**Noname manuscript No.**
(will be inserted by the editor)

# Applying GMRES to the Helmholtz equation with shifted Laplacian preconditioning: What is the largest shift for which wavenumber-independent convergence is guaranteed?

**M. J. Gander · I. G. Graham · E. A. Spence**

December 23, 2014

**Abstract** There has been much recent research on preconditioning discretisations of the Helmholtz operator $\Delta + k^2$ (subject to suitable boundary conditions) using a discrete version of the so-called "shifted Laplacian" $\Delta + (k^2 + \mathrm{i}\varepsilon)$ for some $\varepsilon > 0$. This is motivated by the fact that, as $\varepsilon$ increases, the shifted problem becomes easier to solve iteratively. Despite many numerical investigations, there has been no rigorous analysis of how to chose the shift. In this paper, we focus on the question of how large $\varepsilon$ can be so that the shifted problem provides a preconditioner that leads to $k$-independent convergence of GMRES, and our main result is a sufficient condition on $\varepsilon$ for this property to hold. This result holds for finite element discretisations of both the interior impedance problem and the sound-soft scattering problem (with the radiation condition in the latter problem imposed as a far-field impedance boundary condition). Note that we do not address the important question of how large $\varepsilon$ should be so that the preconditioner can easily be inverted by standard iterative methods.

**Keywords:** Helmholtz equation, iterative method, preconditioning, high frequency, shifted Laplacian preconditioner, GMRES.

## 1 Introduction

The Helmholtz equation is the simplest possible model of wave propagation. Although most applications are concerned with the propagation of waves in exterior domains, it is common to use as a model problem the Helmholtz equation posed in an interior domain with an impedance boundary condition, i.e.

$$\Delta u + k^2 u = -f \quad \text{in } \Omega, \tag{1.1a}$$
$$\partial_n u - \mathrm{i}ku = g \quad \text{on } \Gamma, \tag{1.1b}$$

where $\Omega$ is a bounded Lipschitz domain in $\mathbb{R}^d$ ($d = 2$ or 3) with boundary $\Gamma$, and $f$ and $g$ are prescribed functions. This paper is predominately concerned with the *interior impedance problem* (1.1), but we also consider the exterior Dirichlet problem, with the radiation condition realised as an impedance boundary condition (i.e. a first-order absorbing boundary condition).

The Helmholtz equation is difficult to solve numerically for the following two reasons:

M. J. Gander
Section de Mathématiques, Université de Genève, 2-4 rue du Livre, CP 64, CH-1211 Genève, Switzerland,
E-mail: Martin.Gander@unige.ch

I. G. Graham
Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK,
E-mail: I.G.Graham@bath.ac.uk

E.A. Spence
Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK,
E-mail: E.A.Spence@bath.ac.uk

1. The solutions of the homogeneous Helmholtz equation oscillate on a scale of $1/k$, and so to approximate them accurately one needs the total number of degrees of freedom, $N$, to be proportional to $k^d$ as $k$ increases. Furthermore, the *pollution effect* means that in some cases (e.g. for low-order finite element methods) having $N \sim k^d$ is still not enough to keep the relative error bounded independently of $k$ as $k$ increases. This growth of $N$ with $k$ leads to very large matrices, and hence to large (and sometimes intractable) computational costs.
2. The standard variational formulation of the Helmholtz equation is sign-indefinite (i.e. not coercive). This means that (i) it is hard to prove error estimates for the Galerkin method that are explicit in $k$, and (ii) it is hard to prove anything a priori about how iterative methods behave when solving the Galerkin linear system; indeed, one expects iterative methods to behave extremely badly if the indefinite system is not preconditioned.

Quite a lot of recent research has focused on preconditioning (1.1) using the discretisation of the original Helmholtz problem with a complex shift:

$$\Delta u + (k^2 + \mathrm{i}\varepsilon)u = -f \quad \text{in } \Omega, \tag{1.2a}$$

$$\partial_n u - \mathrm{i}\eta u = g \quad \text{on } \Gamma. \tag{1.2b}$$

The parameter $\eta$ is usually chosen to be either $k$ or $\sqrt{k^2 + \mathrm{i}\varepsilon}$, and the analysis in this paper covers both these choices. It is well-known that, with $k$ fixed, the solution of (1.2) tends to the solution of (1.1) as $\varepsilon \to 0$; this is called the "principle of limited absorption". When used as a preconditioner for (1.1), the problem (1.2) is usually called the "shifted Laplacian preconditioner" (even though the shift is added to the Helmholtz operator itself).

In some ways it is more natural to consider adding absorption to the problem (1.1) by letting $k \mapsto k + \mathrm{i}\delta$ for some $\delta > 0$ (with $\eta$ then usually chosen as either $k$ or $k + \mathrm{i}\delta$). The results in this paper are equally applicable to this preconditioner, however we consider absorption in the form of (1.2) since this form seems to be more prevalent in the literature.

The question then arises, how should one choose the "absorption" (or "shift") parameter $\varepsilon$? In this paper we investigate this question when (1.1) is solved using finite element methods (FEMs) of fixed order.

One of the advantages of the shifted Laplacian preconditioner is that it can be applied when the wavenumber $k$ is variable (i.e. the medium being modelled is inhomogeneous) as was done, for example, in [15], [42], and [54] (with the last paper considering the higher-order case). In the present paper, however, all the theory is for constant $k$ (although Example 5.5 contains an experiment where $k$ is variable).

Recall that the standard variational formulation of (1.2) (for any $\varepsilon \geq 0$) is, given $f \in L^2(\Omega)$, $g \in L^2(\Gamma)$, $\eta > 0$, and $k > 0$,

$$\text{find } u \in H^1(\Omega) \text{ such that} \quad a_\varepsilon(u, v) = F(v) \quad \text{for all } v \in H^1(\Omega), \tag{1.3}$$

where

$$a_\varepsilon(u, v) := \int_\Omega \nabla u \cdot \overline{\nabla v} - (k^2 + \mathrm{i}\varepsilon) \int_\Omega u\overline{v} - \mathrm{i}\eta \int_\Gamma u\overline{v}. \tag{1.4}$$

and

$$F(v) := \int_\Omega f\overline{v} + \int_\Gamma g\overline{v}. \tag{1.5}$$

The original Helmholtz problem that we are interested in solving, (1.1), is therefore (1.3) when $\varepsilon = 0$ and $\eta = k$, and in this case we write $a(u, v)$ instead of $a_\varepsilon(u, v)$.

If $V_N$ is an $N$-dimensional subspace of $H^1(\Omega)$ with basis $\{\phi_i : i = 1, \ldots, N\}$ then the corresponding Galerkin approximation of (1.3) is:

$$\text{find } u_N \in V_N \text{ such that} \quad a_\varepsilon(u_N, v_N) = F(v_N) \quad \text{for all } v_N \in V_N. \tag{1.6}$$

The Galerkin equations (1.6) are equivalent to the $N$-dimensional linear system

$$\mathbf{A}_\varepsilon \mathbf{u} = \mathbf{f}, \quad \text{with} \quad \mathbf{A}_\varepsilon = \mathbf{S} - (k^2 + \mathrm{i}\varepsilon)\mathbf{M} - \mathrm{i}\eta\mathbf{N}, \tag{1.7}$$

where $\mathbf{S}_{\ell,m} = \int_\Omega \nabla\phi_\ell \cdot \nabla\phi_m$ is the stiffness matrix, $\mathbf{M}_{\ell,m} = \int_\Omega \phi_\ell\phi_m$ is the domain mass matrix, and $\mathbf{N}_{\ell,m} = \int_\Gamma \phi_\ell\phi_m$ is the boundary mass matrix. When $\varepsilon = 0$ and $\eta = k$, (1.7) is the discretisation

of the original problem (1.1), in which case we write $\mathbf{A}$ instead of $\mathbf{A}_\varepsilon$ in (1.7). Note that $\mathbf{A}_\varepsilon$ and $\mathbf{A}$ are both symmetric but not Hermitian.

The "shifted Laplacian preconditioner" (applied in left-preconditioning mode) replaces the solution of $\mathbf{A}\mathbf{u} = \mathbf{f}$ with the solution of:

$$\mathbf{A}_\varepsilon^{-1}\mathbf{A}\mathbf{u} = \mathbf{A}_\varepsilon^{-1}\mathbf{f}. \tag{1.8}$$

GMRES works well applied to this problem if $\|\mathbf{I} - \mathbf{A}_\varepsilon^{-1}\mathbf{A}\|_2$ is sufficiently small (and this can be quantified by the Elman estimate recalled in Theorem 1.8 and Corollary 1.9 below).

In practice, $\mathbf{A}_\varepsilon^{-1}$ in (1.8) is replaced with an approximation that is easy to compute (e.g. a multigrid V-cycle). Then, letting $\mathbf{B}_\varepsilon^{-1}$ denote an approximation of $\mathbf{A}_\varepsilon^{-1}$, we replace (1.8) with

$$\mathbf{B}_\varepsilon^{-1}\mathbf{A}\mathbf{u} = \mathbf{B}_\varepsilon^{-1}\mathbf{f}. \tag{1.9}$$

Writing

$$\mathbf{I} - \mathbf{B}_\varepsilon^{-1}\mathbf{A} \;=\; \mathbf{I} - \mathbf{B}_\varepsilon^{-1}\mathbf{A}_\varepsilon \;+\; \mathbf{B}_\varepsilon^{-1}\mathbf{A}_\varepsilon(\mathbf{I} - \mathbf{A}_\varepsilon^{-1}\mathbf{A}), \tag{1.10}$$

we see that a sufficient condition for GMRES to converge in a $k$–independent number of steps is that both $\|\mathbf{I} - \mathbf{A}_\varepsilon^{-1}\mathbf{A}\|_2$ and $\|\mathbf{I} - \mathbf{B}_\varepsilon^{-1}\mathbf{A}_\varepsilon\|_2$ are sufficiently small. We write these two conditions as

(P1) $\mathbf{A}_\varepsilon^{-1}$ is a good preconditioner for $\mathbf{A}$

and

(P2) $\mathbf{B}_\varepsilon^{-1}$ is a good preconditioner for $\mathbf{A}_\varepsilon$.

In other words, the task is to find $\varepsilon$ and $\mathbf{B}_\varepsilon$ so that both properties (P1) and (P2) are satisfied. At this stage, one might already guess that achieving both (P1) and (P2) imposes somewhat contradictory requirements on $\varepsilon$. Indeed, on the one hand, (P1) requires $\varepsilon$ to be sufficiently small (since the ideal preconditioner for $\mathbf{A}$ is $\mathbf{A}^{-1}$, which is $\mathbf{A}_0^{-1}$). On the other hand, the larger $\varepsilon$ is, the less oscillatory the shifted problem is, and the cheaper it will be to construct a good approximation to $\mathbf{A}_\varepsilon^{-1}$ in (P2). These issues have been explored numerically in the literature (see the discussion in §1.1 below), however there are no rigorous results about how to achieve either (P1) or (P2), and hence no theory about the best choice of $\varepsilon$.

In this paper we perform the first step in this analysis by describing rigorously how large one can choose $\varepsilon$ so that (P1) still holds. These results can then be used in conjunction with results concerning (P2) to answer the question of how to choose $\varepsilon$ in (1.9). Indeed, the question of how one should choose $\varepsilon$ for (P2) to hold when $\mathbf{B}_\varepsilon^{-1}$ is constructed using multigrid is considered in the recent preprint [7]. Furthermore, in a subsequent paper [21] we will describe for a class of domain decomposition preconditioners how $\varepsilon$ should be chosen for these so that (P2) holds.

It could be argued that splitting the question of how to choose $\varepsilon$ in (1.9) into (P1) and (P2) is somewhat artificial from a practical point of view. However, it is difficult to see how any rigorous numerical analysis of this question can proceed without this split.

We also mention here that although the discussion above was presented in the context of left preconditioning, it applies equally well to right preconditioning and the main results (Theorems 1.4 and 1.5) are for both approaches.

Before outlining the main results of this paper, we review the literature on the shifted Laplacian preconditioner, focusing on the choices of $\varepsilon$ proposed, and whether these choices are aimed at achieving (P1) or (P2). (Although not all of this previous work concerns finite-element discretisations of the Helmholtz equation, in the discussion below we still use $\mathbf{A}$ to denote the discretisation of the (unshifted) Helmholtz problem, and $\mathbf{A}_\varepsilon^{-1}$ to denote the preconditioner arising from the shifted Helmholtz problem.)

### 1.1 Previous work on the shifted Laplacian preconditioner

Preconditioning the Helmholtz operator with the inverse of the Laplacian was proposed in [2], and preconditioning with $(\Delta - k^2)^{-1}$ was proposed in [32].

Preconditioning the Helmholtz operator with $(\Delta + \mathrm{i}\varepsilon)^{-1}$ was considered in [16] and [17], and then preconditioning with $(\Delta + k^2 + \mathrm{i}\varepsilon)^{-1}$ was considered in [15] and [55]. For both preconditioners, the

authors chose $\varepsilon \sim k^2$, and constructed an approximation to the discrete counterpart of $(\Delta+\mathrm{i}\varepsilon)^{-1}$ or $(\Delta+k^2+\mathrm{i}\varepsilon)^{-1}$ using multigrid. (Using the notation above, preconditioning with the second operator corresponds to choosing $\varepsilon \sim k^2$ and constructing $\mathbf{B}_\varepsilon^{-1}$ using a multigrid V-cycle.) Preconditioning with $(\Delta + k^2 + \mathrm{i}\varepsilon)^{-1}$ and $\varepsilon \sim k^2$ was then further investigated in the context of multigrid in [8] and [49].

The choice $\varepsilon \sim k^2$ was motivated by analysis of the 1-d Helmholtz equation in an interval with Dirichlet boundary conditions in [16, §5], [14, §5.1.2], [15, §3], with this analysis using the fact that in this situation the eigenvalues of the Laplacian are known explicitly. The investigations in [16, §5] and [14, §5.1.2] considered preconditioning the 1-d Helmholtz operator with $(\mathrm{d}^2/\mathrm{d}x^2 + k^2(a+\mathrm{i}b))^{-1}$, and found that, under the restriction that $a \leq 0$, $|\lambda_{\max}|/|\lambda_{\min}|$ was minimised for the operator $(\mathrm{d}^2/\mathrm{d}x^2 + k^2(a+\mathrm{i}b))^{-1}(\mathrm{d}^2/\mathrm{d}x^2 + k^2)$ when $a = 0$ and $b = \pm 1$. The eigenvalues of $(\mathrm{d}^2/\mathrm{d}x^2 + k^2(a+\mathrm{i}b))^{-1}(\mathrm{d}^2/\mathrm{d}x^2 + k^2)$ for this boundary value problem were plotted in [15, §3], and it was found that they were better clustered for $a = 1$ and several choices of $b \sim 1$ than for $a = 0$ and $b = 1$. (This eigenvalue clustering can be seen as partially achieving (P1) at the continuous level).

A more general eigenvalue-analysis was conducted in [55], with this investigation considering a general class of Helmholtz problems (including the interior impedance problem in 2- and 3-d). This investigation hinged on the fact that the field of values of many Helmholtz problems is contained within a closed half-plane (and thus the eigenvalues are also in this closed half-plane). One can see this for the interior impedance problem by noting from (1.4) that, since $\eta \in \mathbb{R}$,

$$\Im a_0(v,v) \leq 0 \quad \text{for all } v \in H^1(\Omega).$$

The investigation in [55] uses the bound on the number of GMRES iterations that (i) assumes that the eigenvalues are enclosed by a circle not containing the origin, and (ii) involves the condition number of the matrix of eigenvectors (see, e.g., [45, Theorem 5], [44, Corollary 6.33]). Because of (i), [55] needs to assume that the wavenumber has a small imaginary part (to prevent the circle enclosing zero), and because of (ii) [55] needs to assume that the matrix of eigenvectors is well-conditioned. Under this strong assumption about the matrix of eigenvectors, it was shown that when the operator $\Delta + \widetilde{k}^2$, with $\widetilde{k} = k + \mathrm{i}\alpha$, $\alpha > 0$, is preconditioned by $(\Delta + c + \mathrm{i}d))^{-1}$ with $c \leq 0$, the best choices for $c$ and $d$ (in terms of minimising the number of GMRES iterations) are $c = 0$ and $d = |\widetilde{k}^2|$ [55, §4.1].

Another eigenvalue-analysis of the Helmholtz equation in 1-d with Dirichlet boundary conditions was conducted in [18]. Here, the eigenvalues of a finite-difference discretisation of this problem were calculated, and it was stated that $\varepsilon < k$ is needed for the eigenvalues to be clustered around one (which partially achieves (P1)). Furthermore, a Fourier analysis of multigrid in this paper showed that $\varepsilon \sim k^2$ is needed for multigrid to converge for $\mathbf{A}_\varepsilon$.

Other uses of the shifted Laplacian preconditioner include its use with $\varepsilon \sim k^2$ in the context of domain decomposition methods in [30], and its use with $\varepsilon \sim k$ in the sweeping preconditioner of Enquist and Ying in [13] (these authors consider preconditioning the Helmholtz equation with $k$ replaced by $k + \mathrm{i}\delta$ with $\delta \sim 1$, and this corresponds to choosing $\varepsilon \sim k$). Finally we note that solving the problem with absorption by preconditioning with the inverse of the Laplacian (i.e. aiming to achieve (P2) with $\varepsilon = 0$) has been investigated in [25], [24].

Two points to note from this literature review are the following.

(i) All the analysis of how to choose $\varepsilon$ has focused on studying the eigenvalues of $\mathbf{A}_\varepsilon^{-1}\mathbf{A}$ (and then trying to either minimise $|\lambda_{\max}|/|\lambda_{\min}|$ or cluster the eigenvalues around the point 1).
(ii) All these investigations, apart from that in [55], consider the Helmholtz equation posed in a 1-d interval with Dirichlet boundary conditions, under the assumption that $k^2$ is not an eigenvalue.

Recall that linear systems involving Hermitian matrices can be solved using the conjugate gradient method, and bounds on the number of iterations can be obtained from information about the eigenvalues of the matrix. However, if the matrix is non-Hermitian, general purpose iterative solvers such as GMRES or BiCGStab are required, and information about the spectrum is usually not enough to provide information about the number of iterations required. Even when $\mathbf{A}$ is Hermitian (as is the case for Dirichlet boundary conditions, but not for impedance boundary conditions), $\mathbf{A}_\varepsilon$ is not Hermitian, and therefore the investigations of the eigenvalues of $\mathbf{A}_\varepsilon^{-1}\mathbf{A}$ discussed above are not sufficient to provide bounds on the number of iterations (with this fact noted in [15]).

### 1.2 Statement of the main results

In this paper we prove several results that give sufficient conditions on $\varepsilon$ for the shifted Laplacian to be a good preconditioner for the Helmholtz equation, in the sense that (P1) above is satisfied. We emphasise again that these results alone are not sufficient to decide how to choose the shift in the design of practical preconditioners for $\mathbf{A}$, since they do not consider the cost of constructing approximations of $\mathbf{A}_\varepsilon^{-1}$, or equivalently the question of when the property (P2) holds.

The boundary value problems for the Helmholtz equation that we consider are

1. the interior impedance problem (1.1), and
2. the truncated sound-soft scattering problem.

By "the truncated sound-soft scattering problem" we mean the exterior Dirichlet problem (with zero Dirichlet boundary conditions on the obstacle) where the radiation condition is imposed via an impedance boundary condition on the boundary of a large domain containing the obstacle (i.e. a first-order absorbing boundary condition); see Problem 2.4 and Figure 2.

We consider solving these boundary value problems with FEMs of fixed order. Although such methods suffer from the pollution effect, they are still highly used in applications. We prove results when

(a) the boundary of the domain is smooth and a quasi-uniform sequence of meshes is used, and
(b) the domain is non-smooth and locally refined meshes are used (under suitable assumptions).

For simplicity, we now state the main results of the paper for the interior impedance problem when (a) holds (Theorems 1.4 and 1.5 below). The analogous result for the interior impedance problem when (b) holds is Theorem 4.4, and the analogous result for the truncated sound-soft scattering problem when (a) holds is Theorem 4.5.

**Notation 1.1** *We use the notation $a \lesssim b$ to mean that there exists a $C > 0$ (independent of all parameters of interest and in particular $k, \varepsilon$, and $h$) such that $a \leq C\, b$. We say that $a \sim b$ if $a \lesssim b$ and $a \gtrsim b$.*

Throughout the paper we make the assumption that

$$\varepsilon \lesssim k^2. \tag{1.11}$$

It is possible to derive analogous results for larger $\varepsilon$, but $\varepsilon \sim k^2$ is the largest value of the shift/absorption usually considered in the literature and we do not expect interesting results for larger $\varepsilon$.

**Definition 1.2 (Star-shaped)**
*(i) The domain $\Omega$ is* star-shaped with respect to the point $\mathbf{x}_0 \in \Omega$ *if the line segment $[\mathbf{x}_0, \mathbf{x}]$ is a subset of $\Omega$ for all $\mathbf{x} \in \Omega$.*

*(ii) The domain $\Omega$ is star-shaped with respect to the ball $B_a(\mathbf{x}_0)$ (with $a > 0$ and $\mathbf{x}_0 \in \Omega$) if $\Omega$ is star-shaped with respect to every point in $B_a(\mathbf{x}_0)$.*

**Remark 1.3 (Remark on star-shapedness)** *If $\Omega$ is Lipschitz (and so has a normal vector at almost every point on the boundary) then $\Omega$ is star-shaped with respect to $\mathbf{x}_0$ if and only if $(\mathbf{x} - \mathbf{x}_0) \cdot \mathbf{n}(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \partial\Omega$ for which $\mathbf{n}(\mathbf{x})$ is defined. Furthermore, $\Omega$ is star-shaped with respect to $B_a(\mathbf{x}_0)$ if and only if $(\mathbf{x} - \mathbf{x}_0) \cdot \mathbf{n}(\mathbf{x}) \geq a$ for all $\mathbf{x} \in \partial\Omega$ for which $\mathbf{n}(\mathbf{x})$ is defined (for proofs of these statements see [36, Lemma 5.4.1] or [27, Lemma 3.1]). Whenever we consider a star-shaped domain (in either sense) in this paper, we assume that $\mathbf{x}_0 = \mathbf{0}$.*

**Theorem 1.4 (Sufficient conditions for $\mathbf{A}_\varepsilon^{-1}$ to be a good preconditioner)** *Suppose that* either *$\Omega$ is a $C^{1,1}$ domain in 2- or 3-d that is star-shaped with respect to a ball* or *$\Omega$ is a convex polygon and suppose that $\mathbf{A}$ and $\mathbf{A}_\varepsilon$ are obtained using $H^1$-conforming polynomial elements of fixed order on a quasi-uniform sequence of meshes. Assume that $\varepsilon \lesssim k^2$ and* either *$\eta = k$ or $\eta = \sqrt{k^2 + \mathrm{i}\varepsilon}$. Then, given any $k_0 > 0$ and $C > 0$, there exist $C_1, C_2, C_3 > 0$ (independent of $h, k$, and $\varepsilon$ but depending on $k_0$ and $C$) such that if $hk^2 \geq C$ and*

$$hk\sqrt{|k^2 - \varepsilon|} \leq C_1 \tag{1.12}$$

*then*

$$\left\|\mathbf{I}-\mathbf{A}_\varepsilon^{-1}\mathbf{A}\right\|_2 \leq C_2\frac{\varepsilon}{k} \tag{1.13}$$

*and*

$$\left\|\mathbf{I}-\mathbf{A}\mathbf{A}_\varepsilon^{-1}\right\|_2 \leq C_3\frac{\varepsilon}{k} \tag{1.14}$$

*for all* $k \geq k_0$.

Therefore, if $\varepsilon/k$ is sufficiently small, $\mathbf{A}_\varepsilon^{-1}$ is a good preconditioner for $\mathbf{A}$. (If absorption is added to the original problem by letting $k \mapsto k+\mathrm{i}\delta$, with corresponding Galerkin matrix $\boldsymbol{\mathcal{A}}_\delta$, then the analogues of (1.13) and (1.14) are $\|\mathbf{I}-\boldsymbol{\mathcal{A}}_\delta^{-1}\mathbf{A}\|_2 \leq C_2\delta$ and $\|\mathbf{I}-\mathbf{A}\boldsymbol{\mathcal{A}}_\delta^{-1}\|_2 \leq C_3\delta$, and thus if $\delta$ is sufficiently small, $\boldsymbol{\mathcal{A}}_\delta^{-1}$ is a good preconditioner for $\mathbf{A}$.) Theorem 1.4 has the following consequence.

**Theorem 1.5 ($k$-independent GMRES estimate)** *If the assumptions of Theorem 1.4 hold and $\varepsilon/k$ is sufficiently small, then when GMRES is applied to either of the equations $\mathbf{A}_\varepsilon^{-1}\mathbf{A}\mathbf{u} = \mathbf{A}_\varepsilon^{-1}\mathbf{f}$ or $\mathbf{A}\mathbf{A}_\varepsilon^{-1}\mathbf{v} = \mathbf{f}$, it converges in a $k$–independent number of iterations.*

These two theorems are proved in §4, along with analogous results for non-quasi-uniform meshes.

*Where do the requirements on $h$ in Theorem 1.4 come from?* The requirement (1.12) ensures that the Galerkin method is quasi-optimal, with constant independent of $k$ and $\varepsilon$, when it is applied to the variational problem (1.3), and the proof of Theorem 1.4 requires this quasi-optimality. (Recall that the best result so far about quasi-optimality of the $h$-FEM is that, under some geometric restrictions, quasi-optimality holds with constant independent of $k$ when $hk^2 \lesssim 1$ [35, Prop. 8.2.7]. The condition (1.12) is the analogue of $hk^2 \lesssim 1$ for the shifted problem; see Lemma 3.5 below for more details.) We discuss the condition (1.12) more in Remark 4.2, but note that if quasi-optimality could be proved under less restrictive conditions, then the bound (1.13) would hold under these conditions too.

When dealing with discretisations of the Helmholtz equation one expects to encounter a condition such as (1.12), however one does not usually expect to encounter a condition such as $hk^2 \geq C$ (although in practice this will always be satisfied). This second condition is only necessary when $\eta = \sqrt{k^2+\mathrm{i}\varepsilon}$ (and not when $\eta = k$), and arises from bounding $\|\mathbf{A}_\varepsilon^{-1}\mathbf{N}\|_2$ and $\|\mathbf{N}\mathbf{A}_\varepsilon^{-1}\|_2$ independently of $k$, $\varepsilon$, and $h$; see §1.3 below and Lemma 4.1.

*How sharp is the bound (1.13)?* Numerical evidence suggests that (1.13) is sharp in the sense that the right-hand side cannot be replaced by $\varepsilon/k^\alpha$ for $\alpha > 1$. Indeed, Figure 1 plots the boundary of the numerical range of $\mathbf{A}_\varepsilon^{-1}\mathbf{A}$ for increasing $k$ for each of the three choices $\varepsilon = k$, $\varepsilon = k^{3/2}$, and $\varepsilon = k^2$ (Recall that the numerical range of a matrix $\mathbf{C}$ is the set $W(\mathbf{C}) := \{(\mathbf{C}\,\mathbf{x},\mathbf{x}) : \mathbf{x} \in \mathbb{C}^N, \|\mathbf{x}\|_2 = 1\}$.) In this example, $\Omega$ is the unit square, $\eta = k$, $f = 1$, $g = 0$, $V_N$ is the standard hat-function basis for conforming $P1$ finite elements on a uniform triangular mesh on $\Omega$, and the mesh diameter $h$ is chosen to decrease proportional to $k^{-2}$. The numerical range is computed using an accelerated version of the algorithm of Cowen and Harel [9] (the algorithm is adapted to sparse matrices and the eigenvalues are estimated by an iterative method, which avoids forming the system matrix).

The figures show that when $\varepsilon = k$ the numerical range remains bounded away from the origin as $k$ increases, whereas when $\varepsilon = k^{3/2}$ or $k^2$ the distance of the numerical range from the origin decreases as $k$ increases. This is consistent with the result of Theorem 1.4 since, when $\|\mathbf{x}\|_2 = 1$,

$$\left|\left(\mathbf{A}_\varepsilon^{-1}\mathbf{A}\mathbf{x},\mathbf{x}\right)\right| = \left|1-\left((\mathbf{I}-\mathbf{A}_\varepsilon^{-1}\mathbf{A})\mathbf{x},\mathbf{x}\right)\right| \geq 1-\left\|\mathbf{I}-\mathbf{A}_\varepsilon^{-1}\mathbf{A}\right\|_2 \geq 1-C_2\frac{\varepsilon}{k},$$

(where $C_2$ is the constant in (1.13)). This bound shows that when $\varepsilon/k$ is small enough, the numerical range is bounded away from the origin, although we cannot quantify "small enough" here, since the value of $C_2$ is unknown (although in principle one could work it out).

Of course, these experiments do not rule out the possibility that a bound such as

$$\left\|\mathbf{I}-\mathbf{A}_\varepsilon^{-1}\mathbf{A}\right\|_2 \leq C_3\frac{\varepsilon}{k^\alpha} \tag{1.15}$$

holds for some $\alpha > 1$ and for some large $C_3$. Nevertheless, in §6 we see that the condition "$\varepsilon/k$ sufficiently small" also arises when one considers how well the solution of the boundary value

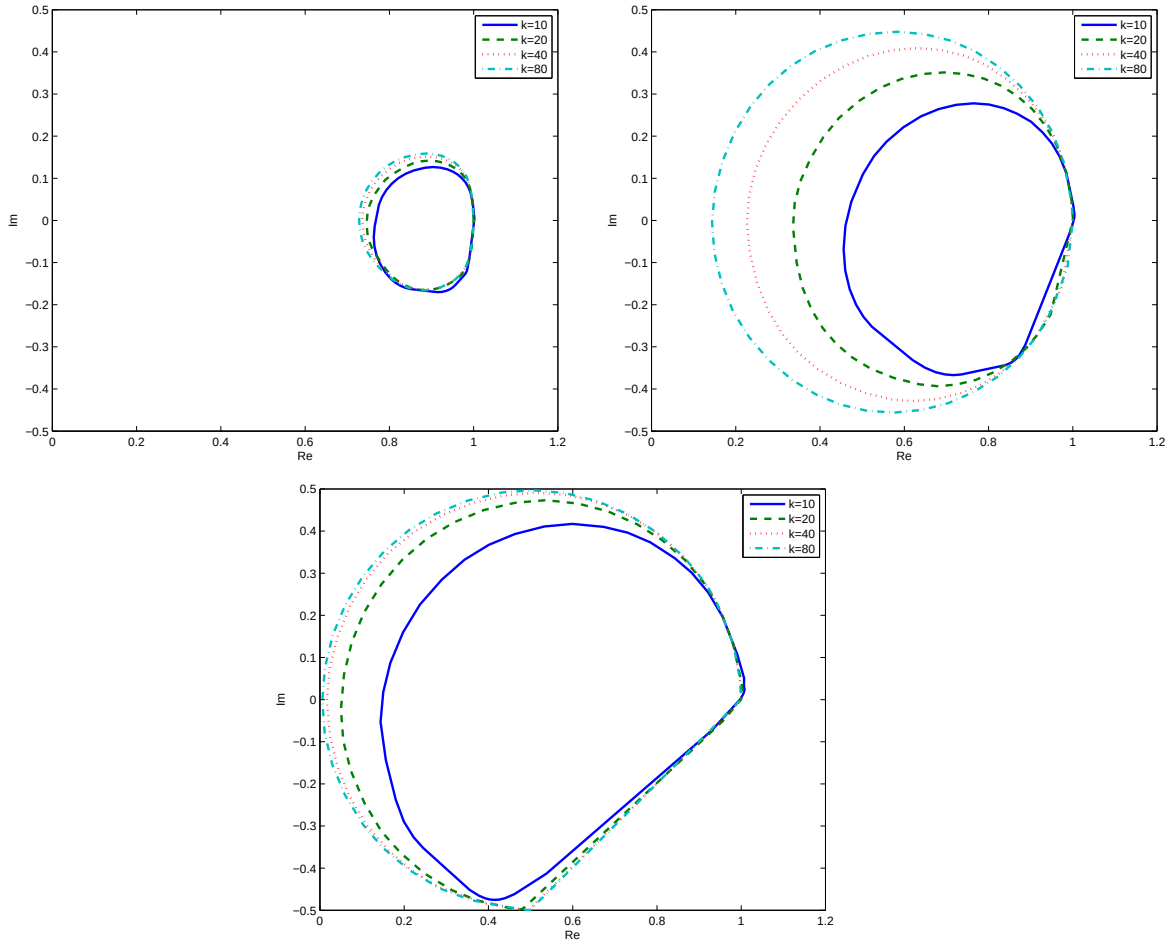


**Fig. 1** The numerical range of $\mathbf{A}_\varepsilon^{-1}\mathbf{A}$, from top left to bottom for $\varepsilon = k$ (top left), $\varepsilon = k^{3/2}$ (top right), and $\varepsilon = k^2$ (bottom), $k = 10, 20, 40, 80$

problem with absorption (1.2) approximates the solution of the boundary value problem without absorption (1.1), independently of any discretisations, and thus we conjecture that (1.15) does not hold for any $\alpha > 1$.

A disadvantage of the bounds in Theorem 1.4 is that they seem to allow for the possibility that $\|\mathbf{I} - \mathbf{A}_\varepsilon^{-1}\mathbf{A}\|_2$ and $\|\mathbf{I} - \mathbf{A}\mathbf{A}_\varepsilon^{-1}\|_2$ might grow with increasing $k$ if $\varepsilon \gg k$. However, we also prove the following result, which rules out any growth.

**Lemma 1.6 (Alternative bound on $\|\mathbf{I} - \mathbf{A}_\varepsilon^{-1}\mathbf{A}\|_2$ and $\|\mathbf{I} - \mathbf{A}\mathbf{A}_\varepsilon^{-1}\|_2$)** *Under the conditions of Theorem 1.4 there exists a $C_4 > 0$ such that*

$$\max\left\{\left\|\mathbf{I} - \mathbf{A}_\varepsilon^{-1}\mathbf{A}\right\|_2, \left\|\mathbf{I} - \mathbf{A}\mathbf{A}_\varepsilon^{-1}\right\|_2\right\} \leq C_4, \tag{1.16}$$

*for all $k \geq k_0$.*

In Table 1 we plot $\mathrm{dist}(0, W(\mathbf{A}_\varepsilon^{-1}\mathbf{A}))$ and also the number of GMRES iterations needed to reduce the initial residual by six orders of magnitude, starting with a zero initial guess, when GMRES is applied to $\mathbf{A}_\varepsilon^{-1}\mathbf{A}\mathbf{x} = \mathbf{A}_\varepsilon^{-1}\mathbf{1}$. The difference between the two sets of results is that results on the left are obtained with $h = k^{-2}$ (in accordance with the conditions of Theorem 1.4), and the results on the right are obtained with the less restrictive condition that $h = k^{-3/2}$; we see that the two sets of results are almost identical.

When $\varepsilon = k$ the number of iterations stays constant as $k$ increases (which is consistent with Theorem 1.5), but when $\varepsilon = k^{3/2}$ or $\varepsilon = k^2$ the number of iterations grows with $k$. The results of more extensive experiments are given in §5, but they all show similar behaviour (i.e. the number of iterations remaining constant as $k$ increases when $\varepsilon = k$, but increasing as $k$ increases for larger $\varepsilon$).

| $k$ | $\varepsilon = k$ | $\varepsilon = k^{3/2}$ | $\varepsilon = k^2$ | $k$ | $\varepsilon = k$ | $\varepsilon = k^{3/2}$ | $\varepsilon = k^2$ |
|---|---|---|---|---|---|---|---|
| 10 | 0.76 (**6**) | 0.46 (**8**) | 0.15 (**13**) | 10 | 0.76 (**6**) | 0.45 (**8**) | 0.14 (**13**) |
| 20 | 0.75 (**6**) | 0.34 (**11**) | 0.055 (**24**) | 20 | 0.75 (**6**) | 0.34 (**11**) | 0.054 (**24**) |
| 40 | 0.74 (**6**) | 0.23 (**14**) | 0.017 (**48**) | 40 | 0.74 (**6**) | 0.23 (**14**) | 0.017 (**48**) |
| 80 | 0.73 (**6**) | 0.14 (**16**) | 0.0060 (**86**) | 80 | 0.73 (**6**) | 0.14 (**16**) | 0.0060 (**86**) |

**Table 1** $\operatorname{dist}(0, W(\mathbf{A}_\varepsilon^{-1}\mathbf{A}))$ and (in bold) the number of GMRES iterations needed to reduce the initial residual by six orders of magnitude starting with a zero initial guess. The results on the left are obtained with $h = k^{-2}$, and the results on the right are obtained with $h = k^{-3/2}$.

### 1.3 The idea behind the proofs of Theorems 1.4 and 1.5

*The idea behind Theorem 1.4.* Considering first the case of left preconditioning and noting that

$$\mathbf{I} - \mathbf{A}_\varepsilon^{-1}\mathbf{A} = \mathbf{A}_\varepsilon^{-1}(\mathbf{A}_\varepsilon - \mathbf{A}) = -\mathrm{i}\varepsilon\mathbf{A}_\varepsilon^{-1}\mathbf{M} - \mathrm{i}(\eta - k)\mathbf{A}_\varepsilon^{-1}\mathbf{N}, \tag{1.17}$$

where $\mathbf{M}$ and $\mathbf{N}$ are as in (1.7), we see that a bound on $\|\mathbf{I} - \mathbf{A}_\varepsilon^{-1}\mathbf{A}\|_2$ can be obtained from bounds on $\|\mathbf{A}_\varepsilon^{-1}\mathbf{M}\|_2$ and $\|\mathbf{A}_\varepsilon^{-1}\mathbf{N}\|_2$. We obtain bounds on $\|\mathbf{A}_\varepsilon^{-1}\mathbf{M}\|_2$ and $\|\mathbf{A}_\varepsilon^{-1}\mathbf{N}\|_2$ in Lemma 4.1 below using an argument that bounds these quantities when $\mathbf{A}_\varepsilon$ is the Galerkin matrix of a general variational problem and one has

(i) a bound on the solution operator of the continuous problem, and
(ii) conditions under which the Galerkin method is quasi-optimal.

In our context, we need the bound (i) and the conditions (ii) (along with the corresponding constant of quasi-optimality) to be explicit in $h$, $k$, and $\varepsilon$.

Regarding (i): proving bounds on the solution of the Helmholtz equation posed in exterior domains is a classic problem, and in particular can be achieved using identities introduced by Morawetz in [39]. Bounds on the solution of the interior impedance problem (1.1) and the truncated sound-soft scattering problem were proved independently (although essentially using Morawetz's identities) in [35], [10], and [26] (see [6, §5.3], [50, §1.2] for discussions of this work). In this paper we use Green's identity to bound the solution of the shifted interior impedance problem (1.2) explicitly in $k$ and $\varepsilon$ when $\varepsilon \gtrsim k$ and $\Omega$ is a general Lipschitz domain, and we use Morawetz's identities to bound the solution (again explicitly in $k$ and $\varepsilon$) when $\varepsilon \lesssim k$ and $\Omega$ is a Lipschitz domain that is star-shaped with respect to a ball. (We also prove analogous results for the truncated sound-soft scattering problem.)

Regarding (ii): $k$-explicit quasi-optimality of the $h$-version of the FEM was proved by Melenk in [35] in the case $\varepsilon = 0$. Indeed Melenk showed that quasi-optimality holds with a quasi-optimality constant independent of $k$ under the condition that $hk^2 \lesssim 1$. This result was obtained using a duality argument that is often attributed to Schatz [47] along with the $k$-explicit bound on the solution discussed in (i). We apply this argument to the case when $\varepsilon > 0$, with the only difference being that the variational formulation of (1.2) is coercive when $\varepsilon > 0$ with coercivity constant $\sim \varepsilon/k^2$ (see Lemma 3.1). Therefore, instead of the mesh threshold $hk^2 \lesssim 1$ we obtain $hk\sqrt{|k^2 - \varepsilon|} \lesssim 1$, reflecting the fact that if $\varepsilon = k^2$ then the uniform coercivity in this case implies that quasi-optimality holds with no mesh threshold.

The argument used to bound $\|\mathbf{A}_\varepsilon^{-1}\mathbf{M}\|_2$ and $\|\mathbf{A}_\varepsilon^{-1}\mathbf{N}\|_2$ in Lemma 4.1 below can also be used to bound $\|\mathbf{A}_\varepsilon^{-1}\|_2$ (when $\mathbf{A}_\varepsilon$ is the Galerkin matrix of a general variational problem) if one has (i) and (ii) above. We have not been able to find this argument explicitly in the literature, although it is alluded to in [31, Last paragraph of §2.4]. Furthermore, put another way, this argument states that if the sesquilinear form satisfies a continuous inf-sup condition and the Galerkin solutions exist, are unique, and are quasi-optimal, then one can obtain a discrete inf-sup condition. When phrased in this way, this result can be seen as a special case of [33, Theorem 3.9].

**Remark 1.7** *The argument for the case of right preconditioning is very similar, in that a bound on* $\|\mathbf{I} - \mathbf{A}\mathbf{A}_\varepsilon^{-1}\|_2$ *can be obtained from bounds on* $\|\mathbf{M}\mathbf{A}_\varepsilon^{-1}\|_2$ *and* $\|\mathbf{N}\mathbf{A}_\varepsilon^{-1}\|_2$. *Then, because* $\mathbf{M}$ *and* $\mathbf{N}$ *are real symmetric matrices,*

$$\left\|\mathbf{M}\mathbf{A}_\varepsilon^{-1}\right\|_2 = \|\left(\mathbf{M}\mathbf{A}_\varepsilon^{-1}\right)^*\|_2 = \left\|(\mathbf{A}_\varepsilon^*)^{-1}\mathbf{M}\right\|_2 \quad \text{and} \quad \left\|\mathbf{N}\mathbf{A}_\varepsilon^{-1}\right\|_2 = \|\left(\mathbf{N}\mathbf{A}_\varepsilon^{-1}\right)^*\|_2 = \left\|(\mathbf{A}_\varepsilon^*)^{-1}\mathbf{N}\right\|_2.$$

*Since the matrix $\mathbf{A}_\varepsilon^*$ is simply the Galerkin matrix corresponding to the adjoint to problem (1.2), and we also have bounds on the solution operator for this problem (as outlined in Remark 2.5), the argument to obtain bounds on $\|\mathbf{A}_\varepsilon^{-1}\mathbf{M}\|_2$ and $\|\mathbf{A}_\varepsilon^{-1}\mathbf{N}\|_2$ in Lemma 4.1 below can be repeated for the adjoint problem, resulting in bounds on $\|\mathbf{M}\mathbf{A}_\varepsilon^{-1}\|_2$ and $\|\mathbf{N}\mathbf{A}_\varepsilon^{-1}\|_2$.*

*The idea behind Theorem 1.5.* Theorem 1.5 follows from Theorem 1.4 by using the Elman estimate for GMRES.

**Theorem 1.8** *If the matrix equation $\mathbf{C}\mathbf{x} = \mathbf{y}$ is solved using GMRES then, for $m \in \mathbb{N}$, the GMRES residual $\mathbf{r}_m := \mathbf{C}\mathbf{x}_m - \mathbf{y}$ satisfies*

$$\frac{\|\mathbf{r}_m\|_2}{\|\mathbf{r}_0\|_2} \leq \sin^m \beta, \qquad \text{where} \quad \cos\beta = \frac{\operatorname{dist}\big(0, W(\mathbf{C})\big)}{\|\mathbf{C}\|_2} \tag{1.18}$$

*(recall that $W(\mathbf{C}) := \{(\mathbf{C}\mathbf{x}, \mathbf{x}) : \mathbf{x} \in \mathbb{C}^N, \|\mathbf{x}\|_2 = 1\}$ is the* numerical range *or* field of values*).*

The bound (1.18) was originally proved in [12] (see also [11, Theorem 3.3]) and appears in the form above in [3, Equation 1.2]. A variant of this theory, where the Euclidean inner product $(\cdot,\cdot)$ and norm $\|\cdot\|_2$ are replaced by a general inner product and norm, is used in [5].

Theorem 1.8 has the following corollary.

**Corollary 1.9** *If $\|\mathbf{I} - \mathbf{C}\|_2 \leq \sigma < 1$, then in (1.18)*

$$\cos\beta \geq \frac{1-\sigma}{1+\sigma} \quad \text{and} \quad \sin\beta \leq \frac{2\sqrt{\sigma}}{(1+\sigma)^2}\,.$$

Theorem 1.5 follows from Theorem 1.4 by applying Corollary 1.9 with $\mathbf{C} = \mathbf{A}_\varepsilon^{-1}\mathbf{A}$. Indeed, Theorem 1.4 shows that if $\varepsilon/k$ is sufficiently small, $\|\mathbf{I} - \mathbf{C}\|_2$ can be bounded below one, independently of $k$ and $\varepsilon$. Therefore GMRES converges and the number of iterations is independent of $k$.

### 1.4 Outline and preliminaries

In Section 2 we prove bounds that are explicit in $k, \eta$, and $\varepsilon$ on the solutions of the shifted interior impedance problem (1.2) and the shifted truncated sound-soft scattering problem. In Section 3 we prove results about the continuity and coercivity of $a_\varepsilon(\cdot,\cdot)$ and obtain sufficient conditions for the Galerkin method applied to $a_\varepsilon(\cdot,\cdot)$ to be quasi-optimal (with all the constants given explicitly in terms of $k$, $\eta$, and $\varepsilon$). In Section 4 we put the results of Sections 2 and 3 together to prove Theorem 1.4 and its analogue for non-quasi-uniform meshes. In Section 5 we illustrate the theory with numerical experiments. Section 6 contains some concluding remarks about approximating the solution of (1.1) by the solution of (1.2), independently of any discretisations.

*Notation and recap of elementary results.* Let $\Omega \subset \mathbb{R}^d$, $d = 2$, or 3, be a bounded Lipschitz domain (where by "domain" we mean a connected open set) with boundary $\Gamma$. We do not introduce any special notation for the trace operator, and thus the trace theorem is simply

$$\|v\|_{H^{1/2}(\Gamma)} \lesssim \|v\|_{H^1(\Omega)} \quad \text{for all } v \in H^1(\Omega) \tag{1.19}$$

(see [34, Theorem 3.38, Page 102]), and the multiplicative trace inequality is

$$\|v\|^2_{L^2(\partial\Omega)} \lesssim \|v\|_{L^2(\Omega)}\,\|v\|_{H^1(\Omega)} \quad \text{for all } v \in H^1(\Omega) \tag{1.20}$$

[22, Theorem 1.5.1.10, last formula on Page 41].

Let $\partial_n$ denote the normal-derivative trace on $\Omega$ (with the convention that the normal vector points out of $\Omega$). Recall that if $u \in H^2(\Omega)$ then $\partial_n u := \mathbf{n}\cdot\nabla u$, and, for $u \in H^1(\Omega)$ with $\Delta u \in L^2(\Omega)$, $\partial_n u$ is defined so that Green's first identity holds (see, e.g., [6, Equation (A.29)]). Denote the surface gradient on $\Gamma$ by $\nabla_\Gamma$; see, e.g., [6, Equation A.14] for the definition of this operator in terms of a parametrisation of the boundary.

Finally, we repeatedly use the inequalities

$$2ab \leq \frac{a^2}{\delta} + \delta b^2 \tag{1.21}$$

and

$$\frac{1}{2}(a+b)^2 \leq a^2 + b^2 \leq (a+b)^2, \tag{1.22}$$

where $a, b$, and $\delta$ are all $> 0$. (Recalling Notation 1.1, we see that (1.22) implies that $a + b \sim \sqrt{a^2 + b^2}$.)

## 2 Bounds on the solution operators to the problems with absorption

In this section we prove bounds that are explicit in $k$, $\eta$, and $\varepsilon$ on the solutions of the shifted interior impedance problem and the shifted truncated sound-soft scattering problem. First, we define precisely what we mean by these problems.

**Problem 2.1 (Interior Impedance Problem with absorption)** *Let $\Omega \subset \mathbb{R}^d$, with $d = 2$ or 3, be a bounded Lipschitz domain with outward-pointing unit normal vector $\mathbf{n}$ and let $\Gamma := \partial\Omega$. Given $f \in L^2(\Omega)$, $g \in L^2(\Gamma)$, $\eta \in \mathbb{C} \setminus \{0\}$ and $\varepsilon \geq 0$, find $u \in H^1(\Omega)$ such that*

$$\Delta u + (k^2 + \mathrm{i}\varepsilon)u = -f \quad \text{in } \Omega, \tag{2.1a}$$

$$\partial_n u - \mathrm{i}\eta u = g \quad \text{on } \Gamma. \tag{2.1b}$$

**Remark 2.2 (Existence and uniqueness)** *One can prove using Green's identity that the solution of the Problem 2.1 (if it exists) is unique; see §2.1.2. One can prove via Fredholm theory (using the fact that $H^1(\Omega)$ is compactly contained in $L^2(\Omega)$) that uniqueness implies existence in exactly the same way as for the problem with $\varepsilon = 0$.*

**Remark 2.3 (The choice of $\eta$)** *If one thinks of the impedance boundary condition as being a first order approximation to the Sommerfeld radiation condition, then for the unshifted problem $\eta$ should be equal to $k$, and for the shifted problem $\eta$ should be equal to $\sqrt{k^2 + \mathrm{i}\varepsilon}$. With $\eta_R$ and $\eta_I$ denoting the real and imaginary parts of $\eta$ respectively, we prove bounds under the assumption that $\eta_R \sim k$ and $0 \leq \eta_I \lesssim k$. These assumptions cover both the case that $\eta = k$ and the case that $\eta = \sqrt{k^2 + \mathrm{i}\varepsilon}$ (recall that we assume that $\varepsilon \lesssim k^2$).*

**Problem 2.4 (Truncated sound-soft scattering problem with absorption)** *Let $\Omega_D$ be a bounded Lipschitz open set in $\mathbb{R}^d$ ($d = 2$ or 3) such that the open complement $\Omega_+ := \mathbb{R}^d \setminus \Omega_D$ is connected. Let $\Omega_R$ be a bounded Lipschitz domain such that $\Omega_D \subset \Omega_R \subset \mathbb{R}^d$ with $d(\Omega_D, \partial\Omega_R) > 0$ (where $d(\cdot,\cdot)$ is the distance function). Let $\Gamma_R := \partial\Omega_R$, $\Gamma_D := \partial\Omega_D$, and $\Omega := \Omega_R \setminus \overline{\Omega_D}$ (thus $\partial\Omega = \Gamma_R \cup \Gamma_D$ and $\Gamma_R \cap \Gamma_D = \emptyset$). Given $f \in L^2(\Omega)$, $g \in L^2(\Gamma_R)$, $\eta \in \mathbb{C} \setminus \{0\}$, and $\varepsilon \geq 0$, find $u \in H^1(\Omega)$ such that*

$$\Delta u + (k^2 + \mathrm{i}\varepsilon)u = -f \quad \text{in } \Omega, \tag{2.2a}$$

$$\partial_n u - \mathrm{i}\eta u = g \quad \text{on } \Gamma_R, \tag{2.2b}$$

$$u = 0 \quad \text{on } \Gamma_D. \tag{2.2c}$$

If $\varepsilon = 0, \eta = k$, $\Omega_R$ is a large ball containing $\Omega_D$, and $f$ and $g$ are chosen appropriately, then the solution of the truncated sound-soft scattering problem is a classical approximation to the solution of the sound-soft scattering problem (see, e.g., [6, Equation (2.16)]); Figure 2 shows $\Omega_R$ and $\Omega_D$ in this case. We use the convention that on $\Gamma_D$ the normal derivative $\partial_n v$ equals $\mathbf{n}_D \cdot \nabla v$ for $v$ that are $H^2$ in a neighbourhood of $\Gamma_D$, and similarly $\partial_n v = \mathbf{n}_R \cdot \nabla v$ on $\Gamma_R$, where $\mathbf{n}_D$ and $\mathbf{n}_R$ are oriented as in Figure 2. Note that Remarks 2.2 and 2.3 also apply to Problem 2.4.

We go through the details of the bounds for Problem 2.1 in §2.1, and then outline in §2.2 the (small) modifications needed to the arguments to prove the analogous bounds for Problem 2.4.

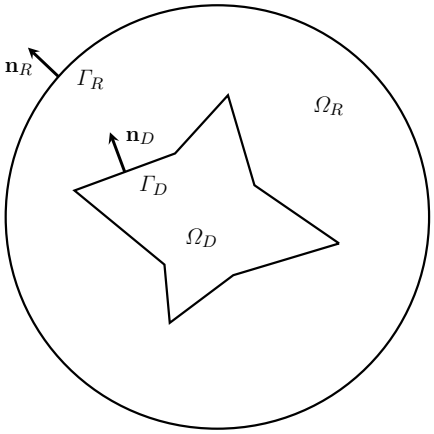


**Fig. 2** An example of the domains $\Omega_D$ and $\Omega_R$ in Problem 2.4.

## 2.1 Bounds on the interior impedance problem with absorption

**Remark 2.5 (The adjoint problem)** *All the bounds on the solution of the interior impedance problem proved in this section are also valid when the signs of $\varepsilon$ and $\eta$ are changed; i.e. the bounds also hold for the solution of*

$$\Delta w + (k^2 - \mathrm{i}\varepsilon) w = -f \quad \text{in } \Omega, \tag{2.3a}$$

$$\partial_n w + \mathrm{i}\eta w = g \quad \text{on } \Gamma \tag{2.3b}$$

*(under the same conditions on $\varepsilon$ and $\eta$). This fact is not immediately obvious (one has to go through the proofs and check).*

**Remark 2.6 (Regularity)** *Let $u$ be the solution of Problem 2.1. Since $f \in L^2(\Omega)$ we have that $\Delta u \in L^2(\Omega)$, and since $g \in L^2(\Gamma)$ we have that $\partial_n u \in L^2(\Gamma)$. These two facts imply that $u \in H^1(\Gamma)$ by a regularity result of Nečas for Lipschitz domains [41, §5.2.1], [34, Theorem 4.24(ii)].*

We now state the two main results of this section.

**Theorem 2.7 (Bound for $\varepsilon > 0$ for general Lipschitz $\Omega$)** *Let $u$ solve Problem 2.1, let $\eta = \eta_R + \mathrm{i}\eta_I$ and assume that $\eta_I \geq 0$, $\eta_R > 0$. Then, given $k_0 > 0$, there exists a $C > 0$, independent of $\varepsilon$, $k$, $\eta_R$, and $\eta_I$, such that*

$$\|\nabla u\|^2_{L^2(\Omega)} + k^2 \|u\|^2_{L^2(\Omega)} \leq C \left[ \frac{k^2}{\varepsilon^2} \left( 1 + \frac{\varepsilon}{k^2} + \left( \frac{\varepsilon}{k^2} \right)^2 \right) \|f\|^2_{L^2(\Omega)} + \frac{k^2}{\varepsilon \eta_R} \left( 1 + \frac{\varepsilon}{k^2} \right) \|g\|^2_{L^2(\Gamma)} \right] \tag{2.4}$$

*for all $k \geq k_0$, $\eta_R > 0$, and $\varepsilon > 0$.*

Assuming that $\varepsilon \lesssim k^2$, we obtain the following corollary.

**Corollary 2.8** *If the conditions in Theorem 2.7 hold and, in addition, $\varepsilon \lesssim k^2$, then*

$$\|\nabla u\|^2_{L^2(\Omega)} + k^2 \|u\|^2_{L^2(\Omega)} \lesssim \left[ \frac{k^2}{\varepsilon^2} \|f\|^2_{L^2(\Omega)} + \frac{k^2}{\varepsilon \eta_R} \|g\|^2_{L^2(\Gamma)} \right] \tag{2.5}$$

*for all $k \geq k_0$, $\eta_R > 0$, and $\varepsilon > 0$. In particular, if $\eta_I \geq 0$, $\eta_R \sim k$, and $\varepsilon \sim k$ then*

$$\|\nabla u\|^2_{L^2(\Omega)} + k^2 \|u\|^2_{L^2(\Omega)} \lesssim \left[ \|f\|^2_{L^2(\Omega)} + \|g\|^2_{L^2(\Gamma)} \right], \tag{2.6}$$

*while if $\eta_I \geq 0$, $\eta_R \sim k$, and $\varepsilon \sim k^2$ then*

$$\|\nabla u\|^2_{L^2(\Omega)} + k^2 \|u\|^2_{L^2(\Omega)} \lesssim \left[ \frac{1}{k^2} \|f\|^2_{L^2(\Omega)} + \frac{1}{k} \|g\|^2_{L^2(\Gamma)} \right]$$

*for all $k \geq k_0$.*

This corollary shows how the $k$-dependence of the bounds on the solution operator improves as $\varepsilon$ is increased from $k$ to $k^2$.

As $\varepsilon \to 0$, the right-hand side of (2.5) blows up. A bound that is valid uniformly in this limit can be obtained by imposing some geometric restrictions on $\Omega$.

**Theorem 2.9 (Bound for $\varepsilon/k$ sufficiently small when $\Omega$ is star-shaped with respect to a ball and Lipschitz)** *Let $\Omega$ be a Lipschitz domain that is star-shaped with respect to a ball (see Definition 1.2), and let $u$ be the solution of Problem 2.1 in $\Omega$. If $\eta_R \sim k$ and $|\eta_I| \lesssim k$ then, given $k_0 > 0$, there exist $c$ and $C$ (independent of $k$, $\varepsilon$, and $\eta$ and $> 0$) such that, if $\varepsilon/k \leq c$ for all $k \geq k_0$, then*

$$\|\nabla u\|^2_{L^2(\Omega)} + k^2 \|u\|^2_{L^2(\Omega)} \leq C \left[\|f\|^2_{L^2(\Omega)} + \|g\|^2_{L^2(\Gamma)}\right] \quad \text{for all} \quad k \geq k_0. \tag{2.7}$$

**Remark 2.10 (The case $\varepsilon = 0$)** *The bound (2.7) for $\varepsilon = 0$ was proved for $d = 2$ in [35, Prop. 8.1.4] and for $d = 3$ in [10, Theorem 1] using essentially the same methods we use here (see Remark 2.16 for more details).*

It is useful for what follows to combine the results of Theorems 2.7 and 2.9 to form the following corollary.

**Corollary 2.11 (Bound for $\varepsilon \lesssim k^2$)** *If $\Omega$ is star-shaped with respect to a ball, $\varepsilon \lesssim k^2$, $\eta_R \sim k$, and $0 \leq \eta_I \lesssim k$, then, given $k_0 > 0$,*

$$\|\nabla u\|^2_{L^2(\Omega)} + k^2 \|u\|^2_{L^2(\Omega)} \lesssim \left[\|f\|^2_{L^2(\Omega)} + \|g\|^2_{L^2(\Gamma)}\right] \tag{2.8}$$

*for all $k \geq k_0$.*

In §3 we find sufficient conditions for the Galerkin method applied to Problem 2.1 to be quasi-optimal. To do this, we need a bound on the $H^2$-norm of the solution (in cases where the solution is in $H^2(\Omega)$), and this can be obtained by combining the following lemma with the bound (2.8).

**Lemma 2.12 (A bound on the $H^2(\Omega)$ norm)** *Let $u$ be the solution of Problem 2.1, and assume further that $g \in H^{1/2}(\Gamma)$. If $\Omega$ is $C^{1,1}$ (in 2- or 3-d) then $u \in H^2(\Omega)$ and there exists a $C$ (independent of $k$ and $\varepsilon$) such that*

$$\|u\|_{H^2(\Omega)} \leq C \left[(1+k)\sqrt{\|\nabla u\|^2_{L^2(\Omega)} + k^2 \|u\|^2_{L^2(\Omega)}} + \|f\|_{L^2(\Omega)} + \|g\|_{H^{1/2}(\Gamma)}\right] \tag{2.9}$$

*for all $k > 0$ and $\varepsilon \geq 0$. Furthermore, if $\Omega$ is a convex polygon and $g \in H^{1/2}_{\rm pw}(\Gamma)$ (i.e. $H^{1/2}$ on each side) then the bound (2.9) also holds, with $\|g\|_{H^{1/2}(\Gamma)}$ replaced by $\|g\|_{H^{1/2}_{\rm pw}(\Gamma)}$ (i.e. the sum of the $H^{1/2}$–norms of $g$ on each side).*

*Proof of Lemma 2.12.* First consider the case when $\Omega$ is $C^{1,1}$. By [22, Theorem 2.3.3.2, Page 106], if $v \in H^1(\Omega)$ with $\Delta v \in L^2(\Omega)$ and $\partial_n v \in H^{1/2}(\Gamma)$ then

$$\|v\|_{H^2(\Omega)} \lesssim \left(\|\Delta v\|_{L^2(\Omega)} + \|v\|_{H^1(\Omega)} + \|\partial_n v\|_{H^{1/2}(\Gamma)}\right). \tag{2.10}$$

The bound (2.9) then follows from (2.10) by using (i) the fact that $u$ satisfies the PDE (2.1a) and boundary conditions (2.1b), and (ii) the trace theorem (1.19).

When $\Omega$ is a convex polygon, the result (2.9) will follow if we can again establish that (2.10) holds (except with the condition that $\partial_n v \in H^{1/2}(\Gamma)$ replaced by $\partial_n v \in H^{1/2}_{\rm pw}(\Gamma)$). (There is a slight subtlety in that we need to show that $\|u\|_{H^{1/2}_{\rm pw}(\Gamma)} \lesssim \|u\|_{H^1(\Omega)}$, but this follows from the trace result for polygons in [22, Part (c) of Theorem 1.5.2.3, Page 43] using the fact that $u$ is continuous at the corners of the polygon. This latter fact follows from the Sobolev embedding theorem [34, Theorem 3.26] and the fact that $u \in H^1(\Gamma)$, which follows from the regularity result of Nečas [34, Theorem 4.24 (ii)] since $u \in H^2(\Omega)$ implies $\partial_n u \in L^2(\Gamma)$.)

The bound (2.10) can be established when $\Omega$ is a convex polygon by combining two results in [22] and performing some additional work as follows. When $\Omega$ is a convex polygon and $v$ is such that $v \in H^1(\Omega)$, $\Delta v \in L^2(\Omega)$, and $\partial_n v = 0$ on $\Gamma$, then

$$\|v\|_{H^2(\Omega)} \lesssim \left(\|\Delta v\|_{L^2(\Omega)} + \|v\|_{L^2(\Omega)}\right) \tag{2.11}$$

by [22, Theorem 4.3.1.4, Page 198]. When $\partial_n v \neq 0$ but is in $H^{1/2}_{\rm pw}(\Gamma)$ then $v \in H^2(\Omega)$ by [22, Corollary 4.4.3.8, Page 233] (note that the sum in [22, Equation 4.4.3.8] is empty since $\Omega$ is convex). Therefore, by linearity, to prove that the bound (2.10) holds when $\Omega$ is a convex polygon we only need to show that for these domains there exists a lifting operator $G : H^{1/2}_{\rm pw}(\Gamma) \to H^2(\Omega)$ with $\partial_n G(g) = g$ and

$$\|G(g)\|_{H^2(\Omega)} \lesssim \|g\|_{H^{1/2}_{\rm pw}(\Gamma)} \tag{2.12}$$

(in fact we show below that this is the case when $\Omega$ is any polygon). Using a partition of unity it is sufficient to construct such an operator when (i) $\Omega$ is a half-space, and (ii) $\Omega$ is an infinite wedge.

For (i), given $g$ define $G(g)$ to be the solution of the Neumann problem for Laplace's equation in $\Omega$ (with Neumann data $g$). The explicit expression for the solution in terms of the Fourier transform shows that (2.12) is satisfied.

For (ii), first consider the case when the wedge angle is $\pi/2$ (i.e. a right-angle). By linearity we can take $g$ to be zero on one side of the wedge. Introduce coordinates $(x_1, x_2)$ so that $g \neq 0$ on the positive $x_1$–axis and $g = 0$ on the positive $x_2$–axis. Extend $g$ to the negative $x_1$–axis by requiring that $g$ is even about $x_1 = 0$; one can then show that this extension is a continuous mapping from $H^{1/2}(\mathbb{R}^+)$ to $H^{1/2}(\mathbb{R})$. The solution of the Neumann problem for Laplace's equation in the half-space $\{(x_1, x_2) : x_2 > 0\}$ then satisfies $\partial_n u = 0$ on the positive $x_2$–axis, and thus this function satisfies the requirements of the lifting. A lifting for a wedge of arbitrary angle can be obtained from a lifting for a right-angled wedge by expressing the function in polar coordinates and rescaling the angular variable. (Note that all our liftings up to this point have satisfied Laplace's equation. Rescaling the angular variable means that the resulting function does not satisfy Laplace's equation, but is still in $H^2(\Omega)$.) ∎

### *2.1.1 Green, Rellich, and Morawetz identities for the Helmholtz equation*

For the proofs of Theorems 2.7 and 2.9 we need the following identities.

**Lemma 2.13 (Green, Rellich, and Morawetz identities for the Helmholtz equation)** *Let $v \in C^2(D)$ for some domain $D \subset \mathbb{R}^d$, and let*

$$\mathcal{L}v := \Delta v + k^2 v, \qquad \mathcal{M}v := \mathbf{x} \cdot \nabla v + \alpha v,$$

*for $k$ and $\alpha \in \mathbb{R}$. Then, on the domain $D$,*

$$\overline{v}\mathcal{L}v = \nabla \cdot \left[\overline{v}\nabla v\right] - |\nabla v|^2 + k^2|v|^2 \qquad \textit{(Green)}, \tag{2.13}$$

$$2\Re\left(\mathbf{x} \cdot \overline{\nabla v}\mathcal{L}v\right) = \nabla \cdot \left[2\Re\left(\mathbf{x} \cdot \overline{\nabla v}\,\nabla v\right) + (k^2|v|^2 - |\nabla v|^2)\mathbf{x}\right] + (d-2)|\nabla v|^2 - dk^2|v|^2 \qquad \textit{(Rellich)}, \tag{2.14}$$

$$2\Re\left(\overline{\mathcal{M}v}\mathcal{L}v\right) = \nabla \cdot \left[2\Re\left(\overline{\mathcal{M}v}\,\nabla v\right) + (k^2|v|^2 - |\nabla v|^2)\mathbf{x}\right] + (d-2-2\alpha)|\nabla v|^2 + (2\alpha - d)k^2|v|^2 \qquad \textit{(Morawetz)}. \tag{2.15}$$

*Proof of Lemma 2.13.* The identities (2.13) and (2.14) can be proved by expanding the divergences on the right-hand sides; for (2.13) this is straightforward, but for (2.14) this is more involved; see, e.g., [51, Lemma 2.1] for the details. The identity (2.15) is then (2.14) plus $2\alpha$ times the real part of (2.13). ∎

**Remark 2.14** *All three of the identities in Lemma 2.13 are formed by multiplying the Helmholtz operator $\mathcal{L}v$ by a function of $v$, say $\overline{\mathcal{N}v}$, and then expressing this quantity as the divergence of something plus some non-divergence terms. The multiplier $\mathcal{N}v = v$ is associated with the name of Green, and (2.13) is a special case of the pointwise form (as opposed to integrated form) of Green's*

*first identity. The multiplier $\mathcal{N}v = \mathbf{x}\cdot\nabla v$ was introduced by Rellich in [43], and identities resulting from multipliers that are derivatives of $v$ are thus often called Rellich identities. The idea of taking $\mathcal{N}v$ to be a linear combination of $v$ and a derivative of $v$ (in general $\mathbf{Z}\cdot\nabla v - \mathrm{i}k\beta v + \alpha v$ for $\mathbf{Z}$ a real vector field and $\beta$ and $\alpha$ real scalar fields) was used extensively by Morawetz in the context of the Helmholtz and wave equations; see [38], [40], and [39]. The identity (2.15) is essentially contained in [39, §I.2] and [40]; see [52, Remark 2.7] for more details. For more discussion of Rellich and Morawetz identities, see [6, §5.3].*

For the proofs of Theorems 2.7 and 2.9, we integrate the indentities (2.13) and (2.15) over $\Omega$.

**Lemma 2.15 (Integrated forms of the Green and Morawetz identities)** *With $\Omega$ as in Problem 2.1, define the space $V$ by*

$$V := \left\{ v : v \in H^1(\Omega),\ \Delta v \in L^2(\Omega),\ \partial_n v \in L^2(\Gamma),\ v \in H^1(\Gamma) \right\}, \tag{2.16}$$

*(note that either of the conditions $\partial_n v \in L^2(\Gamma)$ or $v \in H^1(\Gamma)$ can be dropped from the definition of $V$ by the results of Nečas [41, §5.1.2, 5.2.1], [34, Theorem 4.24]). Then, with $\mathcal{L}v$ and $\mathcal{M}v$ as in Lemma 2.13, if $v \in V$ then*

$$\int_\Omega \overline{v}\mathcal{L}v = \int_\Gamma \overline{v}\,\partial_n v + \int_\Omega k^2|v|^2 - |\nabla v|^2 \tag{2.17}$$

*and*

$$\int_\Omega 2\Re\big(\overline{\mathcal{M}v}\mathcal{L}v\big) = \int_\Gamma 2\Re\big(\overline{\mathcal{M}v}\,\partial_n v\big) + \big(k^2|v|^2 - |\nabla v|^2\big)(\mathbf{x}\cdot\mathbf{n}) + \int_\Omega (d-2-2\alpha)|\nabla v|^2 + (2\alpha - d)k^2|v|^2, \tag{2.18}$$

*where the expression $\nabla v$ in the integral on $\Gamma$ is understood as $\nabla_\Gamma v + \mathbf{n}\partial_n v$.*

*Proof of Lemma 2.15.* Equations (2.17) and (2.18) hold as consequences of the divergence theorem applied to the identities (2.13) and (2.15). Indeed, the divergence theorem $\int_\Omega \nabla\cdot\mathbf{F} = \int_\Gamma \mathbf{F}\cdot\mathbf{n}$ is valid when $\Omega$ is Lipschitz and $\mathbf{F} \in (C^1(\overline{\Omega}))^d$ [34, Theorem 3.34]. Therefore, (2.17) and (2.18) hold for $v \in \mathcal{D}(\overline{\Omega}) := \{U|_\Omega : U \in C_0^\infty(\mathbb{R}^d)\}$. By the density of $\mathcal{D}(\overline{\Omega})$ in the space $V$ [37, Appendix A], (2.17) and (2.18) hold for $v \in V$. ∎

### *2.1.2 Proof of Theorem 2.7*

*Outline* The only ingredients for the proof are the integrated form of Green's identity (2.17), the Cauchy-Schwarz inequality, and the inequality (1.21). By Remark 2.6, the solution $u$ of Problem 2.1 is in the space $V$; therefore, by Lemma 2.15, (2.17) holds with $v$ replaced by $u$. Using the impedance boundary condition (2.1b) and the fact that $\mathcal{L}u = -f - \mathrm{i}\varepsilon u$ in $\Omega$, we obtain

$$(k^2 + \mathrm{i}\varepsilon)\,\|u\|^2_{L^2(\Omega)} - \|\nabla u\|^2_{L^2(\Omega)} + (\mathrm{i}\eta_R - \eta_I)\,\|u\|^2_{L^2(\Gamma)} = -\int_\Omega f\,\overline{u} - \int_\Gamma g\,\overline{u}. \tag{2.19}$$

From here, the proof consists of the following three steps:

1. Use the imaginary part of (2.19) to estimate $\|u\|^2_{L^2(\Omega)}$ and $\|u\|^2_{L^2(\Gamma)}$ by $\|f\|^2_{L^2(\Omega)}$ and $\|g\|^2_{L^2(\Gamma)}$.
2. Use the real part of (2.19) to estimate $\|\nabla u\|^2_{L^2(\Omega)} + k^2\,\|u\|^2_{L^2(\Omega)}$ by $\|u\|^2_{L^2(\Omega)}$, $\|u\|^2_{L^2(\Gamma)}$, $\|f\|^2_{L^2(\Omega)}$, and $\|g\|^2_{L^2(\Gamma)}$.
3. Put the estimates of Steps 1 and 2 together to give the result (2.4).

*Step 1.* Taking the imaginary part of (2.19) and using the Cauchy-Schwarz inequality, we obtain

$$\varepsilon \|u\|^2_{L^2(\Omega)} + \eta_R \|u\|^2_{L^2(\Gamma)} \leq \|f\|_{L^2(\Omega)} \|u\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma)} \|u\|_{L^2(\Gamma)} . \tag{2.20}$$

(Note that this inequality establishes uniqueness of the interior impedance problem with absorption, since if $f$ and $g$ are both zero then the inequality implies that $u$ is zero in $\Omega$.) Using the inequality (1.21) on both terms on the right-hand side, we find that

$$\left(\varepsilon - \frac{\delta_1}{2}\right) \|u\|^2_{L^2(\Omega)} + \left(\eta_R - \frac{\delta_2}{2}\right) \|u\|^2_{L^2(\Gamma)} \leq \frac{1}{2\delta_1} \|f\|^2_{L^2(\Omega)} + \frac{1}{2\delta_2} \|g\|^2_{L^2(\Gamma)} . \tag{2.21}$$

Taking $\delta_1 = \varepsilon$ and $\delta_2 = \eta_R$, we obtain

$$\frac{\varepsilon}{2} \|u\|^2_{L^2(\Omega)} + \frac{\eta_R}{2} \|u\|^2_{L^2(\Gamma)} \leq \frac{1}{2\varepsilon} \|f\|^2_{L^2(\Omega)} + \frac{1}{2\eta_R} \|g\|^2_{L^2(\Gamma)} . \tag{2.22}$$

*Step 2.* Taking the real part of (2.19) yields

$$-k^2 \|u\|^2_{L^2(\Omega)} + \|\nabla u\|^2_{L^2(\Omega)} + \eta_I \|u\|^2_{L^2(\Gamma)} = \Re \int_\Omega f\,\overline{u} + \Re \int_\Gamma g\,\overline{u},$$

and thus (since $\eta_I \geq 0$)

$$\|\nabla u\|^2_{L^2(\Omega)} \leq k^2 \|u\|^2_{L^2(\Omega)} + \|f\|_{L^2(\Omega)} \|u\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma)} \|u\|_{L^2(\Gamma)} .$$

Adding $k^2 \|u\|^2_{L^2(\Omega)}$ to both sides and then using the inequality (1.21) on the terms involving $f$ and $g$, we obtain

$$\|\nabla u\|^2_{L^2(\Omega)} + k^2 \|u\|^2_{L^2(\Omega)} \leq \left(2k^2 + \frac{\delta_1}{2}\right) \|u\|^2_{L^2(\Omega)} + \frac{\delta_2}{2} \|u\|^2_{L^2(\Gamma)} + \frac{1}{2}\left(\frac{1}{\delta_1} \|f\|^2_{L^2(\Omega)} + \frac{1}{\delta_2} \|g\|^2_{L^2(\Gamma)}\right) . \tag{2.23}$$

*Step 3.* We choose $\delta_1 = k^2$ in (2.23) and then use (2.22) to estimate $\|u\|^2_{L^2(\Omega)}$ and $\|u\|^2_{L^2(\Gamma)}$ in terms of $\|f\|^2_{L^2(\Omega)}$ and $\|g\|^2_{L^2(\Gamma)}$ to get

$$\|\nabla u\|^2_{L^2(\Omega)} + k^2 \|u\|^2_{L^2(\Omega)} \lesssim \left(\frac{k^2}{\varepsilon^2} + \frac{\delta_2}{\varepsilon\eta_R} + \frac{1}{k^2}\right) \|f\|^2_{L^2(\Omega)} + \left(\frac{k^2}{\varepsilon\eta_R} + \frac{\delta_2}{\eta_R^2} + \frac{1}{\delta_2}\right) \|g\|^2_{L^2(\Gamma)} .$$

We then choose $\delta_2 = \eta_R$ (to make $1/\delta_2$ and $\delta_2/\eta_R^2$ equal) and obtain the bound (2.4).

### *2.1.3 Proof of Theorem 2.9*

*Outline.* The proof consists of the following two steps:

1. Use the integrated Morawetz identity (2.18) to show that, given $k_0 > 0$, there exist $c$ and $C$ (independent of $k$, $\eta$, and $\varepsilon$ and $> 0$) such that if $\varepsilon \leq ck$ then

$$\|\nabla u\|^2_{L^2(\Omega)} + k^2 \|u\|^2_{L^2(\Omega)} \leq C \left[(k^2 + |\eta|^2) \|u\|^2_{L^2(\Gamma)} + \|f\|^2_{L^2(\Omega)} + \|g\|^2_{L^2(\Gamma)}\right] \tag{2.24}$$

   for all $k \geq k_0$.
2. Use the imaginary part of Green's identity to remove the $(k^2 + |\eta|^2) \|u\|^2_{L^2(\Gamma)}$ term from the right-hand side of (2.24).

We first prove the bound in Step 2 and then prove the bound in Step 1.

*Step 2.* In the proof of Theorem 2.7 we used the imaginary part of Green's identity to obtain the bound (2.22). We could use (2.22) to bound the $(k^2+|\eta|^2)\|u\|^2_{L^2(\Gamma)}$ term in (2.24) by $\|f\|^2_{L^2(\Omega)}$ and $\|g\|^2_{L^2(\Gamma)}$, however the right-hand side of (2.22) blows up if $\varepsilon \to 0$ and we want to be able to include the case when $\varepsilon = 0$.

The bound (2.22) came from (2.21) with $\delta_1 = \varepsilon$ and $\delta_2 = \eta_R$. If we instead keep $\delta_1$ arbitrary we obtain

$$\frac{\eta_R}{2}\|u\|^2_{L^2(\Gamma)} + \varepsilon\|u\|^2_{L^2(\Omega)} \le \frac{1}{2\delta_1}\|f\|^2_{L^2(\Omega)} + \frac{1}{2\eta_R}\|g\|^2_{L^2(\Gamma)} + \frac{\delta_1}{2}\|u\|^2_{L^2(\Omega)} \tag{2.25}$$

Dropping $\varepsilon\|u\|^2_{L^2(\Omega)}$ from the left-hand side of (2.25) and then using the resulting inequality in (2.24) we obtain

$$\begin{aligned}\|\nabla u\|^2_{L^2(\Omega)} + \left(k^2 - \delta_1 C\frac{k^2+|\eta|^2}{\eta_R}\right)\|u\|^2_{L^2(\Omega)} &\le C\left(1+\frac{k^2+|\eta|^2}{\delta_1\eta_R}\right)\|f\|^2_{L^2(\Omega)} \\ &\quad + C\left(1+\frac{k^2+|\eta|^2}{\eta_R^2}\right)\|g\|^2_{L^2(\Gamma)}\,,\end{aligned} \tag{2.26}$$

for all $k \ge k_0$. If $\eta_R \sim k$ and $|\eta_I| \lesssim k$ then

$$\frac{k^2+|\eta|^2}{\eta_R} \le bk, \text{ for some } b>0, \text{ and } \frac{k^2+|\eta|^2}{(\eta_R)^2} \lesssim 1.$$

Therefore, if we let $\delta_1 = k\theta$ (for some $\theta > 0$) then the right-hand side of (2.26) is $\lesssim \|f\|^2_{L^2(\Omega)} + \|g\|^2_{L^2(\Gamma)}$, which is the right-hand side of (2.7) (with the constant $C$ in (2.7) different to the constant $C$ in (2.26)). The left-hand side of (2.26) is then

$$\ge \|\nabla u\|^2_{L^2(\Omega)} + k^2(1-Cb\theta)\|u\|^2_{L^2(\Omega)}$$

and so choosing $\theta$ less than $1/(Cb)$ gives the result (2.7).

*Step 1.* Remark 2.6 implies that $u$ is in the space $V$ defined by (2.16). Lemma 2.15 then implies that the integrated identity (2.18) holds with $v$ replaced by $u$. Recalling that $\nabla u$ on $\Gamma$ is understood as $\nabla_\Gamma u + \mathbf{n}\partial_n u$, we find that the integral over $\Gamma$ in (2.18) can be rewritten as

$$\int_\Gamma 2\Re\left(\overline{(\mathbf{x}\cdot\nabla_\Gamma u + \alpha u)}\partial_n u\right) + \left(|\partial_n u|^2 + k^2|u|^2 - |\nabla_\Gamma u|^2\right)(\mathbf{x}\cdot\mathbf{n}). \tag{2.27}$$

Therefore, using both (2.27) and that fact that $\mathcal{L}u = -f - \mathrm{i}\varepsilon u$, we can rewrite (2.18) as

$$\begin{aligned}&\int_\Omega (2\alpha+2-d)|\nabla u|^2 + (d-2\alpha)k^2|u|^2 + \int_\Gamma |\nabla_\Gamma u|^2(\mathbf{x}\cdot\mathbf{n}) \\ &\quad = 2\Re\int_\Omega \overline{\mathcal{M}u}\, f - 2\varepsilon\Im\int_\Omega \overline{\mathcal{M}u}\, u + \int_\Gamma 2\Re\left(\overline{(\mathbf{x}\cdot\nabla_\Gamma u+\alpha u)}\partial_n u\right) + \left(|\partial_n u|^2 + k^2|u|^2\right)(\mathbf{x}\cdot\mathbf{n}).\end{aligned} \tag{2.28}$$

We now let

$$\delta_- := \operatorname*{ess\,inf}_{\mathbf{x}\in\Gamma}(\mathbf{x}\cdot\mathbf{n}), \quad \delta_+ := \operatorname*{ess\,sup}_{\mathbf{x}\in\Gamma}(\mathbf{x}\cdot\mathbf{n}), \quad R := \operatorname*{ess\,sup}_{\mathbf{x}\in\Gamma}|\mathbf{x}|,$$

and note that $\delta_+ \ge \delta_- > 0$ since $\Omega$ is assumed to be star-shaped with respect to a ball (see Remark 1.3). Using both the definition of $\mathcal{M}u$ and the Cauchy-Schwarz inequality on the right-hand side of (2.28), and writing the integrals as norms, we obtain that

$$\begin{aligned}&(2\alpha+2-d)\|\nabla u\|^2_{L^2(\Omega)} + (d-2\alpha)k^2\|u\|^2_{L^2(\Omega)} + \delta_-\|\nabla_\Gamma u\|^2_{L^2(\Gamma)} \\ &\quad \le 2R\|\nabla u\|_{L^2(\Omega)}\|f\|_{L^2(\Omega)} + 2\alpha\|u\|_{L^2(\Omega)}\|f\|_{L^2(\Omega)} + 2\varepsilon R\|\nabla u\|_{L^2(\Omega)}\|u\|_{L^2(\Omega)} \\ &\qquad + \delta_+\left(\|\partial_n u\|^2_{L^2(\Gamma)} + k^2\|u\|^2_{L^2(\Gamma)}\right) + 2R\|\nabla_\Gamma u\|_{L^2(\Gamma)}\|\partial_n u\|_{L^2(\Gamma)} + 2\alpha\|u\|_{L^2(\Gamma)}\|\partial_n u\|_{L^2(\Gamma)}\,.\end{aligned}$$

(Note that the boundary condition (2.1b) gives us $\partial_n u$ on $\Gamma$ in terms of $u$ and $g$, but we choose not to use this yet.) Next we let $2\alpha = d-1$ so that the coefficients of both $\|\nabla u\|^2_{L^2(\Omega)}$ and $\|u\|^2_{L^2(\Omega)}$ on the left-hand side become equal to one. We now use (1.21) on each of the terms on the right-hand side (with a different $\delta$ each time) to obtain

$$\left(1 - R\delta_3 - \frac{\varepsilon R}{\delta_5}\right)\|\nabla u\|^2_{L^2(\Omega)} + \left(1 - \frac{(d-1)\delta_4}{2k^2} - \frac{\varepsilon R\delta_5}{k^2}\right)k^2\|u\|^2_{L^2(\Omega)} + (\delta_- - R\delta_6)\|\nabla_\Gamma u\|^2_{L^2(\Gamma)}$$
$$\leq \left(\frac{R}{\delta_3} + \frac{d-1}{2\delta_4}\right)\|f\|^2_{L^2(\Omega)} + \left(\delta_+ + \frac{R}{\delta_6} + \frac{d-1}{2\delta_7}\right)\|\partial_n u\|^2_{L^2(\Gamma)} + \left(\delta_+ + \frac{(d-1)\delta_7}{2k^2}\right)k^2\|u\|^2_{L^2(\Gamma)}. \tag{2.29}$$

To prove the bound (2.24) we need to ensure that a) each bracket on the left-hand side is greater than zero and doesn't grow with $k$, and b) each bracket on the right-hand side does not grow with $k$.

We choose $\delta_7 = 1, \delta_6 = \delta_-/(2R)$ (so that the coefficient of $\|\nabla_\Gamma u\|^2_{L^2(\Gamma)}$ on the left-hand side becomes $\delta_-/2$, which is $> 0$), $\delta_4 = k^2/(d-1)$, and $\delta_3 = 1/(2R)$. With these choices, and neglecting the term involving $\|\nabla_\Gamma u\|^2_{L^2(\Gamma)}$ on the left-hand side, we obtain from (2.29) the bound

$$\left(\frac{1}{2} - \frac{\varepsilon R}{\delta_5}\right)\|\nabla u\|^2_{L^2(\Omega)} + \left(\frac{1}{2} - \frac{\varepsilon R\delta_5}{k^2}\right)k^2\|u\|^2_{L^2(\Omega)}$$
$$\leq C'\left(\|\partial_n u\|^2_{L^2(\Gamma)} + \left(1 + \frac{1}{k^2}\right)\left(k^2\|u\|^2_{L^2(\Gamma)} + \|f\|^2_{L^2(\Omega)}\right)\right), \tag{2.30}$$

for some $C' > 0$ (independent of $k$, $\eta$ and $\varepsilon$). The right-hand side of (2.30) is bounded above by

$$C''\left(\|g\|^2_{L^2(\Gamma)} + \left(1 + k^2 + |\eta|^2\right)\|u\|^2_{L^2(\Gamma)} + \left(1 + \frac{1}{k^2}\right)\|f\|^2_{L^2(\Omega)}\right)$$

for some $C'' > 0$ (again independent of $k$, $\eta$ and $\varepsilon$), since the boundary condition (2.1b) and the inequality (1.21) imply that

$$\|\partial_n u\|^2_{L^2(\Gamma)} \leq 2\Big(|\eta|^2\|u\|^2_{L^2(\Gamma)} + \|g\|^2_{L^2(\Gamma)}\Big).$$

Also, given any $k_0 > 0$, there exists a $C''' > 0$ independent of $k$ such that

$$\left(1 + \frac{1}{k^2}\right)\|f\|^2_{L^2(\Omega)} \leq C'''\|f\|^2_{L^2(\Omega)} \qquad \text{for all } k \geq k_0.$$

Therefore, to establish (2.24) we only need to show that the coefficients of $\|\nabla u\|^2_{L^2(\Omega)}$ and $k^2\|u\|^2_{L^2(\Omega)}$ on the left-hand side of (2.30) are bounded away from zero, independently of $k$. If $\varepsilon = 0$ this is immediately true. If $\varepsilon \neq 0$ we choose $\delta_5 = 4\varepsilon R$. The left-hand side of (2.30) then becomes

$$\frac{1}{4}\|\nabla u\|^2_{L^2(\Omega)} + \left(\frac{1}{2} - \frac{4R^2\varepsilon^2}{k^2}\right)k^2\|u\|^2_{L^2(\Omega)}.$$

If $\varepsilon/k \leq 1/(4R)$ then this last expression is

$$\geq \frac{1}{4}\left(\|\nabla u\|^2_{L^2(\Omega)} + k^2\|u\|^2_{L^2(\Omega)}\right)$$

and we are done.

**Remark 2.16** *The earlier proofs of the bound (2.7) when $\varepsilon = 0$ discussed in Remark 2.10 use essentially the same method that we use here, except that to get (2.24) they apply the Rellich (2.14) and Green (2.13) identities separately and then take the particular linear combination that corresponds to the Morawetz identity (2.15) with $2\alpha = d-1$ (whereas we use the Morawetz identity with $2\alpha = d-1$ directly). (In addition, these earlier proofs only consider the case when $\Gamma$ is piecewise smooth, and not Lipschitz.)*

*In the next subsection we obtain the analogue of the bound of Theorem 2.9 for the truncated sound-soft scattering problem (see Theorem 2.18). For the case $\varepsilon = 0$, this bound was obtained in [26, Proposition 3.3] using essentially the same method as we do (but again using a combination of the Rellich and Green identities that is equivalent to using the Morawetz identity).*

### 2.2 Bounds on the truncated sound-soft scattering problem

The following are analogues of Theorems 2.7 and 2.9 for Problem 2.4.

**Theorem 2.17 (Bound for $\varepsilon > 0$ for Lipschitz $\Omega_D$ and $\Omega_R$)** *Let $u$ be the solution of Problem 2.4, and let $\eta = \eta_R + \mathrm{i}\eta_I$ with $\eta_I \geq 0$, $\eta_R > 0$. Then, given $k_0 > 0$, there exists a $C > 0$, independent of $\varepsilon$, $k$, $\eta_R$ and $\eta_I$, such that*

$$\|\nabla u\|^2_{L^2(\Omega)} + k^2 \|u\|^2_{L^2(\Omega)} \leq C \left[ \frac{k^2}{\varepsilon^2} \left( 1 + \frac{\varepsilon}{k^2} + \left( \frac{\varepsilon}{k^2} \right)^2 \right) \|f\|^2_{L^2(\Omega)} + \frac{k^2}{\varepsilon \eta_R} \left( 1 + \frac{\varepsilon}{k^2} \right) \|g\|^2_{L^2(\Gamma)} \right]$$

*for all $k \geq k_0$, $\eta_R > 0$, and $\varepsilon > 0$.*

**Theorem 2.18 (Bound for $\varepsilon/k$ sufficiently small when $\Omega_R$ and $\Omega_D$ are star-shaped)** *Let $u$ be the solution to Problem 2.4 and assume that $\Omega_R$ is star-shaped with respect to a ball centred at the origin and $\Omega_D$ is star-shaped with respect to the origin, i.e.*

$$\operatorname*{ess\,inf}_{\mathbf{x} \in \Gamma_D} (\mathbf{x} \cdot \mathbf{n}_D) \geq 0 \quad \textit{and} \quad \operatorname*{ess\,inf}_{\mathbf{x} \in \Gamma_R} (\mathbf{x} \cdot \mathbf{n}_R) > 0, \tag{2.31}$$

*where $\mathbf{n}_D$ and $\mathbf{n}_R$ are the unit normal vectors to $\Omega_D$ and $\Omega_R$ respectively (oriented as in Figure 2). If $\eta_R \sim k$ and $|\eta_I| \lesssim k$, then, given $k_0 > 0$, there exist $c$ and $C$ (independent of $k$, $\eta$, and $\varepsilon$ and $> 0$) such that, if $\varepsilon/k \leq c$ for all $k \geq k_0$, then*

$$\|\nabla u\|^2_{L^2(\Omega)} + k^2 \|u\|^2_{L^2(\Omega)} \leq C \left[ \|f\|^2_{L^2(\Omega)} + \|g\|^2_{L^2(\Gamma)} \right]$$

*for all $k \geq k_0$.*

*Proof of Theorem 2.17.* This follows the proof of Theorem 2.7 exactly. Indeed, the starting point of Theorem 2.7 was (2.19) (the integrated form of Green's identity with the PDE and boundary conditions imposed on $u$), and this holds for the truncated sound-soft scattering problem with $\Gamma$ replaced by $\Gamma_R$ (since the integral over $\Gamma_D$ that arises when Green's identity is applied in $\Omega$ is zero as $u = 0$ on $\Gamma_D$). ∎

*Proof of Theorem 2.18.* This follows the proof of Theorem 2.9 exactly. Indeed, Step 2 is the same since it depends on Green's identity. For Step 1, we note that applying the integrated Morawetz identity (2.18) in $\Omega$ yields (2.28) with $\Gamma$ replaced by $\Gamma_R$, and the additional term $\int_{\Gamma_D} (\mathbf{x} \cdot \mathbf{n}_D) |\partial_n u|^2$ on the left-hand side. By (2.31), this additional term is non-negative, and the proof proceeds as before. ∎

## 3 Variational formulations and quasi-optimality

In §3.1 we prove results about the continuity and coercivity of $a_\varepsilon(\cdot,\cdot)$, and then we use these in §3.2–§3.4 to obtain sufficient conditions for quasi-optimality of the Galerkin method applied to $a_\varepsilon(\cdot,\cdot)$. In §3.1-3.4 we consider the interior impedance problem, and then in §3.5 we outline the small modifications needed to extend the results to the truncated scattering problem.

### 3.1 Continuity and coercivity of $a_\varepsilon(\cdot,\cdot)$

Recall from §1 the variational formulation of the shifted interior impedance problem (1.3) and its Galerkin approximation (1.6). Define a norm on $H^1(\Omega)$ by

$$\|v\|^2_{1,k,\Omega} := \|\nabla v\|^2_{L^2(\Omega)} + k^2 \|v\|^2_{L^2(\Omega)}\,; \tag{3.1}$$

in what follows we always have $k \geq k_0$ for some $k_0 > 0$ and thus $\|\cdot\|_{1,k,\Omega}$ is indeed a norm and is equivalent to the usual $H^1$-norm.

**Lemma 3.1 (Continuity and coercivity of $a_\varepsilon(\cdot,\cdot)$)**

*(i) If* $|\eta| \lesssim k$ *and* $\varepsilon \lesssim k^2$ *then, given* $k_0 > 0$*, there exists a* $C_c$ *(independent of* $k$*,* $\eta$*, and* $\varepsilon$*) such that*

$$\left|a_\varepsilon(u,v)\right| \le C_c \, \|u\|_{1,k,\Omega} \, \|v\|_{1,k,\Omega}$$

*for all* $k \ge k_0$ *and* $u, v \in H^1(\Omega)$*.*

*(ii) If* $\eta_R$ *and* $\eta_I$ *are both* $\ge 0$ *and* $0 < \varepsilon \lesssim k^2$*, then there exists a constant* $\alpha > 0$ *(independent of* $k$*,* $\eta$*, and* $\varepsilon$*) such that*

$$\left|a_\varepsilon(v,v)\right| \ge \alpha \, \frac{\varepsilon}{k^2} \, \|v\|^2_{1,k,\Omega} \tag{3.2}$$

*for all* $k > 0$ *and* $v \in H^1(\Omega)$*.*

*Proof.* (i) This follows from the Cauchy-Schwarz inequality and the multiplicative trace inequality (1.20).
(ii) Given $k > 0$ and $\varepsilon > 0$, define $p > 0$ and $q > 0$ by

$$p^2 := \frac{k^2 + \sqrt{k^4 + \varepsilon^2}}{2} \quad \text{and} \quad q^2 := \frac{-k^2 + \sqrt{k^4 + \varepsilon^2}}{2},$$

so that $k^2 + \mathrm{i}\varepsilon = (p + \mathrm{i}q)^2$. The definition of $p$ and the fact that $\varepsilon \lesssim k^2$ mean that $k \le p \lesssim k$, and the fact that $2qp = \varepsilon$ then implies that $q \sim \varepsilon/k$. Now

$$a_\varepsilon(v,v) = \|\nabla v\|^2_{L^2(\Omega)} - (p + \mathrm{i}q)^2 \, \|v\|^2_{L^2(\Omega)} - \mathrm{i}\eta \, \|v\|^2_{L^2(\Gamma)},$$

and so

$$(p - \mathrm{i}q) a_\varepsilon(v,v) = (p - \mathrm{i}q) \, \|\nabla v\|^2_{L^2(\Omega)} - (p + \mathrm{i}q)(p^2 + q^2) \, \|v\|^2_{L^2(\Omega)} - \mathrm{i}(p - \mathrm{i}q)\eta \, \|v\|^2_{L^2(\Gamma)}. \tag{3.3}$$

Therefore, taking the imaginary part of each side of (3.3), we have

$$\Im\big[-(p - iq) a_\varepsilon(v,v)\big] = q\left[\|\nabla v\|^2_{L^2(\Omega)} + (p^2 + q^2) \, \|v\|^2_{L^2(\Omega)}\right] + (p\eta_R + q\eta_I)\|v\|^2_{L^2(\Gamma)}.$$

Now, defining $\Theta := -(p - iq)/|p - iq| = -(p - iq)/\sqrt{p^2 + q^2}$, and using the fact that $\eta_R$ and $\eta_I$ are both $\ge 0$ we have

$$\begin{aligned}\left|a_\varepsilon(v,v)\right| = \left|\Theta a_\varepsilon(v,v)\right| &\ge \Im\big[\Theta a_\varepsilon(v,v)\big] \\ &\ge \frac{q}{\sqrt{p^2 + q^2}}\left[\|\nabla v\|^2_{L^2(\Omega)} + (p^2 + q^2) \, \|v\|^2_{L^2(\Omega)}\right].\end{aligned}$$

The result (3.2) follows since $p \sim k$, $q \sim \varepsilon/k$, and $\varepsilon \lesssim k^2$. ∎

**Remark 3.2** *This "trick" of multiplying the sesquilinear form by the complex conjugate of the wavenumber (in the proof above this was* $p - \mathrm{i}q$*) is well known in, for example, the time-domain boundary-integral-equation literature; see [23, Proposition 1].*

Note that (with $\varepsilon \lesssim k^2$) the bound (3.2) is sharp in its $k$- and $\varepsilon$-dependence. Indeed, if $u_j$ is a Dirichlet eigenfunction of $-\Delta$ on $\Omega$ with eigenvalue $\lambda_j$, then

$$\frac{a_\varepsilon(u_j,u_j)}{\|u_j\|^2_{1,k,\Omega}} = \frac{\lambda_j - (k^2 + \mathrm{i}\varepsilon)}{\lambda_j + k^2}.$$

Therefore, if $k = k_j := \sqrt{\lambda_j}$ then

$$\frac{a_\varepsilon(u_j,u_j)}{\|u_j\|^2_{1,k_j,\Omega}} = \frac{-\mathrm{i}\varepsilon}{2k_j^2} \sim \frac{\varepsilon}{k_j^2}.$$

### 3.2 Abstract conditions for quasi-optimality

To state the main result of this section, we need to introduce the solution operator of the adjoint problem. Given $f \in L^2(\Omega)$, define $S^*_{k,\varepsilon} f$ as the solution of the variational problem

$$a_\varepsilon(v, S^*_{k,\varepsilon} f) = (v, f)_{L^2(\Omega)} \quad \text{for all } v \in H^1(\Omega). \tag{3.4}$$

This is the variational formulation of the adjoint problem (2.3) with $g = 0$; i.e. if $w = S^*_{k,\varepsilon} f$ satisfies (3.4) then $w$ is a solution of the weak form of (2.3) with $g = 0$, and vice versa.

**Lemma 3.3 (Quasi-optimality for $a_\varepsilon(\cdot,\cdot)$)** *Assume that $\varepsilon \lesssim k^2$ and $|\eta| \lesssim k$. Let $C_c$ and $\alpha$ be the constants in Lemma 3.1. Let $u_N$ denote the Galerkin solution defined by (1.6). Let*

$$\eta(V_N) := \sup_{f \in L^2(\Omega)} \inf_{v_N \in V_N} \frac{\|S^*_{k,\varepsilon} f - v_N\|_{1,k,\Omega}}{\|f\|_{L^2(\Omega)}}. \tag{3.5}$$

*If*

$$\sqrt{|k^2 - \varepsilon|}\, C_c\, \eta(V_N) \le \sqrt{\frac{\alpha}{2}} \tag{3.6}$$

*then*

$$\|u - u_N\|_{1,k,\Omega} \le \frac{2C_c}{\alpha} \inf_{v_N \in V_N} \|u - v_N\|_{1,k,\Omega}\,. \tag{3.7}$$

The analogue of this result for the Helmholtz equation (i.e. $\varepsilon = 0$) first appeared in the form above as [46, Theorem 2.5], although the argument goes back to Schatz [47] and has been used by several authors since then (see, e.g., the discussion in [19, §4] and the references therein). The only difference in our use of this argument (compared to previous uses) is that, instead of using the fact that $a_\varepsilon(\cdot,\cdot)$ satisfies a Gårding inequality, we use the fact that when $\varepsilon = k^2$ it is coercive with constant independent of $k$. Note that $\eta(V_N)$ in (3.5) is not related to the $\eta$ in the impedance boundary condition (2.1b); we use this notation to be consistent with the other uses of this argument in the literature.

*Proof.* We first prove the bound (3.7) under the assumption that $u_N$ exists. Choosing $v = v_N \in V_N$ in (1.3) and subtracting this from (1.6), we have Galerkin orthogonality:

$$a_\varepsilon(u - u_N, v_N) = 0 \quad \text{for all } v_N \in V_N. \tag{3.8}$$

Coercivity (3.2) and the triangle inequality imply that, for any $v \in H^1(\Omega)$,

$$\alpha \|v\|^2_{1,k,\Omega} \le \left|a_{k^2}(v,v)\right| \le \left|a_\varepsilon(v,v) - \mathrm{i}(k^2 - \varepsilon)\|v\|^2_{L^2(\Omega)}\right| \le \left|a_\varepsilon(v,v)\right| + \left|k^2 - \varepsilon\right| \|v\|^2_{L^2(\Omega)}\,. \tag{3.9}$$

We now apply this last inequality with $v = e_N := u - u_N$ and use that fact that, by Galerkin orthogonality, $a(e_N, e_N) = a(e_N, u - v_N)$ for any $v_N \in V_N$. This yields

$$\alpha \|e_N\|^2_{1,k,\Omega} \le \left|a_\varepsilon(e_N, u - v_N)\right| + \left|k^2 - \varepsilon\right| \|e_N\|^2_{L^2(\Omega)} \tag{3.10}$$

$$\le C_c \|e_N\|_{1,k,\Omega} \|u - v_N\|_{1,k,\Omega} + \left|k^2 - \varepsilon\right| \|e_N\|^2_{L^2(\Omega)} \tag{3.11}$$

(where we have used the continuity of $a_\varepsilon(\cdot,\cdot)$ to obtain the second inequality). If we can show that

$$\left|k^2 - \varepsilon\right| \|e_N\|^2_{L^2(\Omega)} \le \frac{\alpha}{2} \|e_N\|^2_{1,k,\Omega} \tag{3.12}$$

then we obtain the result (3.7).

Now, using the definition of $S^*_{k,\varepsilon}$ (3.4), Galerkin orthogonality (3.8), continuity of $a_\varepsilon(\cdot,\cdot)$, and the definition of $\eta(V_N)$ (3.5), we have

$$\|e_N\|^2_{L^2(\Omega)} = a_\varepsilon(e_N, S^*_{k,\varepsilon} e_N) = a_\varepsilon(e_N, S^*_{k,\varepsilon} e_N - v_N) \le C_c \|e_N\|_{1,k,\Omega} \left(\eta(V_N) \|e_N\|_{L^2(\Omega)}\right),$$

for any $v_N \in V_N$. Therefore

$$\|e_N\|_{L^2(\Omega)} \le C_c\, \eta(V_N) \|e_N\|_{1,k,\Omega}\,, \tag{3.13}$$

and the condition (3.6) is sufficient to ensure that (3.12) holds.

Up to now, we have assumed that $u_N$ (the solution of the variational problem (1.6)) exists. The fact that $u_N$ exists can be established using [33, Theorem 3.9], but here we follow the simpler approach found in, e.g., [4, Theorem 5.7.6]. Since (1.6) is a system of $N$ equations with $N$ unknowns, existence for all right-hand sides is equivalent to uniqueness. Therefore, we only need to show that if $F = 0$ and $N$ is such that the condition (3.6) holds, then (1.6) only has the trivial solution $u_N = 0$. Seeking a contradiction, suppose that $a_\varepsilon(u_N, v_N) = 0$ for all $v_N \in V_N$ for some $u_N \neq 0$. Remark 2.2 implies that $u = 0$, and then (3.7) implies that $u_N = 0$ when $N$ is such that (3.6) holds. Therefore, the solution to (1.6) exists and is unique when $N$ satisfies (3.7). ∎

**Remark 3.4 (Using coercivity for $\varepsilon = \gamma k^2$, for some $\gamma > 0$, instead of for $\varepsilon = k^2$.)** *In the proof of Lemma 3.3 we used the coercivity of $a_{k^2}(\cdot,\cdot)$. Instead, we could have used the coercivity of $a_{\gamma k^2}(\cdot,\cdot)$, with $\gamma$ any positive constant. If we had done this, then the mesh threshold for quasi-optimality would be*

$$\sqrt{|\gamma k^2 - \varepsilon|}\, C_c\, \eta(V_N) \leq \sqrt{\frac{\gamma\alpha}{2}} \tag{3.14}$$

*and the constant of quasi-optimality in (3.7) would be $2C_c/(\gamma\alpha)$.*

### 3.3 Quasi-optimality: smooth domains and convex polygons

In this subsection we consider the case when $\Omega$ is *either* a $C^{1,1}$ 2- or 3-d domain that is star-shaped with respect to a ball *or* a convex polygon. We also assume that $V_N$ has the property that, for all $w \in H^2(\Omega)$,

$$\inf_{v_N \in V_N} \|w - v_N\|_{1,k,\Omega} \lesssim h\, \|w\|_{H^2(\Omega)} + hk\, \|w\|_{H^1(\Omega)}\,; \tag{3.15}$$

this is true, for example, for continuous piecewise-polynomial elements on a triangular mesh by properties of the quasi-interpolant given in [48, Theorem 4.1].

We now use Lemma 3.3 to prove the following result.

**Lemma 3.5 (Quasi-optimality for $a_\varepsilon(\cdot,\cdot)$ for smooth domains and convex polygons)** *Suppose that the variational problem (1.3) is solved using the Galerkin method with $V_N \subset H^1(\Omega)$. Assume that $\varepsilon \lesssim k^2$, $\eta_R \sim k$, and $\eta_I \lesssim k$. Then, given $k_0 > 0$, there exists $C_1 > 0$ (with $C_1$ independent of $h, k$, and $\varepsilon$) such that if $k \geq k_0$ and*

$$hk\sqrt{|k^2 - \varepsilon|} \leq C_1 \tag{3.16}$$

*then (3.7) holds.*

*Proof.* Given $f \in L^2(\Omega)$, let $w := S^*_{k,\varepsilon} f$. By Remark 2.5, given $k_0 > 0$, $\|w\|_{H^1(\Omega)} \lesssim \|f\|_{L^2(\Omega)}$ for all $k \geq k_0$. Moreover, Lemma 2.12 then implies that $\|w\|_{H^2(\Omega)} \lesssim k\|f\|_{L^2(\Omega)}$ for all $k \geq k_0$. Combining these bounds with (3.15) yields

$$\inf_{v_N \in V_N} \|w - v_N\|_{1,k,\Omega} \lesssim hk\, \|f\|_{L^2(\Omega)}\,.$$

Therefore, from the definition of $\eta(V_N)$,

$$\sqrt{|k^2 - \varepsilon|}\, C_c\, \eta(V_N) \leq C_c hk \sqrt{|k^2 - \varepsilon|},$$

and the result follows from Lemma 3.3. ∎

**Remark 3.6** *For arbitrary curved $C^{1,1}$ domains it is not always possible to fit the domain boundary exactly with polynomial elements, and some analysis of non-conforming error is then necessary; since this is very standard, we do not give it here.*

### 3.4 Quasi-optimality: non-smooth domains

In obtaining Lemma 3.5 from Lemma 3.3 we used a bound on the $H^2$-norm of the solution of the adjoint problem to estimate $\eta(V_N)$ and get a mesh-threshold for quasi-optimality. We now consider domains in which the solution to the adjoint problem is not in $H^2(\Omega)$. In this case we can still estimate $\eta(V_N)$ (and thus get conditions for quasi-optimality) under assumptions on the solution and the mesh that we now explain.

**Assumption 3.7** *Let $\Omega$ be a bounded Lipschitz polyhedron in $\mathbb{R}^d$ $(d = 2, 3)$.*

1. *Let $w = S^*_{k,\varepsilon} f$ and let $C_{\rm sol}(k,\varepsilon)$ be such that*

$$\|w\|_{1,k,\Omega} \lesssim C_{\rm sol}(k,\varepsilon)\, \|f\|_{L^2(\Omega)} \tag{3.17}$$

*for all $f \in L^2(\Omega)$ and for all $0 \le \varepsilon \lesssim k^2$. Assume that there exists a weight function $\Phi \in C(\overline{\Omega})$ such that, for any $f \in L^2(\Omega)$,*

$$\sup_{|\alpha|=2} \|\Phi D^\alpha w\|_{L^2(\Omega)} \lesssim k\, C_{\rm sol}(k,\varepsilon)\, \|f\|_{L^2(\Omega)}\,. \tag{3.18}$$

2. *With $\Phi$ as in Part 1, assume that if $v \in H^1(\Omega)$ and $\sup_{\alpha=2} \|\Phi D^\alpha v\|_{L^2(\Omega)} < \infty$ then there exists a shape-regular simplicial mesh sequence so that the corresponding finite element space $V_N$ has dimension $N$, has largest element diameter $(1/N)^{1/d}$, and satisfies*

$$\inf_{v_N \in V_N} \left\{ \left(\frac{1}{N}\right)^{1/d} |v - v_N|_{H^1(\Omega)} + \|v - v_N\|_{L^2(\Omega)} \right\} \lesssim \left(\frac{1}{N}\right)^{2/d} \sup_{|\alpha|=2} \|\Phi D^\alpha v\|_{L^2(\Omega)}\,. \tag{3.19}$$

**Remark 3.8** *Part 2 of Assumption 3.7 holds by results in [1], and Part 1 of Assumption 3.7 holds when $\Omega$ is a polygon in $\mathbb{R}^2$ and $\varepsilon = 0$ by [19, Theorem 3.2] (and we expect similar arguments to apply when $0 < \varepsilon \lesssim k^2$). We now discuss both these sets of results when $\Omega$ is a polygon. The result [19, Theorem 3.2] proves that there exists a weight function $\Phi \in C(\overline{\Omega})$ such that the soluton $u = S^*_{k,0} f$ of (2.3) with $f \in L^2(\Omega)$, $g = 0$, and $\varepsilon = 0$ has a decomposition $u = u_{H^2} + u_{\mathcal{A}}$, where*

$$\|u_{H^2}\|_{H^2(\Omega)} \lesssim C_{\rm sol}(k,0)\, \|f\|_{L^2(\Omega)}\,, \tag{3.20}$$

$$\sum_{|\alpha|=2} \|\Phi D^\alpha u_{\mathcal{A}}\|_{L^2(\Omega)} \lesssim k\, C_{\rm sol}(k,0)\, \|f\|_{L^2(\Omega)}\,. \tag{3.21}$$

*The weight function $\Phi$ can be taken to be one at convex corners, but at a non-convex corner $\Phi(\mathbf{x}) \sim r^\beta$ as $r \to 0$ for $\mathbf{x}$ in a neighbourhood of a corner point $\mathbf{x}_0$ with exterior angle $\omega$, where $r := |\mathbf{x} - \mathbf{x}_0|$ and $\beta > 1 - \pi/\omega$; see [19, Equation (24) and Lemma 3.11] (this decay of the weight function compensates for singularities in the second derivatives of $u_{\mathcal{A}}$). The subsequent verification of (3.19) can be obtained from several references, e.g. [1] and the references therein. Indeed, the existence of a suitably refined shape-regular mesh and corresponding $v_N \in V_N$ satisfying*

$$|v - v_N|_{H^1(\Omega)} \lesssim \left(\frac{1}{N}\right)^{1/d} \sup_{|\alpha|=2} \|\Phi D^\alpha v\|_{L^2(\Omega)} \tag{3.22}$$

*follows from [1, Theorems 3.2 and 3.3] and particularly the estimate [1, Equation (3.19)]. Note that [1] uses very different notation to ours; the weighted norm on the right-hand side of [1, Equation (3.19)] coincides with that on the right-hand side of our equation (3.19), and $H_0$ in [1] equals $\pi/\omega$ in our notation (the statement that $H_0 = \pi/(2\omega_0)$ on [1, Page 68] is a typo). The required complexity of the mesh follows from the discussion in [1, Remark 3.1] and the shape-regularity is [1, Condition (d) on Page 71]. The estimate on $\|v - v_N\|_{L^2(\Omega)}$ in (3.19) is not proved explicitly in [1] but follows using similar arguments.*

**Remark 3.9 (How does $C_{\rm sol}(k,\varepsilon)$ depend on $k$ and $\varepsilon$?)** *By combining Theorems 2.7 and 2.9 (and using Remark 2.5) we see that if $\Omega$ is Lipschitz and star-shaped with respect to a ball, $0 \le \varepsilon \lesssim k^2$, $\eta_R \sim k$, and $0 < \eta_I \lesssim k$, then (3.17) holds with $C_{\rm sol}(k,\varepsilon) \sim 1$; in what follows we only consider this situation.*

The following is the analogue of Lemma 3.5 for non-smooth domains.

**Lemma 3.10 (Quasi-optimality for $a_\varepsilon(\cdot,\cdot)$ for non-smooth domains)** *Suppose that $\Omega$ is such that Assumption 3.7 holds with $C_{\rm sol}(k,\varepsilon)\sim 1$, and suppose that the variational problem (1.3) is solved using the Galerkin method in the space $V_N$. If $\varepsilon \lesssim k^2$, $\eta_R \sim k$, and $\eta_I \lesssim k$ then, given $k_0>0$, there exists a $C_1>0$ (independent of $N,k$, and $\varepsilon$) such that, if $k\geq k_0$ and*

$$N^{-1/d}k\sqrt{|k^2-\varepsilon|}\leq C_1, \tag{3.23}$$

*then (3.7) holds.*

*Proof.* By Lemma 3.3 we only need to estimate $\eta(V_N)$ and ensure that (3.6) holds. With $w=S_k^*f$, Assumption 3.7 implies that there exists a $v_N\in V_N$ such that

$$|w-v_N|_{H^1(\Omega)}\lesssim \left(\frac{1}{N}\right)^{1/d} k\,\|f\|_{L^2(\Omega)} \quad\text{and}\quad \|w-v_N\|_{L^2(\Omega)}\lesssim \left(\frac{1}{N}\right)^{2/d} k\,\|f\|_{L^2(\Omega)}\ ,$$

from which it follows that

$$\|w-v_N\|_{1,k,\Omega}\lesssim \left(\frac{1}{N}\right)^{1/d} k\left[1+\left(\frac{1}{N}\right)^{1/d}k\right]\|f\|_{L^2(\Omega)}\,.$$

Therefore,

$$\sqrt{|k^2-\varepsilon|}\,\eta(V_N)\lesssim \left(\frac{1}{N}\right)^{1/d} k\sqrt{|k^2-\varepsilon|}\left[1+\left(\frac{1}{N}\right)^{1/d}k\right],$$

and this implies that, given $k_0>0$, there exists a $C_1>0$ such that the condition (3.23) is sufficient to ensure that (3.6) holds. ∎

### 3.5 The truncated sound-soft scattering problem with absorption

The variational formulation of Problem 2.4 is almost identical to that of Problem 2.1 except that the Hilbert space is now $V=\{v\in H^1(\Omega): v=0 \text{ on } \Gamma_D\}$, and the integrals over $\Gamma$ in $a_\varepsilon(\cdot,\cdot)$ and $F(\cdot)$ defined in (1.4) and (1.5) respectively are replaced by integrals over $\Gamma_R$. Lemma 3.1 (continuity and coercivity of $a_\varepsilon(\cdot,\cdot)$) holds as before. Lemma 3.5 holds if $\Omega$ is $C^{1,1}$ and satisfies the geometric assumptions in Theorem 2.18. Similarly, if Assumption 3.7 is satisfied with $\Omega=\Omega_R\setminus\overline{\Omega_D}$ and $\Omega_R$ and $\Omega_D$ are as in Theorem 2.18, then Lemma 3.10 holds.

## 4 Proofs of Theorem 1.4 and its analogue for non-quasi-uniform meshes

In §4.1 we consider the interior impedance problem, and in §4.2 we consider the truncated sound-soft scattering problem.

### 4.1 Results about the interior impedance problem

#### *4.1.1 Smooth domains and quasi-uniform meshes (i.e. Proof of Theorem 1.4)*

As discussed in §1.3, we prove Theorem 1.4 by obtaining bounds on $\|\mathbf{A}_\varepsilon^{-1}\mathbf{M}\|_2$, $\|\mathbf{A}_\varepsilon^{-1}\mathbf{N}\|_2$, $\|\mathbf{M}\mathbf{A}_\varepsilon^{-1}\|_2$ and $\|\mathbf{N}\mathbf{A}_\varepsilon^{-1}\|_2$.

**Lemma 4.1** *Under the same conditions as in Theorem 1.4, given $k_0>0$, there exist $C_1,C_2>0, C_3>0$ (independent of $h,k$, and $\varepsilon$ but depending on $k_0$) such that if $hk\sqrt{|k^2-\varepsilon|}\leq C_1$ then*

$$(i)\quad \max\left\{\left\|\mathbf{A}_\varepsilon^{-1}\mathbf{M}\right\|_2,\left\|\mathbf{M}\mathbf{A}_\varepsilon^{-1}\right\|_2\right\}\leq\frac{C_2}{k}\quad\text{and}\quad (ii)\quad \max\left\{\left\|\mathbf{A}_\varepsilon^{-1}\mathbf{N}\right\|_2,\left\|\mathbf{N}\mathbf{A}_\varepsilon^{-1}\right\|_2\right\}\leq\frac{C_3}{h^{1/2}k}. \tag{4.1}$$

*Proof of Lemma 4.1.* We prove below the estimates on $\|\mathbf{A}_\varepsilon^{-1}\mathbf{M}\|_2$ and $\|\mathbf{A}_\varepsilon^{-1}\mathbf{N}\|_2$; the analogous estimates for $\|\mathbf{M}\mathbf{A}_\varepsilon^{-1}\|_2$ and $\|\mathbf{N}\mathbf{A}_\varepsilon^{-1}\|_2$ are obtained as outlined in Remark 1.7 (and also using the fact that a result analogous to Lemma 3.5 holds for the adjoint problem).

Given $v_N \in V_N$, let $\mathbf{v}$ denote the vector of the nodal values of $v_N$. A standard scaling argument for the mass matrix $\mathbf{M}$ yields

$$\|v_N\|^2_{L^2(\Omega)} = (\mathbf{M}\mathbf{v},\mathbf{v})_2 \sim h^d \|\mathbf{v}\|_2^2 . \tag{4.2}$$

Therefore,

$$h^d k^2 \|\mathbf{v}\|_2^2 \sim k^2 \|v_N\|^2_{L^2(\Omega)} \lesssim \|v_N\|^2_{1,k,\Omega} . \tag{4.3}$$

We first prove the bound (i) in (4.1) (i.e. the bound on $\|\mathbf{A}_\varepsilon^{-1}\mathbf{M}\|_2$). Given $\mathbf{f} \in \mathbb{C}^N$, we create a variational problem whose Galerkin discretisation leads to the equation $\mathbf{A}_\varepsilon\widetilde{\mathbf{u}} = \mathbf{M}\,\mathbf{f}$. Indeed, let $\widetilde{f} := \sum_j f_j\phi_j$ and note that $\widetilde{f} \in L^2(\Omega)$. Define $\widetilde{u}$ to be the solution of the variational problem

$$a_\varepsilon(\widetilde{u},v) = (\widetilde{f},v)_{L^2(\Omega)} \quad \text{for all } v \in H^1(\Omega), \tag{4.4}$$

let $\widetilde{u}_N$ be the solution of the finite element approximation of (4.4), i.e.,

$$a_\varepsilon(\widetilde{u}_N,v_N) = (\widetilde{f},v_N)_{L^2(\Omega)} \quad \text{for all } v_N \in V_N, \tag{4.5}$$

and let $\widetilde{\mathbf{u}}$ be the vector of nodal values of $\widetilde{u}_N$. The definition of $\widetilde{f}$ then implies that (4.5) is equivalent to $\mathbf{A}_\varepsilon\widetilde{\mathbf{u}} = \mathbf{M}\,\mathbf{f}$, and so to obtain a bound on $\|\mathbf{A}_\varepsilon^{-1}\mathbf{M}\|_2$ we need to bound $\|\widetilde{\mathbf{u}}\|_2$ in terms of $\|\mathbf{f}\|_2$. Note that the hypotheses imply that the bound on the solution operator (2.8) holds (by Corollary 2.11), and also that if $hk\sqrt{|k^2-\varepsilon|} \le C_1$ then quasi-optimality (3.7) holds (by Lemma 3.5). Starting with (4.3) we then have

$$\begin{aligned} h^{d/2}k\,\|\widetilde{\mathbf{u}}\|_2 \lesssim \|\widetilde{u}_N\|_{1,k,\Omega} &\le \|\widetilde{u}-\widetilde{u}_N\|_{1,k,\Omega} + \|\widetilde{u}\|_{1,k,\Omega} \\ &\lesssim \|\widetilde{u}\|_{1,k,\Omega} + \|\widetilde{u}\|_{1,k,\Omega} \quad \text{(by quasi-optimality)}, \\ &\lesssim \|\widetilde{f}\|_{L^2(\Omega)} \quad \text{(using the bound on the solution operator)}. \end{aligned} \tag{4.6}$$

Finally, (4.2) implies that $\|\widetilde{f}\|_{L^2(\Omega)} \sim h^{d/2}\|\mathbf{f}\|_2$, and using this in (4.6) yields $\|\widetilde{\mathbf{u}}\|_2 \lesssim k^{-1}\|\mathbf{f}\|_2$, which implies the bound (i) in (4.1).

To prove the bound (ii) in (4.1), given $\mathbf{g} \in \mathbb{C}^N$ we create a variational problem whose Galerkin discretisation leads to the equation $\mathbf{A}_\varepsilon\widetilde{\mathbf{u}} = \mathbf{N}\,\mathbf{g}$. Indeed, let

$$\widetilde{g} := \sum_{j\,:\,x_j\in\Gamma} g_j\phi_j,$$

where $x_j$ is the $j$th node of the mesh (note that $\widetilde{g} \in L^2(\Gamma)$). Define $\widetilde{u}$ to be the solution of the variational problem

$$a_\varepsilon(\widetilde{u},v) = (\widetilde{g},v)_{L^2(\Gamma)} \quad \text{for all } v \in H^1(\Omega), \tag{4.7}$$

let $\widetilde{u}_N$ be the solution of

$$a_\varepsilon(\widetilde{u}_N,v_N) = (\widetilde{g},v_N)_{L^2(\Gamma)} \quad \text{for all } v_N \in V_N, \tag{4.8}$$

and let $\widetilde{\mathbf{u}}$ be the vector of nodal values of $\widetilde{u}_N$. Similar to before, $\mathbf{A}_\varepsilon\widetilde{\mathbf{u}} = \mathbf{N}\,\mathbf{g}$, and then, as in (4.6), $h^{d/2}k\|\widetilde{\mathbf{u}}\|_2 \lesssim \|\widetilde{g}\|_{L^2(\Gamma)}$. Imitating the proof of (4.2) we find that $\|\widetilde{g}\|_{L^2(\Gamma)} \sim h^{(d-1)/2}\|\mathbf{g}\|_2$, and then combining these last two inequalities we obtain $\|\widetilde{\mathbf{u}}\|_2 \lesssim h^{-1/2}k^{-1}\|\mathbf{g}\|_2$, implying (ii) in (4.1). ∎

*Proof of Theorem 1.4 using Lemma 4.1.* When $\eta = k$ the bound (1.13) follows immediately from (1.17) using the bound on $\|\mathbf{A}_\varepsilon^{-1}\mathbf{M}\|_2$ in (4.1). When $\eta = \sqrt{k^2+\mathrm{i}\varepsilon}$ it is straightforward to show that $|\eta - k| \lesssim \varepsilon/k$, and then (1.13) follows from inserting both the bounds in (4.1) into (1.17) and using the hypothesis that $hk^2 \ge C$. Identical arguments can be used to prove (1.14). ∎

*Proof of Lemma 1.6.* If in the proof of Lemma 4.1 we use the bound (2.5) instead of the bound (2.8) then we obtain

$$\left\|\mathbf{A}_\varepsilon^{-1}\mathbf{M}\right\|_2 \lesssim \frac{1}{\varepsilon} \quad \text{and} \quad \left\|\mathbf{A}_\varepsilon^{-1}\mathbf{N}\right\|_2 \lesssim \frac{1}{\varepsilon^{1/2}kh^{1/2}}, \tag{4.9}$$

(and analogous estimates in the right preconditioning case). Repeating the proof of Theorem 1.4 but using (4.9) instead of (4.1) (and recalling the hypotheses that $\varepsilon \lesssim k^2$ and $hk^2 \geq C$), we obtain (1.16). ∎

**Remark 4.2 (Stability as opposed to quasi-optimality)** *Inspecting the proof of Lemma 4.1 we see that all it really requires is that the Galerkin solutions to the variational problems (4.4) and (4.7) exist and satisfy*

$$k\,\|\widetilde{u}_N\|_{L^2(\Omega)} \lesssim \|\widetilde{f}\|_{L^2(\Omega)} \quad \text{and} \quad k\,\|\widetilde{u}_N\|_{L^2(\Omega)} \lesssim \|\widetilde{g}\|_{L^2(\Gamma)} \tag{4.10}$$

*respectively. To obtain (4.10) we used quasi-optimality (from Lemma 3.5) and the bound on the solution operator (from Corollary 2.11). For the standard variational formulation of the Helmholtz equation (1.3) (with $\varepsilon = 0$), when $V_N \subset H^1(\Omega)$ consists of piecewise-linear polynomials, existence and uniqueness of the Galerkin solution $u_N$ and the bound (4.10) (but not quasi-optimality) were established under the mesh threshold $hk^{3/2} \lesssim 1$ in [56, Theorem 6.1]. This result was proved by establishing the corresponding result for a class of interior penalty methods, and then taking the limit as the penalty parameter tends to zero (and relying on the fact that the stability bound in Theorem 2.9 holds when $\varepsilon = 0$). If this result could be extended to the problem with absorption then we could establish the bounds (4.1) under the mesh threshold $hk^{3/2} \lesssim 1$.*

*The condition $hk^{3/2} \lesssim 1$ has appeared in other investigations of fixed-order finite element methods for the Helmholtz equation. In particular, Ihlenburg and Babuška [29], [28, Chapter 4] proved that in 1–d this condition is sufficient to keep the relative error in both the $H^1$-semi-norm and the $L^2$-norm bounded independently of $k$ (but is not sufficient for the method to be quasi-optimal); see [20, §1.2.2] for a review of both this and other related work.*

### *4.1.2 Non-smooth domains and shape-regular meshes*

We now consider non-smooth domains satisfying Assumption 3.7. Let $\mathcal{T}$ be any mesh in the sequence of meshes in Part 2 of Assumption 3.7, and let $\tau$ denote a typical simplex in $\mathcal{T}$. For each node $x_i$ of the mesh $\mathcal{T}$, introduce a representative mesh diameter $h_i$ which can be chosen to be the diameter of any of the simplices $\tau$ that touch $x_i$. It is a property of shape-regular meshes that (with the hidden constants independent of the mesh) $h_\tau \sim h_i$ for all $\tau$ that touch node $x_i$. Then let $\mathbf{D}$ be the diagonal matrix with diagonal entries $\mathbf{D}_{ii} = h_i^d$. Furthermore, let $\mathbf{D}_\Gamma$ be the diagonal matrix with $(\mathbf{D}_\Gamma)_{ii} = h_i^{d-1}$ if $x_i \in \Gamma$ and $(\mathbf{D}_\Gamma)_{ii} = 0$ otherwise.

The following result follows from standard scaling arguments using shape-regularity.

**Lemma 4.3** *For all $\mathbf{x} \in \mathbb{C}^N$,*

$$\text{(i)} \quad (\mathbf{M}\,\mathbf{x}, \mathbf{x})_2 \sim (\mathbf{D}\,\mathbf{x}, \mathbf{x})_2, \quad \text{and} \quad \text{(ii)} \quad (\mathbf{N}\,\mathbf{x}, \mathbf{x})_2 \sim (\mathbf{D}_\Gamma\,\mathbf{x}, \mathbf{x})_2.$$

The next theorem is the analogue of Theorem 1.4 for non-smooth domains.

**Theorem 4.4 (Sufficient conditions for $\mathbf{A}_\varepsilon^{-1}$ to be a good preconditioner when $\Omega$ is non-smooth)** *Suppose that $\Omega$ is Lipschitz and star-shaped with respect to a ball, and suppose further that Assumption 3.7 is satisfied. Suppose that both the interior impedance problem and its shifted counterpart are solved using the Galerkin method with $V_N$ corresponding to a mesh satisfying Assumption 3.7. Define $h_\Gamma := \min_{p\in\Gamma} h_p$. Assume that $\varepsilon \lesssim k^2$ and* either $\eta = k$ or $\eta = \sqrt{k^2 + \mathrm{i}\varepsilon}$. *Then, given $k_0 > 0$ and $C > 0$, there exist $C_1$ and $C_2$ (independent of $N, k$, and $\varepsilon$) such that if $k \geq k_0$, $h_\Gamma k^2 \geq C$, and*

$$N^{-1/d} k \sqrt{|k^2 - \varepsilon|} \leq C_1 \tag{4.11}$$

*then*

$$\left\| \mathbf{I} - \mathbf{D}^{1/2} \mathbf{A}_\varepsilon^{-1} \mathbf{A} \mathbf{D}^{-1/2} \right\|_2 \leq C_2 \frac{\varepsilon}{k}. \tag{4.12}$$

*and*

$$\left\| \mathbf{I} - \mathbf{D}^{-1/2} \mathbf{A} \mathbf{A}_\varepsilon^{-1} \mathbf{D}^{1/2} \right\|_2 \leq C_2 \frac{\varepsilon}{k}. \tag{4.13}$$

Recalling Corollary 1.9, we therefore see that (4.12) implies that, after a simple diagonal scaling on the left with $\mathbf{D}^{1/2}$ and on the right with $\mathbf{D}^{-1/2}$, equations involving the left-preconditioned matrix $\mathbf{A}_\varepsilon^{-1}\mathbf{A}$ can be solved with GMRES in a $k$-independent number of iterations when $\varepsilon/k$ is sufficiently small. The same statement is true for right preconditioning by (4.13), but the diagonal scalings should be performed in the opposite order.

*Proof of Theorem 4.4.* Similar to (1.17), we write

$$\begin{aligned}\mathbf{I}-\mathbf{D}^{1/2}\mathbf{A}_\varepsilon^{-1}\mathbf{A}\mathbf{D}^{-1/2} &= \mathbf{D}^{1/2}(\mathbf{I}-\mathbf{A}_\varepsilon^{-1}\mathbf{A})\mathbf{D}^{-1/2} = \mathbf{D}^{1/2}\mathbf{A}_\varepsilon^{-1}(\mathbf{A}_\varepsilon-\mathbf{A})\mathbf{D}^{-1/2}\\ &= -\mathrm{i}\varepsilon\mathbf{D}^{1/2}\mathbf{A}_\varepsilon^{-1}\mathbf{M}\mathbf{D}^{-1/2}-\mathrm{i}(\eta-k)\mathbf{D}^{1/2}\mathbf{A}_\varepsilon^{-1}\mathbf{N}\mathbf{D}^{-1/2}.\end{aligned}$$

The proof of (4.12) then consists of showing that

$$\left\|\mathbf{D}^{1/2}\mathbf{A}_\varepsilon^{-1}\mathbf{M}\,\mathbf{D}^{-1/2}\right\|_2 \lesssim \frac{1}{k} \quad \text{and} \quad \left\|\mathbf{D}^{1/2}\mathbf{A}_\varepsilon^{-1}\mathbf{N}\,\mathbf{D}^{-1/2}\right\|_2 \lesssim \frac{1}{h_\Gamma^{1/2}k}, \tag{4.14}$$

following the proof of the analogous bounds (4.1) in the smooth case.

Once we have proved (4.14), the proof of (4.13) is similar since

$$\left\|\mathbf{D}^{-1/2}\mathbf{M}\mathbf{A}_\varepsilon^{-1}\mathbf{D}^{1/2}\right\|_2 = \left\|\mathbf{D}^{1/2}(\mathbf{A}_\varepsilon^*)^{-1}\mathbf{M}\mathbf{D}^{-1/2}\right\|_2,$$

and

$$\left\|\mathbf{D}^{-1/2}\mathbf{N}\mathbf{A}_\varepsilon^{-1}\mathbf{D}^{1/2}\right\|_2 = \left\|\mathbf{D}^{1/2}(\mathbf{A}_\varepsilon^*)^{-1}\mathbf{N}\mathbf{D}^{-1/2}\right\|_2,$$

and $\mathbf{A}_\varepsilon^*$ is just the Galerkin matrix arising from the adjoint problem (2.3). Therefore, the argument used below to prove the bounds in (4.14) can also be used to prove analogous bounds on $\left\|\mathbf{D}^{1/2}\mathbf{A}_\varepsilon^{-1}\mathbf{M}\,\mathbf{D}^{-1/2}\right\|_2$ and $\left\|\mathbf{D}^{1/2}\mathbf{A}_\varepsilon^{-1}\mathbf{N}\,\mathbf{D}^{-1/2}\right\|_2$.

Returning to the proof of (4.14), given $\mathbf{f}\in\mathbb{C}^N$, we define $\widetilde{f}$, and $\widetilde{u}$ as in the first part of the proof of Lemma 4.1. Then the nodal values $\widetilde{\mathbf{u}}$ of $\widetilde{u}_N$ satisfy the linear system $\mathbf{A}_\varepsilon\widetilde{\mathbf{u}}=\mathbf{M}\,\mathbf{f}$. Moreover, using Lemma 4.3

$$\begin{aligned}k\left\|\mathbf{D}^{1/2}\widetilde{\mathbf{u}}\right\|_2 = k\,(\mathbf{D}\,\widetilde{\mathbf{u}},\widetilde{\mathbf{u}})_2^{1/2} \sim k\,(\mathbf{M}\,\widetilde{\mathbf{u}},\widetilde{\mathbf{u}})_2^{1/2} = k\,\|\widetilde{u}_N\|_{L^2(\Omega)} &\le \|\widetilde{u}_N\|_{1,k,\Omega}\\ &\le \|\widetilde{u}-\widetilde{u}_N\|_{1,k,\Omega}+\|\widetilde{u}\|_{1,k,\Omega}\,.\end{aligned} \tag{4.15}$$

By quasi-optimality (Lemma 3.10), the bound on the solution of the continuous problem (Corollary 2.11), the definition of $\widetilde{f}$, and Part (i) of Lemma 4.3 we have

$$k\left\|\mathbf{D}^{1/2}\widetilde{\mathbf{u}}\right\|_2 \lesssim \|\widetilde{u}\|_{1,k,\Omega} \lesssim \|\widetilde{f}\|_{L^2(\Omega)} = (\mathbf{M}\,\mathbf{f},\mathbf{f})_2^{1/2} \sim (\mathbf{D}\,\mathbf{f},\mathbf{f})_2^{1/2} \sim \left\|\mathbf{D}^{1/2}\mathbf{f}\right\|_2.$$

Remembering that $\mathbf{A}_\varepsilon\widetilde{\mathbf{u}}=\mathbf{M}\,\mathbf{f}$, we have

$$\left\|\mathbf{D}^{1/2}\mathbf{A}_\varepsilon^{-1}\mathbf{M}\,\mathbf{f}\right\|_2 \lesssim \frac{1}{k}\left\|\mathbf{D}^{1/2}\mathbf{f}\right\|_2,$$

and since $\mathbf{f}$ was arbitrary, this implies the first bound in (4.14).

Given $\mathbf{g}\in\mathbb{C}^N$, we define $\widetilde{g}$ and $\widetilde{u}$ as in the second part of the proof of Lemma 4.1; thus $\mathbf{A}_\varepsilon\widetilde{\mathbf{u}}=\mathbf{N}\,\mathbf{g}$. The inequalities in (4.15) hold as before, and then (using quasi-optimality and the bound on the solution of the continuous problem) we have $k\|\mathbf{D}^{1/2}\widetilde{\mathbf{u}}\|_2 \lesssim \|\widetilde{g}\|_{L^2(\Gamma)}$. By Part (ii) of Lemma 4.3, $\|\widetilde{g}\|_{L^2(\Gamma)} = (\mathbf{N}\mathbf{g},\mathbf{g})_2^{1/2} \sim (\mathbf{D}_\Gamma\mathbf{g},\mathbf{g})_2^{1/2} = \|\mathbf{D}_\Gamma^{1/2}\mathbf{g}\|_2$, so

$$k\left\|\mathbf{D}^{1/2}\mathbf{A}_\varepsilon^{-1}\mathbf{N}\mathbf{g}\right\|_2 \lesssim \left\|\mathbf{D}_\Gamma^{1/2}\mathbf{g}\right\|_2.$$

Since $\mathbf{g}$ was arbitrary this implies that

$$k\left\|\mathbf{D}^{1/2}\mathbf{A}_\varepsilon^{-1}\mathbf{N}\mathbf{D}^{-1/2}\mathbf{g}\right\|_2 \lesssim \left\|\mathbf{D}_\Gamma^{1/2}\mathbf{D}^{-1/2}\mathbf{g}\right\|_2. \tag{4.16}$$

Now, the definitions of $\mathbf{D}$, $\mathbf{D}_\Gamma$, and $h_\Gamma$ imply that

$$\left\|\mathbf{D}_\Gamma^{1/2}\mathbf{D}^{-1/2}\mathbf{g}\right\|_2 \le \max_{p\in\Gamma}\left(h_p^{(d-1)/2}h_p^{-d/2}\right)\|\mathbf{g}\|_2 = \max_{p\in\Gamma}\left(h_p^{-1/2}\right)\|\mathbf{g}\|_2 = h_\Gamma^{-1/2}\,\|\mathbf{g}\|_2\,,$$

and using this in (4.16) we find the second bound in (4.14).

■

### 4.2 Results about the truncated sound-soft scattering problem

Repeating the proof of Lemma 4.1, but now using the bounds on the continuous problem in Theorems 2.17 and 2.18 and the results in §3.5, we obtain the following result.

**Theorem 4.5 (Sufficient conditions for $\mathbf{A}_\varepsilon^{-1}$ to be a good preconditioner for the truncated sound-soft scattering problem)** *Suppose that $\Omega_D$ and $\Omega_R$ are both $C^{1,1}$, $\Omega_D$ is star-shaped with respect the origin, and $\Omega_R$ is star-shaped with respect to a ball centred at the origin. Suppose that the Galerkin discretisations of both the truncated sound-soft scattering problem and its shifted counterpart are formed with the finite dimensional subspaces consisting of piecewise polynomials on a quasi-uniform sequence of meshes. If $\varepsilon \lesssim k^2$ and* either $\eta = k$ or $\eta = \sqrt{k^2 + \mathrm{i}\varepsilon}$, *then, given $k_0 > 0$ and $C > 0$, there exist $C_1$ and $C_2$ (independent of $h, k$, and $\varepsilon$) such that if $hk\sqrt{|k^2 - \varepsilon|} \leq C_1$ and $hk^2 \geq C$ then*

$$\max\left\{\left\|\mathbf{I} - \mathbf{A}_\varepsilon^{-1}\mathbf{A}\right\|_2, \left\|\mathbf{I} - \mathbf{A}\mathbf{A}_\varepsilon^{-1}\right\|_2\right\} \leq C_2\frac{\varepsilon}{k} \tag{4.17}$$

*for all $k \geq k_0$.*

Recalling Corollary 1.9, we see that equations involving the matrices $\mathbf{A}_\varepsilon^{-1}\mathbf{A}$ and $\mathbf{A}\mathbf{A}_\varepsilon^{-1}$ can then be solved with GMRES in a $k$-independent number of iterations when $\varepsilon/k$ is sufficiently small.

In the non-smooth case, if $\Omega$ satisfies Assumption 3.7 (along with the geometric conditions in Theorem 4.5) then the analogue of Theorem 4.4 holds and, after a simple diagonal scaling on the left with $\mathbf{D}^{1/2}$ and on the right with $\mathbf{D}^{-1/2}$, problems involving the matrix $\mathbf{A}_\varepsilon^{-1}\mathbf{A}$ can be solved with GMRES in a $k$-independent number of iterations when $\varepsilon/k$ is sufficiently small (an analogous statement also holds for $\mathbf{A}\mathbf{A}_\varepsilon^{-1}$, with the diagonal scalings performed in the opposite order).

## 5 Numerical experiments

In this section we display the results of five numerical experiments. The first two experiments concern the interior impedance problem (Problem 2.1), the next two concern the truncated sound-soft scattering problem (Problem 2.4), and the final experiment concerns the Helmholtz equation in an inhomogeneous medium.

The purpose of these experiments is to investigate the choices of $\varepsilon$ under which the property (P1) of §1 holds (i.e. choices of $\varepsilon$ under which $\mathbf{A}_\varepsilon^{-1}$ is a good preconditioner for $\mathbf{A}$ in the sense that GMRES for $\mathbf{A}_\varepsilon^{-1}\mathbf{A}$ converges independently of $k$). We emphasise that we do *not* consider the time taken to construct good approximations of $\mathbf{A}_\varepsilon^{-1}$ (i.e. the question of choosing $\varepsilon$ and $\mathbf{B}_\varepsilon$ so that the property (P2) holds); in this sense, the investigation in this section is purely theoretical.

Recall from Theorem 1.8 that sufficient conditions for the number of GMRES iterations, $n_{\mathrm{GMRES}}$, needed solve the equation

$$\mathbf{A}_\varepsilon^{-1}\mathbf{A}\mathbf{u} = \mathbf{A}_\varepsilon^{-1}\mathbf{f} \tag{5.1}$$

to be independent of $k$ are

(i) that

$$d := \mathrm{dist}(0, W(\mathbf{A}_\varepsilon^{-1}\mathbf{A})) \tag{5.2}$$

is bounded away from the origin, independently of $k$, and

(ii) that $\|\mathbf{A}_\varepsilon^{-1}\mathbf{A}\|_2$ is bounded above, independently of $k$.

Theorem 1.4 shows that these two conditions are satisfied when (for the interior impedance problem) $\Omega$ is star-shaped with respect to a ball and $\varepsilon/k$ is sufficiently small. Furthermore, Lemma 1.6 shows that (ii) is satisfied if $\varepsilon \lesssim k^2$ (if $\Omega$ is still star-shaped with respect to a ball). Since these results indicate that (i) is a more restrictive condition on $\varepsilon$ than (ii), in the experiments below we do not compute $\|\mathbf{A}_\varepsilon^{-1}\mathbf{A}\|_2$, and we instead concentrate on exploring the behaviour of $d$.

In all five experiments we compute $d$, and in all but the first one we compute $n_{\mathrm{GMRES}}$. For the computation of $n_{\mathrm{GMRES}}$, the vector $\mathbf{f}$ on the right-hand side of (5.1) is taken to be the vector of ones, the initial guess is taken to be zero, and the stopping criterion is the reduction of the initial residual by six orders of magnitude.

All meshes start with an initial mesh, possibly locally refined, but then the meshes are refined uniformly by dividing triangles into four smaller ones, possibly several times over. Finally some mesh smoothing is applied, which modifies the elements slightly. The maximum mesh diameter $h$ is a key indicator of mesh density. However, with $h_{\min}$ denoting the diameter of the smallest element, some meshes have rather large ratio $h/h_{\min}$, in which case the effect of diagonal scaling is also investigated (cf. Theorem 4.4). For mesh refinement as $k$ increases, we explore two choices: (i) $hk \sim 1$ (i.e. a fixed number of grid points per wavelength) and (ii) $h \sim k^{-3/2}$, where the hidden constants are specified below. Although neither of these choices are covered by the theory, recall that there is some prospect of extending the theory to cover the choice $h \sim k^{-3/2}$; see Remark 4.2. Furthermore, Table 1 shows almost identical results arising from the choices $h \sim k^{-3/2}$ and $h \sim k^{-2}$. We study seven choices for $\varepsilon$, namely $\varepsilon = k/4,\ k/2,\ k, 2k, 4k,\ k^{3/2},$ and $k^2$.

**Example 5.1** In this first example we study the finite-difference approximation of the interior impedance problem (with the 5-point Laplacian), on the unit square on a uniform $n \times n$ grid (so $h = 1/n$), where either (i) $n = 2k$ (so that the number of grid points per wavelength is $2k(2\pi/k) = 4\pi \approx 12.57$) or (ii) $n = \lceil k^{3/2} \rceil$ (so that the number of grid points per wavelength is approximately $2\pi k^{1/2}$). The values of $d$ obtained for these two choices of $n$ are given in Tables 2 and 3 respectively. We observe that as long as $\varepsilon \lesssim k$, the value of $d$ remains approximately

| $k$ | $\varepsilon = k/4$ | $\varepsilon = k/2$ | $\varepsilon = k$ | $\varepsilon = 2k$ | $\varepsilon = 4k$ | $\varepsilon = k^{3/2}$ | $\varepsilon = k^2$ |
|---|---|---|---|---|---|---|---|
| 5 | 0.9288 | 0.8626 | 0.7450 | 0.5633 | 0.3441 | 0.5289 | 0.2778 |
| 10 | 0.9288 | 0.8641 | 0.7520 | 0.5808 | 0.3701 | 0.4434 | 0.1384 |
| 20 | 0.9234 | 0.8550 | 0.7384 | 0.5650 | 0.3581 | 0.3256 | 0.0512 |
| 40 | 0.9202 | 0.8492 | 0.7289 | 0.5519 | 0.3438 | 0.2194 | 0.0142 |
| 80 | 0.9181 | 0.8455 | 0.7226 | 0.5426 | 0.3338 | 0.1380 | 0.0054 |
| 160 | 0.9175 | 0.8442 | 0.7201 | 0.5386 | 0.3288 | 0.0842 | 0.0028 |

**Table 2** The values of $d$ for Example 5.1 when $n = 2k$.

| $k$ | $\varepsilon = k/4$ | $\varepsilon = k/2$ | $\varepsilon = k$ | $\varepsilon = 2k$ | $\varepsilon = 4k$ | $\varepsilon = k^{3/2}$ | $\varepsilon = k^2$ |
|---|---|---|---|---|---|---|---|
| 5 | 0.9297 | 0.8642 | 0.7477 | 0.5670 | 0.3478 | 0.5326 | 0.2812 |
| 10 | 0.9307 | 0.8677 | 0.7579 | 0.5888 | 0.3743 | 0.4482 | 0.1391 |
| 20 | 0.9252 | 0.8582 | 0.7436 | 0.5720 | 0.3652 | 0.3326 | 0.0516 |
| 40 | 0.9221 | 0.8525 | 0.7342 | 0.5593 | 0.3514 | 0.2256 | 0.0150 |
| 80 | 0.9199 | 0.8485 | 0.7276 | 0.5494 | 0.3406 | 0.1422 | 0.0057 |

**Table 3** The values of $d$ for Example 5.1 when $n = \lceil k^{3/2} \rceil$.

constant as $k$ increases. However, as soon as $\varepsilon$ grows faster than $k$, $d$ tends to zero.

**Example 5.2** Here we study the linear finite-element approximation of the interior impedance problem on a uniform triangular $n \times n$ grid on the unit square, with the same two regimes for $h$ as in Example 5.1. Tables 4 and 5 give the values of $d$ and (in parentheses) $n_{\text{GMRES}}$. The values of $d$ are similar to those in Example 5.1. The number of iterations, $n_{\text{GMRES}}$, remains constant as $k$ increases in all the cases when $\varepsilon$ is proportional to $k$, but grows in all other cases; this is in line with Theorem 1.5.

| $k$ | $\varepsilon = k/4$ | $\varepsilon = k/2$ | $\varepsilon = k$ | $\varepsilon = 2k$ | $\varepsilon = 4k$ | $\varepsilon = k^{3/2}$ | $\varepsilon = k^2$ |
|---|---|---|---|---|---|---|---|
| 10 | 0.9328(4) | 0.8714(5) | 0.7641(6) | 0.5971(7) | 0.3861(9) | 0.4594(8) | 0.1466(13) |
| 20 | 0.9272(4) | 0.8618(5) | 0.7493(6) | 0.5797(8) | 0.3729(10) | 0.3413(11) | 0.0538(25) |
| 40 | 0.9246(4) | 0.8569(5) | 0.7411(6) | 0.5675(8) | 0.3590(11) | 0.2311(13) | 0.0156(47) |
| 80 | 0.9230(4) | 0.8540(5) | 0.7360(6) | 0.5610(7) | 0.3525(10) | 0.1477(16) | 0.0039(84) |
| 160 | 0.9223(4) | 0.8525(5) | 0.7336(6) | 0.5547(7) | 0.3439(10) | 0.0870(19) | 0.0030(148) |

**Table 4** The values of $d$ (and in parentheses $n_{\text{GMRES}}$) for Example 5.2 when $n = 2k$.

| $k$ | $\varepsilon = k/4$ | $\varepsilon = k/2$ | $\varepsilon = k$ | $\varepsilon = 2k$ | $\varepsilon = 4k$ | $\varepsilon = k^{3/2}$ | $\varepsilon = k^2$ |
|---|---|---|---|---|---|---|---|
| 10 | 0.9323(4) | 0.8706(5) | 0.7627(6) | 0.5943(7) | 0.3812(9) | 0.4550(8) | 0.1432(13) |
| 20 | 0.9260(4) | 0.8595(5) | 0.7458(6) | 0.5749(8) | 0.3704(11) | 0.3367(11) | 0.0525(24) |
| 40 | 0.9226(4) | 0.8535(5) | 0.7358(6) | 0.5609(8) | 0.3529(11) | 0.2275(14) | 0.0150(48) |
| 80 | 0.9201(4) | 0.8490(5) | 0.7283(6) | 0.5504(8) | 0.3417(10) | 0.1443(16) | 0.0056(86) |

**Table 5** The values of $d$ (and in parentheses $n_{\text{GMRES}}$) for Example 5.2 when $n = \lceil k^{3/2} \rceil$.

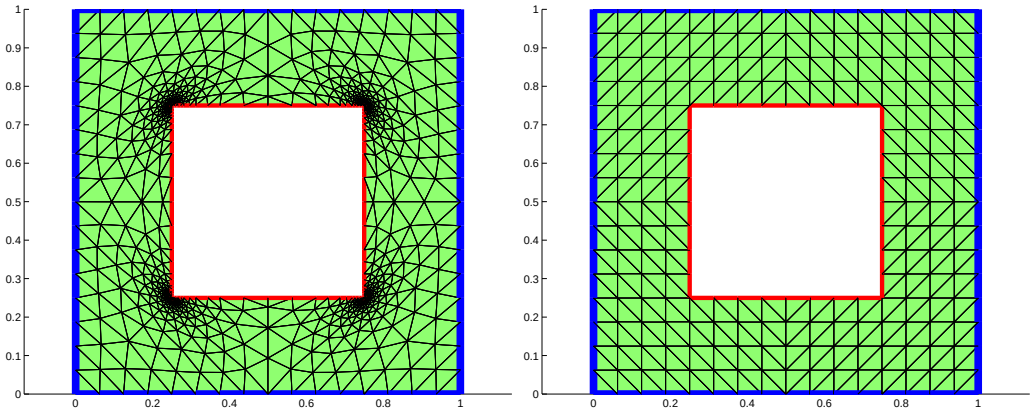

**Fig. 3** Meshes for scattering problem of Example 5.3

**Example 5.3** In this example we study the influence of local mesh refinement. We solve the truncated sound-soft scattering problem (Problem 2.4) when the domain is the region between a unit square and a square obstacle of side length 1/2 placed symmetrically inside; see Figure 3. In order to deal with irregularity near reentrant corners we perform local refinement to obtain an initial mesh with 288 nodes. The ratio $h/h_{\min}$ in this mesh is about 160. We use this mesh for computations with the wave number $k = 10$ and then perform one uniform refinement of this mesh for each doubling of $k$; thus we are working essentially with a quasi-uniform sequence but with a rather high mesh ratio. After three refinements and smoothing $h/h_{\min}$ is about 240. The mesh for $k = 20$ is depicted in Figure 3 (left). We also perform computations on the uniform mesh depicted in Figure 3 (right), which contains initially 240 nodes, and is shown without refinement.

Tables 6 and 7 show the computed values of $d$ (and $n_{\text{GMRES}}$) without and then with diagonal scaling (note that a 0 in the table indicates that the numerical range contains the origin).

| $k$ | $\varepsilon = k/4$ | $\varepsilon = k/2$ | $\varepsilon = k$ | $\varepsilon = 2k$ | $\varepsilon = 4k$ | $\varepsilon = k^{3/2}$ | $\varepsilon = k^2$ |
|---|---|---|---|---|---|---|---|
| 10 | 0.9595(3) | 0.9214(4) | 0.8519(4) | 0.7346(5) | 0.5614(6) | 0.6259(6) | 0.2845(8) |
| 20 | 0.9464(4) | 0.8960(4) | 0.8041(5) | 0.6487(6) | 0.4181(8) | 0.3749(8) | 0(15) |
| 40 | 0.9308(3) | 0.8655(4) | 0.7452(5) | 0.5383(6) | 0.2228(8) | 0(10) | 0(30) |
| 80 | 0.9096(3) | 0.8236(4) | 0.6632(5) | 0.3828(6) | 0(8) | 0(12) | 0(54) |
| 160 | 0.8790(3) | 0.7636(4) | 0.5480(5) | 0.1702(6) | 0(8) | 0(14) | 0(95) |

**Table 6** The values of $d$ (and in parentheses $n_{\text{GMRES}}$) for Example 5.3 without diagonal scaling.

| $k$ | $\varepsilon = k/4$ | $\varepsilon = k/2$ | $\varepsilon = k$ | $\varepsilon = 2k$ | $\varepsilon = 4k$ | $\varepsilon = k^{3/2}$ | $\varepsilon = k^2$ |
|---|---|---|---|---|---|---|---|
| 10 | 0.9612(3) | 0.9247(3) | 0.8579(4) | 0.7450(5) | 0.5792(6) | 0.6407(5) | 0.3255(8) |
| 20 | 0.9528(4) | 0.9089(4) | 0.8298(5) | 0.6995(6) | 0.5165(8) | 0.4840(8) | 0.1042(14) |
| 40 | 0.9450(4) | 0.8942(4) | 0.8035(5) | 0.6582(6) | 0.4600(8) | 0.3224(11) | 0.0258(31) |
| 80 | 0.9433(4) | 0.8909(4) | 0.7973(5) | 0.6466(7) | 0.4443(9) | 0.2119(13) | 0.0024(60) |
| 160 | 0.9420(4) | 0.8885(4) | 0.7931(5) | 0.6405(7) | 0.4380(9) | 0.1337(16) | 0(116) |

**Table 7** The values of $d$ (and in parentheses $n_{\text{GMRES}}$) for Example 5.3 with diagonal scaling.

Whereas diagonal scaling produces values of $d$ that behave similarly to those for the uniform mesh (as, in some sense, predicted by the theory), without diagonal scaling $d$ decreases as $\varepsilon$

| $k$ | $\varepsilon = k/4$ | $\varepsilon = k/2$ | $\varepsilon = k$ | $\varepsilon = 2k$ | $\varepsilon = 4k$ | $\varepsilon = k^{3/2}$ | $\varepsilon = k^2$ |
|---|---|---|---|---|---|---|---|
| 10 | 0.9624(3) | 0.9272(4) | 0.8630(4) | 0.7550(5) | 0.5964(6) | 0.6554(6) | 0.3457(8) |
| 20 | 0.9491(4) | 0.9021(4) | 0.8179(5) | 0.6818(7) | 0.4953(8) | 0.4627(8) | 0.1059(17) |
| 40 | 0.9401(4) | 0.8852(4) | 0.7881(5) | 0.6345(7) | 0.4337(9) | 0.2972(11) | 0.0230(35) |
| 80 | 0.9378(4) | 0.8810(5) | 0.7810(5) | 0.6231(7) | 0.4200(9) | 0.1973(14) | 0.0061(69) |
| 160 | 0.9371(4) | 0.8794(4) | 0.7780(5) | 0.6186(7) | 0.4139(9) | 0.1236(17) | 0.0014(121) |

**Table 8** The values of $d$ (and in parentheses $n_{\text{GMRES}}$) for Example 5.3 with a uniform mesh.

increases, and this even happens when $\varepsilon \sim k$. It is interesting to note that the values of $n_{\text{GMRES}}$ are not substantially altered by the presence of the diagonal scaling, despite the fact that without the diagonal scaling the numerical range often contains the origin. For comparison, we show in Table 8 the results obtained on a uniform mesh sequence, with the initial mesh illustrated in Figure 3 (right), having 240 nodes for the case $k = 10$. The values of $d$ and $n_{\text{GMRES}}$ in this case are similar to those in the case of diagonal scaling (in Table 7).

**Example 5.4** Our next example involves a non-star-shaped scatterer, which is not covered by the theory. This domain is depicted in Figure 4 (top left pane). The outer boundary is a square of size $2L \times 2L$. The obstacle is a square of size $2R \times 2R$ placed symmetrically inside, with an $2a \times 2a$ square removed from one side. The configuration is symmetric about the centre vertical line. For the experiments we use $L = 6$, $R = 3$ and $a = 1$. We solve the truncated sound soft scattering problem (Problem 2.4), with an incident plane wave coming from the bottom left corner of the picture, with the direction of propagation at 45 degrees with the positive $x-$axis.

The scatterer in this example is *trapping*, since there exist closed paths of rays in its exterior (see, e.g., [6, Definition 5.4] for the definition of trapping and nontrapping). In such a domain we cannot expect the solution operator to be bounded independently of $k$, as Theorem 2.18 proves is the case for star-shaped scatterers. Indeed, when $k = m\pi/a$ for $m \in \mathbb{Z}$ there exist *quasimodes* (in some sense, approximate eigenvalues of the operator), and these show that if the bound on the solution operator

$$\|\nabla u\|^2_{L^2(\Omega)} + k^2 \|u\|^2_{L^2(\Omega)} \lesssim C(k)^2 \left[\|f\|^2_{L^2(\Omega)} + \|g\|^2_{L^2(\Gamma)}\right]$$

holds, then $C(k)$ must grow at least linearly with $k$; see [6, Equation 5.38].

Tables 9 and 10 give the values of $d$ and $n_{\text{GMRES}}$ for meshes obtained by uniform refinement of the initial mesh depicted in Figure 4, with $h \sim k^{-1}$ as $k$ is increased. The corresponding rows of Tables 9 and 10 use the same meshes, so that in Table 10 the waves are considerably less well-resolved. The resulting total wave (incident plus scattered) for several choices of $k$ is depicted in Figure 4 (top right and bottom panes) for the well-resolved case corresponding to Table 9. The symbol (>) indicates that GMRES did not converge in fewer than 1000 iterations. It is interesting to compare rows 3-5 of Table 9, with rows 1-3 of Table 10, since these are problems with the same values of $k$. We are therefore solving the same physical problem with the same preconditioner, but in the second table we are largely underresolved, while in the first one the resolution is substantially higher (while still being at a fixed number of points per wave length). We see that the preconditioner works much better when the discretization is finer, which is natural since the theory in the rest of the paper is based on continuous (as opposed to discrete) arguments.

Comparing Table 9 with Tables 6–8, we see that the behaviour of $d$ and $n_{\text{GMRES}}$ for the trapping domain is quite different to the behaviour of $d$ and $n_{\text{GMRES}}$ for the square. Indeed, whereas 0 is never in the numerical range for the square, it is for the trapping domain for the larger values of $\varepsilon$. Furthermore, whereas for the square $n_{\text{GMRES}}$ is fairly constant (as $k$ increases) when $\varepsilon \sim k$, for the trapping domain $n_{\text{GMRES}}$ unequivocally grows for some cases when $\varepsilon \sim k$, e.g. $\varepsilon = k, 2k$, and $4k$, (although for $\varepsilon = k/4$ the number of iterations is still constant for the trapping domain in the well-resolved case).

**Example 5.5** Our final experiment involves an inhomogeneous medium, which is also not covered by the theory. Let $\Omega$ be the rectangular domain shown in Figure 5 on the left, which represents

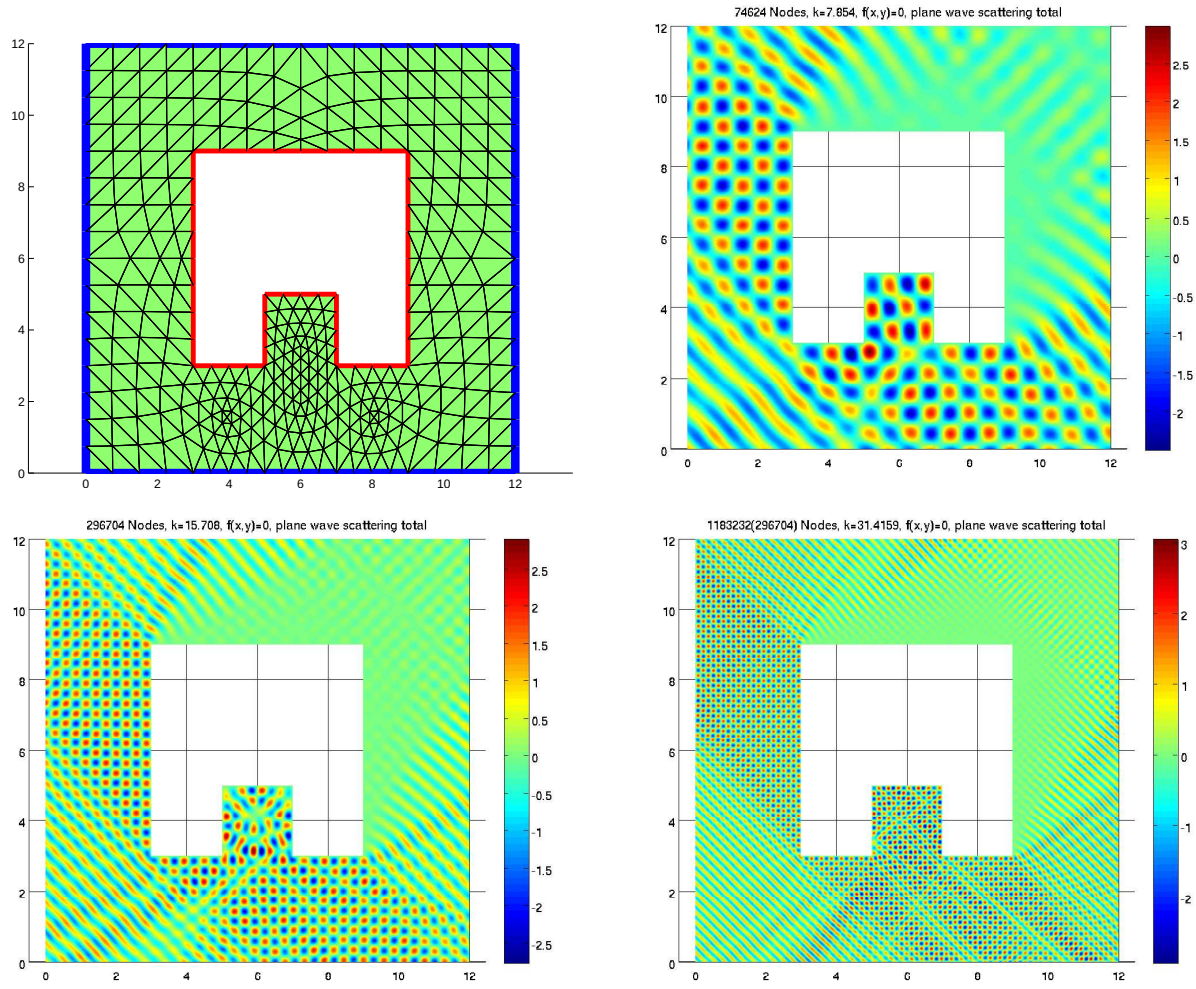


**Fig. 4** Initial mesh (top left) and three converged solutions for Example 5.4, corresponding to the last three lines in Table 9.

| $k$ | $\varepsilon = k/4$ | $\varepsilon = k/2$ | $\varepsilon = k$ | $\varepsilon = 2k$ | $\varepsilon = 4k$ | $\varepsilon = k^{3/2}$ | $\varepsilon = k^2$ |
|---|---|---|---|---|---|---|---|
| $5\pi/8$ | 0.3902(8) | 0.1932(12) | 0.0624 (17) | 0(27) | 0 (41) | 0.0227 (21) | 0 (26) |
| $10\pi/8$ | 0.4385 (8) | 0.2211 (12) | 0.0612 (18) | 0 (29) | 0 (48) | 0 (29) | 0 (47) |
| $20\pi/8$ | 0.3878 (9) | 0.1764 (13) | 0.0319 (19) | 0 (32) | 0 (55) | 0(41) | 0 (95) |
| $40\pi/8$ | 0.3069 (9) | 0.1137 (13) | 0 (21) | 0 (34) | 0 (61) | 0 (60) | 0 (198) |
| $80\pi/8$ | 0.2478 (9) | 0.0762 (14) | 0 (22) | 0 (37) | 0 (66) | 0 (89) | 0 (418) |

**Table 9** The values of $d$ (and in parentheses $n_{\text{GMRES}}$) for Example 5.4 with $n \sim k$

| $k$ | $\varepsilon = k/4$ | $\varepsilon = k/2$ | $\varepsilon = k$ | $\varepsilon = 2k$ | $\varepsilon = 4k$ | $\varepsilon = k^{3/2}$ | $\varepsilon = k^2$ |
|---|---|---|---|---|---|---|---|
| $5\pi/2$ | 0.1610 (15) | 0.0798 (21) | 0.0246 (32) | 0 (54) | 0 (91) | 0(70) | 0 (123) |
| $10\pi/2$ | 0.1199 (20) | 0.0374 (32) | 0.0227 (53) | 0 (94) | 0(177) | 0(176) | 0 (475) |
| $20\pi/2$ | 0 (35) | 0 (57) | 0 (100) | 0 (185) | 0 (355) | 0 (491) | (>) |
| $40\pi/2$ | 0.0266 (48) | 0 (86) | 0 (160) | 0 (306) | 0 (594) | 0 (>) | 0 (>) |
| $80\pi/2$ | 0.0197 (75) | 0 (140) | 0 (267) | 0 (515) | 0 (>) | 0 (> ) | 0 (> ) |

**Table 10** The values of $d$ (and in parentheses $n_{\text{GMRES}}$) for Example 5.4 with $n \sim k$ and a different range of $k$

a 2d cross section of a model of a microwave oven with a chicken in it. We consider the interior Helmholtz problem

$$\begin{aligned} \Delta u + (k^2/c^2)u &= 0 \quad && \text{in } \Omega, \\ \partial_n u - \mathrm{i}ku &= g \quad && \text{on the right boundary}, \\ u &= 0 \quad && \text{on the remaining boundaries}, \end{aligned}$$

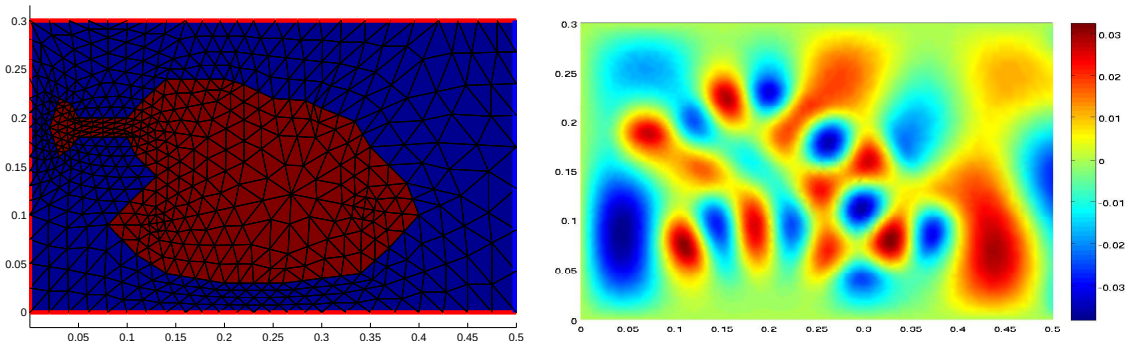

**Fig. 5** Inhomogeneous medium example on the left (chicken in a microwave) with an initial mesh of 597 nodes for $k = 10$. Solution on the right for $k = 40$

| $k$ | $\varepsilon = k/4$ | $\varepsilon = k/2$ | $\varepsilon = k$ | $\varepsilon = 2k$ | $\varepsilon = 4k$ | $\varepsilon = k^{3/2}$ | $\varepsilon = k^2$ |
|---|---|---|---|---|---|---|---|
| | microwave oven with chicken | | | | | | |
| 10 | 0.9423(5) | 0.8769(6) | 0.7352(7) | 0.4671(8) | 0.1323(10) | 0.2387(9) | 0(14) |
| 20 | 0.2725(6) | 0.0910(8) | 0(10) | 0(13) | 0(18) | 0(19) | 0(40) |
| 40 | 0(7) | 0(9) | 0(13) | 0(18) | 0(27) | 0(37) | 0(132) |
| | empty microwave oven | | | | | | |
| 10 | 0.9656(3) | 0.9309(4) | 0.8621(4) | 0.7333(5) | 0.5308(6) | 0.6062(6) | 0.2402(7) |
| 20 | 0.9141(4) | 0.8365(5) | 0.7019(6) | 0.4956(7) | 0.2386(9) | 0.1974(9) | 0(18) |
| 40 | 0.8477(4) | 0.7268(5) | 0.5477(7) | 0.3299(9) | 0.1249(13) | 0.0182(17) | 0(53) |

**Table 11** Chicken in a microwave problem, and also for comparison the corresponding empty microwave oven, but using the same mesh

with a renormalised wave speed of $c = 1$ in air, and $c = 1/\sqrt{5}$ in the chicken. The source on the right is as in a classical microwave oven, and modeled by a Robin condition with $g = 1$. This is a synthetic problem, since the frequency in a microwave oven is given, and we vary it here only for illustrative purposes. The ratio between the two values of $c$ is roughly what one expects physically, and we choose $k = 10, 20, 40$ as in the earlier experiments (with this parameter now corresponding to the frequency).

Table 11 shows the values of $d$ and $n_{\text{GMRES}}$, both for the microwave with the chicken in, and also for an empty microwave (corresponding to $c = 1$ everywhere). We clearly see that the inhomogeneous medium causes difficulties for the preconditioner, and at $k = 40$ the numerical range contains zero for every choice of $\varepsilon$ considered. The number of iterations, $n_{\text{GMRES}}$, grows with $k$ in all cases, but this growth gets faster as $\varepsilon$ increases. The preconditioner works much better in the empty microwave oven, although in this case $d$ decreases with $k$ even when $\varepsilon \sim k$; this decrease is probably caused by the fact that we used the same irregular mesh as in the inhomogeneous case. Despite this decrease in $d$, the number of iterations remain roughly constant as $k$ increases when $\varepsilon = k/4$ and $\varepsilon = k/2$.

## 6 Concluding remarks

The results of this paper show that the shifted Laplacian is a good preconditioner for finite-element discretisations of the Helmholtz equation, in the sense that (P1) of §1 is satisfied, if $\varepsilon/k$ is sufficiently small. We emphasise again that this is not the end of the story, since for practical computations we need both (P1) and (P2) to be satisfied (or at least a compromise reached between the two); this paper, however, contains the first rigorous results on how to achieve one of (P1) or (P2).

In this conclusion we show that the requirement "$\varepsilon/k$ is sufficiently small" naturally appears when one considers how well the solution of the problem with absorption approximates the solution of the problem without absorption (independently of any discretisation). We focus on Problem 2.1 (the interior impedance problem), but note that analogous results hold for Problem 2.4 (the truncated sound-soft scattering problem).

**Theorem 6.1 (Approximating $u$ by $u_\varepsilon$)** *Let $\Omega$ be a Lipschitz domain that is star-shaped with respect to a ball (see Definition 1.2). Given $f \in L^2(\Omega)$ and $g \in L^2(\Gamma)$, let $u$ be the solution of*

*Problem 2.1 with $\varepsilon = 0$ and $\eta = k$ (i.e. $u$ satisfies (1.1)) and let $u_\varepsilon$ be the solution of Problem 2.1 with $\varepsilon \neq 0$ and $\eta = k$ (i.e. $u$ satisfies (1.2) with $\eta = k$).*

*If $\varepsilon \lesssim k^2$ then, given $k_0 > 0$, there exists $C_1$ (independent of $k$ and $\varepsilon$) such that*

$$\|u - u_\varepsilon\|_{1,k,\Omega} \leq C_1 \frac{\varepsilon}{k} \left( \|f\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma)} \right) \tag{6.1}$$

*for all $k \geq k_0$. Furthermore, given $k_0 > 0$ there exist $C_2$ and $C_3$ (independent of $k$ and $\varepsilon$) such that if $\varepsilon \leq C_2 k$ then*

$$\frac{\|u - u_\varepsilon\|_{L^2(\Omega)}}{\|u\|_{L^2(\Omega)}} \leq C_3 \frac{\varepsilon}{k} \tag{6.2}$$

*for all $k \geq k_0$.*

Therefore, if $\varepsilon/k$ is sufficiently small then both the relative $L^2$-error in approximating $u$ by $u_\varepsilon$ and the error relative to the data are small.

Note that the principle of limited absorption states that, with $k$ fixed, $u_\varepsilon \to u$ as $\varepsilon \to 0$ (for a proof of this result for the exterior Dirichlet problem see [53, Chapter 9, Theorem 1.3]). In contrast, here we consider fixing $\varepsilon$ as a function of $k$ and then approximating $u$ by $u_\varepsilon$ for arbitrarily large $k$.

*Proof of Theorem 6.1.* By subtracting (1.2a) from (1.1a) and (1.2b) from (1.1b) we have that

$$(\Delta + k^2)(u - u_\varepsilon) = \mathrm{i}\varepsilon u_\varepsilon \quad \text{in } \Omega \quad \text{and} \quad \partial_n(u - u_\varepsilon) - \mathrm{i}k(u - u_\varepsilon) = 0 \quad \text{on } \Gamma.$$

By using the bound (2.8) with $\varepsilon = 0$ on $u - u_\varepsilon$, we have that, given $k_0 > 0$,

$$\|u - u_\varepsilon\|_{1,k,\Omega} \lesssim \varepsilon \|u_\varepsilon\|_{L^2(\Omega)} \tag{6.3}$$

for all $k \geq k_0$. The bound (2.8) applied to $u_\varepsilon$ then implies that if $\varepsilon \lesssim k^2$ then

$$\varepsilon \|u_\varepsilon\|_{L^2(\Omega)} \lesssim \frac{\varepsilon}{k} \left( \|f\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma)} \right),$$

and then using this in (6.3) we obtain (6.1).

Using the triangle inequality $\|u_\varepsilon\|_{L^2(\Omega)} \leq \|u - u_\varepsilon\|_{L^2(\Omega)} + \|u\|_{L^2(\Omega)}$ in (6.3), we find that there exist $C_2$ and $C_3$ such that if $\varepsilon \leq C_2 k$ then

$$\|u - u_\varepsilon\|_{1,k,\Omega} \leq C_3 \varepsilon \|u\|_{L^2(\Omega)},$$

which implies the result (6.2). ∎

**Remark 6.2 (The choice of $\eta$)** *In Theorem 6.1 we considered the case that $\eta = k$, i.e. the impedance parameter in the shifted problem is that in the unshifted problem. If $\eta = \sqrt{k^2 + \mathrm{i}\varepsilon}$ then it is straightforward to show that (6.1) holds, but we have not been able to prove that (6.2) holds in this case.*

**Remark 6.3 (The case $k \lesssim \varepsilon \lesssim k^2$)** *If $k \lesssim \varepsilon \lesssim k^2$ then one can obtain a bound analogous to (6.1), but one cannot obtain a bound analogous to (6.2) (at least, not with the results in the rest of this paper). Indeed, if $k \lesssim \varepsilon \lesssim k^2$ then we can apply the bound (2.5) (instead of the bound (2.8)) to $u - u_\varepsilon$ and obtain that*

$$\|u - u_\varepsilon\|_{1,k,\Omega} \lesssim \varepsilon \left( \frac{k}{\varepsilon} \right) \|u_\varepsilon\|_{L^2(\Omega)}. \tag{6.4}$$

*Since $u_\varepsilon$ itself satisfies the bound (2.5), we then have that*

$$\|u - u_\varepsilon\|_{1,k,\Omega} \lesssim \left( \frac{k}{\varepsilon} \|f\|_{L^2(\Omega)} + \left( \frac{k}{\varepsilon} \right)^{1/2} \|g\|_{L^2(\Gamma)} \right),$$

*which is analogous to (6.1). If we seek to prove an analogous bound to (6.2), however, we find from (6.4) that*

$$\|u - u_\varepsilon\|_{1,k,\Omega} \lesssim k \|u - u_\varepsilon\|_{L^2(\Omega)} + k \|u\|_{L^2(\Omega)}. \tag{6.5}$$

*One cannot get a bound on the relative $L^2$-error from (6.5), unless the omitted constant is $< 1$. (In principle we could check if this is ever the case, but doing so would be difficult.)*

**Acknowledgements**

The authors thank Lehel Banjai (Heriot-Watt), Robert Kirby (Baylor), Markus Melenk (TU Wien), and Valery Smyshlyaev (University College London) for useful discussions. The authors also thank the referees and the editor for their constructive comments. E.A.S was supported by EPSRC grant EP/1025995/1.

## References

1. T. Apel, A-M. Sändig, and J. R. Whiteman. Graded mesh refinement and error estimates for finite element solutions of elliptic boundary value problems in non-smooth domains. *Mathematical methods in the Applied Sciences*, 19(1):63–85, 1996.
2. A. Bayliss, C. I Goldstein, and E. Turkel. An iterative method for the Helmholtz equation. *Journal of Computational Physics*, 49(3):443–457, 1983.
3. B. Beckermann, S. A. Goreinov, and E. E. Tyrtyshnikov. Some remarks on the Elman estimate for GMRES. *SIAM journal on Matrix Analysis and Applications*, 27(3):772–778, 2006.
4. S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics*. Springer, 2000.
5. X-C. Cai and O. B. Widlund. Domain decomposition algorithms for indefinite elliptic problems. *SIAM Journal on Scientific and Statistical Computing*, 13(1):243–258, 1992.
6. S. N. Chandler-Wilde, I. G. Graham, S. Langdon, and E. A. Spence. Numerical-asymptotic boundary integral methods in high-frequency acoustic scattering. *Acta Numerica*, 21(1):89–305, 2012.
7. P-H. Cocquet and M. Gander. Analysis of multigrid performance for finite element discretizations of the shifted Helmholtz equation. *preprint*, 2014.
8. S. Cools and W. Vanroose. Local Fourier Analysis of the complex shifted Laplacian preconditioner for Helmholtz problems. *Numerical Linear Algebra with Applications*, 20:575–597, 2013.
9. C. C. Cowen and E. Harel. An effective algorithm for computing the numerical range, 1995. *Unpublished manuscript*.
10. P. Cummings and X. Feng. Sharp regularity coefficient estimates for complex-valued acoustic and elastic Helmholtz equations. *Mathematical Models and Methods in Applied Sciences*, 16(1):139, 2006.
11. S. C. Eisenstat, H. C. Elman, and M. H. Schultz. Variational iterative methods for nonsymmetric systems of linear equations. *SIAM Journal on Numerical Analysis*, pages 345–357, 1983.
12. H. C. Elman. *Iterative Methods for Sparse Nonsymmetric Systems of Linear Equations*. PhD thesis, Yale University, 1982.
13. B. Engquist and L. Ying. Sweeping preconditioner for the Helmholtz equation: hierarchical matrix representation. *Comm. Pure Appl. Math.*, 64:697–735, 2011.
14. Y. A. Erlangga. Advances in iterative methods and preconditioners for the Helmholtz equation. *Archives of Computational Methods in Engineering*, 15(1):37–66, 2008.
15. Y. A. Erlangga, C. W. Oosterlee, and C. Vuik. A novel multigrid based precconditioner for heterogeneous Helmholtz problems. *SIAM J. Sci. Comp.*, 27:1471–1492, 2006.
16. Y. A. Erlangga, C. Vuik, and C. W. Oosterlee. On a class of preconditioners for solving the Helmholtz equation. *Applied Numerical Mathematics*, 50(3):409–425, 2004.
17. Y. A Erlangga, C. Vuik, and C. W. Oosterlee. Comparison of multigrid and incomplete LU shifted-Laplace preconditioners for the inhomogeneous Helmholtz equation. *Applied numerical mathematics*, 56(5):648–666, 2006.
18. O. G. Ernst and M. J. Gander. Why it is difficult to solve Helmholtz problems with classical iterative methods. In I. G. Graham, T. Y. Hou, O. Lakkis, and R. Scheichl, editors, *Numerical Analysis of Multiscale Problems*, volume 83 of *Lecture Notes in Computational Science and Engineering*, pages 325–363. Springer, 2012.
19. S. Esterhazy and J. M. Melenk. On stability of discretizations of the Helmholtz equation. In I. G. Graham, T. Y. Hou, O. Lakkis, and R. Scheichl, editors, *Numerical Analysis of Multiscale Problems*, volume 83 of *Lecture Notes in Computational Science and Engineering*, pages 285–324. Springer, 2012.
20. I. G. Graham, M. Löhndorf, J. M. Melenk, and E. A. Spence. When is the error in the $h$-BEM for solving the Helmholtz equation bounded independently of $k$? *BIT Numer. Math.*, to appear, 2014.
21. I. G. Graham, E. A. Spence, and E. Vainikko. Additive Schwarz methods for the Helmholtz equation with and without absorption. *in preparation*, 2014.
22. P. Grisvard. *Elliptic problems in nonsmooth domains*. Pitman, Boston, 1985.
23. T. Ha-Duong. *Topics in computational wave propagation*, volume 31 of *Lecture Notes in Computational Science and Engineering*, chapter On retarded potential boundary integral equations and their discretisation, pages 301–336. Springer, 2003.
24. A. Hannukainen. Field of values analysis of Laplace preconditioners for the Helmholtz equation. *preprint*, 2012.
25. A. Hannukainen. Field of Values Analysis of a Two-Level Preconditioner for the Helmholtz Equation. *SIAM Journal on Numerical Analysis*, 51(3):1567–1584, 2013.
26. U. Hetmaniuk. Stability estimates for a class of Helmholtz problems. *Commun. Math. Sci*, 5(3):665–678, 2007.
27. R. Hiptmair, A. Moiola, and I. Perugia. Stability results for the time-harmonic Maxwell equations with impedance boundary conditions. *Mathematical Models and Methods in Applied Sciences*, 21(11):2263–2287, 2011.
28. F. Ihlenburg. *Finite element analysis of acoustic scattering*, volume 132. Springer Verlag, 1998.

29. F. Ihlenburg and I. Babuška. Finite element solution of the Helmholtz equation with high wave number Part I: The $h$-version of the FEM. *Computers & Mathematics with Applications*, 30(9):9–37, 1995.
30. J-H. Kimn and M. Sarkis. Shifted Laplacian RAS Solvers for the Helmholtz Equation. In *Proceedings of the 20th International Conference on Domain Decomposition Methods in San Diego, California*, 2011.
31. R. C. Kirby. From functional analysis to iterative methods. *SIAM Review*, 52(2):269–293, 2010.
32. A. Laird and M. Giles. Preconditioned iterative solution of the 2D Helmholtz equation. Technical Report NA 02-12, Computing Lab, Oxford University, 2002.
33. M. Löhndorf and J. M. Melenk. Wavenumber-explicit $hp$-BEM for high frequency scattering. *SIAM Journal on Numerical Analysis*, 49(6):2340–2363, 2011.
34. W. C. H. McLean. *Strongly elliptic systems and boundary integral equations*. Cambridge University Press, 2000.
35. J. M. Melenk. *On generalized finite element methods*. PhD thesis, The University of Maryland, 1995.
36. A. Moiola. *Trefftz-discontinuous Galerkin methods for time-harmonic wave problems*. PhD thesis, Seminar for applied mathematics, ETH Zürich, 2011. Available at http://e-collection.library.ethz.ch/view/eth:4515.
37. A. Moiola and E. A. Spence. Is the Helmholtz equation really sign-indefinite? *SIAM Review*, 56(2):274–312, 2014.
38. C. S. Morawetz. The decay of solutions of the exterior initial-boundary value problem for the wave equation. *Communications on Pure and Applied Mathematics*, 14(3):561–568, 1961.
39. C. S. Morawetz. Decay for solutions of the exterior problem for the wave equation. *Communications on Pure and Applied Mathematics*, 28(2):229–264, 1975.
40. C. S. Morawetz and D. Ludwig. An inequality for the reduced wave operator and the justification of geometrical optics. *Communications on Pure and Applied Mathematics*, 21:187–203, 1968.
41. J. Nečas. *Les méthodes directes en théorie des équations elliptiques*. Masson, 1967.
42. C.W. Oosterlee, C. Vuik, W.A. Mulder, and R.-E. Plessix. Shifted-Laplacian preconditioners for heterogeneous Helmholtz problems. In B. Koren and C. Vuik, editors, *Advanced Computational Methods in Science and Engineering*, volume 71 of *Lecture Notes in Computational Science and Engineering*, pages 21–46. Springer, 2010.
43. F. Rellich. Darstellung der Eigenwerte von $\Delta u + \lambda u = 0$ durch ein Randintegral. *Mathematische Zeitschrift*, 46(1):635–636, 1940.
44. Y. Saad. *Iterative methods for sparse linear systems*. SIAM, 2nd edition, 2003.
45. Y. Saad and M. H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, 1986.
46. S. A. Sauter. A refined finite element convergence theory for highly indefinite helmholtz problems. *Computing*, 78(2):101–115, 2006.
47. A. H. Schatz. An observation concerning Ritz-Galerkin methods with indefinite bilinear forms. *Math. Comp*, 28(128):959–962, 1974.
48. L. R. Scott and S. Zhang. Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Mathematics of Computation*, 54(190):483–493, 1990.
49. A. H. Sheikh, D. Lahaye, and C. Vuik. On the convergence of shifted Laplace preconditioner combined with multilevel deflation. *Numerical Linear Algebra with Applications*, 20:645–662, 2013.
50. E. A. Spence. Wavenumber-explicit bounds in time-harmonic acoustic scattering. *SIAM J. Math. Anal.*, 46(4):2987–3024, 2014.
51. E. A. Spence, S. N. Chandler-Wilde, I. G. Graham, and V. P. Smyshlyaev. A new frequency-uniform coercive boundary integral equation for acoustic scattering. *Communications on Pure and Applied Mathematics*, 64(10):1384–1415, 2011.
52. E. A. Spence, I. V. Kamotski, and V. P. Smyshlyaev. Coercivity of combined boundary integral equations in high frequency scattering. *Comm. Pure Appl. Math.*, to appear, 2015.
53. M. E. Taylor. *Partial differential equations II: Qualitative Studies of Linear Equations*. Number 116 in Applied Mathematical Sciences. Springer, 1996.
54. N. Umetani, S. P. MacLachlan, and C. W. Oosterlee. A multigrid-based shifted Laplacian preconditioner for a fourth-order Helmholtz discretization. *Numer. Linear Algebra Appl.*, 16(8):603–626, 2009.
55. M. B. Van Gijzen, Y. A. Erlangga, and C. Vuik. Spectral analysis of the discrete Helmholtz operator preconditioned with a shifted Laplacian. *SIAM Journal on Scientific Computing*, 29(5):1942–1958, 2007.
56. H. Wu. Pre-asymptotic error analysis of CIP-FEM and FEM for the Helmholtz equation with high wave number. Part I: linear version. *IMA Journal of Numerical Analysis*, 34(3):1266–1288, 2014.