$\mathscr{A}(u, v) := \underbrace{(\nabla u, \nabla v) - k^2(u, v)}_{b(u,v)} + \langle \mathscr{G}[u], v\rangle = \int_{\partial D} g\bar{v}\, \mathrm{d}s. \tag{6}$$

In practice, we truncate the DtN map after N terms, and use $\mathscr{G}^N$. This leads to the variational problem:

Find $u_N \in V$ such that for all $v \in V$,

$$\mathscr{A}_N(u_N, v) := b(u_N, v) + \langle \mathscr{G}^N[u_N], v\rangle = \int_{\partial D} g\bar{v}\, \mathrm{d}s. \tag{7}$$

We focus on the consistency error introduced by using $\mathscr{G}^N$ instead of $\mathscr{G}$, since the approximation errors incurred due to using finite element approximations and truncations of Fourier series can be estimated by standard techniques (see [18]). We now observe, using the variational problems defined by Eqs. (6) and (7), that

$$\begin{aligned} |\mathscr{A}_N(u_N - u, v)| &\leqslant |\mathscr{A}_N(u_N, v) - \mathscr{A}(u, v)| + |\mathscr{A}(u, v) - \mathscr{A}_N(u, v)| \\ &= |\mathscr{A}(u, v) - \mathscr{A}_N(u, v)| = |\langle (\mathscr{G} - \mathscr{G}^N)[u], v\rangle|. \end{aligned} \tag{8}$$

In analogy to [7], our analysis will hinge on being able to demonstrate that the bilinear form $\mathscr{A}_N(u, v)$ satisfies an *inf-sup condition* (also called the Ladhyzhenskaya–Babŭska–Brezzi condition [2]), i.e., there exists a constant $\gamma_N > 0$ such that $\gamma_N \|u\|_V \leqslant \sup_{v\in V, v\neq 0}(a(u, v)/\|v\|_V)$. Note that this condition is equivalent to requiring $\inf_{u\in V, v\,0} \sup_{v\in V, v\neq 0}(a(u, v)/\|u\|_V\|v\|_V) \geqslant \gamma_N > 0$, and hence the constant γ_N is called the inf-sup constant. If we can show that (7) satisfies such an inf-sup condition, then estimate (8) will allow us to obtain a bound on $\|u - u_N\|_V$:

Theorem 2.1. *Let $u, u_N \in V$ be solutions of* (6) *and* (7), *respectively. If* $\lim_{N\to\infty} \gamma_N > 0$, *then the following estimate holds*:

$$\|u - u_N\|_V \leqslant C\frac{1}{\gamma_N}\|\mathscr{G} - \mathscr{G}^N\|_{L(H^\alpha, H^{-1/2})}\|u\|_{H^\alpha(\Sigma)},$$

where u has trace $u|_\Sigma \in H^\alpha$, $\alpha \geqslant \frac{1}{2}$.

Note that in [7] the authors prove an error estimate for their finite-infinite element formulation which relies on the discrete inf-sup constant being bounded away from zero. Numerical evidence is provided for this bound. Here we are able to analytically establish that the γ_N are bounded away from zero for N large enough.

3. Well-posedness of variational formulations and the inf-sup condition

It was shown in [14,7] that the variational problem

$$\langle\langle B_0 u, v\rangle\rangle_0 := \int_{\Omega_0} \nabla u \cdot \nabla v - k^2 uv\, \mathrm{d}V_0 + \int_{r=a} \mathscr{G}_0[u]v\, \mathrm{d}s = F(v), \quad \forall v \in H^1(\Omega_0)$$

is well-posed for $F \in (H^1(\Omega_0))'$. Here, Ω_0 is the annular region between ∂D and the circle $r = a$, and $\mathscr{G}_0$ is the DtN map associated with a circular artificial boundary. The result followed by showing that the linear operator $B_0 : H^1(\Omega_0) \to (H^1(\Omega_0))'$ is Fredholm, and that the variational problem has a unique solution in $H^1(\Omega_0)$. (We have denoted by $\langle\langle \cdot, \cdot\rangle\rangle_0$ the duality pairing between $H^1(\Omega_0)$ and its dual.) The proof relies on the spectral characterization of $\mathscr{G}_0$. In particular, if

$$\langle \mathscr{G}_0[\mu], v\rangle = \sum_{p=-\infty}^{\infty} a\lambda_p \hat{\mu}_p \bar{\hat{v}}_p, \quad \lambda_p = -k\frac{\mathrm{d}_z H_p^{(1)}(ka)}{H_p^{(1)}(ka)}, \tag{9}$$

where $\mathrm{d}_z H_p^{(1)}$ is the first derivative of $H_p^{(1)}$ with respect to its argument, then well-posedness follows by showing that the $\mathrm{Im}(\lambda_p) < 0$ are bounded, and that $\mathrm{Re}(\lambda_p) \geqslant 1/a > 0$.

We now sketch the proof of well-posedness of problems (6) and (7). We first show that the operator $S_0 : V \to (V)'$ defined via $\langle\langle S_0 u, v\rangle\rangle := \mathscr{A}_0(u, v)$ is Fredholm with index zero (Theorem 3.1). We then show that the operator $S : V \to V'$ defined by $\langle\langle Su, v\rangle\rangle := \mathscr{A}(u, v)$ satisfies $\|S - S_0\| \leqslant 1/\|S_0^{-1}\|$ for $\delta > 0$ small enough. Thus, by a standard perturbation argument, S is also a Fredholm operator (see, e.g., [1, Theorem 2.3.5]). In [18] it is shown that solutions of (6) are unique, implying the invertibility of S and the well-posedness of this formulation.

We use this argument again to show that S_N defined via $\langle\langle S_N u, v\rangle\rangle := \mathscr{A}_N(u, v)$ has a bounded inverse, since S_N is close in operator norm to S. Hence, the variational problem (7) is also well-posed for sufficiently small perturbation parameter δ. In particular, we can see that $\mathscr{A}_N$ satisfies the discrete inf-sup condition (see e.g., [5, Theorem 3.6]): there exists $\gamma_N > 0$ such that for all $v \in V$,

$$\gamma_N \|v\|_V \leqslant \sup_{w \in V \setminus \{0\}} \frac{\mathscr{A}_N(v, w)}{\|w\|_V}. \tag{10}$$

Proof of Theorem 2.1. In order to conclude the estimate of Theorem 2.1, we need to show that γ_N are strongly bounded away from zero. This follows by observing that, from the theoretical results of [17,18], $\mathscr{G}^N \to \mathscr{G}$ in the operator norm. Clearly, $\gamma_N > 0$ for all $N \geqslant N_0$. Now, for $v, w \in V$, we have

$$\mathscr{A}(v, w) = \mathscr{A}(v, w) - \mathscr{A}_N(v, w) + \mathscr{A}_N(v, w),$$

which implies

$$\frac{|\mathscr{A}(v, w)|}{\|w\|_V} \leqslant c\|(\mathscr{G} - \mathscr{G}^N)\|_{L(H^{1/2}, H^{-1/2})}\|v\|_V + \frac{|\mathscr{A}_N(v, w)|}{\|w\|_V},$$

where $c > 0$ is the trace constant. Thus,

$$\sup_{w \in V, w \neq 0} \frac{|\mathscr{A}(v, w)|}{\|w\|} - c\|(\mathscr{G} - \mathscr{G}^N)\|_{L(H^{1/2}, H^{-1/2})}\|v\|_V \leqslant \sup_{w \in V, w \neq 0} \frac{|\mathscr{A}_N(v, w)|}{\|w\|_V}.$$

If γ is the inf-sup constant for $\mathscr{A}(\cdot, \cdot)$, we can choose $N_1 \geqslant N_0$ such that for all $N > N_1$,

$$\gamma - c\|(\mathscr{G} - \mathscr{G}^N)\|_{L(H^{1/2}, H^{-1/2})} \geqslant \frac{\gamma}{2}.$$

Then, for all $N \geqslant N_1$,

$$\frac{\gamma}{2} \leqslant \inf_{v \in V, v \neq 0} \sup_{w \in V, w \neq 0} \frac{|\mathscr{A}_N(v, w)|}{\|w\|_V \|v\|_V}.$$

This shows that for $N > N_1$, $\gamma_N \geqslant \gamma/2 > 0$.

From (8),

$$|\mathscr{A}_N(u_N - u, v)| \leqslant |\langle(\mathscr{G} - \mathscr{G}^N)[u], v\rangle|,$$

and we can use the inf-sup constant γ_N for $\mathscr{A}_N$ and the trace constant c to obtain

$$\|u - u_N\|_V \leqslant \frac{c}{\gamma_N}\|\mathscr{G} - \mathscr{G}^N\|_{L(H^{1/2}, H^{-1/2})}\|u\|_V \leqslant \frac{2c}{\gamma}\|\mathscr{G} - \mathscr{G}^N\|_{L(H^{1/2}, H^{-1/2})}\|u\|_V.$$

Notice that the Taylor remainder of the DtN map *must* be measured in the weak space $H^{1/2}$ (established in [18]) rather than the smoother space $H^{s+3/2}$ ($s \geqslant 0$) studied in [17]. We also point out that there is a cost associated with this more delicate estimate, namely that the perturbation f must be somewhat smoother, $f \in H^5([0, 2\pi])$. □

The key to the preceding argument is that S_0 is a Fredholm operator of index zero, which is established in the following theorem.

Theorem 3.1. *If $f \in H^5([0, 2\pi])$, then there exists a $\delta_0 > 0$ such that if $0 < \delta < \delta_0$, then $S_0 = A + C$ where the linear operators A and C are, respectively, invertible and compact as maps from V to V'. Hence, S_0 is a Fredholm operator of index zero.*

Proof of Theorem 3.1. It is clear that $\mathscr{A}_0(u, v)$ is a continuous sesquilinear form on $V \times V$. We now define the sesquilinear forms a, d on $V \times V$:

$$a(u, v) := \int_\Omega \nabla u \cdot \nabla \bar{v} + u\bar{v}\,\mathrm{d}V + \mathrm{Re}\left\{\int_\Sigma \mathscr{G}_0[u]\bar{v}\,\mathrm{d}s\right\}, \tag{11a}$$

$$d(u, v) := -\int_\Omega (k^2 + 1)u\bar{v}\,\mathrm{d}V + \mathrm{Im}\left\{\int_\Sigma \mathscr{G}_0[u]\bar{v}\,\mathrm{d}s\right\}. \tag{11b}$$

Clearly $\mathscr{A}_0(u, v) = a(u, v) + d(u, v)$.

By inspection a is continuous; for coercivity we note that

$$a(u, u) = \|u\|_V^2 + \mathrm{Re}\left\{\int_\Sigma \mathscr{G}_0[u]\bar{u}\,\mathrm{d}s\right\}.$$

If we can show $\mathrm{Re}\{\int_\Sigma \mathscr{G}_0[\xi]\bar{\xi}\,\mathrm{d}s\} \geqslant 0$ for all $\xi \in H^{1/2}(\Sigma)$, the coercivity of $a(u, v)$ is established, since the trace of $u \in V$ lies in $H^{1/2}(\Sigma)$. We can describe the arc-length parameter $\mathrm{d}s$ on Σ as

$$(\mathrm{d}s)^2 = [(a + \delta f)^2 + (\delta f')^2]\,(\mathrm{d}\theta)^2,$$

and can thus estimate

$$\begin{aligned}
\mathrm{Re}\left\{\int_\Sigma \mathscr{G}_0[\xi]\bar{\xi}\,\mathrm{d}s\right\} &= \mathrm{Re}\left\{\int_0^{2\pi} \mathscr{G}_0[\xi]\bar{\xi}a\,\mathrm{d}\theta\right\} \\
&\quad + \mathrm{Re}\left\{\int_0^{2\pi} \mathscr{G}_0[\xi]\bar{\xi}([(a + \delta f)^2 + (\delta f')^2]^{1/2} - a)\,\mathrm{d}\theta\right\} \\
&\geqslant (1 - \bar{c}(\delta))\mathrm{Re}\left\{\int_0^{2\pi} \mathscr{G}_0[\xi]\bar{\xi}a\,\mathrm{d}\theta\right\} \\
&= (1 - \bar{c}(\delta))\sum_{p=-\infty}^{\infty} \mathrm{Re}\{\lambda_p\}|\hat{\xi}_p|^2 \geqslant 0,
\end{aligned} \tag{12}$$

for δ sufficiently small, where we used $\mathrm{Re}\{\lambda_p\} \geqslant 0$, from (9). Here

$$1 - \bar{c}(\delta) = 1 - \max_{[0,2\pi]} \left| \left[\left(1 + \delta \frac{f(\theta)}{a}\right)^2 + \delta^2 \frac{f'(\theta)^2}{a^2} \right]^{1/2} - 1 \right| \geqslant 0,$$

provided δ is chosen small enough. Denote by δ_0 the largest such perturbation. We can hence use the coercivity of $a(\cdot, \cdot)$ to define an invertible operator $A : V \to V'$:

$$\langle\langle Au, v\rangle\rangle := a(u, v), \quad \forall u, v \in V,$$

provided $0 \leqslant \delta \leqslant \delta_0$.

We now turn our attention to $d(u, v)$. The continuity of the first term in (11b) is clear, while the continuity of the second follows by the calculation

$$\mathrm{Im}\left\{\int_0^{2\pi} \mathscr{G}_0[\xi]\bar{\sigma}a\,\mathrm{d}\theta\right\} = \sum_{p=-\infty}^{\infty} \mathrm{Im}\left\{\lambda_p\right\}\hat{\xi}_p\bar{\hat{\sigma}}_p.$$

This defines a continuous sesquilinear map on $L^2([0, 2\pi]) \times L^2([0, 2\pi])$, since the $\mathrm{Im}\left\{\lambda_p\right\}$ are bounded for all p, see (9). Therefore, we can also bound $\mathrm{Im}\left\{\int_\Sigma \mathscr{G}_0[\xi]\bar{\sigma}\,\mathrm{d}s\right\}$ for δ smaller than δ_0. The embeddings of $H^1(\Omega)$ in $L^2(\Omega)$ and $H^{1/2}(\Sigma)$ in $L^2(\Sigma)$ are compact, and therefore the sesquilinear form $d(u, v)$ can be used to define a compact operator $C : V \to V'$. Consequently, the variational problem

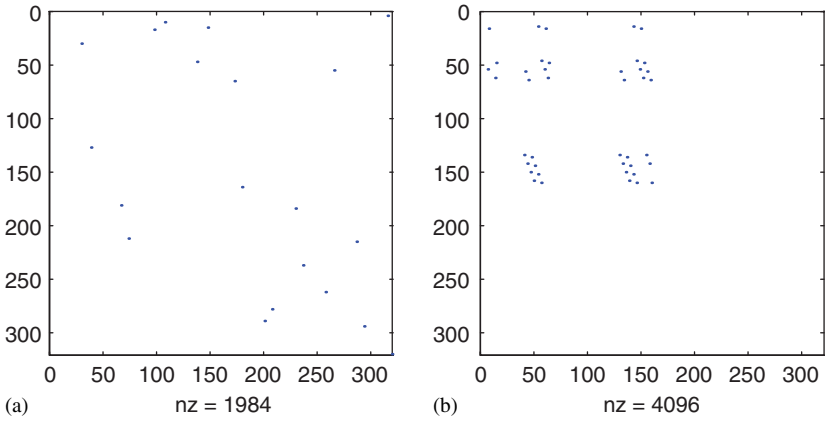

Fig. 2. Sparsity patterns for the matrices K, M and Q: (a) Sparsity pattern of K and M; (b) Starsity pattern of Q.

Find $u^0 \in V$ such that for all $v \in V$,

$$\mathscr{A}_0(u^0, v) = \int_\Gamma g\bar{v}\,\mathrm{d}s$$

can then be written in operator notation as $(A + C)u^0 = F$, for some $F \in V'$, or $(I + A^{-1}C)u^0 = A^{-1}F$, where $A^{-1}C$ is a compact map from $V \to V$. This proves the assertion that S_0 is a Fredholm operator of index zero. □

The uniqueness of solutions to the operator equation $S_0 u_0 = F$ can be proven using Rellich's lemma and an analytic continuation argument as in [14]. This, along with the preceding theorem, gives us the desired invertibility result for S_0.

4. Preconditioning

In this section we present some preliminary results concerning the preconditioning of the discrete system obtained from the variational formulation for u_N. We note that until now, the algorithm has been quite general: any Galerkin method (finite element or spectral element) would be suitable for use in the truncated annular region. For definiteness we consider a finite element approximation of the variational problem (7). To this end, we introduce a mesh T_h, and a finite-dimensional subspace V_h of V. Let $\{\phi_i\}_{i=1}^{N_h}$ be a basis for this subspace. If we approximate u_N by $u_{N,N_h} = \sum_{j=1}^{N_h} \tilde{u}_j \phi_j$, (7) leads to the system of equations:

$$\sum_{j=1}^{N_h} \tilde{u}_j \left[\int_\Omega \nabla\phi_j \cdot \nabla\phi_l - k^2\phi_j\phi_l \,\mathrm{d}V + \int_\Sigma \mathscr{G}^N[\phi_j]\phi_l \,\mathrm{d}s \right] = \int_{\partial D} g\phi_l \,\mathrm{d}s,$$

for $l = 1, \ldots, N_h$. We can write this in matrix notation as

$$L\vec{u} := (K - k^2 M + Q)\vec{u} = \vec{f}, \quad \vec{u} = (\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_{N_h})^{\mathrm{T}}. \tag{13}$$

In order to solve (13) for large N_h, it is crucial to understand the sparsity patterns of the matrices K, M and Q, which depend on the discretization of the computational domain. Fig. 2 gives a sample of these patterns for the matrices K, M and Q.

The number of non-zero entries in K and M will be at most $(d_{\max} + 1)N_h$, where $d_{\max}$ is the largest degree of a vertex in the discretization, and typically remains bounded during successive refinements. Thus, K and M are sparse. On the other hand, if ϕ_j is supported on the boundary Σ, $\mathscr{G}^N[\phi_j]$ will be non-zero along the *entire* boundary, so Q will have dense sub-matrices (see Fig. 2). However, the number of non-zero entries will be at most N_Σ^2, where N_Σ is the number of vertices lying on the boundary. Since $N_\Sigma = \mathrm{o}(N_h)$, the fraction of the entries of Q that are non-zero eventually becomes small, making Q sparse as well. Direct methods, e.g., ones utilizing the LU or QR factorizations,

Table 1
Condition numbers for different preconditioners and discretizations

$h_{\max}$	L	$P_A^{-1}L$	$P_{B_0}^{-1}L$	$P_{B_1}^{-1}L$	$P_{B_2}^{-1}L$	$P_C^{-1}L$	$P_D^{-1}L$	$P_E^{-1}L$
0.3791	121.0	32.3	31.4	8.63	7.15	554.8	3.09	223.5
0.1911	544.5	114.3	84.6	22.6	3.6	946.5	6.44	613.6
0.0991	2887	390.4	264.9	60.4	7.5	2784	15.3	1936

can destroy this sparsity [19], and are therefore not ideal for systems of the form (13). For this reason, iterative methods are favored as they retain and take advantage of the sparsity of the system [11]. In addition, for large-scale problems, their execution time and floating-point operation (FLOP) counts can grow much more slowly as a function of the size of the system when compared to direct methods.

Of course, in practice an iterative method requires an easily computed preconditioner P which transforms the original problem $Lx = b$ into $(P^{-1}L)x = P^{-1}b$, where $P^{-1}L$ approximates the identity in some sense. We now consider a class of preconditioners P defined by

$$P = K - k^2 M + \hat{Q},$$

where $\hat{Q}$ is an "easily computed" approximation to Q.

One may expect that one simple preconditioner may be $P = (K - k^2 M)$; however, as the mesh is refined, we found that the conditioning of P deteriorates quite dramatically. The next natural preconditioner one may conceive of is $P = K - k^2 M + Q^0$, where we approximate $Q_{jl} = (\int_\Sigma \mathscr{G}^N[\phi_j]\phi_l \, \mathrm{d}s)$ by $(Q^0)_{jl} := (\int_\Sigma \mathscr{G}^0[\phi_j]\phi_l \, \mathrm{d}s)$. We will see that this choice indeed, at least experimentally, is the most successful in terms of computational efficiency and performance.

The following five options were explored:

(A) A far-field approximation: $\hat{Q}^{\mathrm{FF}}_{j,l} = (\mathrm{i}k) \int_\Sigma \phi_j \phi_l \, \mathrm{d}s$.
(B) A low-order approximation to Q: $\hat{Q}^{m}_{j,l} = \int_\Sigma \mathscr{G}^m[\phi_j]\phi_l \, \mathrm{d}s$.
(C) Hermitian approximation: $\hat{Q}^{\mathrm{H}} = (\frac{1}{2})(Q + Q^{\mathrm{H}})$.
(D) Symmetric approximation: $\hat{Q}^{\mathrm{S}} = (\frac{1}{2})(Q + Q^{\mathrm{T}})$.
(E) A low-wavenumber approximation: $\hat{Q}^{\mathrm{LW}}_{j,l} = \int_\Sigma \mathscr{G}^N[\phi_j]|_{k=1}\phi_l \, \mathrm{d}s$.

We now present the results of a series of numerical experiments investigating the condition numbers, $\kappa(L) := |L||L^{-1}|$, of these five classes of preconditioners, and their behavior when combined with the GMRES and BiCGStab iterative methods. In all experiments, the wavenumber was set to $k = 3$, and the geometry of the scatterer and artificial boundary are $r = 1 + 0.4 f(\theta)$ and $r = 1.5 + 0.3 f(\theta)$, respectively, where $f(\theta) = \cos(4\theta)$. Approximations to f and the Dirichlet data on Σ, $\xi := u|_\Sigma$, at equally spaced gridpoints are stored in vectors of length $N_\theta = 256$, while the number of Fourier modes retained in approximating f and ξ are $N_f = 4$ and $N_\xi = 8$, respectively. See the Appendix for formulas for the $\mathscr{G}_n$ and [17] for complete details on their numerical implementation. In our simulation of the matrix L, the Taylor series approximation to $\mathscr{G}$ is truncated after $N = 10$ terms, while the tolerance for the iterative methods was set to $\varepsilon = 10^{-12}$. For a complete discussion of these numerical parameters see [17].

In Table 1 we report numerical simulations of the condition numbers of the matrix L (where $\mathscr{G}$ is approximated by $\mathscr{G}^{10}$) and the preconditioned matrices $P^{-1}L$ for the five approximations of $\hat{Q}$ appearing in P. It is clear that the condition number of L is inversely quadratic as a function of $h_{\max}$, as predicted by the standard finite element theory [11]. We do point out that this behavior is unaffected by the value N in $\mathscr{G}^N$ for N sufficiently large (say $N \geqslant 10$). The condition number of the preconditioned system is reduced most significantly when alternative (B) is used. Furthermore, the condition number of $P_{B_j}^{-1}L$ decreases as j increases. This is to be expected as the eigenvalue distribution of $\hat{Q}^j$ should converge to that of Q as $j \to \infty$ when all other parameters are fixed.

In Table 2 we list numerical approximations to the condition numbers and "deviation from normality" for preconditioners of type (B). This non-normal behavior is inherent at the continuous level as well, and is unavoidable. The deviation from normality at the discrete level is measured using the condition number of $V(M)$, the matrix whose rows

Table 2
Condition numbers and deviation from normality for preconditioners of type B. Here, $V(M)$ is a matrix holding the eigenvectors in the diagonalization $M = VDV^{-1}$

j	$\kappa(L_j)$	$\kappa(V(L_j))$	$\kappa(P_{B_j}^{-1}L)$	$\kappa(V(P_{B_j}^{-1}L))$	Build Time (s)
0	720	245	84.6	7.8×10^{17}	9.23
1	567	246	22.6	1.5×10^{16}	13.35
2	543	255	3.6	1.7×10^{16}	18.69
3	540	254	2.06	4.6×10^{14}	24.95
4	545	251	1.57	3.7×10^{15}	31.89
5	408	237	1.38	1.8×10^{16}	40.03
6	544	231	1.20	4.6×10^{14}	50.83
7	544	256	1.19	9.5×10^{16}	60.97
8	544	245	1.16	1.2×10^{16}	72.29
9	544	252	1.11	2.0×10^{14}	84.74

Table 3
Number of iterations of GMRES (10) required by different preconditioners, and the time required to build the preconditioner

Preconditioner	$h_{\max} = 0.191$		$h_{\max} = 0.0991$		$h_{\max} = 0.0505$	
	Time (s)	Iter.	Time (s)	Iter.	Time (s)	Iter.
None	99.2	Fail	243.8	Fail	643.8	Fail
P_A	0.01	19	0.01	19	0.01	19
P_{B_0}	8.1	9	26.6	9	88.7	9
P_{B_1}	12.9	9	38.9	8	126.1	8
P_{B_2}	18.7	7	54.9	6	174.2	6
P_E	95.0	45	228.2	47	632.5	47

Table 4
Number of iterations of BiCStab required by different preconditioners, and the time required to build the preconditioner

Preconditioner	$h_{\max} = 0.191$		$h_{\max} = 0.0991$		$h_{\max} = 0.0505$	
	Time (s)	Iter.	Time (s)	Iter.	Time (s)	Iter.
None	98.5	Fail	246.5	Fail	649.5	Fail
P_A	0.01	Fail	0.01	Fail	0.01	Fail
P_{B_0}	8.2	11	26.6	9	90.0	10
P_{B_1}	12.8	7	39.8	5.5	126.9	7
P_{B_2}	18.6	4.5	55.6	4.5	176.1	4
P_E	95.5	Fail	237.3	Fail	633.8	Fail

contain the eigenvectors of the matrix M. The matrices L_j are approximations to L where $\mathcal{G}$ is approximated by $\mathcal{G}^j$. From these data we deduce that the matrices $P_B^{-1}L$ are highly non-normal, as indicated by the high condition number of their V-matrices in the VDV^{-1}-factorization. On the one hand, this is unfortunate, as error bounds on iterative methods applied to non-normal matrices (such as GMRES) often involve the factor $\kappa(V)$ [11]. On the other hand, such error bounds are seldom tight, and, in fact, our experiments indicate that this is probably the case here.

Regarding alternatives (C) and (D), the cost of forming these preconditioners is as great as that of forming Q itself, rendering them disadvantaged when compared with the other options. However, it should be pointed out that (D) does provide a great reduction in the condition number. As the data in Tables 3 and 4 and Fig. 3 clearly indicate, methods (A)

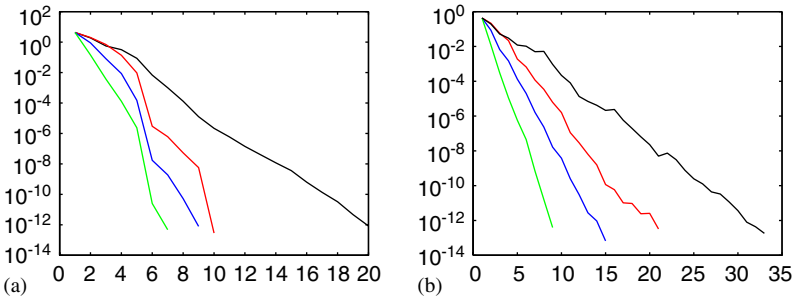

Fig. 3. Semilog plot of error versus the number of iterations for different methods. The black, red, blue and green curves represent methods (A), B_0, B_1 and B_2, respectively: (a) GMRFS (10); (b) BICGStub.

and (E) are inferior to approach (B). Regarding choice (B), as j is increased, the number of iterations required to meet the tolerance ε consistently decreases. However, when the time required to build $\hat{Q}^j$ is factored in, the *total* execution time is actually smallest for $\hat{Q}^0$.

5. Conclusion

We have presented an error analysis for a high-order accurate DtN-FE algorithm for solving two-dimensional exterior scattering problems. This method involves a perturbative algorithm for the enforcement of an *exact* transparent boundary condition at a quite general artificial boundary. Having completed this analysis we described five preconditioning methodologies for a finite element implementation of our algorithm. One method, based upon low-order approximation of the DtN map, stood out as clearly superior for both GMRES and BiCGStab iterative schemes. Future work will include a theoretical justification of our preconditioning strategy, as well as a more efficient way to invert the preconditioner.

Acknowledgments

L.C. gratefully acknowledges summer support from NSERC through a USRA award. D.P.N. gratefully acknowledges support from NSF through Grant no. DMS-0406007. N.N. gratefully acknowledges support from NSERC and the FQRNT.

Appendix. Perturbative calculation of the DtN map

We gather in this section the specific expressions used for the terms $\mathscr{G}_n$ in the perturbation expansion of the DtN map, $\mathscr{G}$, using the method of field expansions [17]. Recall that if the Dirichlet data ξ are given on Σ, the DtN operator $\mathscr{G}$ is defined via $\mathscr{G}[\xi] := \nabla w|_\Sigma \cdot N$, where w solves

$$\Delta w + k^2 w = 0, \quad x \in \mathrm{Ext}(\Sigma) \tag{14a}$$

$$w = \xi, \quad x \in \Sigma \tag{14b}$$

$$\lim_{r\to\infty} \sqrt{r}(\partial_r w - \mathrm{i}kw) = 0. \tag{14c}$$

The method of field expansions is based upon the fact that both the DtN map $\mathscr{G}$ and the solution w are analytic with respect to boundary variations, parameterized by δ. We can therefore expand the field in a perturbation series in δ:

$$w(r, \theta, \delta) = \sum_{n=0}^{\infty} w_n(r, \theta)\delta^n. \tag{15}$$

We then insert this into the defining Helmholtz problem (14), and obtain the PDE satisfied by w_n:

$$\Delta w_n(r,\theta) + k^2 w_n(r,\theta) = 0, \quad r > a, \tag{16a}$$

$$w_n(a,\theta) = \delta_{n,0}\xi(\theta) - \sum_{l=0}^{n-1} \partial_r^{n-l} w_l(a,\theta)\frac{f^{n-l}}{(n-l)!}, \tag{16b}$$

$$\lim_{r\to\infty} r^{1/2}(\partial_r w_n - \mathrm{i}k w_n) = 0, \tag{16c}$$

where $\delta_{n,p}$ is the Kronecker delta. Noting that

$$w_n(r,\theta) = \sum_{p=-\infty}^{\infty} a_{n,p} H_p^{(1)}(kr)\mathrm{e}^{\mathrm{i}p\theta} \tag{17}$$

satisfies (14a) and (14c), we can now compute the nth term in the expansion of the DNO. After some algebra, we get

$$\begin{aligned}\mathscr{G}(\delta f)\xi &= \sum_{n=0}^{\infty} \mathscr{G}_n(f)[\xi]\delta^n = \nabla w(a+\delta f(\theta),\theta)\cdot N_{\delta f}\\ &= \sum_{n=0}^{\infty}\sum_{p=-\infty}^{\infty}\left[-k(a+\delta f)d_z H_p^{(1)}(k(a+\delta f)) + \frac{\delta\partial_\theta f}{(a+\delta f)}(\mathrm{i}p)H_p^{(1)}(k(a+\delta f))\right]a_{n,p}\mathrm{e}^{\mathrm{i}p\theta}\delta^n,\end{aligned}$$

where $N_{\delta f} = (-(a+\delta f), \delta\partial_\theta f)$. This readily yields the following recursion for the field expansion $\mathscr{G}_n(f)$:

$$\begin{aligned}\mathscr{G}_n(f)\xi = &-ka\sum_{l=0}^{n}\sum_{p=-\infty}^{\infty} a_{l,p}\frac{(kf)^{n-l}}{(n-l)!}d_z^{n+1-l}H_p^{(1)}(ka)\mathrm{e}^{\mathrm{i}p\theta}\\ &-\frac{f}{a}\mathscr{G}_{n-1}(f)\xi\\ &-2kf\sum_{l=0}^{n-1}\sum_{p=-\infty}^{\infty} a_{l,p}\frac{(kf)^{n-1-l}}{(n-1-l)!}d_z^{n-l}H_p^{(1)}(ka)\mathrm{e}^{\mathrm{i}p\theta}\\ &-\frac{k}{a}f^2\sum_{l=0}^{n-2}\sum_{p=-\infty}^{\infty} a_{l,p}\frac{(kf)^{n-2-l}}{(n-2-l)!}d_z^{n-1-l}H_p^{(1)}(ka)\mathrm{e}^{\mathrm{i}p\theta}\\ &+\frac{1}{a}(\partial_\theta f)\sum_{l=0}^{n-1}\sum_{p=-\infty}^{\infty} a_{l,p}\frac{(kf)^{n-1-l}}{(n-1-l)!}d_z^{n-1-l}H_p^{(1)}(ka)(\mathrm{i}p)\mathrm{e}^{\mathrm{i}p\theta}.\end{aligned} \tag{18}$$

References

[1] K. Atkinson, W. Han, Theoretical numerical analysis, Texts in Applied Mathematics, vol. 39, Springer, New York, 2001(a functional analysis framework).

[2] I. Babŭska, Error-bounds for the finite element method, Numer. Math. 16 (1971) 322–333.

[3] A. Bayliss, M. Gunzburger, E. Turkel, Boundary conditions for the numerical solution of elliptic equations in exterior regions, SIAM J. Appl. Math. 42 (2) (1982) 430–451.

[4] J.-P. Berenger, A perfectly matched layer for the absorption of electromagnetic waves, J. Comput. Phys. 114 (2) (1994) 185–200.

[5] D. Braess, Finite Elements, second ed., Cambridge University Press, Cambridge, 2001, Theory, Fast Solvers, and Applications in Solid Mechanics (L.L. Schumaker, Trans., from the 1992 German edition).

[6] D. Colton, R. Kress, Inverse Acoustic and Electromagnetic Scattering Theory, second ed., Springer, Berlin, 1998.

[7] L. Demkowicz, F. Ihlenburg, Analysis of a coupled finite-infinite element method for exterior Helmholtz problems, Numer. Math. 88 (1) (2001) 43–73.

[8] R. Djellouli, C. Farhat, A. Macedo, R. Tezaur, Finite element solution of two-dimensional acoustic scattering problems using arbitrarily shaped convex artificial boundaries, J. Comput. Acoust. 8 (1) (2000) 81–99 Finite Elements for Wave Problems, Trieste, 1999.

[9] B. Engquist, A. Majda, Absorbing boundary conditions for the numerical simulation of waves, Math. Comput. 31 (139) (1977) 629–651.
[10] K. Feng, Finite element method and natural boundary reduction, in: Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Warsaw, 1983), Warsaw, 1984, pp. 1439–1453 (PWN).
[11] A. Greenbaum, Iterative Methods for Solving Linear Systems, SIAM, Philadelphia, PA, 1997.
[12] M.J. Grote, J.B. Keller, Nonreflecting boundary conditions for Maxwell's equations, J. Comput. Phys. 139 (2) (1998) 327–342.
[13] H.D. Han, X.N. Wu, Approximation of infinite boundary condition and its application to finite element methods, J. Comput. Math. 3 (2) (1985) 179–192.
[14] I. Harari, T.J.R. Hughes, Analysis of continuous formulations underlying the computation of time-harmonic acoustics in exterior domains, Comput. Methods Appl. Mech. Eng. 97 (1) (1992) 103–124.
[15] C. Johnson, J.-C. Nédélec, On the coupling of boundary integral and finite element methods, Math. Comput. 35 (152) (1980) 1063–1079.
[16] J.B. Keller, D. Givoli, Exact nonreflecting boundary conditions, J. Comput. Phys. 82 (1) (1989) 172–192.
[17] D.P. Nicholls, N. Nigam, Exact non-reflecting boundary conditions on general domains, J. Comput. Phys. 194 (1) (2004) 278–303.
[18] D.P. Nicholls, N. Nigam, Error analysis of a coupled finite element/DtN map algorithm on general domains, 2005, submitted for publication.
[19] L.N. Trefethen, D. Bau, Numerical Linear Algebra, SIAM, Philadelphia, PA, 1997.