Chapter 2

PDE Theory of the Deterministic Helmholtz Equation and Theory of its Discretisation

2.1 Introduction

This chapter has two main foci:

1. Recapping theory for the deterministic Helmholtz equation in heterogeneous media, especially well-posedness results and a priori bounds on the solution, and
2. Recapping and extending theory of the finite-element method (FEM) for the deterministic Helmholtz equation, especially error bounds.

In section 2.2.1 we will define two Helmholtz problems, one on an infinite exterior domain, and the other on a truncated domain, and discuss their physical relevance. In section 2.2.2 we will recap in some detail the well-posedness results and a priori bounds from [34]; these results will be crucial for our analysis of stochastic Helmholtz problems in chapter 3. Then in section 2.2.3 we set these results in their wider context with a review of the literature on well-posedness results and a priori bounds for the Helmholtz equation. We will then move on to the FEM for (1.3); in section 2.3.1 we give the variational formulations of out two Helmholtz problems. In section 2.3.2 we recap basic concepts of the FEM, then in section 2.3.3 we will prove new error bounds for the FEM for the Helmholtz equation. Finally in section 2.3.4 we will give an overview of the literature on error bounds and quasi-optimality for the Helmholtz equation.

2.2 PDE Theory of the Deterministic Helmholtz Equation

We will begin by defining the two deterministic Helmholtz problems that we consider in this thesis; we will consider their stochastic counterparts in subsequent chapters.

2.2.1 Deterministic Helmholtz Problems

We first state the two Helmholtz problems of interest in 'strong form' (that is, where derivatives are understood as distributional derivatives). In this section we largely follow the presentation in [34]

We first establish notation for function spaces; let $L^\infty(D_+;\mathbb{R}^{d\times d})$ be the set of all matrix-valued functions $A : D_+ \to \mathbb{R}^{d\times d}$ such that $A_{i,j} \in L^\infty(D_+;\mathbb{R})$ for all $i,j = 1,\dots,d$. Where the range of functions is $\mathbb{C}$ we suppress the second argument in a function space, e.g. we write $L^2(D_+)$ for $L^2(D_+;\mathbb{C})$.

Problem 2.1 (Exterior Dirichlet Problem). *Let D_- be a bounded Lipschitz open set such that the open complement $D_+ := \mathbb{R}^d \setminus \overline{D_-}$ is connected and let $\Gamma_D := \partial D_-$. Let $\gamma_D \to$ denote the trace operator. Given*

- $k > 0$,
- $f \in L^2(D_+)$ *with compact support,*

- $g_D \in H^{1/2}(\Gamma_D)$,
- $n \in L^\infty(D_+;\mathbb{R})$ *such that* $1-n$ *has compact support and there exist* $0 < n_{\min} < n_{\max} < \infty$ *such that*

$$n_{\min} \leq n(\mathbf{x}) < n_{\max} \text{ for almost every } \mathbf{x} \in D_+,$$

- $A \in L^\infty(D_+;\mathbb{R}^{d\times d})$ *such that* $I - A$ *has compact support,* A *is symmetric, and there exist* $0 < A_{\min} < A_{\max} < \infty$ *such that*

$$A_{\min}|\boldsymbol{\xi}|^2 \leq (A(\mathbf{x})\boldsymbol{\xi}) \cdot \overline{\boldsymbol{\xi}} < A_{\max}|\boldsymbol{\xi}|^2 \text{ for all } \boldsymbol{\xi} \in \mathbb{C}^d \text{ for almost every } \mathbf{x} \in D_+,$$

we say $u \in H^1_{\text{loc}}(D_+)$ *satisfies the* exterior Dirichlet problem *if*

$$\nabla \cdot (A\nabla u) + k^2 n u = -f \text{ in } D_+, \tag{2.1}$$

$$\gamma_D u = g_D, \tag{2.2}$$

and u *satisfies the Sommerfeld radiation condition*

$$\frac{\partial u}{\partial r}(\mathbf{x}) - iku(\mathbf{x}) = o\left(\frac{1}{r^{(d-1)/2}}\right) \text{ as } r := |\mathbf{x}| \to \infty, \text{ uniformly in } \hat{\mathbf{x}} := \mathbf{x}/|\mathbf{x}|. \tag{2.3}$$

A physical interpretation of Problem 2.1 is that u is the acoustic pressure field caused by the scattering of an incoming wave u_i from the scatterer D_-.[1] The Dirichlet boundary condition (2.2) then means D_- corresponds to a sound-soft scatterer, that is, one on which the total field $u + u_i$ vanishes, with $g_D = -\gamma_D u_i$. The function f represents domain-based source terms from the incident wave; if there are no other pressure sources in the domain, and $\Delta u_i + k^2 u_i = 0$, then $f = \nabla \cdot ((A - I)\nabla u_i) + k^2(n-1)u_i$.

The Sommerfeld radiation condition (2.3) ensures that the solutions of Problem 2.1 correspond to physically 'outgoing' waves (see, e.g., [38, Section 1.1.3]), and also guarantees the uniqueness of solution to Problem 2.1, see, e.g., [22, p. 16]. Observe that Problem 2.1 is defined on an infinite spatial domain; if one discretises Problem 2.1 using domain-based methods (such as FEMs) the infiniteness of the domain causes an issue[2]. Therefore a common appproach is to truncated Problem 2.1 with an artificial boundary that is sufficiently large to contain D_- and all the inhomogeneities in A, n, and f. Options for the truncated boundary condition include a perfectly matched layer, first introduced in [8] for Maxwell's equations, which mimics the whole of the external domain, or FEM-BEM coupling (a numerical method, where BEM stands for boundary-element method), as in, e.g., [37], where a boundary element method is used to approximate the solution in the exterior of the truncated domain. However, in this thesis, we will use a simpler approach, imposing an *impedance boundary condition*

$$\partial_\nu u - iku = g_I \tag{2.4}$$

on the truncated boundary. If $g_I = 0$, then (2.4) can be seen as a first-order approximation to (2.3). Truncating with an impedance boundary condition gives rise to the following deterministic Helmholtz problem

Problem 2.2 (Truncated Exterior Dirichlet Problem)**.** *Let* D_- *be a bounded Lipschitz open set such that the open complement* $D_+ := \mathbb{R}^d \setminus \overline{D_-}$ *is connected. Let* $\widetilde{D}$ *be a bounded connected Lipschitz open set such that* $\overline{D_-} \subset\subset \widetilde{D}$. *Let* $D := \widetilde{D} \setminus \overline{D_-}$, $\Gamma_D := \partial D_-$, *and* $\Gamma_I := \partial\widetilde{D}$. *Let* $\gamma_I \to$ *denote the trace operator. Given*

- $k > 0$,
- $f \in L^2(D)$
- $g_D \in H^{1/2}(\Gamma_D)$,
- $g_I \in L^2(\Gamma_I)$
- $n \in L^\infty(D;\mathbb{R})$ *such that* $\operatorname{supp}(1-n) \subset\subset D$ *and there exist* $0 < n_{\min} < n_{\max} < \infty$ *such that*

$$n_{\min} \leq n(\mathbf{x}) < n_{\max} \text{ for almost every } \mathbf{x} \in D,$$

- $A \in L^\infty(D;\mathbb{R}^{d\times d})$ *such that* $\operatorname{supp}(I-A) \subset\subset D$,[8] *$A$ is symmetric, and there exist* $0 < A_{\min} < A_{\max} < \infty$ *such that* EdN:8

$$A_{\min}|\boldsymbol{\xi}|^2 \leq (A(\mathbf{x})\boldsymbol{\xi})\cdot\overline{\boldsymbol{\xi}} < A_{\max}|\boldsymbol{\xi}|^2 \text{ for all } \boldsymbol{\xi}\in\mathbb{C}^d \text{ for almost every } \mathbf{x}\in D,$$

we say $u \in H^1(D)$ satisfies the truncated exterior Dirichlet problem *if*

$$\nabla\cdot(A\nabla u) + k^2 n u = -f \text{ in } D,$$

$$\gamma_D u = g_D, \text{ and}$$

$$\gamma_I \partial_\nu u - ik\gamma_I u = g_I. \tag{2.5}$$

is only an approximation to

Observe that, by construction, $\partial D = \Gamma_I \cup \Gamma_I$ and $\Gamma_D \cap \Gamma_I = \emptyset$.

Whilst the impedance boundary condition (2.5) ~~does not exactly mimic~~ the Sommerfeld radiation condition (2.3), the solutions of Problem 2.2 are still 'wave-like', and we will see below that solutions of Problem 2.2 posses many of the same properties as solutions of Problem 2.1. We also note that a common Helmholtz model problem in the numerical-analysis community is the *interior impedance problem*, which is simply Problem 2.2 in the case $D_- = \emptyset$.

what is "this"? (what exactly is independent of k?)

2.2.2 Well-posedness and a priori bounds

We will now recap the well-posedness results and a priori bounds for Problems 2.1 and 2.2 from [34]; these results will be crucial for proving well-posedness results and a priori bounds for the stochastic analogues of Problems 2.1 and 2.2 in chapter 3. The novelty of the results in [34] is that they hold independently of k, and the a priori bounds we prove are explicit in A, n and k; this explicitness is necessary in order to prove similar a priori bounds for stochastic A and n. We prove these results under conditions on A and n that are, in some sense 'nontrapping'. Informally, a medium is 'nontrapping' if all rays travelling through the medium escape in a uniform time; this definition, and the sense in which our conditions are 'nontrapping', is discussed in section 2.2.3 below.

We first define the classes of A and n for which we will prove well-posedness results and a priori bounds.

require A and n to be

Definition 2.3 (Class of nontrapping media). *Let $A \in C^{0,1}(\overline{D_+};\mathbb{R}^{d\times d})$, $n \in C^{0,1}(\overline{D_+};\mathbb{R})$, and $\mu_1,\mu_2 > 0$. We say that $A \in \mathrm{NT}_{\mathrm{Mat},D_+}(\mu_1)$ if*

$$A(\mathbf{x}) - (\mathbf{x}\cdot\nabla)A(\mathbf{x}) \geq \mu_1$$

in the sense of quadratic forms for almost every $\mathbf{x}\in D_+$. We say that $n \in \mathrm{NT}_{\mathrm{scal},D_+}(\mu_2)$ if

$$n(\mathbf{x}) + \mathbf{x}\cdot\nabla n(\mathbf{x}) \geq \mu_2$$

for almost every $\mathbf{x}\in D_+$.

If D is as in Problem 2.2, then we define $\mathrm{NT}_{\mathrm{Mat},D}(\mu_1)$ and $\mathrm{NT}_{\mathrm{scal},D}(\mu_2)$ analagously.

Remark 2.4 (Definition 2.3 is well-defined). *Observe that the conditions in definition 2.3 are well-defined, as A and n have weak first-order derivatives. For bounded Lipschitz open sets D, $C^{0,1}(D) = W^{1,\infty}(D)$ (see, e.g., [28, Section 4.2.3, Theorem 5]). As $I-A$ and $1-n$ are both compactly supported (from the definition of Problem 2.1) and Lipschitz (from definition 2.3), the set $D = (\operatorname{supp}(I-A)\cup\operatorname{supp}(1-n))_{+1}$ is a bounded Lipshitz open set, where $S_{+1} = \{\mathbf{x}+\boldsymbol{\varepsilon}\in\mathbb{R}^d : \mathbf{x}\in S \text{ and } \boldsymbol{\varepsilon}\in\mathbb{R}^d \text{ with } \|\boldsymbol{\varepsilon}\|<1\}$. Therefore $C^{0,1}(D) = W^{1,\infty}(D)$, and hence $A \in W^{1,\infty}(D;\mathbb{R}^{d\times d})$ and $n \in W^{1,\infty}(D;\mathbb{R})$. Outside D, $A = I$ and $n = 1$, and thus $A \in W^{1,\infty}(\mathbb{R}^d\setminus D;\mathbb{R}^{d\times d})$ and $n \in W^{1,\infty}(\mathbb{R}^d\setminus D;\mathbb{R})$. As $A = I$ and $n = 1$ on the boundary ∂D, it follows that $A \in W^{1,\infty}(\overline{D_+};\mathbb{R}^{d\times d})$ and $n \in W^{1,\infty}(\overline{D_+};\mathbb{R})$.*

[9] EdN:9

Our well-posedness results will require the scatterer D_- to be star-shaped, and our statement of the truncated problem will require the truncation domain to be star-shaped with respect to a ball. We now recall these definitions.

[1] In the literature the scattered field is sometimes denotes u_s, in which case u denotes the total field $u_i + u_s$.

[2] Although, if one was able to compute the Dirichlet-to-Neumann operator for the exterior Dirichlet problem for the *homogeneous* Helmholtz equation, then one could discretise Problem 2.1 exactly. See Problem 2.9 for the variational formulation of Problem 2.1, which is posed on a finite domain and includes the Dirichlet-to-Neumann operator. In practice, the Dirichlet-to-Neumann operator is not computable

[8] EdNote: Euan—in hetero, this requirement is instead phrased 'dist$(\operatorname{supp}(I-A),\Gamma_I) > 0$'. Is there any reason for phrasing it that way?

[9] EdNote: Both—Can you think of better notation for these conditions? I'm not overly keen to write NT_{A} etc., as there is potentially confusion between then function A and the roman A is the subscript. Or am I being overcautious?

expensive to compute and so is approximated.

I think what you have is fine

no reason - your way is better!

Definition 2.5 (Star-shaped, star-shaped with respect to a ball). *We say that D_- is star-shaped with respect to the point $\mathbf{x}_0$ if for all $\mathbf{x} \in D_-$, the line segment $[\mathbf{x}_0, \mathbf{x}] \in D_-$.*

We say that D_- is star-shaped with respect to the ball B if D_- is star shaped with respect to $\mathbf{x}_0$, for all $\mathbf{x}_0 \in B$.

We can now state well-posedness results and a priori bounds for the Helmholtz equation in the class of heterogeneous media we have just defined. We denote the ball of radius R about the point $\mathbf{x}_0$ by $B_R(\mathbf{x}_0)$. We denote $B_R(\mathbf{0})$ by B_R.

Theorem 2.6 (Well-posedness and bound for the EDP). *If D_-, A, n, and f satisfy the requirements in Problem 2.1, $g_D = 0$, D_- is star-shaped with respect to the origin, and there exists $\mu_1, \mu_2 > 0$ such that $A \in \mathrm{NT}_{\mathrm{Mat},D_+}(\mu_1)$ and $n \in \mathrm{NT}_{\mathrm{scal},D_+}(\mu_2)$ then the solution of Problem 2.1 exists and is unique. Furthermore, given $R > 0$ such that $\mathrm{supp}(I - A)$, $\mathrm{supp}(1 - n)$, and $\mathrm{supp}\, f$ are compactly contained in $D_R, = D \cap B_R$, then*

$$\mu_1 \|\nabla u\|^2_{L^2(D_R)} + \mu_2 k^2 \|u\|^2_{L^2(D_R)} \leq C_1 \|f\|^2_{L^2(D_R)},$$

for all $k > 0$, where

$$C_1 := 4\left(\frac{R^2}{\mu_1} + \frac{1}{\mu_2}\left(R + \frac{d-1}{2k}\right)^2\right).$$

For the proof of theorem 2.6, see [34, Theorem 2.5].

The following result is the analogue of theorem 2.6 for the solution of Problem 2.2. However, the statement is slightly more complicated than the statement of theorem 2.6 due to the presence of the impedance boundary Γ_I, and its effect on the solution.

Theorem 2.7 (Well-posedness and bound for the TEDP). *If D_-, A, n, f, and g_I satisfy the requirements in Problem 2.2, $g_D = 0$, D_- is star-shaped with respect to the origin, $\widetilde{D}$, is star-shaped with respect to a ball, and there exists $\mu_1, \mu_2 > 0$ such that $A \in \mathrm{NT}_{\mathrm{Mat},D}(\mu_1)$ and $n \in \mathrm{NT}_{\mathrm{scal},D}(\mu_2)$, then the solution of Problem 2.2 exists and is unique. Let:*

- $L_I := \max_{\mathbf{x} \in \Gamma_I} |\mathbf{x}|$ *and*
- aL_I *be the radius of the ball with respect to which $\widetilde{D}$ is star-shaped.*

Then

$$\mu_1 \|u\|^2_{L^2(D)} + \mu_2 k^2 \|\nabla u\|^2_{L^2(D)} + aL_I \|\nabla_{\Gamma_I} \gamma_I u\|^2_{L^2(\Gamma_I)} + 2L_I k^2 \|\gamma_I u\|^2_{L^2(\Gamma_I)} \leq C_2 \|f\|^2_{L^2(D_R)} + \widetilde{C}_2 \|g_I\|^2_{L^2(\Gamma_I)}$$

for all $k > 0$, where ∇_{Γ_I} is the surface gradient on Γ_I,

$$C_2 := 4\left(\frac{L_I^2}{\mu_1} + \frac{1}{\mu_2}\left(\beta + \frac{d-1}{2k}\right)^2\right),$$

$$\widetilde{C}_2 := 2\left(2\left(1 + \frac{2}{a}\right) + \frac{\beta}{L_I} + \frac{(d-1)^2}{4}\right)L_I,$$

and

$$\beta := L_I\left(2 + \frac{1}{(kL_I)^2} + 2\left(1 + \frac{2}{a}\right)\right).$$

For the proof of theorem 2.7, see [34, Theorem A.6 (i)].

Observe that the above results are stated only in the case that $g_D = 0$. Whilst there is no mathematical difficulty in proving analagous results in the case $g_D \neq 0$, the calculations in this case are more involved, as one must consider the surface gradient on the Dirichlet boundary, and this surface gradient depends on A. In the case $A = I$, these calculations are significantly simplified, and so in the case $A = I$ and $g_D \neq 0$ analagous results to theorems 2.6 and 2.7 are proved in [34, Theorem 2.19(ii)] (for Problem 2.1) and [34, Theorem A.6(iv)] (for Problem 2.2).

We highlight that theorems 2.6 and 2.7 and the similar results in [34] are significant for the following two reasons.

1. These are the first A, n, and k-explicit bounds on the solution of the Helmholtz equation in the case where both A and n are heterogeneous. As will discussed in more detail in section 2.2.3 below, previous results were either not A, n, and k-explicit, or did not have A *and* n varying. The k-explicitness of these results is crucial for understanding how the solution of the Helmholtz equation (and numerical methods for its approximation) behave for large k, and the A-and-n-explicitness is crucial for proving bounds on the stochastic Helmholtz equation, as in chapter 3.

2. These are the first bounds explicit in A and n where the bound *and* the restrictions on A and n are independent of k. Previous results in the literature only proved such bounds by imposing conditions on A and n that became *more stringent* as $k \to \infty$; again, this literature will be more fully discussed in section 2.2.3 below.

Remark 2.8 (Extensions of theorems 2.6 and 2.7). *Theorems 2.6 and 2.7 are extended to wider classes of heterogeneous A and n and to the case $g_D \neq 0$ in [34]. As stated above, the case $g_D \neq 0$ (with $A = I$) is treated in [34, Theorem 2.19(ii)] (for Problem 2.1) and [34, Theorem A.6(iv)] (for Problem 2.2), and the case $n = 1$ is covered in [34, Theorem 2.19(i)] (for Problem 2.1) and [34, Theorem A.6(ii)]. We highlight that when $A = I$ or $n = 1$ the condition on the non-constant coefficient can be slightly weakened from those in definition 2.3. When A and n are discontinuous, [34, Condition 2.6] gives analogues of the conditions in definition 2.3, and then the result corresponding to theorem 2.6 is proved in [34, Theorem 2.7]. Letting A and n be L^∞-perturbations of nontrapping media is discussed in [34, Remark 2.15], and relaxing the Lipschitz assumption on Γ_D is outlined in [34, Remark 2.13], with the caveat that when Γ_D is non-Lipschitz, we instead formulate Problem 2.1 as a variational problem, which is discussed in section 2.3.1 below. The above extensions and generalisations can all be applied to Problem 2.2, as mentioned in [34, p. 2916].*

2.2.3 Discussion of results on well-posedness and a priori bounds for the Helmholtz equation

We will now review the historical development of well-posedness results, and a priori bounds for the Helmholtz equation.

Well-posedness results

By 'well-posedness', we mean that a solution of the Helmholtz equation exists, is unique, and continuously depends on the data f and g_D (and g_I for Problem 2.2).

We note that proving well-posedness results and a priori bounds for the Helmholtz equation is much more involved than proving such results for the stationary diffusion equation

$$\nabla \cdot (A \nabla u) = -f \text{ in } D. \tag{2.6}$$

In (2.6) (where A is heterogeneous but not stochastic), if A is bounded above and bounded away from zero, then the associated bilinear form is bounded and coercive. Then the Lax–Milgram Theorem applies, and one immediately obtains well-posedness and an a priori bound (in $H^1(D)$) that is explicit in A.

However, for (2.1), the situation is much more subtle (as will be discussed in more detail below). Even if A and n are bounded above and bounded away from zero, in general one cannot prove a bound

$$\|u\|_{1,k} \leq C(\|f\| + \|g_D\|), \tag{2.7}$$

where C depends explicitly on A, n, and the wavenumber k. The sesquilinear form a associated with the standard variational formulation of the Helmholtz equation is not coercive; however, a does satisfy a Gårding inequality

$$\Re a(v, v) + k^2(A_{\min} + n_{\max})\|v\|^2_{L^2(D)} \geq A_{\min}\|v\|^2_{1,k}, \tag{2.8}$$

where $\|v\|_{1,k} := \left(\|\nabla v\|_{L^2(D)} + k^2\|v\|_{L^2(D)}\right)^{1/2}$ is the weighted H^1 norm used frequently when studying Helmholtz problems[3] The Gårding inequality means $\Re a(v, v)$ is 'coercive' if an appropriate multiple of the L^2-norm is added.

If the solution of the Helmholtz equation is unique, then existence and an a priori bound on the solution follow from Fredholm Theory (see, e.g. [64, Theorems 5.10 and 5.18]; Fredholm theory can be applied because the sesquilinear form satisfies a Gårding inequality.) Therefore the challenge of proving well-posedness reduces to proving uniqueness. However, we note that the a priori bound one obtains using Fredholm theory is *not* explicit in k, A or n.

For homogeneous problems (with $A = I$ and $n = 1$) uniqueness follows from the Sommerfeld radiation condition; for heterogeneous problems, the Unique Continuation Principle (UCP) gives uniqueness, under some additional smoothness assumptions on A and n (see, e.g., [34, p. 2871] for a discussion of the UCP and [35, Section 2] for the application of the UCP to show uniqueness for heterogeneous Helmholtz problems). Therefore, as well-posedness

[3] The norm $\|\cdot\|_{1,k}$ is used because solutions of the Helmholtz equation typically are of order $1/k$ and have first-order derivatives of order k; therefore the norm $\|\cdot\|_{1,k}$ should contain terms of roughly the same size.

results for the Helmholtz equation are essentially well-understood[4], we now turn our attention to a priori bounds on the solution that are explicit in k, A, and n.

k-, A-, and n-explicit a priori bounds

All these bounds we now discuss will, unless otherwise stated, be for the weighted H^1 norm $\|\cdot\|_{1,k}$. We will only consider the case where the scatterer D_- is compact, and the inhomogeneities in A and n are compactly supported, as in Problem 2.1. Research into so-called rough surface scattering, where either D_- or the inhomogeneities in A and n are not compactly supported, is itself a rich area of research (see, e.g., the literature reviews in [65]), but this area is not the concern of this thesis.

Techniques for proving a priori bounds There are two main classes of techniques for proving a priori bounds on the Helmholtz equation in inhomogeneous media. The first uses techniques from semiclassical analysis (a branch of microlocal analysis), and studies the behaviour of rays through the medium. In order for this approach to be used D_-, A, and n must all be smooth, so that rays and the notion of reflections from the scatterer D_- can be defined (the notion of a reflection is not well-defined if the scatterer has a corner).[10] EdN:10

When one uses microlocal analysis tools to prove a k-independent bound, one first shows that the problem is nontrapping[5]. Once one has proved the problem is nontrapping, one uses further technical tools[6] to conclude a k-independent bound.

The second class of techniques is multiplier techniques, where the PDE (2.1) is multiplied by carefully chosen multiples of, e.g., u and $\mathbf{x}\cdot\nabla u$, and the resulting expression is then integrated by parts and rearranged. Whilst conceptually simpler than semiclassical analysis tools, multiplier methods allow one to prove bounds in situations that are inaccesible to semiclassical analysis, e.g., when the scatterer or coefficients are not smooth. Multiplier methods were first used for wave problems by Morawetz in the 60s for studying energy decay for the wave equation. See [33] for an overview of this, and other aspects of Morawetz's work and [67, Theorem 1.1] for the connection between energy decay for the wave equation and a priori bounds on the Helmholtz equation[7]. However, multiplier techniques typically require more severe restrictions on the geometry of the scatterer (and truncation boundary, in the case of Problem 2.2) than semiclassical analysis techniques[8].

Review of a priori bounds We now summarise the state of the field regarding bounds (2.7), and especially the dependence of the constant C on k, A, n, and D_-. Some of these results are proved in the context of energy decay results for the time-domain wave equation; for simplicity's sake, we do not distinguish in our comments between these results, and those proving bounds (2.7) directly. This section borrows heavily from the literature reviews in [18, Section 1.1] and [34, Sections 1 and 2.4].

We note that trapping behaviour can be caused either by the scatterer D_- (for example, if D_- contains a cavity in which rays can be 'trapped', see, e.g.,, or by the medium, defined by A and n. In the case where A and n are discontinuous, these discontinuities can cause rays to be trapped in a manner analagous to the concept of total internal reflection.

Typically, results in the literature either assume the medium is constant and focus on the behaviour of the wave induced by the scatterer, or assume there is no scatterer (or a 'nice' scatterer that does not cause trapping behaviour, such as a convex scatterer) and consider the behaviour induced by inhomogeneties in the medium; this review will focus only on scattering induced by inhomogeneities in the medium, as this is the focus of the results presented

[4] Observe that if one can prove an a-priori bound of the form (2.7), then one can conclude uniqueness (as the solution of the Helmholtz equation with zero data must therefore be the zero function). Therefore, if one can prove such a priori bounds *without* the restrictions on A and n needed to apply the UCP, one can conclude uniqueness (and well-posedness, as outlined above) in a wider class of media; see [34, pp. 2873, 2883] for more details on how the results in [34] can be used in this way.

[10] EdNote: Euan—I want to add some chat here about when these methods were first introduced, but I'm struggling to figure it out? Is it with Lax & Phillips?

[5] The problem is *nontrapping* if, for any bounded set $S \subseteq D$ there exists a time $t(S)$ such that any ray starting in S and evolving according to the laws of geometrical optics leaves S by time $t(S)$. The rigorous definition is more technical; see [34, Section 6] for an overview. The problem is called *trapping* if it is not nontrapping. We will see below that one can prove k-independent bounds even when rays cannot be defined, typically when D_-, A, and n are not smooth. In such situations, one usually uses the mutiplier techniques discussed below. In an abuse of terminology, we call all situations where a k-independent bound holds 'nontrapping'.

[6] Either: (i) Vainberg's paramatrix argument from [66] and Melrose and Sjöstrand's results on propagation of singularities [46], (ii) Lax–Phillips theory [?] and propagation of singularities, or (iii) Burq's defect-measure argument [14].

[7] As is common in the analysis literature, many papers, such as [67] consider 'cut-off resolvent' bounds, where the resolvent $(\Delta+k^2)^{-1}$ is multiplied by a cut-off function to deal with an infinite domain. Nevertheless, these results are in the same spirit as the a priori bound we discuss below.

[8] One can choose more complicated multipliers to mitigate some of these restrictions, as in [51], but most of the works we discuss below do not.

above (where the scatterer is assume star-shaped) and is also the focus of the corresponding stochastic results in chapter 3, where the medium is stochastic, and not the boundary of the scatterer. We recall that the best behaviour, obtained when the medium is nontrapping, is that the constant C in (2.7) is independent of k (possibly for some $k \geq k_0$). For an overview of results around obstacle scattering, where C can grow logarithmically, polynomially, or exponentially in k depending on the scatterer, we refer the reader to the recent literature reviews in [18, Sections 1.1 and 1.3].

A priori bounds for Problem 2.1 In the worst case, the constant C can depend exponentially on k; i.e.,[11] EdN:11

$$C = C_1 \exp(kC_2), \tag{2.9}$$

for some constants C_1 and C_2 depending on D_-, A, and n. This bound was proved when $n = 1$ for general $A \in C^\infty$ and a general obstacle D_- by Burq [13], using Carleman estimates. The bound (2.9) was shown to be sharp by, e.g., Betcke, Chandler-Wilde, Graham, Langdon, and Lindner in [9, Equation 2.22], who, for Problem 2.1 with constant media and a scatterer whose boundary contains a certain part of an ellipse, construct a sequence of wavenumbers k_m (with corresponding solutions u_m and right-hand sides f_m of (2.1)) such that

$$k\|u_m\|_{L^2(D_+)} \gtrsim \exp(\gamma k_m)\|f_m\|_{L^2(D_+)},$$

for some $\gamma > 0$.

We now review results in the rest of the literature. The worst-case bound (2.9) was proved for $D_- = \emptyset$, $A = I$, and n jumping downwards across a C^∞ interface by Popov and Vodev [56] for A and n piecewise constant, jumping across a common C^∞ interface by Bellasoued [7]; and for $D_- = \emptyset$, $A = I$, and n a Lipschitz perturbation of 1 by Shapiro [62]. Carleman estimates were used in [7], and semiclassical analysis techniques in [56, 62][12]. [13] EdN:12

Capdeboscq, Leadbetter, and Parker [15, 16] studied the transmission problem (A and n jumping across a common interface) for a sphere and consider arbitrary-order Sobolev norms of the solution on spheres outside the transmission boundary. Using an expansion of the solution in terms of Bessel functions they showed that the constant C can grow super-algebraically in k when n jumps downwards, but is bounded when n jump upwards. They also show that, even in the case of n jumping downwards, if the jump is sufficiently small, k is sufficiently small, or the norm is measured sufficiently far away from the origin (where all the definitions of 'sufficient' depend on the size of the jump and the size of the sphere), then C is bounded independently of k. EdN:13

Other results with C bounded independently of k have been obtained for the transmission problem where n jumps up across a strictly convex, smooth boundary by Cardoso, Popov, and Vodev [17, 55] using semiclassical analysis techniques, and for the transmission problem with both A and n jumping by Moiola and Spence [47] using multiplier techniques under assumptions on the jumps in A and n. Also, theorem 2.6 and its related extensions discussed in remark 2.8 prove C is bounded independently of k in a variety of situations. We mention in passing the work of Morawetz and colloborators [48, 50, 49, 51][14] who proved k-independent bounds for a variety of scatterers in homogeneous media. Whilst these results are not concerned with the Helmholtz equation in heterogeneous media, they pioneered the use of the multiplier techniques used to prove theorems 2.6 and 2.7. EdN:14

A priori bounds for Problem 2.2 We now shift our attention to bounds for Problem 2.2 (or, more commonly, the *interior impedance problem* (IIP), that is, Problem 2.2 with no scatterer) with a 'nice' truncation boundary Γ_I—typically star-shaped with respect to a ball or similar. These bounds have mostly been proved by the numerical analysis community, due to Problem 2.2's usage as a model problem for numerical methods (the truncation boundary means that the problem is posed on a finite domain, enabling domain-discretisation methods, such as finite elements, to be used). The vast majority of these results use, in essence, the multiplier techniques introduced by Morawetz.

We briefly recap results for homogeneous media in chronological order: Melenk [43] and Cummings and Feng [23] proved that C is k-independent for the IIP in a domain that is either convex or star-shaped and smooth, and this result was generalised to a mixture of impedance, Dirichlet, and Neumann boundary conditions by Hetmaniuk [36]. Esterhazy and Melenk [27], loosened the restriction on the domain to only being Lipschitz, but could only prove bounds with C increasing at the rate $k^{5/2}$. (The work [27] does not use multiplier techniques, but rather uses properties of the Green's function for the Helmholtz equation from [44].) The rate of growth of C for a general

[11] EdNote: Euan, is [13, Theorem 2] saying that the pole-free region is exponentially (as $k \to \infty$) close to the real axis, and in this region, the bound is exponential? My French isn't that great....

[12] EdNote: Euan—is this distinction between Carleman estimates and semiclassical analysis a good one? Or are Carleman estimates part of the semiclassical/microlocal analysis toolbox?

[13] EdNote: Euan—can you briefly recap for me how one moves from nontrapping/results on location of resonances to an a priori bound?

[14] EdNote: Euan, I'm having a hard time seeing what's different in the first three of these references

(Lipschitz) domain was improved to order k by Spence [63] (using multiplier techniques) and finally eliminated for a general C^∞ domain (using microlocal analysis) by Baskin, Spence, and Wunsch [4]. We note briefly that recently Li and Wu [42] proved a k-independent bound for Problem 2.1 truncated with a perfectly matched layer, using properties of Bessel functions.

We now move on to bounds for the TEDP (or IIP) in heterogeneous media; all of the following bounds were proved using multiplier methods, unless otherwise stated. A bound with C independent of k was proved by Feng, Lin, and Lorton [31] for random media, under the k-dependent assumption that $A = I$ and $n = 1 + \eta$, with η a random field and $\|\eta\|_{L^\infty(D)} \lesssim 1/k$ almost surely. Brown, Gallistl, and Peterseim proved bounds with k-independent C in [12], under conditions related to, but more restrictive than, those in [34]. Barucq, Chaumont-Frelet and Gout [3] proved a k-independent bound for 2-D piecewise-constant media, under the condition that n increases as one moves away from the origin[15]. Ohlberger and Verfürth [54] studied the case where $n = 1$, A is scalar-valued (i.e. $A = \varepsilon I$), and the heterogeneity ε is given by many small inclusions, motivated by the analysis of multiscale methods for the Helmholtz equation. They proved a bound with C proportional to k^3 in this case. Graham and Sauter [35] took a very similar approach to [34], proving bounds for heterogeneous media when $A = I$ under conditions on n that are analogous to those in [34]. EdN:15

In related results, for the 1-dimensional Helmholtz equation in heterogeneous media, Chaumont-Frelet [19, Section 2.1.5, Theorem 3] used multiplier methods with specially-chosen test functions to prove a k-independent bound for piecewise constant media, under assumptions on the media that limit the number of 'pieces'. In contrast, Sauter and Torres [57] used properties of the 1-dimensional Green's function to prove a bound for the 1-dimensional Helmholtz equation with arbitrarily many 'pieces', but with C that grows algebraically in k.

Bounds explicit in the parameters Bounds on the heterogeneous Helmholtz equation that are explicit in all parameters of interest (such as k, A, and n) are crucial for proving k-explicit bounds on the corresponding stochastic Helmholtz equation; such bounds on the stochastic Helmholtz equation are the subject of chapter 3. We observe in passing that the only works that have bounds explicit in all the parameters of interest are those of Moiola and Spence [47]; Galkowski, Spence, and Wunsch [32]; and Graham, the author of this thesis, and Spence [34]. We briefly highlight that the work of Galkowski, Spence, and Wunsch uses semiclassical analysis techniques to show, for nontrapping media, that C depends on the length of the longest ray in the medium. However, calculating this length for a given D_-, A, and n is far from straightforward. In contrast, in [47, 34] C depends on A and n in easily-calculatable ways (e.g., via $\max_{\mathbf{x}\in D_+} n(\mathbf{x})$).

Dependence of trapping behaviour on k Observe that many of the results above that prove a k-independent bound are of the form 'place restrictions on D_-, A, and n to ensure nontrapping, then conclude result'. However, the recent works [47, 41] have shown that even in the worst case of exponential growth in k of the constant C, this behaviour is realised at very few frequencies; [47] provided numerical evidence for the transmission problem through a sphere that showed the realisation of super-algebraic growth is very sensitive to the value of k (in some cases, changing k at the 13th significant figure changed the behaviour completely. More rigorously, [41] used microlocal analysis techniques to show that one can exclude a set of k of arbitrarily small measure, and then obtain merely algebraic growth in C.

2.3 Theory of the Discretisation of the Helmholtz Equation

We will now shift our attention to the numerical analysis of the Helmholtz equation in heterogeneous media; we will study the finite-element method for the Helmholtz equation. We first provide the variational formulations of the Helmholtz equation, define the finite element method, and recall results on the apporximation properties of finite-element spaces, before proving our main result, a new error bound for the finite-element method for the Helmholtz equation in heterogeneous media.

2.3.1 Variational Formulations for the Helmholtz equation

The finite element method is based on the variational formulation of the Helmholtz equation; for simplicity of exposition, we state the variational formulation of Problems 2.1 and 2.2 in the case $g_D = 0$, although these can be generalised to the case $g_D \neq 0$.

[15] EdNote: Ivan/Euan—have I interpreted their condition correctly? This makes sense from the pictures in the article, but one can only reconcile these pictures with [3, Equation (2)] if one takes $\mathbf{n}_r$ to be the normal vector pointing *into* Ω_r, which seems unconventional....

Problem 2.9 (Variational formulation of EDP when $g_D = 0$). *Let D_+, A, n, and f be as in Problem 2.1. Choose $R > 0$ such that* $\operatorname{supp} f$, $\operatorname{supp}(I - A)$, $\operatorname{supp}(1 - n) \subset\subset B_R$, *and define $D_R := D_+ \cap B_R$.*

We say $u \in H^1_{0,\mathrm{D}}(D_R)$ satisfies the variational formulation of the exterior Dirichlet problem *with $g_D = 0$ if*

$$a(u, v) = F(v) \text{ for all } v \in H^1_{0,\mathrm{D}}(D_R),$$

where

$$a(w, v) := \int_{D_R} \left(\langle A \nabla w, \nabla v \rangle_{\mathbb{R}^d} - k^2 n w \overline{v} \right) - \langle T_R \gamma_R w, \gamma_R v \rangle_{\Gamma_R}$$

and

$$F(v) := \int_{D_R} f \overline{v},$$

where $T_R : H^{1/2}(\Gamma_R) \to H^{-1/2}(\Gamma_R)$ is the Dirichlet-to-Neumann map for the homogeneous Helmholtz equation $\Delta u + k^2 u = 0$ combined with the Sommerfeld radiation condition in the exterior of B_R; and $\langle \cdot, \cdot \rangle_{\Gamma_R}$ is the duality pairing on Γ_R.

Lemma 2.10 (Equivalence of formulations for the EDP). *Problems 2.1 and 2.9 are equivalent. If $u \in H^1_{\mathrm{loc}}(D_+)$ solves Problem 2.1 then $u|_{D_R} \in H^1_{0,\mathrm{D}}(D_R)$ and $u|_{D_R}$ solves Problem 2.9 (for R as in Problem 2.9). Conversely, if $u \in H^1_{0,\mathrm{D}}(D_R)$ solves Problem 2.9, then extending u to $H^1_{\mathrm{loc}}(D_+)$ by the solution of the exterior Dirichlet problem for the homogeneous Helmholtz equation with the Sommerfeld radiation condition in the exterior of D_R (with Dirichlet data γu on ∂B_R), u solves Problem 2.1.*

For a proof of lemma 2.10, see [34, Lemma 3.3].

Problem 2.11 (Variational formulation of TEDP when $g_D = 0$). *Let D, A, n, f, and g_I be as in Problem 2.2. We say $u \in H^1_{0,\mathrm{D}}(D_R)$ satisfies the* variational formulation of the truncated exterior Dirichlet problem *with $g_D = 0$ if*

$$a_T(u, v) = F_T(v) \text{ for all } v \in H^1_{0,\mathrm{D}}(D_R),$$

where

$$a_T(w, v) := \int_{D_R} \left(\langle A \nabla w, \nabla v \rangle_{\mathbb{R}^d} - k^2 n w \overline{v} \right) - ik \int_{\Gamma_I} \gamma_I w \gamma_I \overline{v}$$

and

$$F_T(v) := \int_{D_R} f \overline{v} + \int_{\Gamma_I} g_I \gamma_I \overline{v}$$

Lemma 2.12 (Equivalence of formulations for the TEDP). *Problems 2.2 and 2.11 are equivalent, i.e., $u \in H^1_{0,\mathrm{D}}(D_R)$ solves Problem 2.2 if, and only if, u solves Problem 2.11.*

For a proof of lemma 2.12, see [34, Lemma A.7].

2.3.2 Finite-element theory

We now give a brief summary of elementary concepts in finite-element theory. We focus only on those concepts that we be needed to prove the new error bound for finite-elements discretisations of the Helmholtz equation in section 2.3.3 below.

A crucial quantity in our analysis of finite-element methods will be the mesh size of a triangulation of the domain. (Note that in this thesis we use the words 'mesh' and 'triangulation' interchangably; strictly speaking, the term 'triangulation' only makes sense in 2-d, but we ignore this technicality.

Definition 2.13 (Triangulation). *A triangulation of a polygonal domain D is a finite collection of sets $K_i \subseteq \overline{D}$ such that*

1. $\mathring{K}_i \cap \mathring{K}_j = \emptyset$ *for* $i \neq j$,
2. $\bigcup_i K_i = \overline{D}$, *and*
3. *Something about triangles that works in 3-D too.*

Definition 2.14 (Mesh Size). *Given a triangulation $\mathcal{T} = \{K_i\}$ of a domain D, the* mesh size *of $\mathcal{T}$ is defined to be*

$$h = \max_{K_i} \operatorname{diam} K_i.$$

We can now define finite-element spaces associated with a triangulation; we first define the space of polynomials on a set.

Definition 2.15 (Set of polynomials). *For $K \subseteq \mathbb{R}^d$, $\mathcal{P}_p(K)$ is the set of polynomials defined on K with total degree at most p.*

Definition 2.16 (Finite element space of degree p). *Given a triangulation $\{K_i\}$ with mesh size h of a domain D, the* (continuous) piecewise-polynomial finite-element space of degree p *associated with $\{K_i\}$ is*

$$V_{h,p} \coloneqq \left\{ v_h : D \to \mathbb{C} : v_h \in C^0(D) \text{ and } v_h\big|_{K_i} \in \mathcal{P}_p(\overline{K_i}) \text{ for all } K_i \right\}.$$

Throughout this thesis, we only consider the h-finite-element method, i.e., the degree p of the polynomials associated with the space is assumed fixed, and we consider refining h. This is in contrast to the p-FEM, where h is fixed and p is increased, and the hp-FEM, where both h and p are increased according to some rule. hp-FEM methods for the Helmholtz equation were analysed by Melenk and Sauter in [44, 45][16], where they showed one can obtain an exponential rate of convergence with respect to the number of degrees of freedom. EdN:16

The following lemma shows how the approximation properties of the space $V_{h,p}$ depend on h and p:

Lemma 2.17 (Existence of quasi-interpolant). *Let $p \geq 1$ and let $v \in H^m(D)$. Then there exists $C > 0$ independent of v and $\widetilde{v}_{h,p} \in V_{h,p}$ such that for all $s \leq \min\{m, p+1\}$*

$$\|v - \widetilde{v}_{h,p}\|_{L^2(D)} \leq C h^s \|v\|_{H^s(D)}$$

and

$$\|v - \widetilde{v}_{h,p}\|_{H^1(D)} \leq C h^{s-1} \|v\|_{H^s(D)}.$$

For a proof of lemma 2.17 see, e.g., [11, Corollary 4.4.24, Remark 4.4.27].

Remark 2.18 (The function $\widetilde{v}_{h,p}$). *The function $\widetilde{v}_{h,p}$ in lemma 2.17 is constructed using 'averaged Taylor polynomials', see, e.g., [60],[11, Section 4.4] for details*[9].

With the concept of a finite-element space in place, we can now define the finite-element approximation to the variational problems Problems 2.9 and 2.11.

Problem 2.19 (Finite-element approximation of Problem 2.11). *We say that $u_h \in V_{h,p}$ is the finite-element approximation of u (the solution to Problem 2.11) if*

$$a(u_h, v_h) = L(v_h) \text{ for all } v_h \in V_{h,p}.$$

The finite-element approximation of Problem 2.9 is defined analagously.

Remark 2.20 (Not considering variational crimes). *Observe that Problem 2.19 requires $V_{h,p} \subset H^1_{0,\mathrm{D}}(D_R)$. This inclusion is true if D_R can be triangulated; otherwise, we must modify the definition of Problem 2.19 and commit a variational crime by approximating the boundary of D_R by a polygon, or introducing mesh elements with curved boundaries using interpolated boundary conditions or isoparametric finite elements. See, e.g., [11, Chapter 10] for an overview of the additional errors introduced by variational crimes (although not in the context of the Helmholtz equation). In this thesis we will ignore such variational crimes, and the additional errors they induce; such analysis is standard, ~~and orthogonal to the work in this thesis~~.*

[16] EdNote: I couldn't immediately see in these results that one obtains exponential convergence with respect to the number of DOFs—have I missed something, or is this result contained somewhere else?

[9] Observe that in [11], the authors use different notation to us; they use m to denote the regularity of v and p to denote the integrability of v, i.e., in [11], $v \in W^{m,p}(D)$.

2.3.3 Error bound for the heterogeneous IIP

We now move on to present new work—bounds on the error $u_h - u$ between the finite-element approximation and the true solution of Problem 2.11. These bounds, proven for the Helmholtz equation in *heterogeneous* media are generalisations of results already in the literature that the finite-element error is bounded (as $k \to \infty$) provided $h \lesssim k^{-(2p+1)/2p}$. These results will be crucial for our analysis of the multi-level Monte Carlo method for the Helmholtz equation in chapter 5.

The proof of these results uses a so-called 'elliptic projection' technique, where the variational formulation of a PDE related to Problem 2.11 is used as part of the proof. We will only prove results for Problem 2.19, and not for the finite-element approximation of Problem 2.9, as our proof uses properties of the related PDE and these properties have only been proven with impedance boundary conditions (in the recent preprint [21]) and not with an exact Dirichlet-to-Neumann map on the truncation boundary. However, we imagine the results proven for the related PDE with an impedance boundary condition also hold for an exact Dirichlet-to-Neumann boundary condition, and so we anticipate that the results we prove here also hold for finite-element approximations of Problem 2.9.

2.3.4 Discussion of FEM for the Helmholtz equation

We now discuss error bounds for finite-element methods for the Helmholtz equation. We will give some intuition behind these bounds, provide a brief history of their development, and briefly constrast them with error bounds for the stationary diffusion equation.

Intuition for fixed number of points per wavelength Recall from section 1.1.2 that if one takes the mesh size in the finite element method $h \sim 1/k$ (for first-order finite elements), then the interpolation (or best approximation) error is bounded uniformly in k. This restriction is motivated by observing that solutions of the Helmholtz equation typically have $\|u\|_{H^2(D)} \sim k$, and so one can bound the H^1-norm of the interpolation error if $h \sim 1/k$ using lemma 2.17. As explained in section 1.1.2, this restriction ensures there are a fixed number of discretisation points per wavelength of the solution. An alternative motivation for taking $h \sim 1/k$ (for first-order elements) is the Nyquist–Shannon sampling theorem[10] (see, e.g., [2, §5.21]) that states that any function v (in 1-d) whose Fourier transform lies inside $[-\lambda, \lambda]$ (for some $\lambda > 0$) is completely determined (via its Fourier series) by the point values $v(0)$, $v(\pm\mu)$, $v(\pm 2\mu)$, ..., for any $\mu < 1/(2\lambda)$.

Since the solution of the one-dimensional Helmholtz equation with constant coefficients is

$$u(x) = A\sin(kx) + B\cos(kx), \tag{2.10}$$

(for some constants A and B), which has Fourier Transform

$$\hat{u}(\xi) = \frac{A}{2i}\left(\delta\left(\xi - \frac{k}{2\pi}\right) - \delta\left(\xi + \frac{k}{2\pi}\right)\right) + \frac{B}{2}\left(\delta\left(\xi - \frac{k}{2\pi}\right) + \delta\left(\xi + \frac{k}{2\pi}\right)\right).$$

Therefore, if u is sampled at regularly-spaced points that are strictly closer together than π/k (as $\lambda = k/2\pi$ here), then u can be reconstructed perfectly from these samples. In conclusion, one expects the be able to interpolate the solution of the Helmholtz equation with uniform error in k if $h \sim 1/k$.

However, as was stated in section 1.1.2, the finite-element method for the Helmholtz equation suffers from pollution; and $h \sim 1/k$ is not sufficient to keep the finite-element error bounded as $k \to \infty$. Therefore we will now provide an overview of previous work giving mesh conditions under which the finite-element error $\|u - u_h\|_{1.k}$ is bounded as $k \to \infty$, as well as mesh conditions under which the finite-element method is quasi-optimal. It would be more natural to consider the relative error $\|u - u_h\|_{1.k}/\|u\|_{1.k}$; and to investigate the conditions on h which enable the relative error to be bounded as $k \to \infty$. However, current technology only allows us to prove results for the error, not the relative error, and so we investigate the error instead.

Previous results on the finite-element error Bayliss, Goldstein, and Turkel [5, §3] performed computations for $d = 2$ and first-order finite elements with D a square with Neumann boundary conditions on three sides and an impedance-like boundary condition on the other side. The results of the computations suggested the mesh condition $h \lesssim k^{-3/2}$ is sufficient to bound the relative error in the L^2 norm uniformly in k. Ihlenberg and Babuska [39, Theorem 5, §3.4] proved that the error in the H^1 seminorm for a 1-dimensional problem on an interval with a zero Dirichlet boundary condition at one end, an impedance boundary condition at the other end and a uniform

[10] Proved by Shannon in his seminal paper in information theory [61, Theorem 1]

mesh is bounded independently of k provided $h \lesssim k^{-3/2}$, and then concluded that the relative error is similarly bounded. However, the proof in [39] used the fact that the solution of the Helmholtz equation in 1-d is given by (2.10) and so has not been generalised to higher dimensions.

Subsequent work on proving error bounds for the Helmholtz equation focused on first proving quasi-optimality for the finite-element method, and then concluding bounds on the error; we discuss results on quasi-optimality below. However, recent efforts on proving error bounds have used so-called elliptic projection ideas; these ideas are at the heart of our results in section 2.3.5 below.

This idea of using elliptic projections was introduced to prove error bounds for discontinuous Galerkin methods for the TEDP by Feng and Wu [29, 30]. Wu [68] then used elliptic projctions to prove that the error in the first-order finite-element method for the IIP is of the order h^2k^3 if $h \sim k^{-3/2}$ (and therefore the error is bounded); in [68] this result was shown for a discontinuous-Galerkin first-order method for the Helmholtz equation, and the results for the standard finite-element method were obtained as a special case. Du and Wu [24] extended the results in [68] to higher-order finite-elements for the IIP, showing that the error $\sim h^{2p}k^{2p+1}$. Chaumont-Frelet and Nicaise [20] obtained similar results for first-order finite elements for the TEDP when the scatterer induces corner singularities (and therefore [20] includes additional contraints on the mesh arising from the corner singularities that we do not mention here). Wu and Zou [69, Lemma 3.3] obtained the first results for heterogeneous media; they showed the error is bounded if $h \lesssim k^{-3/2}$ for first-order finite elements for the IIP. The results in [69] we obtained for a special class of heterogeneous media as part of an argument proving similar results for a nonlinear Helmholtz equation.

Observe that the results proved above for higher-order finite elements (the error is bounded if $h^{2p}k^{2p+1}$ is bounded) become less stringent as p increases ($h^{2p}k^{2p+1}$ is bounded if $h \sim k^{-1-1/2p}$). In this section we present similar results for general classes of heterogeneous media; that is, we show that the error in the finite-element method is bounded if $h \lesssim k^{-1-1/2p}$ and the underlying PDe is nontrapping.

Previous results on quasi-optimality As mentioned above, one way to prove error bounds for the Helmholtz equation is to first prove the finite-element method is quasi-optimal, and then conclude that the error is bounded. However, as we will show below, the mesh conditions required for the finite-element method to be quasi-optimal are *more* restrictive than those required for the error to be bounded. Recall that the finite-element method is quasi-optimal if there exists $C > 0$, independent of h such that

$$\|u - u_h\| \leq C \inf_{v_h \in V_{h,p}} \|u - v_h\|;$$

i.e., up to a constant, u_h is the best approximation to u in the space $V_{h,p}$ in the norm $\|\cdot\|$.

We first contrast the Helmholtz equation with the stationary diffusion equation (2.6). For the stationary diffusion equation, one immediately obtains quasi-optimality for *any* mesh by Céa's Lemma (and then obtains that the relative error is bounded by properties of particular members of the finite-element space as in, e.g., lemma 2.17. We emphasise again that this result holds for any mesh, with no restriction on h.

However, proving quasi-optimality for the Helmholtz equation is more tricky; Céa's Lemma relies on the coercivity of the sesquilinear form, and the sesquilinear forms arising from standard discretisations of the Helmholtz equation are not coercive for large k. Therefore, to prove quasi-optimality, one instead uses the Aubin–Nitsche duality argument. It was first introduced by Aubin [1] and Nitsche [52] for coercive problems, and applied to problems satisfying a Gårding inequality by Schatz [58].

The Aubin–Nitsche arument was first used for Helmholtz problems by Melenk in his PhD thesis [43, Proposition 8.2.7], he proved that the finite-element method for the Helmholtz equation is quasi-optimal under the very restrictive mesh condition[11] $h \lesssim k^{-2}$. Graham and Sauter [35, Remark 4.4 b.] and Galkowski, Spence, and Wunsch [32, Theorem 3] have obtained analogous conditions for general finite-element spaces for the IIP [35] and first-order finite elements for the EDP [32] respectively.

In conclusion, the error for the finite-element method for Helmholtz problems is bounded if $h \lesssim k^{-1-1/p}$, and the finite-element method is quasi-optimal if $h \lesssim k^{-2}$ (for first-order finite elements). We will now prove analagous results on the boundedness of the error for heterogeneous Helmholtz problems.

2.3.5 New error bounds for the Helmholtz equation in heterogeneous media

In this section, we prove that the finite-element approximation of the solution to the Helmholtz TEDP exists if $h \lesssim k^{-3/2}$. Moreover, we give an expression for the hidden constant that is completely explicit in A and n, and we

[11] Observe that for large values of k, the mesh condition $h \lesssim k^{-2}$ is prohibitive—it would result in linear systems of size, e.g., 10^{12}, for the Helmholtz equation with $k = 100$ in 3-D.