Is the Helmholtz equation really sign-indefinite?

Article

Accepted Version

Is the Helmholtz equation really sign-indefinite?

Andrea Moiola* Euan A. Spence†

July 10, 2013 (revised version)

Abstract

The usual variational (or weak) formulations of the Helmholtz equation are sign-indefinite in the sense that the bilinear forms cannot be bounded below by a positive multiple of the appropriate norm squared. This is often for a good reason, since in bounded domains under certain boundary conditions the solution of the Helmholtz equation is not unique at wavenumbers that correspond to eigenvalues of the Laplacian, and thus the variational problem cannot be sign-definite. However, even in cases where the solution is unique for all wavenumbers, the standard variational formulations of the Helmholtz equation are still indefinite when the wavenumber is large. This indefiniteness has implications for both the analysis and the practical implementation of finite element methods. In this paper we introduce new *sign-definite* (also called *coercive* or *elliptic*) formulations of the Helmholtz equation posed in either the interior of a star-shaped domain with impedance boundary conditions, or the exterior of a star-shaped domain with Dirichlet boundary conditions. Like the standard variational formulations, these new formulations arise just by multiplying the Helmholtz equation by particular test functions and integrating by parts.

Keywords: Helmholtz equation, high frequency, coercivity, sign-definiteness, Morawetz identity, frequency-explicit analysis, finite element method.

AMS subject classification: 35J05, 35J20, 65N30.

1 Introduction

The Helmholtz equation

$$\Delta u + k^2 u = 0, \tag{1.1}$$

with wavenumber $k > 0$, is arguably the simplest possible model of wave propagation. For example, if we look for solutions of the wave equation

$$\frac{\partial^2 U}{\partial t^2} - c^2 \Delta U = 0 \tag{1.2}$$

in the form $U(\mathbf{x}, t) = \Re\{u(\mathbf{x}) \mathrm{e}^{-\mathrm{i}\omega t}\}$, then the function $u(\mathbf{x})$ satisfies the Helmholtz equation (1.1) with $k = \omega/c$. Assuming a similar dependence on time reduces the Maxwell equations to the so-called time-harmonic Maxwell equations, and in certain situations these can be further reduced to the Helmholtz equation. Similarly, the time-harmonic elastic wave equation (often called the Navier equation) also reduces to the Helmholtz equation in certain circumstances. Because the Helmholtz equation is at the heart of linear wave propagation, much research effort has gone into both studying the properties of its solutions (for example their asymptotic behaviour as $k \to \infty$) and designing methods for computing them efficiently.

Many numerical methods for solving the Helmholtz equation are based on its standard variational (or weak) formulations, and these are sign-indefinite when k is large. In the literature, one often finds this sign-indefiniteness attributed to the Helmholtz equation itself; some recent examples of this attribution include the following:

[1]Department of Mathematics and Statistics, University of Reading, Whiteknights, PO Box 220, Reading, RG6 6AX, UK, A.Moiola@reading.ac.uk

[2]Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK, E.A.Spence@bath.ac.uk

"...the Helmholtz operator for scattering problems is a highly indefinite complex-valued linear operator." (2013)

"The main difficulty of the analysis is caused by the strong indefiniteness of the Helmholtz equation ..." (2009)

"Problems in high-frequency scattering of acoustic or electromagnetic waves are highly indefinite." (2013)

The goal of this paper is to introduce new *sign-definite* variational formulations of two frequently-encountered boundary value problems (BVPs) for the Helmholtz equation. These formulations can be obtained by multiplying the PDE by particular test functions and integrating by parts (just like the standard formulations). Thus, we aim to emphasise that, whereas the standard variational formulations of the Helmholtz equation are sign-indefinite, this sign-indefiniteness is *not* an inherent feature of the Helmholtz equation, only of its standard formulations.

1.1 Background: variational formulations of the Helmholtz equation

One of the most common variational problems is the following: given a real Hilbert space $\mathcal{V}$, a bilinear form $a(\cdot,\cdot) : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ and a continuous linear functional $F : \mathcal{V} \to \mathbb{R}$,

$$\text{find } u \in \mathcal{V} \text{ such that } a(u,v) = F(v) \text{ for all } v \in \mathcal{V}. \tag{1.3}$$

The particular variational problem that most mathematicians first encounter is that corresponding to the Dirichlet problem for Poisson's equation: given a bounded domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, and a real, square-integrable function f on Ω, find u such that

$$\Delta u = -f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega. \tag{1.4}$$

(In this paper we will always assume that the domains in which the PDEs are posed are Lipschitz; see, e.g., [39, Definition 1.2.1.1], [49, Definition 3.28] for the definition of a Lipschitz domain.) The variational problem associated with (1.4) is given by (1.3) with

$$a(u,v) := \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}\mathbf{x} \quad \text{and} \quad F(v) := \int_\Omega f\, v \, \mathrm{d}\mathbf{x}, \tag{1.5}$$

where the Hilbert space is $H^1_0(\Omega)$ (informally, functions in the Sobolev space $H^1(\Omega)$ that are zero on $\partial\Omega$) with inner product and norm

$$(u,v)_{H^1_0(\Omega)} := \int_\Omega (\nabla u \cdot \nabla v + uv) \, \mathrm{d}\mathbf{x}, \quad \|v\|^2_{H^1_0(\Omega)} := \|\nabla v\|^2_{L^2(\Omega)} + \|v\|^2_{L^2(\Omega)}.$$

This variational formulation is obtained by multiplying the PDE in (1.4) by a $v \in H^1_0(\Omega)$, integrating over Ω, and using Green's first identity

$$\int_\Omega v \Delta u \, \mathrm{d}\mathbf{x} = -\int_\Omega \nabla v \cdot \nabla u \, \mathrm{d}\mathbf{x} + \int_{\partial\Omega} v \frac{\partial u}{\partial n} \mathrm{d}s, \tag{1.6}$$

i.e. the divergence theorem applied to $v\nabla u$.

Returning to the general variational problem (1.3), ideally one would like to prove that there exist $C_c > 0$ and $\alpha > 0$ such that

$$|a(u,v)| \le C_c \|u\|_{\mathcal{V}} \|v\|_{\mathcal{V}} \quad \text{for all } u, v \in \mathcal{V}, \quad \text{(continuity)}, \tag{1.7}$$

$$|a(v,v)| \ge \alpha \|u\|^2_{\mathcal{V}} \quad \text{for all } v \in \mathcal{V}, \quad \text{(coercivity)}. \tag{1.8}$$

"Sign-definite" is used as a synonym for "coercive" (thus a variational problem is sign-indefinite if and only if it is not coercive). Note that several authors call property (1.8) "$\mathcal{V}$-ellipticity" (see, e.g. [21, §1], [46, §2.4.1], [68, Equation 2.43]) and use the word "coercivity" for the weaker property of satisfying a Gårding inequality ([46, §2.4.3], [68, Definition 2.1.54]).

If the two properties (1.7) and (1.8) can be established then there are three important consequences. The first is that the Lax–Milgram theorem implies that there exists a unique solution to (1.3), and this satisfies

$$\|u\|_{\mathcal{V}} \leq \frac{1}{\alpha} \|F\|_{\mathcal{V}'} . \tag{1.9}$$

The second and third consequences concern the Galerkin discretisation of the variational problem (1.3), namely, given $\mathcal{V}_N$, a finite dimensional subspace of $\mathcal{V}$ (with dimension N),

$$\text{find } u_N \in \mathcal{V}_N \text{ such that } a(u_N, v_N) = F(v_N) \text{ for all } v_N \in \mathcal{V}_N. \tag{1.10}$$

If continuity, (1.7), and coercivity, (1.8), hold then the Lax–Milgram theorem implies that the Galerkin solution u_N exists and is unique, and Céa's lemma implies that u_N satisfies

$$\|u - u_N\|_{\mathcal{V}} \leq \frac{C_c}{\alpha} \inf_{w_N \in \mathcal{V}_N} \|u - w_N\|_{\mathcal{V}} , \tag{1.11}$$

where C_c and α are as in (1.7) and (1.8) respectively (see, e.g., [12, §2.8]); the Galerkin method is then said to be *quasi-optimal*. The third consequence is that the finite dimensional matrix of the Galerkin method, $\mathbf{A}$, inherits analogous continuity and coercivity properties from the bilinear form:

$$|(\mathbf{A}\mathbf{u}, \mathbf{v})| \leq M_2 C_c \|\mathbf{u}\| \|\mathbf{v}\| \quad \text{and} \quad |(\mathbf{A}\mathbf{v}, \mathbf{v})| \geq M_1 \alpha \|\mathbf{v}\|^2, \quad \text{for all } \mathbf{u}, \mathbf{v} \in \mathbb{R}^N, \tag{1.12}$$

where M_1 and M_2 are constants depending on the discretisation (see Section 5.2). Coercivity in particular has important implications for the efficient solution of the linear system involving this matrix.

For the Dirichlet problem for Poisson's equation, (1.4), continuity of $a(\cdot, \cdot)$ follows from the Cauchy–Schwarz inequality, and coercivity follows from the Poincaré–Friedrichs inequality; the latter inequality being that $\|v\|_{H^1(\Omega)} \leq c \|\nabla v\|_{L^2(\Omega)}$ for some $c > 0$ for all $v \in H^1_0(\Omega)$, see, e.g., [12, §5.3], [36, §5.6.1, Theorem 3]. Therefore the variational problem (1.5) has a unique solution. Moreover, the Galerkin equations (1.10) have a unique solution for any subspace $\mathcal{V}_N \subset H^1_0(\Omega)$ and the Galerkin method is quasi-optimal. Furthermore, the fact that $a(\cdot, \cdot)$ is coercive and also symmetric (i.e. $a(u, v) = a(v, u)$) means that the linear system arising from the Galerkin method is positive definite, and thus can be solved efficiently by iterative solvers such as the conjugate gradient method, or multigrid (see, e.g., [11, Chapters 4 and 5], [32, Chapter 2]).

The situation for the Dirichlet problem for the Helmholtz equation, namely

$$\Delta u + k^2 u = -f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega, \tag{1.13}$$

with $k > 0$ and f a given real, square-integrable function, is very different. Indeed, the BVP (1.13) does not have a unique solution if $k^2 = \lambda_j$ for λ_j an eigenvalue of the negative Laplacian in Ω with zero Dirichlet boundary conditions. Proceeding as before, we multiply the PDE in (1.13) by $v \in H^1_0(\Omega)$, integrate over Ω and use Green's first identity, and obtain the variational problem (1.3) with $F(\cdot)$ as in (1.5) but now with $a(\cdot, \cdot)$ replaced by $a_D(\cdot, \cdot)$ defined by

$$a_D(u, v) := \int_\Omega \left(\nabla u \cdot \nabla v - k^2 uv \right) \mathrm{d}\mathbf{x} \tag{1.14}$$

(with the subscript D standing for "Dirichlet"). For the Helmholtz equation it is convenient to use the k-dependent inner product and norm

$$(u, v)_{1,k,\Omega} := \int_\Omega \left(\nabla u \cdot \nabla v + k^2 uv \right) \mathrm{d}\mathbf{x}, \quad \|v\|^2_{1,k,\Omega} := \|\nabla v\|^2_{L^2(\Omega)} + k^2 \|v\|^2_{L^2(\Omega)} \tag{1.15}$$

on the space $H^1_0(\Omega)$. Continuity of $a_D(\cdot, \cdot)$ follows as before using the Cauchy–Schwarz inequality, but now

$$a_D(v, v) = \int_\Omega |\nabla v|^2 \, \mathrm{d}\mathbf{x} - k^2 \int_\Omega |v|^2 \, \mathrm{d}\mathbf{x}.$$

It is clear that $a_D(v, v)$ cannot be bounded below by $\|v\|^2_{1,k,\Omega}$ for all $k > 0$; indeed, if $k^2 = \lambda_j$ (the j-th eigenvalue of the negative Laplacian with Dirichlet boundary conditions) then $a_D(u_j, u_j) = 0$, for u_j

the corresponding eigenfunction. Furthermore, if $k^2 > \lambda_1$ then the bilinear form takes both positive and negative real values. Indeed, if j is such that $\lambda_j > k^2 > \lambda_1$ then $a_D(u_j, u_j) > 0 > a_D(u_1, u_1)$; thus the bilinear form is not coercive by [8, Propositions 3.2 and 3.3].

Although $a_D(\cdot,\cdot)$ is not coercive, it satisfies a Gårding inequality, i.e. adding a multiple of $\|v\|^2_{L^2(\Omega)}$ to $a_D(v,v)$ makes it larger than $\|v\|^2_{1,k,\Omega}$, since

$$a_D(v,v) + 2k^2 \|v\|^2_{L^2(\Omega)} = \|v\|^2_{1,k,\Omega}\,. \tag{1.16}$$

Even though we no longer have coercivity, can we recover any of its three consequences described above (existence and uniqueness, quasi-optimality, and sign-definiteness of the discretised linear system)? Classic Fredholm theory implies that if k^2 is not an eigenvalue of the negative Laplacian, then a solution to the variational problem exists and is unique (this relies on the compact embedding of $H^1_0(\Omega)$ in $L^2(\Omega)$, see, e.g., [36, §6.2.3], [68, Theorem 2.10.4]). However, although this method does give a bound on u in terms of f, this bound is not explicit in k. One can also show that, given a suitable finite dimensional subspace $\mathcal{V}_N$, the Galerkin equations (1.10) have a solution which satisfies

$$\|u - u_N\|_{\mathcal{V}} \leq \widetilde{C} \inf_{w_N \in \mathcal{V}_N} \|u - w_N\|_{\mathcal{V}} \tag{1.17}$$

for some $\widetilde{C} > 0$, *provided the subspace dimension N is large enough* (see, e.g., [68, Theorem 4.2.9]). However, it is very difficult to find out how the threshold for N and constant $\widetilde{C}$ in (1.17) depend on k. Finally, the Galerkin matrix for this problem is still symmetric, as in the Poisson case, but is no longer positive definite, having both positive and negative eigenvalues when k^2 is sufficiently large. This fact, coupled also with difficulties if k^2 is close to an eigenvalue of the Laplacian, mean that it is harder to solve the linear system arising from the Helmholtz bilinear form (1.14) than the Poisson one (1.5).

Although the Helmholtz equation in Ω with Dirichlet boundary conditions is not well-posed for every k, the solution under impedance boundary conditions, i.e.

$$\Delta u + k^2 u = -f \quad \text{in } \Omega, \qquad \frac{\partial u}{\partial n} - \mathrm{i}ku = g \quad \text{on } \partial\Omega, \tag{1.18}$$

where f and g are given square-integrable functions, exists and is unique for every real $k \neq 0$; this is because the eigenvalues of the Laplacian with impedance boundary conditions are not real. How will considering the Helmholtz equation under impedance boundary conditions instead of Dirichlet boundary conditions change the properties of the associated variational formulation? Immediate differences are that, since the boundary conditions involve the imaginary unit "i", the variational formulation of this BVP involves complex-valued Sobolev spaces, a sesquilinear form $a(\cdot,\cdot) : \mathcal{V} \times \mathcal{V} \to \mathbb{C}$ instead of a bilinear form, and an antilinear functional $F(\cdot) : \mathcal{V} \to \mathbb{C}$. Multiplying the PDE in (1.18) by $\overline{v}$, integrating over Ω, and using Green's first identity and the impedance boundary condition, we obtain the variational problem (1.3) with

$$a_I(u,v) := \int_\Omega \left(\nabla u \cdot \overline{\nabla v} - k^2 u\overline{v}\right) \mathrm{d}\mathbf{x} - \mathrm{i}k \int_{\partial\Omega} u\overline{v}\,\mathrm{d}s, \qquad F(v) := \int_\Omega f\overline{v}\,\mathrm{d}\mathbf{x} + \int_{\partial\Omega} g\overline{v}\,\mathrm{d}s. \tag{1.19}$$

The appropriate Hilbert space is now $H^1(\Omega)$ with norm and inner product given by (1.15), replacing v by $\overline{v}$ in the integral. Continuity of $a_I(\cdot,\cdot)$ follows in a similar way to before (although since $a_I(\cdot,\cdot)$ now involves an integral over $\partial\Omega$ we also need to use the continuity of the trace map from $H^1(\Omega)$ to $L^2(\partial\Omega)$). The arguments that show that $a_D(\cdot,\cdot)$ is not coercive also show that $a_I(\cdot,\cdot)$ is not coercive for $k^2 \geq \lambda_1$; this is because the integral over $\partial\Omega$ in $a_I(v,v)$ is zero if v is a Dirichlet eigenfunction of the negative Laplacian. Since

$$\Re a_I(v,v) = \int_\Omega |\nabla v|^2\,\mathrm{d}\mathbf{x} - k^2 \int_\Omega |v|^2\,\mathrm{d}\mathbf{x},$$

the real part of $a_I(\cdot,\cdot)$ satisfies (1.16), and thus $a_I(\cdot,\cdot)$ satisfies a Gårding inequality. Fredholm theory can then be applied, as in the case of the Dirichlet problem, to show that a solution to the variational problem exists, and, furthermore, that given a finite dimensional subspace $\mathcal{V}_N$ the Galerkin solution u_N exits, is unique, and satisfies (1.17), provided that N is greater than some

threshold. Again, this classic theory gives no information about how the constants depend on k, but this dependence has been quantified using more sophisticated techniques in [50, Proposition 8.2.7] (for the h-version of the finite element method) and [53], [54], [52] (for the hp-version). Finally, regarding the linear system: this is sign-indefinite (as in the Dirichlet case) and non-Hermitian (because the boundary condition involves the imaginary unit "i", and therefore $a_I(u,v) \neq \overline{a_I(v,u)}$); thus the eigenvalues are complex and lie on both sides of the imaginary axis. These facts are not the only reasons why it is difficult to solve the linear systems associated with Helmholtz problems, but they contribute strongly to this difficulty; see the reviews [34], [35], [33], [1] and the references therein for more details.

In summary, in moving from Dirichlet boundary conditions to impedance boundary conditions, even though we gain well-posedness of the Helmholtz equation for every k, *we still keep the sign-indefiniteness of the sesquilinear form.* The main aim of this paper is to show that it *is* possible to have a sign-definite, i.e. coercive, formulation of the Helmholtz equation under impedance boundary conditions (at least for a wide class of domains, namely star-shaped domains), if one is prepared to modify the space $H^1(\Omega)$ and the sesquilinear form $a_I(\cdot,\cdot)$ (this formulation is presented in §1.3).

We note at this stage that other coercive formulations of the Helmholtz impedance problem do exist. We discuss these in more detail in §1.2 below, but emphasise here that for these formulations at least one of the following is true: *(i)* the formulation is an integral equation on $\partial\Omega$, *(ii)* the formulation requires restricting $\mathcal{V}$ to include only piecewise-solutions of the homogeneous Helmholtz equation (so-called "operator-adapted" or "Trefftz" spaces), *(iii)* the formulation is a least-squares formulation (under which *any* well-posed linear BVP is coercive). In contrast, the formulation introduced in this paper is a formulation in Ω (not on the boundary $\partial\Omega$), does not require operator-adapted spaces, and is not a least-squares formulation.

1.2 Existing coercive formulations of the Helmholtz equation

In §1.1 we discussed the most basic variational formulation of the interior impedance problem (1.19) and saw that this formulation was sign-indefinite. Of course, there are many different ways of formulating BVPs involving the Helmholtz equation; the vast majority of these, however, are also sign-indefinite. There do in fact exist a few coercive formulations, which we now briefly outline. We also discuss some formulations that are not coercive, but enjoy some of the benefits of coercivity.

Integral equation formulations Since closed-form expressions for the fundamental solution of the Helmholtz equation exist, a popular way of solving Helmholtz BVPs is by reformulating them as integral equations on the boundary of the domain; this is the so-called boundary integral method. This approach is especially popular when considering problems posed in unbounded domains, since it exchanges a problem on a d-dimensional infinite domain for one on a $(d-1)$-dimensional finite domain.

- The standard second-kind integral operator used to solve the Dirichlet problem in the exterior of a bounded obstacle (the so-called "combined potential" or "combined field" operator for this problem) is coercive for a variety of domains when k is large enough [71, Theorem 1.2], [29, Theorems 4.2 and 4.12], [8]. By standard properties of integral equations, this integral operator can also be used to solve the interior impedance problem (1.18) (see, e.g., [17, Corollary 2.28 and Theorem 2.30])

- A modification of the standard combined potential operator for the exterior Dirichlet problem, the so-called "star-combined operator", is coercive for all $k > 0$ for all Lipschitz domains that are star-shaped with respect to a ball [70, Theorem 1.1].

- A modification of the standard combined potential operator for the exterior Neumann problem is coercive for the circle and sphere when k is large enough [10, Theorem 3.6].

- In the case of scattering by a flat screen, the standard first-kind integral equations for both the Dirichlet and Neumann problems are coercive for all $k > 0$ [40, Theorem 2], [18].

Trefftz-discontinuous Galerkin methods Since approximating highly-oscillatory solutions of the Helmholtz equation with piecewise polynomials requires large numbers of degrees of freedom, many methods have been proposed that seek to approximate solutions of the Helmholtz equation with oscillatory basis functions. One of the main classes of these "wave-based" methods are *Trefftz methods*, which use basis functions that are locally (i.e. inside each mesh element) solutions of the Helmholtz equation. One of the main examples of such a method is the Ultra Weak Variational Formulation (UWVF) [15], [16], which can be recast as a special discontinuous Galerkin (DG) method.

For such "Trefftz-discontinuous Galerkin (TDG) methods" applied to either the interior impedance problem (1.18) or the exterior Dirichlet problem as formulated in Definition 4.2 below, the associated sesquilinear form is continuous and coercive in a norm consisting of jumps of functions over element edges/faces [44, §3.1], [56, §4.3], [14, Lemma 3.4] (a slightly weaker result was proved in the original analysis of the UWVF; see [16, Lemma 3.3, Equation 3.30]). Error estimates in a mesh-independent norm (such as the $L^2(\Omega)$ norm) can then be obtained by using a duality argument.

Least-squares methods As we saw in §1.1, the best possible variational problem involves a symmetric, coercive, sesquilinear form, as in the case of the Dirichlet problem for Poisson's equation. Least-squares finite element methods can be viewed as an attempt to recover this situation for non-symmetric or indefinite problems. Indeed, the standard least-squares formulation of *any* well-posed BVP for *any* linear PDE with linear boundary conditions leads to a symmetric, coercive, sesquilinear form [9, §2.2.1, §3.2]. This is not the end of the story, however, since there are then subtle questions about which norms to choose for the least-squares functionals.

In the least-squares framework, second order PDEs are usually converted into first-order systems to reduce the condition number of the discretised problem. A standard first-order system reformulation of the Helmholtz exterior Dirichlet problem was considered in [48]. The authors proved that this formulation was well-posed, and hence coercive, but did not determine how the coercivity constant depends on k; this dependence can in principle be determined using the k-explicit bounds on the solution of the Helmholtz equation that have recently been obtained.

A new variational formulation of the Helmholtz equation as a first-order system was recently introduced in [26]. This "Discontinuous Petrov Galerkin (DPG)" method can be thought of as a least-squares method in a non-standard inner product. Using k-explicit bounds on the solution of the Helmholtz interior impedance problem, a fully k-explicit analysis of the "theoretical" version of this method is given in [26], whereas a k-explicit analysis of the "practical" version is still lacking. For this latter version, the matrix of the Galerkin discretisation is only positive-semidefinite instead of positive-definite.

A quadratic functional for the exterior impedance problem In [27] (see also [28]) Després showed that there exists a quadratic functional that is minimised by the solution of the exterior impedance problem for the Helmholtz equation. This functional acts on solutions of the homogeneous Helmholtz equation in the exterior of the obstacle, and involves the impedance traces of the solution (i.e. $\partial v/\partial n \pm \mathrm{i}kv$), its outgoing and incoming far-field patterns, and the function prescribed in the impedance boundary condition; see [27, Theorem 3.1], [28, Theorem 1], [5, Proposition 3.1]. The analogue of this functional for the corresponding problem for the time-harmonic Maxwell equations was introduced in [23, Theorem 1].

This functional can then be used to define a variational problem satisfied by the solution of the exterior impedance problem, with a sesquilinear form that is continuous and coercive on the space of impedance traces and far-field patterns of Helmholtz solutions. Alternatively, one can think of the far-field patterns as continuous functions of the impedance traces, and obtain a continuous and coercive variational formulation on the space of impedance traces (although in this case it is not clear how the continuity constant depends on k).

T-coercivity Any well-posed variational problem of the form (1.3) is coercive *if* one is allowed to introduce another bounded linear operator into the sesquilinear form. That is, if the variational problem (1.3) has a unique solution that depends continuously on $F(\cdot)$ (or equivalently $a(\cdot,\cdot)$

satisfies an inf-sup condition [68, §2.1.6]), then there exists a $T : \mathcal{V} \to \mathcal{V}$ and an $\alpha' > 0$ such that

$$|a(v, Tv)| \geq \alpha' \|v\|_{\mathcal{V}}^2 \qquad \text{for all } v \in \mathcal{V};$$

see [68, Remark 2.1.48], [20, Theorem 1]. This reformulation only yields the advantages of coercivity, however, if the variational problem is sufficiently simple for T to be known explicitly. In the case of the standard variational formulation of the Helmholtz Dirichlet problem, the operator T can be expressed in terms of eigenspace projectors and thus approximated by discrete operators on sufficiently fine meshes; the size of the meshwidth threshold, however, is not clear [22, §3].

Interior penalty methods Finally, recall that interior penalty methods arise by adding terms to the appropriate sesquilinear forms to penalise jumps of various quantities over interfaces between elements of a mesh. Although the variational formulations of these methods are not coercive, for certain methods some of the consequences of coercivity hold; namely, the Galerkin equations have a unique solution without any constraint on the dimension of the (piecewise polynomial) approximation space, and error estimates can be obtained that are explicit in k, h, and p [37, Remarks 4.3 and 5.1], [38, Remark 3.2], [74, Corollary 3.5, Theorem 4.4]. For the interior penalty discontinuous Galerkin (IPDG) methods introduced in [37] and [38], the penalty terms are added so that the properties just highlighted can be proved using Rellich-type identities. (These methods share a conceptual link with the new variational formulations introduced in this paper, since, as we see in §1.4 below, the new formulations in this paper are designed using closely-related Morawetz-type identities.) Adding a penalty term to the standard variational formulation (1.19) was considered in [74]. For this formulation the properties above are proved for subspaces consisting of piecewise-linear polynomials using the fact that functions in these subspaces satisfy Laplace's equation on each element, and then using Green's identity for Laplace's equation (i.e. (1.6)).

1.3 A new coercive variational formulation of the Helmholtz equation

Given a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, $f \in L^2(\Omega)$, and $g \in L^2(\partial\Omega)$, consider the problem of solving the Helmholtz equation in Ω subject to an impedance boundary condition:

$$\Delta u + k^2 u = -f \qquad \text{in } \Omega, \tag{1.20a}$$

$$\frac{\partial u}{\partial n} - \mathrm{i}ku = g \qquad \text{on } \partial\Omega. \tag{1.20b}$$

We now present a new variational formulation of this interior impedance problem. We also consider in §4 the sound-soft scattering problem for the Helmholtz equation, i.e. (1.20a) posed in the *exterior* of a bounded domain with Dirichlet boundary conditions (see Definitions 4.1 and 4.2 below), and the results outlined below for the interior impedance problem have counterparts for this exterior problem.

As we reviewed in §1.1, a variational formulation has three ingredients: a Hilbert space, a sesquilinear form, and an antilinear functional. The Hilbert space of the new formulation is defined by

$$V := \left\{ v : \, v \in H^1(\Omega), \; \Delta v \in L^2(\Omega), \; v \in H^1(\partial\Omega), \frac{\partial v}{\partial n} \in L^2(\partial\Omega) \right\} \tag{1.21}$$

with norm

$$\begin{aligned} \|v\|_V^2 := {} & k^2 \|v\|_{L^2(\Omega)}^2 + \|\nabla v\|_{L^2(\Omega)}^2 + k^{-2} \|\Delta v\|_{L^2(\Omega)}^2 \\ & + L \left(k^2 \|v\|_{L^2(\partial\Omega)}^2 + \|\nabla_{\partial\Omega} v\|_{L^2(\partial\Omega)}^2 + \left\| \frac{\partial v}{\partial n} \right\|_{L^2(\partial\Omega)}^2 \right), \end{aligned} \tag{1.22}$$

where L is the diameter (or some other characteristic length scale) of the domain and $\nabla_{\partial\Omega}$ is the surface gradient on $\partial\Omega$ (recall that $\nabla_{\partial\Omega}$ is such that if v is differentiable in a neighbourhood of $\partial\Omega$ then

$$\nabla_{\partial\Omega} v = \nabla v - \mathbf{n} \frac{\partial v}{\partial n}$$

on $\partial\Omega$, where $\mathbf{n} = \mathbf{n}(\mathbf{x})$ is the outward-pointing unit normal vector at the point $\mathbf{x} \in \partial\Omega$). We weight the derivatives by k and include L in front of the boundary terms so that, when computed for solutions of the homogeneous Helmholtz equation with wavenumber k, each term of the norm scales in the same way as k and L vary; see Remark 3.8 below.

Although this space may appear strange, standard regularity results imply that if $u \in H^1(\Omega)$ is the solution to (1.20) then $u \in V$; see Proposition 3.2. In addition, we show below that $V \subset H^{3/2}(\Omega)$, and for classical solutions of the homogeneous Helmholtz equation these two spaces are in fact equivalent (i.e. if $v \in C^2(\Omega)$ is such that $\Delta v + k^2 v = 0$, then $v \in H^{3/2}(\Omega)$ implies that $v \in V$); see Remark 3.7.

Define the sesquilinear form $b : V \times V \to \mathbb{C}$ by

$$\begin{aligned} b(u,v) := \int_\Omega \left(\nabla u \cdot \overline{\nabla v} + k^2 u \overline{v} + \left(\mathcal{M}u + \frac{1}{3k^2}\mathcal{L}u \right) \overline{\mathcal{L}v} \right) \mathrm{d}\mathbf{x} \\ - \int_{\partial\Omega} \left(\mathrm{i}ku\,\overline{\mathcal{M}v} + \left(\mathbf{x}\cdot\nabla_{\partial\Omega} u - \mathrm{i}k\beta u + \frac{d-1}{2}u \right) \overline{\frac{\partial v}{\partial n}} + (\mathbf{x}\cdot\mathbf{n}) \left(k^2 u\overline{v} - \nabla_{\partial\Omega} u \cdot \overline{\nabla_{\partial\Omega} v} \right) \right) \mathrm{d}s, \end{aligned} \tag{1.23}$$

and the functional $G : V \to \mathbb{C}$ by

$$G(v) := \int_\Omega \left(\overline{\mathcal{M}v} - \frac{1}{3k^2}\overline{\mathcal{L}v} \right) f \,\mathrm{d}\mathbf{x} + \int_{\partial\Omega} \overline{\mathcal{M}v}\, g \,\mathrm{d}s, \tag{1.24}$$

where β is an arbitrary real constant, d is the spatial dimension,

$$\mathcal{L}u := \Delta u + k^2 u, \quad \text{and} \quad \mathcal{M}u := \mathbf{x}\cdot\nabla u - \mathrm{i}k\beta u + \frac{d-1}{2}u.$$

The sesquilinear form $b(\cdot,\cdot)$ and functional $G(\cdot)$ are defined in this way because if u is the solution to the BVP (1.20), then

$$b(u,v) = G(v) \quad \text{for all } v \in V; \tag{1.25}$$

this is not obvious, and we explain why below (in §1.4 and Proposition 3.2).

Using the Cauchy–Schwarz inequality, it is straightforward to show that the sesquilinear form $b(\cdot,\cdot)$ is continuous on V, i.e. (1.7) holds with $\mathcal{V} = V$. The explicit value of the constant C_c is given in Lemma 3.3 below; in particular, if β is independent of k (as we choose it to be below), $C_c \sim k$ as $k \to \infty$.

The main novelty of $b(\cdot,\cdot)$ is that, for some domains, it is coercive on V:

Theorem 1.1. *Let Ω be a Lipschitz domain with diameter L that is star-shaped with respect to a ball, i.e. there exists a $\gamma > 0$ such that*

$$\mathbf{x}\cdot\mathbf{n}(\mathbf{x}) \geq \gamma L$$

for all $\mathbf{x} \in \partial\Omega$ such that $\mathbf{n}(\mathbf{x})$ exists (see Remark 3.5 for how this is related to the usual definition of star-shapedness). If the arbitrary constant β is chosen such that

$$\beta \geq \frac{L}{2}\left(1 + \frac{4}{\gamma} + \frac{\gamma}{2}\right)$$

then, for any $k > 0$,

$$\Re b(v,v) \geq \frac{\gamma}{4}\|v\|_V^2 \quad \text{for all } v \in V, \tag{1.26}$$

i.e. $b(\cdot,\cdot)$ is coercive on V with constant $\gamma/4$.

Following the discussion in §1.1 we know there are three immediate consequences of this result:

1. The variational problem (1.25) has a unique solution which satisfies $\|u\|_V \leq (4/\gamma)\,\|G\|_{V'}$.
2. The Galerkin method applied to (1.25) has a unique solution for *any* finite dimensional subspace $V_N \subset V$ and is quasi-optimal, with an explicit bound for the constant of quasi-optimality given by $4C_c/\gamma$.

3. The matrix of the linear system resulting from the Galerkin method is also coercive (in the sense of (1.12)) with an explicit value for the coercivity constant. In particular, the inequality (1.26) implies that the Galerkin matrix has positive definite Hermitian part.

Regarding **1.**: this is the least interesting consequence, since we already have existence and uniqueness of the solution to the BVP (1.20) from the standard variational formulation and Fredholm theory (although it is perhaps interesting that we can get these results in this alternative way). It is straightforward to bound $\|G\|_{V'}$ in terms of the L^2-norms of f and g; see Remark 3.6. However, the resulting bound on $\|u\|_V$ was already essentially proved in [50, Proposition 8.1.4] for $d = 2$ and [24, Theorem 1] for $d = 3$.

Regarding **2.**: this is interesting because, as discussed in §1.1, establishing quasi-optimality of the Galerkin method for the standard variational formulation (1.19) with all the constants (including the threshold for quasi-optimality to hold) explicit in k is a challenging problem. Note that for the standard variational formulation (1.19) there are in fact two k-dependent thresholds for the subspace dimension N: one for the bound (1.17) to hold, and one for the best approximation error on the right-hand side to be small (the latter depends on the particular $\mathcal{V}_N$ and is a consequence of the fact that solutions of the Helmholtz equation are highly oscillatory). The new formulation eliminates the first threshold, but the second one still remains (since it is a consequence of approximation theory and independent of the variational formulation).

The main disadvantage of the new formulation is that the space V includes the requirement $\Delta v \in L^2(\Omega)$. This means that the standard C^0 finite element spaces of $H^1(\Omega)$ are not subspaces of V, and in fact, any finite element space that is a subspace of V (i.e. the elements are conforming) must also be a subspace of $C^1(\overline{\Omega})$ (see §5.1). Of course, there are several well-known piecewise-polynomial finite element spaces consisting of C^1-elements (originally designed for solving the biharmonic equation) that could then be used in the new formulation to give a conforming method. We discuss this more in Section 5.

Regarding **3.**: as discussed briefly in §1.1, solving the Helmholtz equation with iterative methods is difficult, and a contributing factor is the sign-indefiniteness of the standard variational formulations. Whether the new formulation can alleviate some of this difficulty remains to be seen and will require a detailed, separate investigation. However, as a start, in §5.2 we investigate whether we can determine anything a priori about how the Generalised Minimal Residual method (GMRES) behaves when it is applied to the linear systems arising from the new formulation (without any preconditioning).

1.4 The idea behind the new formulation

As we saw in §1.1, the standard variational formulation of the interior impedance BVP for the Helmholtz equation (1.20) is based on integrating over Ω the identity

$$\overline{v}\mathcal{L}u = \nabla \cdot \left[\overline{v}\nabla u\right] - \nabla u \cdot \overline{\nabla v} + k^2 u\overline{v}, \tag{1.27}$$

where $\mathcal{L}u := \Delta u + k^2 u$. (This is the differential form, as opposed to the integrated form, of Green's first identity for the Helmholtz equation.)

The new variational formulation (1.25) comes from integrating over Ω the identity

$$\overline{\mathcal{M}v}\mathcal{L}u + \mathcal{M}u\overline{\mathcal{L}v} = \nabla \cdot \left[\overline{\mathcal{M}v}\,\nabla u + \mathcal{M}u\,\overline{\nabla v} + \mathbf{x}(k^2 u\overline{v} - \nabla u \cdot \overline{\nabla v})\right] - \nabla u \cdot \overline{\nabla v} - k^2 u\overline{v}, \tag{1.28}$$

where the multiplier $\mathcal{M}$ is defined by

$$\mathcal{M}v := \mathbf{x} \cdot \nabla v - \mathrm{i}k\beta v + \frac{d-1}{2}v, \tag{1.29}$$

and β is an arbitrary real number.

The key point is the following. When $u = v$, the non-divergence terms of (1.27) equal $-|\nabla v|^2 + k^2|v|^2$, and this expression is *not* single-signed (i.e. for some v it will be positive, and for some v it will be negative). However, when $u = v$, the non-divergence terms of (1.28) equal $-|\nabla v|^2 - k^2|v|^2$ and this expression *is* single-signed. Therefore, just as the identity (1.27) gives

rise to the standard, sign-indefinite, variational formulation of the interior impedance problem, the identity (1.28) can be used as the basis of a new, sign-definite, variational formulation.

Although, to the authors' knowledge, the precise identity (1.28) has not been written down before, it arises naturally from existing ideas, which we now briefly explain. (We focus on the ideas, and give the details of the calculations in §2 below.)

Green's first identity arises from multiplying $\mathcal{L}u$ by $\overline{v}$, for v an arbitrary test function, and Rellich-type identities arise from multiplying $\mathcal{L}u$ by a derivative of $\overline{v}$, most commonly $\mathbf{x}\cdot\overline{\nabla v}$. For the Laplace operator, multiplying by $\mathbf{x}\cdot\overline{\nabla v}$ yields the identity

$$(\mathbf{x}\cdot\overline{\nabla v})\Delta u = \nabla\cdot\Big[(\mathbf{x}\cdot\overline{\nabla v})\,\nabla u\Big] - \nabla u\cdot\overline{\nabla v} - \nabla u\cdot\big((\mathbf{x}\cdot\nabla)\overline{\nabla v}\big), \tag{1.30}$$

which is, in some sense, an analogue of Green's first identity for the Laplace operator (i.e. (1.27) with $k=0$) with a different multiplier. However, (1.27) when $k=0$ and (1.30) differ in the following two important respects. *(i)* When $u=v$ the one non-divergence term on the right-hand side of (1.27) with $k=0$ is single-signed (since it equals $-|\nabla v|^2$). On the other hand, when $u=v$ the non-divergence terms on the right-hand side of (1.30) are not single-singled, since they equal $-|\nabla v|^2 - \nabla v\cdot\big((\mathbf{x}\cdot\nabla)\overline{\nabla v}\big)$. *(ii)* The non-divergence term on the right-hand side of (1.27) when $k=0$ involves only first derivatives of u and v, whereas the second non-divergence term on the right-hand side of (1.30) involves second derivatives of v.

Because of these two considerations, we want to get rid of $-\nabla u\cdot\big((\mathbf{x}\cdot\nabla)\overline{\nabla v}\big)$ on the right-hand side of (1.30). If we add to (1.30) the analogous expression with $\overline{v}$ and u swapped, we can use the identity

$$\nabla u\cdot\big((\mathbf{x}\cdot\nabla)\overline{\nabla v}\big) + \overline{\nabla v}\cdot\big((\mathbf{x}\cdot\nabla)\nabla u\big) = \nabla\cdot\Big[\mathbf{x}\,\nabla u\cdot\overline{\nabla v}\Big] - d\,\nabla u\cdot\overline{\nabla v} \tag{1.31}$$

to express the two undesirable terms as the sum of a divergence and a term with a constant sign when $u=v$. We thus arrive at

$$(\mathbf{x}\cdot\overline{\nabla v})\Delta u + (\mathbf{x}\cdot\nabla u)\overline{\Delta v} = \nabla\cdot\Big[(\mathbf{x}\cdot\overline{\nabla v})\,\nabla u + (\mathbf{x}\cdot\nabla u)\,\overline{\nabla v} - \mathbf{x}\,\nabla u\cdot\overline{\nabla v}\Big] + (d-2)\nabla u\cdot\overline{\nabla v}, \tag{1.32}$$

which, in some sense, is an analogue of Green's second identity for the Laplacian,

$$\overline{v}\Delta u - u\overline{\Delta v} = \nabla\cdot\big[\overline{v}\nabla u - u\overline{\nabla v}\big], \tag{1.33}$$

since it involves both Δu and Δv. (The identity (1.32) appears as [55, Equation 2.5] and its generalisation from the Laplacian to a general 2nd order differential operator $\sum_{i,j}\partial_i(A_{ij}\partial_j)$ and from $\mathbf{x}$ to an arbitrary vector field is given in [49, Lemma 4.22].)

Having obtained the identity (1.32) involving the Laplace operator, it is then relatively straightforward to obtain the following identity involving the Helmholtz operator

$$\begin{aligned}(\mathbf{x}\cdot\overline{\nabla v})\mathcal{L}u + (\mathbf{x}\cdot\nabla u)\overline{\mathcal{L}v} = \nabla\cdot\Big[&(\mathbf{x}\cdot\overline{\nabla v})\,\nabla u + (\mathbf{x}\cdot\nabla u)\,\overline{\nabla v} + \mathbf{x}(k^2u\overline{v} - \nabla u\cdot\overline{\nabla v})\Big]\\ &+ (d-2)\nabla u\cdot\overline{\nabla v} - dk^2u\overline{v}\end{aligned} \tag{1.34}$$

(the details are in §2). This identity with $v=u$ was originally obtained by Rellich [67] and has been used extensively in the analysis of both the Laplace and the Helmholtz equations (with suitable generalisations also used to study higher order elliptic PDEs). For example, Rellich introduced (1.34) with $v=u$ in order to obtain an expression for the eigenvalues of the Laplacian as an integral over $\partial\Omega$ (instead of the usual expression as an integral over Ω used in, e.g., the Rayleigh-Ritz method), and these identities have been used to further study eigenvalues of equations involving the Laplacian in, e.g., [64], [66], [42], [3], [4]. Rellich-type identities have been well-used by the harmonic analysis community (see, e.g., [47, Lemma 2.1.13 and §10 of Chapter 2], [73, Lemma 2.2]), and used more recently by the numerical analysis community to prove k-explicit bounds on the solution to (1.20) and related BVPs (see, e.g., [50, Proposition 8.1.4], [24], [43], [19], [45]); some of this recent work is discussed in Remarks 3.6 and 4.7 below. (The recent review [17, §5.3] explains why Rellich-type identities can be used to do these things.)

Looking to use the identity (1.34) as the basis of a new variational formulation of the Helmholtz equation, we see that the non-divergence terms on the right-hand side of (1.34), namely $(d-2)\nabla u\cdot$

$\overline{\nabla v} - dk^2 u\overline{v}$, involve only first derivatives of u and v, and each term is single-signed when $u = v$. However for $d = 3$ the signs are opposite to each another, and for $d = 2$ we lose the $\nabla u \cdot \overline{\nabla v}$ term and thus have no hope of getting coercivity in a norm involving $|\nabla v|^2$. To remedy these difficulties, we add terms into the multiplier $\mathbf{x} \cdot \overline{\nabla v}$ to obtain the multiplier $\overline{\mathcal{M}v}$ defined by (1.29), and similarly for $\mathbf{x} \cdot \nabla u$, with this process eventually yielding the identity (1.28). Both the non-divergence terms on the right-hand side of (1.28) are now non-zero and have the same sign when $u = v$. This is not the only requirement for coercivity of the resulting sesquilinear form: we also need to control the term involving $\overline{\mathcal{L}v}$ on the left-hand side, as well as the divergence terms (which become integrals over $\partial\Omega$ when (1.28) is integrated over Ω), but these other requirements can ultimately also be achieved (making use of the star-shapedness of Ω); see the proof of Theorem 3.4 for the details.

This idea of adding terms to the $\mathbf{x} \cdot \overline{\nabla v}$ multiplier (which can also be seen as taking certain linear combinations of the Rellich and Green multipliers) goes back to Morawetz (in [58] for the wave equation and in [61] for the Helmholtz equation), and the identity (1.28) with $v = u$ essentially appears in [61] and [60] (see Remark 2.3 for more details). These identities were used by Morawetz to prove bounds on solutions to the wave and Helmholtz equations, and have since been used in a variety of other contexts (see, e.g., [59], [25], [65]), including recently in a numerical analysis context by [70] and [71].

Why did we write the multiplier $\mathcal{M}v$ in the particular form (1.29), with a k multiplying the constant β? The reason is that if our multiplier were $\mathcal{M}v = \mathbf{x} \cdot \nabla v - \mathrm{i}\widetilde{\beta}v + (d-1)v/2$ then we would need to take $\widetilde{\beta} \gtrsim k$ to obtain coercivity with a constant independent of k. Under this restriction, the continuity constant is minimised by $\widetilde{\beta} \sim k$; therefore, it is natural to make this k-dependence of $\widetilde{\beta}$ explicit by letting $\widetilde{\beta} = k\beta$ (with β then chosen to be independent of k for coercivity). The multiplier $\mathcal{M}u$ with β a function of $\mathbf{x}$, can be used to prove bounds on solutions of the Helmholtz equation in exterior domains (see [61]) and in this case β needs to be taken to be independent of k. The reason for this is that, for this application, $\mathcal{M}u$ must be proportional to the first three terms in the large-$|\mathbf{x}|$ asymptotics of solutions of the Helmholtz equation satisfying the Sommerfeld radiation condition (see [61], [70, Remark 2.3]). While this link with the radiation condition explains, to a certain extent, the choice $\beta \sim 1$ for exterior problems, it is less clear why β should be taken to be independent of k to obtain a coercive formulation of the interior impedance problem (without going through the calculations). One possible explanation is that the multiplier should try to, in some sense, mimic the impedance boundary condition (1.20b). Indeed, in Section 3 below we consider the more general impedance boundary condition $\partial u/\partial n - \mathrm{i}k\vartheta u = g$, with ϑ an arbitrary function, and in this case the optimal β is independent of k, but depends on ϑ.

1.5 Outline of paper

In Section 2 we go through the details of deriving the main identity (1.28). In Section 3 we consider the interior impedance problem for the Helmholtz equation, (1.20), and show how the identity (1.28) gives rise to the new coercive variational formulation (1.25). In Section 4 we consider the exterior sound-soft scattering problem for the Helmholtz equation (i.e. the Helmholtz equation posed in the exterior of a bounded domain with Dirichlet boundary conditions) and show that there exists a coercive variational formulation of this problem if the scatterer is star-shaped with respect to a ball. Section 5 begins to investigate some of the implications that the coercivity results have for potential discretisations of the variational formulations. In Section 6 we discuss to what extent the geometric restriction of star-shapedness can be lifted from the new formulations of Sections 3 and 4. We conclude with some remarks in Section 7.

2 Morawetz- and Rellich-type identities

In Lemma 2.1 we prove the identity (1.34), and in Lemma 2.2 we prove a generalisation of the identity (1.28).

Lemma 2.1 (Rellich-type identity). *Let $u, v \in C^2(D)$ for some $D \subset \mathbb{R}^d$, $d \geq 2$, and let $\mathcal{L}v = \Delta v + k^2 v$ where $k \in \mathbb{R}$. Then*

$$(\mathbf{x} \cdot \overline{\nabla v})\mathcal{L}u + (\mathbf{x} \cdot \nabla u)\overline{\mathcal{L}v} = \nabla \cdot \left[(\mathbf{x} \cdot \overline{\nabla v})\,\nabla u + (\mathbf{x} \cdot \nabla u)\,\overline{\nabla v} + \mathbf{x}(k^2 u\overline{v} - \nabla u \cdot \overline{\nabla v})\right]$$

$$+ (d-2)\nabla u \cdot \overline{\nabla v} - dk^2 u\overline{v}. \tag{2.1}$$

Proof. The identity (2.1) is the sum of (1.32) and k^2 times

$$(\mathbf{x}\cdot\overline{\nabla v})u + (\mathbf{x}\cdot\nabla u)\overline{v} = \nabla\cdot[\mathbf{x}\,u\overline{v}] - du\overline{v}. \tag{2.2}$$

The identity (1.32) arises from adding (1.30) to the same expression with u and $\overline{v}$ swapped, and using (1.31). To prove (2.2), (1.30), and (1.31) expand the divergences on the right-hand sides using either the summation convention or the elementary vector calculus identities

$$\nabla\cdot[\mathbf{A}a] = a\nabla\cdot\mathbf{A} + \mathbf{A}\cdot\nabla a, \qquad \nabla(\mathbf{x}\cdot\nabla b) = \nabla b + (\mathbf{x}\cdot\nabla)\nabla b,$$

$$(\mathbf{x}\cdot\nabla)(\mathbf{A}\cdot\mathbf{B}) = \mathbf{B}\cdot\big((\mathbf{x}\cdot\nabla)\mathbf{A}\big) + \mathbf{A}\cdot\big((\mathbf{x}\cdot\nabla)\mathbf{B}\big), \qquad \nabla\cdot\mathbf{x} = d,$$

which hold for any scalar C^1-function a, scalar C^2-function b, and C^1-vector fields $\mathbf{A}$, $\mathbf{B}$. □

Rellich-type identities are most often used (and indeed derived) with $v = u$, i.e. one begins by multiplying $\mathcal{L}u$ by $\mathbf{x}\cdot\overline{\nabla u}$. In this case the "trick" (1.31) for getting rid of the undesirable term $\nabla u\cdot((\mathbf{x}\cdot\nabla)\overline{\nabla u})$ becomes

$$2\Re\big\{\nabla u\cdot\big((\mathbf{x}\cdot\nabla)\overline{\nabla u}\big)\big\} = \nabla\cdot\big[\mathbf{x}\,|\nabla u|^2\big] - d|\nabla u|^2.$$

To use this we need to take the real part of the expression involving $\mathcal{L}u$, and this is the reason that Rellich identities for complex-valued functions always involve $2\Re\{(\mathbf{x}\cdot\overline{\nabla u})\,\mathcal{L}u\}$ (or this expression with a different vector field instead of $\mathbf{x}$).

In §1.4 we sketched how to obtain the identity (1.28), which involved multiplying $\mathcal{L}u$ with $\overline{\mathcal{M}v}$ and $\overline{\mathcal{L}v}$ with $\mathcal{M}u$. Here we derive a slightly more general identity that allows the multiplier involving u to be different from the multiplier involving v. This added generality gives us a bit more flexibility in obtaining a coercive formulation; we discuss this more in §3.

Lemma 2.2 (Morawetz-type identity)**.** *Let u, v be as in Lemma 2.1 and define the operators $\mathcal{M}_j$ by*

$$\mathcal{M}_j v := \mathbf{x}\cdot\nabla v - \mathrm{i}k\beta_j v + \alpha_j v, \quad j = 1, 2, \tag{2.3}$$

where $\beta_j, \alpha_j \in \mathbb{R}$. Then

$$\begin{aligned}\overline{\mathcal{M}_1 v}\mathcal{L}u + \mathcal{M}_2 u\overline{\mathcal{L}v} = \nabla\cdot\Big[\overline{\mathcal{M}_1 v}\,\nabla u + \mathcal{M}_2 u\,\overline{\nabla v} + \mathbf{x}(k^2 u\overline{v} - \nabla u\cdot\overline{\nabla v})\Big] \\ + \big(d - 2 - \alpha_1 - \alpha_2 - \mathrm{i}k(\beta_1 - \beta_2)\big)\nabla u\cdot\overline{\nabla v} + \big(\alpha_1 + \alpha_2 - d + \mathrm{i}k(\beta_1 - \beta_2)\big)k^2 u\overline{v}.\end{aligned} \tag{2.4}$$

(When $\beta_1 = \beta_2 = \beta$ and $\alpha_1 = \alpha_2 = (d-1)/2$, equation (2.4) becomes (1.28).)

Proof. By Green's first identity

$$\overline{v}\mathcal{L}u = \nabla\cdot\big[\overline{v}\nabla u\big] - \nabla u\cdot\overline{\nabla v} + k^2 u\overline{v}, \tag{2.5}$$

$$u\overline{\mathcal{L}v} = \nabla\cdot\big[u\overline{\nabla v}\big] - \nabla u\cdot\overline{\nabla v} + k^2 u\overline{v}, \tag{2.6}$$

and then the identity (2.4) is the Rellich identity (2.1) plus $\mathrm{i}k\beta_1 + \alpha_1$ times (2.5), plus $-\mathrm{i}k\beta_2 + \alpha_2$ times (2.6). □

Remark 2.3 (Relationship to other identities)**.** *If we let $v = u$, $\beta_2 = \beta_1$, and $\alpha_2 = \alpha_1$ in the identity (2.4) then we obtain*

$$2\Re\{\mathcal{M}_1 u\overline{\mathcal{L}u}\} = \nabla\cdot\big[2\Re\{\mathcal{M}_1 u\overline{\nabla u}\} + \mathbf{x}(k^2|u|^2 - |\nabla u|^2)\big] + (d-2-2\alpha_1)|\nabla u|^2 + (2\alpha_1 - d)k^2|u|^2. \tag{2.7}$$

This identity is very similar to [61, Equation A.3] except that the second term in the multiplier in [61] is $-\mathrm{i}k|\mathbf{x}|v$, so the right-hand side of [61, Equation A.3] then contains an extra term from differentiating $|\mathbf{x}|$. The identity (2.7) can be generalised by replacing the vector field $\mathbf{x}$ by an arbitrary vector field, and replacing the constants β_1 and α_1 by arbitrary scalar functions of $\mathbf{x}$. This more general identity was essentially introduced in [60, §I.2] ([60, Lemma 3] contains a particular case of this identity with α_1 chosen to be a certain function of the vector field); the general identity with arbitrary α_1 and β_1 appears in [71, Lemma 2.1].

In Sections 3 and 4 we need the identity (2.4) integrated over a Lipschitz domain Ω when u, v are in the space V defined by (1.21).

Lemma 2.4 (Integrated form of the main identity (2.4))**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain with outward-pointing unit normal $\mathbf{n}$. If $u, v \in V$, where V is defined by (1.21), then*

$$\begin{aligned}\int_\Omega \Big(&\overline{\mathcal{M}_1 v}\mathcal{L}u + \mathcal{M}_2 u\overline{\mathcal{L}v} + \big(2 - d + \alpha_1 + \alpha_2 + \mathrm{i}k(\beta_1 - \beta_2)\big)\nabla u \cdot \overline{\nabla v} \\ &+ \big(d - \alpha_1 - \alpha_2 - \mathrm{i}k(\beta_1 - \beta_2)\big)k^2 u\overline{v}\Big)\mathrm{d}\mathbf{x} \\ = \int_{\partial\Omega} \Big(&\overline{\mathcal{M}_1 v}\,\frac{\partial u}{\partial n} + \mathcal{M}_2 u\,\overline{\frac{\partial v}{\partial n}} + (\mathbf{x}\cdot\mathbf{n})(k^2 u\overline{v} - \nabla u \cdot \overline{\nabla v})\Big)\mathrm{d}s, \end{aligned} \tag{2.8}$$

where $\mathcal{L}u$, $\mathcal{L}v$ are as above, and $\mathcal{M}_j$, $j = 1, 2$, are defined by (2.3).

Proof. This is a consequence of the divergence theorem applied to the identity (2.4). The divergence theorem

$$\int_\Omega \nabla \cdot \mathbf{F}\,\mathrm{d}\mathbf{x} = \int_{\partial\Omega} \mathbf{F}\cdot\mathbf{n}\,\mathrm{d}s$$

is valid when Ω is Lipschitz and $\mathbf{F} \in (C^1(\overline{\Omega}))^d$ [49, Theorem 3.34]. In Appendix A we prove that $\mathcal{D}(\overline{\Omega}) := \{U|_\Omega : U \in C^\infty(\mathbb{R}^d)\}$ is dense in V, and thus (2.8) holds for any $u, v \in V$. (Note that this density result is perhaps not immediately obvious due to the subtleties discussed in Remark 4.6 below.) □

3 Interior impedance problem

Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, be a bounded Lipschitz domain with $L = \mathrm{diam}(\Omega)$, i.e. $L := \max_{\mathbf{x},\mathbf{y}\in\partial\Omega} |\mathbf{x} - \mathbf{y}|$. For a Lipschitz domain the outward-pointing unit normal vector $\mathbf{n}(\mathbf{x})$ is defined for almost every $\mathbf{x} \in \partial\Omega$ by Rademacher's theorem (see, e.g., [36, §5.8.3, Theorem 6]. In what follows, whenever we have an expression on $\partial\Omega$ we just write u instead of introducing any notation for the trace of u; we do this to prevent some of the expressions (e.g. (3.2)) from becoming over-complicated.

We consider a slightly more general impedance boundary condition than in (1.20b), in that we replace $\mathrm{i}k$ by $\mathrm{i}k\vartheta$, where ϑ is some prescribed function.

Definition 3.1 (Interior Impedance Problem)**.** *Given $f \in L^2(\Omega)$, $g \in L^2(\partial\Omega)$, and $\vartheta \in L^\infty(\partial\Omega)$ with ϑ real, independent of k and u, and such that*

$$0 < \vartheta_* := \operatorname*{ess\,inf}_{\mathbf{x}\in\partial\Omega} \vartheta(\mathbf{x}) \le \operatorname*{ess\,sup}_{\mathbf{x}\in\partial\Omega} \vartheta(\mathbf{x}) =: \vartheta^* < \infty,$$

find $u \in H^1(\Omega)$ such that

$$\Delta u + k^2 u = -f \quad \text{in } \Omega, \tag{3.1a}$$

$$\frac{\partial u}{\partial n} - \mathrm{i}k\vartheta u = g \quad \text{on } \partial\Omega. \tag{3.1b}$$

The PDE in (3.1a) is understood as holding in a distributional sense. Recall that, for $u \in H^1(\Omega)$ with $\Delta u \in L^2(\Omega)$, $\partial u/\partial n$ is understood as an element of $H^{-1/2}(\partial\Omega)$ via Green's first identity (see, e.g., [49, Lemma 4.3], [17, Equation A.28]). The boundary condition (3.1b) is then understood as saying that this element of $H^{-1/2}(\partial\Omega)$ is actually in $L^2(\partial\Omega)$ and the equation $\partial u/\partial n = \mathrm{i}ku + g$ holds as an equation in $L^2(\partial\Omega)$. It is then straightforward to show that the statement that u satisfies (3.1) is equivalent to the statement that u satisfies the variational problem (1.3) with $V = H^1(\Omega)$, $a(\cdot,\cdot)$ replaced by $a_I(\cdot,\cdot)$, and both $a_I(\cdot,\cdot)$ and $F(\cdot)$ defined by (1.19).

The solution to the interior impedance problem is unique. This is usually proved by applying Green's first identity (1.6) with $v = u$, imposing the PDE and boundary condition, and taking the imaginary part. The coercivity result below, however, gives an alternative proof of uniqueness in the space V defined by (1.21), under the assumption that Ω is star-shaped with respect to a ball.

Note that if ϑ is chosen to be uniformly negative then all the results below follow in the same way (but we do not consider the cases where ϑ approaches zero, changes sign, or becomes unbounded).

We now define a sesquilinear form that can be used to solve the interior impedance problem. Even when $\vartheta \equiv 1$, this sesquilinear form is more general than the one introduced in §1.3, (1.23); this is because it is based on the identity (2.4), whereas (1.23) is based on the identity (1.28). The reason we introduce this more general sesquilinear form is that we can then obtain a coercive formulation that has two free parameters in it (these will be α_2 and β_2); we anticipate that this additional freedom may prove useful (for example when this formulation is implemented numerically).

Define the sesquilinear form $b : V \times V \to \mathbb{C}$ by

$$\begin{aligned} b(u,v) := \int_\Omega \Big(& \big(2-d+\alpha_1+\alpha_2+\mathrm{i}k(\beta_1-\beta_2)\big)\nabla u\cdot\overline{\nabla v} + \big(d-\alpha_1-\alpha_2-\mathrm{i}k(\beta_1-\beta_2)\big)k^2 u\overline{v} \\ & + \left(\mathcal{M}_2 u + \frac{A}{k^2}\mathcal{L}u\right)\overline{\mathcal{L}v}\Big)\,\mathrm{d}\mathbf{x} \\ - \int_{\partial\Omega} \Big(& \mathrm{i}k\vartheta u\,\overline{\mathcal{M}_1 v} + (\mathbf{x}\cdot\nabla_{\partial\Omega}u - \mathrm{i}k\beta_2 u + \alpha_2 u)\,\overline{\frac{\partial v}{\partial n}} + (\mathbf{x}\cdot\mathbf{n})\left(k^2 u\overline{v} - \nabla_{\partial\Omega}u\cdot\overline{\nabla_{\partial\Omega}v}\right)\Big)\mathrm{d}s, \end{aligned} \tag{3.2}$$

and the functional $G : V \to \mathbb{C}$ by

$$G(v) := \int_\Omega \left(\overline{\mathcal{M}_1 v} - \frac{A}{k^2}\overline{\mathcal{L}v}\right) f\,\mathrm{d}\mathbf{x} + \int_{\partial\Omega} \overline{\mathcal{M}_1 v}\, g\,\mathrm{d}s, \tag{3.3}$$

where $\mathcal{M}_j$, $j = 1, 2$, are defined by (2.3), and $\alpha_1, \alpha_2, \beta_1, \beta_2$, and A are all arbitrary real parameters. If we take $\alpha_1 = \alpha_2 = (d-1)/2$, $\beta_1 = \beta_2 = \beta$, $A = 1/3$, and $\vartheta \equiv 1$ then (3.2) becomes the sesquilinear form defined in §1.3, (1.23).

Proposition 3.2 ($b(\cdot,\cdot)$ can be used to solve the interior impedance problem)**.** *If u solves the interior impedance problem of Definition 3.1, then $u \in V$, where V is defined by (1.21), and*

$$b(u,v) = G(v) \quad \text{for all } v \in V, \tag{3.4}$$

where $b(\cdot,\cdot)$ is given by (3.2) and $G(\cdot)$ by (3.3).

Proof. For the solution of the interior impedance problem, u, to be in V we need to show that $\Delta u \in L^2(\Omega)$, $\partial u/\partial n \in L^2(\partial\Omega)$, and $\nabla_{\partial\Omega} u \in (L^2(\partial\Omega))^d$. From the PDE and boundary conditions (3.1) we have that $\Delta u = -k^2 u - f \in L^2(\Omega)$ and $\partial u/\partial n = \mathrm{i}ku + g \in L^2(\partial\Omega)$, and so, by a regularity result of Nečas [63, §5.2.1], [49, Theorem 4.24 (ii)], $\nabla_{\partial\Omega} u \in (L^2(\partial\Omega))^d$.

Since u and v are both in V, the integrated identity (2.8) holds. From the definition of $\mathcal{M}_2 u$,

$$\mathcal{M}_2 u\,\overline{\frac{\partial v}{\partial n}} = (\mathbf{x}\cdot\mathbf{n})\frac{\partial u}{\partial n}\overline{\frac{\partial v}{\partial n}} + (\mathbf{x}\cdot\nabla_{\partial\Omega}u - \mathrm{i}k\beta_2 + \alpha_2 u)\overline{\frac{\partial v}{\partial n}}.$$

Substituting this last expression into (2.8), then using the PDE (3.1a) and boundary conditions (3.1b), and finally rearranging so that all the terms involving f and g are on one side and all the terms involving u are on the other we obtain

$$\begin{aligned} & \int_\Omega \Big(\big(d-\alpha_1-\alpha_2-\mathrm{i}k(\beta_1-\beta_2)\big)k^2 u\overline{v} + \big(2-d+\alpha_1+\alpha_2+\mathrm{i}k(\beta_1-\beta_2)\big)\nabla u\cdot\overline{\nabla v} + \mathcal{M}_2 u\overline{\mathcal{L}v}\Big)\mathrm{d}\mathbf{x} \\ & - \int_{\partial\Omega} \Big(\overline{\mathcal{M}_1 v}\,\mathrm{i}k\vartheta u + (\mathbf{x}\cdot\nabla_{\partial\Omega}u - \mathrm{i}k\beta_2 u + \alpha_2 u)\overline{\frac{\partial v}{\partial n}} + (\mathbf{x}\cdot\mathbf{n})\left(k^2 u\overline{v} - \nabla_{\partial\Omega}u\cdot\overline{\nabla_{\partial\Omega}v}\right)\Big)\mathrm{d}s \\ & = \int_\Omega \overline{\mathcal{M}_1 v}\, f\,\mathrm{d}\mathbf{x} + \int_{\partial\Omega} \overline{\mathcal{M}_1 v}\, g\,\mathrm{d}s. \end{aligned} \tag{3.5}$$

This is almost $b(u,v) = G(v)$ with $b(\cdot,\cdot)$ and $G(\cdot)$ defined by (3.2) and (3.3) respectively, but not quite. We need to add

$$\int_\Omega \frac{A}{k^2}\mathcal{L}u\overline{\mathcal{L}v}\,\mathrm{d}\mathbf{x}$$

to the left-hand side of (3.5) and

$$-\int_\Omega \frac{A}{k^2} f\,\overline{\mathcal{L}v}\,\mathrm{d}\mathbf{x}$$

to the right with $A \in \mathbb{R}$ arbitrary (these terms are equal to each other by the PDE (3.1a)). We add these terms because it turns out that $b(\cdot,\cdot)$ must contain a $\Delta u\overline{\Delta v}$ term to be coercive in the norm of V, (1.22), and this term is not present in (3.5). (We could have just added a multiple of $\mathcal{L}u\overline{\Delta v}$, but we add $\mathcal{L}u\overline{\mathcal{L}v}$ to make $b(\cdot,\cdot)$ as symmetric as possible.) □

As discussed above, the interior impedance problem has exactly one solution, which is in the space V. Theorem 3.4 below shows that the variational problem (3.4) has exactly one solution in V if Ω is star-shaped with respect to a ball; hence in this case the converse to Proposition 3.2 holds, namely, that if u is a solution to (3.4) then u is a solution to the interior impedance problem.

Lemma 3.3 (Continuity of $b(\cdot,\cdot)$)**.** *For all $u, v \in V$ and for all $k > 0$, the continuity bound*

$$|b(u,v)| \le C_c \,\|u\|_V \,\|v\|_V$$

holds with

$$C_c := \sqrt{3}\max\begin{cases} |d-\alpha_1-\alpha_2| + k|\beta_1-\beta_2| + k|\beta_2| + |\alpha_2| + 2|A| + kL; \\ |2-d+\alpha_1+\alpha_2| + k|\beta_1-\beta_2|; \\ 1 + \frac{\vartheta^*}{kL}(k|\beta_1| + |\alpha_1|); \\ 1 + \vartheta^* + \frac{1}{kL}(k|\beta_2| + |\alpha_2|). \end{cases}$$

If $\alpha_1 = \alpha_2 = (d-1)/2$, $\beta_1 = \beta_2 > 0$ and $A > 0$, then the above expression simplifies to

$$C_c = \sqrt{3}\max\left\{\frac{d+1}{2} + k\beta_1 + 2A + kL;\ 1 + \vartheta^* + \frac{1+\vartheta^*}{kL}\Big(k\beta_1 + \frac{d-1}{2}\Big)\right\}.$$

In particular, if A and β_1 are independent of k (as we choose them to be below), then C_c grows linearly in k.

Proof. Define the vector $\mathbf{m}(u) \in \mathbb{R}^6$ by

$$\mathbf{m}(u) := \left\{k\,\|v\|_\Omega\,;\|\nabla v\|_\Omega\,;k^{-1}\,\|\Delta v\|_\Omega\,;L^{1/2}k\,\|v\|_{\partial\Omega}\,;L^{1/2}\,\|\nabla_{\partial\Omega} v\|_{\partial\Omega}\,;L^{1/2}\left\|\frac{\partial v}{\partial n}\right\|_{\partial\Omega}\right\},$$

so that $\|\mathbf{m}(u)\|_2 = \|u\|_V$, where $\|\cdot\|_2$ denotes the (Euclidean) 2-norm on $\mathbb{R}^6$. By Cauchy–Schwarz,

$$|b(u,v)| \le |\mathbf{m}(u)^T\mathbf{M}\mathbf{m}(v)| \le \|\mathbf{m}(u)\|_2\,\|\mathbf{M}\|_2\,\|\mathbf{m}(v)\|_2 = \|u\|_V\,\|\mathbf{M}\|_2\,\|v\|_V\,,$$

where $\mathbf{M}$ is a 6×6 block-diagonal matrix consisting of two 3×3 blocks $\mathbf{M}_1$ and $\mathbf{M}_2$. Thus $C_c \le \|\mathbf{M}\|_2 = \max\{\|\mathbf{M}_1\|_2\,,\|\mathbf{M}_2\|_2\} \le \sqrt{3}\max\{\|\mathbf{M}_1\|_1\,,\|\mathbf{M}_2\|_1\}$, and the assertion follows from the definition of $b(\cdot,\cdot)$, (3.2), which defines the coefficients of $\mathbf{M}$. □

We now prove that $b(\cdot,\cdot)$ defined by (3.2) is coercive (this theorem therefore includes Theorem 1.1 as a special case).

Theorem 3.4 (Coercivity of $b(\cdot,\cdot)$)**.** *Let $b(\cdot,\cdot)$ be defined by (3.2) and V defined by (1.21). Suppose that Ω is a Lipschitz domain with diameter L that is star-shaped with respect to a ball, i.e. there exists a $\gamma > 0$ such that*

$$\mathbf{x}\cdot\mathbf{n}(\mathbf{x}) \ge \gamma\, L, \tag{3.6}$$

for all $\mathbf{x} \in \partial\Omega$ for which $\mathbf{n}(\mathbf{x})$ is defined (see Remark 3.5 for the geometric significance of γ). If

$$\alpha_1 = \frac{d-1}{2}, \qquad \beta_1 \ge \frac{L}{2\vartheta_*}\left[1 + 4\frac{(\vartheta^*)^2}{\gamma} + \frac{\gamma}{2}\right], \quad \text{and} \quad A = \frac{1}{3}, \tag{3.7}$$

then, for any $k > 0$, and for any $\alpha_2, \beta_2 \in \mathbb{R}$,

$$\Re b(v,v) \ge \frac{\gamma}{4}\|v\|_V^2 \qquad \text{for all } v \in V.$$

Proof. The definition of $b(\cdot,\cdot)$, (3.2), implies that, for all $v \in V$,

$$2\Re b(v,v) = \int_\Omega \left(2(2-d+\alpha_1+\alpha_2)|\nabla v|^2 + 2(d-\alpha_1-\alpha_2)k^2|v|^2 + \frac{2A}{k^2}|\mathcal{L}v|^2 + 2\Re\{\mathcal{M}_2 v\overline{\mathcal{L}v}\}\right)\mathrm{d}\mathbf{x}$$
$$- 2\int_{\partial\Omega}(\mathbf{x}\cdot\mathbf{n})(k^2|v|^2 - |\nabla_{\partial\Omega}v|^2)\mathrm{d}s - 2\Re\int_{\partial\Omega}\left(\mathrm{i}k\vartheta v\,\overline{\mathcal{M}_1 v} + (\mathbf{x}\cdot\nabla_{\partial\Omega}v - \mathrm{i}k\beta_2 v + \alpha_2 v)\,\overline{\frac{\partial v}{\partial n}}\right)\mathrm{d}s. \tag{3.8}$$

Recall that the sesquilinear form (3.2) essentially came from the integrated identity (2.8). We now use this identity with $u = v$ to obtain an expression for the integral over Ω of $2\Re\{\mathcal{M}_2 v\overline{\mathcal{L}v}\}$, which appears on the right-hand side of (3.8). Indeed, (2.8) with $u = v$ implies that

$$\int_\Omega 2\Re\{\mathcal{M}_2 v\overline{\mathcal{L}v}\}\,\mathrm{d}\mathbf{x} = \int_\Omega \left((d-2-2\alpha_2)|\nabla v|^2 + (2\alpha_2 - d)k^2|v|^2\right)\mathrm{d}\mathbf{x}$$
$$+ \int_{\partial\Omega}\left((\mathbf{x}\cdot\mathbf{n})\left(\left|\frac{\partial v}{\partial n}\right|^2 + k^2|v|^2 - |\nabla_{\partial\Omega}v|^2\right) + 2\Re\left\{(\mathbf{x}\cdot\nabla_{\partial\Omega}v - \mathrm{i}k\beta_2 v + \alpha_2 v)\overline{\frac{\partial v}{\partial n}}\right\}\right)\mathrm{d}s.$$

Substituting this into (3.8) and using the definition of $\mathcal{M}_1$, (2.3), we find that

$$\begin{aligned} 2\Re b(v,v) = &\int_\Omega \left((2-d+2\alpha_1)|\nabla v|^2 + (d-2\alpha_1)k^2|v|^2 + \frac{2A}{k^2}|\mathcal{L}v|^2\right)\mathrm{d}\mathbf{x} \\ &+ \int_{\partial\Omega}(\mathbf{x}\cdot\mathbf{n})\left(\left|\frac{\partial v}{\partial n}\right|^2 + |\nabla_{\partial\Omega}v|^2 - k^2|v|^2\right)\mathrm{d}s \\ &- 2\Re\int_{\partial\Omega}\mathrm{i}k\vartheta v\left((\mathbf{x}\cdot\mathbf{n})\overline{\frac{\partial v}{\partial n}} + \mathbf{x}\cdot\overline{\nabla_{\partial\Omega}v} + \mathrm{i}k\beta_1\overline{v} + \alpha_1\overline{v}\right)\mathrm{d}s. \end{aligned} \tag{3.9}$$

We first concentrate on the integral over Ω. Using both the triangle inequality and the inequality

$$2ab \le \frac{a^2}{\varepsilon} + \varepsilon b^2, \qquad \text{for } a,b,\varepsilon > 0, \tag{3.10}$$

with $\varepsilon = 1$ we have that $|a|^2 \le 2|b|^2 + 2|a+b|^2$, and thus $|a+b|^2 \ge \frac{1}{2}|a|^2 - |b|^2$. Using this with $a = \sqrt{2A}\Delta v/k$ and $b = \sqrt{2A}kv$, the integral over Ω in (3.9) is greater than or equal to

$$(2-d+2\alpha_1)\left\|\nabla v\right\|^2_{L^2(\Omega)} + (d-2\alpha_1-2A)k^2\left\|v\right\|^2_{L^2(\Omega)} + \frac{A}{k^2}\left\|\Delta v\right\|^2_{L^2(\Omega)}.$$

If we choose $2\alpha_1 = d-1$ and $A = 1/3$ then the coefficients of $\|\nabla v\|^2_{L^2(\Omega)}$, $k^2\|v\|^2_{L^2(\Omega)}$, and $\|\Delta v\|^2_{L^2(\Omega)}/k^2$ become 1, 1/3, and 1/3 respectively (other choices are possible, but the point is that we make each coefficient greater than zero.)

We now concentrate on the two integrals in (3.9) that are over $\partial\Omega$. Using the inequalities $\gamma L \le \mathbf{x}\cdot\mathbf{n} \le L$ (since $|\mathbf{x}| \le L$) the first integral is greater than or equal to

$$L\left(\gamma\left\|\frac{\partial v}{\partial n}\right\|^2_{L^2(\partial\Omega)} + \gamma\left\|\nabla_{\partial\Omega}v\right\|^2_{L^2(\partial\Omega)} - k^2\left\|v\right\|^2_{L^2(\partial\Omega)}\right).$$

Because α_1 and ϑ are real, the second integral over $\partial\Omega$ in (3.9) equals

$$-2\Re\int_{\partial\Omega}(\mathbf{x}\cdot\mathbf{n})\overline{\frac{\partial v}{\partial n}}\mathrm{i}k\vartheta v\,\mathrm{d}s - 2\Re\int_{\partial\Omega}\mathbf{x}\cdot\overline{\nabla_{\partial\Omega}v}\,\mathrm{i}k\vartheta v\,\mathrm{d}s + 2k^2\beta_1\int_{\partial\Omega}\vartheta|v|^2\,\mathrm{d}s.$$

The final term in this last expression is $\ge 2\beta_1\vartheta_* k^2\|v\|^2_{L^2(\partial\Omega)}$. To deal with the first two terms we use the inequalities (3.10) and $|\mathbf{x}\cdot\nabla_{\partial\Omega}v| \le L|\nabla_{\partial\Omega}v|$ to obtain

$$2\Re\int_{\partial\Omega}(\mathbf{x}\cdot\mathbf{n})\overline{\frac{\partial v}{\partial n}}\,\mathrm{i}k\vartheta v\,\mathrm{d}s \le L\vartheta^*\left(\frac{1}{\varepsilon_1}\left\|\frac{\partial v}{\partial n}\right\|^2_{L^2(\partial\Omega)} + \varepsilon_1 k^2\left\|v\right\|^2_{L^2(\partial\Omega)}\right)$$

and

$$2\Re\int_{\partial\Omega}\mathbf{x}\cdot\overline{\nabla_{\partial\Omega}v}\,\mathrm{i}k\vartheta v\,\mathrm{d}s\leq L\vartheta^*\left(\frac{1}{\varepsilon_2}\left\|\nabla_{\partial\Omega}v\right\|^2_{L^2(\partial\Omega)}+\varepsilon_2k^2\left\|v\right\|^2_{L^2(\partial\Omega)}\right)$$

for any $\varepsilon_1,\varepsilon_2>0$.

Putting everything together results in the inequality

$$\begin{aligned}2\Re b(v,v)\geq{}&\left\|\nabla v\right\|^2_{L^2(\Omega)}+\frac{1}{3}k^2\left\|v\right\|^2_{L^2(\Omega)}+\frac{1}{3k^2}\left\|\mathcal{L}v\right\|^2_{L^2(\Omega)}+L\left(\gamma-\frac{\vartheta^*}{\varepsilon_1}\right)\left\|\frac{\partial v}{\partial n}\right\|^2_{L^2(\partial\Omega)}\\&+L\left(\gamma-\frac{\vartheta^*}{\varepsilon_2}\right)\left\|\nabla_{\partial\Omega}v\right\|^2_{L^2(\partial\Omega)}+L\left(\frac{2\beta_1\vartheta_*}{L}-1-\vartheta^*\varepsilon_1-\vartheta^*\varepsilon_2\right)k^2\left\|v\right\|^2_{L^2(\partial\Omega)}.\end{aligned}\tag{3.11}$$

If we choose $\varepsilon_1=\varepsilon_2=2\vartheta^*/\gamma$ then the norms on $\partial\Omega$ in (3.11) become

$$\frac{\gamma L}{2}\left(\left\|\frac{\partial v}{\partial n}\right\|^2_{L^2(\partial\Omega)}+\left\|\nabla_{\partial\Omega}v\right\|^2_{L^2(\partial\Omega)}\right)+L\left(\frac{2\beta_1\vartheta_*}{L}-1-4\frac{(\vartheta^*)^2}{\gamma}\right)k^2\left\|v\right\|^2_{L^2(\partial\Omega)}.$$

Then, if we choose β_1 as in (3.7) these terms are greater than or equal to

$$\frac{\gamma L}{2}\left(\left\|\frac{\partial v}{\partial n}\right\|^2_{L^2(\partial\Omega)}+\left\|\nabla_{\partial\Omega}v\right\|^2_{L^2(\partial\Omega)}+k^2\left\|v\right\|^2_{L^2(\partial\Omega)}\right).$$

Hence, the inequality (3.11) becomes

$$\begin{aligned}2\Re b(v,v)\geq{}&\left\|\nabla v\right\|^2_{L^2(\Omega)}+\frac{1}{3}k^2\left\|v\right\|^2_{L^2(\Omega)}+\frac{1}{3k^2}\left\|\mathcal{L}v\right\|^2_{L^2(\Omega)}\\&+\frac{\gamma L}{2}\left(\left\|\frac{\partial v}{\partial n}\right\|^2_{L^2(\partial\Omega)}+\left\|\nabla_{\partial\Omega}v\right\|^2_{L^2(\partial\Omega)}+k^2\left\|v\right\|^2_{L^2(\partial\Omega)}\right),\end{aligned}$$

and so $b(\cdot,\cdot)$ is coercive with coercivity constant

$$\frac{1}{2}\min\left\{\frac{1}{3};\,\frac{\gamma}{2}\right\}=\frac{\gamma}{4}$$

(since $\gamma\leq1/2$ by Remark 3.5 below). □

Note that α_2 and β_2 do not play any role in the proof that $b(\cdot,\cdot)$ is coercive.

Although the bound obtained in Theorem 3.4 may appear stronger than that in the definition of coercivity, (1.8), this is not the case. Indeed, for a sesquilinear form $a(\cdot,\cdot)$, if (1.8) holds, i.e. $|a(v,v)|\geq\alpha\|v\|^2_V$ for all $v\in\mathcal{V}$, then there exists a complex number σ with $|\sigma|=1$ such that $\Re\{\sigma a(v,v)\}\geq\alpha\|v\|^2_V$, for all $v\in\mathcal{V}$; this follows from the convexity of the numerical range of the operator associated with $a(\cdot,\cdot)$ and the relationship of the numerical range to the coercivity constant α (see, e.g., [8, Propositions 3.2 and 3.3]).

Remark 3.5 (Geometrical significance of γ in the star-shapedness condition (3.6))**.** *The standard definition of star-shapedness is that Ω is star-shaped with respect to a point $\mathbf{x}_0$ if, whenever $\mathbf{x}\in\Omega$, the segment $[\mathbf{x}_0,\mathbf{x}]\subset\Omega$. Furthermore, Ω is star-shaped with respect to the ball $B_a(\mathbf{x}_0)$:= $\{\mathbf{x}\in\mathbb{R}^d,\|\mathbf{x}-\mathbf{x}_0\|_2<a\}$ if it is star-shaped with respect to every point in $B_a(\mathbf{x}_0)$.*

If Ω is Lipschitz (and so has a normal vector at almost every point on the boundary) then Ω 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 a proof see [56, Lemma 5.4.1] or [45, Lemma 3.1]. Note that in Theorems 3.4 and 4.5 we assume (without loss of generality) that the balls are centred at the origin, i.e. $\mathbf{x}_0=\mathbf{0}$.

Remark 3.6 (Bounding the solution of the BVP)**.** *Combining Theorem 3.4, the estimate*

$$\left\|G\right\|_{V'}\leq\sqrt{3}\max\left\{1;\,\frac{1}{kL}(k|\beta_1|+|\alpha_1|+|A|)\right\}\left(L^2\left\|f\right\|^2_{L^2(\Omega)}+L\left\|g\right\|^2_{L^2(\partial\Omega)}\right)^{1/2},$$

and the consequence of coercivity (1.9), we obtain the bound

$$\|u\|_V \leq \frac{4\sqrt{3}}{\gamma}\left(1+\frac{\beta_1}{L}+\frac{d}{2kL}\right)\left(L^2\,\|f\|^2_{L^2(\Omega)}+L\,\|g\|^2_{L^2(\partial\Omega)}\right)^{1/2}, \tag{3.12}$$

for all $k>0$ (under the condition (3.7) on α_1,β_1 and A). Bounds identical to (3.12) in their k-dependence were proved for $d=2$ in [50, Proposition 8.1.4] and for $d=3$ in [24, Theorem 1], essentially using the identity (2.7) with $\beta_1=0$; see [17, §5.3.2] for more explanation. It is interesting to note that taking β_1 to be non-zero in the multiplier $\mathcal{M}_1 u$ does not help in proving the bounds on the solution, but is crucial for the proof that $b(\cdot,\cdot)$ is coercive (since we need to take β_1 large enough to get coercivity).

Remark 3.7 (Relationship of the space V to $H^{3/2}(\Omega)$)**.** *We now outline how to prove the facts mentioned in §1.3 that* (i) *$V\subset H^{3/2}(\Omega)$, and* (ii) *if $v\in C^2(\Omega)$ is such that $\Delta v+k^2v=0$, then $v\in H^{3/2}(\Omega)$ implies that $v\in V$. The statement* (i) *follows from expressing $v\in V$ via Green's integral representation involving the fundamental solution of the Laplacian [49, Theorem 7.5] and then using mapping properties of the Newtonian potential [49, Theorem 6.1] and the single- and double-layer potentials [17, Theorem 2.15]. The statement* (ii) *follows from [17, Lemma A.10].*

Remark 3.8 (Why the norm in V is scaled with k and L)**.** *If v is a plane wave solution to the Helmholtz equation, i.e. $v(\mathbf{x})=\exp(\mathrm{i}k\mathbf{x}\cdot\widehat{\mathbf{a}})$ for some $\widehat{\mathbf{a}}\in\mathbb{R}^d$ with $\|\widehat{\mathbf{a}}\|_2=1$, then each of the terms in the definition (1.22) of $\|v\|^2_V$ is proportional to k^2L^d. Similarly, if $f=-(\Delta v+k^2v)$ and $g=\partial v/\partial n-\mathrm{i}k\vartheta v$, then the factor involving f and g in (3.12) is also proportional to k^2L^d.*

Remark 3.9 (A first-order system formulation)**.** *The interior impedance problem of Definition 3.1 can be rewritten as the first-order system*

$$\nabla\cdot\boldsymbol{\sigma}-\mathrm{i}ku=-(\mathrm{i}k)^{-1}f \qquad \text{in } \Omega, \tag{3.13a}$$

$$\nabla u-\mathrm{i}k\boldsymbol{\sigma}=\mathbf{0} \qquad \text{in } \Omega, \tag{3.13b}$$

$$\boldsymbol{\sigma}\cdot\mathbf{n}-\vartheta u=(\mathrm{i}k)^{-1}g \qquad \text{on } \partial\Omega. \tag{3.13c}$$

Using a Morawetz-type identity for the system (3.13a)–(3.13b), a new variational formulation of this BVP can be obtained (if the domain Ω satisfies the same assumptions as in Theorem 3.4), where continuity and coercivity hold in the space

$$\left\{(u,\boldsymbol{\sigma})\in H^1(\Omega)\times\left(L^2(\Omega)\right)^d:\ \nabla\cdot\boldsymbol{\sigma}\in L^2(\Omega),\ \boldsymbol{\sigma}\in\left(L^2(\partial\Omega)\right)^d,\ \mathbf{D}\boldsymbol{\sigma}\text{ is symmetric}\right\},$$

where $\mathbf{D}\boldsymbol{\sigma}$ is the (distributional) Jacobian matrix of $\boldsymbol{\sigma}$. In three dimensions the symmetry constraint on $\mathbf{D}\boldsymbol{\sigma}$ corresponds to $\nabla\times\boldsymbol{\sigma}$ equalling zero; this constraint makes conformal discretisations of this formulation difficult.

Remark 3.10 (The analogous problem for the time-harmonic Maxwell equations)**.** *The analogue of the interior impedance problem of Definition 3.1 for electromagnetism is the following BVP for the time-harmonic Maxwell equations (with $d=3$):*

$$\nabla\times(\nabla\times\mathbf{E})-k^2\mathbf{E}=\mathbf{J} \qquad \text{in } \Omega, \tag{3.14a}$$

$$(\nabla\times\mathbf{E})\times\mathbf{n}-\mathrm{i}k\vartheta(\mathbf{n}\times\mathbf{E})\times\mathbf{n}=\mathbf{g} \qquad \text{on } \partial\Omega. \tag{3.14b}$$

If the domain Ω satisfies the same assumptions as in Theorem 3.4, it is possible to obtain a new variational formulation of this BVP that is continuous and coercive in the space

$$\left\{\mathbf{E}\in\left(L^2(\Omega)\right)^3:\ \nabla\times\mathbf{E}\text{ and }\nabla\times(\nabla\times\mathbf{E})\in\left(L^2(\Omega)\right)^3,\ \nabla\cdot\mathbf{E}=0,\ \mathbf{E}\text{ and }\nabla\times\mathbf{E}\in\left(L^2(\partial\Omega)\right)^3\right\}.$$

The identities used to derive the formulation arise from vector Morawetz-type multipliers that generalise the vector Rellich-type multipliers found in [56, §5.3] (see also [45, §3]). A similar result can be proven for an equivalent first-order boundary value problem; in both cases the divergence-free constraint in the space makes conformal discretisations difficult.

4 Sound-soft scattering problem

Let Ω_D be a bounded Lipschitz open set (with the subscript D because we are going to consider Dirichlet boundary conditions) such that the open complement $\Omega_+ := \mathbb{R}^d \setminus \overline{\Omega_D}$ is connected. Let $H^1_{\text{loc}}(\Omega_+)$ denote the set of functions, v, such that v is locally integrable on Ω_+ and $\psi v \in H^1(\Omega_+)$ for every compactly supported $\psi \in C^\infty(\overline{\Omega_+}) := \{\psi|_{\Omega_+} : \psi \in C^\infty(\mathbb{R}^d)\}$.

Definition 4.1 (Sound-soft scattering problem)**.** *Given an incident plane wave $u^I(\mathbf{x}) = \exp(\mathrm{i}k\mathbf{x}\cdot\widehat{\mathbf{a}})$ for some $\widehat{\mathbf{a}} \in \mathbb{R}^d$ with $\|\widehat{\mathbf{a}}\|_2 = 1$, find $u^S \in C^2(\Omega_+) \cap H^1_{\text{loc}}(\Omega_+)$ such that the total field $u^T := u^I + u^S$ satisfies*

$$\begin{aligned} \Delta u^T + k^2 u^T &= 0 \quad \text{in } \Omega_+, \\ u^T &= 0 \quad \text{on } \partial\Omega_+, \end{aligned}$$

and u^S satisfies the Sommerfeld radiation condition

$$\frac{\partial u^S}{\partial r}(\mathbf{x}) - \mathrm{i}k u^S(\mathbf{x}) = o\left(\frac{1}{r^{(d-1)/2}}\right),$$

as $r := |\mathbf{x}| \to \infty$, uniformly in $\widehat{\mathbf{x}} := \mathbf{x}/r$.

(We restrict our attention to the case where the incident wave is a plane wave, but the analysis below easily extends to other incident fields, for example those satisfying [17, Definition 2.11].)

Since Ω_+ is unbounded, standard finite element methods (FEMs) cannot be applied to solve the sound-soft scattering problem. One way around this is to truncate Ω_+, i.e. introduce an artificial boundary Γ_R, and impose a boundary condition on Γ_R that approximates the radiation condition. The design of appropriate boundary conditions has been, and still is, the subject of much research (see [46, Chapter 3] for an introduction), but the simplest option is just to impose an impedance boundary condition on Γ_R. We thus consider the following BVP:

Definition 4.2 (Truncated sound-soft scattering problem)**.** *Given Ω_R and Ω_D, bounded Lipschitz domains such that $\Omega_D \subset \Omega_R \subset \mathbb{R}^d$ with $d(\Omega_D, \partial\Omega_R) > 0$, let $\Gamma_R := \partial\Omega_R$, $\Gamma_D := \partial\Omega_D$, and $\Omega := \Omega_R \setminus \overline{\Omega_D}$ (so $\partial\Omega = \Gamma_R \cup \Gamma_D$ and $\Gamma_R \cap \Gamma_D = \emptyset$). Given $f \in L^2(\Omega_R)$, $g_R \in L^2(\Gamma_R)$, $g_D \in H^1(\Gamma_D)$, and $\vartheta \in L^\infty(\partial\Omega_R)$ with ϑ real, independent of k and u, and such that*

$$0 < \vartheta_* := \operatorname*{ess\,inf}_{\mathbf{x}\in\partial\Omega_R} \vartheta(\mathbf{x}) \le \operatorname*{ess\,sup}_{\mathbf{x}\in\partial\Omega_R} \vartheta(\mathbf{x}) =: \vartheta^* < \infty,$$

find $u \in H^1(\Omega)$ such that

$$\Delta u + k^2 u = -f \quad \text{in } \Omega, \tag{4.1a}$$

$$\frac{\partial u}{\partial n} - \mathrm{i}\vartheta k u = g_R \quad \text{on } \Gamma_R, \tag{4.1b}$$

$$u = g_D \quad \text{on } \Gamma_D. \tag{4.1c}$$

If we set $f = 0$, $\vartheta = 1$, $g_R = 0$, and $g_D = -u^I|_{\Gamma_D}$ then the solution to the BVP in Definition 4.2 is an approximation to u^S in Definition 4.1. The simplest choice for Ω_R is just $B_R(\mathbf{0})$ where R is taken large enough so that the ball includes Ω_D, and Figure 1 shows Ω_D and Ω_R in this case.

With Ω defined as in Definition 4.2, define the Hilbert space V by (1.21) with associated norm

$$\begin{aligned} \|v\|_V^2 &:= \|\nabla v\|^2_{L^2(\Omega)} + k^2\|v\|^2_{L^2(\Omega)} + k^{-2}\|\Delta v\|^2_{L^2(\Omega)} \\ &+ L\left(k^2\|v\|^2_{L^2(\Gamma_R)} + \|\nabla_{\Gamma_R} v\|^2_{L^2(\Gamma_R)} + \left\|\frac{\partial v}{\partial n}\right\|^2_{L^2(\Gamma_R)} + k^2\|v\|^2_{L^2(\Gamma_D)} + \|\nabla_{\Gamma_D} v\|^2_{L^2(\Gamma_D)} + \left\|\frac{\partial v}{\partial n}\right\|^2_{L^2(\Gamma_D)}\right), \end{aligned}$$

where $L = \operatorname{diam}(\Omega)$ and ∇_{Γ_D} and ∇_{Γ_R} are the surface gradients on Γ_D and Γ_R respectively. Let $\mathbf{n}_R$ be the outward-pointing unit normal vector to Ω_R, and let $\mathbf{n}_D$ be the outward-pointing unit normal vector to Ω_D (so $\mathbf{n}_D$ is the *inward* pointing normal to Ω on Γ_D). We use the convention that on Γ_D the normal derivative is $\partial v/\partial n = \mathbf{n}_D \cdot \nabla v$, and similarly $\partial v/\partial n = \mathbf{n}_R \cdot \nabla v$ on Γ_R.

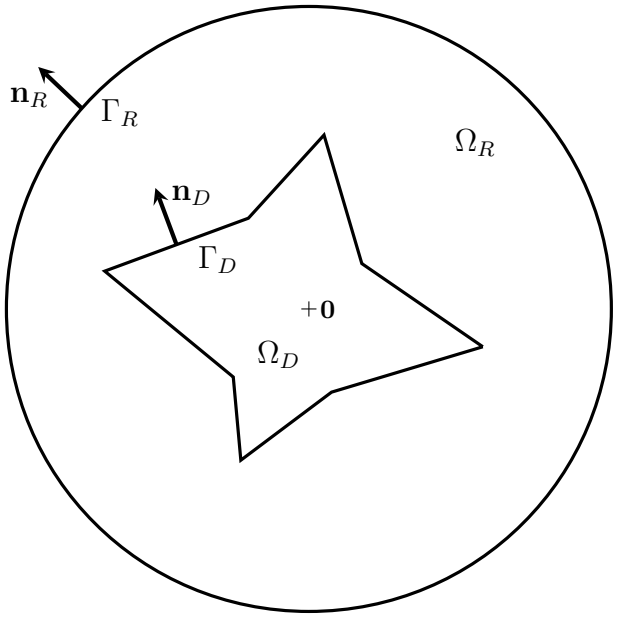

Figure 1: An example of the domains Ω_D and Ω_R in the truncated scattering problem of Definition 4.2.

Define the sesquilinear form $b : V \times V \to \mathbb{C}$ by

$$\begin{aligned} b(u,v) := \int_\Omega \Big(&\big(2-d+\alpha_1+\alpha_2+\mathrm{i}k(\beta_1-\beta_2)\big)\nabla u\cdot\overline{\nabla v} + \big(d-\alpha_1-\alpha_2-\mathrm{i}k(\beta_1-\beta_2)\big)k^2 u\overline{v} \\ &+ \Big(\mathcal{M}_2 u + \frac{A_1}{k^2}\mathcal{L}u\Big)\overline{\mathcal{L}v}\Big)\mathrm{d}\mathbf{x} + \int_{\Gamma_D}\left(\frac{\partial u}{\partial n}\overline{\mathcal{M}_1 v} + LA_2k^2u\overline{v}\right)\mathrm{d}s \\ &- \int_{\Gamma_R}\left(\mathrm{i}k\vartheta u\,\overline{\mathcal{M}_1 v} + (\mathbf{x}\cdot\nabla_{\Gamma_R}u - \mathrm{i}k\beta_2 u + \alpha_2 u)\,\overline{\frac{\partial v}{\partial n}} + (\mathbf{x}\cdot\mathbf{n})\left(k^2u\overline{v} - \nabla_{\Gamma_R}u\cdot\overline{\nabla_{\Gamma_R}v}\right)\right)\mathrm{d}s, \end{aligned} \tag{4.2}$$

and the functional $G : V \to \mathbb{C}$ by

$$\begin{aligned} G(v) := &\int_\Omega \left(\overline{\mathcal{M}_1 v} - \frac{A_1}{k^2}\overline{\mathcal{L}v}\right) f\,\mathrm{d}\mathbf{x} + \int_{\Gamma_R}\overline{\mathcal{M}_1 v}\,g_R\,\mathrm{d}s \\ &- \int_{\Gamma_D}\left((\mathbf{x}\cdot\nabla_{\Gamma_D}g_D - \mathrm{i}k\beta_2 g_D + \alpha_2 g_D)\overline{\frac{\partial v}{\partial n}} + (\mathbf{x}\cdot\mathbf{n})(k^2 g_D\overline{v} - \nabla_{\Gamma_D}g_D\cdot\overline{\nabla_{\Gamma_D}v}) - LA_2k^2 g_D\overline{v}\right)\mathrm{d}s, \end{aligned} \tag{4.3}$$

where $\mathcal{M}_j$, $j = 1, 2$, are defined by (2.3), and $\alpha_1, \alpha_2, \beta_1, \beta_2, A_1$, and A_2 are all arbitrary real parameters. Note that the $b(\cdot,\cdot)$ and $G(\cdot)$ defined by (4.2) and (4.3) respectively are the same as the $b(\cdot,\cdot)$ and $G(\cdot)$ for the interior impedance problem, (3.2) and (3.3) respectively, (identifying Γ_R with $\partial\Omega$) except for extra terms on Γ_D.

Note that we could have formulated the truncated sound-soft scattering problem with a zero Dirichlet boundary condition on Γ_D imposed in the space V (i.e. imposed strongly), as is usually done for the standard variational formulations. (In this case the solution to the truncated problem is an approximation to the total field u^T in Definition 4.1.) We choose not to do this for technical reasons explained below in Remark 4.6.

We now prove the analogues of Proposition 3.2, Lemma 3.3, and Theorem 3.4.

Proposition 4.3 ($b(\cdot,\cdot)$ can be used to solve the truncated sound-soft scattering problem)**.** *If u solves the truncated sound-soft scattering problem of Definition 4.2, then $u \in V$, where V is defined by (1.21), and*

$$b(u,v) = G(v) \quad \text{for all } v \in V,$$

where $b(\cdot,\cdot)$ is given by (4.2) and $G(\cdot)$ by (4.3).

Proof. This is very similar to the proof of Proposition 3.2. The fact that $f \in L^2(\Omega)$ implies that $\Delta u \in L^2(\Omega)$, the fact that $g_R \in L^2(\Gamma_R)$ implies that $\partial u/\partial n \in L^2(\Gamma_R)$, and the fact that $g_D \in H^1(\Gamma_D)$ implies that $u \in H^1(\Gamma_D)$. To show that $u \in V$, we again use the results of Nečas [63, §5.1.2, §5.2.1], [49, Theorem 4.2.4], which show that, for $u \in H^1(\Omega)$ and $\Delta u \in L^2(\Omega)$, the conditions $u \in H^1(\partial\Omega)$ and $\partial u/\partial n \in L^2(\partial\Omega)$ are equivalent. However, the presence of different boundary conditions on Γ_D and Γ_R means that to prove that $\nabla_{\partial\Omega} u \in (L^2(\partial\Omega))^d$ and $\partial u/\partial n \in L^2(\partial\Omega)$ we need to introduce a smooth boundary, Γ^*, between Γ_D and Γ_R and apply the Nečas result first between Γ^* and Γ_D, and then between Γ^* and Γ_R (using interior H^2-regularity of the Laplacian [36, §6.3.1, Theorem 1] and the trace theorem [49, Theorem 3.38] to get $\nabla_{\Gamma^*} u \in (L^2(\Gamma^*))^d$ and $\partial u/\partial n \in L^2(\Gamma^*)$).

To obtain $b(u,v) = G(v)$ we apply the integrated identity (2.8) in Ω and use the boundary conditions (4.1b) and (4.1c). As in the interior case, we add on a multiple of $\mathcal{L}u\overline{\mathcal{L}v}$ to both sides of the identity (involving the constant A_1), but now we also add on a multiple of $u\overline{v}$ on Γ_D (involving the constant A_2); this turns out to be necessary for coercivity. □

Lemma 4.4 (Continuity of $b(\cdot,\cdot)$)**.** *If α_j, β_j, and A_j, $j = 1, 2$, are all independent of k, then*

$$|b(u,v)| \le C_c' \|u\|_V \|v\|_V$$

for all $u, v \in V$ and for all $k > 0$ where

$$C_c' := \max\left\{C_c;\ A_2 + \frac{1}{kL}(k|\beta_1| + |\alpha_1|)\right\},$$

and where C_c is as in Lemma 3.3 with A replaced by A_1.

Proof. This is almost identical to that of Lemma 3.3, except that now there are traces on both Γ_D and Γ_R, and so the vector $\mathbf{m}(u) \in \mathbb{R}^9$. □

Theorem 4.5 (Coercivity of $b(\cdot,\cdot)$)**.** *Let $b(\cdot,\cdot)$ be defined by (4.2) and V defined by (1.21). Suppose that Ω_R is a Lipschitz domain with diameter L that is star-shaped with respect to a ball, i.e. there exists a $\gamma_R > 0$ such that*

$$\mathbf{x} \cdot \mathbf{n}_R(\mathbf{x}) \ge \gamma_R L,$$

for all $\mathbf{x} \in \Gamma_R$ for which $\mathbf{n}_R(\mathbf{x})$ is defined. Suppose that Ω_D is Lipschitz and star-shaped with respect to a ball with the same centre as the previous one, i.e. there exists a $\gamma_D > 0$ such that

$$\mathbf{x} \cdot \mathbf{n}_D(\mathbf{x}) \ge \gamma_D L,$$

for all $\mathbf{x} \in \Gamma_D$ for which $\mathbf{n}_D(\mathbf{x})$ is defined. If

$$\alpha_1 = \alpha_2 = \frac{d-1}{2}, \quad \beta_1 = \beta_2 \ge \frac{L}{2\vartheta_*}\left[1 + 4\frac{(\vartheta^*)^2}{\gamma_R} + \frac{\gamma_R}{2}\right], \quad A_1 = \frac{1}{3}, \quad \text{and} \quad A_2 = 1, \tag{4.4}$$

then, for any $k > 0$,

$$\Re b(v,v) \ge \alpha \|v\|_V^2 \qquad \text{for all } v \in V,$$

with

$$\alpha = \frac{1}{2}\min\left\{\frac{\gamma_R}{2};\ \gamma_D\right\}. \tag{4.5}$$

Proof. This follows the same steps as the proof of Theorem 3.4, with some small differences. As before, the definition of $b(\cdot,\cdot)$ implies that

$$\begin{aligned} 2\Re b(v,v) = &\int_\Omega \left(2(2-d+\alpha_1+\alpha_2)|\nabla v|^2 + 2(d-\alpha_1-\alpha_2)k^2|v|^2 + \frac{2A_1}{k^2}|\mathcal{L}v|^2 + 2\Re\{\mathcal{M}_2 v\overline{\mathcal{L}v}\}\right)\mathrm{d}\mathbf{x} \\ &- 2\int_{\Gamma_R} (\mathbf{x}\cdot\mathbf{n}_R)(k^2|v|^2 - |\nabla_{\Gamma_R} v|^2)\mathrm{d}s - 2\Re\int_{\Gamma_R}\left(\mathrm{i}k\vartheta v\,\overline{\mathcal{M}_1 v} + (\mathbf{x}\cdot\nabla_{\Gamma_R} v - \mathrm{i}k\beta_2 v + \alpha_2 v)\,\overline{\frac{\partial v}{\partial n}}\right)\mathrm{d}s \\ &+ \int_{\Gamma_D}\left(2\Re\left\{\overline{\mathcal{M}_1 v}\frac{\partial v}{\partial n}\right\} + 2LA_2k^2|v|^2\right)\mathrm{d}s. \end{aligned} \tag{4.6}$$

As in the proof of Theorem 3.4 we use the identity (2.8) with $u = v$ to obtain the following expression for the integral over Ω of $2\Re\{\overline{\mathcal{M}_2 v}\mathcal{L}v\}$,

$$\begin{aligned}\int_\Omega 2\Re\{\mathcal{M}_2 v\overline{\mathcal{L}v}\}\,\mathrm{d}\mathbf{x} &= \int_\Omega \Big((d-2-2\alpha_2)|\nabla v|^2 + (2\alpha_2 - d)k^2|v|^2\Big)\mathrm{d}\mathbf{x}\\ &+ \int_{\Gamma_R}\left((\mathbf{x}\cdot\mathbf{n}_R)\left(\left|\frac{\partial v}{\partial n}\right|^2 + k^2|v|^2 - |\nabla_{\Gamma_R} v|^2\right) + 2\Re\left\{\mathbf{x}\cdot\nabla_{\Gamma_R} v - \mathrm{i}k\beta_2 v + \alpha_2 v\right\}\overline{\frac{\partial v}{\partial n}}\right)\mathrm{d}s\\ &- \int_{\Gamma_D}\left(2\Re\left\{\overline{\mathcal{M}_2 v}\frac{\partial v}{\partial n}\right\} + (\mathbf{x}\cdot\mathbf{n}_D)(k^2|v|^2 - |\nabla v|^2)\right)\mathrm{d}s, \end{aligned} \tag{4.7}$$

(recall that $\mathbf{n}_D$ points into Ω and $\mathbf{n}_R$ points out of Ω). We substitute (4.7) into (4.6) and take $\alpha_1 = \alpha_2$, $\beta_1 = \beta_2$, so that $\mathcal{M}_1 v = \mathcal{M}_2 v$ and thus the corresponding terms on Γ_D cancel. We end up with

$$\begin{aligned}2\Re b(v,v) &= \int_\Omega \left((2-d+2\alpha_1)|\nabla v|^2 + (d-2\alpha_1)k^2|v|^2 + \frac{2A_1}{k^2}|\mathcal{L}v|^2\right)\mathrm{d}\mathbf{x}\\ &+ \int_{\Gamma_R}(\mathbf{x}\cdot\mathbf{n}_R)\left(\left|\frac{\partial v}{\partial n}\right|^2 + |\nabla_{\Gamma_R} v|^2 - k^2|v|^2\right)\mathrm{d}s\\ &- 2\Re\int_{\Gamma_R}\left((\mathbf{x}\cdot\mathbf{n}_R)\overline{\frac{\partial v}{\partial n}} + \mathbf{x}\cdot\overline{\nabla_{\Gamma_R} v} + \mathrm{i}k\beta_1\overline{v} + \alpha_1\overline{v}\right)\mathrm{i}k\vartheta v\,\mathrm{d}s\\ &+ \int_{\Gamma_D}\left(2LA_2k^2|v|^2 + (\mathbf{x}\cdot\mathbf{n}_D)\left(\left|\frac{\partial v}{\partial n}\right|^2 + |\nabla_{\Gamma_D} v|^2 - k^2|v|^2\right)\right)\mathrm{d}s. \end{aligned} \tag{4.8}$$

The terms on Γ_R and in Ω are dealt with in exactly the same way as in the proof of Theorem 3.4. The terms on Γ_D are greater than or equal to

$$\gamma_D L\left(\|\nabla_{\Gamma_D} v\|^2_{L^2(\Gamma_D)} + \left\|\frac{\partial v}{\partial n}\right\|^2_{L^2(\Gamma_D)}\right) + L(2A_2 - 1)k^2\|v\|^2_{L^2(\Gamma_D)}.$$

Thus, choosing $A_2 = 1$ and remembering that $\gamma_D \leq 1/2$ (by Remark 3.5) we obtain that $b(\cdot,\cdot)$ is coercive with coercivity constant given by (4.5). □

Note that, in contrast to the interior problem, to prove that $b(\cdot,\cdot)$ is coercive we have had to restrict the values of α_2 and β_2 (i.e. these are no longer free parameters).

Remark 4.6 (Technical considerations)**.** *We formulated the truncated sound-soft scattering problem in terms of the scattered field, with a Dirichlet boundary condition on Γ_D imposed weakly. Instead, we could have formulated the problem in terms of the total field, and imposed the Dirichlet boundary condition in a strong form using the space*

$$V_0 := \left\{v \in H^1(\Omega),\ v|_{\Gamma_D} = 0,\ \Delta v \in L^2(\Omega),\ \nabla v \in \left(L^2(\partial\Omega)\right)^d\right\}.$$

It turns out that the analogous variational formulation is also coercive and continuous on this space, but there is a subtle disadvantage: if D is a Lipschitz polygon or polyhedron with a reentrant corner, then $H^1_0(D,\Delta)\cap H^2(D)$ is not dense in $H^1_0(D,\Delta)$, where $H^1_0(D,\Delta) := \{v \in H^1(D), v|_{\partial D} = 0, \Delta v \in L^2(D)\}$.

Indeed, the fact that, whenever D has reentrant corners, $H^1_0(D,\Delta)\setminus H^2(D)$ is non-empty is well-known (for polygons see [39, Lemma 4.4.3.5], [41, Page 576]). The fact that $H^1_0(D,\Delta)\cap H^2(D)$ is closed in $H^1_0(D,\Delta)$ when D is a 2-d polygon follows from the bound $\|v\|_{H^2(D)} \lesssim \|\Delta v\|_{L^2(\Omega)}$ for all $v \in H^1_0(D,\Delta)\cap H^2(\Omega)$ [39, Theorem 4.3.1.4].

For the truncated sound-soft scattering problem this result implies that if Ω_D has a corner then $V_0 \cap H^2(\Omega)$ is not dense in V_0. However, Lemma 5.1 below implies that any conforming finite element method in V_0 consists of functions that are in $H^2(\Omega)$, and thus these functions are not able to approximate general Helmholtz solutions in V_0. This is analogous to the situation encountered in

the context of the time-harmonic Maxwell equations where $\mathbf{X_N} \cap (H^1(D))^3$ *is not dense in* $\mathbf{X_N}$ *for* D *a non-convex polyhedron, where*

$$\mathbf{X_N} := \left\{ \mathbf{u} \in \left(L^2(D)\right)^3 : \nabla \times \mathbf{u} \in \left(L^2(D)\right)^3, \nabla \cdot \mathbf{u} \in L^2(D), \mathbf{u} \times \mathbf{n} = \mathbf{0} \text{ on } \partial D \right\},$$

see [57, Lemma 3.56]. This is a well-known fact since it prevents H^1*-conforming finite element approximations from converging to singular solutions.*

Remark 4.7 (Bounding the solution of the BVP)**.** *In analogy with the case of the interior problem discussed in Remark 3.6, the coercivity result Theorem 4.5, together with (1.9), gives the following stability bound on the solution of the BVP:*

$$\begin{aligned} \|u\|_V \leq & \frac{4\sqrt{3}\left(2 + \frac{\beta_1}{L} + \frac{d}{2kL}\right)}{\min\{\gamma_R;\ 2\gamma_D\}} \\ & \cdot \left(L^2 \|f\|^2_{L^2(\Omega)} + L \|g_R\|^2_{L^2(\Gamma_R)} + k^2 L \|g_D\|^2_{L^2(\Gamma_D)} + L \|\nabla_{\Gamma_D} g_D\|^2_{L^2(\Gamma_D)} \right)^{1/2} \end{aligned} \tag{4.9}$$

for all $k > 0$*. A bound identical to this one in its* k*-dependence was obtained in [43, Proposition 3.3]. Although only the case* $g_D \equiv 0$ *was considered there, the same method can be used to obtain a bound for the case of non-homogeneous Dirichlet boundary conditions.*

Remark 4.8 (The analogous scattering problems for first-order systems and Maxwell's equations)**.** *The truncated sound-soft scattering problem of Definition 4.2 can be rewritten as a first-order system, and a continuous and coercive variational formulation of this problem exists (similar to the case of the interior impedance problem discussed in Remark 3.9). We have not, however, been able to extend this formulation to the first-order system for the analogous Maxwell BVP (as we could in the interior impedance case—see Remark 3.10). This is consistent with the fact that, to the authors' knowledge, there do not yet exist any wavenumber-explicit stability bounds for this Maxwell BVP. (If we had a continuous and coercive formulation, then we would have such a bound by the consequence of the Lax–Milgram theorem (1.9).)*

5 Implications for numerical methods

The primary aim of this paper is to introduce the new coercive formulations of Sections 3 and 4 as results about the Helmholtz equation itself, independent of potential discretisations. It is not yet clear whether these new formulations will be useful computationally. The property of coercivity, however, immediately implies results about possible Galerkin discretisations of the new formulations, and thus it would seem a shame not to discuss these results here.

In this section, therefore, we briefly begin to investigate potential discretisations of the new variational formulations. The actual implementation of these discretisations, a more thorough study of their properties, and comparison to standard discretisations will be presented elsewhere.

5.1 Conforming finite element methods

We first show that, for Ω a general bounded Lipschitz domain, the requirement in the space V (defined by (1.21)) that $\Delta v \in L^2(\Omega)$ means that any conforming finite element method in this space must use C^1-elements.

Lemma 5.1 (C^1-conformity)**.** *Let* $\mathcal{T}$ *be a triangulation of* Ω *(in the sense of [21, Page 61]). If* $v \in V$ *is also in* $C^2(\overline{K})$ *for each element* $K \in \mathcal{T}$*, then* $v \in C^1(\overline{\Omega}) \cap H^2(\Omega)$*.*

Proof. The fact that a piecewise C^k function belongs to $H^k(\Omega)$ if and only if it belongs to $C^{k-1}(\overline{\Omega})$ is well-known (see, e.g., [11, Theorem II.5.2] or [21, Theorems 5.1 and 30.1]); the proof of this lemma is extremely similar to the proof of the forward implication. Since $V \subset H^1(\Omega)$, for any $v \in V$ that is piecewise C^2 we have that $v \in C(\overline{\Omega})$; thus we only need to show that $\nabla v \in C(\overline{\Omega})$, and then $v \in H^2(\Omega)$ follows from the result mentioned above.

Since $\Delta v \in L^2(\Omega)$ we have that, for any $\phi \in C^\infty_{\rm comp}(\Omega) := \{v \in C^\infty(\Omega), \operatorname{supp} v \subset\subset \Omega\}$,

$$\int_\Omega \phi\, \Delta v \,\mathrm{d}\mathbf{x} = \int_\Omega v\, \Delta\phi \,\mathrm{d}\mathbf{x},$$

and thus

$$\sum_{K\in\mathcal{T}} \int_K (\phi\, \Delta v - v\, \Delta\phi)\mathrm{d}\mathbf{x} = 0.$$

Since $v \in C^2(\overline{K})$, Green's second identity (1.33) applied to each element implies that

$$\sum_{K\in\mathcal{T}} \int_{\partial K} \left(\phi \frac{\partial v}{\partial n} - v\frac{\partial \phi}{\partial n}\right) \mathrm{d}s = 0.$$

Now, since $\phi \in C^\infty_{\rm comp}(\Omega)$ and $v \in C(\overline{\Omega})$, $\sum_{K\in\mathcal{T}} \int_{\partial K} v\, \partial\phi/\partial n \,\mathrm{d}s = 0$ (the integrals over interior edges/faces cancel and ϕ is zero on $\partial\Omega$); thus we are left with

$$\sum_{K\in\mathcal{T}} \int_{\partial K} \phi \frac{\partial v}{\partial n} \mathrm{d}s = 0.$$

Since ϕ is an arbitrary member of $C^\infty_{\rm comp}(\Omega)$, this last equation can only hold if $\partial v/\partial n$ is continuous across each edge/face. Continuity of the tangential part of ∇v follows from the continuity of v across edges, and thus $\nabla v \in C(\overline{\Omega})$. □

For any conforming subspace, the continuity and coercivity properties of the new formulations imply that the corresponding Galerkin methods are quasi-optimal without any constraint on the subspace dimension, albeit with the factor in front of the best approximation error growing with k. For simplicity we state this result for the case of the interior impedance problem of Definition 3.1, but a completely analogous result is valid in the case of the truncated sound-soft scattering problem of Definition 4.2.

Proposition 5.2 (Quasi-optimality)**.** *Suppose that the interior impedance problem of Definition 3.1 is solved using the variational formulation of Proposition 3.2 (with the constants α_j, β_j, and A chosen as in Lemma 3.3 and Theorem 3.4, and β_1 chosen proportional to L), with V_N a finite dimensional subspace of V. Then there exists a $C_{qo} > 0$ (depending only on $d, \vartheta_*, \vartheta^*$ and γ) such that*

$$\|u - u_N\|_V \leq C_{qo}\big(kL + (kL)^{-1}\big) \inf_{v_N\in V_N} \|u - v_N\|_V, \tag{5.1}$$

for any $N > 0$ and for all $k > 0$.

Proof. This is a simple consequence of Céa's lemma (1.11), the coercivity result of Theorem 3.4, and the bound on the continuity constant given in Lemma 3.3. □

That the factor in front of the best approximation error grows with k is somehow expected because of the *pollution effect* [46, §4.6] (which is a special case of the *locking* phenomenon as described in, e.g., [11, §VI.3]). Indeed, if this factor were bounded independently of k then we would have proved that this method did not suffer from the pollution effect (in the sense of [2, Definition 2.1] in the norm $\|\cdot\|_V$) for any choice of V_N. However, it is widely believed that no FEM converging in h (in $d \geq 2$) can be pollution-free, as was proved for a wide class of generalised FEMs in [2, Theorem 4.6].

Since we have established quasi-optimality, we only need to consider the approximation of the solution, i.e. for what subspaces does the best approximation error on the right-hand side of (5.1) tend to zero as $N \to \infty$.

Given a family $\mathbb{V} = \{V_N\}_{N\in\mathbb{N}}$ of finite dimensional subspaces of V, a norm $\|\cdot\|_W$ and $W := \{w \in V \text{ s.t. } \|w\|_W < \infty\}$, we say that $\mathbb{V}$ *approximates* W if

$$\lim_{N\to\infty} \inf_{v_N\in V_N\cap W} \|w - v_N\|_W = 0 \qquad \text{for all } w \in W.$$

Lemma 5.3. *A family of C^1-elements that approximates $H^2(\Omega)$ also approximates V.*

Proof. In Appendix A we prove that $\mathcal{D}(\overline{\Omega})$ is dense in V, and so given $u \in V$ and $\varepsilon > 0$ there exists a $w \in \mathcal{D}(\overline{\Omega})$ such that $\|u - w\|_V < \varepsilon/2$. From the inclusion $H^2(\Omega) \subseteq V$ there exists a constant C such that $\|v\|_V \leq C \|v\|_{H^2(\Omega)}$ for every $v \in H^2(\Omega)$. By Lemma 5.1 and the approximation property in $H^2(\Omega)$, for N large enough we can choose a C^1-element function v_N (which also belongs to $H^2(\Omega)$) such that $\|w - v_N\|_{H^2(\Omega)} < \varepsilon/(2C)$. Then, by the triangle inequality, $\|u - v_N\|_V < \varepsilon$. $\square$

Conditions for polynomial C^1-elements to be dense in $H^2(\Omega)$ are given in [21, Theorem 48.2], and rates of convergence under the assumption of additional regularity are given in [21, Figure 48.1] and [11, Table 3, § II.6]. Note that for the standard variational formulations of Laplace and Helmholtz problems one aims to prove convergence for solutions that belong to $H^{1+s}(\Omega)$ for some $0 < s \leq 1$ and then obtain a convergence rate for solutions in $H^2(\Omega)$ (see e.g. [12, Theorem 5.4.4]). For the new formulation, however, the standard results cited above only give a convergence rate for solutions at least in $H^3(\Omega)$.

An alternative to using piecewise-polynomial basis functions would be to use oscillatory basis functions from the Partition of Unity FEM [51], with the partition of unity chosen so that the elements are C^1-conforming. Convergence rates for plane or spherical wave bases can then be obtained by slightly modifying the proof of [51, Theorem 2.1] and using the approximation results in [50, §8.4], [56, Chapter 3].

A recent interesting development in finite elements has been the introduction of so-called Virtual Element Methods (VEMs); see [7]. The key ideas underpinning the VEM are the use of a piecewise polynomial space enriched with other functions and a choice of the degrees of freedom (DOFs) that allows DOF-based computations; a crucial example of the latter property is that the stiffness matrix can be assembled without computing the non-polynomial basis functions. One of the strengths of these methods is that they allow C^1-conforming discretisations of BVPs involving fourth order operators (such as the biharmonic equation) to be implemented almost as easily as C^0-elements for second order equations; see [13]. The key ingredient for the design of a VEM scheme is a coercive formulation obtained from multiplying the PDE by a test function and integrating by parts; thus the new formulations in this paper seem to be, at least in principle, amenable to this kind of discretisation (and investigations in this direction are currently underway).

5.2 Discrete conditioning and convergence of iterative solvers

Assume that we have a conforming finite element method, with family of subspaces $V_N = \operatorname{span}\{\phi_i : i = 1, \ldots, N\} \subset V$. (We also denote the subspaces V_h when we are explicitly thinking of them as coming from a mesh with meshwidth h.)

Let $b(\cdot, \cdot)$ and $G(\cdot)$ be the sesquilinear form and the antilinear functional introduced either in Section 3 or in Section 4. Define

$$\mathbf{B}_{ij} := b(\phi_j, \phi_i) \quad \text{and} \quad \mathbf{g}_i := G(\phi_i) \quad \text{for } i, j = 1, \ldots, N.$$

Then, if $u_N = \sum_{i=1}^N U_i \phi_i$, $v_N = \sum_{i=1}^N V_i \phi_i$, $\mathbf{u} := (U_i) \in \mathbb{C}^N$, and $\mathbf{v} := (V_i) \in \mathbb{C}^N$, the standard properties of sesquilinear forms imply that

$$(\mathbf{B}\mathbf{u}, \mathbf{v}) = b(u_N, v_N), \tag{5.2}$$

where $(\cdot, \cdot)$ denotes the standard Euclidean inner product for vectors in $\mathbb{C}^N$. The Galerkin method is then equivalent to solving the linear system

$$\mathbf{B}\mathbf{u} = \mathbf{g}.$$

For simplicity we only consider the h-version of the FEM. We use the notation $a \lesssim b$ to mean $a \leq c\, b$, where c is independent of h, k, and L, and $a \sim b$ to mean that both $a \lesssim b$ and $b \lesssim a$.

Proposition 5.4 (Bounds on the discrete condition number)**.** *Let $\mathcal{T}^h, 0 < h \leq 1$, be a quasi-uniform family of triangulations of Ω (in the sense of [21, Pages 61 and 135]) and let $V_h \subset V$ consist of piecewise polynomials of degree $\leq p$, for some fixed p, that are in $C^1(\overline{\Omega})$ with basis*

functions scaled such that $\|v_h\|^2_{L^2(\Omega)} \sim h^d \|\mathbf{v}\|_2^2$ *for all* $v_h \in V_h$. *Then, if* $hk \lesssim 1$, *the condition number* $\kappa(\mathbf{B}) := \|\mathbf{B}\|_2 \, \|\mathbf{B}^{-1}\|_2$ *satisfies*

$$\kappa(\mathbf{B}) \lesssim \frac{1}{h^4 k^2} \left(L + \frac{1}{k} \right) \left(L + \frac{1}{k^2 L} \right). \tag{5.3}$$

Proof. If a sesquilinear form $b(\cdot,\cdot)$ is continuous and coercive with constants C_c and α respectively (as in (1.7) and (1.8)), and $M_1, M_2 > 0$ are such that

$$M_1 \|\mathbf{v}\|_2^2 \leq \|v_h\|_V^2 \leq M_2 \|\mathbf{v}\|_2^2 \qquad \text{for all } \mathbf{v} \in \mathbb{C}^N, \tag{5.4}$$

then the relationship (5.2) implies that, for all $\mathbf{u}, \mathbf{v} \in \mathbb{C}^N$,

$$|(\mathbf{B}\mathbf{u}, \mathbf{v})| \leq M_2 C_c \|\mathbf{u}\|_2 \|\mathbf{v}\|_2 \quad \text{and} \quad |(\mathbf{B}\mathbf{v}, \mathbf{v})| \geq M_1 \alpha \|\mathbf{v}\|_2^2; \qquad \text{thus} \quad \kappa(\mathbf{B}) \leq \frac{M_2 C_c}{M_1 \alpha}. \tag{5.5}$$

The bounds on the continuity constant and coercivity constant of $\mathbf{B}$, given by Lemma 3.3 and Theorem 3.4 respectively, imply that (as in the proof of Proposition 5.2)

$$\frac{C_c}{\alpha} \lesssim kL + \frac{1}{kL}.$$

Therefore, to bound $\kappa(\mathbf{B})$, we only need to find M_1, M_2 such that the norm equivalence (5.4) holds.

Now

$$\|v_h\|_V^2 \geq k^2 \|v_h\|^2_{L^2(\Omega)} \sim k^2 h^d \|\mathbf{v}\|_2^2,$$

so the first inequality in (5.4) holds for some $M_1 \gtrsim k^2 h^d$.

To obtain an upper bound on $\|v_h\|_V^2$ in terms of $\|\mathbf{v}\|_2^2$ we use the inverse estimates (e.g. [11, II.6.8], [12, Lemma 4.5.3], [21, Theorem 17.2])

$$|v_h|_{H^s(\Omega)} \lesssim h^{-s} \|v_h\|_{L^2(\Omega)} \qquad \text{for all } v_h \in V_h, \quad s = 1, 2,$$

(note that this is where we need the assumptions that the mesh is quasi-uniform and that the basis is piecewise-polynomial), the multiplicative trace inequality ([39, Theorem 1.5.1.10, last formula on Page 41], [12, Theorem 1.6.6])

$$\|v\|^2_{L^2(\partial\Omega)} \lesssim \|v\|_{L^2(\Omega)} \left(L^{-1} \|v\|_{L^2(\Omega)} + |v|_{H^1(\Omega)} \right) \qquad \text{for all } v \in H^1(\Omega),$$

and the relation $\|v_h\|^2_{L^2(\Omega)} \sim h^d \|\mathbf{v}\|^2$. The result is that the second inequality in (5.4) holds for some M_2 such that

$$M_2 \lesssim \left(\frac{1}{h^2} + k^2 + \frac{1}{h^4 k^2} + k^2 + k^2 \frac{L}{h} + \frac{1}{h^2} + \frac{L}{h^3} \right) h^d \lesssim \frac{1}{h^4 k} \left(L + \frac{1}{k} \right) h^d,$$

where we used the facts that $h \leq L$ and $hk \lesssim 1$. Combining the bounds on C_c/α, M_1, and M_2 yields the result. □

Remark 5.5 (Discussion of the bound on the condition number)**.** *There are two interesting limits under which to consider the bound (5.3): $h \to 0$ and $k \to \infty$. In the limit $h \to 0$ for fixed k, $\kappa(\mathbf{B}) \sim h^{-4}$; this is expected because of the presence of Δv in the norm and the consequent use of inverse estimates for the H^2-seminorm (compare to FEMs for the biharmonic problem). In the limit $k \to \infty$, we need to tie h to k, since if h is fixed then the best approximation error is not bounded as $k \to \infty$. It is commonly believed that $hk \sim 1$ keeps the relative best approximation error bounded as k increases, although this has only been rigorously proved for certain 1-d problems [46, Equation 4.4.3] and [17, Lemma 6.6]. Under the scaling $hk \sim 1$, $\kappa(\mathbf{B}) \sim k^2$ as $k \to \infty$ (although from Proposition 5.2 we expect some pollution in this limit, and thus some loss of accuracy of the Galerkin solution at large k). There do not yet exist any comparable results about the conditioning of the standard formulation (1.19) to compare this to.*

As discussed in §1, the sign-indefiniteness of the standard variational formulations of the Helmholtz equation is a major issue when solving the resulting linear systems with iterative solvers such as GMRES. We now briefly investigate whether or not we can determine anything a priori about how GMRES behaves when applied to $\mathbf{Bu} = \mathbf{g}$. Of course, linear systems arising from FEMs are usually preconditioned before being solved using GMRES (for a description of the state-of-the-art preconditioners for the Helmholtz equation with large k see the recent reviews [34], [35], [33], [1]), however here we just consider applying GMRES to the unpreconditioned problem.

We use the fact that, for $m \in \mathbb{N}$, the m-th GMRES residual $\mathbf{r}_m$ satisfies

$$\frac{\|\mathbf{r}_m\|_2}{\|\mathbf{r}_0\|_2} \leq \sin^m \beta, \quad \text{where} \quad \cos\beta = \frac{\operatorname{dist}\big(0, W(\mathbf{B})\big)}{\|\mathbf{B}\|_2}, \tag{5.6}$$

where $W(\mathbf{B}) := \{(\mathbf{Bx}, \mathbf{x}) : \mathbf{x} \in \mathbb{C}^N, \|\mathbf{x}\|_2 = 1\}$ is the *numerical range* of $\mathbf{B}$. This bound was originally proved in [31] (see also [30, Theorem 3.3]) and appears in the form above in [6, Equation 1.2].

It follows from (5.5) that $\cos\beta \geq M_1\alpha/(M_2C_c)$ where M_1 and M_2 are as in (5.4). Using the bounds on M_1 and M_2 in the proof of Proposition 5.4, one can then prove that, given $\varepsilon > 0$ and $k_0 > 0$, there exists a $C_1 > 0$ independent of k, h, L, and ε such that

$$m \geq \frac{C_1 L^4}{h^8 k^4} |\log\varepsilon| \quad \text{implies that} \quad \frac{\|\mathbf{r}_m\|_2}{\|\mathbf{r}_0\|_2} \leq \varepsilon, \tag{5.7}$$

for all $k \geq k_0$. To understand this bound better, consider the case $hk \sim 1$ (which, as discussed above, is thought to keep the relative best approximation error under control) and ignore the dependence on L; the bound then becomes $m \gtrsim k^4$. Unfortunately this bound is not practical, since if $hk \sim 1$ then $N \sim k^d$, and (in exact arithmetic) GMRES always converges once the number of iterations, m, reaches the dimension N of the linear system. It is instructive to note that two of the powers of k in $m \gtrsim k^4$ arise from the fact that $C_c/\alpha \sim k$, and two powers come from the norm in V, so even if the method were pollution-free, i.e. if C_c/α were bounded independently of k, then the estimate (5.6) would give $m \gtrsim k^2$, which is still not particularly useful. (Similarly, a hypothetical H^1-conforming scheme with continuity and coercivity properties similar to those of Section 3 would also give $m \gtrsim k^2$.)

In summary, although the bound (5.6) allows us to determine k-explicit, a priori information about the behaviour of (unpreconditioned) GMRES from the continuity and coercivity properties of the new formulation, the resulting bounds do not yield any practical information when k is large.

6 Discussion of the geometric restrictions on the new formulations

The new formulations in Sections 3 and 4 both require that certain domains be star-shaped with respect to a ball. In this section we discuss whether these restrictions can be lifted. This is perhaps more easily understandable for the exterior problem, so we begin here.

6.1 The sound-soft scattering problem

The coercive formulation of the truncated sound-soft scattering problem in Section 4 needed both Ω_D (the obstacle) and Ω_R (the interior of the artificial boundary) to be star-shaped with respect to a ball. Indeed, the proof of coercivity required that $\mathbf{x} \cdot \mathbf{n}_D(\mathbf{x}) > 0$ for $\mathbf{x} \in \Gamma_D$ and $\mathbf{x} \cdot \mathbf{n}_R(\mathbf{x}) > 0$ for $\mathbf{x} \in \Gamma_R$. Replacing the vector field $\mathbf{x}$ in the identity (1.28) by an arbitrary vector field $\mathbf{Z}(\mathbf{x})$, one can show that there exists a coercive formulation of the truncated sound-soft scattering problem, for k sufficiently large, if there exists a $\mathbf{Z}(\mathbf{x})$ such that

$$\mathbf{Z}(\mathbf{x}) \cdot \mathbf{n}_D(\mathbf{x}) > 0 \text{ for } \mathbf{x} \in \Gamma_D, \tag{6.1a}$$

$$\mathbf{Z}(\mathbf{x}) \cdot \mathbf{n}_R(\mathbf{x}) > 0 \text{ for } \mathbf{x} \in \Gamma_R, \text{ and} \tag{6.1b}$$

$$\text{there exists a } \theta > 0 \text{ such that } \Re\{\partial_i \mathbf{Z}_j(\mathbf{x}) \xi_i \overline{\xi_j}\} \geq \theta |\xi|^2 \text{ for all } \xi \in \mathbb{C}^d \text{ and for all } \mathbf{x} \in \Omega, \tag{6.1c}$$

(the last condition ensures positivity of the volume terms of the sesquilinear form). The choice $\mathbf{Z}(\mathbf{x}) = \mathbf{x}$ satisfies these conditions when Ω_D and Ω_R are star-shaped with respect to a ball; for what other domains does such a $\mathbf{Z}(\mathbf{x})$ exist? Note that since the choice of Ω_R is up to us when using the truncated problem to approximate the full scattering problem, we are really interested in obtaining an appropriate $\mathbf{Z}(\mathbf{x})$ for a wider class of Ω_D.

For Helmholtz problems in domains exterior to a bounded obstacle, the key geometric condition is that of *nontrapping*. Roughly speaking, an exterior domain is nontrapping if any ray hitting the obstacle and then reflecting with the angle of incidence equal to the angle of reflection eventually escapes to infinity (after multiple reflections if necessary). For example, one can show that star-shaped domains are nontrapping. In contrast, trapping domains can "trap" certain rays in a neighbourhood of the obstacle for an arbitrary long time. (The review [17, §5.2] contains a more precise discussion of trapping and nontrapping which is aimed at numerical analysts but contains references to the more technical definitions.)

Morawetz, Ralston, and Strauss showed in [62, §4] that if $\Omega_+ := \mathbb{R}^d \setminus \overline{\Omega_D}$ is a 2-dimensional nontrapping domain, then, with Ω_R the ball of radius R for some sufficiently large $R > 0$, there exists a $\mathbf{Z}(\mathbf{x})$ in $\Omega := \Omega_R \setminus \overline{\Omega_D}$ such that

$$\mathbf{Z}(\mathbf{x}) \cdot \mathbf{n}_D(\mathbf{x}) > 0 \text{ for } \mathbf{x} \in \Gamma_D, \tag{6.2a}$$

$$\mathbf{Z}(\mathbf{x}) = \mathbf{x} \text{ for } \mathbf{x} \in \Gamma_R, \text{ and} \tag{6.2b}$$

$$\Re\{\partial_i \mathbf{Z}_j(\mathbf{x}) \xi_i \overline{\xi_j}\} \geq 0 \text{ for all } \xi \in \mathbb{C}^d \text{ and for all } \mathbf{x} \in \Omega. \tag{6.2c}$$

This $\mathbf{Z}(\mathbf{x})$ satisfies (6.1a) and (6.1b), but not quite (6.1c). Although it is not immediately clear whether the construction of the $\mathbf{Z}(\mathbf{x})$ of [62, §4] can be suitably modified to obtain a $\mathbf{Z}(\mathbf{x})$ satisfying the more stringent requirements (6.1), the similarity of the conditions (6.1) and (6.2) indicates that it is reasonable to believe that a coercive formulation of the truncated sound-soft scattering problem exists if Ω_+ is nontrapping (or perhaps satisfies a slightly more restrictive condition). However, although the existence of a $\mathbf{Z}(\mathbf{x})$ satisfying (6.2) is shown constructively in [62, §4], it is not immediately clear how easily this $\mathbf{Z}(\mathbf{x})$ can be evaluated numerically (which would be a requirement if a variational formulation involving a similar $\mathbf{Z}(\mathbf{x})$ were to be implemented practically).

In addition, there is a good reason to believe that coercive formulations *cannot* exist for trapping domains (or at least not formulations that are coercive uniformly in k). Indeed, one of the consequences of coercivity is the bound on the solution (4.9). An analogous bound holds for the sound-soft scattering problem of Definition 4.1 in nontrapping domains, with the norm of the solution (weighted with k as in (1.22)) bounded uniformly by norms of the data [72], [60] (see the discussion in [17, Theorem 5.6 and Remark 5.9]). However, for certain trapping domains the norm of the solution operator can grow exponentially with k (see, e.g., [17, §5.6.2, Page 221]). Thus, if a coercive formulation of the truncated sound soft scattering problem existed for these trapping domains, and $b(\cdot,\cdot)$ and $G(\cdot)$ were normalised so that $\|G\|_{V'} \lesssim \|f\|_{L^2(\Omega)} + \|g_R\|_{L^2(\Gamma_R)} + k\|g_D\|_{L^2(\Gamma_D)} + \|\nabla_{\Gamma_D} g_D\|_{L^2(\Gamma_D)}$, (with the omitted constant independent of k as in the formulation of §4), then the coercivity constant would have to decrease exponentially with k.

6.2 The interior impedance problem

The coercive formulation of the interior impedance problem in Section 3 required the bounded domain Ω to be star-shaped with respect to a ball, with the inequality $\mathbf{x} \cdot \mathbf{n}(\mathbf{x}) > 0$ used often in the proof of coercivity. Similar to above, replacing the vector field $\mathbf{x}$ in the identity (1.28) by $\mathbf{Z}(\mathbf{x})$, one can show that there exists a coercive formulation of the interior impedance problem if there exists a $\mathbf{Z}(\mathbf{x})$ such that

$$\mathbf{Z}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) > 0 \text{ for } \mathbf{x} \in \Gamma, \text{ and} \tag{6.3a}$$

$$\text{there exists a } \theta > 0 \text{ such that } \Re\{\partial_i \mathbf{Z}_j(\mathbf{x}) \xi_i \overline{\xi_j}\} \geq \theta |\xi|^2 \text{ for all } \xi \in \mathbb{C}^d \text{ and for all } \mathbf{x} \in \Omega. \tag{6.3b}$$

The choice $\mathbf{Z}(\mathbf{x}) = \mathbf{x}$ satisfies these conditions for Ω that are star-shaped with respect to a ball. It is not clear, however, how to construct such a $\mathbf{Z}$ for more general domains; although it is straightforward to construct a $\mathbf{Z}$ satisfying (6.3a) (see [39, Lemma 1.5.1.9], [73, Theorem 1.12 (vi)]), satisfying (6.3b) is much more difficult. (Note that the impedance boundary condition corresponds

to the boundary taking energy away from any impinging wave, and thus the concepts of trapping and nontrapping, relying on energy conservation, do not apply to this problem.)

Regarding bounds on the solution in terms of the data, the currently best available ones for the interior impedance problem in general Lipschitz domains have positive powers of k on the right-hand sides [69, Theorem 1.6]. Whether these bounds are sharp is not yet known; if they are sharp, then any formulation that is coercive for general Lipschitz domains would have the coercivity constant decreasing at least polynomially in k (assuming $b(\cdot,\cdot)$ and $G(\cdot)$ are normalised such that $\|G\|_{V'} \lesssim \|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)}$ with the omitted constant independent of k).

7 Concluding remarks

This paper began by questioning whether the Helmholtz equation should be described as "sign-indefinite". The fact remains that the standard variational formulations of the Helmholtz equation are sign-indefinite. However, we hope that by introducing the sign-definite formulations in this paper, which are obtained in a manner similar to how the standard variational formulations are obtained (i.e. by multiplying by a test function and integrating by parts), we will at least make the reader hesitate if they ever find themselves writing "the Helmholtz equation is sign-indefinite"!

Acknowledgements

The authors thank the following for useful comments and discussions: Xavier Antoine (Université de Lorraine), Anthony Ashton (University of Cambridge), Heikko Berninger (Université de Genève), Geoffrey Burton (University of Bath), Simon Chandler-Wilde (University of Reading), Xiaobing Feng (University of Tennessee), Martin Gander (Genève), Mahadevan Ganesh (Colorado School of Mines), Ivan Graham (Bath), David Hewett (Reading), Ralf Hiptmair (ETH Zürich), Ilia Kamotski (University College London), Markus Melenk (TU Wien), Peter Monk (University of Delaware), Roger Moser (Bath), Ilaria Perugia (Università di Pavia), Valery Smyshlyaev (UCL), Alastair Spence (Bath), and Elisabeth Ullmann (Bath).

The authors also thank the anonymous referees for their thoughtful comments that greatly improved the presentation of the paper.

A.M. was supported by SNSF fellowship 137294 and E.A.S. by EPSRC grant EP/1025995/1.

A Appendix: Density of $\mathcal{D}(\overline{\Omega})$ in the space V

Lemma A.1. *Let Ω be a bounded Lipschitz domain. Then $\mathcal{D}(\overline{\Omega}) := \{U|_\Omega : U \in C^\infty(\mathbb{R}^d)\}$ is dense in the space V defined by (1.21).*

Proof. In this proof we use γ to denote the trace operator $H^s(\Omega) \to H^{s-1}(\partial\Omega)$ for $1/2 < s < 3/2$ (see, e.g, [49, Theorem 3.38,]). We also use the notation $\mathcal{D}(\Omega)$ for $C^\infty_{\rm comp}(\Omega) = \{v \in C^\infty(\Omega),\, \operatorname{supp} v \subset\subset \Omega\}$.

Via a partition of unity it is sufficient to consider the case of a Lipschitz hypograph, i.e.

$$\Omega := \big\{(x', x_d) \subset \mathbb{R}^d : x' \in \mathbb{R}^{d-1},\, x_d > f(x')\big\},$$

where $f : \mathbb{R}^{d-1} \to \mathbb{R}$ is in $C^{0,1}(\mathbb{R}^{d-1})$ (for examples of this method of arguing, see, e.g. [17, §A.2] and [49, Page 89 onwards]). Since Ω is now unbounded, we define V as the space of functions u such that $\|u\psi\|_V < \infty$, for any $\psi \in \mathcal{D}(\overline{\Omega})$, where $\|\cdot\|_V$ is defined by (1.22).

The main idea of the proof is that a given $u \in V$ can be approximated by u_t where, for $t > 0$,

$$u_t(\mathbf{x}) := u(\mathbf{x} + t\mathbf{e}_d),$$

where $\mathbf{e}_d$ is the unit vector in the x_d-direction. Thus, for $\mathbf{x} \in \partial\Omega$, $u_t(\mathbf{x})$ is u evaluated on a parallel surface to $\partial\Omega$, at a distance t above. Now, by standard interior regularity results for the Laplacian applied to bounded subsets of Ω (see, e.g., [36, §6.3.1], [49, Theorem 4.16]), we have that $u \in H^2_{\rm loc}(\Omega)$, i.e. $\chi u \in H^2(\Omega)$ for every $\chi \in \mathcal{D}(\Omega)$, and thus $\psi u_t \in H^2(\Omega)$ for every $\psi \in \mathcal{D}(\overline{\Omega})$. The

key point is that u_t and all its derivatives of order ≤ 2 are square-integrable in any bounded subset of Ω (including subsets that share part of their boundary with Ω) but this is not in general true for u.

The main part of the proof consists of showing that $\|(u-u_t)\psi\|_V \to 0$ as $t \to 0$, for any $\psi \in \mathcal{D}(\overline{\Omega})$. Assuming this result holds, we have that given $\varepsilon > 0$ and $\psi \in \mathcal{D}(\overline{\Omega})$ there exists a $t > 0$ such that $\|(u-u_t)\psi\|_V < \varepsilon/2$. Let C_ψ be such that $\|v\psi\|_V \leq C_\psi \|v\psi\|_{H^2(\operatorname{supp}\psi)}$ for every $v \in H^2(\operatorname{supp}\psi)$. Since the restriction of $\mathcal{D}(\overline{\Omega})$ is dense in $H^2(\operatorname{supp}\psi)$ [49, Page 77], there exists a $w \in \mathcal{D}(\overline{\Omega})$ such that $\|(w-u_t)\psi\|_{H^2(\Omega)} < \varepsilon/(2C_\psi)$. Noting that $\|(w-u_t)\psi\|_V \leq C_\psi \|(w-u_t)\psi\|_{H^2(\Omega)} < \varepsilon/2$, we then have that $\|(u-w)\psi\|_V < \varepsilon$ by the triangle inequality, and so we are done.

Therefore, we need only prove that, for all $\psi \in \mathcal{D}(\overline{\Omega})$, $\|(u-u_t)\psi\|_V \to 0$ as $t \to 0$, and we do this by considering each of the terms in $\|(u-u_t)\psi\|_V$ separately.

We first show that $\|(u-u_t)\psi\|_{H^1(\Omega)} \to 0$ as $t \to 0$. To do this, choose $v \in H^1(\mathbb{R}^d)$ such that $v|_{\Omega_1} = u$, where $\Omega_1 := \{\mathbf{x} \in \Omega : \operatorname{dist}(\mathbf{x}, \operatorname{supp}\psi) < 1\}$. Then define $v_t(\mathbf{x}) := v(\mathbf{x} + t\mathbf{e}_d)$ for $t > 0$, and thus $v_t|_{\operatorname{supp}\psi} = u_t$ for any $0 < t < 1$. These definitions immediately imply that, for $0 < t < 1$ and some $C > 0$,

$$\|(u-u_t)\psi\|_{H^1(\Omega)} \leq C\,\|u-u_t\|_{H^1(\operatorname{supp}\psi)} \leq C\,\|v-v_t\|_{H^1(\mathbb{R}^d)}\,. \tag{A.1}$$

If $v \in C^\infty_{\text{comp}}(\mathbb{R}^d)$ then $\|v-v_t\|_{H^1(\mathbb{R}^d)} \to 0$ as $t \to 0$, and thus, by the density of $C^\infty_{\text{comp}}(\mathbb{R}^d)$ in $H^1(\mathbb{R}^d)$, this is also true for $v \in H^1(\mathbb{R}^d)$. The inequality (A.1) then implies that $\|(u-u_t)\psi\|_{H^1(\Omega)} \to 0$ as $t \to 0$.

In order to show that $\|\Delta((u-u_t)\psi)\|_{L^2(\Omega)} \to 0$ as $t \to 0$, we only need to show that $\|(\Delta u - \Delta u_t)\psi\|_{L^2(\Omega)} \to 0$, since the terms involving $(u-u_t)\Delta\psi$ and $\nabla(u-u_t)\cdot\nabla\psi$ are bounded by $\|u-u_t\|_{H^1(\operatorname{supp}\psi)}$, which tends to zero by the previous paragraph. Therefore, we need to show that

$$\int_\Omega (\Delta u - \Delta u_t)\psi\phi\,\mathrm{d}\mathbf{x} \to 0 \quad \text{for all } \phi \in L^2(\Omega),$$

and since $\mathcal{D}(\Omega)$ is dense in $L^2(\Omega)$ we only need prove this for $\phi \in \mathcal{D}(\Omega)$. By the definition of the weak derivative and the Cauchy–Schwarz inequality, for $\phi \in \mathcal{D}(\Omega)$,

$$\begin{aligned}\left|\int_\Omega (\Delta u - \Delta u_t)\psi\phi\,\mathrm{d}\mathbf{x}\right| &= \left|\int_\Omega (u-u_t)\Delta(\psi\phi)\,\mathrm{d}\mathbf{x}\right| \\ &= \left|\int_{\operatorname{supp}\psi} (u-u_t)\Delta(\psi\phi)\,\mathrm{d}\mathbf{x}\right| \leq \|u-u_t\|_{L^2(\operatorname{supp}\psi)}\|\Delta(\psi\phi)\|_{L^2(\operatorname{supp}\psi)},\end{aligned}$$

which tends to zero as $t \to 0$.

Moving to the terms on the boundary, we have that the L^2-norm of the trace of $(u-u_t)\psi$ converges by the continuity of the trace operator:

$$\big\|\gamma\big((u-u_t)\psi\big)\big\|_{L^2(\partial\Omega)} \leq \big\|\gamma\big((u-u_t)\psi\big)\big\|_{H^{1/2}(\partial\Omega)} \leq \|(u-u_t)\psi\|_{H^1(\Omega)} \to 0. \tag{A.2}$$

To show that $\|\nabla_{\partial\Omega}\gamma((u-u_t)\psi)\|_{L^2(\partial\Omega)} \to 0$ we only need to show that $\|\psi\nabla_{\partial\Omega}(\gamma(u-u_t))\|_{L^2(\partial\Omega)} \to 0$, since the $\gamma(u-u_t)\nabla_{\partial\Omega}\psi$ term can be controlled using the mapping properties of the trace operator in a manner similar to that in (A.2).

In order to prove that $\|\psi\nabla_{\partial\Omega}(\gamma(u-u_t))\|_{L^2(\partial\Omega)} \to 0$, we only need to show that

$$\int_{\partial\Omega} \nabla_{\partial\Omega}(\gamma(u-u_t))\cdot\psi\boldsymbol{\phi}\,\mathrm{d}s \to 0 \quad \text{as } t \to 0, \quad \text{for all } \boldsymbol{\phi} \in L^2_t(\partial\Omega),$$

where $L^2_t(\partial\Omega) := \{\boldsymbol{\phi} \in (L^2(\partial\Omega))^d : \mathbf{n}\cdot\boldsymbol{\phi} = 0\}$. Let $\nabla^*_{\partial\Omega} : L^2_t(\partial\Omega) \to (H^1(\partial\Omega))^*$ denote the adjoint of $\nabla_{\partial\Omega} : H^1(\partial\Omega) \to L^2_t(\partial\Omega)$. There exists a dense subspace X_t of $L^2_t(\partial\Omega)$ such that $\nabla^*_{\partial\Omega}(X_t) \subset L^2(\partial\Omega)$ (for explicit constructions of X_t in 2- and 3-d see [17, §A.3, Page 278]). Using this fact, and noting that the range of integration can be changed to $\operatorname{supp}\psi \cap \partial\Omega$, we only need to show that

$$\int_{\operatorname{supp}\psi\cap\partial\Omega} \gamma(u-u_t)\cdot\,\nabla^*_{\partial\Omega}(\psi\boldsymbol{\phi})\,\mathrm{d}s \to 0 \quad \text{as } t \to 0, \quad \text{for all } \boldsymbol{\phi} \in X_t.$$

However, this integral is bounded by $\|\gamma(u-u_t)\|_{L^2(\operatorname{supp}\psi\cap\partial\Omega)}\|\nabla^*_{\partial\Omega}(\psi\boldsymbol{\phi})\|_{L^2(\operatorname{supp}\psi\cap\partial\Omega)}$, which tends to zero as $t\to 0$ using arguments identical to those used in (A.2).

The last term to control is $\|(\partial((u-u_t)\psi)/\partial n\|_{L^2(\partial\Omega)}$. The regularity result of Nečas [63, §5.1.2], [49, Theorem 4.24 (i)] implies that this term can be bounded by a sum of all the previous terms. Indeed, this result (with the differential operator equal to the Laplacian) applied to the function $(u-u_t)\psi$ on the domain $\Omega' := \Omega\cap B_R$, with $R>0$ chosen such that $\operatorname{supp}\psi\subset B_R$, implies that, for some $C>0$,

$$\begin{aligned}\left\|\frac{\partial}{\partial n}\big((u-u_t)\psi\big)\right\|_{L^2(\partial\Omega)} &= \left\|\frac{\partial}{\partial n}\big((u-u_t)\psi\big)\right\|_{L^2(\partial\Omega')}\\ &\le C\Big(\|(u-u_t)\psi\|_{H^1(\Omega')} + \big\|\Delta\big((u-u_t)\psi\big)\big\|_{L^2(\Omega')} + \big\|\gamma\big((u-u_t)\psi\big)\big\|_{H^1(\partial\Omega')}\Big),\end{aligned}$$

which tends to zero as $t\to 0$; thus the proof is complete. □

References

[1] X. Antoine and M. Darbas, *Numerical Methods for Acoustics Problems*, Saxe-Coburg Publications, 2013, ch. Integral Equations and Iterative Schemes for Acoustic Scattering Problems.

[2] I. M. Babuška and S. A. Sauter, *Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers?*, SIAM Review, (2000), pp. 451–484.

[3] A. H. Barnett and A. Hassell, *Boundary quasi-orthogonality and sharp inclusion bounds for large Dirichlet eigenvalues*, SIAM Journal on Numerical Analysis, 49 (2011), pp. 1046–1063.

[4] ——, *Fast computation of high frequency Dirichlet eigenmodes via the spectral flow of the interior Neumann-to-Dirichlet map*, Communications on Pure and Applied Mathematics, to appear (2013). doi: 10.1002/cpa.21458.

[5] N. Bartoli and F. Collino, *Integral equations via saddle point problem for 2d electromagnetic problems*, ESAIM: Mathematical Modelling and Numerical Analysis, 34 (2000), pp. 1023–1049.

[6] 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 (2006), pp. 772–778.

[7] L. Beirão da Veiga, F. Brezzi, A. Cangiani, G. Manzini, L. D. Marini, and A. Russo, *Basic principles of virtual element methods*, Math. Models Methods Appl. Sci., 23 (2013), pp. 199–214.

[8] T. Betcke and E. A. Spence, *Numerical estimation of coercivity constants for boundary integral operators in acoustic scattering*, SIAM Journal on Numerical Analysis, 49 (2011), pp. 1572–1601.

[9] P. B. Bochev and M. D. Gunzburger, *Least-squares finite element methods*, vol. 166, Springer Verlag, 2009.

[10] Y. Boubendir and C. Turc, *Wave-number estimates for regularized combined field boundary integral operators in acoustic scattering problems with Neumann boundary conditions*, IMA Journal of Numerical Analysis, (2013).

[11] D. Braess, *Finite elements*, Cambridge University Press, Cambridge, third ed., 2007. Theory, fast solvers, and applications in elasticity theory, Translated from the German by Larry L. Schumaker.

[12] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, vol. 15 of Texts in Applied Mathematics, Springer, 2000.

[13] F. Brezzi and L. D. Marini, *Virtual element methods for plate bending problems*, Comput. Methods Appl. Mech. Engrg., 253 (2013), pp. 455–462.

[14] A. Buffa and P. Monk, *Error estimates for the ultra weak variational formulation of the Helmholtz equation*, M2AN Math. Model. Numer. Anal., 42 (2008), pp. 925–940.

[15] O. Cessenat, *Application d'une nouvelle formulation variationnelle aux équations d'ondes harmoniques, Problèmes de Helmholtz 2D et de Maxwell 3D*, PhD thesis, Université Paris IX Dauphine, 1996.

[16] O. Cessenat and B. Després, *Application of an Ultra Weak Variational Formulation of elliptic PDEs to the two-dimensional Helmholtz problem*, SIAM Journal on Numerical Analysis, 35 (1998), pp. 255–299.

[17] 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 (2012), pp. 89–305.

[18] S. N. Chandler-Wilde and D. P. Hewett, *Frequency-explicit continuity and coercivity estimates for integral equation methods for acoustic screen problems*, In preparation, (2013).

[19] S. N. Chandler-Wilde and P. Monk, *Wave-number-explicit bounds in time-harmonic scattering*, SIAM Journal on Mathematical Analysis, 39 (2008), pp. 1428–1455.

[20] L. Chesnel and P. Ciarlet, *T-coercivity and continuous Galerkin methods: application to transmission problems with sign changing coefficients*, Numer. Math., 124 (2013), pp. 1–29.

[21] P. G. Ciarlet, *Basic error estimates for elliptic problems*, in Handbook of numerical analysis, Vol. II, Handb. Numer. Anal., II, North-Holland, Amsterdam, 1991, pp. 17–351.

[22] P. Ciarlet Jr., *T-coercivity: Application to the discretization of Helmholtz-like problems*, Comput. Math. Appl., 64 (2012), pp. 22–34.

[23] F. Collino and B. Després, *Integral equations via saddle point problems for time-harmonic maxwell's equations*, Journal of computational and applied mathematics, 150 (2003), pp. 157–192.

[24] 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 (2006), pp. 139–160.

[25] M. Dafermos and I. Rodnianski, *The red-shift effect and radiation decay on black hole spacetimes*, Communications on Pure and Applied Mathematics, 62 (2009), pp. 859–919.

[26] L. Demkowicz, J. Gopalakrishnan, I. Muga, and J. Zitelli, *Wavenumber explicit analysis for a DPG method for the multidimensional Helmholtz equation*, Comput. Methods Appl. Mech. Engrg., (2012), pp. 126–138.

[27] B. Després, *Fonctionnelle quadratique et equations integrales pour les problemes d'onde harmonique en domaine exterieur*, Modélisation mathématique et analyse numérique, 31 (1997), pp. 679–732.

[28] ——, *Quadratic functional and integral equations for harmonic wave equations*, in Mathematical and Numerical Aspects of Wave Propagation (Golden, CO), SIAM, Philadelphia, 1998, pp. 56–64.

[29] V. Domínguez, I. G. Graham, and V. P. Smyshlyaev, *A hybrid numerical-asymptotic boundary integral method for high-frequency acoustic scattering*, Numerische Mathematik, 106 (2007), pp. 471–510.

[30] S. C. Eisenstat, H. C. Elman, and M. H. Schultz, *Variational iterative methods for nonsymmetric systems of linear equations*, SIAM Journal on Numerical Analysis, (1983), pp. 345–357.

[31] H. C. Elman, *Iterative Methods for Sparse Nonsymmetric Systems of Linear Equations*, PhD thesis, Yale University, 1982.

[32] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*, Oxford University Press,, 2005.

[33] B. Engquist and L. Ying, *Fast algorithms for high frequency wave propagation*, in Numerical Analysis of Multiscale Problems, I. G. Graham, T. Y. Hou, O. Lakkis, and R. Scheichl, eds., vol. 83 of Lecture Notes in Computational Science and Engineering, Springer, 2012, pp. 127–161.

[34] Y. Erlangga, *Advances in iterative methods and preconditioners for the Helmholtz equation*, Archives of Computational Methods in Engineering, 15 (2008), pp. 37–66.

[35] O. G. Ernst and M. J. Gander, *Why it is difficult to solve Helmholtz problems with classical iterative methods*, in Numerical Analysis of Multiscale Problems, I. G. Graham, T. Y. Hou, O. Lakkis, and R. Scheichl, eds., vol. 83 of Lecture Notes in Computational Science and Engineering, Springer, 2012, pp. 325–363.

[36] L. C. Evans, *Partial differential equations*, American Mathematical Society Providence, RI, 1998.

[37] X. Feng and H. Wu, *Discontinuous Galerkin methods for the Helmholtz equation with large wave number*, SIAM Journal on Numerical Analysis, 47 (2009), pp. 2872–2896.

[38] ——, *hp-Discontinuous Galerkin methods for the Helmholtz equation with large wave number*, Mathematics of computation, 80 (2011), pp. 1997–2024.

[39] P. Grisvard, *Elliptic problems in nonsmooth domains*, vol. 24, Pitman, Boston, 1985.

[40] T. Ha-Duong, *On the boundary integral equations for the crack opening displacement of flat cracks*, Integral Equations and Operator Theory, 15 (1992), pp. 427–453.

[41] M. S. Hanna and K. T. Smith, *Some remarks on the Dirichlet problem in piecewise smooth domains*, Communications on Pure and Applied Mathematics, 20 (1967), pp. 575–593.

[42] A. Hassell and T. Tao, *Upper and lower bounds for normal derivatives of Dirichlet eigenfunctions*, Mathematical Research Letters, 9 (2002), pp. 289–305.

[43] U. Hetmaniuk, *Stability estimates for a class of Helmholtz problems*, Commun. Math. Sci, 5 (2007), pp. 665–678.

[44] R. Hiptmair, A. Moiola, and I. Perugia, *Plane wave discontinuous Galerkin methods for the 2D Helmholtz equation: analysis of the p-version*, SIAM J. Numer. Anal., 49 (2011), pp. 264–284.

[45] ——, *Stability results for the time-harmonic Maxwell equations with impedance boundary conditions*, Mathematical Models and Methods in Applied Sciences, 21 (2011), pp. 2263–2287.

[46] F. Ihlenburg, *Finite element analysis of acoustic scattering*, vol. 132, Springer Verlag, 1998.

[47] C. E. Kenig, *Harmonic analysis techniques for second order elliptic boundary value problems*, American Mathematical Society, 1994.

[48] B. Lee, T. A. Manteuffel, S. F. McCormick, and J. Ruge, *First-order system least-squares for the Helmholtz equation*, SIAM Journal on Scientific Computing, 21 (2000), pp. 1927–1949.

[49] W. C. H. McLean, *Strongly elliptic systems and boundary integral equations*, Cambridge University Press, 2000.

[50] J. M. Melenk, *On generalized finite element methods*, PhD thesis, The University of Maryland, 1995.

[51] J. M. Melenk and I. Babuška, *The partition of unity finite element method: Basic theory and applications*, Comput. Method Appl. M., 139 (1996), pp. 289–314.

[52] J. M. Melenk, A. Parsania, and S. Sauter, *General DG-methods for highly indefinite Helmholtz problems*, Journal of Scientific Computing, (2013), pp. 1–46.

[53] J. M. Melenk and S. Sauter, *Convergence analysis for finite element discretizations of the Helmholtz equation with Dirichlet-to-Neumann boundary conditions*, Math. Comp, 79 (2010), pp. 1871–1914.

[54] ——, *Wavenumber explicit convergence analysis for Galerkin discretizations of the Helmholtz equation*, SIAM J. Numer. Anal., 49 (2011), pp. 1210–1243.

[55] E. Mitidieri, *A Rellich type identity and applications*, Communications in Partial Differential Equations, 18 (1993), pp. 125–151.

[56] 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.

[57] P. Monk, *Finite element methods for Maxwell's equations*, Oxford University Press, 2003.

[58] 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 (1961), pp. 561–568.

[59] ——, *Time decay for the nonlinear Klein-Gordon equation*, Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, 306 (1968), pp. 291–296.

[60] ——, *Decay for solutions of the exterior problem for the wave equation*, Communications on Pure and Applied Mathematics, 28 (1975), pp. 229–264.

[61] 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 (1968), pp. 187–203.

[62] C. S. Morawetz, J. V. Ralston, and W. A. Strauss, *Decay of solutions of the wave equation outside nontrapping obstacles*, Communications on Pure and Applied Mathematics, 30 (1977), pp. 447–508.

[63] J. Nečas, *Les méthodes directes en théorie des équations elliptiques*, Masson, 1967.

[64] L. E. Payne, *Inequalities for eigenvalues of membranes and plates*, J. Rat. Mech. Anal, 4 (1955), p. 529.

[65] B. Perthame and L. Vega, *Morrey–Campanato estimates for Helmholtz equations*, Journal of functional analysis, 164 (1999), pp. 340–355.

[66] S. I. Pohozaev, *Eigenfunctions of the equation* $\Delta u + \lambda f(u) = 0$, Soviet Math. Dokl, 6 (1965), pp. 1408–1411.

[67] F. Rellich, *Darstellung der Eigenwerte von* $\Delta u + \lambda u = 0$ *durch ein Randintegral*, Mathematische Zeitschrift, 46 (1940), pp. 635–636.

[68] S. A. Sauter and C. Schwab, *Boundary Element Methods*, Springer-Verlag, Berlin, 2011.

[69] E. A. Spence, *Wavenumber-explicit bounds in time harmonic acoustic scattering*, Preprint, (2013).

[70] 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 (2011), pp. 1384–1415.

[71] E. A. Spence, I. V. Kamotski, and V. P. Smyshlyaev, *Coercivity of combined boundary integral equations in high frequency scattering*, Preprint, (2013).

[72] B. R. Vainberg, *On the short wave asymptotic behaviour of solutions of stationary problems and the asymptotic behaviour as $t \to \infty$ of solutions of non-stationary problems*, Russian Mathematical Surveys, 30 (1975), pp. 1–58.

[73] G. Verchota, *Layer potentials and regularity for the Dirichlet problem for Laplace's equation in Lipschitz domains*, Journal of Functional Analysis, 59 (1984), pp. 572–611.

[74] H. Wu, *Continuous interior penalty finite element methods for the Helmholtz equation with large wave number*, arXiv preprint, arXiv:1106.4079, (2011).