Acta Numerica (2026), pp. 1–171
doi:10.1017/S0962492925100275

Numerical analysis of the high-frequency Helmholtz equation using semiclassical analysis

Jeffrey Galkowski
Department of Mathematics, University College London,
London, WC1H 0AY, UK
E-mail: j.galkowski@ucl.ac.uk

Euan A. Spence
Department of Mathematics, University of Bath,
Bath, BA2 7AY, UK
E-mail: e.a.spence@bath.ac.uk

We consider the numerical solution of high-frequency scattering problems modelled by the Helmholtz equation with a bounded obstacle. Although the analysis of this problem dates back at least 50 years, over the past decade or so, tools and techniques from *semiclassical analysis* have provided a new perspective and been used to settle several long-standing open problems in this area. Semiclassical analysis works in phase space (i.e. position and frequency) and describes rigorously the extent to which solutions of high-frequency PDEs are dictated by the properties of the corresponding geometric optics rays.

The goals of the article are (i) to give a introduction to semiclassical analysis aimed at non-experts and (ii) to showcase some of the numerical analysis results about finite element methods, boundary element methods and domain decomposition methods obtained using semiclassical techniques.

2020 Mathematics Subject Classification: 35J05, 65N30, 65N38, 65N55

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

CONTENTS

1 Introduction: numerical solution of the Helmholtz equation 2
2 Summary of the contents of this article 4
3 Informal statement of the main results 10
4 Definition of the scattering problem and its approximation by a radial PML 24
5 A primer on semiclassical analysis 30
6 Preasymptotic error estimates for the h-FEM (**R1**) 49
7 The hp-FEM does not suffer from the pollution effect (**R2**) 67
8 Pollution and preasymptotic error estimates for the h-BEM (**R3**) 78
9 h-FEM with non-uniform meshes defined by ray-dynamics (**R4**) 100
10 The parallel overlapping Schwarz method with PMLs on the subdomain boundaries (**R5**) 111
A h- and p-explicit polynomial approximation 121
B Definition of the standard boundary integral operators 156
C Precise statement of Informal Theorem 5.5 158
References 160

1. Introduction: numerical solution of the Helmholtz equation

The Helmholtz equation

$$(-\Delta - k^2)u = 0 \tag{1.1}$$

with wavenumber $k > 0$ is the simplest model of time-harmonic wave propagation. Indeed, if $u(x)$ satisfies (1.1), then

$$U(x,t) = u(x)e^{-ikt} \tag{1.2}$$

is a time-harmonic solution of the wave equation (with speed one)

$$\frac{\partial^2 U}{\partial t^2} - \Delta U = 0. \tag{1.3}$$

This article is focused on the Helmholtz equation posed outside some compact scatterer (e.g. a penetrable or impenetrable obstacle, or a combination of the two) with data corresponding to an incident wave. (For precise definitions of these boundary value problems, see Section 4.) We are specifically interested in the case when the wavenumber $k \gg 1$; because k dictates the frequency of the waves in the time domain via (1.2), this limit is known as the *high-frequency limit*.

The main computational goal considered in this article. Given a specified accuracy and a fixed $k \gg 1$, compute solutions of the Helmholtz equation (1.1) (and its variable-coefficient generalization) to that specified accuracy.

We highlight immediately that the $k \to \infty$ limit of a Helmholtz solution is governed by the behaviour of the geometric optics rays (which are straight lines for

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

constant wave speed). Furthermore, there is a large body of research into methods that directly compute the high-frequency limit of the Helmholtz equation – methods based on ray-tracing and/or solving the Eikonal and transport equations; see e.g. the *Acta Numerica* article by Engquist and Runborg (2003). The accuracy of these methods – by design – increases with k, but for any fixed k the error is bounded below; that is, it cannot be decreased arbitrarily. In addition, because asymptotic methods rely on a well-chosen asymptotic ansatz for the solution, these methods are intractable in complicated geometries.

Therefore, while for Helmholtz problems with extremely large k and simple geometries, asymptotic methods give solutions accurate enough for some practical applications, there is a large class of Helmholtz problems arising from applications where achieving the goal above is necessary.

To compute an approximation to the solution of the Helmholtz equation, broadly speaking one has to make the following three choices.

Choice 1: Formulation. When the wave speed is (piecewise) constant, one can reduce solving the PDE (1.1) to solving an integral equation on the boundary of the scatterer. When the wave speed is variable, one usually truncates the domain exterior to the scatterer – with some form of local *absorbing boundary condition* (Engquist and Majda 1977*a*,*b*) or *perfectly matched layer* (PML) (Berenger 1994) to minimize reflections of the waves from this artificial boundary – and consider some variational formulation of (1.1) posed in this truncated domain.

Choice 2: Finite-dimensional approximation space. Both the finite element method (FEM) and the boundary element method (BEM) (i.e. solving boundary integral equations using the FEM) usually seek the solution as a piecewise polynomial: a mesh is imposed on the domain, and the solution is represented on each mesh element as a polynomial. Alternatively, one can try to tailor the space to the wave nature of the solution; a classic choice here is to use plane waves on each element, i.e. $e^{ikx\cdot a_j}$ for some directions a_j, with such methods that use problem-adapted basis functions called *Trefftz methods*. We note that Choices 1 and 2 cannot always be made independently; for example, since continuity across element interfaces cannot be imposed directly for many Trefftz basis functions (such as piecewise plane waves), they cannot be used in the standard Helmholtz variational formulation; see the discussion of variational formulations suitable for Trefftz methods in Hiptmair, Moiola and Perugia (2016, §2).

Choice 3: How to solve the linear system. Choices 1 and 2 result in a linear system $A\mathbf{x} = \mathbf{b}$, where $\mathbf{x} \in \mathbb{C}^N$ contains the coefficients of the approximation to the solution in the space of Choice 2. This linear system is large, indefinite, non-self-adjoint and generally non-normal; solving such a system is a very challenging problem in numerical linear algebra. The linear system is large because accurately approximating an arbitrary function oscillating with frequency $\lesssim k$ in d dimensions requires $\sim k^d$ degrees of freedom (DoF) (in other words, the space of functions

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

oscillating at frequency $\lesssim k$ in d dimensions has dimension $\sim k^d$). Furthermore, the *pollution effect* means that when the polynomial degree is fixed, having k^d DoF in the FEM is *not* enough to maintain accuracy as $k \to \infty$; we discuss this in much more detail below. For the BEM, the linear systems require at least $\sim k^{d-1}$ DoF (due to the dimension reduction of working on the boundary), but are dense, as opposed to the sparse FEM matrices. The linear system is non-self-adjoint because of the condition that the scattered wave propagates away from the scatterer (i.e. the radiation condition). This non-self-adjointness means that the conjugate gradient method cannot be used to solve the system, and the method of choice is the generalized minimum residual method (GMRES) (Saad and Schultz 1986). The linear system of the standard variational formulation is indefinite (roughly speaking, has eigenvalues with both positive and negative real parts), and iterative solvers famously struggle with this feature (see e.g. Ernst and Gander 2012). On top of all these difficulties, any ill-conditioning at the continuous level (e.g. due to trapped waves) is inherited by the discrete linear system; see e.g. Marchand, Galkowski, Spence and Spence (2022) for an investigation of this in the context of BEM. Because the linear system is difficult to solve, we seek a *preconditioner*, i.e. a matrix B^{-1} whose action is relatively cheap to compute such that $B^{-1}A\mathbf{x} = B^{-1}\mathbf{b}$ is easier to solve than $A\mathbf{x} = \mathbf{b}$. Popular choices of preconditioners are those based on *domain decomposition* (DD), where an approximate inverse for A is obtained by dividing the computational domain into subdomains and solving Helmholtz problems on these subdomains.

2. Summary of the contents of this article

2.1. The results

This paper contains five main results.

R1 Upper bounds on the error for the h-version of the FEM (i.e. with piecewise polynomials of fixed degree p) showing how the k-dependence of the mesh-width h dictates the accuracy as $k \to \infty$, extending the results in Galkowski and Spence (2025). (See Informal Theorem 3.1.) These bounds are observed empirically to be optimal when the mesh is uniform.

R2 Upper bounds on the error for the hp-version of the FEM showing how the k-dependence of h and p dictate the accuracy as $k \to \infty$, extending the results in Lafontaine, Spence and Wunsch (2022*b*) and Galkowski, Lafontaine, Spence and Wunsch (2023*b*, 2024*c*). In particular, these estimates show that k-uniform accuracy can be maintained with $\sim k^d$ degrees of freedom; that is, the hp-FEM *does not suffer from the pollution effect*. (See Informal Theorem 3.2.) As a component of this analysis, Appendix A gives a self-contained treatment of p- and h-explicit polynomial approximation with novel aspects including precise Sobolev space estimates for polynomial approximation of high-regularity functions.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

R3 Upper and lower bounds on the error for the h-version of the BEM showing how h must depend on k to maintain accuracy as $k \to \infty$ (Galkowski and Spence 2023, Galkowski, Rachh and Spence 2025). In particular, these prove that, in many situations, $\sim k^{d-1}$ degrees of freedom is *not* enough to maintain accuracy as $k \to \infty$, that is, the h-BEM *suffers from the pollution effect*. (See Informal Theorems 3.3 and 3.5.)

R4 Non-uniform h-FEM meshes for scattering problems with PML truncation, where the meshes are defined by ray dynamics (Averseng, Galkowski and Spence 2025). These meshes violate the conditions on h given by **R1**, and involve coarser meshes away from trapping and in the PML while still maintaining accuracy as $k \to \infty$. (See Table 3.1.) The meshes are designed using bounds on the Galerkin error showing how varying the meshwidth in one region affects errors both in that region and elsewhere.

R5 Analysis of overlapping Schwarz methods for the Helmholtz problem, truncated by a PML, where the subdomains have an absorbing layer at their boundaries, such as a PML, and the number of iterations in the convergence results is specified in terms of the interaction between the ray dynamics and the geometry of the subdomains (Galkowski *et al.* 2024*a*). (See Informal Theorem 3.7.)

These results are stated informally (i.e. with minimum technicalities) in Section 3, along with a discussion of the history and context of each result. Later in the article we give the precise statements, along with complete proofs of **R1**, **R2** and the upper bounds in **R3**, and sketch proofs of the lower bounds in **R3** and the results **R4** and **R5**.

Therefore, in terms of the three choices in the previous subsection, the results of this paper are mainly concerned with the standard variational formulation posed in a truncated domain exterior to the obstacle, with piecewise polynomial approximation spaces, and (one-level) domain decomposition used as a solver.

2.2. *How these results were obtained: semiclassical analysis*

The unifying feature of the above five results is that they were obtained using *semiclassical analysis* (SCA). SCA is a branch of PDE theory that, in particular, describes rigorously the extent to which Helmholtz solutions in the $k \to \infty$ limit are governed by the rays. In our context, the adjective 'semiclassical' essentially means 'high-frequency', but comes from the fact that these techniques can be used to show how particle motion in classical mechanics arises from wave functions in quantum mechanics; see Zworski (2012, §1.2). SCA is part of the wider field of *microlocal analysis*, a generalization of Fourier analysis: Fourier analysis splits functions up into components oscillating at different frequencies, whereas microlocal analysis localizes in both frequency and space (to the extent allowed by the uncertainty principle). The combination of physical and Fourier space is then known as *phase space*.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Two key concepts from SCA used to obtain results **R1**–**R5** above are *ellipticity* and *propagation*. We now illustrate these in the constant-coefficient case using the Fourier transform. Given $u\colon \mathbb{R}^d \to \mathbb{C}$, let

$$\mathcal{F}_k(u)(\xi) := \int_{\mathbb{R}^d} e^{-ik\langle \xi, x\rangle} u(x)\, dx, \tag{2.1}$$

that is, $\mathcal{F}_k(u)$ is the Fourier transform of u, with frequency variable scaled by k (so that frequency k corresponds to $|\xi| = 1$). If u solves the Helmholtz equation (1.1), then

$$(|\xi|^2 - 1)\mathcal{F}_k(u)(\xi) = 0. \tag{2.2}$$

When $||\xi| - 1| \geq c > 0$ (i.e. at unscaled frequencies far from k), the symbol $|\xi|^2 - 1$ is invertible; thus the inverse of the Helmholtz operator acts like the inverse of an elliptic operator (e.g. the inverse of the Laplacian) on such frequencies. However, when $|\xi|$ is close to 1 (i.e. the unscaled frequency is close to k), SCA shows that the inverse of the Helmholtz operator with $k \gg 1$ encodes propagation along geometric optics rays.

We highlight that the results **R1** and **R2** – the analyses of the h-FEM and hp-FEM – *only use ellipticity at high frequencies*, that is, no information about propagation is required. The results **R3**, **R4** and **R5** then use information about *both* ellipticity *and* propagation.

The following four results are not included in this article, but also use SCA ideas, with the first using propagation and the last three using ellipticity.

- Sharp upper and lower bounds on the error incurred by truncating the unbounded domain exterior to an obstacle by a local absorbing boundary condition (Galkowski, Lafontaine and Spence 2024*b*), with these bounds showing that the relative error is bounded away from zero, independent of the frequency, and regardless of the geometry of the artificial boundary.
- Bounds showing that for general Helmholtz scattering problems (specifically, those in the so-called 'black-box scattering framework': Sjöstrand and Zworski 1991), the error incurred by truncation by a PML is exponentially small in frequency, the width of the PML and the strength of the scaling in the PML region (Galkowski, Lafontaine and Spence 2023*a*); these results are summarized in Section 4.5 below.
- New finite-dimensional approximation spaces for the high-frequency Helmholtz problems with smooth variable coefficients based on coherent states using $O(k^{d-1/2})$ degrees of freedom (Chaumont-Frelet, Dolean and Ingremeau 2024*a*).
- Convergence theory for two-level hybrid Schwarz domain decomposition preconditioners for high-frequency Helmholtz problems (Galkowski and Spence 2026).

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Remark 2.1 (Hybrid numerical-asymptotic methods). When one knows the asymptotic form of the solution as $k \to \infty$, up to amplitudes that have a k-independent rate of oscillation, one can build this information into the approximation space. These spaces allow for methods that require only $O(1)$ degrees of freedom as $k \to \infty$.

There is a long history of such methods, going back to the 1970s (see the *Acta Numerica* review article by Chandler-Wilde, Graham, Langdon and Spence 2012, and in particular their literature review in Section 1), and in recent years these methods have come to be known as *hybrid numerical-asymptotic methods*.

These approximation spaces have mostly been developed for boundary integral methods (because the dimension reduction makes the asymptotic form of solutions more tractable). Two situations where these methods can be rigorously analysed are scattering of highly structured incident fields (e.g. plane waves or point sources) by

(i) smooth, strictly convex obstacles, where the asymptotic form of the solution was obtained in Melrose and Taylor (1985) using microlocal methods, and

(ii) certain non-trapping polygons where the asymptotic form of the solution was obtained by the method of images together with representations of the solution near corners using separation of variables; see Chandler-Wilde and Langdon (2007, Theorems 3.2 and 3.3) and Chandler-Wilde, Hewett, Langdon and Twigger (2015, Theorem 3.6). The form of the diffracted wave can also be obtained from a microlocal approach; see Cheeger and Taylor (1982*a*,*b*).

While hybrid numerical-asymptotic methods therefore fall into the category of numerical methods/analyses that (at least morally) use semiclassical analysis, they require detailed *a priori* knowledge of both the data and the scatterer. In this article, we focus on methods/analyses using semiclassical ideas that are applicable without such detailed *a priori* knowledge (with the resulting methods then requiring more degrees of freedom than hybrid numerical-asymptotic methods).

2.3. *From Fourier multipliers to pseudodifferential operators*

In the previous section, we saw that the constant-coefficient Helmholtz operator is a Fourier multiplier, that is,

$$(-k^{-2}\Delta - 1)u(x) := \mathcal{F}_k^{-1}((|\bullet|^2 - 1)\mathcal{F}_k(u)(\bullet))(x).$$

In motivating the use of pseudodifferential operators to study more general Helmholtz operators, it is instructive to first consider the PDE

$$(-k^{-2}\Delta + 1)v = f \quad \text{on } \mathbb{R}^d \tag{2.3}$$

(sometimes called the *modified Helmholtz equation*), which becomes

$$(|\xi|^2 + 1)\mathcal{F}_k(v)(\xi) = \mathcal{F}_k(f)(\xi)$$

after taking the (k-weighted) Fourier transform (2.1). Since $|\xi|^2 + 1 \geq 1 > 0$ for all ξ, both the partial differential operator $(-k^{-2}\Delta + 1)$ and its inverse $(-k^{-2}\Delta + 1)^{-1}$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

are Fourier multipliers, and the solution v to (2.3) can be written in terms of the multiplier via

$$v(x) = \mathcal{F}_k^{-1}((|\bullet|^2 + 1)^{-1}\mathcal{F}_k(f)(\bullet))(x). \tag{2.4}$$

One can understand how the Fourier transform arrives at (2.4) in two ways.

(i) Since the partial differential operator $(-k^{-2}\Delta + 1)$ is translation-invariant (and hence is a convolution operator), the Fourier transform diagonalizes $(-k^{-2}\Delta + 1)$ and thus gives the full description of how $(-k^{-2}\Delta + 1)$ and its inverse act in a (generalized) basis of L^2.

(ii) The Fourier transform decomposes functions into its different frequencies, and $(-k^{-2}\Delta + 1)$ acts by multiplication by $\sigma(\xi) := |\xi|^2 + 1$ on each frequency. Thus, since $|\sigma(\xi)| \geq 1 > 0$, multiplication by $1/\sigma(\xi)$ gives the inverse operator.

For constant-coefficient PDEs on $\mathbb{R}^d$, points (i) and (ii) are actually the same phenomenon. However, their generalizations to more complicated operators, e.g. the variable-coefficient generalization of (2.3), namely

$$-k^{-2}\operatorname{div}(A\nabla v) + nv = f \quad \text{on } \mathbb{R}^d, \tag{2.5}$$

are different.

(i) The analogue of the first approach is to expand in eigenfunctions of a self-adjoint operator and use the functional calculus/spectral theorem to decompose the operator. For this to be effective, one needs an appropriate self-adjoint operator, and information about its eigenfunctions.

(ii) The analogue of the second approach is to use microlocal analysis, i.e. to decompose in both space and frequency and observe that the PDE (2.5) approximately acts by multiplication by $\sigma(x, \xi) := \langle A(x)\xi, \xi\rangle + n(x)$. Multiplication by $(\sigma(x, \xi))^{-1}$ is then a good approximation to the action of the inverse of the PDE. At a technical level, *semiclassical pseudodifferential operators* enable us to make this second approach work.

The result **R1** is, in fact, obtained using only approach (i), i.e. expanding in eigenfunctions of a self-adjoint operator, whilst still 'thinking semiclassically'. **R2** and **R4** intertwine approaches (i) and (ii), using the both expansions in eigenfunctions and the tools of semiclassical analysis, while **R3** and **R5** use only semiclassical tools.

The standard (i.e. non-semiclassical) calculus of pseudodifferential operators was introduced in the 1960s to obtain approximate inverses, modulo smooth errors, for elliptic variable-coefficient PDEs. The semiclassical pseudodifferential calculus then builds in the large parameter k to produce approximate inverses to operators such as (2.3) and (2.5) (i.e. *semiclassically* elliptic operators), modulo an error that is both smoothing *and* smaller than any algebraic power of k. Semiclassical

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

pseudodifferential operators take the following form: given a function $a(x,\xi)$ on phase space,

$$(\widetilde{\mathrm{Op}_k}(a)u)(x) := \frac{k^d}{(2\pi)^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{ik\langle x-y,\xi\rangle}a(x,\xi)u(y)\,dy\,d\xi \tag{2.6}$$

is a semiclassical pseudodifferential operator with symbol a (the reason for our use of the tilde is explained in Section 5.1). Observe that if $a(x,\xi)=a(\xi)$, then

$$(\widetilde{\mathrm{Op}_k}(a)v)(x) = \mathcal{F}_k^{-1}(a(\bullet)(\mathcal{F}_k v)(\bullet))(x),$$

that is, $\widetilde{\mathrm{Op}_k}(a)$ is a Fourier multiplier.

As discussed in Section 2.2 (after (2.2)), the Helmholtz operator is semiclassically elliptic only at high frequencies (i.e. at frequencies $> k$), and thus its inverse is *not* a semiclassical pseudodifferential operator. In this article, we obtain information about the inverse of the Helmholtz operator via propagation estimates (see Section 5.6), which describe how the phase-space structure of the Helmholtz solution is governed by the ray dynamics.

The discussion above described how results relying on semiclassical ellipticity (such as **R1** and **R2**) can sometimes be obtained without explicit use of SCA. Furthermore, for some special geometries and coefficients (e.g. star-shaped obstacles and coefficients satisfying certain monotonicity properties), bounds on the solution operator (which implicitly contain information about propagation) can also be obtained without explicit use of SCA; see the discussion in Section 5.6.1 and Remark 5.22. However, we know of no other way to obtain the results relying on propagation (and depending on ray dynamics) in **R3**, **R4** and **R5** without using the semiclassical machinery.

2.4. *The technical overhead required to use SCA, and how we deal with this in this article*

One of the main obstacles in using semiclassical/microlocal techniques is the technical overhead in stating and proving results about pseudodifferential operators. In this article we deal with this in the following ways.

Section 5 contains the results from SCA used in the article. While these results are stated rigorously, their proofs are sketched using informal statements of the fundamental properties of semiclassical pseudodifferential operators; see Informal Theorem 5.5 below.

The precise statement of Informal Theorem 5.5 is then given in an appendix (see Theorem C.1), referencing the complete treatments of SCA in Dimassi and Sjöstrand (1999), Zworski (2012) and Dyatlov and Zworski (2019, Appendix E). The sketch proofs can then be made rigorous by repeating the arguments with Informal Theorem 5.5 replaced by Theorem C.1 (although we do not do this here).

Our hope is that this approach provides an accessible introduction to the ideas behind semiclassical analysis and the 'phase-space point of view'. We emphasize that

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

it is this phase-space point of view that gives the conceptual/heuristic ideas underpinning the results in this article, with the technical details of the pseudodifferential calculus then used in the rigorous implementation of these ideas. Furthermore, even in rigorous proofs, it is often possible to use Theorem C.1 as a 'black box' without needing the details of how these fundamental properties of pseudodifferential operators are obtained.

3. Informal statement of the main results

The scattering problem is stated precisely in Section 4; we state it informally here as: given a bounded open set $\Omega_- \subset \mathbb{R}^d$, $d \geq 1$, with connected open complement Ω_+, variable coefficients A and n such that $A \equiv I$ and $n \equiv 1$ outside a compact set, and a compactly supported source f, find u satisfying the Sommerfeld radiation condition and such that

$$Pu := -k^{-2}\mathrm{div}(A\nabla u) - nu = f \text{ in } \Omega_+, \quad Bu|_{\partial\Omega_+} = 0, \tag{3.1}$$

where $B = I$ or $B = \partial_{\nu,A}$ (the conormal derivative).

We work in norms where each derivative is scaled by k^{-1}; that is, for an open set $\Omega \subset \mathbb{R}^d$, we define for $m \in \mathbb{N}$

$$\|u\|^2_{H^m_k(\Omega)} := \sum_{\substack{|\alpha|\leq m \\ \alpha\in\mathbb{N}^d}} \frac{1}{|\alpha|!\alpha!}\|(k^{-1}\partial)^\alpha u\|^2_{L^2(\Omega)}. \tag{3.2}$$

In this section, the α-dependence of the constants in (3.2) is not important, but it *is* important for the results on p-explicit polynomial approximation in Appendix A and hence also for obtaining the hp-FEM result **R2**.

For $m \in \mathbb{N}$ and $\Omega \subset \mathbb{R}^d$ open, we also define

$$H^m_0(\Omega) := \overline{C^\infty_c(\Omega)}^{\|\cdot\|_{H^m_k(\Omega)}},$$

and $H^{-m}_k(\Omega) := (H^m_k(\Omega) \cap H^m_0(\Omega))^*$ (where * denotes the dual space, and the norm is defined as a linear functional on $H^m_k(\Omega)$).

Let $\mathcal{R}$ be the solution operator for the problem (3.1) (i.e. the map $f \mapsto u$, aka the resolvent). Given $\chi \in C^\infty_c(\mathbb{R}^d)$ with $\chi \equiv 1$ on $\mathrm{supp}(A - I) \cup \mathrm{supp}(n - 1) \cup \Omega_-$, let

$$\rho(k) := \|\chi\mathcal{R}(k)\chi\|_{L^2\to L^2}, \tag{3.3}$$

that is, $\rho(k)$ is the $L^2 \to L^2$ norm of the cut-off resolvent. One can show that up to a k-independent constant, $\rho(k)$ does not depend on the choice of χ.

With this definition, $\rho(k) \sim k$ for non-trapping problems (i.e. when all geometric optics rays escape a neighbourhood of the scatterer) and $\rho(k) \gg k$ for trapping problems (see the references in Section 4.3 below). Furthermore, $\rho(k)$ is polynomially bounded in k for 'most' frequencies (see Section 4.3).

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

The advantage of scaling each derivative with k^{-1} – both in the Helmholtz equation (3.1) and in the norm (3.2) – is that the norm of $\mathcal{R}(k)$ has the same k-dependence between any spaces on which it is defined, e.g. given $\widetilde{\chi} \in C_c^\infty(\mathbb{R}^d)$ with $\operatorname{supp}\widetilde{\chi} \cap \operatorname{supp}(1-\chi) = \emptyset$, $\|\widetilde{\chi}\mathcal{R}(k)\chi\|_{L^2 \to H^1_k} \leq C(1+\rho(k))$. This can be proved by Green's identity; see e.g. Spence (2014, Lemma 2.2).

3.1. ***R1***: *k-explicit error analysis of the h-FEM*

Informal Theorem 3.1 (Preasymptotic error estimates for the h-FEM). Suppose that the Helmholtz problem (3.1) – with the Sommerfeld radiation condition approximated by *either* a radial perfectly matched layer *or* the exact Dirichlet-to-Neumann map on the boundary of a ball, *or* an impedance boundary condition – is solved using the Galerkin FEM with degree p polynomials (with approximation space denoted by $\mathcal{V}_h$). If the coefficients are piecewise $C^{p-1,1}$, the boundary of the domain and the boundaries across which the coefficients jump are all $C^{p,1}$, and

$$\rho(k)(hk)^{2p} \text{ is sufficiently small,}$$

then the Galerkin solution u_h exists, is unique, and satisfies

$$\|u-u_h\|_{H^1_k(\Omega)} \leq C(1+\rho(k)(hk)^p) \min_{v_h \in \mathcal{V}_h} \|u-v_h\|_{H^1_k(\Omega)} \tag{3.4}$$

and, for $n \in \{1,\dots,p\}$,

$$\|u-u_h\|_{H^{1-n}_k(\Omega)} \leq C((hk)^n + \rho(k)(hk)^p) \min_{v_h \in \mathcal{V}_h} \|u-v_h\|_{H^1_k(\Omega)}. \tag{3.5}$$

Furthermore, if the data f has frequency $\lesssim k$, then

$$\frac{\|u-u_h\|_{H^1_k(\Omega)}}{\|u\|_{H^1_k(\Omega)}} \leq C(1+\rho(k)(hk)^p)(hk)^p. \tag{3.6}$$

The rigorous statement of Informal Theorem 3.1 (in the specific case of PML truncation) is Theorem 6.27 below.

The context of Informal Theorem 3.1. Informal Theorem 3.1 contains the following two results:

(i) if $\rho(k)(hk)^p$ is sufficiently small then the Galerkin solution is *quasioptimal* (i.e. the Galerkin error is bounded by a multiple of the best approximation error), uniformly in k, via (3.4), and

(ii) for physically relevant data, the relative H^1_k-error can be made controllably small by making $\rho(k)(hk)^{2p}$ sufficiently small via (3.6).

Regarding result (i) (quasioptimality): the Schatz duality argument (Schatz 1974) (recapped in Lemma 6.9 below) combined with H^2-regularity of the solution implies quasioptimality for any $p \geq 1$ when $\rho(k)hk$ is sufficiently small. This was first recognized for one-dimensional problems in Aziz, Kellogg and Stephens

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

(1988), with a k-explicit bound on the solution operator obtained using Rellich–Morawetz identities (discussed in Section 5.6.1 and Remark 5.22 below). Other uses of this argument on one-dimensional problems included Douglas Jr, Santos, Sheen and Bennethum (1993, Lemma 2.6), Ihlenburg and Babuška (1995*a*, Theorem 3), Makridakis, Ihlenburg and Babuška (1996, Theorem 3.2) and on the two-dimensional interior impedance problem in Melenk (1995, Proposition 8.2.7).

Result (i) was then proved for certain constant-coefficient Helmholtz problems (Melenk and Sauter 2010, 2011) as part of their hp-FEM analysis (discussed in Section 3.2 below). One notable feature of this work is that (as far as the authors are aware) it was the first use in a numerical analysis context of the fact that the Helmholtz solution operator is well-behaved on high frequencies (which we understood in Section 2.2 in terms of semiclassical ellipticity).

Result (i) for general Helmholtz problems satisfying essentially the same assumptions as the rigorous version of Informal Theorem 3.1 (see Corollary 6.26 and Theorem 6.27 below) was then proved in Chaumont-Frelet and Nicaise (2020). These assumptions are essentially that the solution operator satisfies the natural elliptic-regularity shift: that is, if the data $f \in H^{m-1}$ then $u \in H^{m+1}$ for $0 \leq m \leq p$ (see Assumptions 6.19 and 6.23 below).

Regarding result (ii) (controllably small relative error): the threshold '$k(hk)^{2p}$ sufficiently small' for controllably small relative error was first identified for $p = 1$ in Ihlenburg and Babuška (1995*a*) and for general $p \in \mathbb{Z}^+$ in Ihlenburg and Babuška (1997) for constant-coefficient one-dimensional problems (where the solution can be written down explicitly). These papers also introduced the terminology that the solution is quasioptimal in the *asymptotic regime*. The complement of this is the *preasymptotic regime*, and the fact that the asymptotic regime is not 'hk sufficiently small' (and thus $\gg k^d$ degrees of freedom is required to maintain accuracy) is the *pollution effect*; see also Babuška and Sauter (1997) and Definition 3.4 below. The condition '$k(hk)^{2p}$ sufficiently small' also appears from so-called *dispersion analysis*, namely analysis of the FEM linear system for the constant-coefficient Helmholtz equation (Harari and Hughes 1991, Ihlenburg and Babuška 1995*b*,*a*, Ihlenburg 1998, Deraemaeker, Babuška and Bouillard 1999, Ainsworth 2004).

Result (ii) when $p = 1$ was proved

- for the interior impedance problem for discontinuous Galerkin methods in Feng and Wu (2009, 2011) and for continuous interior penalty methods and the standard FEM in Wu (2014), and
- for truncation via the exact Dirichlet-to-Neumann map (Lafontaine, Spence and Wunsch 2022*a*, Theorem 4.1), truncation via a radial, k-independent PML (Li and Wu 2019, Theorem 4.4) and truncation via a radial, k-dependent PML (Chaumont-Frelet, Gallistl and Tomezyk 2022, Theorem 7.2).

All these results used the *elliptic projection* argument introduced in Feng and Wu (2009); we discuss this argument in detail in Section 6.4 below.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Result (ii) for general $p \in \mathbb{Z}^+$ was then proved for the interior impedance problem in Du and Wu (2015, Theorem 5.1) (for constant coefficients) and Pembery (2020, Theorem 2.39) (for variable coefficients), and then for general Helmholtz problems satisfying an elliptic-regularity shift in Galkowski and Spence (2025) (as described in Informal Theorem 3.1).

3.2. ***R2***: *k-explicit error analysis of the hp-FEM*

Informal Theorem 3.2 (The hp-FEM does not suffer from the pollution effect). Suppose that the Helmholtz problem (3.1) with a Dirichlet boundary condition and the Sommerfeld radiation condition approximated by a radial PML is solved using the Galerkin FEM with degree p polynomials.

Suppose that the coefficients are analytic near Ω_- and smooth elsewhere, Ω_- has analytic boundary, the PML truncation boundary is $C^{1,1}$, and $\rho(k)$ is polynomially bounded in k. Then there exists $C > 0$ such that, given $k_0, \epsilon > 0$, there exists $c > 0$ such that, if $k \geq k_0$,

$$\frac{hk}{p} \leq c \quad \text{and} \quad p \geq 1 + \epsilon \log k,$$

then the Galerkin solution u_h exists, is unique, and satisfies

$$\|u - u_h\|_{H^1_k(\Omega)} \leq C \min_{v_h \in \mathcal{V}_h} \|u - v_h\|_{H^1_k(\Omega)}. \tag{3.7}$$

The rigorous statement of Informal Theorem 3.2 is Theorem 7.2 below.

The context of Informal Theorem 3.2. With a uniform mesh, the number of degrees of freedom of the piecewise polynomial approximation space $\sim (p/h)^d$; thus Informal Theorem 3.2 implies that with $hk/p = c$, k-uniform quasioptimality is achieved with the number of degrees of freedom $\sim k^d$, that is, the hp-FEM does not suffer from the pollution effect.

The first results of this form were proved for constant-coefficient Helmholtz problems in Melenk and Sauter (2010, 2011), Esterhazy and Melenk (2012) and Melenk, Parsania and Sauter (2013). As mentioned in the discussion in Section 3.1, these results used the fact that Helmholtz operator is well-behaved on high frequencies, with the splitting achieved via the Fourier transform in $\mathbb{R}^d$ and then combined with extension operators to define splittings on Lipschitz domains (see Melenk and Sauter 2011).

Results about quasioptimality of the hp-FEM applied to variable-coefficient problems were then obtained via two approaches:

(i) Lafontaine *et al.* (2022*b*), Galkowski *et al.* (2023*b*, 2024*c*), and

(ii) Bernkopf, Chaumont-Frelet and Melenk (2025),

with the rigorous version of Informal Theorem 3.2 proved in Galkowski *et al.* (2024*c*).

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Both approaches (i) and (ii) use the fact that the Helmholtz solution operator is well-behaved on high frequencies. In Lafontaine *et al.* (2022*b*) and Galkowski *et al.* (2023*b*, 2024*c*) this is understood and implemented via semiclassical ellipticity, with the expository paper by Spence (2023) reproducing the result in Melenk and Sauter (2010) using this approach. Bernkopf *et al.* (2025) uses semiclassical ellipticity implicitly via the fact that $(-k^{-2}\Delta - 1)^{-1}$ and $(-k^{-2}\Delta + 1)^{-1}$ act similarly on high-frequency data; see the discussion in Bernkopf *et al.* (2025, §2.5). The approach of Lafontaine *et al.* (2022*b*) and Galkowski *et al.* (2023*b*, 2024*c*) allows for smooth coefficients (with no obstacles) and the addition of an analytic Dirichlet obstacle, provided that the coefficients are analytic in a neighbourhood of the obstacle. The approach of Bernkopf *et al.* (2025) considers piecewise analytic coefficients and analytic boundaries/interfaces.

In this article, Informal Theorem 3.2 is proved via establishing, first, a new abstract result about quasioptimality of the hp-FEM under abstract regularity-type assumptions on the natural 'low frequencies' associated with the operator (see Theorem 7.1 below); this result is conceptually similar to – and simpler than – the analysis in Galkowski *et al.* (2023*b*, 2024*c*) that reduces proving quasioptimality of the hp-FEM applied to 'black box' scattering problems (in the sense of Sjöstrand and Zworski 1991) to proving certain elliptic-regularity-type statements.

The regularity assumptions in this abstract result are then shown to be satisfied in the set-up of Informal Theorem 3.2 (see Section 7.5 below). In fact, we expect these regularity-type assumptions to also hold for coefficients that are piecewise Gevrey in a neighbourhood of a Gevrey interface (which would then link – and extend – the approaches (i) and (ii) above), but, to the best of the authors' knowledge, the elliptic-regularity results necessary are not immediately available in the literature.

3.3. ***R3***: *Optimal error estimates for the h-BEM*

When $A \equiv I$ and $n \equiv 1$ the scattering problem (3.1) can be reformulated as a boundary integral equation (BIE) on $\Gamma_- := \partial\Omega_-$ (see Appendix B). We consider the standard second-kind BIE formulations for the Helmholtz scattering problems, which for the Dirichlet problem involve the operators

$$A'_k := \tfrac{1}{2}I + K'_k - ikV_k \quad \text{and} \quad A_k := \tfrac{1}{2}I + K_k - ikV_k, \tag{3.8}$$

and for the Neumann problem

$$B_{\rm reg} := i\big(\tfrac{1}{2}I - K_k\big) + V_{ik}H_k \quad \text{and} \quad B'_{\rm reg} := i\big(\tfrac{1}{2}I - K'_k\big) + H_kV_{ik}, \tag{3.9}$$

where the operators V_k, K_k, K'_k and H_k are the single-, double-, adjoint-double-layer and hypersingular operators defined by (B.4) and (B.5) below. For the history of these particular second-kind formulations, see Remark 8.1 below.

The BIEs involving these operators can all be written in the form: given $f \in L^2(\Gamma_-)$, find $v \in L^2(\Gamma_-)$ such that

$$\mathcal{A}v = f, \quad \mathcal{A} := c_0(I + \mathcal{L}), \quad c_0 \in \mathbb{C} \setminus \{0\}, \tag{3.10}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

where the operator $\mathcal{L}\colon L^2(\Gamma_-) \to L^2(\Gamma_-)$ is compact; see Theorem B.2 below. Let

$$\rho_{\mathcal{A}}(k) \coloneqq \|\mathcal{A}^{-1}\|_{L^2(\Gamma_-)\to L^2(\Gamma_-)},$$

and observe that the compactness of $\mathcal{L}$ implies that $\rho_{\mathcal{A}}(k) \geq c_0^{-1}$ (see Lemma B.3 below).

For the Dirichlet problem,

$$\rho_{A_k}(k) = \rho_{A'_k}(k) \leq Ck^{-1}\rho(k), \tag{3.11}$$

where $\rho(k)$ is defined by (3.3); see Chandler-Wilde and Monk (2008), Baskin, Spence and Wunsch (2016) and Chandler-Wilde, Spence, Gibbs and Smyshlyaev (2020, Lemma 6.2). For the Neumann problem, $\rho_{B_{\mathrm{reg}}}(k) = \rho_{B'_{\mathrm{reg}}}(k)$, but the relationship with the Helmholtz solution operator is more complicated; see Galkowski, Marchand and Spence (2022, equation 7.8).

We study the Galerkin approximation to v with $a\colon L^2(\Gamma_-)\times L^2(\Gamma_-) \to \mathbb{C}$ defined by

$$a(v,w) \coloneqq \langle \mathcal{A}v, w\rangle_{L^2(\Gamma_-)}.$$

Informal Theorem 3.3 (Preasymptotic error estimates for the h-BEM). Suppose that Γ_- is smooth and $\rho_{\mathcal{A}}(k)$ is polynomially bounded in k. Suppose that (3.10) is solved using the Galerkin BEM with degree p polynomials (with approximation space $\mathcal{V}_h$). If

$$(hk)^{2p+2}\rho_{\mathcal{A}}(k) \text{ is sufficiently small,} \tag{3.12}$$

then for all $v \in L^2(\Gamma_-)$, the Galerkin approximation v_h to v in $\mathcal{V}_h$ exists, is unique, and satisfies

$$\|v-v_h\|_{L^2(\Gamma_-)} \leq \left(1+C(hk)^{p+1}\rho_{\mathcal{A}}(k)+Chk+Ck^{-1}\right) \min_{w_h\in\mathcal{V}_h} \|v-w_h\|_{L^2(\Gamma_-)}. \tag{3.13}$$

Furthermore, if the right-hand side f has frequency $\lesssim k$ (e.g. in plane-wave scattering), then

$$\frac{\|v-v_h\|_{L^2(\Gamma_-)}}{\|v\|_{L^2(\Gamma_-)}} \leq \left(1+C(hk)^{p+1}\rho_{\mathcal{A}}(k)+Chk+Ck^{-1}\right)C(hk)^{p+1}. \tag{3.14}$$

Informal Theorem 3.3 and its rigorous version, Theorem 8.12 below, show that the Galerkin solution is quasioptimal, uniformly in k, if $(hk)^{p+1}\rho_{\mathcal{A}}(k)$ is sufficiently small, and that the relative error can be made controllably small if $(hk)^{2p+2}\rho_{\mathcal{A}}(k)$ is sufficiently small (compare to the h-FEM result of Informal Theorem 3.1). By (3.11), $\rho_{A_k}(k) \leq C$ when the problem is non-trapping, and thus these results show that the h-BEM for the Dirichlet BIEs does not suffer from the pollution effect for non-trapping problems (with this first proved in Galkowski and Spence 2023). However, since $\rho_{A_k}(k) \gg 1$ for trapping problems and $\rho_{B_{\mathrm{reg}}}(k) \gg 1$ even for non-trapping problems, Informal Theorem 3.3/Theorem 8.12 leave open the possibility that the h-BEM suffers from the pollution effect.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

To define pollution carefully, we recall that for a scattering problem, the solution v to the BIE (3.10) oscillates at frequency $\lesssim k$ with few additional properties. The Weyl law (see e.g. Zworski 2012, Chapter 14) or the Nyquist–Shannon–Whittaker sampling theorem (Whittaker 1915, Shannon 1949; see e.g. Bachman, Narici and Beckenstein 2000, Theorem 5.21.1) then implies that the dimension of this space is $\sim k^{d-1}$. Thus it is certainly not possible to achieve accuracy for general scattering data with a space of dimension $\ll k^{d-1}$. Motivated by this, we say that a numerical BIE method *suffers from the pollution effect* if $N \gg k^{d-1}$ is required to maintain accuracy of the computed solutions as $k \to \infty$. We make this precise via the following definition.

Definition 3.4 (The pollution effect in $L^2(\Gamma_-)$). Let $\mathcal{V}$ be a collection of subspaces of $L^2(\Gamma_-)$. For $V \in \mathcal{V}$ and $k > 0$, let $\mathcal{A}_V^{-1}(k)\colon L^2(\Gamma_-) \to V$ be an approximation of $\mathcal{A}^{-1}(k)$. The pair $(\mathcal{V}, \{\mathcal{A}_V^{-1}\}_{V\in\mathcal{V}})$ *suffers from the pollution effect in* $L^2(\Gamma_-)$ if

$$\inf_{\Lambda>0} \limsup_{k\to\infty} \sup_{\substack{V\in\mathcal{V}\\ \dim V\geq \Lambda k^{d-1}}} \sup_{f\in V} C_{\mathrm{qo}}(f,V) = \infty, \tag{3.15}$$

where

$$C_{\mathrm{qo}}(f,V) := \inf\left\{M>0\colon \left\|\mathcal{A}^{-1}f - \mathcal{A}_V^{-1}f\right\|_{L^2(\Gamma_-)} \leq M \min_{w\in V}\left\|\mathcal{A}^{-1}f - w\right\|_{L^2(\Gamma_-)}\right\}.$$

To unpack this definition, observe that if the right-hand side of (3.15) is finite, then there exists k_0, Λ and $M < \infty$ such that for all $k \geq k_0$, $V \in \mathcal{V}$ with $\dim V \geq \Lambda k^{d-1}$ and $f \in L^2(\Gamma_-)$,

$$\left\|\mathcal{A}^{-1}f - \mathcal{A}_V^{-1}f\right\|_{L^2(\Gamma_-)} \leq M \min_{w\in V}\left\|\mathcal{A}^{-1}f - w\right\|_{L^2(\Gamma_-)}, \tag{3.16}$$

that is, k-uniform quasi-optimality is achieved (for all possible data) with a choice of subspace dimension proportional to k^{d-1}.

Informal Theorem 3.5 (The h-BEM suffers from the pollution effect). Let $\mathcal{A} \in \{A_k, A_k'\}$. There exists Γ_- smooth such that if (3.10) is solved using the Galerkin BEM with degree p polynomials (with approximation space $\mathcal{V}_h$), then given $\epsilon > 0$ there are sequences of $h \to 0$ and $k \to \infty$ with $hk \leq \epsilon$ and a sequence of $v \in L^2(\Gamma_-)$ such that the Galerkin approximation v_h to v in $\mathcal{V}_h$, if it exists, satisfies

$$\|v - v_h\|_{L^2(\Gamma_-)} \geq \epsilon^{-1} \min_{w_h\in\mathcal{V}_h} \|v - w_h\|_{L^2(\Gamma_-)}. \tag{3.17}$$

Furthermore, if $\Omega_- = B(0,1)$ and $\mathcal{A} \in \{B_{\mathrm{reg}}, B_{\mathrm{reg}}'\}$, then there exists $c > 0$ such that if (3.10) is solved using the Galerkin BEM with degree p polynomials (with approximation space $\mathcal{V}_h$), there exists $v \in L^2(\Gamma_-)$ such that the Galerkin approximation v_h to v in $\mathcal{V}_h$, if it exists, satisfies

$$\|v - v_h\|_{L^2(\Gamma_-)} \geq c\left(1 + \min\left\{(hk)^{p+1}\rho_{\mathcal{A}}(k), (hk)^{-p-1}\right\}\right) \min_{w_h\in\mathcal{V}_h} \|v - w_h\|_{L^2(\Gamma_-)}. \tag{3.18}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

The bound (3.17) implies that the h-BEM with $\mathcal{A} \in \{A_k, A_k'\}$ suffers from the pollution effect (in the sense of Definition 3.4). Moreover, since for $\mathcal{A} \in \{B_{\text{reg}}, B_{\text{reg}}'\}$ and $\Omega_- = B(0,1)$, $\rho_{\mathcal{A}}(k) \sim k^{1/3}$ (Galkowski *et al.* 2025, §14), if $hk = \epsilon$, then (3.18) implies that the quasioptimality constant is *at least* $c\epsilon^{-p-1}$, and hence that decreasing ϵ can actually make the quasioptimality constant worse until the point that $\epsilon^{2p+2}k^{1/3} \lesssim 1$. The bound (3.18) also shows that Informal Theorem 3.3 is optimal over its full range of validity (3.12).

In the rigorous versions of Informal Theorem 3.5 – Theorems 8.18, 8.20 and 8.21 below – we quantify (3.17) and show that the phenomenon of decreasing hk resulting in *increasing* the quasioptimality constant also occurs for certain Ω_- with $\mathcal{A} \in \{A_k, A_k'\}$. Furthermore, Figure 8.2 gives a numerical illustration of pollution for the h-BEM, with further numerical results demonstrating pollution in h-BEM in Galkowski *et al.* (2025, §2.3).

The context of Informal Theorem 3.3. This was proved in Galkowski *et al.* (2025). Of the previous investigations into proving quasioptimality of the Galerkin method with piecewise polynomials applied to A_k' and A_k, namely Banjai and Sauter (2007), Löhndorf and Melenk (2011), Graham, Löhndorf, Melenk and Spence (2015), Spence, Kamotski and Smyshlyaev (2015), Galkowski, Müller and Spence (2019) and Galkowski and Spence (2023), the best results are the following.

- The result of Galkowski and Spence (2023, Lemma 3.1) that if Γ_- is smooth and $\rho_{\mathcal{A}}(k)(hk)$ is sufficiently small, then the Galerkin method is quasioptimal with constant proportional to $\rho_{\mathcal{A}}(k)$.
- The result of Löhndorf and Melenk (2011, Theorem 3.17) that if Γ is analytic, then the Galerkin method applied to A_k' is quasioptimal (with constant independent of k) if $(hk)^{p+1}k^5\rho_{\mathcal{A}}(k)$ is sufficiently small (Löhndorf and Melenk 2011, equation 3.22); a similar result holds for A_k with k^5 replaced by k^6 (Löhndorf and Melenk 2011, equation 3.26).

The quasioptimality constants in these two sets of results are larger than that in (3.13), and the thresholds for existence are more restrictive than that in (3.12).

As far as the authors are aware, there are no other results stated in the literature about quasioptimality of the Galerkin method applied to either B_{reg} or B_{reg}' defined by (3.9).

Remark 3.6 (Analysis of collocation and Nyström methods). We highlight that Galkowski *et al.* (2025) contains results analogous to Informal Theorem 3.3 for the Galerkin method in two dimensions with trigonometric polynomials, the collocation method in two dimensions with *either* piecewise polynomials *or* trigonometric polynomials, and for a two-dimensional Nyström method – i.e. a fully discrete method – based on trigonometric polynomials. For brevity, we do not state these results here, but note that for all the methods involving trigonometric polynomials, up to potential factors of k^ϵ for any $\epsilon > 0$, there is no pollution, even for trapping obstacles (see Galkowski *et al.* 2025, Table 2.1 for an overview of these results).

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

The context of Informal Theorem 3.5. As discussed in Section 3.1, the pollution effect was famously identified in Ihlenburg and Babuška (1995*a*, 1997) for the FEM applied to constant-coefficient Helmholtz problems in one dimension, and has since been observed in many numerical experiments involving the FEM in higher dimensions.

In contrast, there is a common belief in both engineering (Marburg 2002, 2016, 2017) and numerical analysis (Banjai and Sauter 2007, Graham *et al.* 2015, Löhndorf and Melenk 2011) that boundary integral methods do not suffer from the pollution effect. Indeed, it is standard in engineering to ask whether a specific number of points per wavelength (e.g. six: Marburg 2002) is sufficient to obtain accurate solutions. Informal Theorem 3.5 disproves this common belief, showing that no fixed number of points per wavelength is sufficient. In fact, to the best of the authors' knowledge, Informal Theorem 3.5 (or rather its rigorous analogues Theorems 8.18, 8.20 and 8.21) is the first rigorous proof of pollution for a numerical method applied to a Helmholtz problem with non-empty scatterer.

3.4. ***R4**: Non-uniform h-FEM meshes for scattering problems with PML truncation, where the meshes are defined by ray dynamics*

We consider the scattering problem in (3.1) with smooth coefficients, smooth obstacle, the radiation condition approximated by a radial PML, and under the assumption that $\rho(k)$ is polynomially bounded in k.

Theorem 3.1 shows that the h-FEM solution

- is quasioptimal if $\rho(k)(hk)^p$ is sufficiently small (see (3.4)), and
- has controllably small H^1_k relative error (for data with frequency $\lesssim k$) if $\rho(k)(hk)^{2p}$ is sufficiently small (see (3.6)).

We now show that the properties of quasioptimality and controllably small relative error can be achieved using certain non-uniform meshes that violate the conditions above on h and involve coarser meshes away from trapping and in the perfectly matched layer (PML).

Subsets of Ω defined by ray dynamics. We define billiard trajectories (i.e. the natural analogue of geometric optics rays in the presence of a boundary) to be geodesics for the metric $g^{-1} = n^{-1}A$ in Ω_+ continued by reflection with respect to g at the boundary of Ω_+; when $A = I$ and $n = 1$, these are straight-line paths continued using the Snell–Descartes law at the boundary. Next, we define the *cavity* $\mathcal{K} \subset \overline{\Omega}_+$ as the set of points $x \in \overline{\Omega}_+$ such that there is a billiard trajectory passing over x that remains in a compact set for all positive and negative times. We also define the *visible set* $\mathcal{V} \subset \overline{\Omega}_+$ as those points $x \in \overline{\Omega}_+$ such that there is a billiard trajectory passing over x that remains in a compact set for all positive times or all negative times. Finally, we define the *invisible set* $\mathcal{I} := \overline{\Omega}_+ \setminus (\mathcal{V} \cup \mathcal{K})$ (the adjectives visible and invisible are relative to the cavity). Let $\Gamma_{\rm tr} := \partial\Omega \setminus \Gamma_-$ be the truncation boundary, and let $\Omega_{\mathcal{P}} \subset \Omega$ be an open neighbourhood of $\Gamma_{\rm tr}$ that is

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

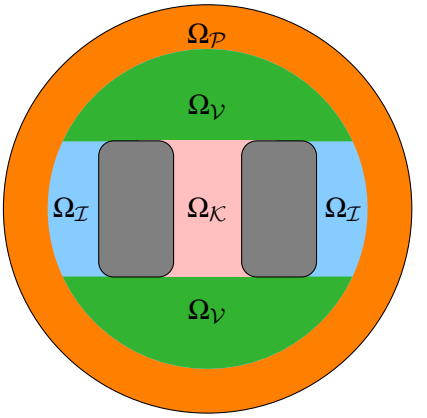


Figure 3.1. The domains $\Omega_{\mathcal{K}}, \Omega_{\mathcal{V}}, \Omega_{\mathcal{I}}$ and $\Omega_{\mathcal{P}}$, when Ω_- consists of two (rounded) aligned rectangles

strictly contained in the PML. Next, let $\Omega_{\mathcal{K}}$, $\Omega_{\mathcal{V}}$ and $\Omega_{\mathcal{I}}$ be open neighbourhoods of the intersections with $\overline{\Omega}$ of, respectively, $\mathcal{K}$, $\mathcal{V} \setminus (\mathcal{K} \cup \Omega_{\mathcal{P}})$ and $\mathcal{I} \setminus \Omega_{\mathcal{P}}$ in the subspace topology of $\overline{\Omega}$ such that $\Omega_{\mathcal{K}} \cap \Gamma_{\rm tr} = \Omega_{\mathcal{V}} \cap \partial\Omega_{\rm tr} = \Omega_{\mathcal{I}} \cap \Gamma_{\rm tr} = \emptyset$; for an illustration of these domains, see Figure 3.1.

The finite element space. Given a triangulation $\mathcal{T}$ of Ω, we define

$$h_\star := \sup\{h_T : T \in \mathcal{T},\, T \cap \Omega_\star \neq \emptyset\}, \quad \text{with } \star \in \{\mathcal{K}, \mathcal{V}, \mathcal{I}, \mathcal{P}\}, \tag{3.19}$$

where h_T is the width of the element T. We also define $h := \max_{T\in\mathcal{T}} h_T$.

Results about the h-FEM applied with non-uniform meshes. Theorem 9.1 below details how varying the meshwidth in one of the regions $\Omega_{\mathcal{K}}, \Omega_{\mathcal{V}}, \Omega_{\mathcal{I}}$ and $\Omega_{\mathcal{P}}$ affects errors both in that region and elsewhere. Table 3.1 gives some important special cases of Theorem 9.1, showing, in particular, how the conditions '$\rho(k)(hk)^p$ sufficiently small' for quasioptimality and '$\rho(k)(hk)^{2p}$ is sufficiently small for controllably small H^1-error can be violated away from $\Omega_{\mathcal{K}}$.

In Table 3.1, *k-QO* stands for k-uniform quasioptimality and means that there exists $C > 0$ independent of k such that

$$\|u - u_h\|_{H^1_k(\Omega)} \leq C \inf_{w\in\mathcal{V}_h} \|u - w_h\|_{H^1_k(\Omega)},$$

k-QO away from trapping means that if U is any open set not intersecting $\mathcal{K}$, then

$$\|u - u_h\|_{H^1_k(U)} \leq C \inf_{w_h\in\mathcal{V}_h} \|u - w_h\|_{H^1_k(\Omega)},$$

CRE stands for controllably small relative error and means that if $\|u\|_{H^{p+1}_k(\Omega)} \leq C\|u\|_{H^1_k(\Omega)}$, i.e. u is oscillating at frequency $\lesssim k$, then by making c in the leftmost column small as a function ϵ, one can obtain

$$\|u - u_h\|_{H^1_k(\Omega)} \leq \epsilon\|u\|_{H^1_k(\Omega)},$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Table 3.1. Summary of the special cases of Theorem 9.1 with $\mathcal{K} \neq \emptyset$ (i.e. the problem is trapping). In all cases we require $h_P k = c$ (not shown in the table for brevity), which does not contribute to the asymptotic number of degrees of freedom (DoFs). The notations k-QO, k-QO away from trapping, CRE, and CRE away from trapping are explained in the text.

fewer DoFs ↕ more DoFs

Mesh threshold	Asymptotic DoFs	Theoretical guarantee
$(h_{\mathcal{K}}k)^p\rho(k) + (h_{\mathcal{V}}k)^p\rho(k) + (h_{\mathcal{I}}k)^p\rho(k) = c$	$\mathrm{vol}(\Omega)k^d(\rho(k))^{d/p}$	k-QO
$(h_{\mathcal{K}}k)^p\rho(k) + (h_{\mathcal{V}}k)^p\sqrt{k\rho(k)} + (h_{\mathcal{I}}k)^p k = c$	$\mathrm{vol}(\Omega_{\mathcal{K}})k^d(\rho(k))^{d/p}$	k-QO
$(h_{\mathcal{K}}k)^p\sqrt{k\rho(k)} + (h_{\mathcal{V}}k)^p k + (h_{\mathcal{I}}k)^p k = c$	$\mathrm{vol}(\Omega_{\mathcal{K}})k^{d+1/(2p)}(\rho(k))^{d/(2p)}$	k-QO away from trapping
$(h_{\mathcal{K}}k)^{2p}\rho(k) + (h_{\mathcal{V}}k)^{2p}\rho(k) + (h_{\mathcal{I}}k)^{2p}\rho(k) = c$	$\mathrm{vol}(\Omega)k^d(\rho(k))^{d/(2p)}$	CRE
$(h_{\mathcal{K}}k)^{2p}\rho(k) + (h_{\mathcal{V}}k)^{2p}\sqrt{k\rho(k)} + (h_{\mathcal{I}}k)^{2p} k = c$	$\mathrm{vol}(\Omega_{\mathcal{K}})k^d(\rho(k))^{d/(2p)}$	CRE
$(h_{\mathcal{K}}k)^{2p}\rho(k) + (h_{\mathcal{V}}k)^p k + (h_{\mathcal{I}}k)^p k = c$	$\mathrm{vol}(\Omega_{\mathcal{K}})k^d(\rho(k))^{d/(2p)}$	CRE away from trapping

and, similarly, *CRE away from trapping* means that if u is oscillating at frequency k, then for any U not intersecting $\mathcal{K}$, one can obtain

$$\|u - u_h\|_{H^1_k(U)} \leq \epsilon \|u\|_{H^1_k(\Omega)},$$

by making c small.

The context of the results summarized in Table 3.1. Estimates on the local FEM error for second-order linear elliptic PDEs were pioneered in Nitsche and Schatz (1974); see also Descloux (1975), Schatz and Wahlbin (1977), Schatz and Wahlbin (1982), Wahlbin (1991, Chapter 9) and Demlow, Guzmán and Schatz (2011). The standard interpretation of these results is that the FEM solution is locally quasioptimal up to a lower-order term that allows error to propagate in from the rest of the domain, sometimes called the 'slush' term in the literature.

The k-explicit analogues of these classic local-FEM-error results were then obtained in Averseng, Galkowski and Spence (2024), with the result being, essentially, that if $\Omega_0 \subset \Omega_1 \subset \Omega$, then

$$\|u - u_h\|_{H^1_k(\Omega_0)} \leq C\Big(\min_{w_h \in V_h} \|u - w_h\|_{H^1_k(\Omega_1)} + \|u - u_h\|_{(H^p_k(\Omega_1)^*} \Big). \tag{3.20}$$

The interpretation of (3.20) is that the 'slush' term (i.e. the second term on the right-hand side) consists of low frequencies, since a bound on a high k-weighted Sobolev norm of a function in terms of a low k-weighted Sobolev norm, with the constant independent of k, implies that the function is controlled by its frequencies $\lesssim k$, that is, the Helmholtz FEM solution is locally quasioptimal, modulo low frequencies. However, the results of Averseng *et al.* (2024) do not give any information about what impacts the low-frequency term in (3.20).

To the best of the authors' knowledge, Theorem 9.1 (and its more sophisticated analogue, Theorem 3.11 of Averseng *et al.* 2025) is the first result specifying in a k-explicit way how the low-frequency errors propagate around the domain (see the discussion in Section 9.1.3 below), and then using this information to design k-dependent non-uniform finite element meshes that maintain accuracy (either in the sense of k-QO or CRE) as $k \to \infty$. In particular, one can interpret the estimates in Theorem 9.1 (and their more sophisticated analogues in Theorem 3.11 of Averseng *et al.* 2025) as effectively estimating the low-frequency term in (3.20).

3.5. ***R5**: Analysis of overlapping Schwarz methods with PMLs at the subdomain boundaries*

Let P be as in (3.1) with $A \equiv I$ and $\Omega_- = \emptyset$. Given a hyper-rectangle Ω, let $P_{\rm s}$ be the operator P with the Sommerfeld radiation condition approximated by a Cartesian PML near $\partial\Omega$. We consider the problem: given $f \in H^{-1}(\Omega)$, find $u \in H^1_0(\Omega)$ such that $P_{\rm s} u = f$.

We split Ω into overlapping hyper-rectangles Ω_j, and let $P^j_{\rm s}$ be the operator $P_{\rm s}$ on Ω_j with additional Cartesian PMLs contained in the overlap between Ω_j and its neighbours. Let $\{\chi_j\}_{j=1}^N$ be a partition of unity with $\chi_j \in C^\infty_c(\Omega_j)$ and χ_j

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

supported away from the PML regions of P_{s}^j that are not PML regions of P_{s}. The parallel overlapping Schwarz method is then: let $u_0 = 0$ and for $n \geq 0$ define $u^{n+1} \in H_0^1(\Omega)$ by

$$u^{n+1} := u^n + \sum_{j=1}^{N} \chi_j (P_{\mathrm{s}}^j)^{-1} (f - P_{\mathrm{s}} u^n)|_{\Omega_j} \tag{3.21}$$

(see e.g. Dolean, Jolivet and Nataf 2015, §1.1).

Informal Theorem 3.7. Suppose that P is non-trapping. Let $\mathcal{N}$ be the maximum number of subdomains (counted with multiplicity) through which a geometric optics ray can travel before leaving Ω. After $\mathcal{N}$ iterations, the error $u - u^{\mathcal{N}}$ is both smooth and $O(k^{-\infty})$ (i.e. superalgebraically small in k).

The rigorous statement of Informal Theorem 3.7 is Theorem 10.5 below. We make the following immediate remarks.

- The number $\mathcal{N}$ is defined more precisely in Section 10.2 below, but we note here that when $n \equiv 1$ and $\{\Omega_j\}_{j=1}^N$ is a decomposition of Ω into N strips, then $\mathcal{N} = N$.
- Galkowski *et al.* (2024*a*), who originally proved Informal Theorem 3.7, also provide analogous results about sequential overlapping Schwarz methods, but for brevity we do not state these here.
- All the results in Galkowski *et al.* (2024*a*), while initially stated for Cartesian PML truncation on Ω and Cartesian PMLs on the subdomains, apply to truncation methods based on complex absorption and other types of PML; see the discussion in Galkowski *et al.* (2024*a*, Appendix A).

The context of Informal Theorem 3.7. The design and analysis of DD methods for solving the Helmholtz equation is a very active area; see e.g. the reviews by Erlangga (2008), Ernst and Gander (2012), Lahaye, Tang and Vuik (2017) and Gander and Zhang (2019, 2022).

A key question is: What boundary conditions should be imposed on the DD subdomains? It is known that the optimal boundary condition on a particular DD subdomain is the Dirichlet-to-Neumann map for the Helmholtz equation posed in the exterior of that subdomain (as a subset of the whole domain); see Nataf, Rogier and de Sturler (1994, §2), Engquist and Zhao (1998, §2) and Dolean *et al.* (2015, §2.4). However, complete knowledge of these maps is equivalent to knowing the solution operator for the original problem. The design of good, practical subdomain boundary conditions is then the goal of *optimized Schwarz methods*; see e.g. Gander (2006).

Using PML as a subdomain boundary condition for Helmholtz DD was first advocated for in Toselli (1998), and there are now many DD methods for Helmholtz using PML on subdomain boundaries with impressive empirical performance; see e.g. Engquist and Ying (2011), Stolk (2013), Chen and Xiang (2013), Zepeda-

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Núñez and Demanet (2016), Taus, Zepeda-Núñez, Hewett and Demanet (2019), Leng and Ju (2019), Royer, Geuzaine, Béchet and Modave (2022), Bouziani, Nataf and Tournier (2023) and the reviews by Gander and Zhang (2019, 2022).

It is now understood in a k-explicit way how PML approximates the Dirichlet-to-Neumann map associated with the Sommerfeld radiation condition. Indeed, when $A \equiv I$, $n \equiv 1$ and $d = 2$, the error between the true Helmholtz solution of (3.1) and the Cartesian PML approximation decreases exponentially in k, the width of the PML, and the strength of the scaling by Chen and Xiang (2013, Lemma 3.4); this result also holds for general scatterers in $d \geq 2$ with a radial PML, by Galkowski *et al.* (2023*a*) (see also Theorem 4.3). For general smooth n and any $d \geq 2$, this error is smaller than any negative power of k, by Galkowski *et al.* (2024*a*, Theorem 4.6).

However, to the best of the authors' knowledge, Informal Theorem 3.7 (or rather its rigorous statement Theorem 10.5) is the first result to provide a rigorous understanding of how well PML approximates the (more complicated) Dirichlet-to-Neumann maps corresponding to the optimal DD boundary conditions for general decompositions and non-empty scatterers.

For example, in the particular case of a strip decomposition with N subdomains, the parallel overlapping Schwarz method with a strip decomposition with N subdomains converges in N iterations with the optimal boundary conditions; see Nataf *et al.* (1994, Proposition 2.4) and Engquist and Zhao (1998, Lemma 2.2). For such a strip decomposition with $n \equiv 1$, Informal Theorem 3.7/Theorem 10.5 shows that the parallel method with PML at the subdomain boundaries has $O(k^{-\infty})$ error after N iterations; that is, PML approximates well the optimal boundary condition in this case. However, we point out that $\mathcal{N}$ can be much larger than N when $n \not\equiv 1$ and the rays are not straight lines (e.g. if the rays 'turn around' at one end of the domain), and PML does not approximate well the optimal boundary condition in this case.

The only other existing k-explicit rigorous convergence results are for the sequential 'source transfer' DD methods (which Gander and Zhang 2019 showed can be considered as a particular type of optimized Schwartz method) applied to (3.1) with $A \equiv I$, $n \equiv 1$ and $\Omega_- = \emptyset$ (i.e. no scattering). These methods involve subdomains that only overlap via the PMLs (i.e. the physically relevant parts of the subdomains do not overlap). For these methods applied with $A \equiv I$, $n \equiv 1$ and $\Omega_- = \emptyset$ (i.e. no scattering), k-explicit convergence of the DD method at the continuous level can be obtained from a k-explicit result about how well (Cartesian) PML approximates the Sommerfeld radiation condition; this is precisely because the optimal boundary conditions on the subdomains in this case are the Dirichlet-to-Neumann maps associated with the Sommerfeld radiation condition. Therefore accuracy of Cartesian PML when $A \equiv I$, $n \equiv 1$ and $\Omega_- = \emptyset$ (Chen and Xiang 2013, Lemma 3.4) is then the heart of the k-explicit convergence proofs of the source-transfer-type methods in Chen and Xiang (2013), Leng and Ju (2019), Du and Wu (2020) and Leng and Ju (2021, 2022) in the case of no scattering.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

4. Definition of the scattering problem and its approximation by a radial PML

4.1. The scattering problem

We consider scattering by a combination of an impenetrable obstacle and a penetrable obstacle (the latter corresponding to variable coefficients in the PDE). Let $\Omega_- \Subset \mathbb{R}^d$ be a bounded Lipschitz open set with connected open complement, $\Omega_+ := \mathbb{R}^d \setminus \overline{\Omega}_-$. Let A be an L^∞ symmetric, positive-definite matrix with real coefficients and $n \in L^\infty(\overline{\Omega_+}; \mathbb{R}_+)$ such that $\operatorname{supp}(A - I) \cup \operatorname{supp}(n-1) \Subset \overline{\Omega}_+$. Recall that the *conormal derivative* $\partial_{n,A}$ is such that $\partial_{n,A} u := n \cdot (A \nabla u)$ when $u \in H^2(\Omega)$, and that $\partial_{n,A} u$ can be defined for $u \in H^1(\Omega)$ with $\operatorname{div}(A\nabla u) \in L^2(\Omega)$ by Green's identity; see e.g. McLean (2000, Lemma 4.3).

Given $f \in L^2_{\text{comp}}(\overline{\Omega_+})$, the scattering problem is to find

$$u \in H^1_{\text{loc}}(\overline{\Omega_+}) := \{v : \chi v \in H^1(\Omega_+) \text{ for all } \chi \in C_c^\infty(\mathbb{R}^d)\}$$

such that

$$\begin{aligned} Pu := -k^{-2} \operatorname{div}(A\nabla u) - nu = f \text{ in } \Omega_+, \quad (Bu)|_{\partial\Omega_+} = 0, \\ \|(k^{-1}\partial_r - i)u\|_{L^\infty(|x|=r)} = o_{r\to\infty}(r^{(1-d)/2}), \end{aligned} \tag{4.1}$$

where $Bu = u$ in the sound-soft case and $Bu = \partial_{\nu,A} u$ in the sound-hard case. The condition as $r \to \infty$ in (4.1) is the *Sommerfeld radiation condition*; this condition ensures that the corresponding solutions of the wave equation created via (1.2) travel away from the obstacle as time increases.

4.2. Existence and uniqueness of the solution of the scattering problem

Since the boundary value problem (4.1) is Fredholm with index zero (see e.g. Dyatlov and Zworski 2019, Theorem 4.4), existence of a solution to (4.1) follows from uniqueness. Rellich's uniqueness theorem – see Rellich (1943), Colton and Kress (1983, §3.3), Dyatlov and Zworski (2019, Theorem 4.17) – shows that any two solutions to (4.1) must agree away from $\operatorname{supp}(A - I) \cup \operatorname{supp}(n-1) \cup \Omega_-$. Uniqueness can then be shown via the unique continuation principle (UCP), that is, any compactly supported solution of (4.1) with $f \equiv 0$ must be 0. The standard tool used to prove such a result is a *Carleman estimate*, going back to Carleman (1939) (where a UCP is proved for the case $d = 2$, $A \equiv I$ and $n \in L^\infty$).

There is a large literature on proving UCPs under progressively weaker conditions on A and n. Furthermore, given a UCP for A and n with certain regularity, Ball, Capdeboscq and Tsering-Xiao (2012) have proved uniqueness for the problem when A and n satisfy that regularity 'piecewise'; this argument is formulated in Ball *et al.* (2012) for the time-harmonic Maxwell equations, but holds more generally, in particular for the Helmholtz equation.

The upshot is that in two dimensions, the UCP holds when $A \in L^\infty$ and $n \in L^p$ with $p > 1$ (Alessandrini 2012). In three dimensions, the UCP holds when

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

A is Lipschitz (and hence piecewise Lipschitz by Ball *et al.* 2012) and $n \in L^\infty$ (Hörmander 1959). Much more singular n are allowed (although this is not important for this article); see Jerison and Kenig (1985), Wolff (1992) and the surveys in Lerner (2019, Chapter 8) and Wolff (1993). Examples of $A \in C^{0,\alpha}$ for all $\alpha < 1$ for which the solution of (4.1) is not unique at a particular $k > 0$ (and thus the UCP fails) are given in Plis (1963), Miller (1974) and Filonov (2001); see also Chaumont-Frelet (2025).

4.3. *Recap of the k-dependence of the solution operator to* (4.1)

Let $\mathcal{R}$ be the solution operator for the problem (4.1) (i.e. the map $f \mapsto u$, also known as the resolvent). Let R_{scat} be such that

$$\operatorname{supp}(A - I) \cup \operatorname{supp}(n - 1) \cup \Omega_- \Subset B(0, R_{\text{scat}}), \tag{4.2}$$

let $\chi \in C_c^\infty(\mathbb{R}^d)$ with $\chi \equiv 1$ on $B(0, R_{\text{scat}})$, and let $\rho(k)$ be defined by (3.3).

By considering solutions of the form $e^{ikx_1}\psi(x)$ for $\psi \in C_c^\infty(\mathbb{R}^d)$ with $\operatorname{supp}\psi \subset B(0, R_{\text{scat}}) \setminus \operatorname{supp}(A - I) \cup \operatorname{supp}(n - 1) \cup \Omega_-$, one can show that there exists $c > 0$ such that $\rho(k)$ defined by (3.3) satisfies

$$\rho(k) \geq ck \quad \text{for all } k > 0.$$

If the problem is non-trapping (i.e. all geometric optics rays escape to infinity), then given $k_0 > 0$ there exists $C > 0$ such that

$$\rho(k) \leq Ck \quad \text{for all } k \geq k_0$$

(Morawetz 1975, Vainberg 1975, Melrose and Sjöstrand 1982, Burq 2002, Chandler-Wilde and Monk 2008 and Dyatlov and Zworski 2019, Theorem 4.43), with the constant C proportional to the length of the longest ray (Galkowski, Spence and Wunsch 2020).

When the problem is trapping, $\rho(k)$ grows faster than k (Bony, Burq and Ramond 2010 and Dyatlov and Zworski 2019, Theorem 7.1), but if the boundary and coefficients are sufficiently smooth, then given $k_0 > 0$ there exist $C_1, C_2 > 0$ such that

$$\rho(k) \leq C_1 \exp(C_2 k) \quad \text{for all } k \geq k_0$$

(Burq 1998, Vodev 2000, Bellassoued 2003). Furthermore, without any smoothness requirements, the resolvent can, in fact, grow arbitrarily fast (Chandler-Wilde and Sadeghi 2025).

Nevertheless, for 'most' frequencies, the resolvent is polynomially bounded in k: for all $\delta, k_0, \varepsilon > 0$, there exists $C > 0$ and a set $\mathcal{J} \subset \mathbb{R}$ with Lebesgue measure $\leq \delta$ such that

$$\rho(k) \leq Ck^{5d/2+\varepsilon} \quad \text{for all } k \in [k_0, \infty) \setminus \mathcal{J} \tag{4.3}$$

(Lafontaine, Spence and Wunsch 2021).

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

4.4. The PML problem

We approximate the Sommerfeld radiation condition via a radial perfectly matched layer (PML). With $R_{\rm scat}$ defined by (4.2), let $R_{\rm PML} > R_{\rm scat}$ and let $\Omega_{\rm tr}$ be such that $B_{R_{\rm PML}} \Subset \Omega_{\rm tr}$. Let $\Omega := \Omega_+ \cap \Omega_{\rm tr}$ and $\Gamma_{\rm tr} := \partial\Omega_{\rm tr}$.

For $0 \le \theta < \pi/2$ (the *scaling angle*), let the PML scaling function $g_\theta \in C^\infty([0,\infty);\mathbb{R})$ be defined by $g_\theta(r) := g(r)\tan\theta$ for some g satisfying

$$\{g(r) = 0\} = \{g'(r) = 0\} = \{r \le R_{\rm PML}\}, \quad g'(r) \ge 0, \tag{4.4}$$

that is, the scaling 'turns on' at $r = R_{\rm PML}$. Given $g(r)$, let

$$\alpha(r) := 1 + ig'_\theta(r) \quad \text{and} \quad \beta(r) := 1 + ig_\theta(r)/r.$$

We now define two possible PML problems; both are formed by first replacing Δ in (4.1) by

$$\begin{aligned}\Delta_\theta &= \left(\frac{1}{1+ig'_\theta(r)}\frac{\partial}{\partial r}\right)^2 + \frac{d-1}{(r+ig_\theta(r))(1+ig'_\theta(r))}\frac{\partial}{\partial r} + \frac{1}{(r+ig_\theta(r))^2}\Delta_\omega \\ &= \frac{1}{(1+ig'_\theta(r))(r+ig_\theta(r))^{d-1}}\frac{\partial}{\partial r}\left(\frac{(r+ig_\theta(r))^{d-1}}{1+ig'_\theta(r)}\frac{\partial}{\partial r}\right) + \frac{1}{(r+ig_\theta(r))^2}\Delta_\omega\end{aligned}$$

(with Δ_ω the surface Laplacian on S^{d-1}) and then *either* multiplying by $\alpha\beta^{d-1}$ *or* not.

In both cases, the PML problem has the form: given $f \in L^2(\Omega)$, find $u \in H^1(\Omega)$ such that

$$P_\theta u := -k^{-2}\operatorname{div}(A_\theta \nabla u) + k^{-2} b_\theta \cdot \nabla u - n_\theta u = f \text{ in } \Omega,\ (Bu)|_{\partial\Omega_+} = 0,\ u|_{\partial\Omega_{\rm tr}} = 0, \tag{4.5}$$

with the associated sesquilinear form

$$a(u,v) := \int_\Omega \left(k^{-2}A_\theta(x)\nabla u(x)\cdot\overline{\nabla v(x)} + k^{-2}\langle b_\theta(x),\nabla u\rangle\overline{v(x)} - n_\theta(x)u(x)\overline{v(x)}\right)dx. \tag{4.6}$$

Finally, since $\alpha = \beta = 1$ for $r \le R_{\rm PML}$, when the data f in (4.1) is supported in this region, the right-hand sides of the two different PML problems are the same.

Comparison of the two different formulations. Before we give the precise definitions of the two different formulations, we discuss their advantages and disadvantages.

The multiplication by $\alpha\beta^{d-1}$ ensures that $b_\theta = 0$ in (4.5), that is, the resulting operator is in divergence form. However, additional assumptions are required for P_k to satisfy the Gårding inequality: there exists $\omega \in \mathbb{R}$ such that for all $k_0 > 0$, there are $c, C > 0$ such that for all $k \ge k_0$,

$$\operatorname{Re}\left(e^{i\omega}a_k(u,u)\right) \ge c\|u\|^2_{H^1_k(\Omega)} - C\|u\|^2_{L^2(\Omega)} \quad \text{for all } u \in H^1(\Omega). \tag{4.7}$$

In particular, Lemma 2.3 of Galkowski *et al.* (2024*c*) shows that (4.7) holds for any

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

$g_\theta(r)$ satisfying the above assumptions in $d = 2$ and holds in $d = 3$ when $g_\theta(r)/r$ is, in addition, non-decreasing, and Remark 2.1 of Galkowski *et al.* (2024*c*) shows that such an additional assumption is needed.

If we instead integrate by parts the complex-scaled PDE (i.e. without multiplying by $\alpha\beta^{d-1}$), then the resulting sesquilinear form satisfies the Gårding inequality after multiplication by $e^{i\omega}$ for some suitable constant ω (Galkowski *et al.* 2024*b*, Lemma A.6).

We note that in other papers on PMLs, the scaled variable, which in our case is $r + ig_\theta(r)$, is often written as $r(1 + i\widetilde{\sigma}(r))$ with $\widetilde{\sigma}(r) = \sigma_0$ for r sufficiently large; see e.g. Hohage, Schmidt and Zschiedrich (2003, §4) and Bramble and Pasciak (2007, §2). Therefore, to convert from our notation, set $\widetilde{\sigma}(r) = g_\theta(r)/r$ and $\sigma_0 = \tan\theta$. In this alternative notation, the assumption that $g_\theta(r)/r$ is non-decreasing is therefore that $\widetilde{\sigma}$ is non-decreasing; see Bramble and Pasciak (2007, §2).

The sesquilinear form after multiplication by $\alpha\beta^{d-1}$. Let $b_\theta := 0$,

$$A_\theta := \begin{cases} A & \text{in } \Omega, \\ HDH^\top & \text{in } (B_{R_{\mathrm{PML}}})^c \end{cases} \quad \text{and} \quad n_\theta := \begin{cases} n & \text{in } \Omega \cap B_{R_{\mathrm{PML}}}, \\ \alpha(r)\beta(r)^{d-1} & \text{in } (B_{R_{\mathrm{PML}}})^c, \end{cases} \tag{4.8}$$

where, in polar coordinates (r, φ),

$$D := \begin{pmatrix} \beta(r)\alpha(r)^{-1} & 0 \\ 0 & \alpha(r)\beta(r)^{-1} \end{pmatrix} \quad \text{and} \quad H := \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} \text{ for } d = 2,$$

and, in spherical polar coordinates (r, φ, ϕ),

$$D := \begin{pmatrix} \beta(r)^2\alpha(r)^{-1} & 0 & 0 \\ 0 & \alpha(r) & 0 \\ 0 & 0 & \alpha(r) \end{pmatrix} \text{ and } H := \begin{pmatrix} \sin\varphi\cos\phi & \cos\varphi\cos\phi & -\sin\phi \\ \sin\varphi\sin\phi & \cos\varphi\sin\phi & \cos\phi \\ \cos\varphi & -\sin\varphi & 0 \end{pmatrix}$$

for $d = 3$; observe that then $A = I$ and $n = 1$ when $r = R_{\mathrm{PML}}$, and thus A_θ and n_θ are continuous at $r = R_{\mathrm{PML}}$.

Lemma 4.1 (Galkowski *et al.* 2024*c*, Lemma 2.3). Let g_θ satisfy (4.4) and the additional assumption when $d = 3$ that $g_\theta(r)/r$ is non-decreasing. Given $\epsilon > 0$, there exists $c > 0$ such that for all $\epsilon \le \theta \le \pi/2 - \epsilon$, A_θ defined by (4.8) satisfies

$$\operatorname{Re}(A_\theta(x)\xi, \xi)_2 \ge c\|\xi\|_2^2 \quad \text{for all } \xi \in \mathbb{C}^d \text{ and } x \in \Omega;$$

thus the Gårding inequality (4.7) holds with $\omega = 0$.

The sesquilinear form without multiplication by $\alpha\beta^{d-1}$. Let

$$A_\theta := \begin{cases} A & \text{in } \Omega, \\ HDH^\top & \text{in } (B_{R_{\mathrm{PML}}})^c \end{cases} \quad \text{and} \quad n_\theta := \begin{cases} n & \text{in } \Omega \cap B_{R_{\mathrm{PML}}}, \\ 1 & \text{in } (B_{R_{\mathrm{PML}}})^c, \end{cases} \tag{4.9}$$

where, in polar coordinates (r, φ),

$$D := \begin{pmatrix} \alpha(r)^{-2} & 0 \\ 0 & \beta(r)^{-2} \end{pmatrix} \quad \text{and} \quad H := \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} \text{ for } d = 2,$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

and, in spherical polar coordinates (r, φ, ϕ),

$$D := \begin{pmatrix} \alpha(r)^{-2} & 0 & 0 \\ 0 & \beta(r)^{-2} & 0 \\ 0 & 0 & \beta(r)^{-2} \end{pmatrix} \quad \text{and} \quad H := \begin{pmatrix} \sin\varphi\cos\phi & \cos\varphi\cos\phi & -\sin\phi \\ \sin\varphi\sin\phi & \cos\varphi\sin\phi & \cos\phi \\ \cos\varphi & -\sin\varphi & 0 \end{pmatrix}$$

for $d = 3$. (Since $A_\theta = I$ and $n_\theta = 1$ when $r = R_{\text{PML}}$, A and n are continuous at $r = R_{\text{PML}}$.) In addition, for $d = 2$,

$$b_\theta(r) := \begin{cases} 0 & \text{in } \Omega \cap B_{R_{\text{PML}}}, \\ H \begin{pmatrix} \alpha^{-2}(\log(\alpha\beta))' \\ 0 \end{pmatrix} & \text{in } (B_{R_{\text{PML}}})^c, \end{cases}$$

and for $d = 3$,

$$b_\theta(r) := \begin{cases} 0 & \text{in } \Omega \cap B_{R_{\text{PML}}}, \\ H \begin{pmatrix} \alpha^{-2}(\log(\alpha\beta^2))' \\ 0 \\ 0 \end{pmatrix} & \text{in } (B_{R_{\text{PML}}})^c. \end{cases}$$

Lemma 4.2 (Galkowski *et al.* 2023*a*, Lemma A.6). Let g_θ satisfy (4.4). Given $\epsilon > 0$, there exists $\omega \in \mathbb{R}, c > 0$ such that for all $\epsilon \le \theta \le \pi/2 - \epsilon$, A_θ defined by (4.9) satisfies

$$\operatorname{Re}(e^{i\omega} A_\theta(x)\xi, \xi)_2 \ge c\|\xi\|_2^2 \quad \text{for all } \xi \in \mathbb{C}^d \text{ and } x \in \Omega;$$

thus the Gårding inequality (4.7) holds with $\omega = 0$.

4.5. *The PML problem is exponentially accurate at high frequency and inherits the bound on the solution operator of the scattering problem*

We define the exponential rate of growth for the solution operator away from $\mathcal{J} \subset \mathbb{R}$:

$$\Lambda(P, \mathbb{R} \setminus \mathcal{J}) := \limsup_{\substack{k\to\infty \\ k\in\mathbb{R}\setminus\mathcal{J}}} \frac{1}{k} \log \rho(k) = \limsup_{\substack{k\to\infty \\ k\in\mathbb{R}\setminus\mathcal{J}}} \frac{1}{k} \log \|\chi\mathcal{R}(k)\chi\|_{L^2\to L^2}. \tag{4.10}$$

We write $\Lambda(P)$ for $\Lambda(P, \mathbb{R})$. The bounds on $\rho(k)$ recapped in Section 4.3 imply that

- if the boundary and coefficients are sufficiently smooth, $\Lambda(P) < \infty$,
- if the problem is non-trapping then $\Lambda(P) = 0$,
- for all $\delta > 0$ there is a set $\mathcal{J} \subset \mathbb{R}$ with $|\mathcal{J}| \le \delta$ such that $\Lambda(P, \mathbb{R} \setminus \mathcal{J}) = 0$.

The following are simplified versions of Galkowski *et al.* (2023*a*, Theorems 1.5 and 1.6). Both results involve a quantity $\theta_0(P, \mathcal{J}, R_{\text{tr}})$ defined in Galkowski *et al.* (2023*a*, equation 1.10). For our purposes, we only need that $\theta_0(P, \mathcal{J}, R_{\text{tr}}) < \pi/2$ with $\theta_0(P, \mathcal{J}, R_{\text{tr}}) = 0$ if $\Lambda(P, \mathbb{R} \setminus \mathcal{J}) = 0$.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Theorem 4.3 (The PML error decreases exponentially in k, the PML width and the scaling angle). Let $\mathcal{J} \subset \mathbb{R}$ and suppose that P is such that $\Lambda(P, \mathbb{R} \setminus \mathcal{J}) < \infty$. Let $\chi \in C_c^\infty(B(0, R_{\mathrm{PML}}))$ with $\chi \equiv 1$ in a neighbourhood of $B(0, R_{\mathrm{scat}})$, and $\epsilon > 0$. Then there are $C, c, k_0 > 0$ such that for all $R_{\mathrm{tr}} > R_{\mathrm{PML}} + \epsilon$, Ω_{tr} with $B(0, R_{\mathrm{tr}}) \subset \Omega_{\mathrm{tr}} \subset \mathbb{R}^d$ and Lipschitz boundary, $\theta_0(P, \mathcal{J}, R_{\mathrm{tr}}) + \epsilon < \theta < \pi/2 - \epsilon$, $f \in L^2$, $k > k_0$ and $k \notin \mathcal{J}$, the solution v of $P_\theta v = \chi f$ exists, is unique, and satisfies

$$\begin{aligned}&\|\chi(u - v)\|_{H_k^1} + \|(1 - \chi)(u - v)\|_{H_k^2(B(0,R_{\mathrm{PML}}))} \\ &\quad \leq C \exp(-kc(R_{\mathrm{tr}} - R_1 + \epsilon) \tan\theta - 2\Lambda(P, \mathbb{R} \setminus \mathcal{J}))\|\chi f\|_{L^2},\end{aligned}$$

where u is the solution of $Pu = \chi f$.

Theorem 4.4 (The PML problem inherits the solution-operator bound of the scattering problem). Let $J \subset \mathbb{R}$ and suppose that P is such that $\Lambda(P, J) < \infty$. Let $\chi \in C_c^\infty(B(0, R_{\mathrm{PML}}))$ with $\chi \equiv 1$ in a neighbourhood of $B(0, R_{\mathrm{scat}})$, and $\epsilon > 0$. Then there are $C, k_0 > 0$ such that the following holds. For all $R_{\mathrm{tr}} > R_1 + \epsilon$, Ω_{tr} with $B(0, R_{\mathrm{tr}}) \subset \Omega_{\mathrm{tr}} \subset \mathbb{R}^d$ and Lipschitz boundary, $\theta_0(P, J, R_{\mathrm{tr}}) + \epsilon < \theta < \pi/2 - \epsilon$, all $f \in L^2$ with $\operatorname{supp} f \subset \Omega_{\mathrm{tr}}$, all $k > k_0$ and $k \notin \mathcal{J}$, the solution v to $P_\theta v = f$ exists, is unique, and satisfies

$$\|v\|_{H_k^1(\Omega_{\mathrm{tr}})} \leq C\|\chi\mathcal{R}(k)\chi\|_{L^2 \to L^2}\|f\|_{L^2}. \tag{4.11}$$

Remark 4.5 (Results on the accuracy of PML truncation at fixed k). Theorem 4.3 shows that for a PML with any positive width, the PML solution exists and is unique for sufficiently large k and sufficiently large θ, with θ only needing to be > 0 for non-trapping problems.

For a fixed k, the PML solution exists and is unique when the width is sufficiently large (see Lassas and Somersalo 1998, Theorem 2.1, Lassas and Somersalo 2001, Theorem A and Hohage *et al.* 2003, Theorem 5.8), and the H_k^1-errors decrease exponentially in the width (see Lassas and Somersalo 1998, Theorem 2.1, Lassas and Somersalo 2001, Theorem A and Hohage *et al.* 2003, Theorem 5.8).

4.6. Definition of the Helmholtz PML problem considered in this article

We collect here the definitions of the Helmholtz PML problem considered elsewhere in the article.

We write $u \in \overline{C^{\ell,1}}(\Omega)$ if, for all connected components Ω_j of Ω, there exists $U_j \in C^{\ell,1}(\mathbb{R}^d)$ such that $U_j|_{\Omega_j} = u|_{\Omega_j}$ (i.e. u is $C^{\ell,1}$ 'up to the boundary').

Definition 4.6 (H^ℓ-Helmholtz problem). The problem (4.1) is an *H^ℓ-Helmholtz problem* if Ω_- is $C^{\ell-1,1}$, and there exists $\Gamma_{\mathrm{p}} \Subset \Omega_+$ a $C^{\ell-1,1}$ closed, embedded hypersurface such that $A, n \in \overline{C^{\ell-2,1}}(\Omega_+ \setminus \Gamma_{\mathrm{p}})$. The problem (4.1) is an H^∞-Helmholtz problem if it is an H^ℓ-Helmholtz problem for all ℓ.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

The whole point of this notation is that if (4.1) is an H^ℓ*-Helmholtz problem* then, by elliptic regularity, $u \in H^\ell(\Omega_+ \setminus \Gamma_{\rm p})$; see e.g. McLean (2000, Theorems 4.18 and 4.20).

Definition 4.7 (H^ℓ-Helmholtz problem truncated using a PML). An H^ℓ-Helmholtz problem is truncated using a $C^{n,1}$ PML with a $C^{r,1}$ truncation boundary if $g_\theta \in C^{n,1}$ and $\Gamma_{\rm tr} \in C^{r,1}$ and, for the formulation corresponding to multiplying by $\alpha\beta^{d-1}$, $g_\theta(r)/r$ is non-decreasing.

5. A primer on semiclassical analysis

5.1. Symbols, quantization and semiclassical pseudodifferential operators on $\mathbb{R}^d$

It is standard in the analysis of partial differential equations (PDEs) to use the position variables, $x \in \mathbb{R}^d$, to decompose the solution of a partial differential equation into component parts localized at different positions (via e.g. a partition of unity). Similarly, Fourier analysis allows us to do the same in the momentum/frequency variables, $\xi \in \mathbb{R}^d$. Microlocal analysis – or, more specifically, since we work in the high-frequency regime, semiclassical analysis – allows us to do both of these things simultaneously, up to the limit imposed by the Heisenberg uncertainty principle.[1]

As described in Section 2.2, the combination of physical and Fourier variables is known as *phase space*. In this section we consider $x \in \mathbb{R}^d$, and treat phase space as $\mathbb{R}^{2d}$; strictly speaking $(x,\xi) \in T^*\mathbb{R}^d := \mathbb{R}^d \times (\mathbb{R}^d)^*$, i.e. $\mathbb{R}^d$ times its dual, but this distinction is not important in what follows. We work with the small parameter $\hbar = k^{-1}$; typically in SCA the small parameter is denoted by h, but we instead use $\hbar$ to distinguish it from the FEM/BEM meshwidth.

The operators that allow us to work microlocally are 'quantizations' of symbols, where a symbol is a function of (x,ξ); this terminology comes from quantum mechanics, where the symbols are classical observables (i.e. functions of position and momentum) and their quantizations are quantum measurements (i.e. operators acting on wavefunctions).

We work with quantizations of symbols from the following classes: given $m \in \mathbb{R}$,

$$S^m(\mathbb{R}^{2d}) := \left\{ a \in C^\infty(\mathbb{R}^{2d}) \colon \forall \alpha,\beta \in \mathbb{N}^d \sup_{\substack{(x,\xi)\in\mathbb{R}^{2d}\\ 0<\hbar<1}} \langle\xi\rangle^{-m+|\beta|} |\partial_x^\alpha \partial_\xi^\beta a(x,\xi)| < \infty \right\},$$

where $\langle\xi\rangle := (1+|\xi|^2)^{1/2}$. That is, for a symbol in S^m, taking a derivative in ξ results in one power of $\langle\xi\rangle$ faster decay in ξ, but taking a derivative in x does not change the decay in ξ. Note that we allow the symbols $a(x,\xi)$ to depend on $\hbar$ but do not write this dependence explicitly in the notation.

[1] The mathematical formulation of this principle is actually not important in the rest of this article, but can be found in e.g. Zworski (2012, Theorem 3.9).

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Example 5.1 (Examples of symbols).

(i) $a(x,\xi) := \sum_{|\alpha|\le m} a_\alpha(x)\xi^\alpha$, where $m \in \mathbb{N}$, $a_\alpha \in C^\infty$ and $\partial^\gamma a_\alpha \in L^\infty$ for all γ and α, is in S^m.

(ii) $\langle\xi\rangle^{-m} := (1+|\xi|^2)^{-m/2} \in S^{-m}$.

(iii) If $\chi \in C_c^\infty(\mathbb{R}^{2d})$, then $\chi \in S^{-N}$ for every $N \ge 1$.

Given a symbol $a(x,\xi)$, the most basic quantization map defines an operator for each $\hbar > 0$ by

$$\begin{aligned}(\widetilde{\mathrm{Op}_\hbar}(a)u)(x) &:= (2\pi\hbar)^{-d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{\frac{i}{\hbar}\langle x-y,\xi\rangle}a(x,\xi)u(y)\,dy\,d\xi \\ &= (2\pi\hbar)^{-d}\int_{\mathbb{R}^d} e^{\frac{i}{\hbar}\langle x,\xi\rangle}a(x,\xi)(\mathcal{F}_\hbar u)(\xi)\,d\xi \end{aligned} \tag{5.1}$$

for $x \in \mathbb{R}^d$ and $u \in \mathcal{S}(\mathbb{R}^d)$, with $\widetilde{\mathrm{Op}_\hbar}(a)$ acting on $\mathcal{S}'(\mathbb{R}^d)$ then defined by duality.

Example 5.2 (Examples of quantizations).

(i) If $a(x,\xi) = 1$, then $(\widetilde{\mathrm{Op}_\hbar}(a)v)(x) = v(x)$, i.e. $\widetilde{\mathrm{Op}_\hbar}(1) = I$.

(ii) If $a(x,\xi) = a(x)$, then $(\widetilde{\mathrm{Op}_\hbar}(a)v)(x) = a(x)v(x)$.

(iii) If $a(x,\xi) = a(\xi)$, then $(\widetilde{\mathrm{Op}_\hbar}(a)v)(x) = \mathcal{F}_\hbar^{-1}(a(\cdot)(\mathcal{F}_\hbar v)(\cdot))(x)$, i.e. $\widetilde{\mathrm{Op}_\hbar}(a)$ is a Fourier multiplier.

(iv) If $a(x,\xi) := \sum_{|\alpha|\le m} a_\alpha(x)\xi^\alpha$, where $m \in \mathbb{N}$, $a_\alpha \in C^\infty$, then $(\widetilde{\mathrm{Op}_\hbar}(a)v)(x) = \sum_{|\alpha|\le m} a_\alpha(x)(\hbar D)^\alpha v$, where $D := -i\partial$.

As special cases of (iv), we see that $-\hbar^2\Delta - 1 = \widetilde{\mathrm{Op}_\hbar}(|\xi|^2 - 1)$ and

$$P := -\hbar^2\sum_{i,j=1}^d \partial_{x^i}A^{ij}(x)\partial_{x^j} - n(x) = \widetilde{\mathrm{Op}_\hbar}(\langle A\xi,\xi\rangle - i\hbar\xi_j\partial_{x^i}A^{ij}(x) - n(x)). \tag{5.2}$$

Unfortunately, $\widetilde{\mathrm{Op}_\hbar}(a)$ does not have the property that it is *properly supported*; recall that an operator B is properly supported if and only if, for any $\chi \in C_c^\infty(\mathbb{R}^d)$, there exist $\chi_1, \chi_2 \in C_c^\infty(\mathbb{R}^d)$ such that

$$\chi B = \chi B\chi_1 \quad \text{and} \quad B\chi = \chi_2 B\chi.$$

Having the quantization be properly supported is helpful when dealing with functions that are only locally in L^2 (such as Helmholtz solutions). We therefore let $\psi \in C_c^\infty(\mathbb{R})$ with $0 \notin \mathrm{supp}(1-\psi)$ and use the quantization map

$$(\mathrm{Op}_\hbar(a)u)(x) := (2\pi\hbar)^{-d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{\frac{i}{\hbar}\langle x-y,\xi\rangle}a(x,\xi)\psi(|x-y|)u(y)\,dy\,d\xi. \tag{5.3}$$

With this quantization, parts (i), (ii) and (iv) of Example 5.2 still hold. Moreover, by repeatedly integrating by parts the integral defining $\widetilde{\mathrm{Op}_\hbar}(a) - \mathrm{Op}_\hbar(a)$, we find

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

that for all $a \in S^m(\mathbb{R}^{2d})$ and all $N > 0$, there exists $C > 0$ such that for $0 < \hbar < 1$,

$$\|\widetilde{\mathrm{Op}_\hbar}(a) - \mathrm{Op}_\hbar(a)\|_{H_k^{-N}(\mathbb{R}^d)\to H_k^N(\mathbb{R}^d)} \leq C\hbar^N, \tag{5.4}$$

where the spaces $H_k^s(\mathbb{R}^d)$ are defined in Section 5.2 below.

We say that an ($\hbar$-dependent) operator $B\colon C_c^\infty(\mathbb{R}^d) \to \mathcal{D}'(\mathbb{R}^d)$ is $O(\hbar^\infty)_{\Psi_h^{-\infty}}$ if, for all $N > 0$, there exists $C > 0$ such that

$$\|B\|_{H_k^{-N}(\mathbb{R}^d)\to H_k^N(\mathbb{R}^d)} \leq C\hbar^N. \tag{5.5}$$

In particular, $\widetilde{\mathrm{Op}_\hbar}(a) = \mathrm{Op}_\hbar(a) + O(\hbar^\infty)_{\Psi_h^{-\infty}}$ by (5.4).

We then define the class of semiclassical pseudodifferential operator of order m, $\Psi_\hbar^m(\mathbb{R}^d)$, by

$$\Psi_\hbar^m(\mathbb{R}^d) := \left\{\mathrm{Op}_\hbar(a) + O(\hbar^\infty)_{\Psi_\hbar^{-\infty}} \colon a \in S^m(\mathbb{R}^{2d})\right\} \tag{5.6}$$

(without the inclusion of the residual term in (5.6), this class of operators cannot contain, for example, $(-\hbar^2\Delta + 1)^{-1}$, since this operator is not properly supported).

Remark 5.3 (Why do we work in S^m?). S^m – known as the Kohn–Nirenberg symbol class (Kohn and Nirenberg 1965) – is a natural class of symbols that includes both polynomials and, when they are invertible, the inverse of polynomials. This results in the fact that the quantization of S^m includes (up to residual terms) both the (semiclassical) differential operators and, when the differential operator is elliptic in an appropriate sense, the inverse of such a differential operator.

Much wider classes of (smooth) symbols exist; see e.g. Taylor (1996, Chapter 7, §1), Zworski (2012, §4.4.1) and Dyatlov and Zworski (2019, equation E.1.48).

Remark 5.4 (Other symbol classes). There are many calculi of (semiclassical) pseudodifferential operators that allow for the treatment of various geometric singularities; e.g. boundaries, corners, cone points, etc. In this article we discuss only the standard semiclassical calculus, although we use some results from the calculus associated with boundaries in Section 9 (when stating the analogues of Theorems 5.23 and 5.24 below for the PML problem: see Table 9.1 below).

In general, non-smooth symbol classes are more technically complicated and have worse remainder terms than smooth symbol classes; see e.g. Taylor (2023, §13.9).

5.2. Semiclassical Sobolev spaces

For $s \geq 0$, let

$$|u|^2_{H_k^s(\mathbb{R}^d)} := \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d} |\xi|^{2s}|\mathcal{F}_k(u)(\xi)|^2\, d\xi$$

(where $\mathcal{F}_k$ – the semiclassical Fourier transform – is defined by (2.1) above) and

$$\|u\|^2_{H_k^s(\mathbb{R}^d)} := \sum_{\ell=0}^{s} \frac{1}{(\ell!)^2}|u|^2_{H_k^\ell(\mathbb{R}^d)} \tag{5.7}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

(compare to (3.2)). Let $H_k^{-s}(\mathbb{R}^d) := (H_k^s(\mathbb{R}^d))^*$, with the corresponding norm defined by duality.

Given $s_0 > 0$, for all $|s| \leq s_0$ these norms are then equivalent – up to s_0- and d-dependent constants – to the more standard Sobolev norms on $\mathbb{R}^d$ defined by the Fourier transform:

$$|||u|||^2_{H_k^s(\mathbb{R}^d)} := (2\pi\hbar)^{-d} \int_{\mathbb{R}^d} \langle \xi \rangle^{2s} |\mathcal{F}_\hbar u(\xi)|^2 \, d\xi. \tag{5.8}$$

Throughout this article we work with the definition (3.2)/(5.7), but the precise values of the constants in the norm are only important for the results on h- and p-explicit polynomial approximation in Appendix A (and thus, in particular, not important in this section).

5.3. *Fundamental properties of semiclassical pseudodifferential operators*

Informal Theorem 5.5 (Properties of semiclassical pseudodifferential calculus).

(i) *Quantization.* The *quantization* map

$$\mathrm{Op}_\hbar \colon S^m(\mathbb{R}^{2d}) \to \Psi_\hbar^m(\mathbb{R}^d)$$

defined by (5.3) is such that $\mathrm{Op}_\hbar(a)$ is properly supported, $\mathrm{Op}_\hbar(1) = I$, and $[\mathrm{Op}_\hbar(a)u](x) = a(x)u(x)$ when a is independent of ξ.

(ii) *Principal symbol.* There is a *principal symbol* map

$$\sigma_\hbar \colon \Psi_\hbar^m(\mathbb{R}^d) \to S^m(\mathbb{R}^d)$$

such that

$$\sigma_\hbar \circ \mathrm{Op}_\hbar = I + \text{l.o.t.},$$

where 'l.o.t.' stands for 'lower-order terms' – both smaller in $\hbar$ and more smoothing.

(iii) *Boundedness in Sobolev norms.* Given $a \in S^m(\mathbb{R}^{2d})$, for all $s \in \mathbb{R}$ there exists $C_s > 0$ such that

$$\|\mathrm{Op}_\hbar(a)\|_{H_k^s(\mathbb{R}^d) \to H_k^{s-m}(\mathbb{R}^d)} \leq C_s,$$

that is, the quantization of a symbol of order m decreases Sobolev regularity by m. Furthermore, for $a \in S^0(\mathbb{R}^{2d})$,

$$\|\mathrm{Op}_\hbar(a)\|_{L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)} \leq \sup_{(x,\xi) \in \mathbb{R}^{2d}} |a(x,\xi)| + \text{l.o.t.}$$

(iv) *Composition.*

$$\mathrm{Op}_\hbar(a)\,\mathrm{Op}_\hbar(b) = \mathrm{Op}_\hbar(ab) + \text{l.o.t.}$$

Furthermore, $\mathrm{Op}_\hbar(a)\,\mathrm{Op}_\hbar(b) = \mathrm{Op}_\hbar(c) + O(\hbar^\infty)_{\Psi_\hbar^{-\infty}}$ with $\operatorname{supp} c \subset \operatorname{supp} a \cap \operatorname{supp} b$.

(v) *Adjoints.* $(\mathrm{Op}_\hbar(a))^* = \mathrm{Op}_\hbar(\overline{a}) + \text{l.o.t.}$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

(vi) *Commutators.*

$$i\hbar^{-1}[\mathrm{Op}_\hbar(a), \mathrm{Op}_\hbar(b)] := i\hbar^{-1}(\mathrm{Op}_\hbar(a)\,\mathrm{Op}_\hbar(b) - \mathrm{Op}_\hbar(b)\,\mathrm{Op}_\hbar(a)) = \mathrm{Op}_\hbar(\{a, b\}) + \text{l.o.t.},$$

where the Poisson bracket $\{\cdot, \cdot\}$ is defined by

$$\{a, b\} := \langle \partial_\xi a, \partial_x b\rangle - \langle \partial_\xi b, \partial_x a\rangle.$$

(vii) *Sharp Gårding inequality.* If $a \in S^m(\mathbb{R}^{2d})$, $w \in S^{m/2}(\mathbb{R}^{2d})$ and $a \geq w^2$, then for all $u \in H_k^{m/2}(\mathbb{R}^d)$,

$$\langle \mathrm{Op}_\hbar(a)u, u\rangle \geq \|\mathrm{Op}_\hbar(w)u\|^2_{L^2(\mathbb{R}^d)} + \text{l.o.t.} \tag{5.9}$$

The main way in which Informal Theorem 5.5 is non-rigorous is that we have not been precise about what the lower-order terms are. The precise statement of these lower-order terms is given in Theorem C.1 below. However, we emphasize that the sketch proofs in the rest of this section can be understood using only Informal Theorem 5.5.

As an example of point (ii) of Informal Theorem 5.5,

$$p(x, \xi) := \sigma_\hbar(P)(x, \xi) = \langle A\xi, \xi\rangle - n(x) = \sum_{i,j=1}^{d} A^{ij}(x)\xi_i\xi_j - n(x), \tag{5.10}$$

that is, the term on the right-hand side of (5.2) involving $\hbar$ has been neglected.

Remark 5.6 (Pseudodifferential operators on a manifold). In Section 8 we use pseudodifferential operators defined on the boundary of Ω_- (with the boundary assumed smooth). We remark that the analogue of Informal Theorem 5.5 (and its precise version Theorem C.1) holds on any smooth, compact manifold, M, with $\mathbb{R}^{2d}$ replaced by the cotangent bundle T^*M, and $\mathbb{R}^d$ replaced by M (Dyatlov and Zworski 2019, Appendix E), except that now the quantization map is defined by (5.3) on coordinate charts (see Dyatlov and Zworski 2019, equation E.1.31).

5.4. Ellipticity

Ellipticity is the notion of 'microlocal invertibility': at elliptic points, the solution operator for P is very well-behaved, in that it both has good estimates and is (almost) local.

From now on, to keep the notation concise, we abbreviate $\|\cdot\|_{H_k^s(\mathbb{R}^d)}$ to $\|\cdot\|_{H_k^s}$ and $\|\cdot\|_{L^2(\mathbb{R}^d)}$ to $\|\cdot\|_{L^2}$.

The next result involves an arbitrary pseudodifferential operator $Q = \mathrm{Op}_\hbar(q)$. We emphasize that in almost all applications of this result in this article, Q is taken to be either P (5.2) or P^*.

Theorem 5.7 (Elliptic estimate). Let $m_1, m_2 \in \mathbb{R}$ and let $q \in S^{m_1}(\mathbb{R}^{2d})$, $a \in S^{m_2}(\mathbb{R}^{2d})$ and $b \in S^0(\mathbb{R}^{2d})$ be such that there exists $c > 0$ such that

$$|q(x, \xi)| \geq c\langle\xi\rangle^{m_1}, \quad (x, \xi) \in \mathrm{supp}\, a, \tag{5.11}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

and $\operatorname{supp} a \cap \operatorname{supp}(1-b) = \emptyset$. Then, for all $\chi \in C_c^\infty(\mathbb{R}^d)$, there are $e \in S^{m_2-m_1}(\mathbb{R}^{2d})$, $\psi \in C_c^\infty(\mathbb{R}^d)$ such that for all $N > 0$, there exists $C > 0$ such that for $0 < \hbar < 1$ and $u \in \mathcal{D}'(\mathbb{R}^d)$,

$$\|\chi(\operatorname{Op}_\hbar(a) - \operatorname{Op}_\hbar(e)\operatorname{Op}_\hbar(b)\operatorname{Op}_\hbar(q))u\|_{H_k^N} \le C_N \hbar^N \|\psi u\|_{H_k^{-N}}. \tag{5.12}$$

In particular,

$$\|\chi \operatorname{Op}_\hbar(a)u\|_{H_k^s} \le C\|\operatorname{Op}_\hbar(b)\operatorname{Op}_\hbar(q)u\|_{H_k^{s+m_2-m_1}} + C\hbar^N \|\psi u\|_{H_k^{-N}}. \tag{5.13}$$

Sketch of proof.

(1) By (5.11), $e := a/q \in S^{m_2-m_1}$ and $a/q = ab/q$. Hence, by the composition (iv) and principal symbol properties (ii) from Informal Theorem 5.5,

$$\operatorname{Op}_\hbar(a) = \operatorname{Op}_\hbar\left(\tfrac{a}{q}bq\right) + \text{l.o.t.} = \operatorname{Op}_\hbar\left(\tfrac{a}{q}\right)\operatorname{Op}_\hbar(b)\operatorname{Op}_\hbar(q) + \text{l.o.t.}$$

The result (5.12) then follows by iterating this argument (to improve the regularity and size of the remainder term) and using the fact that

$$\begin{aligned}&\chi\left(\operatorname{Op}_\hbar\left(\tfrac{a}{q}bq\right) - \operatorname{Op}_\hbar\left(\tfrac{a}{q}\right)\operatorname{Op}_\hbar(b)\operatorname{Op}_\hbar(q)\right)\\ &\quad = \chi\left(\operatorname{Op}_\hbar\left(\tfrac{a}{q}bq\right) - \operatorname{Op}_\hbar\left(\tfrac{a}{q}\right)\operatorname{Op}_\hbar(b)\operatorname{Op}_\hbar(q)\right)\psi,\end{aligned}$$

for some ψ, since the quantization is properly supported.

(2) The bound (5.13) follows from (5.12) since

$$\|\operatorname{Op}_\hbar(e)\|_{H_k^{s+m_2-m_1} \to H_k^s} \le C$$

by the boundedness property (iii) in Informal Theorem 5.5. □

Example 5.8. The prototypical semiclassically elliptic differential operator is $\widetilde{P} := -\hbar^2\Delta + 1$. Indeed, $\widetilde{P} = \operatorname{Op}_\hbar(\widetilde{p})$ with $\widetilde{p}(x,\xi) = |\xi|^2 + 1 = \langle\xi\rangle^2$. Since this operator is a Fourier multiplier, directly by the Fourier transform,

$$\|u\|_{H_k^s} \le C\|\widetilde{P}u\|_{H_k^{s-2}},$$

which is analogous to the elliptic estimate (5.13) with $a = b = 1$, no cut-offs, no remainder term, and $\operatorname{Op}_\hbar(q) = \widetilde{P}$.

5.5. *The Hamiltonian flow*

In Section 2.2 we stated that SCA describes how Helmholtz solutions in the limit $k \to \infty$ are governed by the rays. The rays appear via the commutator property (vi) of Informal Theorem 5.5, which implies that

$$i\hbar^{-1}[P, \operatorname{Op}_\hbar(a)] = \operatorname{Op}_\hbar(H_p a) + \text{l.o.t.}, \tag{5.14}$$

where l.o.t. stands for lower-order terms and

$$H_p a := \{p, a\} = \langle \partial_\xi p(x,\xi), \partial_x a\rangle - \langle \partial_x p(x,\xi), \partial_\xi a\rangle.$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

The key point is that

$$\begin{aligned} &H_p a \text{ is the derivative of } a(x,\xi) \text{ when } (x,\xi) \text{ evolve according} \\ &\text{to Hamilton's equations with Hamiltonian } p(x,\xi). \end{aligned} \tag{5.15}$$

To see this, recall that given a function $\mathscr{H}(x,\xi)$ (the *Hamiltonian*), Hamilton's equations are

$$\frac{dx_i}{dt}(t) = \frac{\partial}{\partial \xi_i}\mathscr{H}(x(t),\xi(t)), \quad \frac{d\xi_i}{dt}(t) = -\frac{\partial}{\partial x_i}\mathscr{H}(x(t),\xi(t)). \tag{5.16}$$

By the chain rule and the definition of the Poisson bracket $\{\cdot,\cdot\}$,

$$\frac{d}{dt}(f(x(t),\xi(t);t)) = \left(\{\mathscr{H},f\} + \frac{\partial f}{\partial t}\right)(x(t),\xi(t);t). \tag{5.17}$$

Given (x_0,ξ_0), we define the Hamiltonian flow

$$\varphi_t(x_0,\xi_0) := (x(t),\xi(t)), \tag{5.18}$$

where $(x(t),\xi(t))$ is the solution of (5.16) with initial condition $(x(0),\xi(0)) = (x_0,\xi_0)$. (Another notation for φ_t is $\exp(tH_p)$.) With this notation, given $a(x,\xi)$, (5.17) can be rewritten as

$$\frac{d}{dt}(a\circ\varphi_t) = \{\mathscr{H},a\}\circ\varphi_t. \tag{5.19}$$

When $\mathscr{H} = p$, (5.19) therefore justifies the statement (5.15).

When P is given by (5.2) with $A \equiv I$ and $n \equiv 1$, by (5.10) $p(x,\xi) = |\xi|^2 - 1$, and Hamilton's equations (5.16) become

$$\dot{x}_i = 2\xi_i, \quad \dot{\xi}_i = 0,$$

with solution

$$x = x_0 + 2t\xi_0, \quad \xi = \xi_0,$$

i.e. straight-line motion with speed $2\xi_0$. We see below that special importance is played by the set $\Sigma := \{p = 0\}$; on this set $|\xi_0| = 1$ so that the flow restricted to Σ has speed 2.

Remark 5.9 (Generalized broken bicharacteristics). When the domain has a boundary, the Hamiltonian trajectories need to be replaced by *generalized broken bicharacteristics* (Hörmander 1985, Section 24.3), which, roughly speaking, extend trajectories that intersect the boundary by the Snell–Descartes law of reflection. When $A \equiv I$ and $n \equiv 1$, these are the usual billiard trajectories, except possibly when the trajectory is tangent to a point on the boundary that is flat to infinite order.

5.6. Propagation

5.6.1. From energy estimates to commutators

In PDE theory, *energy estimates* usually refer to bounds obtained by multiplying the PDE by a combination of the solution and its derivatives and integrating by

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

parts. If P is self-adjoint and V is a vector field, then

$$2\mathrm{Re}\,\langle Pu,(V-V^*)u\rangle = \langle (V-V^*)u,Pu\rangle + \langle Pu,(V-V^*)u\rangle = \langle [P,V-V^*]u,u\rangle, \tag{5.20}$$

so that if

$$[P,V-V^*] \geq c(V-V^*)^2, \tag{5.21}$$

then one can bound $(V-V^*)u$ in terms of Pu. In other words, standard arguments used to obtain energy estimates can/should be thought of as *positive commutator arguments*.

For example, if $V = x\cdot\nabla + d/2$ then $V^* = -x\cdot\nabla - d/2$, so that $V - V^* = 2x\cdot\nabla$, and then $[V-V^*,\Delta] = 2[x\cdot\nabla,\Delta] = -4\Delta$ and (5.21) is satisfied; this is the basis of so-called Rellich–Morawetz identities (Rellich 1940, Morawetz and Ludwig 1968, Morawetz 1975), which have been the main way the numerical analysis community has obtained propagation information about Helmholtz solutions (see Remark 5.22 below for more discussion on this).

In semiclassical and microlocal analysis, propagation estimates can be thought of as 'microlocalized energy estimates', where one multiplies by the quantization of certain carefully chosen symbols. If A and P are both self-adjoint, then

$$-2i\mathrm{Im}\,\langle Pu,Au\rangle = (\langle Au,Pu\rangle - \langle Pu,Au\rangle) = \langle [P,A]u,u\rangle. \tag{5.22}$$

By the key relation (5.14), the commutator term in (5.22) has a sign (up to l.o.t.) provided that the symbol $a = \sigma_\hbar(A)$ either increases or decreases along the flow with Hamiltonian $p = \sigma_\hbar(P)$.

5.6.2. *The basic propagation argument*

The elliptic estimate from Theorem 5.7 shows that Helmholtz solutions are well-behaved at high frequency (since P is semiclassically elliptic there). For simplicity, therefore, we only consider propagation estimates with compactly supported symbols (i.e. compact support in both space and frequency). For propagation estimates with more general symbols, see Dyatlov and Zworski (2019, Appendix E.4).

Theorem 5.10 (Basic propagation estimate). Suppose that P is given by (5.2) and let $\varphi_t := \exp(tH_p)$ be the Hamiltonian flow of $p = \sigma_\hbar(P)$ (given by (5.10)). Let $T\in\mathbb{R}$, $(x_0,\xi_0)\in\mathbb{R}^{2d}$ and $(y_0,\eta_0) := \varphi_{-T}(x_0,\xi_0)$. Let U be a neighbourhood of (y_0,η_0) and U_1 be a neighbourhood of

$$\bigcup_{t=0}^{T}\varphi_{-t}(x_0,\xi_0).$$

Then there is a neighbourhood, V, of (x_0,ξ_0) and $\chi\in C_c^\infty(\mathbb{R}^d)$ such that for all $a\in C_c^\infty(V;\mathbb{R})$ and $b, b_1\in C_c^\infty(\mathbb{R}^{2d})$ with $\mathrm{supp}(1-b)\cap U = \emptyset$, $\mathrm{supp}(1-b_1)\cap U_1 = \emptyset$ and all $N>0$, there exists $C>0$ such that for $0<\hbar<1$ and for all $u\in\mathcal{D}'(\mathbb{R}^d)$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

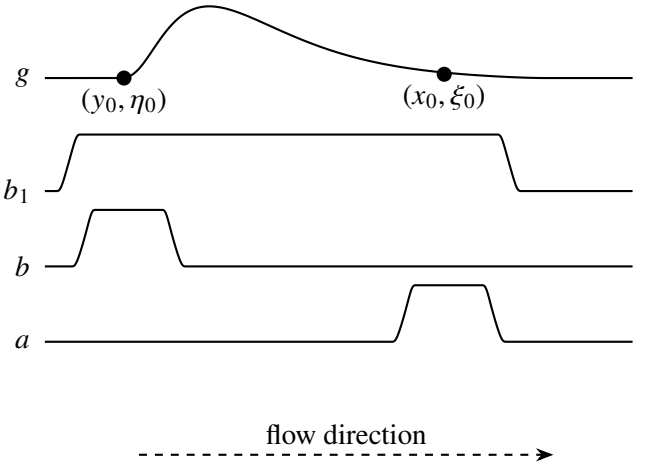


Figure 5.1. Profile of the escape function, g, the control function, b, and the observation function, a, along the flow φ_t used in the propagation estimate of Theorem 5.10 (when the assumptions hold with $T > 0$).

with $\mathrm{Op}_\hbar(b_1)Pu \in L^2(\mathbb{R}^d)$

$$\|\mathrm{Op}_\hbar(a)u\|_{L^2} \leq C\hbar^{-1}\|\mathrm{Op}_\hbar(b_1)Pu\|_{L^2} + C\|\mathrm{Op}_\hbar(b)u\|_{L^2} + C\hbar^N\|\chi u\|_{H_k^{-N}}. \quad (5.23)$$

Remark 5.11 (Direction of propagation). Theorem 5.10 involves propagation in both directions; this is possible since $\mathrm{Im}\, p = 0$. More generally, if $\mathrm{Im}\,\sigma_\hbar(P) \leq 0$ then one can propagate forwards (i.e. $T \geq 0$) under the Hamiltonian flow defined by $\mathrm{Re}\,\sigma_\hbar(P)$, and if $\mathrm{Im}\,\sigma_\hbar(P) \geq 0$ then one can propagate backwards under this flow (i.e. $T \leq 0$).

Sketch proof of Theorem 5.10.

(1) Without loss of generality, we assume that $T \geq 0$. If $T < 0$, then we can apply the proof below to $-P$, noting that the backward flow for $-P$ equals the forward flow for P.

By Dyatlov and Zworski (2019, Lemma E.48), for example, there exists an 'escape function' $g \in C_c^\infty(\mathbb{R}^{2d}; \mathbb{R})$ such that there exists $c > 0$ satisfying

$$g \geq 0, \quad H_p g^2 \leq -cg^2 + b^2, \quad g^2 \geq ca^2, \quad \mathrm{supp}\, g \subset U_1 \quad (5.24)$$

that is, g is decreasing along the flow apart from on the support of b and is supported where b_1 is 1. Figure 5.1 gives a sketch of the functions a, b, b_1 and g in the flow direction.

(2) Let $G := \mathrm{Op}_\hbar(g)$. By (5.14),

$$i\hbar^{-1}[P, G^*G] = \mathrm{Op}_\hbar(H_p g^2) + \text{l.o.t.}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

and, by the self-adjointness of P,

$$2\hbar^{-1}\mathrm{Im}\,\langle Pu, G^*Gu\rangle = i\hbar^{-1}(\langle G^*Gu, Pu\rangle - \langle Pu, G^*Gu\rangle)$$
$$= i\hbar^{-1}\langle [P, G^*G]u, u\rangle = \langle \mathrm{Op}_\hbar(H_p g^2)u, u\rangle + \text{l.o.t.}, \quad (5.25)$$

where the boundary terms in the integration by parts vanish since G is compactly supported.

(3) By the sharp Gårding inequality (5.9) (with $w = 0$) together with (5.24),

$$\langle \mathrm{Op}_\hbar(b^2 - cg^2 - H_p g^2)u, u\rangle \geq \text{l.o.t.}$$

Thus, by the adjoint (v) and composition properties (iv) of Informal Theorem 5.5,

$$\langle \mathrm{Op}_\hbar(H_p g^2)u, u\rangle \leq -c\|Gu\|_{L^2}^2 + \|\mathrm{Op}_\hbar(b)u\|_{L^2}^2 + \text{l.o.t.} \quad (5.26)$$

(4) The combination of (5.25), (5.26) and the Cauchy–Schwarz inequality implies that

$$-2\hbar^{-1}\|GPu\|_{L^2}\|Gu\|_{L^2} \leq 2\hbar^{-1}\mathrm{Im}\,\langle Pu, G^*Gu\rangle$$
$$\leq -c\|Gu\|_{L^2}^2 + \|\mathrm{Op}_\hbar(b)u\|_{L^2}^2 + \text{l.o.t.}$$

(5) Combining the result of step (4) with the weighted Young's inequality,

$$2ab \leq \epsilon a^2 + \epsilon^{-1}b^2 \quad \text{for all } a, b, \epsilon > 0, \quad (5.27)$$

gives

$$\|Gu\|_{L^2}^2 \leq C\hbar^{-2}\|GPu\|_{L^2}^2 + C\|\mathrm{Op}_\hbar(b)u\|_{L^2}^2 + \text{l.o.t.} \quad (5.28)$$

(6) Since $g^2 \geq ca^2$, the sharp Gårding inequality (5.9) (with $w = 0$) implies that $\langle \mathrm{Op}_\hbar(g^2)u, u\rangle \geq c\langle \mathrm{Op}_\hbar(a^2)u, u\rangle + \text{l.o.t.}$, from which

$$c\|\mathrm{Op}_\hbar(a)u\|_{L^2}^2 \leq \|Gu\|_{L^2}^2 + \text{l.o.t.} \quad (5.29)$$

by the composition (iv) and adjoint properties (v) of Informal Theorem 5.5. The combination of (5.28) and (5.29) implies that

$$\|\mathrm{Op}_\hbar(a)u\|_{L^2} \leq C\hbar^{-1}\|GPu\|_{L^2} + C\|\mathrm{Op}_\hbar(b)u\|_{L^2} + \text{l.o.t.}$$

(7) Next, Theorem 5.7 and the support properties of g imply

$$\|GPu\|_{L^2} \leq \|\mathrm{Op}_\hbar(b_1)Pu\|_{L^2} + \text{l.o.t.}$$

and hence

$$\|\mathrm{Op}_\hbar(a)u\|_{L^2} \leq C\hbar^{-1}\|\mathrm{Op}_\hbar(b_1)Pu\|_{L^2} + C\|\mathrm{Op}_\hbar(b)u\|_{L^2} + \text{l.o.t.} \quad (5.30)$$

(compare to (5.23)).

(8) By iterating steps (1)–(6), one can make the remainder both smoother and smaller in terms of $\hbar$. Finally, the error term is compactly supported since a, b and b_1 are compactly supported. □

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

5.7. Wavefront set

For our treatment of **R5** (the result about the parallel overlapping Schwarz method) in Section 10, it is useful to have the notion of semiclassical wavefront set. We highlight that this notion is only used in Section 10 (for readers wanting to skip this concept initially).

Definition 5.12 ($\hbar$-tempered distributions). An $\hbar$-dependent family of distributions $v_\hbar \in \mathcal{D}'(\mathbb{R}^d)$ is *$\hbar$-tempered* if, for all $\chi \in C_c^\infty(\mathbb{R}^d)$, there exist $C, M > 0$ such that

$$\|\chi v_\hbar\|_{H_\hbar^{-M}(\mathbb{R}^d)} \leq C\hbar^{-M}.$$

The wavefront set of an $\hbar$-tempered distribution u, $\mathrm{WF}_\hbar(u)$, should be understood as the points in phase space at which u has non-trivial mass. One precise version of this is given in the following definition.

Definition 5.13 (Wavefront set). Suppose that u is $\hbar$-tempered. Then $(x_0, \xi_0) \in \mathbb{R}^{2d}$ is *not in the wavefront set of u*, written $(x_0, \xi_0) \notin \mathrm{WF}_\hbar(u)$, if there exists $a \in C_c^\infty(\mathbb{R}^{2d})$ such that $a(x_0, \xi_0) = 1$ and $M > 0$ such that

$$\|\mathrm{Op}_\hbar(a)u\|_{H_k^{-M}(\mathbb{R}^d)} = O(\hbar^\infty).$$

Example 5.14 (Simple examples of wavefront sets). By integration by parts and/or stationary phase,

(i) for $\xi_0 \in \mathbb{R}^d$ and $u(x) := \exp(i\hbar^{-1}\langle x, \xi_0\rangle)$,

$$\mathrm{WF}_\hbar(u) = \{(x, \xi_0) \colon x \in \mathbb{R}^d\},$$

(ii) for $(x_0, \xi_0) \in \mathbb{R}^{2d}$, and $v(x) := \exp\big(i\hbar^{-1}\langle x - x_0, \xi_0\rangle - \frac{1}{2}\hbar^{-1}|x - x_0|^2\big)$,

$$\mathrm{WF}_\hbar(v) = \{(x_0, \xi_0)\}.$$

Lemma 5.15 (Properties of $\mathrm{WF}_\hbar$). Let u be $\hbar$-tempered. If $(x_0, \xi_0) \notin \mathrm{WF}_\hbar(u)$, then there is a neighbourhood U of (x_0, ξ_0) such that for all $a \in C_c^\infty(U)$ and all $N > 0$,

$$\|\mathrm{Op}_\hbar(a)u\|_{H_k^N(\mathbb{R}^n)} = O(\hbar^\infty).$$

Observe that Lemma 5.15 implies that $\mathrm{WF}_\hbar(u)$ is closed.

Proof of Lemma 5.15. Let $(x_0, \xi_0) \notin \mathrm{WF}_\hbar(u)$. By the definition of $\mathrm{WF}_\hbar(u)$, there are $M \geq 0$, and $b \in C_c^\infty(\mathbb{R}^{2d})$ with $b(x_0, \xi_0) = 1$ such that

$$\|\mathrm{Op}_\hbar(b)u\|_{H_k^{-M}(\mathbb{R}^d)} = O(\hbar^\infty). \tag{5.31}$$

Let

$$U := \big\{(x, \xi) \colon |b(x, \xi)| > \tfrac{1}{2}\big\}.$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

By the bound (5.13) from Theorem 5.7, for any $a \in C_c^\infty(U)$, there are $e \in S^{-\infty}(\mathbb{R}^{2d})$ and $\chi_1, \chi_2 \in C_c^\infty(\mathbb{R}^d)$ such that

$$\chi_1 \operatorname{Op}_\hbar(a) = \operatorname{Op}_\hbar(a) \tag{5.32}$$

and

$$\|\chi_1 \operatorname{Op}_\hbar(a)u\|_{H_k^N} \leq \|\operatorname{Op}_\hbar(b)u\|_{H_k^{-N}} + C_N \hbar^N \|\chi u\|_{H_k^{-N}(\mathbb{R}^d)}. \tag{5.33}$$

The result then follows from combining (5.31), (5.32), (5.33), the boundedness property (iii) of Informal Theorem 5.5, and the fact that u is $\hbar$-tempered. □

We record the following corollaries of Theorems 5.7 and 5.10.

Corollary 5.16 (Ellipticity and the wavefront set). Suppose that $u \in \mathcal{D}'(\mathbb{R}^d)$ is $\hbar$-tempered, P is given by (5.2) and p is given by (5.10). Then

$$\operatorname{WF}_\hbar(u) \subset \{p = 0\} \cup \operatorname{WF}_\hbar(Pu).$$

Proof. It is sufficient to prove that if $p(x_0, \xi_0) \neq 0$ and $(x_0, \xi_0) \notin \operatorname{WF}_\hbar(Pu)$, then $(x_0, \xi_0) \notin \operatorname{WF}_\hbar(u)$. Since $(x_0, \xi_0) \notin \operatorname{WF}_\hbar(Pu)$, by the definition of $\operatorname{WF}_\hbar(Pu)$ there exists $b \in C_c^\infty(\mathbb{R}^{2d})$ with $b(x_0, \xi_0) = 1$ such that for all N,

$$\|\operatorname{Op}_\hbar(b)Pu\|_{H_k^N} = O(\hbar^\infty).$$

Furthermore, since $p(x_0, \xi_0) \neq 0$, there exists $a \in C_c^\infty(\mathbb{R}^{2d})$ with $a(x_0, \xi_0) = 1$ such that $\operatorname{supp} a \subset \{|p| > 0\} \cap \{|b| > 0\}$. Then, by Theorem 5.7, there are $\chi, \psi \in C_c^\infty(\mathbb{R}^d)$ with $\chi(x_0) = 1$ such that for all N,

$$\|\chi \operatorname{Op}_\hbar(a)\|_{L^2} \leq C\|\operatorname{Op}_\hbar(b)Pu\|_{H_k^{-N}} + C\hbar^N \|\psi u\|_{H_k^{-N}}.$$

Therefore, since u is $\hbar$-tempered,

$$\|\chi \operatorname{Op}_\hbar(a)\|_{L^2} = O(\hbar^\infty).$$

Finally, since $\sigma_\hbar(\chi \operatorname{Op}_\hbar(a)) = \chi a$ (by the composition (iv) and principal symbol properties (ii) of Informal Theorem 5.5) and $\chi(x_0)a(x_0, \xi_0) = 1$, $(x_0, \xi_0) \notin \operatorname{WF}_\hbar(u)$. □

Corollary 5.17 (Propagation and the wavefront set). Suppose that $u \in \mathcal{D}'(\mathbb{R}^d)$ is $\hbar$-tempered and P is as in (5.2). If there exists $T \in \mathbb{R}$ such that $\varphi_{-T}(x_0, \xi_0) \notin \operatorname{WF}_\hbar(u)$ and

$$\bigcup_{t=0}^{T} \varphi_{-t}(x_0, \xi_0) \cap \operatorname{WF}_\hbar(Pu) = \emptyset,$$

then $(x_0, \xi_0) \notin \operatorname{WF}_\hbar(u)$.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Proof. By Lemma 5.15 and compactness of $\bigcup_{t=0}^{T} \varphi_{-t}(x_0, \xi_0)$, there are $b, b_1 \in C_c^\infty(\mathbb{R}^{2d})$ such that $\varphi_{-T}(x_0, \xi_0) \notin \operatorname{supp}(1-b)$,

$$\bigcup_{t=0}^{T} \varphi_{-t}(x_0, \xi_0) \cap \operatorname{supp}(1-b_1) = \emptyset,$$

and

$$\|\mathrm{Op}_\hbar(b)u\|_{H_k^N} + \|\mathrm{Op}_\hbar(b_1)Pu\|_{H_k^N} = O(\hbar^\infty).$$

By Theorem 5.10, there are $a \in C_c^\infty(\mathbb{R}^{2d})$ and $\chi \in C_c^\infty(\mathbb{R}^d)$ such that $a(x_0, \xi_0) = 1$ and

$$\|\mathrm{Op}_\hbar(a)u\|_{L^2} \le C\hbar^{-1}\|\mathrm{Op}_\hbar(b_1)Pu\|_{L^2} + C\|\mathrm{Op}_\hbar(b)u\|_{L^2} + C\hbar^N\|\chi u\|_{H_k^{-N}}.$$

Therefore

$$\|\mathrm{Op}_\hbar(a)u\|_{L^2} = O(\hbar^\infty),$$

and the result follows from the definition of $\mathrm{WF}_\hbar(u)$. □

5.8. *Application of ellipticity and propagation to solutions of the Helmholtz equation*

5.8.1. *Outgoing solutions of the Helmholtz equation*

Solutions of the Helmholtz equation satisfying the Sommerfeld radiation condition are semiclassically outgoing; that is, far from the scatterer they are microlocally concentrated on trajectories travelling towards infinity. This fact is encoded in the following lemma via the support condition on a.

Lemma 5.18 (The semiclassical outgoing property). For all $R > 0$, there are $R' > R$, $c > 0$ such that for all $a \in S^0(\mathbb{R}^{2d})$ with

$$\operatorname{supp} a \subset \left\{\left\langle \frac{x}{|x|}, \xi \right\rangle \le c,\ |x| > R'\right\} \cup \left\{\big||\xi| - 1\big| > c > 0, |x| > R'\right\}, \tag{5.34}$$

and all $N > 0$, there are $\psi \in C_c^\infty(\mathbb{R}^d)$ and $C > 0$ such that if $u \in H^2_{\mathrm{loc}}(\mathbb{R}^d)$ satisfies

$$(-\hbar^2\Delta - 1)u = 0, \quad |x| > R, \quad \|(\hbar D_r - 1)u\|_{L^\infty(|x|=r)} = o_{r\to\infty}\left(r^{(1-d)/2}\right),$$

then

$$\|\mathrm{Op}_\hbar(a)u\|_{L^2(\mathbb{R}^d)} \le C\hbar^N\|\psi u\|_{H_k^{-N}(\mathbb{R}^d)}. \tag{5.35}$$

In addition, there exists $N_0 > 0$ such that for $\chi \in C_c^\infty(\mathbb{R}^d)$ with $\operatorname{supp}\chi \cap B(0, R) = \emptyset$,

$$\|\chi u\|_{L^2(\mathbb{R}^d)} \le C\hbar^{-N_0}\|\psi u\|_{H_k^2(\mathbb{R}^d)}. \tag{5.36}$$

A much sharper bound than (5.36) holds, but (5.36) is sufficient for the later proofs in this section. Indeed, using the result of Theorem 5.20 below in the proof of Lemma 5.18, we see that (5.36) holds with N_0 replaced by 1.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Proof of Lemma 5.18. Let

$$(R_0(\hbar)f)(x) \coloneqq \int_{\mathbb{R}^d} \Phi_\hbar(x,y)\, f(y)\, dy, \tag{5.37}$$

where $\Phi_\hbar$ is the outgoing fundamental solution satisfying $(-\hbar^2\Delta - 1)\Phi_\hbar(x,y) = \delta(x-y)$, that is,

$$\Phi_\hbar(x,y) \coloneqq \hbar^{-2}\frac{i}{4}\left(\frac{\hbar^{-1}}{2\pi|x-y|}\right)^{(d-2)/2} H^{(1)}_{d/2-1}\big(\hbar^{-1}|x-y|\big) \tag{5.38}$$

(see e.g. McLean 2000, Chapter 9). Observe that

$$u = R_0(-\hbar^2\Delta - 1)u =: R_0 f,$$

with $f \in L^2(\mathbb{R}^d)$, $\operatorname{supp} f \subset B(0,R)$. Since $f = Pu$, for any s,

$$\|f\|_{H^s_k} \le C\|u\|_{H^{s+2}_k(B(0,R))}. \tag{5.39}$$

Using standard Hankel function asymptotics for large argument, for $|z| > R+1$,

$$R_0 f(z) = \hbar^{-(d-3)/2-2}\int_{B(0,R)} e^{\frac{i}{\hbar}|z-y|} q(|z-y|) f(y)\, dy, \tag{5.40}$$

where

$$|\partial^\alpha q(s)| \le C_\alpha |s|^{\frac{1-d}{2}-|\alpha|}.$$

Thus, by the Cauchy–Schwarz inequality,

$$\|\chi u\|_{L^2(\mathbb{R}^d)} \le C\hbar^{-(d-3)/2-2}\|f\|_{L^2}, \tag{5.41}$$

and then (5.36) follows by combining (5.39) (with $s=0$) and (5.41).

Let ψ be the cut-off function appearing in the definition of $\operatorname{Op}_\hbar$ (5.3). To prove (5.35), observe that for R' large enough, $\operatorname{supp}(a(x,\xi)\psi(|x-\bullet|)) \subset \{|\bullet| > R+1\}$, and thus, by (5.40),

$$\begin{aligned}[\operatorname{Op}_\hbar(a)u](x) &= [\operatorname{Op}_\hbar(a)R_0 f](x)\\ &= \frac{\hbar^{-2-\frac{d-3}{2}}}{(2\pi\hbar)^d}\int_{|z|>R+1,\,|y|<R} e^{\frac{i}{\hbar}\Phi(z,\xi;x,y)} a(x,\xi)\psi(|x-z|)q(|z-y|)f(y)\, d\xi\, dz\, dy.\end{aligned}$$

where

$$\Phi(z,\xi;x,y) \coloneqq \langle x-z,\xi\rangle + |z-y|.$$

The critical points of the phase in (z,ξ) occur when $\partial_z\Phi = \partial_\xi\Phi = 0$, i.e. when

$$\xi = \frac{z-y}{|z-y|} \quad \text{and} \quad x = z,$$

so that $\xi = (x-y)/|x-y|$. We now claim that the support properties of a (5.34) imply that there are no critical points of the phase function on the support of the integrand. Indeed, when $|x| > R'$ for large enough R', $(x-y)/|x-y|$ is pointing

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

'outwards' (i.e. close to the direction $x/|x|$) and has modulus one, while, by (5.34), on supp a *either* ξ is pointing 'inwards' (or rather 'not too outward' because of the presence of $c > 0$) *or* $|\xi|$ is bounded away from one. Repeatedly integrating by parts in z and ξ then brings down arbitrarily many powers of $\hbar$ and $\langle\xi\rangle^{-1}$, proving the result. □

5.8.2. Trapping and non-trapping
Define $\pi_x\colon \mathbb{R}^{2d} \to \mathbb{R}^d$ by $\pi_x(x_0, \xi_0) = x_0$. Then

$$\Gamma_{\mathrm{f}} := \{(x,\xi)\colon |(\pi_x(\varphi_t(x,\xi))| \nrightarrow \infty \text{ as } t \to \infty\}, \tag{5.42}$$

i.e. Γ_{f} is the *forward trapped set*, and

$$\Gamma_{\mathrm{b}} := \{(x,\xi)\colon |(\pi_x(\varphi_t(x,\xi))| \nrightarrow \infty \text{ as } t \to -\infty\}, \tag{5.43}$$

i.e. Γ_{b} is the *backward trapped set*. Let $K := \Gamma_{\mathrm{f}} \cap \Gamma_{\mathrm{b}}$, i.e. K is the *trapped set*.

Lemma 5.19 (Properties of $\Gamma_{\mathrm{f}}, \Gamma_{\mathrm{b}}$ and K).

(i) $\Gamma_{\mathrm{f}} \neq \emptyset$ if and only if $\Gamma_{\mathrm{b}} \neq \emptyset$.

(ii) $\Gamma_{\mathrm{f}} \neq \emptyset$ if and only if $K \neq \emptyset$.

(iii) Γ_{f}, Γ_{b} and thus K are closed.

References for the proof. Part (i) follows by time-reversal of the flow, part (ii) is proved in Dyatlov and Zworski (2019, Proposition 6.4), for example, and part (iii) is proved in Dyatlov and Zworski (2019, Proposition 6.3). □

Recall that $\mathcal{R}$ is the solution operator for the problem (4.1).

Theorem 5.20 (Non-trapping cut-off resolvent bound). Let P be given by (5.2)/(4.1) with $\Omega_- = \emptyset$, A and n smooth, and $\mathrm{supp}(A - I)$ and $\mathrm{supp}(n-1)$ compact. If $K = \emptyset$, then for all $\chi \in C_c^\infty(\mathbb{R}^d)$, there exists $C > 0$ such that for all $0 < \hbar < 1$,

$$\|\chi\mathcal{R}\chi\|_{L^2 \to H^2_{\hbar}} \leq C\hbar^{-1}. \tag{5.44}$$

Sketch of proof.

(1) *Set-up.* Since $\|\chi\mathcal{R}\chi\|_{L^2\to H^2_{\hbar}}$ is a continuous function of $\hbar$ and bounded for every fixed $\hbar$, it is sufficient to prove (5.44) for $\hbar$ sufficiently small. Without loss of generality, we can assume that

$$\mathrm{supp}(A - I) \cup \mathrm{supp}(n-1) \subset \mathrm{supp}\,\chi.$$

Let $u := \mathcal{R}\chi f$, $R > 0$ such that $\mathrm{supp}\,\chi \subset B(0,R)$, and $\chi_1 \in C_c^\infty(\mathbb{R}^d)$ with $\mathrm{supp}(1-\chi_1) \cap B(0,R) = \emptyset$.

(2) *The elliptic region.* By the composition property (iv) in Informal Theorem 5.5, given $a \in S^0$ with $\mathrm{supp}\, a \cap \{p = 0\} = \emptyset$ and χ_1 as above, there exists a' such that $\mathrm{Op}_\hbar(a)\chi_1 = \mathrm{Op}_\hbar(a')$ and, up to an $O(\hbar^\infty)\Psi_\hbar^{-\infty}$ error, $\mathrm{supp}\, a' \subset \mathrm{supp}\, a$. Then, by

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

the fact that $\mathrm{Op}_\hbar(a)$ is properly supported and the elliptic estimate (Theorem 5.7, applied with a replaced by a'), there are $\psi_1, \psi \in C_c^\infty(\mathbb{R}^d)$ such that for any $M > 0$,

$$\begin{aligned}\|\mathrm{Op}_\hbar(a)\chi_1 u\|_{H_k^2} &= \|\psi_1 \mathrm{Op}_\hbar(a)\chi_1 u\|_{H_k^2} \\ &= \|\psi_1 \mathrm{Op}_\hbar(a')u\|_{H_k^2} \\ &\leq C\|Pu\|_{L^2} + C\hbar^M\|\psi u\|_{L^2}. \end{aligned} \tag{5.45}$$

(3) *The propagative region.* Since $K = \emptyset$, $\Gamma_\mathrm{b} = \emptyset$ (by Lemma 5.19(ii)), and thus for any $R' > R$ and all $(x_0, \xi_0) \in \{p = 0\} \cap \operatorname{supp} \chi_1$, there exists $t > 0$ such that

$$\varphi_{-t}(x_0, \xi_0) \in \left\{(x, \xi) \colon |x| > R', \left\langle \frac{x}{|x|}, \xi \right\rangle \leq 0\right\}.$$

By the propagation estimate (Theorem 5.10), there are $a, b \in C_c^\infty(\mathbb{R}^{2d})$ and $\psi \in C_c^\infty(\mathbb{R}^d)$, with $a(x_0, \xi_0) = 1$ and

$$\operatorname{supp} b \subset \left\{(x, \xi) \colon |x| > R', \left\langle \frac{x}{|x|}, \xi \right\rangle \leq 0\right\},$$

such that given $M > 0$, there exists $C > 0$ such that

$$\|\mathrm{Op}_\hbar(a)u\|_{L^2} \leq C\hbar^{-1}\|Pu\|_{L^2} + \|\mathrm{Op}_\hbar(b)u\|_{L^2} + C\hbar^M\|\psi u\|_{L^2}.$$

By the outgoing property (5.35) and the support properties of b, we obtain

$$\|\mathrm{Op}_\hbar(a)u\|_{L^2} \leq C\hbar^{-1}\|Pu\|_{L^2} + C\hbar^M\|\psi u\|_{L^2}. \tag{5.46}$$

(4) *Conclusion.* Equation (5.45) and (5.46) together with the compactness of $\{p = 0\} \cap \operatorname{supp} \chi_1$ and a partition of unity argument imply that

$$\|\chi_1 u\|_{H_k^2} \leq C\hbar^{-1}\|Pu\|_{L^2} + C\hbar^M\|\psi u\|_{L^2}$$

for some $\psi \in C_c^\infty(\mathbb{R}^d)$ and any $M > 0$. Absorbing part of the rightmost term into the left-hand side, we obtain

$$\|\chi_1 u\|_{H_k^2} \leq C\hbar^{-1}\|Pu\|_{L^2} + C\hbar^M\|\psi(1 - \chi_1)u\|_{L^2}. \tag{5.47}$$

Since $\psi(1 - \chi_1) \in C_c^\infty$ with support away from $B(0, R)$, we can use the crude bound (5.36) to bound the last term in (5.47). Then, with $M = N + N_0$,

$$\|\chi_1 u\|_{H_k^2} \leq C\hbar^{-1}\|Pu\|_{L^2} + C\hbar^N\|u\|_{H_k^2(B(0,R))};$$

the result then follows by absorbing the rightmost term into the left-hand side. □

Remark 5.21 (History and context of Theorem 5.20). Theorem 5.20, and its generalization to either a Dirichlet or Neumann obstacle, was proved in Vainberg (1975) using the propagation results of Melrose and Sjöstrand (1978, 1982) (see also Dyatlov and Zworski 2019, Theorem 4.43). Alternative proofs appear in Melrose (1979) and Burq (2002), using, respectively, Lax–Phillips scattering theory (Lax and Phillips 1989) and *defect measures* (another way of encoding propagation information; see Zworski 2012, §5).

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Remark 5.22 (Proving cut-off resolvent bounds using vector fields). When Ω_- is star-shaped and A and n satisfy certain monotonicity properties, the bound (5.44) can be proved using the vector field $V = x \cdot \nabla$ in (5.20), i.e. using Rellich–Morawetz identities (see Section 5.6.1). When $A \equiv I$ and $n \equiv 1$, this result goes back to Morawetz and Ludwig (1968) and Morawetz (1975); for the case of variable A and n, see Bloom (1973), Bloom and Kazarinoff (1977), Perthame and Vega (1999), Graham, Pembery and Spence (2019) and Moiola and Spence (2019). The advantage of the vector-field method is that when it works, the boundary and coefficients can be very rough (see Chandler-Wilde and Monk 2008, Graham *et al.* 2019), but it cannot prove the bound for a general smooth non-trapping obstacle (unlike propagation estimates). For cut-off resolvent estimates proved using non-radial vector fields – and hence managing to break out of the 'star-shaped paradigm' slightly – see Bloom and Kazarinoff (1974), Morawetz (1975), Morawetz, Ralston and Strauss (1977, §4), Chandler-Wilde *et al.* (2020) and Chaumont-Frelet and Spence (2023).

The evolution of propagation estimates relying on positive commutators from vector-field-type arguments is discussed in Morawetz *et al.* (1977) (where the term 'escape function' originated) and Bardos, Lebeau and Rauch (1992, §1). The relationship of the vector-field method to symmetries of the underlying PDE is given in Olver (1993, §4.4).

We emphasize that the vector field $x \cdot \nabla$ has often been used by the numerical analysis community to prove bounds on the solution operator of the interior impedance problem (a ubiquitous model problem in the numerical analysis of the Helmholtz equation) (Aziz *et al.* 1988, Douglas Jr *et al.* 1993, Makridakis *et al.* 1996, Melenk 1995, Proposition 8.2.7, Cummings and Feng 2006, Hetmaniuk 2007, Chaumont Frelet 2015, Chapter 2, Brown, Gallistl and Peterseim 2017, Barucq, Chaumont-Frelet and Gout 2017, Ohlberger and Verfurth 2018, Graham and Sauter 2020) and also to obtain new, coercive variational formulations of this problem (Moiola and Spence 2014, Ganesh and Morgenstern 2020).

5.8.3. Improved estimates with the measurement and/or data away from trapping
The next result shows that (provided the cut-off resolvent is polynomially bounded in $\hbar$) the behaviour of the solution operator is improved when the solution is measured away from the trapped set K.

Theorem 5.23 (Estimates with the measurement away from trapping). Let P be given by (5.2)/(4.1) with $\Omega_- = \emptyset$, A and n smooth, and $\mathrm{supp}(A - I)$ and $\mathrm{supp}(n - 1)$ compact. Suppose that $a \in C_c^\infty(\mathbb{R}^{2d})$ with $\mathrm{supp}\, a \cap K = \emptyset$. Then there exists $\chi \in C_c^\infty(\mathbb{R}^d)$ such that for all $N > 0$, there exists $C > 0$ such that, for $0 < \hbar < 1$ and all $u \in L^2_{\mathrm{loc}}$, with u satisfying the Sommerfeld radiation condition and $\mathrm{supp}(1 - \chi) \cap \mathrm{supp}\, Pu = \emptyset$,

$$\begin{aligned}\|\mathrm{Op}_\hbar(a)u\|_{L^2}^2 &\le C\hbar^{-1}\|Pu\|_{L^2}\|\chi u\|_{L^2} + C\hbar^{-2}\|Pu\|_{L^2}^2 + C\hbar^N\|\chi u\|_{L^2}^2 \\ &\le C\hbar^{-1}\left(\|\chi\mathcal{R}\chi\|_{L^2\to L^2} + \hbar^{-1} + \hbar^N\|\chi\mathcal{R}\chi\|_{L^2\to L^2}^2\right)\|Pu\|_{L^2}^2. \quad (5.48)\end{aligned}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Therefore, if $\|\chi\mathcal{R}\chi\|_{L^2\to L^2} \le C\hbar^{-M}$ for some $M > 0$, then

$$\|\mathrm{Op}_\hbar(a)\mathcal{R}\chi\|_{L^2\to L^2} \le C\sqrt{\hbar^{-1}\|\chi\mathcal{R}\chi\|_{L^2\to L^2}}.$$

As recalled in Section 4.3, $\rho := \|\chi\mathcal{R}\chi\|_{L^2\to L^2} \gg \hbar^{-1}$ when $K \neq \emptyset$, so that $\sqrt{\hbar^{-1}\rho} \ll \rho$ in this case; that is, Lemma 5.23 indeed shows that the solution operator is improved when the solution is measured away from K.

Sketch proof of Theorem 5.23.

(1) Let $w \in C_c^\infty(-2,2)$ $w \equiv \frac{1}{2}$ on $[-1,1]$ and $\int w^2 \equiv 1$. Then set

$$\psi_R(x) := w_R(|x|), \quad w_R(r) := 1 - \int_0^r w^2(s-R)\,ds$$

(i.e. $\psi_R \equiv 1$ on $r \le R-2$ and $\psi_R \equiv 0$ on $r \ge R+2$). Let $R > 0$ be sufficiently large that $\mathrm{supp}(1-\psi_R) \cap (\mathrm{supp}(A-I) \cup \mathrm{supp}(n-1) \cup \mathrm{supp}(Pu)) = \emptyset$. Then, by (5.10), $p(x,\xi) = |\xi|^2 - 1$ on $\mathrm{supp}\,\psi_R$ so that

$$H_p\psi_R = -2\left\langle \xi, \frac{x}{|x|}\right\rangle w^2(|x|-R).$$

(2) Since u is outgoing, Lemma 5.18 implies that all of its energy/mass in phase space over $\mathrm{supp}\, w^2(|\cdot|-R)$ is contained where $\langle \xi, x/|x|\rangle \ge c/2 > 0$ and $|\xi| \le 2$. Since

$$H_p\psi_R \le -cw^2(|x|-R) \text{ on } \left\{\left\langle \xi, \frac{x}{|x|}\right\rangle \ge c/2 > 0\right\} \cup \{|\xi| \le 2\},$$

by the sharp Gårding inequality (5.9) and Lemma 5.18,

$$\langle \mathrm{Op}_\hbar(H_p\psi_R)u, u\rangle \le -c\|w(|\bullet|-R)u\|_{L^2}^2 + \text{l.o.t.}$$

(3) Arguing as in (5.22)/(5.25), for real-valued $g \in C_c^\infty$,

$$2\hbar^{-1}\mathrm{Im}\,\langle Pu, \mathrm{Op}_\hbar(g)u\rangle = i\hbar^{-1}\langle [P, \mathrm{Op}_\hbar(g)]u, u\rangle = \langle \mathrm{Op}_\hbar(H_p g)u, u\rangle + \text{l.o.t.}, \tag{5.49}$$

so that

$$-2\hbar^{-1}|\mathrm{Im}\,\langle Pu, \psi_R u\rangle| \le -c\|w(|\bullet|-R)u\|_{L^2}^2 + \text{l.o.t.} \tag{5.50}$$

Although ψ_R is not compactly supported in ξ and so (5.49) cannot immediately be applied with $g = \psi_R$, the elliptic estimate shows that the high frequencies of u are well-behaved, and so this step can be rigorously implemented by inserting low-frequency cut-offs.

(4) By the Cauchy–Schwarz inequality, if χ is such that $\mathrm{supp}(1-\chi) \cap \mathrm{supp}\,\psi_R = \emptyset$,

$$\|w(|\bullet|-R)u\|_{L^2}^2 \le C\hbar^{-1}\|Pu\|_{L^2}\|\chi u\|_{L^2} + \text{l.o.t.}$$

(5) The result (5.48) now follows by applying Theorem 5.10, the fact that $w(|x|-R) = \frac{1}{2}$ on $\{R-1 \le |x| \le R+1\}$, and the fact that the Hamiltonian trajectory from

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

any point in $\operatorname{supp} a$ reaches $\{R-1 \le |x| \le R+1\}$ in finite time (under one of the forward or backward flow) since $\operatorname{supp} a \cap K = \emptyset$. As in the proof of Theorem 5.10, we iterate the argument to show that the lower-order terms can be made $O(\hbar^\infty)$. □

The next result shows that (provided that the resolvent is polynomially bounded in $\hbar$) if Pu is supported away from the trapped set and u is measured away from trapping, then we recover the non-trapping cut-off resolvent bound.

Theorem 5.24 (Estimates with measurement and data away from trapping). Let P be given by (5.2)/(4.1) with $\Omega_- = \emptyset$, A and n smooth, and $\operatorname{supp}(A-I)$ and $\operatorname{supp}(n-1)$ compact. Suppose that $a, b \in S^0(\mathbb{R}^{2d})$ with $(\operatorname{supp} a \cup \operatorname{supp} b) \cap K = \emptyset$. Then there exists $\chi \in C_c^\infty(\mathbb{R}^d)$ such that for all $N > 0$, there exists $C > 0$ such that, for $0 < \hbar < 1$ and all $u \in L^2_{\mathrm{loc}}(\mathbb{R}^d)$, with u satisfying the Sommerfeld radiation condition and $\operatorname{supp}(1-\chi) \cap \operatorname{supp} Pu = \emptyset$,

$$\|\operatorname{Op}_\hbar(a)u\|_{L^2}^2 \le C\hbar^{-2}\|Pu\|_{L^2}^2 + C\hbar^{-1}\|\operatorname{Op}_\hbar(1-b)Pu\|_{L^2}\|\chi u\|_{L^2} + C\hbar^N\|\chi u\|_{L^2}^2 \tag{5.51}$$

$$\le C\hbar^{-2}\big(1+\hbar^N\|\chi\mathcal{R}\chi\|_{L^2\to L^2}^2\big)\|Pu\|_{L^2}^2 + C\hbar^{-1}\|\operatorname{Op}_\hbar(1-b)Pu\|_{L^2}\|\chi\mathcal{R}\chi\|_{L^2\to L^2}\|Pu\|_{L^2}. \tag{5.52}$$

Moreover, if $\|\chi\mathcal{R}\chi\|_{L^2\to L^2} \le C\hbar^{-M}$ for some $M > 0$, then

$$\|\operatorname{Op}_\hbar(a)\mathcal{R}_P \operatorname{Op}_\hbar(b)\|_{L^2\to L^2} \le C\hbar^{-1}. \tag{5.53}$$

Sketch of proof.

(1) We begin from (5.50). As in the proof of Theorem 5.23, let χ be such that $\operatorname{supp}(1-\chi)\cap\operatorname{supp}\psi_R = \emptyset$. The inequality (5.50) together with the Cauchy–Schwarz inequality implies that

$$\begin{aligned}\|w(|x|-R)u\|_{L^2}^2 &\le C|\hbar^{-1}\operatorname{Im}\langle \operatorname{Op}_\hbar(b)Pu, \psi_R u\rangle| \\ &\quad + C\hbar^{-1}\|\operatorname{Op}_\hbar(1-b)Pu\|_{L^2}\|\chi u\|_{L^2} + \text{l.o.t.} \\ &= C|\hbar^{-1}\operatorname{Im}\langle Pu, \operatorname{Op}_\hbar(\bar b)u\rangle| \\ &\quad + C\hbar^{-1}\|\operatorname{Op}_\hbar(1-b)Pu\|_{L^2}\|\chi u\|_{L^2} + \text{l.o.t.} \\ &\le C\hbar^{-1}\|Pu\|_{L^2}\|\operatorname{Op}_\hbar(\bar b)u\|_{L^2} \\ &\quad + C\hbar^{-1}\|\operatorname{Op}_\hbar(1-b)Pu\|_{L^2}\|\chi u\|_{L^2} + \text{l.o.t.}\end{aligned} \tag{5.54}$$

(2) By Theorem 5.10 and the fact that $\operatorname{supp} b \cap K = \emptyset$ (and hence that the Hamiltonian trajectory from any point in $\operatorname{supp} b$ reaches $\{R-1 \le |x| \le R+1\}$ in finite time),

$$\|\operatorname{Op}_\hbar(\bar b)u\|_{L^2} \le C\hbar^{-1}\|Pu\|_{L^2} + \|w(|x|-R)u\|_{L^2} + \text{l.o.t.} \tag{5.55}$$

(3) Combining (5.54) and (5.55) yields

$$\|w(|x|-R)u\|_{L^2}^2 \le C\hbar^{-2}\|Pu\|_{L^2}^2 + C\hbar^{-1}\|Pu\|_{L^2}\|w(|x|-R)u\|_{L^2} + C\hbar^{-1}\|\operatorname{Op}_\hbar(1-b)Pu\|_{L^2}\|\chi u\|_{L^2} + \text{l.o.t.} \tag{5.56}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

(4) Use the weighted Young's inequality (5.27) on the second term on the right-hand side of (5.56), and absorbing the resulting $C\epsilon\|w(|x|-R)u\|^2_{L^2}$ on the left-hand side yields

$$\|w(|x|-R)u\|^2_{L^2} \leq C\hbar^{-2}\|Pu\|^2_{L^2} + C\hbar^{-1}\|\mathrm{Op}_\hbar(1-b)Pu\|_{L^2}\|\chi u\|_{L^2} + \text{l.o.t.}$$

(5) The result (5.51) follows by applying Theorem 5.10 as in step (5) of the proof of Theorem 5.23.

To obtain (5.53), given $b \in C_c^\infty$ with $\operatorname{supp} b \cap K = \emptyset$, since K is closed (by Lemma 5.19(iii)), there exists $\widetilde{b} \in C_c^\infty$ with $\operatorname{supp}\widetilde{b} \cap K = \emptyset$ and $\operatorname{supp}(1-\widetilde{b}) \cap \operatorname{supp} b = \emptyset$ (i.e. $\widetilde{b}$ is 'a bit bigger' than b). Applying (5.52) with b replaced by $\widetilde{b}$ and $u = R_P \operatorname{Op}_\hbar(b)f$, and using both the boundedness property (iii) in Theorem 5.5 and the polynomial bound on the resolvent, we obtain

$$\begin{aligned}&\|\mathrm{Op}_\hbar(a)R_P \operatorname{Op}_\hbar(b)f\|^2_{L^2}\\ &\quad\leq C\hbar^{-2}\|f\|^2_{L^2} + C\hbar^{-1-M}\|\mathrm{Op}_\hbar(1-\widetilde{b})\operatorname{Op}_\hbar(b)\|_{L^2\to L^2}\|f\|^2_{L^2}.\end{aligned}$$

By the composition property (iv) in Theorem 5.5 $\|\mathrm{Op}_\hbar(1-\widetilde{b})\operatorname{Op}_\hbar(b)\|_{L^2\to L^2} = \text{l.o.t.}$; in fact one can show that it is $O(\hbar^N)$ for any $N > 0$, and the result (5.53) follows. □

Remark 5.25 (History and context of Theorems 5.23 and 5.24). The result of Theorem 5.23 first appeared in Datchev and Vasy (2012*a*), and the result of Theorem 5.24 first appeared in Datchev and Vasy (2012*b*). The result of Theorem 5.24 with $\mathrm{Op}_\hbar(a)$ and $\mathrm{Op}_\hbar(1-b)$ replaced by spatial cut-offs supported far away from $\operatorname{supp}(A-I) \cup \operatorname{supp}(n-1)$, but without the assumption that $\|\chi\mathcal{R}\chi\|_{L^2\to L^2}$ is polynomially bounded, was proved in Burq (2002) and Cardoso and Vodev (2002) (see also Dyatlov and Zworski 2019, Theorem 6.22). The analogues of Theorems 5.23 and 5.24 for the Helmholtz PML problem were proved in Averseng *et al.* (2025, Appendix C); these are used to prove **R4** in Section 9.

6. Preasymptotic error estimates for the h-FEM (R1)

In this section we review and slightly improve the results of Galkowski and Spence (2025) concerning preasymptotic error estimates in the h-FEM, with the main result applied to the Helmholtz PML problem, Theorem 6.27 below.

The argument is based on the fact that Helmholtz operators are semiclassically elliptic at high frequencies. As discussed in Section 2.3, it is often possible to prove results based on this property without explicit use of SCA, and in this section all SCA has been removed in favour of the functional calculus of a self-adjoint operator. Nevertheless, we emphasize that the idea of the proof was obtained via 'thinking semiclassically', with this process illustrated via a model problem in Section 6.4.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

6.1. *Abstract set-up*

Let $\mathcal{H} \subset \mathcal{H}_0 \subset \mathcal{H}^*$ be Hilbert spaces with $\mathcal{H}_0$ identified with its dual, and let $a\colon \mathcal{H} \times \mathcal{H} \to \mathbb{C}$ be a sesquilinear form.

Definition 6.1 (Galerkin approximation of u for a). Let $\mathcal{V}_h \subset \mathcal{H}$ and $u \in \mathcal{H}$. A *Galerkin approximation of u in $\mathcal{V}_h$* is an element $u_h \in \mathcal{V}_h$ such that

$$a(u_h, v_h) = a(u, v_h) \quad \text{for all } v_h \in \mathcal{V}_h \tag{6.1}$$

(with (6.1) known as *Galerkin orthogonality*).

The goal of this section is to analyse the error in approximating u by u_h.

We use the subscript to h to indicate that the object comes from the approximation space (e.g. $\mathcal{V}_h$, u_h). This notation is motivated by the fact that the concrete spaces to which this abstract theory is applied in this article are piecewise polynomial spaces on a triangulation with width h.

Definition 6.2 (Norm of a sesquilinear form). For $a\colon \mathcal{H} \times \mathcal{H} \to \mathbb{C}$ a sesquilinear form,

$$\|a\| := \sup_{\substack{u,v\in\mathcal{H}\\ u,v\neq 0}} \frac{|a(u,v)|}{\|u\|_{\mathcal{H}}\|v\|_{\mathcal{H}}}.$$

If $\|a\| < \infty$, we say that a is *bounded*. We also need the notion of injectivity, surjectivity, and invertibility for sesquilinear forms.

Definition 6.3 (Injectivity, surjectivity, and invertibility). A sesquilinear form is *injective* if there is an operator $\mathcal{R}^*\colon \mathcal{H}^* \to \mathcal{H}$ such that for all $v \in \mathcal{H}^*$,

$$a(w, \mathcal{R}^* v) = \langle w, v\rangle \quad \text{for all } w \in \mathcal{H}; \tag{6.2}$$

a is *surjective* if there is an operator $\mathcal{R}\colon \mathcal{H}^* \to \mathcal{H}$ such that for all $v \in \mathcal{H}^*$,

$$a(\mathcal{R}v, w) = \langle v, w\rangle \quad \text{for all } w \in \mathcal{H};$$

a is *invertible* if it is both surjective and injective. In this case, $\mathcal{R}^* = (\mathcal{R})^*$.

Remark 6.4. Recall that a sesquilinear form $a\colon \mathcal{H} \times \mathcal{H} \to \mathbb{C}$ defines a unique operator $P\colon \mathcal{H} \to \mathcal{H}^*$ such that

$$\langle Pu, v\rangle = a(u, v) \quad \text{for all } u, v \in H$$

(note that $\|a\| = \|P\|_{\mathcal{H}\to\mathcal{H}^*}$). The existence of $\mathcal{R}^*$ as in Definition 6.3 implies that P^* has a right inverse and hence is surjective. In particular, since $\mathcal{H}$ is a Hilbert space (indeed, a reflexive Banach space), this implies that P is injective. Likewise, the existence of $\mathcal{R}$ implies that P is surjective.

6.2. *Coercivity and Céa's lemma*

It is straightforward to analyse the Galerkin method when it is applied to a sesquilinear form that is coercive in the following sense.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Definition 6.5 (Coercive sesquilinear form). A sesquilinear form a is *coercive* if

$$\mathfrak{C}(a) := \inf_{\substack{u\in\mathcal{H}\\ u\neq 0}} \frac{|a(u,u)|}{\|u\|_{\mathcal{H}}^2} > 0.$$

By the Lax–Milgram lemma, if a is bounded and coercive, then it is invertible (in the sense of Definition 6.3). In addition, we have the following estimate for the Galerkin approximation known as Céa's lemma (Céa 1964).

Lemma 6.6 (Céa's lemma). Suppose that $a\colon \mathcal{H}\times\mathcal{H}\to\mathbb{C}$ is a coercive sesquilinear form and $\mathcal{V}_h\subset\mathcal{H}$ is a finite-dimensional subspace. Then, for all $u\in\mathcal{H}$, the Galerkin approximation u_h to u in $\mathcal{V}_h$ exists, is unique, and satisfies

$$\|u-u_h\|_{\mathcal{H}} \leq (\mathfrak{C}(a))^{-1}\|a\|\|(I-\Pi_h)u\|_{\mathcal{H}}, \tag{6.3}$$

where $\Pi_h\colon \mathcal{H}\to\mathcal{V}_h$ denotes the orthogonal projection.

Proof. Suppose that a Galerkin approximation u_h to u in $\mathcal{V}_h$ exists. Then, by coercivity, Galerkin orthogonality and the definition of $\|a\|$,

$$\begin{aligned}\mathfrak{C}(a)\|u-u_h\|_{\mathcal{H}}^2 &\leq |a(u-u_h,u-u_h)|\\ &= |a(u-u_h,(I-\Pi_h)u)|\\ &\leq \|a\|\|u-u_h\|_{\mathcal{H}}\|(I-\Pi_h)u\|_{\mathcal{H}},\end{aligned}$$

and (6.3) holds. Now, if $u=0$ then (6.3) implies that any Galerkin approximation is zero. Hence, since $\mathcal{V}_h$ is finite-dimensional, the Galerkin approximation exists and is unique for any u, and the proof is complete. □

6.3. *The Schatz duality argument*

The standard variational formulations of Helmholtz problems are not coercive for large k. However, they are coercive up to lower-order terms in the following sense.

Assumption 6.7 (Gårding inequality). There are $c_{\mathrm{G}}>0$ and $C_{\mathrm{G}}>0$ such that for all $u\in\mathcal{H}$,

$$|a(u,u)| \geq c_{\mathrm{G}}\|u\|_{\mathcal{H}}^2 - C_{\mathrm{G}}\|u\|_{\mathcal{H}_0}^2.$$

The Schatz duality argument then gives sufficient conditions for quasioptimality in terms of the following adjoint approximability constant.

Definition 6.8 (Adjoint approximability). Let $\mathcal{V}_h\subset\mathcal{H}$, let a be an injective sesquilinear form, and let $\mathcal{W}$ be a Hilbert space with $\mathcal{W}\subset\mathcal{H}^*$. The *adjoint approximability constant from $\mathcal{W}$ to $\mathcal{H}$* is defined by

$$\eta_{\mathcal{W}\to\mathcal{H}}(\mathcal{V}_h) := \|(I-\Pi_h)\mathcal{R}^*\|_{\mathcal{W}\to\mathcal{H}}. \tag{6.4}$$

Lemma 6.9 (Schatz duality argument). Let $a\colon \mathcal{H}\times\mathcal{H}\to\mathbb{C}$ be an injective sesquilinear form satisfying Assumption 6.7, and let $0<\epsilon<1$. Then, for all

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

$u \in \mathcal{H}$ and $\mathcal{V}_h \subset \mathcal{H}$ finite-dimensional subspaces satisfying

$$C_{\mathrm{G}}\|a\|^2\big(\eta_{\mathcal{H}_0\to\mathcal{H}}(\mathcal{V}_h)\big)^2 \le (1-\epsilon)c_{\mathrm{G}}, \tag{6.5}$$

the Galerkin approximation u_h to u in $\mathcal{V}_h$ exists, is unique, and satisfies

$$\|u-u_h\|_{\mathcal{H}} \le \epsilon^{-1}c_{\mathrm{G}}^{-1}\|a\|\|(I-\Pi_h)u\|_{\mathcal{H}}. \tag{6.6}$$

Proof. Suppose that u_h is a Galerkin approximation to u in $\mathcal{V}_h$.

(1) *Estimates in $\mathcal{H}_0$.* Let $v \in \mathcal{H}_0$. By the definition of $\mathcal{R}^*$ (6.2), Galerkin orthogonality (6.1), the definition of $\|a\|$ and the definition of $\eta_{\mathcal{H}_0\to\mathcal{H}}$,

$$\begin{aligned}|\langle u-u_h, v\rangle| &= |a(u-u_h, \mathcal{R}^*v)|\\ &= |a(u-u_h, (I-\Pi_h)\mathcal{R}^*v)|\\ &\le \|a\|\|u-u_h\|_{\mathcal{H}}\|(I-\Pi_h)\mathcal{R}^*v\|_{\mathcal{H}}\\ &\le \|a\|\|u-u_h\|_{\mathcal{H}}\eta_{\mathcal{H}_0\to\mathcal{H}}\|v\|_{\mathcal{H}_0}.\end{aligned}$$

Hence

$$\|u-u_h\|_{\mathcal{H}_0} \le \|a\|\eta_{\mathcal{H}_0\to\mathcal{H}}(\mathcal{V}_h)\|u-u_h\|_{\mathcal{H}}. \tag{6.7}$$

(2) *Estimates in $\mathcal{H}$ using the Gårding inequality.* By Assumption 6.7, (6.1) and (6.7),

$$\begin{aligned}c_{\mathrm{G}}\|u-u_h\|_{\mathcal{H}}^2 &\le |a(u-u_h, u-u_h)| + C_{\mathrm{G}}\|u-u_h\|_{\mathcal{H}_0}^2\\ &\le |a(u-u_h, (I-\Pi_h)u)| + C_{\mathrm{G}}\|a\|^2\big(\eta_{\mathcal{H}_0\to\mathcal{H}}(\mathcal{V}_h)\big)^2\|u-u_h\|_{\mathcal{H}}^2\\ &\le \|a\|\|u-u_h\|_{\mathcal{H}}\|(I-\Pi_h)u\|_{\mathcal{H}} + C_{\mathrm{G}}\|a\|^2\big(\eta_{\mathcal{H}_0\to\mathcal{H}}(\mathcal{V}_h)\big)^2\|u-u_h\|_{\mathcal{H}}^2.\end{aligned}$$

Therefore, using (6.5), and subtracting the last term on the right-hand side to the left-hand side, we obtain

$$\epsilon c_{\mathrm{G}}\|u-u_h\|_{\mathcal{H}}^2 \le \|a\|\|u-u_h\|_{\mathcal{H}}\|(I-\Pi_h)u\|_{\mathcal{H}},$$

and thus (6.6) holds. Now, if $u=0$, then this implies that any Galerkin approximation is zero. Hence, since $\mathcal{V}_h$ is finite-dimensional, the Galerkin approximation exists and is unique for any u, and the proof is complete. □

Remark 6.10 (The history of the Schatz argument). The bound (6.7) when $a(\cdot,\cdot)$ is coercive was proved by Nitsche (1968) and Aubin (1967, Theorem 3.1), and appears in the form (6.7) in, for example, Ciarlet (2002, Theorem 3.2.4). Schatz was then the first to use this argument in conjunction with a Gårding inequality to prove quasioptimality; see Schatz (1974) and Schatz and Wang (1996, Lemma 4). The notation $\eta(\mathcal{V}_h)$ was introduced – and the importance of adjoint approximability emphasized – by Sauter (2006).

6.4. The elliptic projection

The idea behind the elliptic projection argument, as first introduced in Feng and Wu (2009, 2011), is to replace a sesquilinear form a satisfying a Gårding inequality by

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

one that is coercive with the cost of this replacement controlled. In Galkowski and Spence (2025), the original argument is improved by substantially decreasing the cost of this replacement. To explain the argument, we consider a model problem.

Model problem. As a model problem, we consider the Galerkin solution for the problem: given $f \in H^{-1}(\mathbb{T}^d)$, find $u \in H^1(\mathbb{T}^d)$ such that

$$(-k^{-2}\Delta - n(x) - ik^{-1})u = f \quad \text{on } \mathbb{T}^d$$

(where the imaginary term in the PDE is introduced just to ensure that the solution is unique for all k). More precisely, let $\mathcal{H}_0 := L^2(\mathbb{T}^d)$ and $\mathcal{H} := H^1_k(\mathbb{T}^d)$, and let $a_k : \mathcal{H} \times \mathcal{H} \to \mathbb{C}$ be defined by

$$a_k(u, v) := \langle k^{-1}\nabla u, k^{-1}\nabla v\rangle - \langle n(x)u, v\rangle - ik^{-1}\langle u, v\rangle. \tag{6.8}$$

By this definition, the following Gårding inequality holds:

$$\begin{aligned}\operatorname{Re} a_k(u, u) &\geq \|k^{-1}\nabla u\|^2_{L^2(\mathbb{T}^d)} - \|n\|_{L^\infty}\|u\|^2_{L^2(\mathbb{T}^d)}\\ &\geq \|u\|^2_{H^1_k(\mathbb{T}^d)} - (1 + \|n\|_{L^\infty})\|u\|^2_{L^2(\mathbb{T}^d)}.\end{aligned}$$

In particular, this implies that

$$\widetilde{a}_k(u, v) := a_k(u, v) + \langle \widetilde{S}u, v\rangle, \quad \text{with } \widetilde{S} = (1 + \|n\|_{L^\infty})I,$$

is coercive and hence Céa's lemma applies. This is the key observation (applied to this particular model) used in Feng and Wu (2009, Lemma 5.1) and Feng and Wu (2011, Lemma 4.1) (albeit there in the context of discontinuous Galerkin methods).

To produce the estimates from Galkowski and Spence (2025), we need to reduce the 'cost' of obtaining a coercive operator. It is not possible to reduce the norm of $\|\widetilde{S}\|_{\mathcal{H}_0\to\mathcal{H}_0}$, but it *is* possible to improve its regularity properties.

Lemma 6.11 (Construction of S_k for the model problem with constant coefficients). Let a_k be as in (6.8) with $n \equiv 1$. For all $k_0 > 0$, there exists $c > 0$ such that for $k \geq k_0$, there exists S_k for which

$$a_{S_k}(u, v) := a_k(u, v) + \langle S_k u, v\rangle \text{ is coercive with } \mathfrak{C}(a_{S_k}) \geq c. \tag{6.9}$$

Furthermore, for all $N > 0$ there exists $C_N > 0$ such that for all $k \geq k_0$,

$$\|S_k\|_{H^{-N}_k \to H^N_k} \leq C_N. \tag{6.10}$$

Proof. Let $\chi \in C^\infty_c((-2, 2); [0, 1])$ with $\operatorname{supp}(1 - \chi) \cap [-1, 1] = \emptyset$ and define $S : \cup_N H^{-N}_k(\mathbb{T}^d) \to \cap_N H^N_k(\mathbb{T}^d)$ by

$$\widehat{S_k u}(m) := 2\chi(k^{-1}|m|)\widehat{u}(m), \quad m \in \mathbb{Z}^d,$$

where

$$\widehat{u}(m) := \int_{\mathbb{T}^d} e^{-imx}u(x)\,dx.$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

The bound (6.10) follows easily from this definition of S_k. To prove (6.9), we express a_k using Fourier series. By Plancherel's theorem,

$$a_k(u,v) = (2\pi)^{-d}\big(\langle \widehat{u}(m) i k^{-1} m, \widehat{v}(m) i k^{-1} m\rangle_{\ell^2(\mathbb{Z}^d)} - (1 + ik^{-1})\langle \widehat{u}, \widehat{v}\rangle_{\ell^2(\mathbb{Z}^d)}\big),$$

so that

$$\operatorname{Re} a_{S_k}(u,u) = (2\pi)^{-d}\langle (k^{-2}|m|^2 - 1 + 2\chi(k^{-1}|m|))\widehat{u}, \widehat{u}\rangle_{\ell^2(\mathbb{Z}^d)}.$$

Observe that since $\operatorname{supp}(1-\chi)\cap[-1,1] = \emptyset$,

$$\inf_{x\in[0,2]}(x^2 - 1 + 2\chi(x)) > 0, \qquad \inf_{x\in[2,\infty)}(1+x^2)^{-1}(x^2-1) > 0,$$

and hence

$$\operatorname{Re} a_{S_k}(u,u) \geq c\langle (1 + k^{-2}|m|^2)\widehat{u}, \widehat{u}\rangle_{\ell^2(\mathbb{Z}^d)} = c\|u\|^2_{H^1_k(\mathbb{T}^d)},$$

i.e. $\mathfrak{C}(a_{S_k}) \geq c$. □

When the wave speed is not constant, the Fourier analysis used in Lemma 6.11 does not apply. One route to proving the analogue of Lemma 6.11 is to use a pseudodifferential operator S_k rather than a Fourier multiplier.

Indeed, in the variable wave speed case, one could replace S_k by

$$S_k := \operatorname{Op}_\hbar(2n(x)\chi(n(x)^{-1}|\xi|^2))$$

where $\chi \in C_c^\infty(\mathbb{R};[0,1])$, $\operatorname{supp}(1-\chi)\cap[-1,1] = \emptyset$. Since

$$\sigma_\hbar(S_k) = 2n(x) \quad \text{when } |\xi|^2 \leq n(x),$$

there exists $\mu > 0$ such that

$$\operatorname{Re}\sigma_\hbar(-k^{-2}\Delta - n(x) - ik^{-1} + S_k) = |\xi|^2 - n(x) + \sigma_\hbar(S_k) \geq \mu\langle\xi\rangle^2,$$

so that by the sharp Gårding inequality (5.9) from Informal Theorem 5.5,

$$\operatorname{Re}\langle(-k^{-2}\Delta - n(x) - ik^{-1} + S_k)u, u\rangle \geq \frac{\mu}{2}\|u\|^2_{H^1_k}.$$

Furthermore, since $\chi \in C_c^\infty(\mathbb{R})$,

$$\|S_k\|_{H^{-N}_k \to H^N_k} \leq C_N$$

by the boundedness property (iii) from Informal Theorem 5.5.

However, we show in Section 6.6 (see Theorem 6.21) that the procedure in the proof in Lemma 6.11, applied there for constant coefficients, can also be implemented in much more general situations by using the functional calculus of a self-adjoint operator, that is, we do not need to use pseudodifferential operators to generalize to variable coefficients. Before doing this, we provide an analysis of the Galerkin error when there exists S_k such that (6.9) holds.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

6.5. *An abstract elliptic projection set-up*

To demonstrate the main idea behind the h-FEM preasymptotic error estimates, we introduce the notion of coercivity up to a subspace $\mathcal{Z} \subset \mathcal{H}$. This notion replaces the standard Gårding-type inequality in the elliptic projection argument.

Assumption 6.12 (Coercivity up to $\mathcal{Z}^*$). Let $a\colon \mathcal{H} \times \mathcal{H} \to \mathbb{C}$ be a sesquilinear form and $\mathcal{Z} \subset \mathcal{H}$. There exists $S\colon \mathcal{Z}^* \to \mathcal{Z}$ such that

$$\mathfrak{C}(a_S) > 0, \quad a_S(u, v) \coloneqq a(u, v) + \langle Su, v\rangle. \tag{6.11}$$

Remark 6.13. Notice that Assumption 6.12 implies the weaker Gårding-type inequality

$$|a(u, u)| \geq c_{\mathrm{G}}\|u\|^2_{\mathcal{H}} - C_{\mathrm{G}}\|u\|^2_{\mathcal{Z}^*}, \tag{6.12}$$

hence the terminology 'coercivity up to $\mathcal{Z}^*$'.

Remark 6.14. We show in Section 6.6 that, under elliptic-regularity-type assumptions on a, (6.12) holds with $\mathcal{Z} = H^\ell_k$ for some large ℓ. In this context, one should think of the $\|\bullet\|_{\mathcal{Z}^*}$-norm as measuring 'low frequencies' and hence that (6.12) states that a is coercive up to low frequencies (as discussed in Section 2.2).

If a satisfies Assumption 6.12 and $\|a\| < \infty$, then by the Lax–Milgram lemma, there is a unique operator $\mathcal{R}_S\colon \mathcal{H}^* \to \mathcal{H}$ such that

$$a_S(\mathcal{R}_S g, w) = \langle g, w\rangle \quad \text{for all } w \in \mathcal{H},\ g \in \mathcal{H}^*. \tag{6.13}$$

Theorem 6.15 (Abstract elliptic projection argument). Let $a\colon \mathcal{H} \times \mathcal{H} \to \mathbb{C}$ be an injective sesquilinear form satisfying Assumption 6.12, let $\mathcal{V}_h$ be a finite-dimensional subspace of $\mathcal{H}$, and $\epsilon > 0$. Suppose that

$$\|S\|_{\mathcal{Z}^*\to\mathcal{Z}}(\mathfrak{C}(a_S))^{-1}\|a_S\|^2\|(I - \Pi_h)\mathcal{R}_S\|_{\mathcal{Z}\to\mathcal{H}}\eta_{\mathcal{Z}\to\mathcal{H}}(\mathcal{V}_h) \leq 1 - \epsilon. \tag{6.14}$$

Then, for all $u \in \mathcal{H}$, the Galerkin approximation u_h to u in $\mathcal{V}_h$ exists, is unique, and, for all $\mathcal{Y} \subset \mathcal{Z}$, satisfies

$$\begin{aligned}&\|u - u_h\|_{\mathcal{H}}\\ &\quad\leq (\mathfrak{C}(a_S))^{-1}\Big(\|a\| + \epsilon^{-1}\sqrt{2(\mathfrak{C}(a_S))^{-1}\|S\|_{\mathcal{Z}^*\to\mathcal{Z}}}\|a_S\|^2\eta_{\mathcal{Z}\to\mathcal{H}}(\mathcal{V}_h)\Big)\|(I - \Pi_h)u\|_{\mathcal{H}},\end{aligned} \tag{6.15}$$

$$\begin{aligned}&\|u - u_h\|_{\mathcal{Y}^*}\\ &\quad\leq \Big(1 + \epsilon^{-1}\|S\|_{\mathcal{Z}^*\to\mathcal{Z}}(\mathfrak{C}(a_S))^{-1}\|a_S\|^2\|(I - \Pi_h)\mathcal{R}_S\|_{\mathcal{Z}\to\mathcal{H}}\eta_{\mathcal{Z}\to\mathcal{H}}(\mathcal{V}_h)\Big)\\ &\qquad\times \|a_S\|^2(\mathfrak{C}(a_S))^{-1}\eta_{\mathcal{Y}\to\mathcal{H}}(\mathcal{V}_h)\|(I - \Pi_h)u\|_{\mathcal{H}}\end{aligned} \tag{6.16}$$

and

$$\|u - u_h\|_{\mathcal{Z}^*} \leq \epsilon^{-1}\|a_S\|^2(\mathfrak{C}(a_S))^{-1}\eta_{\mathcal{Z}\to\mathcal{H}}(\mathcal{V}_h)\|(I - \Pi_h)u\|_{\mathcal{H}}. \tag{6.17}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Remark 6.16. If $\mathcal{Y} = \mathcal{H}^*$, then (6.16) gives a bound on $\|u-u_h\|_{\mathcal{H}}$ complementary to that in (6.15). When applied to concrete Helmholtz problems, these two bounds both give (3.4), the only difference being the value of the constant C.

Proof.

(1) *Estimates for a_S.* By Assumption 6.12, a_S is coercive. Hence, by Céa's lemma (Lemma 6.6), there exists $\Pi_S : \mathcal{H}^* \to \mathcal{V}_h$ such that

$$a_S(w_h, \Pi_S g) = a_S(w_h, g) \quad \text{for all } w_h \in \mathcal{V}_h,\ g \in \mathcal{H}^*.$$

Moreover,

$$\|(I - \Pi_S)w\|_{\mathcal{H}} \le (\mathfrak{C}(a_S))^{-1}\|a_S\|\|(I - \Pi_h)w\|_{\mathcal{H}}. \tag{6.18}$$

Next we improve (6.18) when $(I - \Pi_S)w$ is measured in $\mathcal{Z}^*$ using the Schatz duality argument (Lemma 6.9). For $v \in \mathcal{Z}$, by (6.13), Galerkin orthogonality and (6.18),

$$\begin{aligned}
|\langle v, (I - \Pi_S)w\rangle| &= |a_S(\mathcal{R}_S v, (I - \Pi_S)w\rangle| \\
&= |a_S((I - \Pi_h)\mathcal{R}_S v, (I - \Pi_S)w)| \\
&\le \|a_S\|\|(I - \Pi_h)\mathcal{R}_S\|_{\mathcal{Z}\to\mathcal{H}}\|(I - \Pi_S)w\|_{\mathcal{H}}\|v\|_{\mathcal{Z}} \\
&\le (\mathfrak{C}(a_S))^{-1}\|a_S\|^2\|(I - \Pi_h)\mathcal{R}_S\|_{\mathcal{Z}\to\mathcal{H}}\|(I - \Pi_h)w\|_{\mathcal{H}}\|v\|_{\mathcal{Z}}.
\end{aligned}$$

Hence

$$\|(I - \Pi_S)w\|_{\mathcal{Z}^*} \le (\mathfrak{C}(a_S))^{-1}\|a_S\|^2\|(I - \Pi_h)\mathcal{R}_S\|_{\mathcal{Z}\to\mathcal{H}}\|(I - \Pi_h)w\|_{\mathcal{H}}. \tag{6.19}$$

(2) *Proof of existence, uniqueness and* (6.17). Suppose that u_h is a Galerkin approximation to u in $\mathcal{V}_h$ and let $v \in \mathcal{Z}$. Then, by (6.2), Galerkin orthogonality (6.1), and (6.11),

$$\begin{aligned}
|\langle u - u_h, v\rangle| &= |a(u - u_h, \mathcal{R}^* v)| \\
&= |a(u - u_h, (I - \Pi_S)\mathcal{R}^* v)| \\
&= |a_S((I - \Pi_h)u, (I - \Pi_S)\mathcal{R}^* v) - \langle S(u - u_h), (I - \Pi_S)\mathcal{R}^* v\rangle| \\
&\le \|a_S\|\|(I - \Pi_h)u\|_{\mathcal{H}}\|(I - \Pi_S)\mathcal{R}^* v\|_{\mathcal{H}} \\
&\quad + \|S\|_{\mathcal{Z}^*\to\mathcal{Z}}\|u - u_h\|_{\mathcal{Z}^*}\|(I - \Pi_S)\mathcal{R}^* v\|_{\mathcal{Z}^*}.
\end{aligned} \tag{6.20}$$

Hence, by (6.18) and (6.19),

$$\begin{aligned}
&|\langle u - u_h, v\rangle| \\
&\quad \le \|a_S\|\|(I - \Pi_h)u\|_{\mathcal{H}}(\mathfrak{C}(a_S))^{-1}|\|a_S\|\|(I - \Pi_h)\mathcal{R}^*\|_{\mathcal{Z}\to\mathcal{H}}\|v\|_{\mathcal{Z}} \\
&\qquad + \|S\|_{\mathcal{Z}^*\to\mathcal{Z}}\|u - u_h\|_{\mathcal{Z}^*}(\mathfrak{C}(a_S))^{-1}\|a_S\|^2 \\
&\qquad\quad \times \|(I - \Pi_h)\mathcal{R}_S\|_{\mathcal{Z}\to\mathcal{H}}\|(I - \Pi_h)\mathcal{R}^*\|_{\mathcal{Z}\to\mathcal{H}}\|v\|_{\mathcal{Z}}
\end{aligned}$$

so condition (6.14) implies that (6.17) holds. Since (6.17) implies uniqueness of the Galerkin solution and $\mathcal{V}_h$ is finite-dimensional, the Galerkin solution exists, is unique, and satisfies (6.17)

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

(3) *Proof of* (6.16). We begin from (6.20), this time using (6.17), (6.18) and (6.19), to obtain

$$\begin{aligned}&|\langle u-u_h, v\rangle|\\ &\quad\le \big(1+\epsilon^{-1}\|(I-\Pi_h)\mathcal{R}_S\|_{\mathcal{Z}\to\mathcal{H}}\eta_{\mathcal{Z}\to\mathcal{H}}(\mathcal{V}_h)\|S\|_{\mathcal{Z}^*\to\mathcal{Z}}(\mathfrak{C}(a_S))^{-1}\|a_S\|^2\big)\\ &\qquad\times\|a_S\|^2(\mathfrak{C}(a_S))^{-1}\eta_{\mathcal{Y}\to\mathcal{H}}(\mathcal{V}_h)\|(I-\Pi_h)u\|_{\mathcal{H}}\|v\|_{\mathcal{Y}},\end{aligned}$$

which implies (6.16).

(4) *Proof of* (6.15). By (6.11) and Galerkin orthogonality (6.1),

$$\begin{aligned}\|u-u_h\|_{\mathcal{H}}^2 &\le (\mathfrak{C}(a_S))^{-1}|a_S(u-u_h,u-u_h)|\\ &\le (\mathfrak{C}(a_S))^{-1}\big(|a(u-u_h,(I-\Pi_h)u)|+|\langle u-u_h,S(u-u_h)\rangle|\big)\\ &\le (\mathfrak{C}(a_S))^{-1}\big(\|a\|\|u-u_h\|_{\mathcal{H}}\|(I-\Pi_h)u\|_{\mathcal{H}}+\|S\|_{\mathcal{Z}^*\to\mathcal{Z}}\|u-u_h\|_{\mathcal{Z}^*}^2\big)\\ &\le \big(\delta\|u-u_h\|_{\mathcal{H}}^2+\tfrac14\|a\|^2(\mathfrak{C}(a_S))^{-2}\delta^{-1}\|(I-\Pi_h)u\|_{\mathcal{H}}^2\\ &\quad+(\mathfrak{C}(a_S))^{-1}\|S\|_{\mathcal{Z}^*\to\mathcal{Z}}\epsilon^{-2}\|a_S\|^4\|(I-\Pi_h)u\|_{\mathcal{H}}^2(\mathfrak{C}(a_S))^{-2}\eta_{\mathcal{Z}\to\mathcal{H}}^2(\mathcal{V}_h)\big).\end{aligned}$$

Optimizing in δ, and using $\sqrt{a^2+b^2}\le a+b$, we obtain (6.15). □

6.6. *Ellipticity for abstract Helmholtz operators*

In this section we consider a *family* of sesquilinear operators $a_k\colon \mathcal{H}\times\mathcal{H}\to\mathbb{C}$, $k\ge 0$, where the norms on all spaces may implicitly depend on k.

Assumption 6.17 (Uniform boundedness). For all $k_0>0$,

$$\sup_{k>k_0}\|a_k\|<\infty.$$

Assumption 6.18 (Uniform Gårding inequality). For all $k_0>0$, there are $c_{\mathrm{G}}>0$ and $C_{\mathrm{G}}>0$ such that for $k>k_0$,

$$\operatorname{Re} a(u,u)\ge c_{\mathrm{G}}\|u\|_{\mathcal{H}}^2-C_{\mathrm{G}}\|u\|_{\mathcal{H}_0}^2.$$

Assumption 6.19 (Abstract elliptic regularity for a_k and $\operatorname{Re} a_k$). Let $\mathcal{Z}_0:=\mathcal{H}_0$, $\mathcal{Z}_1:=\mathcal{H}$, and $\mathcal{Z}_j\subset\mathcal{Z}_{j-1}$ for $j=2,3,\dots,\ell+1$ such that $\mathcal{Z}_j$ is dense in $\mathcal{Z}_{j-1}$. For all $k_0\ge 0$, there are $C_{\mathrm{ell}}>0$ such that for all $u\in\mathcal{H}$ and $k>k_0$,

$$\|u\|_{\mathcal{Z}_j}\le C_{\mathrm{ell}}\left(\|u\|_{\mathcal{H}_0}+\sup_{v\in\mathcal{H},\,\|v\|_{(\mathcal{Z}_{j-2})^*}=1}|a_k(u,v)|\right),\quad j=2,\dots,\ell+1,\tag{6.21}$$

and

$$\|u\|_{\mathcal{Z}_j}\le C_{\mathrm{ell}}\left(\|u\|_{\mathcal{H}_0}+\sup_{v\in\mathcal{H},\,\|v\|_{(\mathcal{Z}_{j-2})^*}=1}|(\operatorname{Re}a_k)(u,v)|\right),\quad j=2,\dots,\ell+1,\tag{6.22}$$

where $(\operatorname{Re}a_k)(u,v)=\frac12(a_k(u,v)+\overline{a_k(v,u)})$.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Remark 6.20. In Assumption 6.19, we mean that if either of the right-hand sides of the bounds (6.21) and (6.22) are finite, then $u \in \mathcal{Z}_j$ and the bounds hold.

We now state the main theorem of this section.

Theorem 6.21. Suppose that $a_k : \mathcal{H} \times \mathcal{H} \to \mathbb{C}$ is a sesquilinear form satisfying Assumptions 6.17, 6.18 and 6.19 for some $\ell \in \mathbb{Z}^+$. Then, for all $\epsilon > 0$ and $k_0 > 0$, there are $C > 0$ and S_k such that for $k > k_0$, a_k satisfies Assumption 6.12 with $S = S_k$, and for all $k > k_0$,

$$\mathfrak{C}(a_{k,S_k}) \geq (1-\epsilon)c_{\mathrm{G}}, \tag{6.23}$$

$$\|S_k\|_{(\mathcal{Z}_{\ell+1})^* \to \mathcal{Z}_{\ell+1}} \leq C, \tag{6.24}$$

$$\|\mathcal{R}_{S_k}\|_{\mathcal{Z}_{j-1} \to \mathcal{Z}_{j+1}} \leq C, \quad j = 0, 1, \dots, \ell, \tag{6.25}$$

where $\mathcal{R}_{S_k}$ is the operator defined by (6.13) with $a = a_k$ and $S = S_k$.

Remark 6.22. In fact the proof shows that there exists $\chi \in C_c^\infty(\mathbb{R})$ such that $S_k = \chi(\mathcal{P})$, where $\mathcal{P}$ is the self-adjoint operator generated by $\operatorname{Re} a_k$. This operator is also used in the proofs of the results **R2** and **R4**.

Proof.

(1) *Construction of S_k.* By Assumptions 6.17 and 6.18 and the Friedrichs extension theorem (see e.g. Reed and Simon 1980, Theorem VIII.15, Grubb 2008, Theorem 12.24, with the original paper by Friedrichs 1934), there is a (potentially unbounded) self-adjoint operator $\mathcal{P} : \mathcal{H}_0 \to \mathcal{H}_0$ with domain

$$\mathcal{D}(\mathcal{P}) = \left\{ u \in \mathcal{H} : \sup_{\substack{v \in \mathcal{H} \\ v \neq 0}} \frac{|\operatorname{Re} a_k(u,v)|}{\|v\|_{\mathcal{H}_0}} < \infty \right\} \tag{6.26}$$

such that

$$\operatorname{Re} a_k(u,v) = \langle \mathcal{P}u, v \rangle \quad \text{for } u \in \mathcal{D}(\mathcal{P}).$$

By the spectral theorem for unbounded operators (see e.g. Reed and Simon 1980, Chapter VII), there is a unitary operator $U : \mathcal{H}_0 \to L^2(\mathbb{R}; \mu)$ for some Borel measure μ such that for $u \in \mathcal{D}(\mathcal{P})$,

$$\mathcal{P}u = U^* M_\lambda U u,$$

where, for $f : \mathbb{R} \to \mathbb{C}$ μ measurable, $(M_f w)(\lambda) := f(\lambda) w(\lambda)$. In addition, $u \in \mathcal{D}(P)$ if and only if $M_\lambda U u \in L^2(\mu)$.

Therefore, for $u \in \mathcal{H}$,

$$\operatorname{Re} a_k(u,u) = \langle \lambda U u, U u \rangle_{L^2(d\mu)} \geq c_{\mathrm{G}} \|u\|_{\mathcal{H}}^2 - C_{\mathrm{G}} \|u\|_{\mathcal{H}_0}^2 \geq (c_{\mathrm{G}} - C_{\mathrm{G}}) \|u\|_{\mathcal{H}_0}^2,$$

which implies that $\operatorname{supp} \mu \subset [c_{\mathrm{G}} - C_{\mathrm{G}}, \infty)$.

Let $M \geq C_{\mathrm{G}}$, to be chosen large depending on ϵ, and $\chi \in C_c^\infty(\mathbb{R}; [0,1])$ with $\operatorname{supp}(1-\chi) \cap [-C_{\mathrm{G}}, M] = \emptyset$, and define

$$S_k u := C_{\mathrm{G}} \chi(\mathcal{P}),$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

where for $f \in C^0(\mathbb{R}, d\mu)$, we define $f(\mathcal{P}) \coloneqq U^* M_f U$. Note that when f is real-valued, $f(\mathcal{P})$ is then self-adjoint. Thus, by the definition of χ and the support properties of μ,

$$\begin{aligned}\operatorname{Re} a_{k,S_k}(u,u) &= \langle (\lambda + C_{\mathrm{G}}\chi(\lambda))Uu, Uu\rangle_{L^2(d\mu)} \\ &\geq \left\langle \frac{\lambda + C_{\mathrm{G}}\chi(\lambda)}{\lambda + C_{\mathrm{G}}}(\lambda + C_{\mathrm{G}})Uu, Uu \right\rangle_{L^2(d\mu)} \\ &\geq \inf_{\lambda \geq M} \frac{\lambda}{\lambda + C_{\mathrm{G}}} \langle (\lambda + C_{\mathrm{G}})Uu, Uu\rangle_{L^2(d\mu)} \\ &\geq \inf_{\lambda \geq M} \frac{\lambda}{\lambda + C_{\mathrm{G}}} c_{\mathrm{G}} \|u\|^2_{\mathcal{H}}.\end{aligned}$$

For M large enough, this implies (6.23).

(2) *Estimates for S_k.* To prove (6.24), we claim that for any $\psi \in C_c^\infty(\mathbb{R})$ there exists $C > 0$ such that

$$\|\psi(\mathcal{P})\|_{(\mathcal{Z}_{\ell+1})^* \to \mathcal{Z}_{\ell+1}} \leq C; \tag{6.27}$$

since $S = C_{\mathrm{G}}\chi(\mathcal{P})$ for some $\chi \in C_c^\infty$, this implies (6.24).

To prove (6.27) we use Assumption 6.19. First, observe that for all $m \geq 0$ and any $\psi \in C_c^\infty(\mathbb{R})$,

$$\|\lambda^m \psi(\lambda)\|_{L^\infty} \leq C.$$

Hence

$$\|\mathcal{P}^m \psi(\mathcal{P})\|_{\mathcal{H}_0 \to \mathcal{H}_0} \leq C.$$

Now, by (6.22), for $2 \leq j \leq \ell + 1$,

$$\|\psi(\mathcal{P})\|_{\mathcal{H}_0 \to \mathcal{Z}_j} \leq C_{\mathrm{ell}} \left(\|\psi(\mathcal{P})\|_{\mathcal{H}_0 \to \mathcal{H}_0} + \|\mathcal{P}\psi(\mathcal{P})\|_{\mathcal{H}_0 \to \mathcal{Z}_{j-2}} \right).$$

Hence, by induction,

$$\|\psi(\mathcal{P})\|_{\mathcal{H}_0 \to \mathcal{Z}_{\ell+1}} \leq C_\ell \sum_{j=0}^{\lceil (\ell+1)/2 \rceil} \|\mathcal{P}^j \psi(\mathcal{P})\|_{\mathcal{H}_0 \to \mathcal{H}_0} \leq C.$$

Now, let $\tilde{\psi} \in C_c^\infty(\mathbb{R})$ with $\operatorname{supp}\psi \cap \operatorname{supp}(1 - \tilde{\psi}) = \emptyset$. Then, since $\tilde{\psi}(\mathcal{P})$ is self-adjoint,

$$\|\tilde{\psi}(\mathcal{P})\|_{(\mathcal{Z}_{\ell+1})^* \to \mathcal{H}_0} \leq C$$

and hence, since $\psi(\mathcal{P}) = \psi(\mathcal{P})\tilde{\psi}(\mathcal{P})$,

$$\begin{aligned}\|\psi(\mathcal{P})\|_{(\mathcal{Z}_{\ell+1})^* \to \mathcal{Z}_{\ell+1}} &= \|\psi(\mathcal{P})\tilde{\psi}(\mathcal{P})\|_{(\mathcal{Z}_{\ell+1})^* \to \mathcal{Z}_{\ell+1}} \\ &\leq \|\psi(\mathcal{P})\|_{\mathcal{H}_0 \to \mathcal{Z}_{\ell+1}} \|\tilde{\psi}(\mathcal{P})\|_{(\mathcal{Z}_{\ell+1})^* \to \mathcal{H}_0} \\ &\leq C,\end{aligned}$$

which proves (6.27).

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

(3) *Estimates for* $\mathcal{R}_{S_k}$. By (6.23) and the Lax–Milgram lemma, $\mathcal{R}_{S_k}$, defined by

$$a_{k,S_k}(\mathcal{R}_{S_k}g, v) = \langle g, v\rangle, \quad g \in \mathcal{H}^*,$$

satisfies

$$\|\mathcal{R}_{S_k}g\|_{\mathcal{H}} \leq \mathfrak{C}(a_{k,S_k})^{-1}\|g\|_{\mathcal{H}^*}.$$

In particular, for $g \in \mathcal{Z}_{j-2}$, $j = 2, \dots, \ell+1$, $v \in \mathcal{H}$,

$$\begin{aligned}
|a_k(\mathcal{R}_{S_k}g, v)| &= |a_{k,S_k}(\mathcal{R}_{S_k}g, v) - \langle S_k\mathcal{R}_{S_k}g, v\rangle| \\
&= |\langle g, v\rangle - \langle S_k g, v\rangle| \\
&\leq \big(\|g\|_{\mathcal{Z}_{j-2}} + \|S_k\mathcal{R}_{S_k}g\|_{\mathcal{Z}_{j-2}}\big)\|v\|_{(\mathcal{Z}_{j-2})^*} \\
&\leq \big(\|g\|_{\mathcal{Z}_{j-2}} + C\|\mathcal{R}_{S_k}g\|_{\mathcal{H}_0}\big)\|v\|_{(\mathcal{Z}_{j-2})^*} \\
&\leq \big(\|g\|_{\mathcal{Z}_{j-2}} + C\|g\|_{\mathcal{H}^*}\big)\|v\|_{(\mathcal{Z}_{j-2})^*} \\
&\leq C\|g\|_{\mathcal{Z}_{j-2}}\|v\|_{(\mathcal{Z}_{j-2})^*},
\end{aligned}$$

which together with (6.21) implies (6.25). □

Next we bound the adjoint approximability constant under the following additional elliptic-regularity assumption.

Assumption 6.23 (Abstract adjoint elliptic regularity for a_k). Let $\mathcal{Z}_0 := \mathcal{H}_0$, $\mathcal{Z}_1 := \mathcal{H}$, and $\mathcal{Z}_j \subset \mathcal{Z}_{j-1}$ for $j = 2, 3, \dots, \ell+1$ such that $\mathcal{Z}_j$ is dense in $\mathcal{Z}_{j-1}$. For all $k_0 \geq 0$, there are $C_{\text{ell}} > 0$ such that for all $u \in \mathcal{H}$ and $k > k_0$,

$$\|u\|_{\mathcal{Z}_j} \leq C_{\text{ell}}\left(\|u\|_{\mathcal{H}_0} + \sup_{v\in\mathcal{H},\, \|v\|_{(\mathcal{Z}_{j-2})^*}=1} |a_k(v,u)|\right), \quad j = 2, \dots, \ell+1. \tag{6.28}$$

Lemma 6.24 (Bound on η (6.4)). Suppose that $a_k : \mathcal{H}\times\mathcal{H} \to \mathbb{C}$ is a sesquilinear form satisfying Assumptions 6.17, 6.18, 6.19 and 6.23 for some $\ell \in \mathbb{Z}^+$. Let $k_0 > 0$. Then there exists $C > 0$ such that for all $k > k_0$ with a_k invertible, all $\mathcal{V}_h \subset \mathcal{H}$, and $m \in \{-1, \dots, \ell\}$,

$$\eta_{\mathcal{Z}_{m+1}\to\mathcal{H}}(\mathcal{V}_h) \leq C\big(\|(I-\Pi_h)\|_{\mathcal{Z}_{\max\{m+3,\ell+1\}}\to\mathcal{H}} + \|(I-\Pi_h)\|_{\mathcal{Z}_{\ell+1}\to\mathcal{H}}\|\mathcal{R}\|_{\mathcal{Z}_{\ell+1}\to(\mathcal{Z}_{m+1})^*}\big) \tag{6.29}$$

and

$$\eta_{\mathcal{H}^*\to\mathcal{H}}(\mathcal{V}_h) \leq C\big(1 + \|(I-\Pi_h)\|_{\mathcal{Z}_{\ell+1}\to\mathcal{H}}\|\mathcal{R}\|_{\mathcal{Z}_{\ell+1}\to\mathcal{H}}\big).$$

Proof. Let $\epsilon > 0$ and let S_k be the operator constructed in Theorem 6.21. With P_k the operator associated with a_k via Remark 6.4,

$$(P_k^* + S_k)\mathcal{R}^* = I + S_k\mathcal{R}^*,$$

so that

$$\mathcal{R}^* = \mathcal{R}_{S_k}^* + \mathcal{R}_{S_k}^* S_k\mathcal{R}^*. \tag{6.30}$$

Arguing as in step (3) of the proof of Theorem 6.21, but using Assumption 6.23

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

(instead of Assumption 6.19), yields

$$\|\mathcal{R}^*_{S_k}\|_{\mathcal{Z}_j\to\mathcal{Z}_{j+2}} \le C, \quad 0 \le j \le \ell - 1.$$

Using this in (6.30), we obtain

$$\begin{aligned}
&\eta_{\mathcal{Z}_{m+1}\to\mathcal{H}}(\mathcal{V}_h)\\
&\quad = \|(I-\Pi_h)\mathcal{R}^*\|_{\mathcal{Z}_{m+1}\to\mathcal{H}}\\
&\quad \le C\|(I-\Pi_h)\|_{\mathcal{Z}_{\max\{m+3,\ell+1\}}\to\mathcal{H}} + C\|(I-\Pi_h)\|_{\mathcal{Z}_{\ell+1}\to\mathcal{H}}\|\mathcal{R}^*\|_{\mathcal{Z}_{m+1}\to(\mathcal{Z}_{\ell+1})^*}\\
&\quad = C\|(I-\Pi_h)\|_{\mathcal{Z}_{\max\{m+3,\ell+1\}}\to\mathcal{H}} + C\|(I-\Pi_h)\|_{\mathcal{Z}_{\ell+1}\to\mathcal{H}}\|\mathcal{R}\|_{\mathcal{Z}_{\ell+1}\to(\mathcal{Z}_{m+1})^*}
\end{aligned}$$

as claimed. The proof of the bound on $\eta_{\mathcal{H}^*\to\mathcal{H}}(\mathcal{V}_h)$ is very similar, using that $\|\mathcal{R}^*_{S_k}\|_{\mathcal{H}^*\to\mathcal{H}} \le C$ by coercivity of a_{k,S_k} (6.23). □

Lemma 6.25 (Oscillatory data implies oscillatory solution). Suppose that $a_k\colon \mathcal{H}\times\mathcal{H}\to\mathbb{C}$ is a sesquilinear form satisfying Assumption 6.19 for some $\ell\in\mathbb{Z}^+$. If $f\in\mathcal{Z}_{\ell-1}$ satisfies

$$\|f\|_{\mathcal{Z}_{\ell-1}} \le C_{\mathrm{osc}}\|f\|_{\mathcal{H}^*},$$

then the solution $u\in\mathcal{H}$ to $a_k(u,v)=\langle f,v\rangle$ for all $v\in\mathcal{H}$ satisfies

$$\|u\|_{\mathcal{Z}_\ell} \le C_{\mathrm{ell}}(1+C_{\mathrm{osc}}\|a_k\|)\|u\|_{\mathcal{H}}.$$

Proof. By (6.21) and the fact that $a(u,v)=\langle g,v\rangle$ for all $v\in\mathcal{H}$,

$$\begin{aligned}
\|u\|_{\mathcal{Z}_{\ell+1}} \le C_{\mathrm{ell}}\big(\|u\|_{\mathcal{H}_0}+\|f\|_{\mathcal{Z}_{\ell-1}}\big) &\le C_{\mathrm{ell}}(\|u\|_{\mathcal{H}_0}+C_{\mathrm{osc}}\|f\|_{\mathcal{H}^*})\\
&\le C_{\mathrm{ell}}(\|u\|_{\mathcal{H}_0}+C_{\mathrm{osc}}\|a_k\|\|u\|_{\mathcal{H}})\\
&\le C_{\mathrm{ell}}(1+C_{\mathrm{osc}}\|a_k\|)\|u\|_{\mathcal{H}}.
\end{aligned}$$

□

Corollary 6.26 (Abstract preasymptotic error bound for the Galerkin method). Suppose that $a_k:\mathcal{H}\times\mathcal{H}\to\mathbb{C}, k\ge 0$ is a family of injective sesquilinear forms satisfying Assumptions 6.17, 6.18, 6.19 and 6.23 for some $\ell\in\mathbb{Z}^+$ and spaces $\mathcal{Z}_0:=\mathcal{H}_0$, $\mathcal{Z}_1:=\mathcal{H}$ and $\mathcal{Z}_j\subset\mathcal{Z}_{j-1}$ for $j=2,3,\dots,\ell+1$ such that $\mathcal{Z}_j$ is dense in $\mathcal{Z}_{j-1}$.

There exists $c,C>0$ such that the following is true. Let $\mathcal{V}_h$ be a finite-dimensional subspace of $\mathcal{H}$. If

$$\|(I-\Pi_h)\|^2_{\mathcal{Z}_{\ell+1}\to\mathcal{H}}\big(1+\|\mathcal{R}\|_{\mathcal{Z}_{\ell+1}\to(\mathcal{Z}_{\ell+1})^*}\big) \le c, \tag{6.31}$$

then for all $u\in\mathcal{H}$, the Galerkin approximation u_h to u in $\mathcal{V}_h$ exists, is unique, and, for all $m\in\{-1,\dots,\ell\}$,

$$\|u-u_h\|_{\mathcal{H}} \le C\big(1+\|(I-\Pi_h)\|_{\mathcal{Z}_{\ell+1}\to\mathcal{H}}\big(1+\|\mathcal{R}\|_{\mathcal{Z}_{\ell+1}\to(\mathcal{Z}_{\ell+1})^*}\big)\big)\|(I-\Pi_h)u\|_{\mathcal{H}} \tag{6.32}$$

and

$$\begin{aligned}
&\|u-u_h\|_{(\mathcal{Z}_{m+1})^*}\\
&\quad \le C\big(\|(I-\Pi_h)\|_{\mathcal{Z}_{\max\{m+3,\ell+1\}}\to\mathcal{H}} + \|(I-\Pi_h)\|_{\mathcal{Z}_{\ell+1}\to\mathcal{H}}\big(1+\|\mathcal{R}\|_{\mathcal{Z}_{\ell+1}\to(\mathcal{Z}_{m+1})^*}\big)\big)\\
&\qquad \times\|(I-\Pi_h)u\|_{\mathcal{H}}.
\end{aligned} \tag{6.33}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Furthermore, given $C_{\mathrm{osc}} > 0$, there exists $C' > 0$ such that with $P\colon \mathcal{H} \to \mathcal{H}^*$ the operator defined by a_k, if

$$\|Pu\|_{\mathcal{Z}_{\ell-1}} \leq C_{\mathrm{osc}}\|Pu\|_{\mathcal{H}^*},$$

then

$$\frac{\|u-u_h\|_{\mathcal{H}}}{\|u\|_{\mathcal{H}}} \leq C'\big(1+\|(I-\Pi_h)\|_{\mathcal{Z}_{\ell+1}\to\mathcal{H}}\big(1+\|\mathcal{R}\|_{\mathcal{Z}_{\ell+1}\to(\mathcal{Z}_{\ell+1})^*}\big)\big)\|(I-\Pi_h)\|_{\mathcal{Z}_{\ell+1}\to\mathcal{H}}. \tag{6.34}$$

Proof. Theorem 6.15 (applied with $\mathcal{Z} = \mathcal{Z}_{\ell+1}$ and $\mathcal{Y} = \mathcal{Z}_{m+1}$, $m \in \{-1,\dots,\ell\}$) and Theorem 6.21 imply that there exist $c, C > 0$ such that if

$$\|(I-\Pi_h)\|_{\mathcal{Z}_{\ell+1}\to\mathcal{H}}\,\eta_{\mathcal{Z}_{\ell+1}\to\mathcal{H}}(\mathcal{V}_h) \leq c, \tag{6.35}$$

then

$$\|u-u_h\|_{\mathcal{H}} \leq C\big(1+\eta_{\mathcal{Z}_{\ell+1}\to\mathcal{H}}(\mathcal{V}_h)\big)\|(I-\Pi_h)u\|_{\mathcal{H}}, \tag{6.36}$$

$$\|u-u_h\|_{(\mathcal{Z}_{m+1})^*} \leq C\eta_{\mathcal{Z}_{m+1}\to\mathcal{H}}(\mathcal{V}_h)\|(I-\Pi_h)u\|_{\mathcal{H}}, \tag{6.37}$$

$$\|u-u_h\|_{(\mathcal{Z}_{\ell+1})^*} \leq C\eta_{\mathcal{Z}_{\ell+1}\to\mathcal{H}}(\mathcal{V}_h)\|(I-\Pi_h)u\|_{\mathcal{H}}, \tag{6.38}$$

where (6.36), (6.37) and (6.38) come from (6.15), (6.16) and (6.17) respectively. Observe that the bound (6.38) is contained in the bound (6.37) when $m = \ell$. The results (6.32) and (6.33) then follow from combining (6.36) and (6.37) with the bound on $\eta_{\mathcal{Z}_{m+1}\to\mathcal{H}}(\mathcal{V}_h)$ from (6.29).

The relative error bound (6.34) then follows from combining Lemma 6.25 with (6.36). □

6.7. *Application to the h-FEM in scattering problems*

The limiting factor in applying the abstract results of Theorems 6.15 and 6.21 (combined in Corollary 6.26) is checking the abstract elliptic-regularity assumptions (Assumptions 6.19, 6.23). We now do this for the Helmholtz PML problem of Definition 4.7, but we emphasize that these elliptic-regularity assumptions are also satisfied for Helmholtz problems *either* with impedance boundary conditions *or* when the exact Dirichlet-to-Neumann map is imposed on the boundary of a large ball; see Galkowski and Spence (2025, §3).

Let

$$\mathcal{H} := H^1_0(\Omega) \quad \text{or} \quad \{v \in H^1(\Omega)\colon v = 0 \text{ on } \Gamma_{\mathrm{tr}}\}, \tag{6.39}$$

depending on whether Γ_- has Dirichlet or Neumann boundary conditions, and let

$$\mathcal{Z}_j := H^j(\Omega\setminus\Gamma_{\mathrm{p}}) \cap \mathcal{H} \tag{6.40}$$

with norm

$$\|v\|_{\mathcal{Z}_j} := \|v\|_{H^j_k(\Omega\setminus\Gamma_{\mathrm{p}})}.$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Theorem 6.27 (Rigorous statement of R1: k-explicit error analysis of the h-FEM). Let $p \in \mathbb{Z}^+$, and let a be the sesquilinear form defining an H^{p+1}-Helmholtz problem approximated using a $C^{p,1}$ PML with $C^{p,1}$ truncation boundary (in the sense of Definitions 4.6 and 4.7). Given $k_0, \epsilon > 0$ and $\Upsilon > 0$, there exist $c, C > 0$ such that for all affine-conforming C^{p+1}-simplicial triangulations $\mathcal{T}$ with constant $\Upsilon > 0$ (in the sense of Definition A.5) such that

$$\partial\Omega \cup \Gamma_{\mathrm{p}} \subset \bigcup_{T \in \mathcal{T}} \partial T,$$

$k \geq k_0$ and $\epsilon \leq \theta \leq \pi/2 - \epsilon$, the following is true. If $h \coloneqq h(\mathcal{T})$ satisfies

$$(hk)^{2p}\rho(k) \leq c, \tag{6.41}$$

then for all $u \in H^1_k$, the Galerkin approximation u_h to u in $\mathcal{P}^p_{\mathcal{T}}$ (defined by (A.3)) exists, is unique, and satisfies

$$\|u - u_h\|_{H^1_k(\Omega)} \leq C(1 + \rho(k)(hk)^p) \min_{v_h \in \mathcal{P}^p_{\mathcal{T}}} \|u - v_h\|_{H^1_k(\Omega)} \tag{6.42}$$

and, for $n \in \{1, \dots, p\}$,

$$\|u - u_h\|_{(\mathcal{Z}_{n-1})^*} \leq C((hk)^n + \rho(k)(hk)^p) \min_{v_h \in \mathcal{P}^p_{\mathcal{T}}} \|u - v_h\|_{H^1_k(\Omega)}. \tag{6.43}$$

Furthermore, given $C_{\mathrm{osc}} > 0$, there exists $C' > 0$ such that if $P\colon H^1_k(\Omega) \to (H^1_k(\Omega))^*$ is the operator defined by a, and if

$$\|Pu\|_{H^{p-1}_k(\Omega\setminus\Gamma_{\mathrm{p}})} \leq C_{\mathrm{osc}}\|Pu\|_{(H^1_k(\Omega))^*}, \tag{6.44}$$

then

$$\frac{\|u - u_h\|_{H^1_k(\Omega)}}{\|u\|_{H^1_k(\Omega)}} \leq C'(1 + \rho(k)(hk)^p)(hk)^p. \tag{6.45}$$

Remark 6.28. Note that $H^j_0(\Omega) \subset \mathcal{Z}_j$, and for all $k_0 > 0$ there exists $C > 0$ such that for $k > k_0$, for $u \in H^j_0$, we have

$$\|u\|_{H^j_k} \leq C\|u\|_{\mathcal{Z}_j}.$$

Therefore $\|u\|_{H^{-j}_k(\Omega)} \leq C\|u\|_{(\mathcal{Z}_j)^*}$ and the bound (6.43) implies bounds on $\|u - u_h\|_{H^{1-n}_k(\Omega)}$ (consistent with (6.17)).

To prove Theorem 6.27, we first recap local elliptic-regularity results near a Dirichlet or Neumann boundary, and for the transmission problem.

Given coefficients A, b and n, let

$$Lu \coloneqq -k^{-2}\operatorname{div}(A\nabla u) - k^{-1}b \cdot \nabla u - nu. \tag{6.46}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Assumption 6.29. For all $x \in \Omega$, $A_{j\ell}(x) = A_{\ell j}(x)$ and

$$\operatorname{Re}(A(x)\xi,\xi)_2 = \operatorname{Re}\sum_{j=1}^{d}\sum_{\ell=1}^{d} A_{j\ell}(x)\xi_\ell\overline{\xi_j} \geq c|\xi|^2 \quad \text{for all } \xi \in \mathbb{C}^d.$$

Theorem 6.30 (Local elliptic regularity near a Dirichlet or Neumann boundary). Let Ω be a Lipschitz domain and let G_1, G_2 be open subsets of $\mathbb{R}^d$ with $G_1 \Subset G_2$ and $G_1 \cap \partial\Omega \neq \emptyset$. Let

$$\Omega_j := G_j \cap \Omega, \quad j = 1,2, \quad \text{and} \quad \Gamma_2 := G_2 \cap \partial\Omega. \tag{6.47}$$

Suppose that A satisfies Assumption 6.29, $A, b, n \in C^{m,1}(\overline{\Omega_2})$, and $\Gamma_2 \in C^{m+1,1}$ for some $m \in \mathbb{N}$. Given $k_0 > 0$, there exists $C > 0$ such that if $k \geq k_0$, $u \in H^1(\Omega_2)$, $Lu \in H^m(\Omega_2)$ and either $u = 0$ or $\partial_{n,A}u = 0$ on Γ_2, then

$$\|u\|_{H^{m+2}_k(\Omega_1)} \leq C\big(\|u\|_{H^1_k(\Omega_2)} + \|Lu\|_{H^m_k(\Omega_2)}\big). \tag{6.48}$$

Proof. The bound in unweighted norms

$$\|u\|_{H^{m+2}(\Omega_1)} \leq C\big(\|u\|_{H^1(\Omega_2)} + \|Lu\|_{H^m(\Omega_2)}\big)$$

is proved in McLean (2000, Theorems 4.7 and 4.18), for example; the proof of the bound

$$\|u\|_{H^{m+2}_k(\Omega_1)} \leq C\big(\|u\|_{H^1_k(\Omega_2)} + \|Lu\|_{H^m_k(\Omega_2)}\big) \tag{6.49}$$

is then very similar. By considering the pairing $\langle Lu, \chi u\rangle_\Omega$, where $\chi \in C^\infty_c(\mathbb{R}^d)$ and $\chi \equiv 1$ on Ω_1, integrating by parts and using Assumption 6.29, we obtain

$$\|u\|_{H^1_k(\Omega_1)} \leq C\big(\|u\|_{L^2(\Omega_2)} + \|Lu\|_{L^2(\Omega_2)}\big).$$

This bound then allows us to replace $\|u\|_{H^1_k(\Omega_2)}$ in (6.48) by $\|u\|_{L^2(\Omega_2)}$ (by introducing an intermediate domain between Ω_1 and Ω_2), and hence obtain the result (6.48). □

Theorem 6.31 (Local elliptic regularity for the transmission problem). Let $\Omega_{\rm in}$ be a Lipschitz domain, and let $\Omega_{\rm out} := \mathbb{R}^d \setminus \overline{\Omega_{\rm in}}$. Let G_1, G_2 be open subsets of $\mathbb{R}^d$ with $G_1 \Subset G_2$ and $G_1 \cap \partial\Omega_{\rm in} \neq \emptyset$. Let

$$\Omega_{\rm in/out,j} := G_j \cap \Omega_{\rm in/out}, \quad j = 1,2, \quad \text{and} \quad \Gamma_2 := G_2 \cap \partial\Omega_{\rm in}.$$

Suppose that A satisfies Assumption 6.29,

$$A|_{\Omega_{\rm in/out,2}}, \quad b|_{\Omega_{\rm in/out,2}}, \quad n|_{\Omega_{\rm in/out,2}} \in C^{m,1}(\overline{\Omega_{\rm in/out,2}}) \quad \text{and} \quad \Gamma_2 \in C^{m+1,1}$$

for some $m \in \mathbb{N}$. Given $k_0 > 0$, there exists $C > 0$ such that if $k \geq k_0$, $u_{\rm in/out} \in H^1(\Omega_{\rm in/out})$, $Lu \in H^m(\Omega_{\rm in/out,2})$, and $u_{\rm in} = u_{\rm out}$ and $\partial_{n,A}u_{\rm in} = \partial_{n,A}u_{\rm out}$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

on Γ_2, then

$$\begin{aligned}&\|u_{\mathrm{in}}\|_{H_k^{m+2}(\Omega_{\mathrm{in},1})} + \|u_{\mathrm{out}}\|_{H_k^{m+2}(\Omega_{\mathrm{out},1})}\\ &\quad\le C\big(\|u_{\mathrm{in}}\|_{H_k^{1}(\Omega_{\mathrm{in},2})} + \|u_{\mathrm{out}}\|_{H_k^{1}(\Omega_{\mathrm{out},2})} + \|Lu_{\mathrm{in}}\|_{H_k^{m}(\Omega_{\mathrm{in},2})} + \|Lu_{\mathrm{out}}\|_{H_k^{m}(\Omega_{\mathrm{out},2})}\big).\end{aligned} \tag{6.50}$$

Proof. The analogue of (6.50) in unweighted norms and with H^1-norms on the right-hand side (instead of L^2-norms) is, for example, Costabel, Dauge and Nicaise (2010, Theorem 5.2.1(i)) and McLean (2000, Theorems 4.7 and 4.20). The proof of the bound in the weighted norms is very similar, and then the H^1_k-norms can be replaced by L^2-norms by considering $\langle Lu, \chi u\rangle_\Omega$, where $\chi \in C_c^\infty(\mathbb{R}^d)$ and $\chi \equiv 1$ on $\Omega_{\mathrm{out},1} \cap \Omega_{\mathrm{in},1}$ (similar to the proof of Theorem 6.30). □

Proof of Theorem 6.27. The plan is to apply Corollary 6.26 (with $\ell = p$) to the sesquilinear form corresponding to the PML problem (i.e. defined by (4.6)); we start by verifying the assumptions of these results, namely that the sesquilinear form is continuous (Assumption 6.17), satisfies a Gårding inequality (Assumption 6.18) and that the abstract elliptic-regularity assumptions (Assumptions 6.19 and 6.23) hold. Without loss of generality we can assume that a is injective, since otherwise $\rho = \infty$ and the theorem is trivially true.

Let $\mathcal{H}$ and $\mathcal{Z}_j$ be defined by (6.39) and (6.40), respectively.

The sesquilinear form (4.6) of the PML problem is continuous, uniformly in k. By Lemmas 4.1 and 4.2, there exists $\omega \in \mathbb{R}$ such that $e^{i\omega}a$ satisfies the Gårding inequality of Assumption 6.18 (with $\omega = 0$ for the first PML formulation in Section 4.4). Since, with $\omega \in \mathbb{R}$,

$$a_k(u - u_h, v_h) = 0 \quad \text{if and only if} \quad e^{i\omega}a_k(u - u_h, v_h) = 0,$$

without loss of generality we can assume that $\omega = 0$.

In both PML formulations, the matrix A_θ (defined by (4.8) and (4.9), respectively) is symmetric. Therefore Assumption 6.29 is satisfied for both P_θ and P_θ^* by Lemmas 4.1 and 4.2 (again noting that we can take $\omega = 0$ without loss of generality).

Definitions 4.6 and 4.7 imply that $\partial\Omega$ is $C^{p,1}$, the coefficients of the PDE are $C^{p-1,1}$ and that Γ_{p} is $C^{p,1}$.

The elliptic-regularity assumptions for a_k (6.21) and its adjoint (6.28) with $\ell = p$ then follow by combining Theorem 6.30 (used near $\partial\Omega_-$ and Γ_{tr}) and Theorem 6.31 (used near Γ_{p}), both with $m = p - 1$. The operator associated with the sesquilinear form $\operatorname{Re} a_k$ is

$$\begin{aligned}&\left(\frac{P + P^*}{2}\right)u\\ &\quad= -k^{-2}\operatorname{div}\left(\frac{A_\theta + \overline{A_\theta}^{\top}}{2}\nabla u\right) + k^{-2}\left(\frac{\nabla\cdot(b_\theta u) - b_\theta\cdot\nabla u}{2}\right) - \left(\frac{n_\theta + \overline{n_\theta}}{2}\right)u.\end{aligned}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Since $A_\theta^\top = A_\theta$ (as noted above), the sesquilinear form $\operatorname{Re} a_k$ also satisfies Assumption 6.29, and the elliptic-regularity assumption for $\operatorname{Re} a_k$ (6.22) holds by the combination of Theorems 6.30 and 6.31 used above.

Finally, since c in both Lemma 4.1 and 4.2 is independent of $\epsilon \le \theta \le \pi/2-\epsilon$ and since the coefficients A_θ, b_θ and n_θ can be bounded uniformly in $\epsilon \le \theta \le \pi/2-\epsilon$, the constant C_ℓ in Assumptions 6.19 and 6.23, and the constants c_{G} and C_{G} in Assumption 6.18, can be taken to be independent of $\epsilon \le \theta \le \pi/2-\epsilon$. Furthermore, the bound on $\|a_k\|$ in Assumption 6.17 can also be taken to be independent of $\epsilon \le \theta \le \pi/2-\epsilon$.

We apply Corollary 6.26 with $\mathcal{V}_h = \mathcal{P}^p_{\mathcal{T}}$. The assumptions on $\mathcal{T}$ and Theorem A.11 imply that there exists $C > 0$ such that for all $m \in \{0,\dots,p\}$,

$$\|(I-\Pi_h)\|_{\mathcal{Z}_{m+1}\to\mathcal{H}} \le C(hk)^m.$$

Furthermore, for $m \in \{-1,\dots,\ell\}$,

$$\|\mathcal{R}\|_{\mathcal{Z}_{\ell+1}\to(\mathcal{Z}_{m+1})^*} \le \|\mathcal{R}\|_{\mathcal{H}_0\to\mathcal{H}_0} = \|\mathcal{R}\|_{L^2(\Omega)\to L^2(\Omega)} \le C\rho,$$

by Theorem 4.4. By these last two displayed inequalities, the bound (6.32) then becomes (6.42). The bound (6.43) with $n \in \{1,\dots,p\}$ then follows from (6.33) applied with $m = n-2$ (so that $m \in \{-1,\dots,p-2\}$ corresponds to $n \in \{1,\dots,p\}$). □

Remark 6.32 (Analysis of the h-FEM including geometric error). Theorem 6.27 assumes that the triangulations resolve both $\partial\Omega$ and any boundaries across which A and n jump. Chaumont-Frelet and Spence (2025) consider the case when these interfaces are approximated to order q by the mesh (so that $q = 1$ for straight elements and $q = p$ for isoparametric elements) and prove that if

$$\rho(k)(hk)^{2p} + \rho(k)h^q \quad \text{is sufficiently small} \tag{6.51}$$

(compare to (6.41)), then the Galerkin solution u_h exists, is unique, and satisfies, for k-oscillatory data (in the sense of (6.44)),

$$\frac{\|u-u_h\|_{H^1_k(\Omega)}}{\|u\|_{H^1_k(\Omega)}} \le C[(1+\rho(k)(hk)^p)(hk)^p + \rho(k)h^q] \tag{6.52}$$

(compare to (6.45)); this result is obtained via proving an abstract analogue of Corollary 6.26 allowing for a variational crime (Chaumont-Frelet and Spence 2025, Theorem 3.2). In (6.52), near $\partial\Omega$, the error $\|u-u_h\|_{H^1_k(\Omega)}$ is measured on the subset of Ω where both u and u_h are well-defined; when A and n are discontinuous, defining precisely where the error $\|u-u_h\|_{H^1_k(\Omega)}$ is measured is more involved (see Chaumont-Frelet and Spence 2025, Theorem 4.11). Observe that (6.51) and (6.52) imply that when the problem is non-trapping (i.e. $\rho(k) \le C$), even with straight elements ($q = 1$), in the limit $k \to \infty$ with h chosen to control the pollution error (via (6.41)), the geometric error is smaller than the pollution error.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Remark 6.33. The analogue of Theorem 6.27 for the PML problem for the time-harmonic Maxwell equations is proved in Chaumont-Frelet, Galkowski and Spence (2024*b*).

7. The hp-FEM does not suffer from the pollution effect (R2)

In this section we analyse the Galerkin error for the hp-FEM. In this case, we require strong regularity assumptions on the coefficients and boundaries in the Helmholtz problem, as well as on the mesh defining the piecewise polynomial spaces. However, under these strong assumptions, one obtains k-uniform quasi-optimality without pollution for piecewise polynomial spaces with high degree.

We saw in Section 6 that both the Schatz duality argument (Lemma 6.9) and the elliptic projection argument from Theorem 6.15 give sufficient conditions for quasioptimality. We use the former here; using the latter would lead to slightly better constants.

The results in this section build on the hp-FEM analyses in Lafontaine *et al.* (2022*b*) and Galkowski *et al.* (2023*b*) (for truncation by the exact Dirichlet-to-Neumann map on a ball) and Galkowski *et al.* (2024*c*) (for PML truncation), although the details of the arguments are different, drawing on the results of Galkowski and Spence (2025) (appearing in Section 6) and Averseng *et al.* (2025).

7.1. An abstract hp-FEM convergence result

The main goal of this section is the following abstract estimate on the adjoint approximability constant. Let

$$\|u\|^2_{H^p_k(\Omega)} := \sum_{\ell=0}^{p} \frac{1}{(\ell!)^2} |u|^2_{H^\ell_k(\Omega)}, \tag{7.1}$$

where

$$|u|^2_{H^p_k(\Omega)} := \sum_{\substack{|\alpha|=p \\ \alpha\in\mathbb{N}^d}} \frac{p!}{\alpha!} \|(k^{-1}D_{x_1})^{\alpha_1}(k^{-1}D_{x_2})^{\alpha_2}\cdots(k^{-1}D_{x_d})^{\alpha_d} u\|^2_{L^2(\Omega)}$$

(observe that this definition is the same as in (3.2)). The key point about the constants in the seminorm definition is that

$$|u|^2_{H^p_k(\mathbb{R}^d)} := \frac{1}{(2\pi)^d} \||\xi|^p \mathcal{F}_k(u)(\xi)\|^2_{L^2(\mathbb{R}^d)}, \tag{7.2}$$

and similarly for $|\cdot|_{H^p(\mathbb{R}^d)}$. We let $|\cdot|_{H^p(\Omega)} := |\cdot|_{H^p_1(\Omega)}$ and $\|\cdot\|_{H^p(\Omega)} := \|\cdot\|_{H^p_1(\Omega)}$.

Theorem 7.1 (hp-FEM quasioptimality under an abstract assumption on low frequencies near interfaces). Suppose a is the sesquilinear form defining an H^∞-Helmholtz problem approximated by a C^∞ PML with $C^{1,1}$ truncation boundary (in the sense of Definitions 4.6 and 4.7). There exists $\psi \in C^\infty_c(\mathbb{R})$ such that for all

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

$W \subset \overline{\Omega}$ open in the subspace topology of $\overline{\Omega}$ with $\Gamma_{\mathrm{p}} \cup \Gamma_{-} \Subset W$, given $M, N, k_0 > 0$ and $\Upsilon > 0$, there exist $C, h_0 > 0$ such that for all affine-conforming C^ω-simplicial triangulations $\mathcal{T}$ with constant $\Upsilon > 0$ (in the sense of Definition A.5) such that

$$\partial\Omega \cup \Gamma_{\mathrm{p}} \subset \bigcup_{T\in\mathcal{T}} \partial T, \tag{7.3}$$

$p \geq 1$, $0 < h(\mathcal{T}) < h_0$, and $k \geq k_0$,

$$\begin{aligned}
&\eta_{L^2\to H^1_k}(\mathcal{P}^p_{\mathcal{T}}) \\
&\quad \leq C\left[\frac{hk}{p} + \left(\frac{hk}{p}\right)^M k^{-N}\rho(k)\right. \\
&\qquad \left. + C^p (hk)^p k^{-p-1}\left(\max\left\{\frac{k}{p+1}, 1\right\}^{p+1} + \|\psi(\mathcal{P})\|_{L^2\to H^{p+1}(W\setminus\Gamma_{\mathrm{p}})\cap H^1_k}\right)\rho(k)\right],
\end{aligned} \tag{7.4}$$

where $h := h(\mathcal{T})$.

The ideas of the proof. The estimate (7.4) is obtained by splitting the adjoint solution operator $\mathcal{R}^*$ into its action on high and low frequencies and using the estimates from Theorem A.11:

$$\|I - \mathcal{C}^p_{\mathcal{T}}\|_{H^m_k(\Omega\setminus\Gamma_{\mathrm{p}})\cap H^1_k\to H^1_k} \leq C_m\left(\frac{hk}{p}\right)^{m-1}, \tag{7.5}$$

$$\|I - \mathcal{C}^p_{\mathcal{T}}\|_{H^{p+1}(\Omega\setminus\Gamma_{\mathrm{p}})\cap H^1_k\to H^1_k} \leq C^{p+1}(hk)^p k^{-p-1}, \tag{7.6}$$

where $\mathcal{C}^p_{\mathcal{T}}$ is a triangulation-preserving operator in the sense of Definition A.10 below.

The main idea is that P is semiclassically elliptic at high frequencies and hence $\mathcal{R}^*\colon L^2 \to H^2_k$ is bounded uniformly in k there. On the other hand, functions with frequency $\lesssim k$ are highly regular (with quantitative bounds on the derivatives) and thus should be well approximated by piecewise polynomials.

The key question is then: How should one define high and low frequencies to make this heuristic argument rigorous? Below, we use two different answers to this question.

Answer 1. The first answer (and the best answer when it is available) is that a function with low frequencies is any function whose (semiclassical) Fourier transform is compactly supported; that is, there exists $\psi \in C^\infty_c(\mathbb{R})$ such that $u = \psi(|\hbar D|)u$, where $D := -i\partial$. For $\psi \in C^\infty_c(\mathbb{R})$, by (7.2) with $k = 1$,

$$(\psi(|\hbar D|)u)(x) := \mathcal{F}^{-1}_\hbar(\psi(|\bullet|)(\mathcal{F}u)(\bullet))(x) \tag{7.7}$$

satisfies

$$|\psi(|\hbar D|)u|_{H^\ell(\mathbb{R}^d)} \leq (Ck)^\ell \|u\|_{L^2(\mathbb{R}^d)},$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

so that by the definition of $\|\cdot\|_{H^p(\mathbb{R}^d)}$,

$$\|\psi(|\hbar D|)\|_{L^2\to H^{p+1}} \le \sqrt{\sum_{\ell=0}^{p+1}\frac{1}{(\ell!)^2}(Ck)^{2\ell}} \le \sum_{\ell=0}^{p+1}\frac{1}{\ell!}(Ck)^{\ell}. \tag{7.8}$$

We show below (in the proof of Theorem 7.1) that by Stirling's approximation,

$$\|\psi(|\hbar D|)\|_{L^2\to H^{p+1}} \le C^{p+1}\max\left\{\frac{k}{p+1},1\right\}^{p+1}. \tag{7.9}$$

Combining this with (7.6), we obtain

$$\begin{aligned}\|(I-\Pi_h^p)\psi(|\hbar D|)\|_{L^2\to H^1_k} &\le C^{p+1}(hk)^p k^{-p-1}\max\left\{\frac{k}{p+1},1\right\}^{p+1}\\ &\le C^{p+1}\left(\frac{hk}{p+1}\right)^p \quad \text{when } p+1\le k;\end{aligned}$$

that is, low frequencies are very well approximated by piecewise polynomials with both h and hk/p small and p large. The main issue to resolve is then: Is there $\psi\in C_c^\infty(\mathbb{R})$ such that $(1-\psi(|\hbar D|))\mathcal{R}^*$ is uniformly bounded? When there are no boundaries or interfaces, semiclassical ellipticity (together with a bound of the form $\rho(k)\le Ck^L$ for some L) provides a positive answer to this question, even when all coefficients are only smooth (with this first shown in Lafontaine *et al.* 2022*b*). However, in the presence of boundaries or interfaces, one cannot work with Fourier multipliers and one needs a second definition of low frequencies.

Answer 2. When the Fourier transform is unavailable, it is natural to define low frequencies in terms of the operator P, or more precisely its real part, $\mathcal{P}$. In this case we say a function is low-frequency if there exists $\psi\in C_c^\infty(\mathbb{R})$ such that $u=\psi(\mathcal{P})u$, that is, u can be written as a sum $\sum_{j=1}^{N_k}a_j u_{\lambda_j}$, of eigenfunctions of $\mathcal{P}$ with eigenvalue, $\lambda_j\le C$. In this case, it is relatively straightforward to check that there exists $\psi\in C_c^\infty(\mathbb{R})$ such that $(1-\psi(\mathcal{P}))\mathcal{R}^*$ is uniformly bounded.

To prove Theorem 7.1 we combine Answers 1 and 2, using Answer 1 away from all boundaries and interfaces, and Answer 2 in a neighbourhood of all non-PML boundaries or interfaces. Finally, near the truncation boundary, we use that for $\chi\in\overline{C^\infty}(\Omega)$ supported inside the PML, $\chi\mathcal{R}^*\colon L^2\to H^2_k$ is uniformly bounded.

7.2. *Application of Theorem 7.1*

To apply Theorem 7.1 to concrete settings, one must estimate $\|\psi(\mathcal{P})\|_{L^2\to H^{p+1}(W)}$. We show below in Theorem 7.7 that if $\Gamma_{\mathrm{p}}=\emptyset$, $\Gamma_-\in C^\omega$, and A and n are analytic in a neighbourhood of Γ_-, then for $\psi\in C_c^\infty(\mathbb{R})$,

$$\|\psi(\mathcal{P})\|_{L^2\to H^{p+1}(W)} \le \left(C\max\left\{\frac{k}{p+1},1\right\}\right)^{p+1}, \tag{7.10}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

i.e. the same bound as in (7.9). Combining (7.10) with (7.4), we obtain the rigorous statement of Informal Theorem 3.2.

Theorem 7.2 (Rigorous statement of R2: hp-FEM quasioptimality for analytic obstacles with variable coefficients). Suppose a is the sesquilinear form defining an H^∞-Helmholtz problem approximated using a C^∞ PML with $C^{1,1}$ truncation boundary (in the sense of Definitions 4.6 and 4.7) and, in addition, suppose that $\Gamma_{\rm p} = \emptyset$, $\Gamma_- \in C^\omega$, a Dirichlet condition is imposed on Γ_-, and there is a neighbourhood, U of Γ_-, such that $A, n \in C^\omega(U)$.

(i) Given $M, N, k_0 > 0$ and $\Upsilon > 0$, there exists $C, h_0 > 0$ such that for all affine-conforming C^ω-simplicial triangulations $\mathcal{T}$ with constant $\Upsilon > 0$ (in the sense of Definition A.8) such that

$$\partial\Omega \subset \bigcup_{T\in\mathcal{T}} \partial T, \tag{7.11}$$

$p \geq 1$, $0 < h(\mathcal{T}) < h_0$, and $k \geq k_0$,

$$\eta_{L^2\to H^1_k}(\mathcal{P}^p_{\mathcal{T}}) \leq C\left[\frac{hk}{p} + \left(\frac{hk}{p}\right)^M k^{-N}\rho(k) + C^p(hk)^p k^{-p-1}\max\left\{\frac{k}{p+1}, 1\right\}^{p+1}\rho(k)\right],$$

where $h := h(\mathcal{T})$.

(ii) Given $k_0 > 0$ $\epsilon > 0$, and $\Upsilon > 0$, for any $\mathcal{J}$ and $L > 0$ such that

$$\sup\{k^{-L}\rho(k)\colon k \in (k_0,\infty)\setminus\mathcal{J}\} < \infty,$$

there exist $C, c, h_0 > 0$ such that for all $k \geq k_0$, $k \notin \mathcal{J}$, C^ω-simplicial triangulations $\mathcal{T}$ with constant $\Upsilon > 0$ such that (7.11) holds,

$$0 < h(\mathcal{T}) < h_0, \quad \frac{h(\mathcal{T})k}{p} < c \quad \text{and} \quad p \geq 1 + \epsilon\log k,$$

then for all $u \in H^1_k(\Omega)$, the Galerkin approximation u_h in $\mathcal{P}^p_{\mathcal{T}}$ satisfies

$$\|u - u_h\|_{H^1_k(\Omega)} \leq C \min_{w_h\in\mathcal{P}^p_{\mathcal{T}}} \|u - w_h\|_{H^1_k(\Omega)}.$$

7.3. Recap of pseudolocality results from Averseng et al. (2025)

To prove Theorem 7.1, we need results on pseudolocality of functions of a self-adjoint operator. To do this, with P_θ defined by (4.5) and $\mathcal{P} = \operatorname{Re} P_\theta$, we fix $\psi_0 \in C^\infty_c(\mathbb{R})$ such that there exists $c > 0$ such that, with $S_k := \psi_0(\mathcal{P})$,

$$\operatorname{Re}\langle (P_\theta + S_k)v, v\rangle \geq c\|v\|^2_{H^1_k}, \quad \text{for all } v \in H^1_k(\Omega).$$

As in (6.13), let $\mathcal{R}_{S_k} := (P_\theta + S_k)^{-1}$.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Before we state the pseudolocality results, we introduce the following notation.

Definition 7.3. Let X and Y be Banach spaces, $f\colon (0,\infty)\to(0,\infty)$, and $B_k\colon X\to Y$ a k-dependent operator. Then

$$B_k = O(f(k))_{X\to Y}$$

if, for all $k_0>0$, there exists $C>0$ such that

$$\|B_k\|_{X\to Y}\le Cf(k),\quad k>k_0.$$

Furthermore,

$$B_k = O(k^{-\infty})_{X\to Y}$$

if, for all $k_0>0$ and $N>0$, there exists $C>0$ such that

$$B_k = O(k^{-N})_{X\to Y}.$$

Lemma 7.4 (Spatial pseudolocality (Averseng *et al.* 2025, Theorems 5.27 and 6.2)). Suppose that $\psi\in\mathcal{S}(\mathbb{R})$ and $\chi_1,\chi_2\in\overline{C^\infty}(\Omega)$ with $\operatorname{supp}\chi_1\cap\operatorname{supp}\chi_2=\emptyset$. Then, for all $N>0$,

$$\begin{aligned}\chi_1\psi(\mathcal{P})\chi_2 &= O(k^{-\infty})_{(H^N_k(\Omega\setminus\Gamma_{\mathrm{p}}))^*\to H^N_k(\Omega\setminus\Gamma_{\mathrm{p}})},\\ \chi_1\mathcal{R}_{S_k}\chi_2 &= O(k^{-\infty})_{(H^N_k(\Omega\setminus\Gamma_{\mathrm{p}}))^*\to H^N_k(\Omega\setminus\Gamma_{\mathrm{p}})},\\ \chi_1\mathcal{R}^*_{S_k}\chi_2 &= O(k^{-\infty})_{(H^N_k(\Omega\setminus\Gamma_{\mathrm{p}}))^*\to H^N_k(\Omega\setminus\Gamma_{\mathrm{p}})}.\end{aligned}$$

Lemma 7.5 (Pseudolocality in frequency (Averseng *et al.* 2025, Theorem 5.37 and Lemmas 6.7 and 6.9)). Let $\psi_1,\psi_2\in C^\infty(\mathbb{R})$ with either $\psi_1\in C^\infty_c(\mathbb{R})$ or $\psi_2\in C^\infty_c(\mathbb{R})$, and $\operatorname{supp}\psi_1\cap\operatorname{supp}\psi_2=\emptyset$. Suppose that $\chi\in C^\infty(\Omega)$ and $\operatorname{supp}\nabla\chi\cap(\partial\Omega\cup\Gamma_{\mathrm{p}})=\emptyset$, and $\operatorname{supp}\chi\cap\Gamma_{\mathrm{tr}}=\emptyset$. Then, for all $N>0$,

$$\begin{aligned}\psi_1(\mathcal{P})\chi\psi_2(\mathcal{P}) &= O(k^{-\infty})_{\mathcal{D}(\mathcal{P}^{-N})\to\mathcal{D}(\mathcal{P}^N)},\\ \psi_1(\mathcal{P})\chi\mathcal{R}_{S_k}\chi\psi_2(\mathcal{P}) &= O(k^{-\infty})_{\mathcal{D}(\mathcal{P}^{-N})\to\mathcal{D}(\mathcal{P}^N)},\\ \psi_1(\mathcal{P})\chi\mathcal{R}^*_{S_k}\chi\psi_2(\mathcal{P}) &= O(k^{-\infty})_{\mathcal{D}(\mathcal{P}^{-N})\to\mathcal{D}(\mathcal{P}^N)}.\end{aligned}$$

The ideas behind the proofs of Lemmas 7.4 and 7.5. Theorems 5.27 and 5.37, respectively, of Averseng *et al.* (2025) provide abstract conditions under which S_k, $\mathcal{R}_{S_k}$ and $\mathcal{R}^*_{S_k}$ are pseudolocal in space and for which an operator B acts pseudolocally in frequency. These conditions boil down to estimating repeated commutators of either spatial cut-offs with P_θ and P^*_θ (i.e. commutators of the form $[\chi,[\chi,[\chi,\dots[\chi,P_\theta]]]]$) in the case of spatial pseudolocality, and estimating repeated commutators of P_θ and P^*_θ with B (i.e. commutators of the form $[P_\theta,[P_\theta,[P_\theta,\dots[P_\theta,B]]]]$) in the case of frequency pseudolocality. These abstract conditions are checked in Averseng *et al.* (2025, Lemma 6.2) for spatial pseudolocality and in Averseng *et al.* (2025, Lemmas 6.7 and 6.9) for frequency pseudolocality.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

To see why commutators arise in the case of spatial pseudolocality of $S_k := \psi_0(\mathcal{P})$, first notice that if $\operatorname{supp}\chi_1 \cap \operatorname{supp}\chi_2 = \emptyset$, then there exists $\chi \in \overline{C^\infty}(\Omega)$ such that $\operatorname{supp}\chi \cap \operatorname{supp}\chi_2 = \emptyset$ and $\operatorname{supp}(1-\chi) \cap \operatorname{supp}\chi_1 = \emptyset$, and hence

$$\chi_1 S_k \chi_2 = \chi_1 \chi S_k \chi_2 = \chi_1 [\chi, S_k] \chi_2.$$

Repeating this argument, we have

$$\chi_1 S_k \chi_2 = \chi_1 [\chi, [\chi, [\chi, \dots [\chi, S_k]]]] \chi_2.$$

Next, recall from the Helffer–Sjöstrand formula (see e.g. Dimassi and Sjöstrand 1999, Theorem 8.1) that for all $f \in \mathcal{S}(\mathbb{R})$, then

$$f(\mathcal{P}) = \frac{1}{\pi} \int_{\mathbb{C}} \partial_{\bar z} \tilde f(z) (\mathcal{P} - z)^{-1} \, dm_{\mathbb{C}}(z),$$

where $dm_{\mathbb{C}}(x+iy) = dx\, dy$, $\partial_{\bar z} := \frac{1}{2}(\partial_x + i\partial_y)$ and $\tilde f$ is an almost analytic extension of f in the sense of Zworski (2012, Theorem 3.6).

Then, since

$$[\chi, (\mathcal{P} - z)^{-1}] = -(\mathcal{P} - z)^{-1} [\chi, \mathcal{P}] (\mathcal{P} - z)^{-1}, \tag{7.12}$$

and $\mathcal{P} = \frac{1}{2}(P_\theta + P_\theta^*)$, iterated commutators of χ with P_θ and P_θ^* appear naturally.

Analogously, frequency pseudolocality involves repeated commutators of a frequency cut-off, $\psi(\mathcal{P})$, with the operator B and hence, again via the Helffer–Sjöstrand formula, can be reduced to iterated commutators with $(\mathcal{P} - z)^{-1}$ and finally to iterated commutators with $\mathcal{P}$ via (7.12) with χ replaced by B.

7.3.1. Bound on the high frequencies of the solution operator

We now use the pseudolocality results of the previous subsection (i.e. Lemmas 7.4 and 7.5) to show that the (adjoint) solution operator behaves well on high frequencies as measured by $\psi(\mathcal{P})$.

Lemma 7.6 (High-frequency elliptic-type estimate). Given $\psi_0 \in C_c^\infty$, let $\psi \in C_c^\infty(\mathbb{R})$ be such that $\operatorname{supp}(1-\psi) \cap \operatorname{supp}\psi_0 = \emptyset$ and $\chi \in \overline{C^\infty}(\Omega)$ such that $\operatorname{supp}\chi \cap \Gamma_{\mathrm{tr}} = \emptyset$, $\operatorname{supp}\nabla\chi \cap (\Gamma_{\mathrm{p}} \cup \Gamma_-) = \emptyset$. Then, for all $N > 0$ $k_0 > 0$, there exists $C > 0$ such that for all $k > k_0$,

$$\|(1 - \psi(\mathcal{P}))\chi \mathcal{R}^*\|_{L^2 \to H_k^2(\Omega\setminus\Gamma_{\mathrm{p}}) \cap H_k^1(\Omega)} \le C\left(1 + k^{-N} \|\mathcal{R}^*\|_{L^2 \to L^2}\right). \tag{7.13}$$

Proof. By (6.30),

$$\mathcal{R}^* = \mathcal{R}_{S_k}^* + \mathcal{R}_{S_k}^* \psi_0(\mathcal{P}) \mathcal{R}^*,$$

so that

$$(1 - \psi(\mathcal{P}))\chi \mathcal{R}^* = (1 - \psi(\mathcal{P}))\chi \mathcal{R}_{S_k}^* + (1 - \psi(\mathcal{P}))\chi \mathcal{R}_{S_k}^* \psi_0(\mathcal{P}) \mathcal{R}^*. \tag{7.14}$$

Let $\tilde\chi \in \overline{C^\infty}(\Omega)$ with $\operatorname{supp}\nabla\tilde\chi \cap (\Gamma_{\mathrm{p}} \cup \partial\Omega) = \emptyset$, $\operatorname{supp}(1-\tilde\chi) \cap \operatorname{supp}\chi = \emptyset$ and

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

$\operatorname{supp}\tilde{\chi} \cap \Gamma_{\mathrm{tr}} = \emptyset$. Then, by Lemma 7.4,

$$\begin{aligned}&(1-\psi(\mathcal{P}))\chi\mathcal{R}^*_{S_k}\psi_0(\mathcal{P})\mathcal{R}^*\\&\quad=(1-\psi(\mathcal{P}))\chi\mathcal{R}^*_{S_k}\tilde{\chi}\psi_0(\mathcal{P})\mathcal{R}^*+(1-\psi(\mathcal{P}))O(k^{-\infty})_{L^2\to H^N_k(\Omega\setminus\Gamma_{\mathrm{p}})}\psi_0(\mathcal{P})\mathcal{R}^*.\end{aligned}\tag{7.15}$$

Now let $\psi_1 \in C_c^\infty(\mathbb{R})$ be such that $\operatorname{supp}\psi_0 \cap \operatorname{supp}(1-\psi_1) = \emptyset$ and $\operatorname{supp}\psi_1 \cap \operatorname{supp}(1-\psi) = \emptyset$. By Lemma 7.5 and then Lemma 7.4,

$$\begin{aligned}(1-\psi(\mathcal{P}))\chi\mathcal{R}^*_{S_k}\tilde{\chi}\psi_0(\mathcal{P})\mathcal{R}^* &= (1-\psi(\mathcal{P}))\chi(1-\psi_1(\mathcal{P}))\mathcal{R}^*_{S_k}\tilde{\chi}\psi_0(\mathcal{P})\mathcal{R}^*\\&\quad+O(k^{-\infty})_{\mathcal{D}(\mathcal{P}^{-N})\to\mathcal{D}(\mathcal{P}^N)}\mathcal{R}^*_{S_k}\tilde{\chi}\psi_0(\mathcal{P})\mathcal{R}^*\\&=(1-\psi(\mathcal{P}))\chi(1-\psi_1(\mathcal{P}))\tilde{\chi}\mathcal{R}^*_{S_k}\tilde{\chi}\psi_0(\mathcal{P})\mathcal{R}^*\\&\quad+O(k^{-\infty})_{L^2\to H^N_k(\Omega\setminus\Gamma_{\mathrm{p}})}\mathcal{R}^*_{S_k}\tilde{\chi}\psi_0(\mathcal{P})\mathcal{R}^*\end{aligned}\tag{7.16}$$

(where we have combined the remainder terms). Finally, by Lemma 7.5,

$$\begin{aligned}&(1-\psi(\mathcal{P}))\chi(1-\psi_1(\mathcal{P}))\tilde{\chi}\mathcal{R}^*_{S_k}\tilde{\chi}\psi_0(\mathcal{P})\mathcal{R}^*\\&\quad=(1-\psi(\mathcal{P}))\chi O(k^{-\infty})_{\mathcal{D}(\mathcal{P}^{-N})\to\mathcal{D}(\mathcal{P}^N)}\psi_0(\mathcal{P})\mathcal{R}^*.\end{aligned}\tag{7.17}$$

The result then follows by combining (7.14), (7.15), (7.16) and (7.17). □

7.4. *Proof of Theorem 7.1*

Proof. Recall from the text underneath (7.9) that the plan is to split into high and low frequencies using (i) the Fourier transform away from all boundaries and interfaces, and (ii) $\psi(\mathcal{P})$ near all non-PML boundaries and interfaces. The contribution from the PML is dealt with using ellipticity of P_θ (and hence good solution-operator bounds) in the PML region.

Let $\varphi_P \in \overline{C^\infty}(\Omega)$ (recall that this notation is defined in Section 4.6) with $\operatorname{supp}\varphi_P$ contained inside the PML and $\operatorname{supp}(1-\varphi_P) \cap \Gamma_{\mathrm{tr}} = \emptyset$. Let $\chi \in C_c^\infty(\Omega)$ with $\operatorname{supp}(1-\chi) \cap \operatorname{supp}(1-\varphi_P) \subset W$ and $\operatorname{supp}\chi \cap \Gamma_p = \emptyset$. Let also $\psi \in C_c^\infty(\mathbb{R})$ with $\operatorname{supp}(1-\psi) \cap [-M, M] = \emptyset$ for some $M > 0$ to be chosen large enough. We then write

$$\mathcal{R}^* = \chi\mathcal{R}^* + (1-\chi)\mathcal{R}^* = ((1-\psi(|\hbar D|))\chi\mathcal{R}^* + \psi(|\hbar D|)\chi\mathcal{R}^* + (1-\chi)\mathcal{R}^*.$$

Since the coefficients of P_θ are smooth on $\operatorname{supp}\chi \Subset \Omega$, by the semiclassical ellipticity of P_θ on high frequencies and the elliptic estimate (Theorem 5.7),

$$(1-\psi(|\hbar D|))\chi\mathcal{R}^* = \mathcal{H} + \mathcal{L},\tag{7.18}$$

where

$$\|\mathcal{H}\|_{L^2(\Omega)\to H^2_k(\Omega)} \le C \quad\text{and}\quad \|\mathcal{L}\|_{L^2(\Omega)\to H^N_k(\Omega)} \le C\hbar^N\rho(k).$$

Indeed, let $\chi_\Omega \in C_c^\infty(\mathbb{R}^d)$ with $\operatorname{supp}(1-\chi_\Omega) \cap \Omega = \emptyset$ and $\chi_i \in C_c^\infty(\Omega)$, $i = 1, 2$ with $\operatorname{supp}(1-\chi_i) \cap \operatorname{supp}\chi_{i-1} = \emptyset$ and $\chi_0 := \chi$. Since a Fourier multiplier (such

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

as $\psi(|\hbar D|)$ (7.7)) can be written in terms of the quantization $\widetilde{\mathrm{Op}_\hbar}$ (5.1), the bound (5.4) implies that

$$(1-\psi(|\hbar D|))\chi = \mathrm{Op}_\hbar(a) + O(\hbar^\infty)_{\Psi_\hbar^{-\infty}}, \tag{7.19}$$

where $a \in C_c^\infty(\mathbb{R}^{2d})$ with $\operatorname{supp} a \subset \{(x,\xi)\colon x \in \operatorname{supp}\chi,\ \xi \in \operatorname{supp}(1-\psi)\}$. Then, by Theorem 5.7 applied with P_θ replaced by P_θ^*, $b = \chi_1$, and u replaced by $\chi_2\mathcal{R}^* f$, there exists $e \in S^{-2}(\mathbb{R}^d)$ such that

$$\|\chi_\Omega(\mathrm{Op}_\hbar(a) - \mathrm{Op}_\hbar(e)\chi_1 P_\theta^*)\chi_2\mathcal{R}^* f\|_{H_k^N} \le C_N\hbar^N\|\chi_2\mathcal{R}^* f\|_{H_k^{-N}}. \tag{7.20}$$

Now, with $\mathscr{R} = O(\hbar^\infty)_{\Psi_\hbar^{-\infty}}$, by (7.20),

$$\begin{aligned}\chi_\Omega(1-\psi(|\hbar D|))\chi\mathcal{R}^* f &= \chi_\Omega(1-\psi(|\hbar D|))\chi\chi_2\mathcal{R}^* f\\ &= \chi_\Omega\,\mathrm{Op}_\hbar(a)\chi_2\mathcal{R}^* f + \chi_\Omega\mathscr{R}\chi_2\mathcal{R}^* f.\end{aligned}$$

Therefore, if $\mathcal{H} \coloneqq \mathrm{Op}_\hbar(e)\chi_1$ and

$$\mathcal{L} \coloneqq (\mathrm{Op}_\hbar(a)\chi_2\mathcal{R}^* - \mathrm{Op}_\hbar(e)\chi_1) + \mathscr{R}\chi_2\mathcal{R}^*,$$

then the mapping properties in (7.18) hold by (7.19) and the boundedness property (iii) of Informal Theorem 5.5.

Combining (7.18) with (7.6), (7.5) and (7.8), we obtain

$$\begin{aligned}&\|(I-\Pi_h^p)\chi\mathcal{R}^*\|_{L^2\to H_k^1}\\ &\quad\le \|(I-\Pi_h^p)\mathcal{H}\|_{L^2\to H_k^1} + \|(I-\Pi_h^p)\mathcal{L}\|_{L^2\to H_k^1} + \|(I-\Pi_h^p)\psi(|\hbar D|)\chi\mathcal{R}^*\|_{L^2\to H_k^1}\\ &\quad\le C\|(I-\Pi_h^p)\|_{H_k^2\to H_k^1} + C_{N,M}k^{-N}\rho(k)\|(I-\Pi_h^p)\|_{H_k^M\to H_k^1}\\ &\qquad + \|(I-\Pi_h^p)\|_{H^{p+1}\to H_k^1}\rho(k)\sum_{\ell=0}^{p+1}\frac{1}{\ell!}(Ck)^\ell\\ &\quad\le C\frac{hk}{p} + C_{N,M}k^{-N}\rho(k)\left(\frac{hk}{p}\right)^M + C^{p+1}(hk)^p k^{-p-1}\rho(k)\sum_{\ell=0}^{p+1}\frac{1}{\ell!}(Ck)^\ell.\end{aligned}$$

We now show that, as claimed in (7.9) above,

$$\sum_{\ell=0}^{p+1}\frac{1}{\ell!}(Ck)^\ell \le C^{p+1}\max\left\{\frac{k}{p+1},1\right\}^{p+1}, \tag{7.21}$$

so that

$$\begin{aligned}&\|(I-\Pi_h^p)\chi\mathcal{R}^*\|_{L^2\to H_k^1}\\ &\quad\le C\frac{hk}{p} + C_{N,M}k^{-N}\rho(k)\left(\frac{hk}{p}\right)^M\\ &\qquad + C^{p+1}(hk)^p k^{-p-1}C^{p+1}\max\left\{\frac{k}{p+1},1\right\}^{p+1}\rho(k).\end{aligned} \tag{7.22}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

First, by Stirling's approximation,

$$\sum_{\ell=0}^{p+1} \frac{1}{\ell!} C^\ell k^\ell \le \sum_{\ell=0}^{p+1} \left(\frac{Ck}{\ell}\right)^\ell.$$

Next, for $Ck/\ell \ge e$ (which is ensured if $Ck \ge p$, with a larger value of C), the function $(Ck/\ell)^\ell$ is increasing. Therefore, if $Ck \ge p$, then

$$\sum_{\ell=0}^{p+1} \left(\frac{Ck}{\ell}\right)^\ell \le (p+2)\left(\frac{Ck}{p+1}\right)^{p+1} \le \left(\frac{Ck}{p+1}\right)^{p+1} \tag{7.23}$$

(where we have increased C in the last inequality). On the other hand, for any $p \ge Ck$,

$$\sum_{\ell=0}^{p+1} \frac{1}{\ell!} C^\ell k^\ell \le \sum_{\ell=0}^{\infty} \frac{1}{\ell!} (Ck)^\ell = e^{Ck} \le C^{p+1},$$

and then (7.21) holds.

To handle $(1-\chi)\mathcal{R}^*$, we write

$$(1-\chi)\mathcal{R}^* = \varphi_P(1-\chi)\mathcal{R}^* + (1-\varphi_P)(1-\chi)\mathcal{R}^*.$$

By ellipticity of P_θ in the PML region (see e.g. Averseng *et al.* 2025, Theorem 4.2 and §C.3),

$$\|\varphi_P(1-\chi)\mathcal{R}^*\|_{L^2\to H^2_k} \le C. \tag{7.24}$$

Thus, by (7.6),

$$\|(I-\Pi_h^p)\varphi_P(1-\chi)\mathcal{R}^*\|_{L^2\to H^1_k} \le C\frac{hk}{p}, \tag{7.25}$$

and it remains to bound $(I-\Pi_h^p)(1-\varphi_P)(1-\chi)\mathcal{R}^*$. To do this, observe that by (7.13), for M large enough,

$$\|(1-\psi(\mathcal{P}))(1-\varphi_P)(1-\chi)\mathcal{R}^*\|_{L^2\to H^2_k(\Omega\setminus\Gamma_{\mathrm{p}})\cap H^1_k(\Omega)} \le C, \tag{7.26}$$

where we have used that $\operatorname{supp}(1-\varphi_P)(1-\chi) \cap \Gamma_{\mathrm{tr}} = \emptyset$ and $(1-\varphi_P)(1-\chi)$ is constant on $\Gamma_{\mathrm{p}} \cup \Gamma_-$. By (7.26) and (7.6),

$$\|(I-\Pi_h^p)(1-\psi(\mathcal{P}))(1-\varphi_P)(1-\chi)\mathcal{R}^*\|_{L^2\to H^1_k} \le C\frac{hk}{p}, \tag{7.27}$$

and it now remains to bound $(I-\Pi_h^p)\psi(\mathcal{P})(1-\varphi_P)(1-\chi)\mathcal{R}^*$. To do this, let $N \ge 0$ and $\chi_0 \in \overline{C^\infty}(\Omega)$ with

$$\operatorname{supp}\chi_0 \subset W \quad \text{and} \quad \operatorname{supp}(1-\chi_0) \cap \operatorname{supp}(1-\chi) \cap \operatorname{supp}(1-\varphi_P) = \emptyset.$$

Then, by Lemma 7.4, for any N and M,

$$\|(1-\chi_0)\psi(\mathcal{P})(1-\varphi_P)(1-\chi)\|_{L^2\to H^{M+1}_k} \le Ck^{-N}. \tag{7.28}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Next, define

$$W_{\mathcal{T}} := \bigcup_{T\in\mathcal{T}\,:\,T\subset W} T.$$

Then let $h_0 > 0$ be small enough such that for $0 < h < h_0$,

$$\operatorname{supp}\chi_0 \subset W_{\mathcal{M}}.$$

By Theorem A.11, applied with $\Omega_{\mathcal{T}_1} = \overline{W_{\mathcal{T}}^c}$, $\Omega_{\mathcal{T}_2} = W_{\mathcal{T}}$, $t = M+1$ and $s = 1$, together with (7.28), we obtain

$$\begin{aligned}
&\|(I-\Pi_h^p)\psi(\mathcal{P})(1-\varphi_P)(1-\chi)\|_{L^2\to H^1_k(\Omega)}\\
&\quad\le C\left(\frac{hk}{p}\right)^M k^{-M-1}\|\psi(\mathcal{P})(1-\varphi_P)(1-\chi)\|_{L^2\to H^{M+1}(W_{\mathcal{T}}^c)}\\
&\qquad + (Chk)^p k^{-p-1}\|\psi(\mathcal{P})\|_{L^2\to H^{p+1}(W_{\mathcal{T}}\setminus\Gamma_{\mathrm{p}})\cap H^1(W_{\mathcal{T}})}\\
&\quad\le C\left(\frac{hk}{p}\right)^M \|(1-\chi_0)\psi(\mathcal{P})(1-\varphi_P)(1-\chi)\|_{L^2\to H^{M+1}_k(W_{\mathcal{T}}^c)}\\
&\qquad + (Chk)^p k^{-p-1}\|\psi(\mathcal{P})\|_{L^2\to H^{p+1}(W\setminus\Gamma_{\mathrm{p}})\cap H^1(W)}\\
&\quad\le C\left(\frac{hk}{p}\right)^M k^{-N} + (Chk)^p k^{-p-1}\|\psi(\mathcal{P})\|_{L^2\to H^{p+1}(W\setminus\Gamma_{\mathrm{p}})\cap H^1(W)}.
\end{aligned}$$

Together with (7.22), (7.25) and (7.27), this implies (7.4).

Strictly speaking, when applying Theorem A.11 we should split the domain into more than two sets – depending on Γ_{p} – and use (7.3). The argument, however, is identical. □

7.5. *Analytic regularity estimates for $\psi(\mathcal{P})$*

Theorem 7.7. Suppose a is the sesquilinear form defining an H^∞-Helmholtz problem approximated using a C^∞ PML with $C^{1,1}$ truncation boundary (in the sense of Definitions 4.6 and 4.7) and, in addition, suppose that $\Gamma_{\mathrm{p}} = \emptyset$, $\Gamma_- \in C^\omega$, a Dirichlet condition is imposed on Γ_-, and there is a neighbourhood, U of Γ_- such that $A, n \in C^\omega(U)$. Then, for all $W \Subset U$, $\psi \in C_c^\infty(\mathbb{R})$ and $k_0 > 0$, there exists $C > 0$ such that

$$\|\psi(\mathcal{P})\|_{L^2\to H^{p+1}(W)} \le \left(C\max\left(\frac{k}{p+1},1\right)\right)^{p+1}. \tag{7.29}$$

Proof. Let $\chi \in C_c^\infty(U)$ with $\operatorname{supp}(1-\chi)\cap W = \emptyset$. Since ψ is compactly supported,

$$\|\chi\partial^\alpha\psi(\mathcal{P})\|_{L^2\to L^2} = \|\chi\partial^\alpha e^{-t\mathcal{P}}e^{t\mathcal{P}}\psi(\mathcal{P})\|_{L^2\to L^2} \le \|\chi\partial^\alpha e^{-t\mathcal{P}}\|_{L^2\to L^2} C e^{Ct}. \tag{7.30}$$

By the heat-equation bounds of Escauriaza, Montaner and Zhang (2017) (as summarized in Galkowski *et al.* 2023*b*, Theorem 4.3), for $0 \le t \le k^2$ and $\tau \in [0,1]$,

$$\|\chi\partial^\alpha e^{-t\mathcal{P}}\|_{L^2\to L^2} \le \exp((tk^{-2})^{-\tau})|\alpha|!C^{|\alpha|}(tk^{-2})^{(\tau-1)|\alpha|/2}. \tag{7.31}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

We now prove that

$$\inf_{t\in[0,k^2]} \|\chi\partial^\alpha\, e^{-t\mathcal{P}}\|_{L^2\to L^2}\, e^{Ct} \le C^{|\alpha|}\max(|\alpha|^{|\alpha|}, k^{|\alpha|}). \tag{7.32}$$

When $|\alpha| \ge k$, let $\tau = 1$ and $t = k$. Then, by (7.31),

$$\|\chi\partial^\alpha\, e^{-t\mathcal{P}}\|_{L^2\to L^2}\, e^{Ct} \le C\, e^{Ck}|\alpha|!C^{|\alpha|}\, e^{Ck},$$

which implies (7.32) when $|\alpha| \ge k$. When $|\alpha| \le k$, we seek t and τ such that

$$(k^{-2}t)^{(\tau-1)|\alpha|/2} = k^{|\alpha|}|\alpha|^{-|\alpha|} \quad \text{and} \quad t = (k^{-2}t)^{-\tau} \tag{7.33}$$

(with the second equality making the arguments of the two exponentials equal). The second equality in (7.33) is that

$$\tau = \frac{\log|\alpha|}{\log k^2|\alpha|^{-1}},$$

and inputting this into the first equality in (7.33) implies that $t = |\alpha|$. With these choices, (7.32) and Stirling's approximation imply that

$$\|\chi\partial^\alpha\, e^{-t\mathcal{P}}\|_{L^2\to L^2}\, e^{Ct} \le C\, e^{C|\alpha|}C^{|\alpha|}k^{|\alpha|},$$

which implies (7.32) when $|\alpha| \le k$. With

$$t = \min(k, |\alpha|) \quad \text{and} \quad \tau = \min\left(\frac{\log|\alpha|}{\log k^2|\alpha|^{-1}}, 1\right),$$

this bound implies that

$$C\|\chi\partial^\alpha\, e^{-t\mathcal{P}}\|_{L^2\to L^2}\, e^{Ct} \le C^{|\alpha|}\max(|\alpha|^{|\alpha|}, k^{|\alpha|}).$$

Combining (7.30), (7.32), the definition of $\|\cdot\|_{H^{p+1}(W)}$ (i.e. (7.1) with $k = 1$), the equality

$$\sum_{|\alpha|=\ell} \frac{1}{\alpha!} = \frac{d^\ell}{\ell!} \tag{7.34}$$

(proved using the multinomial formula) and Stirling's approximation, we obtain

$$\begin{aligned}
\|\psi(\mathcal{P})\|^2_{L^2\to H^{p+1}(W)} &\le \sum_{\ell=0}^{p+1}\sum_{|\alpha|=\ell}\frac{1}{\ell!\alpha!}\|\chi\partial^\alpha\psi(\mathcal{P})\|^2_{L^2\to L^2} \\
&\le \sum_{\ell=0}^{p+1}\sum_{|\alpha|=\ell}\frac{1}{\ell!\alpha!}C^{2\ell}\max(\ell^{2\ell}, k^{2\ell}) \\
&\le \sum_{\ell=0}^{p+1} C^{2\ell}\max\left(1, \frac{k^{2\ell}}{\ell^{2\ell}}\right).
\end{aligned} \tag{7.35}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Now, by arguing as in (7.23), if $k \geq p+1$ then

$$\sum_{\ell=0}^{p+1} C^{2\ell} \max\left(1, \frac{k^{2\ell}}{\ell^{2\ell}}\right) \leq \sum_{\ell=0}^{p+1} C^{2\ell} \frac{k^{2\ell}}{\ell^{2\ell}} \leq C^{2(p+1)} \left(\frac{k}{p+1}\right)^{2(p+1)}. \tag{7.36}$$

Furthermore, if $k \leq p+1$, then

$$\sum_{\ell=0}^{p+1} C^{2\ell} \max\left(1, \frac{k^{2\ell}}{\ell^{2\ell}}\right) = \sum_{\ell=0}^{k} C^{2\ell} \left(\frac{k}{\ell}\right)^{2\ell} + \sum_{\ell=k+1}^{p+1} C^{2\ell} \leq C^{2k} + C^{2(p+1)} \leq C^{2(p+1)}. \tag{7.37}$$

The result (7.29) then follows by combining (7.35), (7.36) and (7.37). □

8. Pollution and preasymptotic error estimates for the h-BEM (R3)

8.1. Second-kind BIEs for Helmholtz scattering problems

In this section we consider the h-BEM for the standard second-kind boundary integral equation formulations for the Helmholtz scattering problems: given $f \in L^2(\Gamma)$, find $v \in L^2(\Gamma)$ such that

$$\mathcal{A}v = f, \quad \mathcal{A} := c_0(I + \mathcal{L}), \quad c_0 \in \mathbb{C} \setminus \{0\}, \tag{8.1}$$

where the operator $\mathcal{L}\colon L^2(\Gamma_-) \to L^2(\Gamma_-)$ is compact. Theorem B.2 below recaps the standard result that the Helmholtz exterior Dirichlet and Neumann problems can be reformulated as integral equations of the form (8.1), where f is given in terms of the known Dirichlet/Neumann boundary data and $\mathcal{A}$ is one of the boundary integral operators A_k', A_k, $B_{\rm reg}$ or $B_{\rm reg}'$ defined in (3.8) and (3.9). Standard mapping properties of these operators (see e.g. Chandler-Wilde *et al.* 2012, Theorems 2.17 and 2.18) imply that A_k', A_k, $B_{\rm reg}$, $B_{\rm reg}'$ are bounded on $L^2(\Gamma_-) \to L^2(\Gamma_-)$. Furthermore, each of A_k', A_k, $B_{\rm reg}$, $B_{\rm reg}'$ is invertible on $L^2(\Gamma_-) \to L^2(\Gamma_-)$. Moreover, they are each equal to a multiple of the identity plus a compact operator on $L^2(\Gamma_-) \to L^2(\Gamma_-)$; see e.g. Chandler-Wilde *et al.* (2012, §2.6) for A_k', A_k and Galkowski *et al.* (2022, Theorem 2.2) for $B_{\rm reg}$, $B_{\rm reg}'$.[2] As before, we use the notation

$$\rho_{\mathcal{A}}(k) := \|\mathcal{A}^{-1}\|_{L^2(\Gamma_-)\to L^2(\Gamma_-)}, \tag{8.2}$$

and observe that since $\mathcal{L}$ is compact, $\rho_{\mathcal{A}}(k) \geq c_0^{-1}$; see Lemma B.3 below.

We study the Galerkin approximation to u with $a\colon L^2(\Gamma_-)\times L^2(\Gamma_-) \to \mathbb{C}$ defined by

$$a(v, w) := \langle \mathcal{A}v, w\rangle_{L^2(\Gamma_-)}$$

under assumptions on $\mathcal{A}$ that allow us to include any of A_k, A_k', $B_{\rm reg}$ or $B_{\rm reg}'$.

[2] In the real-valued $L^2(\Gamma_-)$ inner product, A_k' and A_k are each other's adjoints (hence the $'$ notation); similarly for $B_{\rm reg}$, $B_{\rm reg}'$.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Since $\mathcal{A}\colon L^2(\Gamma_-) \to L^2(\Gamma_-)$, for $\mathcal{V} \subset L^2(\Gamma_-)$, and $u \in L^2(\Gamma_-)$, the Galerkin approximation v_h to v in $\mathcal{V}$ satisfies

$$\Pi_{\mathcal{V}}\mathcal{A}v_h = (I + \Pi_{\mathcal{V}}\mathcal{L})v_h = \Pi_{\mathcal{V}}\mathcal{A}v, \tag{8.3}$$

where $\Pi_{\mathcal{V}}$ is the L^2-orthogonal projector onto $\mathcal{V}$. In fact, we prove error bounds for the solution, $v_h \in \mathcal{V}_h$, to

$$(I + P_{\mathcal{V}}\mathcal{L})v_h = P_{\mathcal{V}}\mathcal{A}v, \tag{8.4}$$

where $P_{\mathcal{V}}\colon L^2(\Gamma) \to \mathcal{V}$ is any projector onto $\mathcal{V}$; that is, in principle, we consider any projection method for approximating v.

Remark 8.1 (The history of the operators A'_k, A_k, $B_{\rm reg}$, $B'_{\rm reg}$). The BIEs involving the operators A'_k and A_k (3.8) were introduced in Brakhage and Werner (1965), Leis (1965) and Panich (1965). The subscript 'reg' on the Neumann boundary integral operators (3.9) indicates that these are not the 'combined-field' (or 'combined-potential') Neumann BIEs introduced by Burton and Miller (1971) (denoted by B'_k and B_k in e.g. Chandler-Wilde *et al.* 2012, §2.6); the operators introduced in Burton and Miller (1971) are given by (3.9) with S_{ik} removed, and thus are not bounded on $L^2(\Gamma_-)$ because $H_k\colon L^2(\Gamma_-) \to H^{-1}(\Gamma)$. The idea of preconditioning H_k with an order -1 operator goes back to Burton (1976) (see e.g. the discussion in Amini and Harris 1990), with the use of S_{ik} proposed in Bruno, Elling and Turc (2012), and then advocated for in Boubendir and Turc (2013) and Vico, Greengard and Gimbutas (2014) (for more details, see e.g. the discussion in Galkowski *et al.* 2022, §2.1.1). The results in this section hold for a wider class of regularizing operators, of which S_{ik} is the prototypical example; see Galkowski *et al.* (2022, Assumption 1.1).

Remark 8.2 (Implementing the operator product for the Neumann BIEs (3.9)). Because of the operator product $S_{ik}H_k$ in $B_{\rm reg}$, in practice, approximations to the solution of $B_{\rm reg}v = f$ are computed by applying the projection method, not to $B_{\rm reg}v = f$, but to the system

$$\begin{pmatrix} i\left(\frac{1}{2}I - K\right) & S_{ik} \\ H_k & -1 \end{pmatrix} \begin{pmatrix} v \\ w \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix},$$

and similarly for $B'_{\rm reg}$. For simplicity, when studying $B_{\rm reg}$ and $B'_{\rm reg}$ we consider the idealized situation of (8.4), that is, we ignore the issue of discretizing the operator product in Galerkin or collocation methods. We emphasize, however, that Galkowski *et al.* (2025) *do* analyse discretizing this operator product in the Nyström method.

8.2. Assumptions on $\mathcal{A}$ and the functional calculus of $-\Delta_{\Gamma_-}$

Recall that since Γ_- is a smooth, compact hypersurface in $\mathbb{R}^d$, the tangential Laplacian, $-\Delta_{\Gamma_-}$ (i.e. the Laplacian induced on Γ_- from the flat Laplacian on $\mathbb{R}^d$) is a self-adjoint, non-negative, elliptic, second-order differential operator on L^2

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

with domain $H^2(\Gamma_-)$. In particular, there are $0 = \lambda_1 < \lambda_2 \leq \cdots$ with $\lambda_j \to \infty$ and an orthonormal basis of $L^2(\Gamma_-)$, $\{\phi_{\lambda_j}\}_{j=1}^{\infty}$ satisfying

$$(-\Delta_{\Gamma_-} - \lambda_j^2)\phi_{\lambda_j} = 0.$$

For $f \in L^\infty(\mathbb{R})$ we define

$$f(-k^{-2}\Delta_{\Gamma_-})u := \sum_{j=1}^{\infty} f(k^{-2}\lambda_j^2)\langle u, \phi_{\lambda_j}\rangle_{L^2(\Gamma_-)}\phi_{\lambda_j}.$$

We recall that elliptic regularity in H_k^s implies that for any basis of vector fields $\{X_i\}_{i=1}^N$ on Γ_- and $j \geq 0$, there exists $C > 0$ such that

$$\begin{aligned}\tfrac{1}{C}\|(-k^{-2}\Delta_{\Gamma_-} + 1)^{j/2}u\|_{L^2(\Gamma_-)} &\leq \sum_{|\alpha|\leq j} \|k^{-|\alpha|}X_1^{\alpha_1}\cdots X_N^{\alpha_N}u\|_{L^2(\Gamma_-)} \\ &\leq C\|(-k^{-2}\Delta_{\Gamma_-} + 1)^{j/2}u\|_{L^2(\Gamma_-)}.\end{aligned}$$

Therefore we define

$$\|u\|_{H_k^s(\Gamma_-)} := \|(-k^{-2}\Delta_{\Gamma_-} + 1)^{s/2}u\|_{L^2(\Gamma_-)}. \tag{8.5}$$

We then have the following estimate for functions of the Laplacian.

Lemma 8.3. If $f \in L^\infty_{\text{comp}}(\mathbb{R})$, then for all $N > 0$ there exists $C > 0$ such that

$$\|f(-k^{-2}\Delta_{\Gamma_-})\|_{H_k^{-N}(\Gamma_-)\to H_k^N(\Gamma_-)} \leq C.$$

Moreover, for all $f \in L^\infty(\mathbb{R})$ and $s \in \mathbb{R}$,

$$\|f(-k^{-2}\Delta_{\Gamma_-})\|_{H_k^s(\Gamma_-)\to H_k^s(\Gamma_-)} \leq \|f\|_{L^\infty}.$$

We make the following assumption on $\mathcal{A}$.

Assumption 8.4 (Assumptions on $\mathcal{A}$ and its high-frequency components).

(i) For all $k > 0$, the map $k \mapsto \mathcal{L}$ is continuous, $\mathcal{A} := c_0(I + \mathcal{L})$ is bounded and invertible on $H^s(\Gamma_-)$ for all $s \in \mathbb{R}$, and $\mathcal{L}\colon H^s(\Gamma_-) \to H^{s+1}(\Gamma_-)$.

(ii) For any $\chi \in C_c^\infty(\mathbb{R}; [0,1])$ with $\operatorname{supp}(1-\chi)\cap[-1,1] = \emptyset$, $(I - \chi(-k^{-2}\Delta_\Gamma))\mathcal{L}$ and $\mathcal{L}(I - \chi(-k^{-2}\Delta_\Gamma))$ are both in $\Psi_\hbar^{-1}(\Gamma)$.

(iii) There exists $L_{\max} > 0$ such that for any $\chi \in C_c^\infty(\mathbb{R}; [0,1])$ with $\operatorname{supp}(1-\chi)\cap[-1,1] = \emptyset$,

$$|\sigma_\hbar(I + (I - \chi(-k^{-2}\Delta_{\Gamma_-}))\mathcal{L})| \geq \tfrac{1}{L_{\max}}.$$

Remark 8.5. Since multiplication by c_0 maps $\mathcal{V}$ to $\mathcal{V}$ for any space of functions, $\mathcal{V}$, we assume without loss of generality that $c_0 = 1$ from now on.

Under Assumption 8.4, the calculus of semiclassical pseudodifferential operators has the following three consequences. Roughly speaking, the first (Lemma 8.6) states that, uniformly in k, the operator $\mathcal{L}$ increases Sobolev regularity by one on

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

high frequencies, the third (Lemma 8.8) states that $(I+(1-\psi(-k^{-2}\Delta_{\Gamma_-}))\mathcal{L})^{-1}$ does not move frequencies, and the second (Lemma 8.7) states that the symbol of this operator controls its norm.

Lemma 8.6. Suppose Assumption 8.4 holds. Then, for all $\psi \in C_c^\infty(\mathbb{R};[0,1])$, with $\operatorname{supp}(1-\psi)\cap[-1,1]=\emptyset$ and all $k_0>0$, $s\in\mathbb{R}$, there exists $C>0$ such that, for $k>k_0$,

$$\begin{aligned}\|\mathcal{L}(I-\psi(-k^{-2}\Delta_{\Gamma_-}))\|_{H_k^s(\Gamma_-)\to H_k^{s+1}(\Gamma_-)}&\le C,\\ \|(I-\psi(-k^{-2}\Delta_{\Gamma_-}))\mathcal{L}\|_{H_k^s(\Gamma_-)\to H_k^{s+1}(\Gamma_-)}&\le C.\end{aligned}\tag{8.6}$$

Proof. This follows from Assumption 8.4(ii) and the boundedness property (iii) of Informal Theorem 5.5 (or, more precisely, its analogue for semiclassical pseudo-differential operators on a smooth manifold; see Remark 5.6). □

Lemma 8.7. Suppose Assumption 8.4 holds. Then, for all $\psi \in C_c^\infty(\mathbb{R};[0,1])$, with $\operatorname{supp}(1-\psi)\cap[-1,1]=\emptyset$, there exists $k_0>0$ such that for all $s\in\mathbb{R}$, there exists $C>0$ such that, for $k>k_0$,

$$(I+(I-\psi(-k^{-2}\Delta_{\Gamma_-}))\mathcal{L})^{-1}\in\Psi_\hbar^0(\Gamma_-)$$

with

$$\|(I+(I-\psi(-k^{-2}\Delta_{\Gamma_-}))\mathcal{L})^{-1}\|_{H_k^s(\Gamma_-)\to H_k^s(\Gamma_-)}\le L_{\max}+Ck^{-1}.\tag{8.7}$$

Proof. By Theorem 5.7 (or rather its analogue on a manifold) together with Assumption 8.4(ii, iii) and the composition property (iv) of (the manifold analogue of) Informal Theorem 5.5, there exists $E\in\Psi_\hbar^0(\Gamma_-)$ with $|\sigma_\hbar(E)|\le L_{\max}$ such that

$$E(I+(I-\chi(-k^{-2}\Delta_{\Gamma_-}))\mathcal{L})=I+R,\quad \text{with } R=O(k^{-\infty})_{\Psi_\hbar^{-\infty}}.$$

Therefore, for $\hbar$ small enough (and hence $k_0>0$ large enough),

$$(I+(1-\psi(-k^{-2}\Delta_{\Gamma_-}))\mathcal{L})^{-1}=(I+R)^{-1}E=E-R(I+R)^{-1}E=E+O(k^{-\infty})_{\Psi_\hbar^{-\infty}}.$$

Hence $(I+(1-\psi(-k^{-2}\Delta_{\Gamma_-}))\mathcal{L})^{-1}\in\Psi_\hbar^0(\Gamma_-)$ and $|\sigma_\hbar((I+(1-\psi(-k^{-2}\Delta_{\Gamma_-}))\mathcal{L})^{-1})|\le L_{\max}$. The bound (8.7) then follows from the boundedness property (iii) of Informal Theorem 5.5, using that the $L^2\to L^2$ bound in terms of the symbol recorded there can be converted into an analogous $H_k^s\to H_k^s$ bound; see Galkowski *et al.* (2025, Lemma 3.2). □

Lemma 8.8. Suppose Assumption 8.4 holds and let $\psi\in C_c^\infty(\mathbb{R};[0,1])$ with $\operatorname{supp}(1-\psi)\cap[-1,1]=\emptyset$. Then there exists $k_0>0$ such that for all $N>0$, $\psi_1,\psi_2\in C_c^\infty(\mathbb{R})$, with $\operatorname{supp}\psi_1\cap\operatorname{supp}(1-\psi_2)=\emptyset$ and $\operatorname{supp}(1-\psi_1)\cap[-1,1]=\emptyset$, there exists $C>0$ such that for $k>k_0$ we have

$$\begin{aligned}&\|\psi_1(-k^{-2}\Delta_\Gamma)\mathcal{L}(1-\psi_2(-k^{-2}\Delta_{\Gamma_-}))\|_{H_k^{-N}(\Gamma_-)\to H_k^N(\Gamma_-)}\le Ck^{-N},\\ &\|\psi_1(-k^{-2}\Delta_\Gamma)(I+(1-\psi(-k^{-2}\Delta_{\Gamma_-}))\mathcal{L})^{-1}(1-\psi_2(-k^{-2}\Delta_{\Gamma_-})\|_{H_k^{-N}(\Gamma_-))\to H_k^N(\Gamma_-)}\\ &\qquad\qquad\le Ck^{-N},\end{aligned}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

$$\|(1-\psi_2(-k^{-2}\Delta_\Gamma))(I+(1-\psi(-k^{-2}\Delta_{\Gamma_-}))\mathcal{L})^{-1}\psi_1(-k^{-2}\Delta_{\Gamma_-})\|_{H_k^{-N}(\Gamma_-))\to H_k^N(\Gamma_-)} \leq Ck^{-N}. \tag{8.8}$$

Sketch proof of Lemma 8.8. To prove the result, we use a notion of frequency pseudolocality similar to that in Lemma 7.5. In particular, one can check using *either* Averseng *et al.* (2025, Theorem 5.37) *or* the fact that by e.g. Galkowski *et al.* (2025, Lemma 3.9), for $\varphi \in C_c^\infty(\mathbb{R})$,

$$\varphi(-k^{-2}\Delta_{\Gamma_-}) = \mathrm{Op}_\hbar(a) + O(k^{-\infty})_{\Psi_\hbar^{-\infty}}$$

with

$$\operatorname{supp} a \subset \{(x',\xi') \in T^*\Gamma_- : |\xi'|^2_{g_{\Gamma_-}} \in \operatorname{supp}\varphi\},$$

then the composition property (iv) of Informal Theorem 5.5 implies that for any $m \in \mathbb{R}$, any ψ_1, ψ_2 as in the statement, and any $A \in \Psi_\hbar^m$, we have

$$\psi_1(-k^{-2}\Delta_{\Gamma_-})A(1-\psi_2(-k^{-2}\Delta_{\Gamma_-}) = O(k^{-\infty})_{\Psi_\hbar^{-\infty}}$$

and

$$(1-\psi_2(-k^{-2}\Delta_{\Gamma_-})A\psi_1(-k^{-2}\Delta_{\Gamma_-}) = O(k^{-\infty})_{\Psi_\hbar^{-\infty}}.$$

The last two bounds in (8.8) immediately follow since $(I+(1-\psi(-k^{-2}\Delta_{\Gamma_-}))\mathcal{L})^{-1} \in \Psi_\hbar^0$ by Lemma 8.7. For the first bound in (8.8), let $\psi_3 \in C_c^\infty(\mathbb{R})$ with $\operatorname{supp}(1-\psi_3) \cap [-1,1] = \emptyset$, and $\operatorname{supp}\psi_3 \cap \operatorname{supp}(1-\psi_2) = \emptyset$. These support properties imply that

$$\begin{aligned}&\psi_1(-k^{-2}\Delta_{\Gamma_-})\mathcal{L}(1-\psi_2(-k^{-2}\Delta_{\Gamma_-})\\ &\quad= \psi_1(-k^{-2}\Delta_{\Gamma_-})\mathcal{L}(1-\psi_3(-k^{-2}\Delta_{\Gamma_-})(1-\psi_2(-k^{-2}\Delta_{\Gamma_-}),\end{aligned}$$

and then the first bound in (8.8) follows since $\mathcal{L}(1-\psi_3(-k^{-2}\Delta_{\Gamma_-})) \in \Psi_\hbar^{-1}$ by Assumption 8.4(ii). □

We now show that the high-frequency components of $(I+\mathcal{L})^{-1}$ are well-behaved; this result can be viewed as a boundary integral analogue of the fact that the Helmholtz operator is semiclassically elliptic on high frequencies (discussed in Section 2.2).

Lemma 8.9. Suppose Assumption 8.4 holds and let $\chi \in C_c^\infty(\mathbb{R};[0,1])$ with $\operatorname{supp}(1-\chi) \cap [-1,1] = \emptyset$. Then, for all $k_0 > 0$, $s \in \mathbb{R}$, and $N > 0$, there exists $C > 0$ such that for all $k > k_0$,

$$\|(I+\mathcal{L})^{-1}\|_{H_k^s \to H_k^s} \leq C\rho_{\mathcal{A}}(k), \tag{8.9}$$

$$\|(I+\mathcal{L})^{-1}(I-\chi(-k^{-2}\Delta_{\Gamma_-}))\|_{H_k^s(\Gamma_-)\to H_k^s(\Gamma_-)} \leq L_{\max} + Ck^{-1} + Ck^{-N}\rho_{\mathcal{A}}(k), \tag{8.10}$$

$$\|(I-\chi(-k^{-2}\Delta_{\Gamma_-}))(I+\mathcal{L})^{-1}\|_{H_k^s(\Gamma_-)\to H_k^s(\Gamma_-)} \leq L_{\max} + Ck^{-1} + Ck^{-N}\rho_{\mathcal{A}}(k). \tag{8.11}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Before proving Lemma 8.9, we remark that the rest of this section only uses Lemmas 8.6 and 8.9 and the following corollary of these two results, that is, Lemmas 8.7 and 8.8 are only used to prove Lemma 8.9.

Corollary 8.10. Suppose Assumption 8.4 holds and let $\chi \in C_c^\infty(\mathbb{R}; [0,1])$ with $\operatorname{supp}(1-\chi) \cap [-1,1] = \emptyset$. Then, for all $k_0 > 0$ and $s, N > 0$, there exists $C > 0$ such that for all $k > k_0$,

$$\|\mathcal{L}(I+\mathcal{L})^{-1}(I-\chi(-k^{-2}\Delta_{\Gamma_-}))\|_{H_k^s(\Gamma_-)\to H_k^{s+1}(\Gamma_-)} \le C + Ck^{-N}\rho_{\mathcal{A}}(k).$$

Proof. Since $I = (I-\chi(-k^{-2}\Delta_{\Gamma_-})) + \chi(-k^{-2}\Delta_{\Gamma_-})$,

$$\begin{aligned}
&\mathcal{L}(I+\mathcal{L})^{-1}(I-\chi(-k^{-2}\Delta_{\Gamma_-})) \\
&\quad = (I-\chi(-k^{-2}\Delta_{\Gamma_-}))\mathcal{L}(I+\mathcal{L})^{-1}(I-\chi(-k^{-2}\Delta_{\Gamma_-})) \\
&\qquad + \chi(-k^{-2}\Delta_{\Gamma_-})\mathcal{L}(I+\mathcal{L})^{-1}(I-\chi(-k^{-2}\Delta_{\Gamma_-})) \\
&\quad = (I-\chi(-k^{-2}\Delta_{\Gamma_-}))\mathcal{L}(I+\mathcal{L})^{-1}(I-\chi(-k^{-2}\Delta_{\Gamma_-})) \\
&\qquad + \chi(-k^{-2}\Delta_{\Gamma_-})(I-(I+\mathcal{L})^{-1})(I-\chi(-k^{-2}\Delta_{\Gamma_-})),
\end{aligned}$$

and then the result follows from Lemma 8.6, (8.10) and Lemma 8.3. □

Proof of Lemma 8.9. We first prove (8.9). First observe that by Assumption 8.4(i) it is sufficient to prove the result for k_0 sufficiently large. Next observe that once (8.9) is proved for $s \ge 0$, since

$$\|(I+\mathcal{L})^{-1}\|_{H_\hbar^s(\Gamma)\to H_\hbar^s(\Gamma)} = \|((I+\mathcal{L})^*)^{-1}\|_{H_\hbar^{-s}(\Gamma)\to H_\hbar^{-s}(\Gamma)},$$

the result for $s < 0$ follows by applying the result for $s \ge 0$ to $(I+\mathcal{L})^*$.

Given $g \in H_k^s$, let $\phi \in H_k^s$ satisfy $(I+\mathcal{L})\phi = g$. We need to show that

$$\|\phi\|_{H_k^s} \le C\rho_{\mathcal{A}}(k)\|g\|_{H_k^s}. \tag{8.12}$$

Let $\widetilde{\chi} \in C_c^\infty(\mathbb{R}; [0,1])$ with $\operatorname{supp}(1-\widetilde{\chi}) \cap [-1,1] = \emptyset$ and $\operatorname{supp}(1-\widetilde{\chi}) \cap \operatorname{supp}\chi = \emptyset$. By Lemma 8.3 and the definition of $\rho_{\mathcal{A}}(k)$ (8.2),

$$\|\widetilde{\chi}(-k^{-2}\Delta_{\Gamma_-})\phi\|_{H_k^s} \le C\|\phi\|_{L^2} \le C\rho_{\mathcal{A}}(k)\|g\|_{L^2} \le C\rho_{\mathcal{A}}(k)\|g\|_{H_k^s}. \tag{8.13}$$

We now claim that

$$\|(I-\widetilde{\chi}(-k^{-2}\Delta_{\Gamma_-})\phi\|_{H_k^s} \le C\|(I-\chi(-k^{-2}\Delta_{\Gamma_-}))(I+\mathcal{L})\phi\|_{H_k^s} + Ck^{-M}\|\phi\|_{H_k^s}. \tag{8.14}$$

Once this is established, the bound (8.12) (and hence (8.9)) for k_0 sufficiently large follows by combining (8.13) with (8.14).

To establish (8.14), let $\psi \in C_c^\infty(\mathbb{R}; [0,1])$ with $\operatorname{supp}(1-\psi) \cap [-1,1] = \emptyset$ and $\operatorname{supp}(1-\chi) \cap \operatorname{supp}\psi = \emptyset$. By Assumption 8.4(iii), $I + (I-\psi(-k^{-2}\Delta_{\Gamma_-})\mathcal{L}$ is elliptic, and thus

$$(I-\chi(-k^{-2}\Delta_{\Gamma_-}))(I+(I-\psi(-k^{-2}\Delta_{\Gamma_-})\mathcal{L}) = (I-\chi(-k^{-2}\Delta_{\Gamma_-}))(I+\mathcal{L}) \tag{8.15}$$

is elliptic on the support of $I-\widetilde{\chi}(-k^{-2}\Delta_{\Gamma_-})$. More precisely, by Galkowski *et al.*

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

(2025, Lemma 3.9),

$$I - \tilde{\chi}(-k^{-2}\Delta_{\Gamma_-}) = \mathrm{Op}_\hbar(\tilde{a}) + O(k^{-\infty})_{\Psi_\hbar^{-\infty}},$$
$$I - \chi(-k^{-2}\Delta_{\Gamma_-}) = \mathrm{Op}_\hbar(a) + O(k^{-\infty})_{\Psi_\hbar^{-\infty}}$$

for some $a, \tilde{a} \in S^0(T^*\Gamma)$ with $\operatorname{supp}\tilde{a} \subset \{|a| > \frac{1}{2}\}$. Therefore, by the composition property (iv) of Informal Theorem 5.5,

$$(I - \chi(-k^{-2}\Delta_{\Gamma_-}))(I + \mathcal{L}) = \mathrm{Op}_\hbar(b) + O(k^{-\infty})_{\Psi_\hbar^{-\infty}}$$

with $\{|a| > \frac{1}{2}\} \subset \{|b| > c > 0\}$ and hence $\operatorname{supp}\tilde{a} \subset \{|b| > c > 0\}$. The bound (8.14) then follows from the manifold analogue of the elliptic estimate of Theorem 5.7.

Next observe that the bound (8.9) implies both (8.10) and (8.11) for all $k_0 < k \leq k_1$ for any $k_1 > 0$ fixed. Thus it remains to check (8.10) and (8.11) for $k > k_1$.

To do this, let $\psi \in C_c^\infty(\mathbb{R})$ with $\operatorname{supp}(1-\psi) \cap [-1,1] = \emptyset$, and $\operatorname{supp}\psi \cap \operatorname{supp}(1-\chi) = \emptyset$. Then, by Lemmas 8.7 and 8.8, there are $k_1 > 0$ and $C > 0$ such that (8.7) and (8.8) hold for $k > k_1$.

Now, since $I + \mathcal{L} = I + (1 - \psi(-k^{-2}\Delta_{\Gamma_-}))\mathcal{L} + \psi(-k^{-2}\Delta_{\Gamma_-})\mathcal{L}$,

$$\begin{aligned}
&(I+\mathcal{L})(I + (1-\psi(-k^{-2}\Delta_{\Gamma_-}))\mathcal{L})^{-1}(I - \chi(-k^{-2}\Delta_{\Gamma_-})) \\
&\quad = I - \chi(-k^{-2}\Delta_{\Gamma_-}) \\
&\qquad + \psi(-k^{-2}\Delta_{\Gamma_-})\mathcal{L}(I + (1-\psi(-k^{-2}\Delta_{\Gamma_-}))\mathcal{L})^{-1}(I - \chi(-k^{-2}\Delta_{\Gamma_-})) \\
&\quad = I - \chi(-k^{-2}\Delta_{\Gamma_-}) + O(k^{-\infty})_{H_k^{-N}\to H_k^N},
\end{aligned}$$

where we have used (8.8) in the last equality. Hence

$$\begin{aligned}
&\|(I + (1-\psi(k^{-2}\Delta_{\Gamma_-}))\mathcal{L})^{-1}(I - \chi(-k^{-2}\Delta_{\Gamma_-})) \\
&\qquad - (I+\mathcal{L})^{-1}(I - \chi(-k^{-2}\Delta_{\Gamma_-}))\|_{H_k^{-N}\to H_k^N} \\
&\qquad\quad \leq Ck^{-N}\rho_{\mathcal{A}}(k).
\end{aligned}$$

Combining this with (8.7) implies (8.10) for $k > k_1$ Similarly,

$$\begin{aligned}
&(I - \chi(-k^{-2}\Delta_{\Gamma_-}))(I + (1-\psi(-k^{-2}\Delta_{\Gamma_-}))\mathcal{L})^{-1}(I+\mathcal{L}) \\
&\quad = (I - \chi(-k^{-2}\Delta_{\Gamma_-}))(I + (1-\psi(-k^{-2}\Delta_{\Gamma_-}))\mathcal{L})^{-1}\psi(-k^{-2}\Delta_{\Gamma_-})\mathcal{L} \\
&\qquad + (I - \chi(-k^{-2}\Delta_{\Gamma_-})) \\
&\quad = (I - \chi(-k^{-2}\Delta_{\Gamma_-})) + O(k^{-\infty})_{H_k^{-N}\to H_k^N},
\end{aligned}$$

which implies

$$\begin{aligned}
&\|(I - \chi(-k^{-2}\Delta_{\Gamma_-}))(I + (1-\psi(-k^{-2}\Delta_{\Gamma_-}))\mathcal{L})^{-1} \\
&\qquad - (I - \chi(-k^{-2}\Delta_{\Gamma_-}))(I+\mathcal{L})^{-1}\|_{H_k^{-N}\to H_k^N} \\
&\qquad\quad \leq C\rho_{\mathcal{A}}(k)k^{-N}.
\end{aligned}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Combining this with (8.7) implies (8.11) for $k > k_1$. □

8.3. Preasymptotic error estimates for the h-BEM

Let $s, t \in \mathbb{R}$, $\mathcal{V} \subset H^s_k(\Gamma_-)$, and $P_\mathcal{V} \colon H^t_k(\Gamma_-) \to \mathcal{V}$ be a projector onto $\mathcal{V}$. Let

$$\pi_{t\to s}(P_\mathcal{V}) := \|I - P_\mathcal{V}\|_{H^t_k(\Gamma_-)\to H^s_k(\Gamma_-)}. \tag{8.16}$$

Theorem 8.11 (Abstract result on convergence of the projection method). Let $s \geq 0$, $M > 0$ $k_0 > 0$, suppose that $\mathcal{A}$ satisfies Assumption 8.4, and $\mathcal{J} \subset (0,\infty)$ satisfies

$$\sup\{k^{-M}\rho_\mathcal{A}(k) \colon k \in (k_0,\infty) \setminus \mathcal{J}\} < \infty.$$

For all $N \geq s$, $\epsilon > 0$, there are $c, C > 0$ such that the following holds. If $k > k_0$, $k \notin \mathcal{J}$, $\mathcal{V} \subset H^s_k(\Gamma_-)$, and $P_\mathcal{V} \colon H^s_k(\Gamma_-) \to \mathcal{V}$ is a projector onto $\mathcal{V}$ satisfying

$$\big(\pi_{N\to -N}(P_\mathcal{V}) + \pi_{N\to s}(P_\mathcal{V})\pi_{s+1\to -N}(P_\mathcal{V})\big)\rho_\mathcal{A}(k) + \pi_{s+1\to s}(P_\mathcal{V}) < c, \tag{8.17}$$

then for all $v \in H^s_k(\Gamma_-)$, the solution $v_h \in \mathcal{V}$ to (8.4) exists, is unique, and satisfies

$$\|1_{[-1-\epsilon,1+\epsilon]}(k^{-2}\Delta_\Gamma)(v - v_h)\|_{H^N_k(\Gamma_-)} \leq C\pi_{s\to s}\pi_{s\to -N}\rho_\mathcal{A}(k)\|(I - P_\mathcal{V})v\|_{H^s_k(\Gamma_-)}, \tag{8.18}$$

$$\begin{aligned}\|(I - 1_{[-1-\epsilon,1+\epsilon]}(k^{-2}\Delta_\Gamma))(v - v_h)\|_{H^s_k(\Gamma_-)} \\ \leq (L_{\max} + Ck^{-1})\pi_{s\to s}\Big(\pi_{s\to s} + C\big[\pi_{N\to s}\pi_{s\to -N}\rho_\mathcal{A}(k) \\ + \pi_{s\to s}\big((\pi_{N\to -N} + \pi_{N\to s}\pi_{s+1\to -N})\rho_\mathcal{A}(k) + \pi_{s+1\to s}\big)\big]\Big)\|(I - P_\mathcal{V})v\|_{H^s_k(\Gamma_-)}.\end{aligned} \tag{8.19}$$

Observe that (8.18) is a bound on the low frequencies of the Galerkin error, and (8.19) is a bound on the high frequencies.

In the concrete setting of the h-BEM, with piecewise polynomials Theorem 8.11 has the following consequence.

Theorem 8.12 (Rigorous statement of Informal Theorem 3.3: k-explicit upper bounds on the h-BEM error). Let $\Omega_- \Subset \mathbb{R}^d$ with smooth boundary, $M > 0$, $\Upsilon > 0$ $p \geq 0$, $k_0 > 0$. Suppose that $\mathcal{A} \in \{A_k, A'_k, B_{\mathrm{reg}}, B'_{\mathrm{reg}}\}$ and $\mathcal{J} \subset (0,\infty)$ satisfies

$$\sup\{k^{-M}\rho_\mathcal{A}(k) \colon k \in (k_0,\infty) \setminus \mathcal{J}\} < \infty.$$

Then, for all $\epsilon > 0$, there are $c, C > 0$ such that for all $k > k_0$, $k \notin \mathcal{J}$, and all C^{p+1}-triangulations $\mathcal{T}$ with constant Υ (in the sense of Definition A.5), satisfying

$$(hk)^{2p+2}\rho_\mathcal{A}(k) + hk < c \tag{8.20}$$

(where $h := h(\mathcal{T})$), and for all $v \in L^2(\Gamma_-)$, the Galerkin approximation v_h to v in $\mathcal{P}^p_\mathcal{T}$ exists, is unique, and satisfies

$$\|1_{[-1-\epsilon,1+\epsilon]}(k^{-2}\Delta_\Gamma)(v - v_h)\|_{L^2(\Gamma_-)} \leq C(hk)^{p+1}\rho_\mathcal{A}(k) \min_{w_h\in\mathcal{P}^p_\mathcal{T}} \|v - w_h\|_{L^2(\Gamma_-)} \tag{8.21}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

and

$$\|(I - 1_{[-1-\epsilon,1+\epsilon]}(k^{-2}\Delta_\Gamma))(v - v_h)\|_{L^2(\Gamma_-)} \\ \leq (1 + Ck^{-1})(1 + C(hk)^{2p+2}\rho_{\mathcal{A}}(k) + Chk) \min_{w_h \in \mathcal{P}^p_{\mathcal{T}}} \|v - w_h\|_{L^2(\Gamma_-)}. \quad (8.22)$$

The bounds (8.21) and (8.22) together imply (3.13). The bound (3.14) on the relative error then follows from the combination of (3.13), the polynomial-approximation result of Theorem A.12 and the following result.

Lemma 8.13 (Oscillatory behaviour of BIE solution under plane-wave scattering). Suppose that $\mathcal{A}$ is one of A_k, A'_k, $B_{\rm reg}$ or $B'_{\rm reg}$, and the right-hand side f is as described in Theorem B.2. Then $v \in H^t(\Gamma)$ for all $t \geq 0$, and given $t \geq s \geq 0$ and $k_0 > 0$, there exists $C > 0$ such that

$$\|v\|_{H^t_k(\Gamma)} \leq C\|v\|_{H^s_k(\Gamma)} \quad \text{for all } k \geq k_0. \quad (8.23)$$

We do not prove Lemma 8.13 but instead give a sketch (with references).

Sketch of proof. The steps of the proof are the following.

(1) Show that the high-frequency components of f are $O(k^{-\infty})$, via an explicit calculation using integration by parts/stationary phase; see Galkowski *et al.* (2025, Lemma 7.3).

(2) Show that the high-frequency components of v are $O(k^{-\infty})$ using the result of step (1) and the fact that $(I - \chi(-k^{-2}\Delta_{\Gamma_-}))(I + \mathcal{L})$ is elliptic on high frequencies via Assumption 8.4(iii) and (8.15); see Galkowski *et al.* (2025, Lemma 7.4).

(3) Obtain the bound (8.23) from the result of step (2) and the smoothing property of low-frequency cut-offs (from Lemma 8.3, as in (8.38) below); see Galkowski *et al.* (2025, Theorem 7.2). □

8.4. *Proof of Theorem 8.11*

We start with the following lemma.

Lemma 8.14 (Quasioptimality in terms of the norm of the discrete inverse). Let $s \geq 0$, $\mathcal{V} \subset H^s_k(\Gamma_-)$ and suppose that $P_{\mathcal{V}} \colon H^s_k(\Gamma_-) \to \mathcal{V}$ is a projection onto $\mathcal{V}$ such that $I + P_{\mathcal{V}}\mathcal{L} \colon H^s_k(\Gamma_-) \to H^s_k(\Gamma_-)$ is invertible. Then, for all $v \in H^s_k(\Gamma_-)$, the Galerkin approximation v_h to v in $\mathcal{V}$ exists, is unique, and satisfies

$$v - v_h = (I + P_{\mathcal{V}}\mathcal{L})^{-1}(I - P_{\mathcal{V}})(I - P_{\mathcal{V}})v. \quad (8.24)$$

Proof. We first consider the equation (8.4) as an equation in $H^s_k(\Gamma_-)$. Since $I + P_{\mathcal{V}}\mathcal{L} \colon H^s_k(\Gamma) \to H^s_k(\Gamma)$ is invertible, the solution v_h to (8.4) in $H^s_k(\Gamma)$ exists and is unique as an element of $H^s_k(\Gamma)$. Applying $(I - P_{\mathcal{V}})$ to (8.4), we see that $(I - P_{\mathcal{V}})v_h = 0$ and thus $v_h \in \mathcal{V}$, that is, the equation (8.4) has a unique solution

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

in $\mathcal{V}$. Then (8.4) implies

$$\begin{aligned}(I + P_{\mathcal{V}}\mathcal{L})(v - v_h) &= (I + P_{\mathcal{V}}\mathcal{L})v - P_{\mathcal{V}}(I + \mathcal{L})v \\ &= v + P_{\mathcal{V}}\mathcal{L}v - P_{\mathcal{V}}(I + \mathcal{L})v \\ &= (I - P_{\mathcal{V}})v.\end{aligned}$$

Therefore, since $(I - P_{\mathcal{V}}) = (I - P_{\mathcal{V}})^2$,

$$(v - v_h) = (I + P_{\mathcal{V}}\mathcal{L})^{-1}(I - P_{\mathcal{V}})v = (I + P_{\mathcal{V}}\mathcal{L})^{-1}(I - P_{\mathcal{V}})(I - P_{\mathcal{V}})v. \qquad \square$$

To obtain conditions under which $I + P_{\mathcal{V}}\mathcal{L}\colon H^s_k(\Gamma) \to H^s_k(\Gamma)$ is invertible (and thus under which we can use Lemma 8.14), we write

$$(I + P_{\mathcal{V}}\mathcal{L}) = I + \mathcal{L} + (P_{\mathcal{V}} - I)\mathcal{L} = (I + (P_{\mathcal{V}} - I)\mathcal{L}(I + \mathcal{L})^{-1})(I + \mathcal{L}), \tag{8.25}$$

and study when $I + (P_{\mathcal{V}} - I)\mathcal{L}(I + \mathcal{L})^{-1}$ is invertible. We then use that

$$(I + (P_{\mathcal{V}} - I)\mathcal{L}(I + \mathcal{L})^{-1})^{-1}(I - P_{\mathcal{V}}) = (I - P_{\mathcal{V}})(I + (P_{\mathcal{V}} - I)\mathcal{L}(I + \mathcal{L})^{-1})^{-1}(I - P_{\mathcal{V}}),$$

that is, we can add an $(I - P_{\mathcal{V}})$ on the left. Indeed, if $(I + (P_{\mathcal{V}} - I)\mathcal{L}(I + \mathcal{L})^{-1})v = (I - P_{\mathcal{V}})f$, then $P_{\mathcal{V}}v = 0$ so that $v = (I - P_{\mathcal{V}})v$. Therefore the operator on the right-hand side of (8.24) can be written as

$$(I+P_{\mathcal{V}}\mathcal{L})^{-1}(I-P_{\mathcal{V}}) = (I+\mathcal{L})^{-1}(I-P_{\mathcal{V}})(I+(P_{\mathcal{V}}-I)\mathcal{L}(I+\mathcal{L})^{-1})^{-1}(I-P_{\mathcal{V}}). \tag{8.26}$$

The most naïve way to show that $I + (P_{\mathcal{V}} - I)\mathcal{L}(I + \mathcal{L})^{-1}$ is invertible is to show that

$$\|(P_{\mathcal{V}} - I)\mathcal{L}(I + \mathcal{L})^{-1}\|_{H^s_k(\Gamma_-)\to H^s_k(\Gamma_-)} < \frac{1}{2}, \tag{8.27}$$

in which case $\|(I + (P_N - I)\mathcal{L}(I + \mathcal{L})^{-1})^{-1}\|_{H^s_k(\Gamma_-)\to H^s_k(\Gamma_-)} \le 2$.

To understand when the condition (8.27) holds, given $\epsilon > 0$, we define low- and high-frequency projectors,

$$\Pi_{\mathscr{L}} := 1_{[-1-\epsilon,1+\epsilon]}(-k^{-2}\Delta_{\Gamma_-}) \quad \text{and} \quad \Pi_{\mathscr{H}} := (I - \Pi_{\mathscr{L}}), \tag{8.28}$$

and decompose H^s_k into an L^2-orthogonal sum $\Pi_{\mathscr{L}}H^s_k \oplus \Pi_{\mathscr{H}}H^s_k$. That is,

$$\begin{aligned}&(P_{\mathcal{V}} - I)\mathcal{L}(I + \mathcal{L})^{-1} \\ &\quad = \begin{pmatrix} \Pi_{\mathscr{L}}(P_{\mathcal{V}} - I)\mathcal{L}(I + \mathcal{L})^{-1}\Pi_{\mathscr{L}} & \Pi_{\mathscr{L}}(P_N - I)\mathcal{L}(I + \mathcal{L})^{-1}\Pi_{\mathscr{H}} \\ \Pi_{\mathscr{H}}(P_{\mathcal{V}} - I)\mathcal{L}(I + \mathcal{L})^{-1}\Pi_{\mathscr{L}} & \Pi_{\mathscr{H}}(P_{\mathcal{V}} - I)\mathcal{L}(I + \mathcal{L})^{-1}\Pi_{\mathscr{H}} \end{pmatrix}.\end{aligned} \tag{8.29}$$

Definition 8.15. We write

$$\|A\|_{H^s_k(\Gamma_-)\to H^s_k(\Gamma_-)} \le C_1 \begin{pmatrix} a_{\mathscr{L}\mathscr{L}} & a_{\mathscr{L}\mathscr{H}} \\ a_{\mathscr{H}\mathscr{L}} & a_{\mathscr{H}\mathscr{H}} \end{pmatrix}$$

if

$$A = \begin{pmatrix} \Pi_{\mathscr{L}}A\Pi_{\mathscr{L}} & \Pi_{\mathscr{L}}A\Pi_{\mathscr{H}} \\ \Pi_{\mathscr{H}}A\Pi_{\mathscr{L}} & \Pi_{\mathscr{H}}A\Pi_{\mathscr{H}} \end{pmatrix}, \tag{8.30}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

with

$$\|\Pi_{\mathscr{L}} A \Pi_{\mathscr{L}}\|_{H^s_k(\Gamma_-)\to H^s_k(\Gamma_-)} \le C_1 a_{\mathscr{L}\mathscr{L}}, \quad \|\Pi_{\mathscr{H}} A \Pi_{\mathscr{L}}\|_{H^s_k(\Gamma_-)\to H^s_k(\Gamma_-)} \le C_1 a_{\mathscr{H}\mathscr{L}},$$
$$\|\Pi_{\mathscr{L}} A \Pi_{\mathscr{H}}\|_{H^s_k(\Gamma_-)\to H^s_k(\Gamma_-)} \le C_1 a_{\mathscr{L}\mathscr{H}}, \quad \|\Pi_{\mathscr{H}} A \Pi_{\mathscr{H}}\|_{H^s_k(\Gamma_-)\to H^s_k(\Gamma_-)} \le C_1 a_{\mathscr{H}\mathscr{H}}.$$

One can then easily check that if $\|A\| \le C_1 \mathscr{A}$ and $\|B\| \le C_2 \mathscr{B}$, then $\|AB\| \le C_1 C_2 \mathscr{A}\mathscr{B}$.

Lemma 8.16. Let $s \ge 0$, $k_0 > 0$ and suppose $\mathcal{A}$ satisfies Assumption 8.4. Then, for any $N \in \mathbb{R}$, there exists $C > 0$ such that for all $k > k_0$, $\mathcal{V} \subset H^s_k(\Gamma_-)$ and $P_{\mathcal{V}} \colon H^s_k(\Gamma_-) \to \mathcal{V}$ projectors onto $\mathcal{V}$,

$$\|(P_{\mathcal{V}} - I)\mathcal{L}(I+\mathcal{L})^{-1}\|_{H^s_k(\Gamma_-)\to H^s_k(\Gamma_-)} \le C \begin{pmatrix} \pi_{N\to -N}\rho_{\mathcal{A}}(k) & \pi_{s+1\to -N} \\ \pi_{N\to s}\rho_{\mathcal{A}}(k) & \pi_{s+1\to s} \end{pmatrix}. \tag{8.31}$$

We prove Lemma 8.16 below, but first we use it to understand when (8.27) holds. Indeed, (8.27) holds if all entries on the right-hand side of (8.31) are $\ll 1$. However, this imposes too strong a restriction on the space $\mathcal{V}$, and we can instead change the norm on H^s_k using an invertible operator $\mathcal{C} \colon H^s_k(\Gamma_-) \to H^s_k(\Gamma_-)$ to improve the situation. More precisely, we observe that

$$I + (P_{\mathcal{V}} - I)\mathcal{L}(I+\mathcal{L})^{-1} = \mathcal{C}^{-1}(I + \mathcal{C}(P_{\mathcal{V}} - I)\mathcal{L}(I+\mathcal{L})^{-1}\mathcal{C}^{-1})\mathcal{C}.$$

We now choose $\mathcal{C}$ to be a diagonal matrix with entries chosen so that the bottom-left and top-right entries of $\mathcal{C}(P_{\mathcal{V}} - I)\mathcal{L}(I+\mathcal{L})^{-1}\mathcal{C}^{-1}$ are equal; that is, compared to $(P_{\mathcal{V}} - I)\mathcal{L}(I+\mathcal{L})^{-1}$, we make the bottom-left entry smaller, at the cost of making the top-right entry bigger. The choice of $\mathcal{C}$ that achieves this is

$$\mathcal{C} := \begin{pmatrix} I & 0 \\ 0 & \pi_{N\to s}^{-1/2}\pi_{s+1\to -N}^{1/2}\rho_{\mathcal{A}}(k)^{-1/2} I \end{pmatrix}, \tag{8.32}$$

with then

$$\begin{aligned} &\|\mathcal{C}(P_{\mathcal{V}} - I)\mathcal{L}(I+\mathcal{L})^{-1}\mathcal{C}^{-1}\|_{H^s_k(\Gamma_-)\to H^s_k(\Gamma_-)} \\ &\quad \le C \begin{pmatrix} \pi_{N\to -N}\rho_{\mathcal{A}}(k) & \pi_{N\to s}^{1/2}\pi_{s+1\to -N}^{1/2}\rho_{\mathcal{A}}(k)^{1/2} \\ \pi_{N\to s}^{1/2}\pi_{s+1\to -N}^{1/2}\rho_{\mathcal{A}}(k)^{1/2} & \pi_{s+1\to s} \end{pmatrix}. \end{aligned} \tag{8.33}$$

Inequality (8.17) then implies that

$$\|\mathcal{C}(P_{\mathcal{V}} - I)\mathcal{L}(I+\mathcal{L})^{-1}\mathcal{C}^{-1}\|_{H^s_k(\Gamma_-)\to H^s_k(\Gamma_-)} < \frac{1}{2}, \tag{8.34}$$

and so $I + (P_{\mathcal{V}} - I)\mathcal{L}(I+\mathcal{L})^{-1}$ is invertible and

$$(I + (P_{\mathcal{V}} - I)\mathcal{L}(I+\mathcal{L})^{-1})^{-1} = \mathcal{C}^{-1}(I + \mathcal{C}(P_{\mathcal{V}} - I)\mathcal{L}(I+\mathcal{L})^{-1}\mathcal{C}^{-1})^{-1}\mathcal{C}. \tag{8.35}$$

We now prove Theorem 8.11, assuming both Lemma 8.16 and the following lemma.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Lemma 8.17. Let $s \geq 0$, $k_0 > 0$, and suppose $\mathcal{A}$ satisfies Assumption 8.4. Then, for any $N \in \mathbb{R}$, there exists $C > 0$ such that for all $k > k_0$, $\mathcal{V} \subset H^s_k(\Gamma_-)$ and $P_\mathcal{V} \colon H^s_k(\Gamma_-) \to \mathcal{V}$ projectors onto $\mathcal{V}$,

$$\|(I+\mathcal{L})^{-1}(I-P_\mathcal{V})\|_{H^s_k(\Gamma_-)\to H^s_k(\Gamma_-)} \leq \begin{pmatrix} C\pi_{N\to -N}\rho_\mathcal{A}(k) & C\pi_{s\to -N}\rho_\mathcal{A}(k) \\ C\pi_{N\to s} & (L_{\max}+Ck^{-1})\pi_{s\to s} \end{pmatrix}. \tag{8.36}$$

Proof of Theorem 8.11. By (8.24), we need to bound the high- and low-frequency components of $(I+P_\mathcal{V}\mathcal{L})^{-1}(I-P_\mathcal{V})$. Furthermore, by the combination of (8.26) and (8.35),

$$\begin{aligned}
&(I+P_\mathcal{V}\mathcal{L})^{-1}(I-P_\mathcal{V}) \\
&\quad = (I+\mathcal{L})^{-1}(I-P_\mathcal{V})(I+(P_\mathcal{V}-I)\mathcal{L}(I+\mathcal{L})^{-1})^{-1}(I-P_\mathcal{V}) \\
&\quad = (I+\mathcal{L})^{-1}(I-P_\mathcal{V})\mathcal{C}^{-1}(I+\mathcal{C}(P_\mathcal{V}-I)\mathcal{L}(I+\mathcal{L})^{-1}\mathcal{C}^{-1})^{-1}\mathcal{C}(I-P_\mathcal{V}).
\end{aligned} \tag{8.37}$$

By (8.33), under the assumption (8.17),

$$\begin{aligned}
&\|(I+\mathcal{C}(P_\mathcal{V}-I)\mathcal{L}(I+\mathcal{L})^{-1}\mathcal{C}^{-1})^{-1}\|_{H^s_k\to H^s_k} \\
&\quad \leq \sum_{j=0}^{\infty} \|[\mathcal{C}(P_\mathcal{V}-I)\mathcal{L}(I+\mathcal{L})^{-1}\mathcal{C}^{-1}]\|^j_{H^s_k\to H^s_k} \\
&\quad \leq \sum_{j=0}^{\infty} C^j \begin{pmatrix} \pi_{N\to -N}\rho_\mathcal{A}(k) & \pi^{1/2}_{N\to s}\pi^{1/2}_{s+1\to -N}\rho_\mathcal{A}(k)^{1/2} \\ \pi^{1/2}_{N\to s}\pi^{1/2}_{s+1\to -N}\rho_\mathcal{A}(k)^{1/2} & \pi_{s+1\to s} \end{pmatrix}^j \\
&\quad = \left(I - C \begin{pmatrix} \pi_{N\to -N}\rho_\mathcal{A}(k) & \pi^{1/2}_{N\to s}\pi^{1/2}_{s+1\to -N}\rho_\mathcal{A}(k)^{1/2} \\ \pi^{1/2}_{N\to s}\pi^{1/2}_{s+1\to -N}\rho_\mathcal{A}(k)^{1/2} & \pi_{s+1\to s} \end{pmatrix} \right)^{-1} \\
&\quad = \frac{1}{\big(1-C\pi_{N\to -N}\rho_\mathcal{A}(k)\big)\big(1-C\pi_{s+1\to s}\big) - C^2\pi_{N\to s}\pi_{s+1\to -N}\rho_\mathcal{A}(k)} \\
&\qquad \times \begin{pmatrix} 1-C\pi_{s+1\to s} & C\pi^{1/2}_{N\to s}\pi^{1/2}_{s+1\to -N}\rho_\mathcal{A}(k)^{1/2} \\ C\pi^{1/2}_{N\to s}\pi^{1/2}_{s+1\to -N}\rho_\mathcal{A}(k)^{1/2} & 1-C\pi_{N\to -N}\rho_\mathcal{A}(k) \end{pmatrix} \\
&\quad \leq I + C\mathcal{D}_1,
\end{aligned}$$

where

$$\mathcal{D}_1 = \begin{pmatrix} \big(\pi_{N\to -N}+\pi_{N\to s}\pi_{s+1\to -N}\big)\rho_\mathcal{A}(k)+\pi_{s+1\to s} & \pi^{1/2}_{N\to s}\pi^{1/2}_{s+1\to -N}\rho_\mathcal{A}(k)^{1/2} \\ \pi^{1/2}_{N\to s}\pi^{1/2}_{s+1\to -N}\rho_\mathcal{A}(k)^{1/2} & \big(\pi_{N\to -N}+\pi_{N\to s}\pi_{s+1\to -N}\big)\rho_\mathcal{A}(k)+\pi_{s+1\to s} \end{pmatrix}.$$

In the last step we use the fact that $Ca < 1$, $(1-Ca)^{-1} \leq 1 + C'a$ in the diagonal entries and $(1-Ca)^{-1} \leq C'$ in the off-diagonal entries. Therefore, by (8.35) and

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

the definition of $\mathcal{C}$ (8.32),

$$\|(I + (P_{\mathcal{V}} - I)\mathcal{L}(I + \mathcal{L})^{-1})^{-1}\|_{H^s_k \to H^s_k} \leq I + C\mathcal{D}_2,$$

where

$$\mathcal{D}_2 = \begin{pmatrix} \left(\pi_{N\to -N} + \pi_{N\to s}\pi_{s+1\to -N}\right)\rho_{\mathcal{A}}(k) + \pi_{s+1\to s} & \pi_{s+1\to -N} \\ \pi_{N\to s}\rho_{\mathcal{A}}(k) & \left(\pi_{N\to -N} + \pi_{N\to s}\pi_{s+1\to -N}\right)\rho_{\mathcal{A}}(k) + \pi_{s+1\to s} \end{pmatrix}.$$

Now, with $\Pi_{\mathscr{L}}$ and $\Pi_{\mathscr{H}}$ defined by (8.28), by Lemma 8.3,

$$\|\Pi_{\mathscr{L}}\|_{H^{-N}_\hbar \to H^N_\hbar} \leq C \quad \text{and} \quad \|\Pi_{\mathscr{H}}\|_{H^t_\hbar \to H^t_\hbar} \leq 1, \tag{8.38}$$

and thus, by the notation (8.16),

$$\|I - P_{\mathcal{V}}\|_{H^s_k \to H^s_k} \leq \begin{pmatrix} C\pi_{N\to -N} & C\pi_{s\to -N} \\ C\pi_{N\to s} & \pi_{s\to s} \end{pmatrix}.$$

Therefore, under the assumption (8.17),

$$\|(I + (P_{\mathcal{V}} - I)\mathcal{L}(I + \mathcal{L})^{-1})^{-1}(I - P_{\mathcal{V}})\|_{H^s_k \to H^s_k} \leq \mathcal{D}_3,$$

where $\mathcal{D}_3$ is too large to display, but has components given by

$$\begin{aligned}
\mathcal{D}_3(1,1) &= C\left(\pi_{N\to -N} + \pi_{N\to s}\pi_{s+1\to -N}\right), \\
\mathcal{D}_3(1,2) &= C\left(\pi_{s\to -N} + \pi_{s+1\to -N}\pi_{s\to s}\right), \\
\mathcal{D}_3(2,1) &= C\pi_{N\to s}, \\
\mathcal{D}_3(2,2) &= C\pi_{N\to s}\pi_{s\to -N}\rho_{\mathcal{A}}(k) \\
&\quad + \pi_{s\to s}\left(1 + C\left(\pi_{N\to -N} + \pi_{N\to s}\pi_{s+1\to -N}\right)\rho_{\mathcal{A}}(k) + C\pi_{s+1\to s}\right).
\end{aligned}$$

Finally, combining this last displayed inequality with (8.37), Lemma 8.17, the assumption (8.17), and the fact that $\pi_{N\to -N} \leq \pi_{s\to -s}$ for $s \leq N$ by the definition (8.16), we obtain

$$\begin{aligned}
&\|\Pi_{\mathscr{L}}(I + P_{\mathcal{V}}\mathcal{L})^{-1}(I - P_{\mathcal{V}})\Pi_{\mathscr{L}}\|_{H^s_k(\Gamma)\to H^s_k(\Gamma)} \\
&\quad \leq C\pi_{N\to -N}\rho_{\mathcal{A}}(k)\left(\pi_{N\to -N} + \pi_{N\to s}\pi_{s+1\to -N}\right) + C\pi_{N\to s}\pi_{s\to -N}\rho_{\mathcal{A}}(k),
\end{aligned} \tag{8.39}$$

$$\begin{aligned}
&\|\Pi_{\mathscr{L}}(I + P_{\mathcal{V}}\mathcal{L})^{-1}(I - P_{\mathcal{V}})\Pi_{\mathscr{H}}\|_{H^s_k(\Gamma)\to H^s_k(\Gamma)} \\
&\quad \leq C\pi_{N\to -N}\rho_{\mathcal{A}}(k)\left(\pi_{s\to -N} + \pi_{s+1\to -N}\pi_{s\to s}\right) \\
&\qquad + C\pi_{s\to -N}\rho_{\mathcal{A}}(k)\left[C\pi_{N\to s}\pi_{s\to -N}\rho_{\mathcal{A}}(k)\right. \\
&\qquad \left. + \pi_{s\to s}\left(1 + C\left(\pi_{N\to -N} + \pi_{N\to s}\pi_{s+1\to -N}\right)\rho_{\mathcal{A}}(k) + C\pi_{s+1\to s}\right)\right] \\
&\quad \leq C\pi_{N\to -N}\rho_{\mathcal{A}}(k)\left(\pi_{s\to -N} + \pi_{s+1\to -N}\pi_{s\to s}\right) + C\pi_{s\to s}\pi_{s\to -N}\rho_{\mathcal{A}}(k) \\
&\quad \leq C\pi_{s\to s}\pi_{s\to -N}\rho_{\mathcal{A}}(k),
\end{aligned} \tag{8.40}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

$$\|\Pi_{\mathscr{H}}(I+P_{\mathcal{V}}\mathcal{L})^{-1}(I-P_{\mathcal{V}})\Pi_{\mathscr{L}}\|_{H_k^s(\Gamma)\to H_k^s(\Gamma)} \le C\pi_{N\to s}\left(\pi_{N\to s}\pi_{s+1\to -N}+\pi_{s\to s}\right), \tag{8.41}$$

$$\begin{aligned}
&\|\Pi_{\mathscr{H}}(I+P_{\mathcal{V}}\mathcal{L})^{-1}(I-P_{\mathcal{V}})\Pi_{\mathscr{H}}\|_{H_k^s(\Gamma)\to H_k^s(\Gamma)}\\
&\quad\le (L_{\max}+Ck^{-1})\pi_{s\to s}\big[C\pi_{N\to s}\pi_{s\to -N}\rho_{\mathcal{A}}(k)\\
&\qquad+\pi_{s\to s}\left(1+C(\pi_{N\to -N}+\pi_{N\to s}\pi_{s+1\to -N})\rho_{\mathcal{A}}(k)+C\pi_{s+1\to s}\right)\big]\\
&\quad\le (L_{\max}+Ck^{-1})\pi_{s\to s}\left(\pi_{s\to s}+C\big[\pi_{N\to s}\pi_{s\to -N}\rho_{\mathcal{A}}(k)\right.\\
&\qquad\left.+\pi_{s\to s}\left((\pi_{N\to -N}+\pi_{N\to s}\pi_{s+1\to -N})\rho_{\mathcal{A}}(k)+\pi_{s+1\to s}\right)\big]\right).
\end{aligned} \tag{8.42}$$

The bound (8.18) then follows from (8.39) and (8.40), using that $\pi_{N\to -N}\le\pi_{s\to -N}$, $\pi_{N\to s}\le\pi_{s\to s}$ and $\rho_{\mathcal{A}}(k)\ge c$. The bound (8.19) then follows from (8.41) and (8.42) in a similar way, using that $\pi_{s\to s}\ge 1$ (since $P_{\mathcal{V}}$ is a projection). □

It therefore remains to prove Lemmas 8.16 and 8.17.

Proof of Lemma 8.16. For the top-left entry of the matrix corresponding to $(P_{\mathcal{V}}-I)\mathcal{L}(I+\mathcal{L})^{-1}$ via (8.30), by (8.38) and (8.9),

$$\begin{aligned}
&\|\Pi_{\mathscr{L}}(P_{\mathcal{V}}-I)\mathcal{L}(I+\mathcal{L})^{-1}\Pi_{\mathscr{L}}\|_{H_k^s\to H_k^s}\\
&\quad=\|\Pi_{\mathscr{L}}(P_{\mathcal{V}}-I)(I-(I+\mathcal{L})^{-1})\Pi_{\mathscr{L}}\|_{H_k^s\to H_k^s}\\
&\quad\le\|\Pi_{\mathscr{L}}\|_{H_k^{-N}\to H_k^s}\pi_{N\to -N}\|I-(I+\mathcal{L})^{-1}\|_{H_k^N\to H_k^N}\|\Pi_{\mathscr{L}}\|_{H_k^s\to H_k^N}\\
&\quad\le C\rho_{\mathcal{A}}(k)\pi_{N\to -N}.
\end{aligned}$$

For the top-right entry, let $\chi\in C_c^\infty((-1-\epsilon,1+\epsilon))$ with $\operatorname{supp}(1-\chi)\cap[-1,1]=\emptyset$ and observe that by the definition of $\Pi_{\mathscr{H}}$ (8.28), $\Pi_{\mathscr{H}}=(I-\chi(-k^{-2}\Delta_\Gamma))\Pi_{\mathscr{H}}$. Then, by Corollary 8.10 and (8.38),

$$\begin{aligned}
&\|\Pi_{\mathscr{L}}(P_{\mathcal{V}}-I)\mathcal{L}(I+\mathcal{L})^{-1}\Pi_{\mathscr{H}}\|_{H_k^s\to H_k^s}\\
&\quad\le\|\Pi_{\mathscr{L}}\|_{H_k^{-N}\to H_k^s}\pi_{s+1\to -N}\|\mathcal{L}(I+\mathcal{L})^{-1}(I-\chi(-k^{-2}\Delta_\Gamma))\|_{H_k^s\to H_k^{s+1}}\|\Pi_{\mathscr{H}}\|_{H_k^s\to H_k^s}\\
&\quad\le C\pi_{s+1\to -N}.
\end{aligned}$$

For the bottom-left entry, by (8.38) and (8.9),

$$\begin{aligned}
&\|\Pi_{\mathscr{H}}(P_{\mathcal{V}}-I)\mathcal{L}(I+\mathcal{L})^{-1}\Pi_{\mathscr{L}}\|_{H_k^s\to H_k^s}\\
&\quad=\|\Pi_{\mathscr{H}}(P_{\mathcal{V}}-I)(I-(I+\mathcal{L})^{-1})\Pi_{\mathscr{L}}\|_{H_k^s\to H_k^s}\\
&\quad\le\|\Pi_{\mathscr{H}}\|_{H_k^s\to H_k^s}\pi_{N\to s}\|I-(I+\mathcal{L})^{-1}\|_{H_k^N\to H_k^N}\|\Pi_{\mathscr{L}}\|_{H_k^s\to H_k^N}\\
&\quad\le C\rho_{\mathcal{A}}(k)\pi_{N\to s}.
\end{aligned}$$

Finally, for the bottom-right entry, by (8.38) and Corollary 8.10,

$$\begin{aligned}
&\|\Pi_{\mathscr{H}}(P_{\mathcal{V}}-I)\mathcal{L}(I+\mathcal{L})^{-1}\Pi_{\mathscr{H}}\|_{H_k^s\to H_k^s}\\
&\quad\le\|\Pi_{\mathscr{H}}\|_{H_k^s\to H_k^s}\pi_{s+1\to s}\|\mathcal{L}(I+\mathcal{L})^{-1}(I-\chi(-k^{-2}\Delta_\Gamma))\|_{H_k^s\to H_k^{s+1}}\|\Pi_{\mathscr{H}}\|_{H_k^s\to H_k^s}\\
&\quad\le C\pi_{s+1\to s}.
\end{aligned}$$

□

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Proof of Lemma 8.17. For the top-left entry of the matrix corresponding to $(I + L)^{-1}(I - P_{\mathcal{V}})$ via (8.30), by (8.38) and (8.9),

$$\begin{aligned}
&\|\Pi_{\mathscr{L}}(I+\mathcal{L})^{-1}(I-P_{\mathcal{V}})\Pi_{\mathscr{L}}\|_{H^s_k\to H^s_k}\\
&\quad\le \|\Pi_{\mathscr{L}}\|_{H^{-N}_k\to H^s_k}\|(I+\mathcal{L})^{-1}\|_{H^{-N}_k\to H^{-N}_k}\pi_{N\to -N}\|\Pi_{\mathscr{L}}\|_{H^s_k\to H^N_k}\\
&\quad\le C\rho_{\mathcal{A}}(k)\pi_{N\to -N}.
\end{aligned}$$

For the top-right entry, again by (8.38) and (8.9),

$$\begin{aligned}
&\|\Pi_{\mathscr{L}}(I+\mathcal{L})^{-1}(I-P_{\mathcal{V}})\Pi_{\mathscr{H}}\|_{H^s_k\to H^s_k}\\
&\quad\le \|\Pi_{\mathscr{L}}\|_{H^{-N}_k\to H^s_k}\|(I+\mathcal{L})^{-1}\|_{H^{-N}_k\to H^{-N}_k}\pi_{s\to -N}\|\Pi_{\mathscr{H}}\|_{H^s_k\to H^s_k}\\
&\quad\le C\rho_{\mathcal{A}}(k)\pi_{s\to -N}.
\end{aligned}$$

For the bottom-left entry, let $\chi\in C_c^\infty((-1-\epsilon,1+\epsilon))$ with $\operatorname{supp}(1-\chi)\cap[-1,1]=\emptyset$ and observe that by the definition of $\Pi_{\mathscr{H}}$ (8.28), $\Pi_{\mathscr{H}} = \Pi_{\mathscr{H}}(I-\chi(-k^{-2}\Delta_\Gamma))$. Using this, along with (8.38) and (8.11), we obtain

$$\begin{aligned}
&\|\Pi_{\mathscr{H}}(I+\mathcal{L})^{-1}(I-P_{\mathcal{V}})\Pi_{\mathscr{L}}\|_{H^s_k\to H^s_k}\\
&\quad\le \|\Pi_{\mathscr{H}}\|_{H^s_k\to H^s_k}\|(I-\chi(-k^{-2}\Delta_\Gamma))(I+\mathcal{L})^{-1}\|_{H^s_k\to H^s_k}\pi_{N\to s}\|\Pi_{\mathscr{L}}\|_{H^s_k\to H^N_k}\\
&\quad\le C\pi_{N\to s}.
\end{aligned}$$

Finally, for the bottom-right entry, by (8.38) and (8.11) again,

$$\begin{aligned}
&\|\Pi_{\mathscr{H}}(I+\mathcal{L})^{-1}(I-P_{\mathcal{V}})\Pi_{\mathscr{H}}\|_{H^s_k\to H^s_k}\\
&\quad\le \|\Pi_{\mathscr{H}}\|_{H^s_k\to H^s_k}\|(I-\chi(-k^{-2}\Delta_\Gamma))(I+\mathcal{L})^{-1}\|_{H^s_k\to H^s_k}\pi_{s\to s}\|\Pi_{\mathscr{H}}\|_{H^s_k\to H^s_k}\\
&\quad\le (L_{\max}+Ck^{-1})\pi_{s\to s}.
\end{aligned}$$

□

8.5. *Proof of Theorem 8.12: references for Assumption 8.4*

Given Theorem 8.11, to prove Theorem 8.12 it only remains to check that Assumption 8.4 holds and to obtain estimates on $\pi(\Pi_{\mathcal{P}^p_{\mathcal{T}}})_{s\to t}$ when $\Pi_{\mathcal{P}^p_{\mathcal{T}}}\colon L^2(\Gamma_-)\to L^2(\Gamma_-)$ is the orthogonal projector onto $\mathcal{P}^p_{\mathcal{T}}$.

Using the results of Galkowski *et al.* (2022, Theorem 4.3) and Galkowski (2019, Chapter 4), §5.1.1 of Galkowski *et al.* (2025) shows that Assumption 8.4 holds for A_k and A'_k, and §5.1.2 of Galkowski *et al.* (2025) shows that Assumption 8.4 holds for B_{reg} and B'_{reg}, in all cases with $L_{\max}=1$.

For the $H^t_k\to L^2$ mapping properties of $\Pi_{\mathcal{P}^p_{\mathcal{T}}}$, we refer to Appendix A. Indeed, by Theorem A.12 and the fact that $\Pi_{\mathcal{P}^p_{\mathcal{T}}}$ is the L^2-orthogonal projector, there exists $C>0$ such that for $0\le t\le p+1$,

$$\|(I-\Pi_{\mathcal{P}^p_{\mathcal{T}}})\|_{H^t_k(\Gamma_-)\to L^2(\Gamma_-)} = \|(I-\Pi_{\mathcal{P}^p_{\mathcal{T}}})\|_{L^2(\Gamma_-)\to H^{-t}_k(\Gamma_-)} \le C(hk)^t.$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Hence

$$\begin{aligned} \pi_{1\to 0} &\le Chk, & \pi_{0\to 0} &= 1,\\ \pi_{p+1\to 0} &\le C(hk)^{p+1}, & \pi_{0\to -p-1} &\le C(hk)^{p+1},\\ \pi_{1\to -p-1} &\le \pi_{1\to 0}\pi_{0\to -p-1} \le C(hk)^{p+2},\\ \pi_{p+1\to -p-1} &\le \pi_{p+1\to 0}\pi_{0\to -p-1} \le C(hk)^{2p+2}. \end{aligned}$$

Theorem 8.12 now follows by inserting these bounds into Theorem 8.11 with $N = p + 1$ and $s = 0$, and recalling that $L_{\max} = 1$.

8.6. Pollution in the Galerkin h-BEM with piecewise polynomials

Pollution in the h-FEM (i.e. the fact that $\gg k^d$ degrees of freedom are required to accurately approximate a scattering solution at frequency k using the h-FEM) was famously identified in Ihlenburg and Babuška (1995*a*, 1997). However, it was long conjectured that the h-BEM does not suffer from the pollution effect. Theorem 8.12 shows that, indeed, there is no pollution for the h-BEM applied with a sound-soft, non-trapping obstacle (with this result first proved in Galkowski and Spence 2023). However, the rigorous versions of Informal Theorem 3.5, namely Galkowski *et al.* (2025, Theorems 2.10, 2.12, 2.13, 2.14 and 2.16), show that the h-BEM *does* suffer from pollution when applied to many trapping, sound-soft obstacles and even for sound-hard scattering by the disk.

We do not attempt to give complete proofs of these results here. Instead, we opt for a sketch of the arguments.

(1) *Quasimodes imply pollution.* By Lemma 8.14, the Galerkin approximation v_h to v satisfies

$$v - v_h = \left(I + \Pi_{\mathcal{P}^p_\mathcal{T}}\mathcal{L}\right)^{-1}\left(I - \Pi_{\mathcal{P}^p_\mathcal{T}}\right)\left(I - \Pi_{\mathcal{P}^p_\mathcal{T}}\right)v$$

(see (8.24) below). Therefore, to prove pollution, we need to show that

$$\left(I + \Pi_{\mathcal{P}^p_\mathcal{T}}\mathcal{L}\right)^{-1}\left(I - \Pi_{\mathcal{P}^p_\mathcal{T}}\right)$$

is large. A simple calculation shows that if

$$v := (I + \mathcal{L})^{-1}\left(I - \Pi_{\mathcal{P}^p_\mathcal{T}}\right)\widetilde{f}, \tag{8.43}$$

then

$$\left(I + \Pi_{\mathcal{P}^p_\mathcal{T}}\mathcal{L}\right)v = \left(I - \Pi_{\mathcal{P}^p_\mathcal{T}}\right)v = \left(I - \Pi_{\mathcal{P}^p_\mathcal{T}}\right)^2 v$$

that is, we need to find v with $\|v\|_{L^2} = 1$ such that (8.43) holds and $\left(I - \Pi_{\mathcal{P}^p_\mathcal{T}}\right)v$ is 'small'.

To do this, we assume that there is a *quasimode*, u for $\mathcal{A}$, that is, there are $u, f \in L^2(\Gamma_-)$ with

$$(I + \mathcal{L})u = f, \quad \|u\|_{L^2(\Gamma_-)} = 1 \quad \text{and} \quad \|f\|_{L^2(\Gamma_-)} \ll 1. \tag{8.44}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Moreover, we assume that both f and u are oscillating at frequency $\lesssim k$, but not frequency $\ll k$, that is, there is $\chi \in C_c^\infty(\mathbb{R})$ with $0 \notin \operatorname{supp}\chi$ such that

$$\|(I - \chi(-k^{-2}\Delta_{\Gamma_-}))f\|_{L^2} + \|(I - \chi(-k^{-2}\Delta_{\Gamma_-}))u\|_{L^2} \ll 1, \tag{8.45}$$

and thus by Theorem A.12,

$$\left\|\left(I - \Pi_{\mathcal{T}}^p\right)u\right\|_{L^2(\Gamma_-)} \leq C(hk)^{p+1}\|u\|_{L^2(\Gamma_-)}.$$

If there exists $\tilde{f}$ such that $f = \left(I - \Pi_{\mathcal{T}}^p\right)\tilde{f}$, then

$$\left\|\left(I + \Pi_{\mathcal{T}}^p\mathcal{L}\right)^{-1}\left(I - \Pi_{\mathcal{T}}^p\right)\right\|_{L^2(\Gamma_-)\to L^2(\Gamma_-)} \geq c(hk)^{-p-1}.$$

However, this cannot be true for all h since it contradicts the upper bound on the quasioptimality constant in Theorem 8.12, that is, we cannot solve $f = \left(I - \Pi_{\mathcal{T}}^p\right)\widetilde{f}$.

Motivated by this discussion, our goal is to find $\widetilde{f}$ that is as close as possible to satisfying $\left(I - \Pi_{\mathcal{T}}^p\right)\widetilde{f} = f$. Notice that since $v = (I + \mathcal{L})^{-1}\left(I - \Pi_{\mathcal{T}}^p\right)\tilde{f}$ and $(I + \mathcal{L})^{-1}$ is $O(1)$ on high frequencies (see Lemma 8.9), low-frequency errors cost more than high-frequency errors. Therefore we aim to solve $\left(I - \Pi_{\mathcal{T}}^p\right)\tilde{f} = f + e$, where e has as few low frequencies as possible.

The lower bounds on Galerkin piecewise polynomial approximations of k-oscillating functions from Galkowski (2025) (see more specifically Galkowski *et al.* 2025, Lemma 12.7 and Corollary 12.8) imply that there exists $\widetilde{f}$ such that

$$\left\|\left(I - \Pi_{\mathcal{T}}^p\right)\widetilde{f}\right\|_{L^2(\Gamma_-)} \leq C(hk)^{-p-1}\|f\|_{L^2(\Gamma_-)} \tag{8.46}$$

and

$$\left(I - \Pi_{\mathcal{T}}^p\right)\widetilde{f} = f + \chi_{\mathscr{L}}\left(I - \Pi_{\mathcal{T}}^p\right)\widetilde{f} + \chi_{\mathscr{H}}\left(I - \Pi_{\mathcal{T}}^p\right)\widetilde{f}.$$

Here $\chi_{\mathscr{L}}$ is a finite-rank operator (which can be taken to be zero when $p = 0$) and

$$\|\chi_{\mathscr{L}}\|_{H^{-M}(\Gamma)\to H^M(\Gamma)} \leq C_M \quad \text{and} \quad \|(I + \mathcal{L})^{-1}\chi_{\mathscr{H}}\|_{L^2(\Gamma_-)\to L^2(\Gamma_-)} \leq C$$

(i.e. $\chi_{\mathscr{L}}$ and $\chi_{\mathscr{H}}$ are low- and high-frequency cut-offs, respectively). Thus

$$v = (I + \mathcal{L})^{-1}f + (I + \mathcal{L})^{-1}\chi_{\mathscr{L}}\left(I - \Pi_{\mathcal{T}}^p\right)\widetilde{f} + (I + \mathcal{L})^{-1}\chi_{\mathscr{H}}\left(I - \Pi_{\mathcal{T}}^p\right)\widetilde{f}. \tag{8.47}$$

In the rest of this sketch we ignore the contribution from $\chi_{\mathscr{L}}$; including this contribution leads to a finite-dimensional possible obstruction to pollution. Using (8.47), (8.46) and the fact that $u = (I + L)^{-1}f$ (with $\|u\|_{L^2(\Gamma)} = 1$), we see that $\|v\|_{L^2(\Gamma_-)} \geq 1/2$ if $(hk)^{-p-1}\|f\|_{L^2(\Gamma_-)} \ll 1$. Finally, by (8.43),

$$\left(I - \Pi_{\mathcal{T}}^p\right)v = \left(I - \Pi_{\mathcal{T}}^p\right)\chi_{\mathrm{H}}(I + \mathcal{L})^{-1}\left(I - \Pi_{\mathcal{T}}^p\right)\widetilde{f} + \left(I - \Pi_{\mathcal{T}}^p\right)\chi_{\mathrm{L}}v,$$

where χ_{H} and χ_{L} are high- and low-frequency cut-offs, so that

$$\left\|\left(I - \Pi_{\mathcal{T}}^p\right)v\right\|_{L^2(\Gamma_-)} \leq C(hk)^{-p-1}\|f\|_{L^2(\Gamma_-)} + (hk)^{p+1}\|v\|_{L^2(\Gamma_-)}.$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

In summary,

$$\begin{aligned}
&\left\|\left(I+\Pi_{\mathcal{P}^p_{\mathcal{T}}}\mathcal{L}\right)^{-1}\left(I-\Pi_{\mathcal{P}^p_{\mathcal{T}}}\right)\right\|_{L^2(\Gamma_-)}\\
&\quad\geq \frac{\|v\|_{L^2(\Gamma_-)}}{\left\|\left(I-\Pi_{\mathcal{P}^p_{\mathcal{T}}}\right)v\right\|_{L^2(\Gamma_-)}}\\
&\quad\geq c\begin{cases}(hk)^{-p-1}, & \|f\|_{L^2(\Gamma_-)}\leq (hk)^{2(p+1)},\\ (hk)^{p+1}\|f\|^{-1}_{L^2(\Gamma_-)}, & (hk)^{2(p+1)}\leq \|f\|_{L^2(\Gamma_-)}\ll (hk)^{p+1}.\end{cases}
\end{aligned}$$

Since $\|v\|_{L^2(\Gamma_-)}/\|\left(I-\Pi_{\mathcal{P}^p_{\mathcal{T}}}\right)v\|_{L^2(\Gamma_-)}\geq 1$, by reducing c if necessary,

$$\begin{aligned}
&\left\|\left(I+\Pi_{\mathcal{P}^p_{\mathcal{T}}}\mathcal{L}\right)^{-1}\left(I-\Pi_{\mathcal{P}^p_{\mathcal{T}}}\right)\right\|_{L^2(\Gamma_-)}\\
&\quad\geq \frac{1}{2}+c\begin{cases}(hk)^{-p-1}, & 1\leq (hk)^{2(p+1)}\|f\|^{-1}_{L^2(\Gamma_-)},\\ (hk)^{p+1}\|f\|^{-1}_{L^2(\Gamma_-)}, & \|f\|^{-1}_{L^2(\Gamma_-)}(hk)^{2(p+1)}\leq 1,\end{cases}
\end{aligned}$$

which, if $\|f\|_{L^2(\Gamma_-)}\ll 1$, implies that the Galerkin h-BEM suffers from pollution.

(2) *Construction of quasimodes.* The proof that pollution occurs for the h-BEM is then complete if we can construct u, f satisfying (8.44) and (8.45). The existence of u and f satisfying (8.44) is guaranteed whenever $\rho_{\mathcal{A}}(k_j)\to\infty$ for some $k_j\to\infty$. However, the requirement (8.45) does not automatically hold for these quasimodes. In the case of the unit disk, the fact that the boundary layer operators are diagonal in the basis of trigonometric polynomials enables the use of classical Bessel function asymptotics together for the construction of quasimodes, leading to the following theorem.

Theorem 8.18 (Pollution for the Neumann BIEs on the unit disk). Let $\Upsilon>0$, $p\geq 0$, and $\Omega_-=B(0,1)\subset\mathbb{R}^2$, and $k_0>0$. Then there exists $c>0$ such that for all $k>k_0$ and all C^{p+1}-simplicial triangulations $\mathcal{T}$ with constant Υ (in the sense of Definition A.5), there exists $v\in L^2(\Gamma_-)$ such that the Galerkin approximation v_h to v in $\mathcal{P}^p_{\mathcal{T}}$ with $\mathcal{A}$ given by B_{reg} or B'_{reg}, if it exists, satisfies

$$\frac{\|v_h-v\|_{L^2(\Gamma_-)}}{\left\|\left(I-\Pi_{\mathcal{P}^p_{\mathcal{T}}}\right)v\right\|_{L^2(\Gamma_-)}}\geq \frac{1}{2}+c\begin{cases}(hk)^{-p-1}, & 1\leq (hk)^{2(p+1)}k^{1/3},\\ (hk)^{p+1}k^{1/3}, & k^{1/3}(hk)^{2(p+1)}\leq 1,\end{cases}$$

where $h:=h(\mathcal{T})$. In particular, Theorem 8.12 is optimal in this case.

However, except in the case of the unit ball, constructing quasimodes satisfying (8.45) in addition to (8.44) is accomplished by working with the PDE rather than the BIEs. Indeed, in Galkowski *et al.* (2025, §13), compactly supported functions $u\in H^2_k(\Omega_+)\cap H^1_0(\Omega_+)$ are constructed such that

$$(-k^{-2}\Delta-1)u=g,\quad \|g\|_{L^2}\ll k^{-1},\quad \|u\|_{L^2(\Omega_+)}=1,$$

and such that the microlocal concentration properties of v and g are well understood.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

One then shows that the right-hand side g can be solved away at the cost of boundary data that is k-oscillating. It is then possible to use the description of $\mathcal{A}^{-1}$ in terms of the outgoing Dirichlet-to-Neumann map and the interior impedance to Dirichlet map (see Galkowski *et al.* 2025, equation 13.1 or Chandler-Wilde *et al.* 2012, Theorem 2.33 and Galkowski *et al.* 2022, Lemma 7.4) to find u and f satisfying (8.44) and (8.45).

As an example of this, we consider the following type of domain.

Definition 8.19 (Four-diamond domain). Ω_- is a *four-diamond domain* if there exists $0 < \epsilon < \pi/2$ such that $\Omega_- = \cup_{i=1}^4 \Omega_i \Subset \mathbb{R}^2$, where Ω_i are open, convex, disjoint, and have smooth boundary so that

$$\partial((-\pi,\pi)\times(-\pi,\pi)) \setminus (\cup_{a,b\in\pm 1} B((a\pi,b\pi),\epsilon)) \subset \Gamma_- \subset \mathbb{R}^2 \setminus (\cup_{a,b\in\pm 1} B((a\pi,b\pi),\epsilon/2))$$

and

$$\Omega_- \cap (-\pi,\pi)\times(-\pi,\pi) = \emptyset.$$

Figure 8.1 shows an example of a four-diamond domain.

Using the method of images, it is easy to construct a compactly supported function $u_1 \in H_0^1(\Omega_+)$ that is supported in a small neighbourhood of the square with corners at $(0,\pm\pi)$ and $(\pm\pi,0)$ such that u_1 is microlocally concentrated near this square and with momentum tangent to the square, and u satisfies

$$\|(-k^{-2}\Delta - 1)u_1\|_{L^2(\Omega_+)} \leq Ck^{-2}, \quad \|u_1\|_{L^2(\Omega_+)} = 1, \quad \|k^{-1}\partial_\nu u_1\|_{L^2} \geq c > 0.$$

Using this microlocal information, one can construct $\tilde{u} \in H^2(\Omega_+)$ satisfying the Sommerfeld radiation condition such that

$$(-k^{-2}\Delta - 1)\tilde{u} = (-k^{-2}\Delta - 1)u_1,$$

with

$$\|\tilde{u}|_{\Gamma_-}\|_{L^2} + \|k^{-1}\partial_\nu \tilde{u}|_{\Gamma_-}\|_{L^2} \leq Ck\|(-k^{-2}\Delta - 1)u_1\|_{L^2} \leq Ck^{-1}$$

and $\tilde{u}_{\Gamma_-}$ is microlocally concentrated at frequencies $|\xi| \sim \sqrt{2}/2$. This is done using a propagation of singularities argument for an auxiliary Helmholtz-type problem with additional damping.

Since $u := u_1 - \tilde{u}$ satisfies the Sommerfeld radiation condition and

$$(-k^{-2}\Delta - 1)u = 0, \quad \frac{\|u|_{\Gamma_-}\|_{L^2}}{\|k^{-1}\partial_\nu u|_{\Gamma_-}\|_{L^2}} \geq ck.$$

This, together with Galkowski *et al.* (2025, equation 13.1), is immediately enough to construct a quasimode v satisfying (8.44) and (8.45) in the case of A_k. Slightly more analysis is necessary in the case of A_k' because one needs to find data for the impedance problem in Ω_- so that the resulting Dirichlet data is, in some sense, close to $\tilde{u}|_{\Gamma_-}$. This is again done using a propagation argument for an auxiliary Helmholtz-type problem with additional damping in Ω_-.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

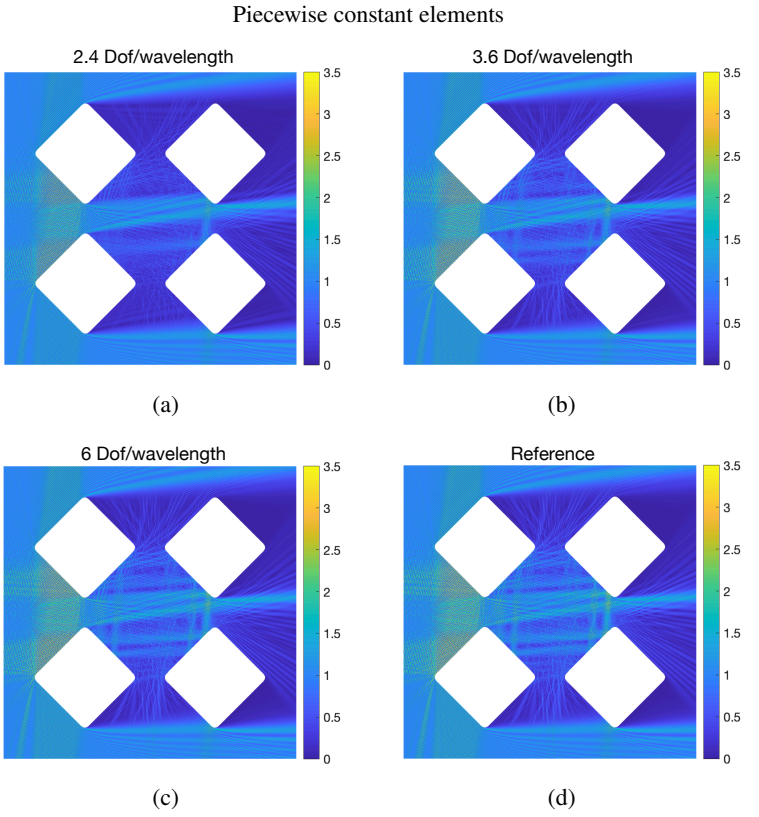
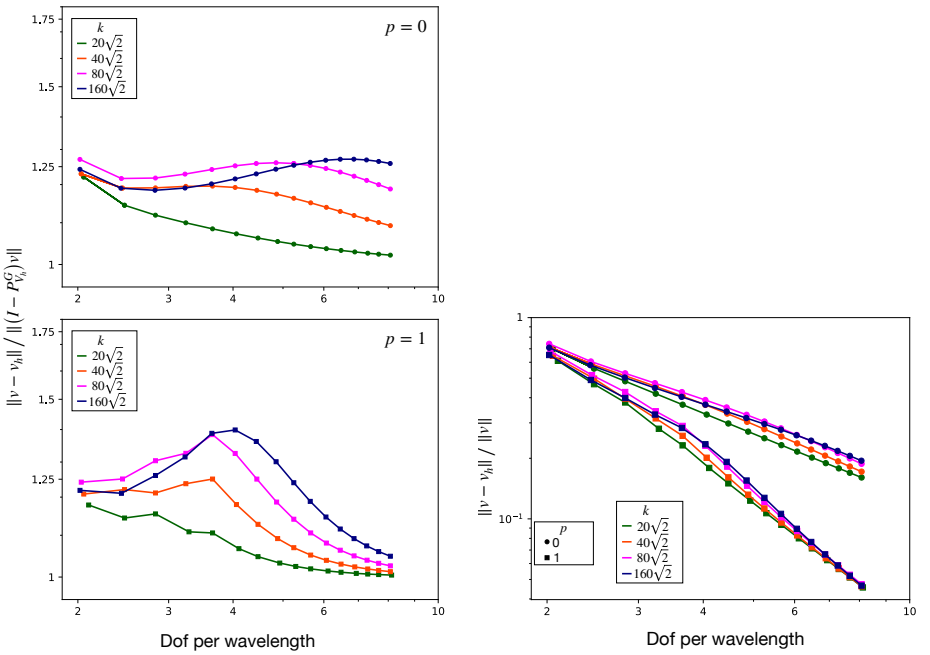


Figure 8.1. Absolute value of the total field at $k = 40\sqrt{2}$ when the plane wave $e^{ik\langle\omega,x\rangle}$, with $\omega = (\cos(5\pi/180), \sin(5\pi/180))$, is incident on a sound-soft domain consisting of four nearly square obstacles. The plot shows the Galerkin solutions for the BIE involving A_k at (a) 2.4, (b) 3.6 and (c) 6 points per wavelength and piecewise constant elements (i.e. $p = 0$), with the reference solution (d) computed with $p = 11$. These numbers of degrees of freedom per wavelength correspond to before, on and after the peak of the quasioptimality constant in Figure 8.2. Many of the features of the Galerkin and reference plots are similar. However, in the Galerkin solutions with low numbers of degrees of freedom per wavelength, the trapped rays are much less pronounced. The plots for piecewise linear elements (i.e. $p = 1$) look qualitatively similar. Figure from Galkowski *et al.* (2025).

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press



(a) (b)

Figure 8.2. (a) The implied quasioptimality constant for piecewise constant (top) and piecewise linear (bottom) elements when the plane wave $e^{ik\langle\omega,x\rangle}$ with $\omega = (\cos(5\pi/180), \sin(5\pi/180))$ is incident on a sound-soft domain consisting of four nearly square obstacles (see Figure 8.1), with this problem solved using the BIE involving A_k. (b) The corresponding L^2 relative error in the density. The reference and computed scattered solutions at $k = 40\sqrt{2}$ are shown in Figure 8.1. The presence of pollution in this example can be seen from both plots. In (a) we observe a peak in the implied quasioptimality constant that shifts up and to the right as k increases. In (b) the error increases for a fixed number of degrees of freedom per wavelength as k increases. Figure from Galkowski *et al.* (2025).

The results of this analysis for the four-diamond domains are as follows.

Theorem 8.20 (Pollution for the four-diamond domain when $p = 0$). Let $p = 0$, $k_0 > 0$, $\Upsilon > 0$, Ω_- be a four-diamond domain, $M > 0$, $\mathcal{J} \subset \mathbb{R}$ such that

$$\sup\{k^{-M}\rho_{\mathcal{A}}(k)\colon k > k_0,\ k \notin \mathcal{J}\} < \infty$$

with $\mathcal{A} = A_k$ or A_k', and $k_n := \sqrt{2}n$. Then there exists $c > 0$ such that for all $k \in \mathbb{R} \setminus \mathcal{J}$ and n such that $k_n^{-1} \le \delta \le 1$, $|k - k_n| < \delta^{-1}k_n^{-1}$, and all C^1-simplicial triangulations $\mathcal{T}$ with constant Υ (in the sense of Definition A.5), there

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

exists $v \in L^2(\Gamma)$ such that the Galerkin approximation v_h to v in $\mathcal{P}^0_{\mathcal{T}}$, if it exists, satisfies

$$\frac{\|v_h - v\|_{L^2(\Gamma_-)}}{\left\|\left(I - \Pi_{\mathcal{P}^p_{\mathcal{T}}}\right)v\right\|_{L^2(\Gamma_-)}} \geq \frac{1}{2} + c \begin{cases} (hk_n)^{-1}, & 1 \leq (hk_n)^2 \delta k_n, \\ (hk_n)\delta k_n, & \delta k_n (hk_n)^2 \leq 1, \end{cases}$$

where $h := h(\mathcal{T})$.

Theorem 8.21 (Pollution for the four-diamond domain when $p \geq 1$). Let $\Upsilon > 0$, $p \geq 1$, Ω_- be a four-diamond domain $\mathcal{A} = A_k$ or A'_k, and $k_n := \sqrt{2}n$.

If there are $C_1 > 0, k_0 > 0$ such that

$$\rho_{\mathcal{A}}(k) \leq C_1 k^2 \quad \text{for all } k \geq k_0, \tag{8.48}$$

then there exist $c, C > 0$ such that for all $k \geq k_0$ and n such that $Ck_n^{-1} \leq \delta \leq 1$ and $|k - k_n| < \delta^{-1} k_n^{-1}$, and all C^{p+1}-simplicial triangulations $\mathcal{T}$ with constant Υ (in the sense of Definition A.5), there exists $v \in L^2(\Gamma)$ such that the Galerkin approximation v_h to v in $\mathcal{P}^p_{\mathcal{T}}$, if it exists, satisfies

$$\frac{\|v_h - v\|_{L^2(\Gamma_-)}}{\left\|\left(I - \Pi_{\mathcal{P}^0_{\mathcal{T}}}\right)v\right\|_{L^2(\Gamma_-)}} \geq \frac{1}{2} + c \begin{cases} (hk_n)^{-p-1}, & 1 \leq (hk_n)^{2(p+1)} \delta k_n, \\ (hk_n)^{p+1}\delta k_n, & \delta k_n (hk_n)^{2(p+1)} \leq 1, \end{cases}$$

where $h := h(\mathcal{T})$.

Remark 8.22 (Discussion of Theorems 8.20 and 8.21).

(i) When Ω_- consists of two aligned squares (or rounded squares) – where the trapping is the same nature (*parabolic*) as for a four-diamond domain (albeit on a smaller set in phase space) – the bound (8.48) is proved in Chandler-Wilde *et al.* (2020, Corollary 1.14). Furthermore, from both quasimode considerations and the bounds of Christianson and Wunsch (2013) and Christianson (2018) in the setting of scattering by metrics of revolution, Chandler-Wilde *et al.* (2020) conjecture that for Ω_- as in Theorem 8.20, and $\mathcal{A} \in \{A_k, A'_k\}$

$$\rho_{\mathcal{A}}(k) \leq Ck. \tag{8.49}$$

If the bound (8.49) holds, then Theorems 8.20 and 8.21 show that Theorem 8.12 is optimal in this case and $\mathcal{J} = \emptyset$ in Theorem 8.20.

(ii) We emphasize that the phenomenon behind Theorems 8.20 and 8.21 is *not* sparse in frequency. Indeed, both Theorems 8.20 and 8.21 hold with $\delta = 1$ for a set of k with infinite measure (since $\sum_{n=1}^{\infty} n^{-1} = \infty$). Furthermore, although we prove Theorem 8.20 with $k_n = \sqrt{2}n$, it is easy to generalize our construction in the proof of Theorem 8.20 to take $k_n = \sqrt{m_n^2 + \ell_n^2}$ for any $m_n, \ell_n \in \mathbb{Z}$ with $c < |m_n/\ell_n| < C$ and such that if $m_n/\ell_n = p_n/q_n$ with p_n and q_n relatively prime, then $|q_n| \leq C$.

(iii) Theorems 8.20 and 8.21 are illustrated by numerical experiments in Figures 8.1 and 8.2 (taken from Galkowski *et al.* 2025).

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

9. h-FEM with non-uniform meshes defined by ray-dynamics (R4)

In this section we consider an H^∞-Helmholtz problem (in the sense of Definition 4.6) with $\Gamma_{\mathrm{p}} = \emptyset$. The goal is to present a simple version of the results of Averseng *et al.* (2025). We do not give complete proofs; instead we give a sketch proof, indicating where extra input is needed.

To begin we recall the notation $\Omega_{\mathcal{K}}$, $\Omega_{\mathcal{V}}$, $\Omega_{\mathcal{I}}$ and $\Omega_{\mathcal{P}}$, respectively, for a neighbourhood of the cavity, the visible set, the invisible set and the PML from Section 3.4 (see also Figure 3.1).

Given a triangulation $\mathcal{T}$ of Ω, we also recall the meshwidth parameters $h_\star$, $\star \in \{\mathcal{K}, \mathcal{V}, \mathcal{I}, \mathcal{P}\}$ and $h = \max_{T\in\mathcal{T}} h_T$ from (3.19). In addition, we define the following measure of local uniformity of the triangulation at scale $\epsilon > 0$:

$$U(\mathcal{T}, \epsilon) := \sup_{x\in\Omega} \sup_{\substack{T_1,T_2\in\mathcal{T} \\ T_1\cap B(x,\epsilon)\neq\emptyset \\ T_2\cap B(x,\epsilon)=\emptyset}} \frac{\operatorname{diam}(T_1)}{\operatorname{diam}(T_2)}.$$

A triangulation $\mathcal{T}$ is *quasiuniform at scale ϵ with constant $\gamma_0 > 0$* if $U(\mathcal{T}, \epsilon) \leq \gamma_0^{-1}$.

Define the matrices

$$\mathcal{C} := \begin{pmatrix} \rho(k) & \sqrt{k\rho(k)} & 0 & 0 \\ \sqrt{k\rho(k)} & k & k & 0 \\ 0 & k & k & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathcal{H} := \begin{pmatrix} h_{\mathcal{K}} & 0 & 0 & 0 \\ 0 & h_{\mathcal{V}} & 0 & 0 \\ 0 & 0 & h_{\mathcal{I}} & 0 \\ 0 & 0 & 0 & h_{\mathcal{P}} \end{pmatrix},$$

$$\mathscr{F} := \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix},$$

$$\mathscr{T} := \begin{pmatrix} 1 & (h_{\mathcal{V}}k)^{2p}\sqrt{k\rho(k)} & (h_{\mathcal{V}}k)^{2p}\sqrt{k\rho(k)}(h_{\mathcal{I}}k)^{2p}k & 0 \\ (h_{\mathcal{K}}k)^{2p}\sqrt{k\rho(k)} & 1 & (h_{\mathcal{I}}k)^{2p}k & 0 \\ (h_{\mathcal{K}}k)^{2p}\sqrt{k\rho}(h_{\mathcal{V}}k)^{2p}k & (h_{\mathcal{V}}k)^{2p}k & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Conceptually, $\mathcal{C}$ is the norm of the localized data-to-solution map (with $\mathcal{C}$ standing for 'communication') and $\mathscr{T}$ controls the propagation of Galerkin errors between subdomains according to the graph in Figure 9.2 (with a simplified version, Figure 9.1, given in the sketch of the proof in Section 9.1.2).

Theorem 9.1. Suppose a is the sesquilinear form defining an H^∞-Helmholtz problem approximated using a C^∞ PML with $C^{1,1}$ truncation boundary. Let $k_0, N, \Upsilon > 0$, $p \in \mathbb{N}\setminus\{0\}$, $\mathcal{J} \subset \mathbb{R}_+$ such that

$$\sup\{k^{-N}\rho(k)\colon k \in (k_0,\infty)\setminus\mathcal{J}\} < \infty,$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

and let $\Omega'_\star$ be compactly contained in $\Omega_\star$ with respect to the subspace topology of $\overline{\Omega}$, $\star \in \{\mathcal{K}, \mathcal{V}, \mathcal{I}, \mathcal{P}\}$.

There exist $c, C > 0$ such that for all $k \in (k_0, \infty) \setminus \mathcal{J}$, all affine-conforming C^{p+1}-simplicial triangulations $\mathcal{T}$ with constant Υ (in the sense of Definition A.8) that are quasiuniform at scale $(\Upsilon + k)^{-1}$ and satisfy

$$(h_{\mathcal{K}}k)^{2p}\rho(k) + (h_{\mathcal{V}}k)^{2p}k + (h_{\mathcal{I}}k)^{2p}k + (h_{\mathcal{P}}k)^{2p} \leq c \tag{9.1}$$

for all $w_{h,\star} \in V^p_{\mathcal{T}_k}$, with $\star \in \{\mathcal{K}, \mathcal{V}, \mathcal{I}, \mathcal{P}\}$, and all $u \in H^1_k(\Omega)$, then the Galerkin approximation $u_h \in \mathcal{P}^p_{\mathcal{T}}$ to u exists, is unique, and satisfies, for $0 \leq m \leq p$,

$$\begin{pmatrix} \|u-u_h\|_{H^{1-m}_k(\Omega'_{\mathcal{K}})} \\ \|u-u_h\|_{H^{1-m}_k(\Omega'_{\mathcal{V}})} \\ \|u-u_h\|_{H^{1-m}_k(\Omega'_{\mathcal{I}})} \\ \|u-u_h\|_{H^{1-m}_k(\Omega'_{\mathcal{P}})} \end{pmatrix} \leq C\left[(\mathcal{H}k)^m + \mathscr{T}\mathcal{C}(\mathcal{H}k)^p + k^{-N}(hk)^m\mathscr{F}\right] \begin{pmatrix} \|u-w_{h,\mathcal{K}}\|_{H^1_k(\Omega_{\mathcal{K}})} \\ \|u-w_{h,\mathcal{V}}\|_{H^1_k(\Omega_{\mathcal{V}})} \\ \|u-w_{h,\mathcal{I}}\|_{H^1_k(\Omega_{\mathcal{I}})} \\ \|u-w_{h,\mathcal{P}}\|_{H^1_k(\Omega_{\mathcal{P}})} \end{pmatrix}, \tag{9.2}$$

where the inequality in (9.2) is understood componentwise.

Remark 9.2. From the estimate (9.2) and the interpretation of $\mathscr{T}$ as the propagation of Galerkin errors, the matrix $\mathcal{C}(\mathcal{H}k)^p$ should be viewed as mapping best approximation errors to Galerkin errors; we discuss this interpretation in Section 9.1.3.

Remark 9.3. Numerical experiments illustrating the consequences of Theorem 9.1 summarized in Table 3.1 can be found in Averseng *et al.* (2025, §2).

For a uniform mesh ($h_{\mathcal{K}} = h_{\mathcal{V}} = h_{\mathcal{I}} = h_{\mathcal{P}} = h$), (9.2) recovers the result of Theorem 6.27. Indeed, for a uniform mesh with $(hk)^{2p}\rho$ sufficiently small, all the elements of the matrix $\mathscr{T}$ are bounded by a constant, and all the elements of $\mathscr{C}(\mathcal{H}k)^p$ are bounded by $\rho(k)(hk)^p$. However, Theorem 9.1 provides much more information than Theorem 6.27: it describes how the best approximation errors and local width in each region affect the Galerkin error in all other regions.

Theorem 9.1 is most interesting when $\rho(k) \gg k$, which is equivalent to the problem being trapping, i.e. $\mathcal{K} \neq \emptyset$ (see Bony *et al.* 2010, Dyatlov and Zworski 2019, Theorem 7.1). In particular, Theorem 9.1 shows that in the trapping case there exist triangulations with $(hk)^p\rho(k) \gg 1$ whose finite element solutions have guaranteed k-uniform quasioptimality. Even when $\mathcal{K} = \emptyset$, Theorem 9.1 gives new information including that one needs only a fixed number of points per wavelength in the PML.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

To compare with the estimate in Theorem 6.27 on relative error (i.e. (6.45)), we state the following corollary of Theorem 9.1, which follows from standard piecewise polynomial approximation estimates (see e.g. Appendix B).

Corollary 9.4. Suppose a is the sesquilinear form defining an H^∞-Helmholtz problem approximated using a C^∞ PML with $C^{1,1}$ truncation boundary. Let $k_0, N, \Upsilon, p \in \mathbb{N} \setminus \{0\}$, $\mathcal{J} \subset \mathbb{R}_+$ such that

$$\sup\{k^{-N}\rho(k)\colon\ k \in (k_0,\infty) \setminus \mathcal{J}\} < \infty,$$

and let $\Omega'_\star$ be compactly contained in $\Omega_\star$ with respect to the subspace topology of $\overline{\Omega}$, $\star \in \{\mathcal{K}, \mathcal{V}, \mathcal{I}, \mathcal{P}\}$.

There exist $c, C > 0$ such that for all $k \in (k_0,\infty)\setminus\mathcal{J}$, all affine-conforming C^{p+1}-simplicial triangulations $\mathcal{T}$ with constant Υ that are quasiuniform at scale $(\Upsilon + k)^{-1}$ with constant Υ and satisfy (9.1) for all $w_{h,\star} \in V^p_{\mathcal{T}_k}$, with $\star \in \{\mathcal{K}, \mathcal{V}, \mathcal{I}, \mathcal{P}\}$, and all $u \in H^1_k(\Omega)$, then the Galerkin approximation $u_h \in \mathcal{P}^p_{\mathcal{T}}$ to u exists, is unique, and satisfies, for $0 \le m \le p$,

$$\begin{pmatrix} \|u-u_h\|_{H^{1-m}_k(\Omega'_{\mathcal{K}})} \\ \|u-u_h\|_{H^{1-m}_k(\Omega'_{\mathcal{V}})} \\ \|u-u_h\|_{H^{1-m}_k(\Omega'_{\mathcal{I}})} \\ \|u-u_h\|_{H^{1-m}_k(\Omega'_{\mathcal{P}})} \end{pmatrix} \le C\left[(\mathcal{H}k)^m + \mathscr{T}\mathcal{C}(\mathcal{H}k)^p + k^{-N}(hk)^m\mathscr{F}\right](\mathcal{H}k)^p \begin{pmatrix} \|u\|_{H^{p+1}_k(\Omega_{\mathcal{K}})} \\ \|u\|_{H^{p+1}_k(\Omega_{\mathcal{V}})} \\ \|u\|_{H^{p+1}_k(\Omega_{\mathcal{I}})} \\ \|u\|_{H^{p+1}_k(\Omega_{\mathcal{P}})} \end{pmatrix}. \quad (9.3)$$

When the data, f, is k-oscillatory in the sense of Lemma 6.25, so is the solution u, and in this case,

$$\|u\|_{H^{p+1}_k(U')} \le C\|u\|_{H^1_k(U)} \quad \text{for } U' \Subset U.$$

Hence one can use (9.3) to find triangulations with $(hk)^{2p}\rho(k) \gg 1$ that nevertheless have guaranteed control on the relative error.

Remark 9.5 (Improvements in Theorem 3.11 (Averseng *et al.* 2025)). Theorem 3.11 in Averseng *et al.* (2025) is stronger than Theorem 9.1 in that it considers arbitrary covers of Ω, and bounds the high ($\gg k$) and low ($\lesssim k$) frequencies of the Galerkin error separately. Two situations in which a more complicated cover is advantageous are the following. (i) There are two or more cavities that are dynamically separated, i.e. for which there is no billiard trajectory whose closure insects both cavities. (ii) We have *a priori* information about the data and/or solution and hence can obtain good control on the right-hand side of (9.2). Even when $\mathcal{K} = \emptyset$,

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

such information combined with Theorem 9.1 allows us to define triangulations with *a priori* improved accuracy in some regions, without the need to choose a small meshwidth everywhere.

9.1. Discussion of the ideas behind Theorem 9.1 and a sketch of the proof

9.1.1. The ideas behind Theorem 9.1

The following two important phenomena motivate Theorem 9.1.

(1) *The solution operator $\mathcal{R}$ encodes the billiard dynamics in Ω.* In particular, for $\chi_1, \chi_2 \in C^\infty(\Omega)$ the operator $\chi_1 \mathcal{R} \chi_2$ behaves differently depending on the locations of $\operatorname{supp} \chi_j$; for example, $\|\mathcal{R}\| \gg k$ when $\Omega_{\mathcal{K}} \neq \emptyset$, but if both χ_1 and χ_2 are away from $\Omega_{\mathcal{K}}$ then $\|\chi_1 \mathcal{R} \chi_2\|_{L^2 \to L^2} \lesssim k$.

(2) *The Galerkin error propagates.* The best possible situation would be local quasioptimality, that is, there exists $C > 0$ such that the Galerkin solution u_h satisfies, for every $U \subset \Omega$,

$$\|u - u_h\|_{H^1_k(U)} \leq C \inf_{w_h \in V^p_{\mathcal{T}}} \|u - w_h\|_{H^1_k(U)}. \tag{9.4}$$

In this case, since approximation of oscillatory functions by piecewise polynomials is well understood (see Galkowski 2025 for lower bounds and Appendix A for upper bounds), the properties of the data and behaviour of $\mathcal{R}$ would dictate the meshwidth in each region. Unfortunately (9.4) cannot hold for general triangulations. Indeed, suppose that (9.4) holds and let $\phi, \phi_1, \phi_2 \in C^\infty(\Omega)$ be such that $\operatorname{supp} \phi \subset \{\phi_1 \equiv 1\}$, $\phi \neq 0$ and $\phi_1 + \phi_2 \equiv 1$ on Ω. Then

$$\phi(u - u_h) = \sum_{j=1}^{2} \phi \mathcal{R} \phi_j P_\theta (u - u_h). \tag{9.5}$$

We now consider a situation where $\mathcal{T}$ has arbitrarily small elements on $\operatorname{supp} \phi_1 =: \Omega_1$, so that by (9.4),

$$\|u - u_h\|_{H^1_k(\Omega_1)} \leq C \inf_{w_h \in V^p_{\mathcal{T}}} \|u - w_h\|_{H^1_k(\Omega_1)} \ll 1.$$

In particular,

$$\|\phi(u - u_h)\|_{H^1_k} + \|\mathcal{R} \phi_1 P_\theta (u - u_h)\|_{H^1_k} \ll 1$$

(by continuity of P_θ and $\mathcal{R}$ and locality of P_θ). Then (9.5) implies that

$$\|\phi \mathcal{R} \phi_2 P_\theta (u - u_h)\|_{H^1_k} \ll 1,$$

which cannot be true unless the meshwidth is also sufficiently small on Ω_2 or $\phi \mathcal{R} \phi_2 \approx 0$. By item (1), the latter is not the case whenever $\operatorname{supp} \phi$ and $\operatorname{supp} \phi_2$ are connected by a billiard trajectory. For a striking illustration of this propagation of error, see Averseng *et al.* (2024, Figure 3).

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

This argument indicates not only that the Galerkin error propagates but also that the norm of the operator $\phi\mathcal{R}\phi_2$ determines the strength of propagation from supp ϕ_2 to supp ϕ.

Item (1) motivates varying the meshwidth from one location to another, but item (2) shows that, to be effective, this strategy must take into account the global behaviour of billiard trajectories. In particular, by item (2), the error in the cavity is *not* just dictated by the meshwidth in the cavity: the meshwidth also needs to be sufficiently small away from the cavity to control the propagating error.

9.1.2. Sketch proof of Theorem 9.1

For simplicity, we consider here the bound (9.2) with $m = p$ and ignore improvements that are possible in the overlaps between subdomains, in the PML region, and by splitting the frequencies of the Galerkin error into those $\gg k$ and $\lesssim k$.

The proof of Theorem 9.1 is, at heart, a localized version of the elliptic projection-type argument used to prove Theorem 6.21, and we first recap this argument. The key insight is the existence of a self-adjoint smoothing operator S_k so that $P_{S_k} := P_\theta + S_k$ is coercive (uniformly in k), and for all N there exists $C > 0$ such that for $k \geq k_0$,

$$\|S_k\|_{H_k^{-N}\to H_k^N} \leq C \tag{9.6}$$

(see Theorem 6.21 for the definition of the operator S_k). Since P_{S_k} is coercive, by Lemma 6.6 there is an *elliptic projection* $\Pi_{S_k} : H_k^1 \to \mathcal{P}_\mathcal{T}^p$ such that

$$\langle P_{S_k} w_h, (I - \Pi_{S_k})u\rangle = 0 \quad \text{for all } w_h \in \mathcal{P}_\mathcal{T}^p, \tag{9.7}$$

and there exists $C > 0$ such that for all $k > k_0$,

$$\|(I - \Pi_{S_k})v\|_{H_k^1} \leq C \inf_{w_h\in\mathcal{P}_\mathcal{T}^p} \|v - w_h\|_{H_k^1} \tag{9.8}$$

(i.e. Π_{S_k} is the adjoint Galerkin projection associated with P_{S_k}). Moreover, by an Aubin–Nitsche-type duality argument (see (6.19)),

$$\|(I - \Pi_{S_k})v\|_{(H_k^{p-1})^*} \leq C(hk)^p \inf_{w_h\in\mathcal{P}_\mathcal{T}^p} \|v - w_h\|_{H_k^1}. \tag{9.9}$$

By (9.7) and Galerkin orthogonality (6.1), for all $w_h \in \mathcal{P}_\mathcal{T}^p$, $v \in H_k^{p-1}$,

$$\begin{aligned}
\langle u - u_h, v\rangle &= \langle P_\theta(u - u_h), \mathcal{R}^* v\rangle \\
&= \langle P_\theta(u - u_h), (I - \Pi_{S_k})\mathcal{R}^* v\rangle \\
&= \langle P_{S_k}(u - u_h), (I - \Pi_{S_k})\mathcal{R}^* v\rangle - \langle S_k(u - u_h), (I - \Pi_{S_k})\mathcal{R}^* v\rangle \\
&= \langle P_{S_k}(u - w_h), (I - \Pi_{S_k})\mathcal{R}^* v\rangle - \langle S_k(u - u_h), (I - \Pi_{S_k})\mathcal{R}^* v\rangle.
\end{aligned} \tag{9.10}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

By (9.8), (9.9) and the mapping properties $S_k\colon \left(H_k^{p-1}\right)^* \to H_k^{p-1}$ and $P_{S_k}\colon H_k^1 \to (H_k^1)^*$,

$$\begin{aligned}&|\langle u-u_h, v\rangle|\\ &\quad\le C\eta_{H_k^{p-1}\to H_k^1}(\mathcal{P}_{\mathcal{T}}^p)\Big(\inf_{w_h\in\mathcal{P}_{\mathcal{T}}^p}\|u-w_h\|_{H_k^1}+(hk)^p\|u-u_h\|_{H_k^{-p+1}}\Big)\|v\|_{H_k^{p-1}},\end{aligned} \tag{9.11}$$

where $\eta_{H_k^{p-1}\to H_k^1}(\mathcal{P}_{\mathcal{T}}^p)$ is defined by (6.4). By duality, (9.11) implies

$$\begin{aligned}\|u-u_h\|_{(H_k^{p-1})^*} &\le C\Big(b\inf_{w_h\in\mathcal{P}_{\mathcal{T}}^p}\|u-w_h\|_{H_k^1}+\omega\|u-u_h\|_{(H_k^{p-1})^*}\Big),\\ b &:= \eta_{H_k^{p-1}\to H_k^1}(\mathcal{P}_{\mathcal{T}}^p),\quad \omega := (hk)^p\eta_{H_k^{p-1}\to H_k^1}(\mathcal{P}_{\mathcal{T}}^p).\end{aligned} \tag{9.12}$$

By Lemma 6.24 and the fact that $\rho(k)\ge ck$,

$$\eta_{H_k^{p-1}\to H_k^1}(\mathcal{P}_{\mathcal{T}}^p)\le C(hk)^p\left(1+\|\mathcal{R}\|_{L^2\to L^2}\right)=C(hk)^p(1+\rho(k))\le C(hk)^p\rho(k).$$

Thus, from (9.1.2), when $(hk)^{2p}\rho(k)$ is sufficiently small, $(1-C\omega)^{-1}$ exists and is positive, and then

$$\begin{aligned}\|u-u_h\|_{(H_k^{p-1})^*} &\le C(1-C\omega)^{-1}b\inf_{w_h\in\mathcal{P}_{\mathcal{T}}^p}\|u-w_h\|_{H_k^1}\\ &\le C(hk)^p\rho(k)\inf_{w_h\in\mathcal{P}_{\mathcal{T}}^p}\|u-w_h\|_{H_k^1},\end{aligned} \tag{9.13}$$

which is the preasymptotic estimate (6.43) for $n=p$.

We now sketch the localized version of the above argument, which is used to prove Theorem 9.1.

Pseudolocality for S and Π_{S_k}. By Lemma 7.4, for $\chi,\psi\in\overline{C^\infty}(\Omega)$ with $\operatorname{supp}\chi\cap\operatorname{supp}\psi=\emptyset$,

$$\|\chi S_k\psi\|_{(H_k^N)^*\to H_k^N}+\|\chi\mathcal{R}S\psi\|_{(H_k^N)^*\to H_k^N}=O(k^{-\infty}). \tag{9.14}$$

We also need pseudolocality of Π_{S_k}; that is, for $\chi,\psi\in\overline{C^\infty}(\Omega)$ with $\operatorname{supp}\chi\cap\operatorname{supp}\psi=\emptyset$,

$$\|\chi(I-\Pi_{S_k})\psi u\|_{H_k^1}\le Ck^{-N}\|(I-\Pi_{S_k})\psi u\|_{(H_k^p)^*}. \tag{9.15}$$

The proof of (9.15) is given in Averseng *et al.* (2025, §7) and we only review the steps of the argument here.

(1) Show that if $\chi_+\in\overline{C^\infty}(\Omega)$ with $\operatorname{supp}\chi\cap\operatorname{supp}(1-\chi_+)=\emptyset$ and $\operatorname{supp}\chi_+\cap\operatorname{supp}\psi=\emptyset$, then

$$\|\chi(I-\Pi_{S_k})\psi u\|_{H_k^1}\le Ck^{-N}\left(\|\chi_+(I-\Pi_{S_k})\psi u\|_{L^2}+\|(I-\Pi_{S_k})\psi u\|_{H_k^{-N}}\right)$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

(Averseng *et al.* 2025, Lemma 7.4). The proof of this fact follows by using coercivity of P_{S_k}, the definition of Π_{S_k}, superapproximation and inverse estimates in $\mathcal{P}^p_{\mathcal{T}}$, pseudolocality of S, and the boundedness of $S\colon \left(H^N_k\right)^* \to H^N_k$.

(2) Show that for $\chi, \psi \in \overline{C^\infty}(\Omega)$ with $\operatorname{supp}\chi \cap \operatorname{supp}\psi = \emptyset$,

$$\|\chi(I - \Pi_{S_k})\psi u\|_{L^2} \leq C\|(I - \Pi_{S_k})\psi u\|_{H^{-N}_k}$$

(Averseng *et al.* 2025, Lemma 7.5). The proof of this estimate uses a Schatz duality-type argument for P_{S_k} and a more sophisticated version of (9.15) that controls how $\|\chi(I - \Pi_{S_k})\psi u\|_{H^1_k}$ behaves when $k^{-1} \leq d(\operatorname{supp}\chi, \operatorname{supp}\psi) \ll 1$.

Sketch proof of Theorem 9.1. To localize the elliptic projection argument, we introduce an open cover $\{\Omega_j\}_{j=1}^M$ of Ω and a partition of unity $\{\phi_j\}_{j=1}^M \subset \overline{C^\infty}(\Omega)$ subordinate to this cover. In Theorem 9.1, $M = 4$ and $(\Omega_1, \Omega_2, \Omega_3, \Omega_4) := (\Omega_{\mathcal{K}}, \Omega_{\mathcal{V}}, \Omega_{\mathcal{I}}, \Omega_{\mathcal{P}})$.

Next, let $\chi_j \in \overline{C^\infty}(\Omega)$, $j = 1, \ldots, M$ such that

$$\operatorname{supp}\chi_j \subset \Omega_j \cup \partial\Omega, \quad \chi_j \equiv 1 \text{ in a neighbourhood of } \operatorname{supp}\phi_j.$$

Arguing as in (9.10), for all $w_{h,j} \in V_k$, $j = 1, \ldots, M$, and $v \in H^{p-1}_k$, we obtain

$$\begin{aligned}
&\langle \chi_i(u - u_h), v\rangle \\
&\quad = \langle P_\theta(u - u_h), \mathcal{R}^*\chi_i v\rangle \\
&\quad = \langle P_{S_k}(u - u_h), (I - \Pi_{S_k})\mathcal{R}^*\chi_i v\rangle - \langle S_k(u - u_h), (I - \Pi_{S_k})\mathcal{R}^*\chi_i v\rangle \\
&\quad = \sum_{j=1}^M \left(\langle P_{S_k}(u - w_{h,j}), (I - \Pi_{S_k})\phi_j\mathcal{R}^*\chi_i v\rangle - \langle S_k(u - u_h), (I - \Pi_{S_k})\phi_j\mathcal{R}^*\chi_i v\rangle\right) \\
&\quad = \sum_{j=1}^M \Big(\langle P_{S_k}\chi_j(u - w_{h,j}), \chi_j(I - \Pi_{S_k})\phi_j\mathcal{R}^*\chi_i v\rangle \\
&\qquad\quad - \langle S_k\chi_j(u - u_h), \chi_j(I - \Pi_{S_k})\phi_j\mathcal{R}^*\chi_i v\rangle \\
&\qquad\quad + Ck^{-N}(\|u - u_h\|_{(H^N_k)^*} + \|u - w_{h,j}\|_{H^1_k})\|(I - \Pi_{S_k})\phi_j\mathcal{R}^*\chi_i v\|_{(H^p_k)^*}\Big).
\end{aligned} \tag{9.16}$$

Next, we need a local analogue of (6.19). We claim that

$$\|\chi_j(I - \Pi_{S_k})v\|_{(H^{p-1}_k)^*} \leq C((h_j k)^p + k^{-N}(hk)^p) \inf_{w_h \in \mathcal{P}^p_{\mathcal{T}}} \|v - w_h\|_{H^1_k}, \tag{9.17}$$

where

$$h_j := \max_{\substack{K \in \mathcal{T} \\ K \cap \Omega_j \neq \emptyset}} h_K.$$

To prove (9.17), let $\chi_j^+ \in C^\infty_c(\Omega_j \cup \partial\Omega)$ with $\operatorname{supp}(1 - \chi_j^+) \cap \operatorname{supp}\chi_j = \emptyset$. Then, by (9.7), pseudolocality of $\mathcal{R}_S$ (9.14), the fixed-regularity polynomial-approximation

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

results from e.g. Theorem A.11, and (9.8),

$$\begin{aligned}
&|\langle w, \chi_j(I-\Pi_{S_k})v\rangle| \\
&\quad = |\langle P_{S_k}\mathcal{R}_{S_k}\chi_j w, (I-\Pi_{S_k})v\rangle| \\
&\quad \le |\langle P_{S_k}\chi_j^+\mathcal{R}_{S_k}w, (I-\Pi_{S_k})v\rangle| + |\langle P_{S_k}(1-\chi_j^+)\mathcal{R}_{S_k}\chi_j w, (I-\Pi_{S_k})v\rangle| \\
&\quad = |\langle P_{S_k}(\chi_j^+\mathcal{R}_{S_k}w - w_h), (I-\Pi_{S_k})v\rangle| \\
&\qquad + |\langle P_{S_k}((1-\chi_j^+)\mathcal{R}_{S_k}\chi_j w - w_{h,1}), (I-\Pi_{S_k})v\rangle| \\
&\quad \le C((h_jk)^p + k^{-N}(hk)^p)\|(I-\Pi_{S_k})v\|_{H_k^1}\|w\|_{H_k^{p-1}} \\
&\quad \le C((h_jk)^p + k^{-N}(hk)^p)\inf_{w_h\in\mathcal{P}_\mathcal{T}^p}\|v-w_h\|_{H_k^1}\|w\|_{H_k^{p-1}},
\end{aligned}$$

and hence (9.17) follows by duality.

By (9.16), (9.17) and the mapping properties $S_k\colon \left(H_k^{p-1}\right)^* \to H_k^{p-1}$ (9.6) and $P_{S_k}\colon H_k^1 \to (H_k^1)^*$,

$$\begin{aligned}
&|\langle \chi_i(u-u_h), v\rangle| \\
&\quad \le C\sum_j \eta_p(j\to i)\Big(\inf_{w_{h,j}\in\mathcal{P}_\mathcal{T}^p}\left(\|\chi_j(u-w_{h,j})\|_{H_k^1} + k^{-N}(hk)^p\|u-w_{h,j}\|_{H_k^1}\right) \\
&\qquad\qquad + ((h_jk)^p + k^{-N}(hk)^p)\|\chi_j(u-u_h)\|_{(H_k^{p-1})^*} \\
&\qquad\qquad + k^{-N}(hk)^p\|u-u_h\|_{(H_k^N)^*}\Big)\|v\|_{H_k^{p-1}},
\end{aligned} \tag{9.18}$$

where

$$\eta_p(j\to i) := \sup_{0\neq v\in H_k^{m-1}}\inf_{w_h\in\mathcal{P}_\mathcal{T}^p}\frac{\|\chi_j\mathcal{R}^*\chi_i v - w_h\|_{H_k^1}}{\|v\|_{H_k^{p-1}}}$$

(compare to (9.11)). By duality, (9.18) implies that

$$\begin{aligned}
&\|\chi_i(u-u_h)\|_{(H_k^{p-1})^*} \\
&\quad \le C\sum_j \eta_p(j\to i)\Big(\inf_{w_{h,j}\in\mathcal{P}_\mathcal{T}^p}\left(\|\chi_j(u-w_{h,j})\|_{H_k^1} + k^{-N}(hk)^p\|u-w_{h,j}\|_{H_k^1}\right) \\
&\qquad\qquad + \left((h_jk)^p + k^{-N}(hk)^p\right)\|\chi_j(u-u_h)\|_{(H_k^{p-1})^*} \\
&\qquad\qquad + k^{-N}(hk)^p\|u-u_h\|_{(H_k^N)^*}\Big).
\end{aligned}$$

We then use a local version of the frequency-splitting argument used to prove Lemma 6.24 (see Averseng *et al.* 2025, Lemma 8.5), to obtain

$$\eta_p(j\to i) \le C(h_jk)^p\|\chi_j\mathcal{R}^*\chi_i\|_{L^2\to L^2} + C1_{\{\Omega_i\cap\Omega_j\neq\emptyset\}}(h_{ij}k)^p,$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Table 9.1. Bounds on $\|\psi\mathcal{R}^*\chi\|_{L^2\to L^2}$ (up to k-independent constants) proved in Averseng *et al.* (2025, §4) for $k \notin \mathcal{J}$. For the scattering problem, i.e. without PML truncation, the upper bounds by $\sqrt{k\rho(k)}$ are given by Theorem 5.23, and the upper bounds by k are given by Theorem 5.24.

$\operatorname{supp}\psi \setminus \operatorname{supp}\chi$	$\Omega_{\mathcal{K}}$	$\Omega_{\mathcal{V}}$	$\Omega_{\mathcal{I}}$	$\Omega_{\mathcal{P}}$
$\Omega_{\mathcal{K}}$	$\rho(k)$	$\sqrt{k\rho(k)}$	$O(k^{-\infty})$	$O(k^{-\infty})$
$\Omega_{\mathcal{V}}$	$\sqrt{k\rho(k)}$	k	k	1
$\Omega_{\mathcal{I}}$	$O(k^{-\infty})$	k	k	1
$\Omega_{\mathcal{P}}$	$O(k^{-\infty})$	1	1	1

where $h_{ij} := \min(h_i, h_j)$. In particular, this yields the system of inequalities

$$\begin{aligned}
&\big(\|\chi_i(u-u_h)\|_{(H_k^{p-1})^*}\big)_{i=1}^M \\
&\quad \le CB\Big(\inf_{w_{h,j}\in\mathcal{P}_{\mathcal{T}}^p}\big(\|\chi_j(u-w_{h,j})\|_{H_k^1} + k^{-N}(hk)^p\|u-w_{h,j}\|_{H_k^1}\big)\Big)_{j=1}^M \\
&\qquad + CW\big(\|\chi_j(u-u_h)\|_{(H_k^{p-1})^*}\big)_{j=1}^M,
\end{aligned} \tag{9.19}$$

where

$$\begin{aligned}
B_{ij} &:= \eta_p(j\to i) \le C(h_jk)^p\|\chi_j\mathcal{R}^*\chi_i\|_{L^2\to L^2} + C1_{\{\Omega_i\cap\Omega_j\neq\emptyset\}}(h_{ij}k)^p, \\
W_{ij} &:= (h_jk)^p\eta_p(j\to i) + k^{-N}(hk)^p\sum_{\ell}\eta_p(\ell\to i) \\
&\le C(h_jk)^{2p}\|\chi_j\mathcal{R}^*\chi_i\|_{L^2\to L^2} + C(h_jk)^p1_{\{\Omega_i\cap\Omega_j\neq\emptyset\}}(h_{ij}k)^p + Ck^{-N}(hk)^{2p}
\end{aligned}$$

(compare to (9.1.2)). If

$$\sum_{n=0}^{\infty}(CW)^n < \infty, \tag{9.20}$$

then $(I-CW)^{-1}$ exists and has non-negative entries. Hence (9.19) implies that

$$\|\underline{u-u_h}\|_{H_k^{-p+1}} \le C(I-CW)^{-1}B\|\underline{u-w_h}\|_{H_k^1}$$

(compare to (9.13)).

To understand when $\sum_n(CW)^n$ converges, consider W as the weighted adjacency matrix of a directed graph with M nodes representing $\{\Omega\}_{j=1}^M$. Observe that the entry in the ith row and jth column of $(CW)^\ell$ is given by C^ℓ times the sum of the weights over all paths of length ℓ from j to i in this graph. Hence the sum converges if, for any i and j, the sum of the weights of all paths from j to i multiplied by $C^{\text{path length}}$ is finite. Using elementary graph analysis, this condition can be reduced

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

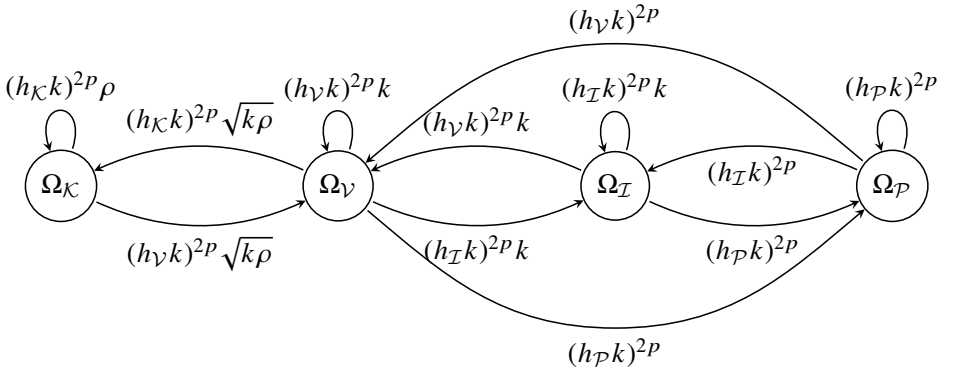


Figure 9.1. Graph showing propagation of errors for the decomposition into $\Omega_{\mathcal{K}}$, $\Omega_{\mathcal{V}}$, $\Omega_{\mathcal{I}}$ and $\Omega_{\mathcal{P}}$ in the simplified set-up of Section 9.1.2. Note that this can be improved in several ways using the analysis in Averseng *et al.* (2025, §8). The graph corresponding to Theorem 9.1 is shown in Figure 9.2.

to the requirement that all the sum of such weights for all non-self-intersecting loops is less than 1 (see Averseng *et al.* 2025, Appendix B).

In the setting of Theorem 9.1, $M = 4$ and $(\Omega_1, \Omega_2, \Omega_3, \Omega_4) \coloneqq (\Omega_{\mathcal{K}}, \Omega_{\mathcal{V}}, \Omega_{\mathcal{I}}, \Omega_{\mathcal{P}})$. For $k \notin \mathcal{J}$, Averseng *et al.* (2025, §4, Appendix C) obtain the bounds of Table 9.1 on $\psi\mathcal{R}^*\chi$ according to the support of ψ and χ. As a result, the graph corresponding to W is the one in Figure 9.1, and the requirement that the sum of weights on all non-self-intersecting loops be less than 1 reduces to (9.1).

To obtain Theorem 9.1, two further improvements to the analysis described above are necessary.

(i) One must treat the PML more carefully to see that, effectively, nothing propagates in or out of the PML (see the differences between Figures 9.1 and 9.2).

(ii) The analysis given above obtains estimates in the lowest available norm, $\left(H_k^{p-1}\right)^*$, whereas Theorem 9.1 obtains estimates in all norms up to H_k^1. Unfortunately, changing the space in which we take v in (9.18) is not quite sufficient. Instead, we separate out the high frequencies of $u - u_h$ on each domain and treat them as an additional part of the decomposition of u.

9.1.3. Interpretation as error propagation

To properly interpret the matrices appearing in (9.2), we return to (9.16), which (modulo $O(k^{-\infty})$ errors) is equivalent to

$$\chi_i(u-u_h) = \sum_{j=1}^{M} \chi_i R_k \phi_j \left((I-\Pi_{S_k})^* \chi_j P_{S_k} \chi_j (u-w_{h,j}) - (I-\Pi_{S_k})^* \chi_j S_k \chi_j (u-u_h)\right). \tag{9.21}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

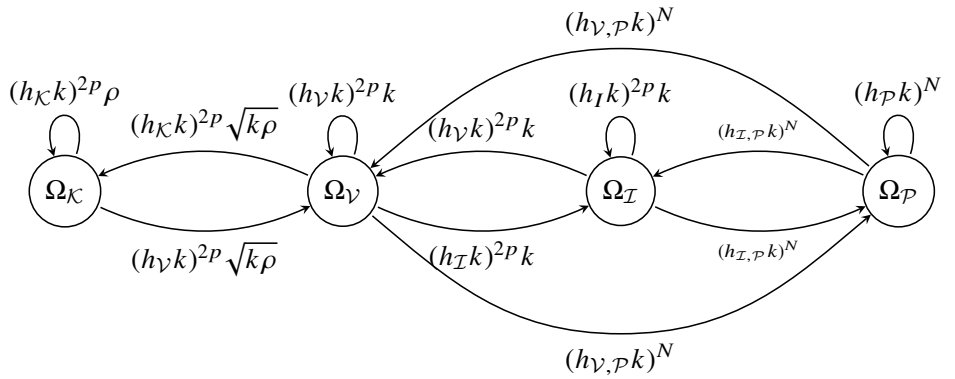


Figure 9.2. Graph showing propagation of errors for the decomposition into $\Omega_{\mathcal{K}}$, $\Omega_{\mathcal{V}}$, $\Omega_{\mathcal{I}}$. Recall that $h_{\mathcal{V},\mathcal{P}} = \min(h_{\mathcal{V}}, h_{\mathcal{P}})$ and $h_{\mathcal{I},\mathcal{P}} = \min(h_{\mathcal{I}}, h_{\mathcal{P}})$.

We are interested in $\|\chi_i(u-u_h)\|_{(H_k^{p-1})^*}$, which we think of as the low frequencies of $\chi_i(u-u_h)$; these low frequencies are captured by $S_k\chi_i(u-u_h)$. For the purposes of this discussion, we assume S_k commutes with $\chi_i R_k\phi_j$. This is not quite true, but (away from the PML), since

$$\|S_k\chi_i R_k\phi_j\|_{(H_k^{p-1})^*\to L^2} \le C\|\chi_i R_k\phi_j\|_{L^2\to L^2},$$

$S_k\chi_i R_k\phi_j$ acts like $\chi_i R_k\phi_j\mathcal{L}$, where $\mathcal{L}$ is a low-pass filter. Averseng *et al.* (2025, Theorem 4.2) use semiclassical ellipticity of the operator in the PML to show that near the PML there is no propagation; we therefore ignore the PML here.

With these caveats, (9.21) implies

$$\begin{aligned} S_k\chi_i(u-u_h) = \sum_{j=1}^{M} \chi_i R_k\phi_j \big(& S_k\chi_j(I-\Pi_{S_k})^*\chi_j P_{S_k}\chi_j(u-w_{h,j}) \\ & - S_k\chi_j(I-\Pi_{S_k})^*\chi_j\tilde{S}_k S_k\chi_j(u-u_h)\big). \end{aligned}$$

The operator $\chi_i R_k\phi_j$ has the effect of propagating between domains. The operator $(I-\Pi_{S_k})^*$ essentially takes the best approximation in the H_k^1-norm and S_k then returns only the frequency $\lesssim k$ components. This process is represented in the graph in Figure 9.3.

To find $S_k\chi_i(u-u_h)$ in terms of the local best approximations to u, one inserts $\chi_j P_{S_k}\chi_j(u-w_{h,j})$ at node 1 in Figure 9.3 and follows the cycle to node 4, producing the first approximation to $S_k\chi_i(u-u_h)$. One then continues around the cycle arbitrarily many times, adding $\chi_j P_{S_k}\chi_j(u-w_{h,j})$ in each cycle. This process converges under the condition (9.20) and the final result at node 4 is $(S_k\chi_i(u-u_h))_i$. The W and B matrices in (9.19) are respectively one full cycle from node 4 to node 4 and a path from node 1 to node 4 in Figure 9.3.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

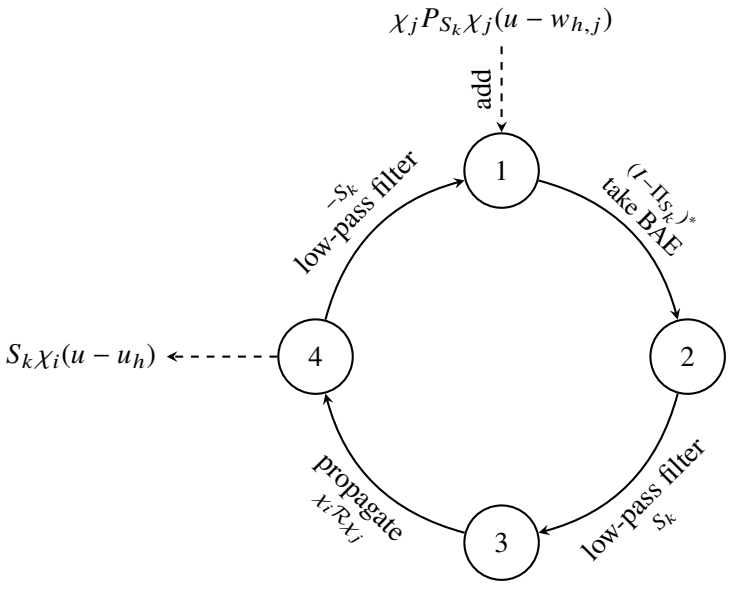


Figure 9.3. Graph showing the process of error propagation when determining the low frequencies of $u - u_h$ from the local best approximation errors. The arrows are labelled first with the type of operation (low-pass filter etc.) and then with the operator whose action gives this effect. These operators are applied multiplicatively.

10. The parallel overlapping Schwarz method with PMLs on the subdomain boundaries (R5)

10.1. Definition of the Helmholtz Cartesian PML problem and the parallel overlapping Schwarz method

Definition of the subdomains. Let the d-dimensional hyper-rectangular domain Ω_{int} be given by the Cartesian product

$$\Omega_{\text{int}} := \prod_{1 \le \ell \le d} (0, L_\ell).$$

Let $\{\Omega_{\text{int},j}\}_{j=1}^N$ be an overlapping decomposition of Ω_{int} in hyper-rectangular subdomains:

$$\Omega_{\text{int},j} := \prod_{1 \le \ell \le d} \left(a_\ell^j, b_\ell^j\right). \tag{10.1}$$

We extend Ω_{int} and each $\Omega_{\text{int},j}$, $1 \le j \le N$, by adding a PML layer to each, to form the domains Ω and Ω_j. Namely, let $\kappa > 0$ and $\kappa_0 > 0$ denote respectively the PML width on Ω and the interior PML width, and let

$$\Omega := \prod_{1 \le \ell \le d} (-\kappa, L_\ell + \kappa), \quad \Omega_j := \prod_{1 \le \ell \le d} \left(a_\ell^j - \kappa_\ell^j, b_\ell^j + \kappa_\ell^j\right),$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

where $\kappa_\ell^j = \kappa_0$ if the corresponding edge of $\Omega_{\mathrm{int},j}$ belongs to the interior of Ω, and κ otherwise; that is, edges of subdomains that touch the boundary have the same PML width as Ω (i.e. κ), and interior subdomain edges have PML width κ_0, which can be different to κ.

The scaled operators. We now define standard Cartesian PMLs of width κ at the boundary of Ω and width κ_0 at the boundary of each Ω_j. For simplicity, we assume that the same PML scaling function is used inside every PML; however, this assumption can be easily removed, with, say, one function used in the PMLs of width κ and another used in the PMLs of width κ_0, at the cost of introducing more notation.

Let $f_{\mathrm{s}} \in C^\infty(\mathbb{R})$ (with subscript s standing for 'scaling') be such that

$$\{x\colon f_{\mathrm{s}}(x) = 0\} = \{x\colon f_{\mathrm{s}}'(x) = 0\} = \{x : x \le 0\},$$
$$f_{\mathrm{s}}'(x) > 0 \text{ for } x > 0, \text{ and } f_{\mathrm{s}}''(x) = 0 \text{ for } x \ge \kappa_{\mathrm{lin}}$$

for some $\kappa_{\mathrm{lin}} < \kappa$ (observe that f_{s} is then linear for $x \ge \kappa_{\mathrm{lin}}$). This assumption is to avoid technical issues about propagation of singularities; see Galkowski *et al.* (2024*a*, Remark 2.5).

For any $1 \le \ell \le d$, we define the following scaling functions in the ℓ-direction $g_\ell \in C^\infty(\mathbb{R}^d)$ by

$$g_\ell(x_\ell) := \begin{cases} f_{\mathrm{s}}(x_\ell - L_\ell) & \text{if } x_\ell \ge L_\ell, \\ 0 & \text{if } x_\ell \in (0, L_\ell), \\ -f_{\mathrm{s}}(-x_\ell) & \text{if } x_\ell \le 0, \end{cases}$$

and, similarly, the subdomain scaling functions in the ℓ-direction $g_{\ell,j} \in C^\infty(\mathbb{R}^d)$ for $1 \le j \le N$ by

$$g_{\ell,j}(x_\ell) := \begin{cases} f_{\mathrm{s}}\big(x_\ell - b_\ell^j\big) & \text{if } x_\ell \ge b_\ell^j, \\ 0 & \text{if } x_\ell \in \big(a_\ell^j, b_\ell^j\big), \\ -f_{\mathrm{s}}\big(a_\ell^j - x_\ell\big) & \text{if } x_\ell \le a_\ell^j. \end{cases}$$

We now define the scaled operators Δ_{s} and $\Delta_{\mathrm{s},j}$ by

$$\Delta_{\mathrm{s}} := \sum_{\ell=1}^d \left(\frac{1}{1 + i g_\ell'(x_\ell)} \partial_{x_\ell} \right)^2$$

and

$$\Delta_{\mathrm{s},j} := \sum_{\ell=1}^d \left(\frac{1}{1 + i g_{\ell,j}'(x_\ell)} \partial_{x_\ell} \right)^2 \quad \text{for all } 1 \le j \le N. \tag{10.2}$$

Given $n \in C^\infty(\mathbb{R}^d)$ that is strictly positive and such that $\mathrm{supp}(1-n) \subset \Omega_{\mathrm{int}}$, let

$$P_{\mathrm{s}} := -k^{-2}\Delta_{\mathrm{s}} - n, \quad P_{\mathrm{s}}^j := -k^{-2}\Delta_{\mathrm{s},j} - n \quad \text{for all } 1 \le j \le N. \tag{10.3}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

These operators are defined on $H^1(\mathbb{R}^d)$, but we consider P_{s} as a operator $H^1(\Omega) \to H^{-1}(\Omega)(= (H^1_0(\Omega))^*)$, and P^j_{s} as an operator $H^1(\Omega_j) \to H^{-1}(\Omega_j)(= (H^1_0(\Omega_j))^*)$. If Dirichlet boundary data in $H^{1/2}$ is prescribed on the corresponding boundaries, these operators are then invertible.

In the proofs of our main results, a key region is the following subset of Ω_j:

$$\operatorname{supp}\left(P^j_{\mathrm{s}} - P_{\mathrm{s}}\right) := \Omega_j \cap \left(\bigcup_{\ell=1}^{d} \overline{\{x \in \mathbb{R}^d : g_{\ell,j}(x_\ell) \neq g_\ell(x_\ell)\}}\right);$$

$\operatorname{supp}(P^j_{\mathrm{s}} - P_{\mathrm{s}})$ is the part of the PML region of Ω_j where the PML in Ω_j differs from the PML in Ω (i.e. where $P^j_{\mathrm{s}} \neq P_{\mathrm{s}}$). Note that when Ω_j is an interior subdomain, $\operatorname{supp}(P^j_{\mathrm{s}} - P_{\mathrm{s}}) = \Omega_j \setminus \Omega_{\mathrm{int,j}}$.

The Helmholtz problem. Given $f \in H^{-1}(\Omega)$, find $u \in H^1_0(\Omega)$ such that

$$P_{\mathrm{s}} u = f. \tag{10.4}$$

Theorem 4.3 above proved that the solution of the radial PML exists, is unique, and is exponentially accurate for sufficiently large k. An analogous result for P non-trapping is proved in Galkowski *et al.* (2024*a*, Lemma 4.5 and Theorem 4.6), except with exponential accuracy replaced by superalgebraic accuracy (i.e. $O(k^{-\infty})$).

The existence and uniqueness results for Cartesian PML at fixed k analogous to those for radial PMLs discussed in Remark 4.5 are Theorem 5.5 of Kim and Pasciak (2010) and Theorem 5.7 of Bramble and Pasciak (2013).

The partition of unity. Let $\{\chi_j\}_{j=1}^N$ be a partition of unity subordinate to

$$\left\{\Omega_j \setminus \operatorname{supp}(P^j_{\mathrm{s}} - P_{\mathrm{s}})\right\}_{j=1}^N,$$

that is, $\{\chi_j\}_{j=1}^N$ is a family of non-negative elements of $C^\infty(\mathbb{R}^d)$ such that for all $x \in \mathbb{R}^d$,

$$\sum_{j=1}^{N} \chi_j(x) = 1, \quad \operatorname{supp} \chi_j \cap \Omega \subset \left(\Omega_j \setminus \operatorname{supp}(P^j_{\mathrm{s}} - P_{\mathrm{s}})\right) \quad \text{for all } j. \tag{10.5}$$

Additionally, in a neighbourhood of $\partial\Omega$, each χ_j does not vary in the normal direction to the boundary of Ω (this last assumption is for technical reasons, and can easily be achieved in practice).

The parallel overlapping Schwarz method. Given $u^n \in H^1_0(\Omega)$, for $n \geq 0$ and any $1 \leq j \leq N$, let $u^{n+1}_j \in H^1(\Omega_j)$ be the solution to

$$\begin{cases} P^j_{\mathrm{s}} u^{n+1}_j = P^j_{\mathrm{s}}(u^n|_{\Omega_j}) - (P_{\mathrm{s}} u^n)|_{\Omega_j} + f|_{\Omega_j} \in H^{-1}(\Omega_j), \\ u^{n+1}_j = u^n \in H^{1/2}(\partial\Omega_j), \end{cases} \tag{10.6}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

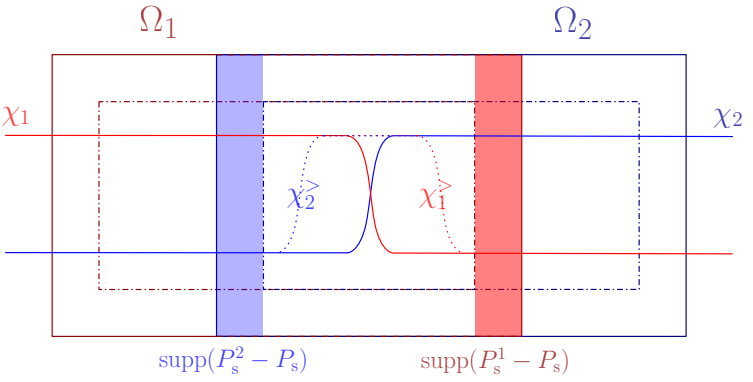


Figure 10.1. The functions $\{\chi_j^>\}$ and $\{\chi_j\}$ for two subdomains.

and then set

$$u^{n+1} := \sum_{j=1}^{N} \chi_j u_j^{n+1}; \tag{10.7}$$

this method is equivalent to (3.21) and its discrete-level analogue can be understood as a natural PML-variant of the well-known restricted additive Schwarz (RAS) method; see Galkowski *et al.* (2024*a*, §8.1) and Dolean *et al.* (2015, Lemma 1.6).

10.2. Definition of $\mathcal{N}$

Recall that φ_t is the Hamiltonian flow (see (5.18)).

Definition 10.1. Given an interval, I, and $(x,\xi) \in \mathbb{R}^{2d}$ such that $p(x,\xi) = 0$, define

$$\gamma_I(x,\xi) := \{\varphi_t(x,\xi),\ t \in I\}. \tag{10.8}$$

When $I = \mathbb{R}$, we say γ_I is a trajectory for P and usually denote it by γ with no subscript.

We now define a set of cut-offs that are bigger than the χ_j. Let $\{\chi_j^>\}_{j=1}^N$ be such that for each j, $\chi_j^> \in C^\infty(\mathbb{R}^d)$, $\mathrm{supp}\,\chi_j^> \cap \Omega \subset \Omega_j$, $\mathrm{supp}(1-\chi_j^>) \cap \mathrm{supp}\,\chi_j = \emptyset$, and $\mathrm{supp}\,\chi_j^> \cap \mathrm{supp}(P_\mathrm{s}^j - P_\mathrm{s}) = \emptyset$; for an illustration of $\{\chi_j^>\}$ and $\{\chi_j\}$ in the case of two subdomains, see Figure 10.1.

Let $\mathcal{W}$ be the set of finite sequences of elements of $\{1,\dots,N\}$ (i.e. the indices of subdomains) such that for any $w = \{w_i\}_{1\le i\le n} \in \mathcal{W}$, $w_i \neq w_{i+1}$ for all $1 \le i \le n$ (i.e. the $(i+1)$th component of w corresponds to a different subdomain to the ith component). We call the elements of $\mathcal{W}$ *words*.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

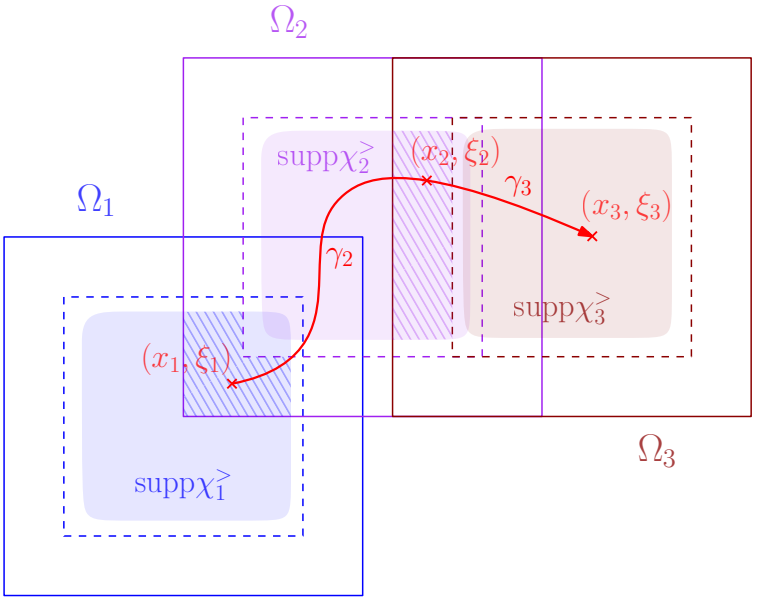


Figure 10.2. A trajectory (in red) following the word $(1,2,3)$ (here $\Omega_1, \Omega_2, \Omega_3$ are interior subdomains of Ω). The blue hatched shading indicates the domain $\operatorname{supp}\chi_1^> \cap \operatorname{supp}(P_s^2 - P_s)$, and the purple hatched shading indicates the domain $\operatorname{supp}\chi_2^> \cap \operatorname{supp}(P_s^3 - P_s)$. (The points (x_j,ξ_j), $j = 1,2,3$, and sub-trajectories γ_2, γ_3 are used in the precise definition of *follow* in Definition 10.2 below.)

Definition 10.2. A trajectory γ for P *follows* a word $w \in \mathcal{W}$ of length $n \geq 2$ if

$$\gamma = \prod_{2 \leq j \leq n} \gamma_j,$$

where the product stands for concatenation, $\gamma_j = \gamma_{(0,T_j]}(x_{j-1},\xi_{j-1})$ for some $T_j > 0$ and $(x_j,\xi_j) \in \mathbb{R}^{2d}$ with the following conditions.

(i) For $1 \leq j \leq n-1$, $x_j \in \operatorname{supp}\chi_{w_j}^> \cap \operatorname{supp}\left(P_s^{w_{j+1}} - P_s\right)$.

(ii) For $2 \leq j \leq n$, define $(x_j(t),\xi_j(t)) \coloneqq \gamma_j(t)$. Then $x_j(t) \in \overline{\Omega_{w_j}}$ for $t \in (0,T_j]$.

(iii) If $(x_n,\xi_n) \coloneqq \varphi_{T_n}(x_{n-1},\xi_{n-1})$, then $x_n \in \operatorname{supp}\chi_{w_n}^>$.

Figure 10.2 illustrates a simple example of a trajectory following a word.

The flow associated with P only allows passage between certain sequences of subdomains, that is, there are words that will not be followed by any trajectory. For example, if $n \equiv 1$, the trajectories of the flow are straight lines, and thus with domains in Figure 10.2 there is no trajectory following the word $(1,2,1)$. This motivates the following definition.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Definition 10.3. A word $w \in \mathcal{W}$ is *allowed* if there exists a trajectory for P that follows w.

Definition 10.4.

$$\mathcal{N} := \sup_{w \text{ allowed}} |w|. \tag{10.9}$$

Since the sets

$$\operatorname{supp} \chi^{>}_{w_1} \cap \operatorname{supp}(P_s^{w_2} - P_s), \ \ldots, \ \operatorname{supp} \chi^{>}_{w_{n-1}} \cap \operatorname{supp}(P_s^{w_n} - P_s), \ \operatorname{supp} \chi^{>}_{w_n} \tag{10.10}$$

are compact and disjoint, the assumption that P is non-trapping implies that the length of allowed words is bounded above, i.e. $\mathcal{N} < \infty$.

10.3. Error estimates for the parallel overlapping Schwarz method

Theorem 10.5 (Rigorous statement of R5). Assume that P is non-trapping. Then there exists $k_0 > 0$ such that for all $M, s > 0$, there exists $C > 0$ such that the following is true for any $f \in H^{-1}(\Omega)$ and $k \geq k_0$. If u is the solution to (10.4), and $\{u^n\}_{n\geq 0}$ is a sequence of iterates for the parallel overlapping Schwarz method (defined by (10.6) and (10.7)), then

$$\|u - u^{\mathcal{N}}\|_{H^s_k(\Omega)} \leq C k^{-M} \|u - u^0\|_{H^1_k(\Omega)}.$$

Remark 10.6. The requirement that k_0 be sufficiently large in Theorem 10.5 is used to guarantee that the solution to all subdomain problems exists. Alternatively, we could allow arbitrary $k_0 > 0$ provided that we exclude an open neighbourhood of the discrete set of k where P_s^j may not be invertible.

10.4. Concrete values of $\mathcal{N}$ for strips and checkerboards when $n \equiv 1$

For $N_\ell \in \mathbb{Z}^+$, $\ell = 1, \ldots, d$, we say that $\{\Omega_j\}_{j=1}^N$ is a *checkerboard* of size $N_1 \times \ldots \times N_\ell$ if the following hold.

(i) For $\ell = 1, \ldots, d$, there exist

$$0 = y^\ell_0 < \ldots < y^\ell_{N_\ell} = L_\ell$$

such that

$$\overline{\Omega_{\text{int}}} = \bigcup_{j=1}^N \overline{U_j}, \quad U_j \cap U_i = \emptyset \quad \text{if } j \neq i,$$

with each U_j of the form $\prod_{\ell=1}^d \left(y^\ell_{m_\ell - 1}, y^\ell_{m_\ell}\right)$ for some $m_\ell \in \{1, \ldots, N_\ell\}$, $\ell = 1, \ldots, d$.

(ii) The overlapping decomposition $\{\Omega_{\text{int},j}\}_{j=1}^N$ of Ω_{int} comes from extending each U_j in each coordinate direction (apart from at $\partial\Omega$).

(iii) The extensions in (ii) are such that for all j, Ω_j overlaps only with $\Omega_{j'}$, where Ω_j and $\Omega_{j'}$ are extensions of adjacent non-overlapping subdomains.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

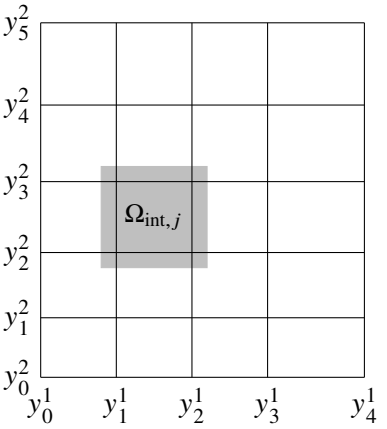


Figure 10.3. Grid formed by the points $\{y_m^\ell\}$ that is used in the definition of a particular 2-checkerboard of size 4×5, and the subdomain $\Omega_{\mathrm{int},j}$ formed by extending $\left(y_1^1, y_2^1\right) \times \left(y_2^2, y_3^2\right)$

We say that $\{\Omega_j\}$ is a $\mathfrak{d}$-checkerboard if $\mathfrak{d} := |\{\ell,\ N_\ell \neq 1\}|$, that is, $\mathfrak{d}$ is the effective dimensionality of the checkerboard.

We say that $\{\Omega_j\}_{j=1}^N$ is a *strip* if it is a 1-checkerboard. To simplify notation, we then assume that the subdomains of a strip are ordered monotonically (i.e. so that Ω_i only overlaps Ω_{i-1} and Ω_{i+1}).

Figure 10.3 shows an example of the grid formed by the points $\{y_m^\ell\}$ that is used in the definition of a particular 2-checkerboard of size 4×5, along with a particular subdomain $\Omega_{\mathrm{int},j}$.

Lemma 10.7 ($\mathcal{N}$ for strips and checkerboards). Suppose that $n \equiv 1$. If $\{\Omega_j\}$ is a $\mathfrak{d}$-checkerboard, then $\mathcal{N} = N_1 + \cdots + N_{\mathfrak{d}} - (\mathfrak{d} - 1)$. In particular, if $\{\Omega_j\}$ is a strip, then $\mathcal{N} = N$.

Lemma 10.7 is proved in Galkowski *et al.* (2024*a*, §7.3).

10.5. Sketch proof of Theorem 10.5

The proof of Theorem 10.5 boils down to understanding how the error $u - u^n$ at one step influences the error $u - u^{n+1}$ at the next step. The two main steps in the proof of Theorem 10.5 are the following.

(1) Write down the Helmholtz problems satisfied by the errors on each subdomain, and hence obtain *the physical error propagation matrix* $\mathbf{T}$ which maps the vector of errors on each subdomain at step n to the same at step $n + 1$ (see (10.17) below).

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

(2) Write $\mathbf{T}^n$ in terms of the operator products

$$\mathscr{T}_w := (\mathbf{T})_{w_n,w_{n-1}}(\mathbf{T})_{w_{n-1},w_{n-2}}\cdots(\mathbf{T})_{w_2,w_1} \tag{10.11}$$

(corresponding to the error propagating from Ω_{w_i} to $\Omega_{w_{i+1}}$ for $i = 1,\dots,n-1$) and show using SCA that $\|\mathscr{T}_w\|_{H^1_k(\Omega)\to H^s_k(\Omega)} = O(k^{-\infty})$ if w is not allowed.

We note that $\mathcal{N}$ appears because no words of length $> \mathcal{N}$ are allowed.

(1) *The error propagation matrix* $\mathbf{T}$ *for the parallel method.* For any $n \geq 1$ and $1 \leq j \leq N$, we define the global and local errors

$$e^n := u - u^n \in H^1_0(\Omega) \quad \text{and} \quad e^n_j := u|_{\Omega_j} - u^n_j \in H^1(\Omega_j). \tag{10.12}$$

These definitions, the definition of the iterate u^n (10.7), and the fact that $\{\chi_j\}_{j=1}^N$ is a partition of unity (10.5) imply that

$$e^n = \sum_{j=1}^N \chi_j e^n_j. \tag{10.13}$$

We introduce the notation

$$\epsilon^n_j := \chi_j e^n_j \quad \text{so that} \quad e^n = \sum_{j=1}^N \epsilon^n_j. \tag{10.14}$$

We can interpret $\epsilon^n_j = \chi_j e^n_j$, after extension by zero from Ω_j, as an element of $H^1_0(\Omega)$, since e^n_j is zero on $\partial\Omega_j \cap \partial\Omega$ and $\chi_j = 0$ on $\partial\Omega_j \setminus \partial\Omega$. We call ϵ^n_j the *local physical error*. We highlight that the fact that e^n is expressed in (10.14) in terms of $\epsilon^n_j = \chi_j e^n_j$ and not the whole local error e^n_j as expected. This is because e^n_j contains contributions from the PML layers of Ω_j that are not PML layers of Ω, and hence not part of the original PDE problem (10.4), and hence 'unphysical'.

For $1 \leq i \leq N$ and $v \in H^1_0(\Omega)$, let $\mathcal{T}_i v \in H^1(\Omega_j)$ be the solution to

$$\begin{cases} P^i_{\mathrm{s}}\mathcal{T}_i v = P^i_{\mathrm{s}}(v|_{\Omega_i}) - (P_{\mathrm{s}}v)|_{\Omega_i} & \text{in } \Omega_i, \\ \mathcal{T}_i v = v & \text{on } \partial\Omega_i. \end{cases} \tag{10.15}$$

Recall that $\{\chi^>_j\}_{j=1}^N$ are a set of cut-off functions that have bigger support than the χ_j.

We define the *physical error propagation matrix* $\mathbf{T}$ as

$$\mathbf{T} := (\chi_i \mathcal{T}_i \chi^>_j)_{1\leq i,j\leq N} : (H^1_0(\Omega))^N \mapsto (H^1(\Omega))^N \tag{10.16}$$

and, for any $n \geq 0$, the physical errors vector $\boldsymbol{\epsilon}^n \in (H^1_0(\Omega))^N$ as

$$(\boldsymbol{\epsilon}^n)_j := \epsilon^n_j, \quad 1 \leq j \leq N.$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Lemma 10.8 (Physical error propagation for the parallel method). The operator $\mathbf{T}$ defined in (10.16) satisfies $\mathbf{T}\colon (H_0^1(\Omega))^N \mapsto (H_0^1(\Omega))^N$ and

$$\boldsymbol{\epsilon}^{n+1} = \mathbf{T}\boldsymbol{\epsilon}^n \quad \text{for all } n \geq 0. \tag{10.17}$$

Lemma 10.8 is proved in Galkowski *et al.* (2024*a*, §5.1).

(2) *Relating powers of* $\mathbf{T}$ *to trajectories of the flow defined by* P. By (10.17), $\boldsymbol{\epsilon}^n = \mathbf{T}^n\boldsymbol{\epsilon}^0$. By the definition of $\mathbf{T}$ (10.16), powers of $\mathbf{T}$ involve compositions of the maps $\chi_i\mathcal{T}_i\chi_j^>$, and these compositions applied to ϵ^0 can be interpreted as the error travelling from subdomain to subdomain through the iterations.

We make this more precise via the definition of the composite map $\mathscr{T}_w$ (10.11), defined for any $w \in \mathcal{W}$ of length $|w| := n \geq 2$. The presence of the cut-offs in the definition of $(\mathbf{T})_{i,j}$ (10.16) means that $\mathscr{T}_w$ is zero unless Ω_{w_j} and $\Omega_{w_{j+1}}$ overlap for all j.

We make the following remarks.

- To see where the sets in (10.10) (i.e. the sets in item (i) of Definition 10.2) come from, observe that by the definition of $\mathcal{T}_i$ (10.15), the functions $\mathcal{T}_i\chi_j^> v$ appearing in the action of $\mathbf{T}$ (10.16) satisfy a PDE on Ω_i with the operator P_{s}^i and right-hand side supported in $\operatorname{supp}\chi_j^> \cap \operatorname{supp}(P_{\mathrm{s}}^i - P_{\mathrm{s}})$.
- $\operatorname{supp}\chi_{w_i} \cap \operatorname{supp}(P_s^{w_i+1} - P_s)$ is the part of the PML of $\Omega_{w_{i+1}}$ that intersects $\operatorname{supp}\chi_{w_i}$; because χ_j is zero on the support of $P_{\mathrm{s}}^j - P_{\mathrm{s}}$, the sets in (10.10) are disjoint.

The main goal of step (2) is thus to prove the following lemma.

Lemma 10.9. If $w \in \mathcal{W}$ is not allowed, then for any $s \geq 1$,

$$\|\mathscr{T}_w\|_{H_k^1(\Omega)\to H_k^s(\Omega)} = O(k^{-\infty}).$$

Proof of Theorem 10.5 using Lemma 10.9. By the definition of the error propagation matrix $\mathbf{T}$ (10.16) and the composite map $\mathscr{T}_w$ (10.11), for any $m \in \mathbb{Z}^+$, all entries of $\mathbf{T}^m$ are sums of operators $\mathscr{T}_w$ over words w of size $m+1$. Since words of size $\mathcal{N}+1$ are not allowed by the definition of $\mathcal{N}$ (10.9), the result that $\|\mathbf{T}^{\mathcal{N}}\|_{(H_k^1(\Omega))^N\to(H_k^s(\Omega))^N} = O(k^{-\infty})$ follows from Lemma 10.9. □

We now sketch the proof of Lemma 10.9.

Informal Lemma 10.10. For any $v \in H_0^1(\Omega)$,

$$\mathrm{WF}_\hbar(\mathscr{T}_w v) \subset \{\gamma_n(T_n)\colon \gamma \text{ follows } w\},$$

where we recall from Definition 10.2 that $\gamma_n(T_n)$ is the endpoint of the trajectory γ_n.

Informal Lemma 10.10 shows that if w is not allowed, then for any $v \in H_0^1(\Omega)$, $\mathrm{WF}_\hbar(\mathscr{T}_w v) = \emptyset$. Combining this with the definition of wavefront set (Definition 5.13) and using semiclassical ellipticity to handle very large frequencies (see Galkowski *et al.* 2024*a*, Lemma B.8) proves Lemma 10.9.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Informal Lemma 10.10 is proved using the following propagation result.

Informal Lemma 10.11. Let $v \in H(\Omega_j)$ be the solution to

$$P_{\mathrm{s}}^j v = f \ \text{ in } \Omega_j, \quad v = g 1_{\partial\Omega_j \setminus \partial\Omega} \ \text{ on } \partial\Omega_j$$

If P is non-trapping, then for $\chi \in \overline{C^\infty}(\Omega)$ with $\operatorname{supp}\chi \subset (\Omega_j \cup \partial\Omega)$,

$$\mathrm{WF}_\hbar(\chi v) \subset \operatorname{supp}\chi \cap \left(\mathrm{WF}_\hbar(f) \cup \bigcup_{t\geq 0} \varphi_t\left(\mathrm{WF}_\hbar(f) \cap \{p_s^j = 0\}\right)\right). \tag{10.18}$$

Observe that (10.18) controls the wavefront set of v away from the boundaries of Ω_j that are not also boundaries of Ω.

Both Informal Lemmas 10.10 and 10.11 are stated as informal lemmas because we have not dealt carefully with the boundaries $\partial\Omega$ and $\partial\Omega_j$, respectively. The rigorous statements of these results are Lemmas 4.1 and 6.2, respectively, of Galkowski *et al.* (2024*a*). Our sketches below similarly ignore these issues.

Sketch proof of Informal Lemma 10.10 using Informal Lemma 10.11. Given $v \in H_0^1(\Omega)$, by the definitions of $(\mathbf{T})_{i,j}$ (10.16) and $\mathcal{T}_j$ (10.15), $(\mathbf{T})_{w_{i+1},w_i} v$ equals $\chi_{w_{i+1}}$ multiplied by a Helmholtz solution on $\Omega_{w_{i+1}}$ with data $(P_{\mathrm{s}}^{w_{i+1}} - P_{\mathrm{s}})\chi_{w_i}^{>} v$. Therefore, by Lemma 10.10,

$$\begin{aligned}
&\mathrm{WF}_\hbar\left((\mathbf{T})_{w_{i+1},w_i} v\right) \\
&\quad \subset \operatorname{supp}\chi_{w_{i+1}} \cap \\
&\qquad \left(\mathrm{WF}_\hbar\left((P_{\mathrm{s}}^{w_{i+1}} - P_{\mathrm{s}})\chi_{w_i}^{>} v\right) \cup \bigcup_{t\geq 0} \varphi_t\left(\mathrm{WF}_\hbar\left((P_{\mathrm{s}}^{w_{i+1}} - P_{\mathrm{s}})\chi_{w_i}^{>} v\right) \cap \{p_{\mathrm{s}}^{i+1} = 0\}\right)\right);
\end{aligned} \tag{10.19}$$

in applying Lemma 10.10 we have used that $v = 0$ on $\partial\Omega$ (so that the support condition in g holds) and we need only estimate $\chi_{w_{i+1}} u$ for a Helmholtz solution u because of the multiplication by χ_{i+1} in the definition of $\mathbf{T}$, (10.16). In other words, in the $k \to \infty$ limit, the mass of $(\mathbf{T})_{w_{i+1},w_i} v$ in phase space comes only from $\operatorname{supp}\chi_{w_i}^{>} \cap \operatorname{supp}(P_{\mathrm{s}}^{w_{i+1}} - P_{\mathrm{s}})$, i.e. the first domain in (10.10).

Next, combining (10.19) with $i = 1$ and $i = 2$, the fact that

$$\bigcap_{i=1,2} \operatorname{supp}\chi_{w_{i+1}} \cap \operatorname{supp}(P_{\mathrm{s}}^{w_i} - P_{\mathrm{s}}) \cap \operatorname{supp}\chi_{w_i}^{>} = \emptyset,$$

the definition of following a word (Definition 10.2), and the fact that $\{p_s^j = 0\} \subset \{p = 0\}$, we see that

$$\mathrm{WF}_\hbar\left(\mathbf{T}_{w_3,w_2}\mathbf{T}_{w_2,w_1} v\right) \subset \{p_{\mathrm{s}} = 0\} \cap \{\gamma_3(T_3) \colon \gamma_3 \text{ follows } w_1 w_2 w_3\}.$$

In other words, in the $k \to \infty$ limit, the mass of $(\mathbf{T})_{w_3,w_2}(\mathbf{T})_{w_2,w_1} v$ in phase space comes only from the mass in $\operatorname{supp}\chi_{w_2} \cap \operatorname{supp}(P_{\mathrm{s}}^{w_3} - P_{\mathrm{s}})$ that in turn came from $\operatorname{supp}\chi_{w_1} \cap \operatorname{supp}(P_{\mathrm{s}}^{w_2} - P_{\mathrm{s}})$.

The result then follows by repeating this argument for longer words. □

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Sketch proof of Lemma 10.11. By ellipticity (i.e. Corollary 5.16 above),

$$\mathrm{WF}_\hbar(v) \subset \{p_s^j = 0\} \cup \mathrm{WF}_\hbar(f). \tag{10.20}$$

We now use propagation, almost in the form of Corollary 5.17, except that since P_s^j is not self-adjoint but does have $\operatorname{Im}\sigma(P_s^j) \leq 0$ near $\sigma(P_s^j) = 0$, one may only propagate forward in time under the flow defined by $\operatorname{Re}\sigma(P_s^j)$ (see Remark 5.11); that is, Corollary 5.17 holds with $T \geq 0$.

The fact that P is non-trapping means that every trajectory of p starting from $\operatorname{supp}\chi$ encounters $\{p_s^j \neq 0\}$ backward in time (see Galkowski *et al.* 2024*a*, Lemma 4.4). It is enough to consider trajectories of the flow defined by p rather than those of the flow defined by $\operatorname{Re} p_s^j$ since $H_{\operatorname{Re} p_s^j} = H_p$ on $\{p_s^j = 0\}$ (see Galkowski *et al.* 2024*a*, Lemma 4.2), and (10.20) allow us to understand $\mathrm{WF}_\hbar(\chi v) \cap \{p_s^j \neq 0\}$.

Therefore, applying Corollary 5.17 with $T \geq 0$ (as discussed above), we see that if

$$(x_0, \xi_0) \in \{p_s^j = 0\} \cap \left(\mathrm{WF}_\hbar(f) \cup \bigcup_{t \geq 0} \varphi_t(\mathrm{WF}_\hbar(f) \cap \{p_s^j = 0\}) \right)^c,$$

then there exists $T > 0$ such that $\varphi_{-T}(x_0, \xi_0) \in \{p_s^j \neq 0\}$ and

$$\left(\bigcup_{t=0}^{T} \varphi_{-t}(x_0, \xi_0) \right) \cap \mathrm{WF}_\hbar(f) = \emptyset.$$

Hence, since $\varphi_{-T}(x_0, \xi_0) \notin \mathrm{WF}_\hbar(v)$, we obtain $(x_0, \xi_0) \notin \mathrm{WF}_\hbar(v)$ and hence also $(x_0, \xi_0) \notin \mathrm{WF}_\hbar(\chi v)$. □

A. h- and p-explicit polynomial approximation

There is a large literature on using interpolants (i.e. polynomials defined via point values of a function) to produce both H^1-conforming and non-conforming approximations by piecewise polynomials; see e.g. Brenner and Scott (2008, Chapter 4), Sauter and Schwab (2011, Chapter 4), Apel and Melenk (2015) and Ern and Guermond (2021, Part III). However, it is well known that designing interpolants that behave well in the large polynomial-degree limit can be subtle (e.g. due to Runge's phenomenon) and thus there has been much work on obtaining p-explicit approximants that are not necessarily interpolants; see the discussion and references in Section A.3. However, the authors are unaware of a single source collecting all the ingredients for the proofs of these p-explicit approximation results.

The main goal of this section is therefore to give a self-contained treatment of p- and h-explicit polynomial approximation in L^2-based Sobolev spaces. The main results are given in Theorems A.11 and A.12 below, and cover approximation by arbitrary-degree piecewise polynomials of *either* fixed-regularity functions *or* high-regularity functions. Furthermore, our results for the high-regularity functions,

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

giving precise Sobolev space estimates, appear to be new (see the discussion in Section A.3).

A.1. Definition of shape-regular sequences of triangulations

Definition A.1 (Triangulation). A finite collection of closed sets $\mathcal{T}$ is a *triangulation* of $\Omega \subset \mathbb{R}^d$ if:

(i) for each $T \in \mathcal{T}$, T is closed, $\mathring{T}$ is non-empty and connected,

(ii) $\overline{\Omega} = \cup_{T\in\mathcal{T}} T$,

(iii) if $T_1, T_2 \in \mathcal{T}$ and $\mathring{T}_1 \cap \mathring{T}_2 \neq \emptyset$ then $T_1 = T_2$,

(iv) for each $T \in \mathcal{T}$, T has Lipschitz boundary and

$$\partial T = \bigcup_{j=1}^{d} \bigcup_{i=1}^{N_{j,d}} F_{j,i}(T),$$

with $F_{j,i}$ a smooth, open, embedded submanifold of dimension $d-j$ such that $\partial F_{j,i} = \partial\overline{F_{j,i}}$,

(v) if $T_1, T_2 \in \mathcal{T}$ and $F_{j_1,i_1}(T_1) \cap F_{j_2,i_2}(T_2) \neq \emptyset$, then $F_{j_1,i_1}(T_1) = F_{j_2,i_2}(T_2)$.

We sometimes use the word mesh as a synonym for triangulation.

Definition A.2 (m-simplex). Let $m > 0$. A set $K \Subset \mathbb{R}^d$ is an *m-simplex* if there are $\{x_j\}_{j=0}^m \subset \mathbb{R}^d$ such that

$$K = \text{convex hull}\big(\{x_j\}_{j=0}^m\big).$$

K is an *open m-simplex* if $K = (\widetilde{K})^\circ$, where $\widetilde{K}$ is an m-simplex and $K \neq \emptyset$.

Definition A.3 (Reference element and element maps). $\widehat{T} \Subset \mathbb{R}^d$ is a reference element for a triangulation $\mathcal{T}$ if $\text{diam}(\widehat{T}) = 1$ and there exists a family of bi-Lipschitz maps $\{\mathscr{F}_T\}_{T\in\mathcal{T}}$ such that $T = \mathscr{F}_T(\widehat{T})$ for all $T \in \mathcal{T}$. $\mathcal{T}$ is *simplicial* if $\widehat{T}$ is a d-simplex.

We use the following measure of how star-shaped an open subset of $\mathbb{R}^d$ is.

Definition A.4 (Chunkiness parameter). Let $\Omega \subset \mathbb{R}^d$ and define

$$\rho_{\max}(\Omega) := \sup\{\rho > 0 : \Omega \text{ is star-shaped with respect to a ball of radius } \rho\}.$$

The *chunkiness parameter of* Ω is defined by

$$\gamma(\Omega) := \frac{\text{diam}(\Omega)}{\rho_{\max}(\Omega)}.$$

Observe that $\gamma(\Omega) = 2$ when Ω is a ball.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Definition A.5 (C^r-triangulation). A triangulation is C^r with constant $\Upsilon > 0$ if $\gamma(\widehat{T}) \leq \Upsilon$ and for each $T \in \mathcal{T}$ there exists $h_T > 0$ such that the element map $\mathscr{F}_T$ can be written as $R_T \circ A_T$, where A_T is an affine map and

$$\begin{aligned} &\|\partial A_T\|_{L^\infty} \leq \Upsilon h_T, \quad \|(\partial A_T)^{-1}\|_{L^\infty} \leq \Upsilon h_T^{-1}, \quad \|(\partial R_T)^{-1}\|_{L^\infty} \leq \Upsilon, \\ &\|\partial^\alpha R_T\|_{L^\infty} \leq \Upsilon^{1+|\alpha|}\alpha! \quad \text{for all } |\alpha| \leq r. \end{aligned} \tag{A.1}$$

Let $h(\mathcal{T}) \coloneqq \sup_{T\in\mathcal{T}} h_T$ be the *width of the triangulation* $\mathcal{T}$. Furthermore, a triangulation is C^ω if (A.1) holds for all $|\alpha| < \infty$.

Remark A.6. Given a piecewise C^r-domain, a C^r-triangulation is constructed in Bernardi (1989, §6), building on the two-dimensional results of Scott (1973) and Zlámal (1973) and the isoparametric elements in general dimension of Lenoir (1986).

Remark A.7 (Shape-regularity). A collection of triangulations $\mathscr{T}$ is *shape-regular* in the sense of Ern and Guermond (2021, Definition 11.2) and Brenner and Scott (2008, Definition 4.4.13) if

$$\sup_{\mathcal{T}\in\mathscr{T}} \sup_{T\in\mathcal{T}} \frac{h_T}{r_{\max}(T)} < \infty,$$

where

$$r_{\max}(T) \coloneqq \sup\{r > 0 \colon \text{there is } x_0 \in T \text{ with } B(x_0, r) \subset T\}.$$

Given $\Upsilon > 0$, the collection of C^1-triangulations with this Υ is then shape-regular. Indeed, the bounds on $\partial\mathscr{F}_T$ and $\partial(\mathscr{F}_T)^{-1}$ from (A.1) and the fact that $\operatorname{diam}(\widehat{T}) = 1$ imply that

$$\Upsilon^{-2} h_T \leq \operatorname{diam}(T) \leq \Upsilon^2 h_T$$

and

$$\Upsilon^{-2} h_T r_{\max}(\widehat{T}) \leq r_{\max}(T) \leq \Upsilon^2 h_T r_{\max}(\widehat{T}),$$

and hence, since $\rho_{\max}(\widehat{T}) \leq r_{\max}(\widehat{T})$,

$$\frac{h_T}{r_{\max}(T)} \leq \frac{\Upsilon^2}{r_{\max}(\widehat{T})} \leq \frac{\Upsilon^2}{\rho_{\max}(\widehat{T})} = \Upsilon^2 \gamma(\widehat{T}).$$

Definition A.8 (Affine-conforming triangulation). A C^r-triangulation (in the sense of Definition A.5) is *affine-conforming* if, whenever $F_{j_1,i_1}(T_1) \cap F_{j_2,i_2}(T_2) \neq \emptyset$, there is an affine isomorphism $\kappa \colon \mathscr{F}_{T_1}^{-1}(F_{j_1,i_1}(T_1)) \to \mathscr{F}_{T_2}^{-1}(F_{j_2,i_2}(T_2))$ such that

$$\mathscr{F}_{T_1}|_{\overline{\mathscr{F}_{T_1}^{-1}(F_{j_1,i_1}(T_1))}} = \mathscr{F}_{T_2}|_{\overline{\mathscr{F}_{T_2}^{-1}(F_{j_2,i_2}(T_2))}} \circ \kappa. \tag{A.2}$$

We use the adjective *conforming* in Definition A.8 because the additional property (A.2) is used to generate an H^1-conforming polynomial approximant in A.11 below.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Remark A.9. Without a condition on the element maps such as Definition A.8, the space of mapped polynomials on two adjacent mesh elements may not even intersect non-trivially when restricted to a common face. If the intersection is trivial, there cannot be any non-trivial, globally H^1-mapped polynomials.

Conditions such as Definition A.8 are therefore common in the literature. For example, the Lagrange $\mathbb{P}_{k,d}$ elements in Ern and Guermond (2021) are affine-conforming (in the sense of Definition A.8) by Ern and Guermond (2021, Definition 8.1 and Exercise 20.1).

Given a triangulation $\mathcal{T}$ with element maps $\{\mathscr{F}_T\}_{T\in\mathcal{T}}$ and reference element $\widehat{T}$, the corresponding spaces of mapped polynomials are defined by

$$\mathcal{P}^p(\mathcal{T},\{\mathscr{F}_T\}) \coloneqq \{v\in H^1(\Omega)\colon \text{for all } T\in\mathcal{T}\colon v|_T\circ\mathscr{F}_T\in\mathbb{P}^p(\widehat{T})\}, \tag{A.3}$$

where $\mathbb{P}^p(\widehat{T})$ denotes the space of polynomials of degree $\le p$ on the reference element. We often abbreviate $\mathcal{P}^p(\mathcal{T},\{\mathscr{F}_T\})$ to $\mathcal{P}^p_{\mathcal{T}}$.

Definition A.10. An operator $\mathcal{I}\colon C^\infty(\Omega)\to L^2(\Omega)$ is *triangulation-preserving* if, for each $T\in\mathcal{T}$,

$$u|_T=0 \quad\Rightarrow\quad (\mathcal{I}u)|_T=0,$$

and for any codimension 1 boundary face $F_{1,i}$ of T,

$$u|_{\overline{F_{1,i}}}=0 \quad\Rightarrow\quad (\mathcal{I}u)|_{\overline{F_{1,i}}}=0.$$

A.2. Statement of the polynomial-approximation results

To state results about approximation of high-regularity functions by arbitrary degree polynomials, we need to choose carefully the constants in the high-regularity Sobolev spaces. For an open set $\Omega\subset\mathbb{R}^d$, we define

$$\begin{aligned}
\|u\|^2_{H^p_k(\Omega)} &\coloneqq \sum_{\ell=0}^{p}\frac{1}{(\ell!)^2}|u|^2_{H^\ell_k(\Omega)},\\
|u|^2_{H^p_k(\Omega)} &\coloneqq \sum_{\substack{|\alpha|=p\\ \alpha\in\mathbb{N}^d}}\frac{p!}{\alpha!}\|(k^{-1}D_{x_1})^{\alpha_1}(k^{-1}D_{x_2})^{\alpha_2}\cdots(k^{-1}D_{x_d})^{\alpha_d}u\|^2_{L^2(\Omega)}.
\end{aligned} \tag{A.4}$$

We also denote $\|\cdot\|_{H^p(\Omega)}\coloneqq\|\cdot\|_{H^p_1(\Omega)}$. The rationale for these choices is given in Section A.5 below.

Theorem A.11 (Approximation in H^1_k). Given $\Upsilon>0$ and $t>d/2$, there exists $C>0$ such that for all $p\ge\max(t-1,1)$, all affine-conforming C^{p+1}-simplicial triangulations with constant Υ, and all partitions of $\mathcal{T}$ into $\mathcal{T}_1\sqcup\mathcal{T}_2$, there exists a triangulation-preserving operator $\mathcal{C}^p_{\mathcal{T}}\colon H^1(\Omega)\cap(\oplus_{T\in\mathcal{T}}H^t(T))\to\mathcal{P}^p_{\mathcal{T}}\cap H^1(\Omega)$ such

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

that for all $0 \le s \le 1$ and $u \in H^1(\Omega) \cap (\oplus_{T\in\mathcal{T}} H^t(T))$,

$$\left\|\left(I - \mathcal{C}_{\mathcal{T}}^p\right)u\right\|_{H_k^s(\Omega_{\mathcal{T}_1})} \le C\left(\frac{h(\mathcal{T})k}{p}\right)^{t-s} k^{-t}\left(\sum_{T\in\mathcal{T}_1}\|u\|^2_{H^t(T)}\right)^{1/2},$$

$$\left\|\left(I - \mathcal{C}_{\mathcal{T}}^p\right)u\right\|_{H_k^s(\Omega_{\mathcal{T}_2})} \le (Ch(\mathcal{T})k)^{p+1-s} k^{-p-1}\left(\sum_{T\in\mathcal{T}_2}\|u\|^2_{H^{p+1}(T)}\right)^{1/2},$$

where

$$\Omega_{\mathcal{T}_i} := \bigcup_{T\in\mathcal{T}_i} T,$$

and for $u \in \mathcal{P}_{\mathcal{T}}^p \cap H^1(\Omega)$, $\mathcal{C}_{\mathcal{T}}^p u = u$. In particular,

$$\begin{aligned}\left\|\left(I - \Pi^{H^1}_{\mathcal{P}_{\mathcal{T}}^p}\right)u\right\|_{H_k^s(\Omega)} \le{}& C\left(\frac{h(\mathcal{T})k}{p}\right)^{t-s} k^{-t}\left(\sum_{T\in\mathcal{T}_1}\|u\|^2_{H^t(T)}\right)^{1/2}\\ &+ (Ch(\mathcal{T})k)^{p+1-s} k^{-p-1}\left(\sum_{T\in\mathcal{T}_2}\|u\|^2_{H^{p+1}(T)}\right)^{1/2},\end{aligned}$$

where $\Pi^{H^1}_{\mathcal{P}_{\mathcal{T}}^p}$ is the H^1_k-orthogonal projector onto $\mathcal{P}_{\mathcal{T}}^p \cap H^1$.

We emphasize that by taking *either* $\mathcal{T}_2 = \emptyset$ *or* $\mathcal{T}_1 = \emptyset$, one obtains from Theorem A.11 separate results on approximating *either* fixed-regularity functions *or* high-regularity functions, respectively. However, Theorem A.11 is the stronger result that, given any partition of the triangulation, there exists a single H^1-conforming and triangulation-preserving operator $\mathcal{C}_{\mathcal{T}}^p$ achieving the fixed-regularity approximation on the first part of the triangulation and the high-regularity approximation on the second part.

Next, we record the non-conforming analogue of Theorem A.11.

Theorem A.12 (Approximation in L^2). Given $\Upsilon > 0$ and $t \ge 0$, there exists $C > 0$ such that for all $p \ge 0$ and all C^{p+1}-triangulations with constant Υ, there exists a triangulation-preserving operator $\mathcal{C}_{\mathcal{T}}^p : (\oplus_{T\in\mathcal{T}} H^t(T)) \to \mathcal{P}_{\mathcal{T}}^p$ such that for all $u \in (\oplus_{T\in\mathcal{T}} H^t(T))$,

$$\left\|\left(I - \mathcal{C}_{\mathcal{T}}^p\right)u\right\|_{L^2(\Omega_{\mathcal{T}_1})} \le C\left(\frac{h(\mathcal{T})k}{p}\right)^{t} k^{-t}\left(\sum_{T\in\mathcal{T}_1}\|u\|^2_{H^t(T)}\right)^{1/2},$$

$$\left\|\left(I - \mathcal{C}_{\mathcal{T}}^p\right)u\right\|_{L^2(\Omega_{\mathcal{T}_2})} \le (Ch(\mathcal{T})k)^{p+1} k^{-p-1}\left(\sum_{T\in\mathcal{T}_2}\|u\|^2_{H^{p+1}(T)}\right)^{1/2},$$

where

$$\Omega_{\mathcal{T}_i} := \bigcup_{T\in\mathcal{T}_i} T,$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

and for $u \in \mathcal{P}^p_{\mathcal{T}} \cap H^1(\Omega)$, $\mathcal{C}^p_{\mathcal{T}} u = u$. In particular,

$$\left\| \left(I - \Pi^{L^2}_{\mathcal{P}^p_{\mathcal{T}}}\right) u \right\|_{H^s_k(\Omega)} \leq C \left(\frac{h(\mathcal{T})k}{p} \right)^t k^{-t} \left(\sum_{T \in \mathcal{T}_1} \|u\|^2_{H^t(T)} \right)^{1/2} + (Ch(\mathcal{T})k)^{p+1} k^{-p-1} \left(\sum_{T \in \mathcal{T}_2} \|u\|^2_{H^{p+1}(T)} \right)^{1/2},$$

where $\Pi^{L^2}_{\mathcal{P}^p_{\mathcal{T}}}$ is the L^2-orthogonal projector onto $\mathcal{P}^p_{\mathcal{T}}$.

A.3. Relationship to existing results in the literature

In two and three dimensions, fixed-regularity results similar to those in Theorems A.11 and A.12 appear in Melenk and Sauter (2010, Appendix B); earlier related work includes that of Babuška and Suri (1987), Babuška, Craig, Mandel and Pitkäranta (1991), Muñoz-Sola (1997) and Guo and Zhang (2009); see the discussion in Apel and Melenk (2015, §3.4). Compared to Melenk and Sauter (2010), our presentation stresses the precise assumptions on the triangulation used to convert approximants on reference elements into H^1-conforming approximants on the triangulation of the full domain, Ω. Furthermore, we give a detailed proof of the construction of such an H^1-conforming approximant on all of Ω starting from an appropriate approximant on the reference element (having a modification of the 'element-by-element construction' in the sense of Melenk and Sauter (2010, Definition 5.3); see Definition A.20, Proposition A.21, (A.39) and Remark A.22 below).

We are not aware of the high-regularity estimates in Theorems A.11 and A.12 in the existing literature. While there are approximation results in spaces of analytic functions (Melenk 2004, Melenk and Sauter 2010, 2021, Bernkopf *et al.* 2025), these do not lead to a norm bound on $I - \Pi_{\mathcal{P}^p_h}$ between Sobolev spaces (as in the high-regularity results in Theorems A.11 and A.12).

This appendix presents an almost self-contained proof of Theorems A.11 and A.12. We use without proof only basic properties of averaged Taylor polynomials from Brenner and Scott (2008, Chapter 4) (precisely, Corollary 4.2.18 and Propositions 4.1.17 and 4.2.8).

A.4. Outline of the proofs

In Section A.5 we motivate the choice of Sobolev norms in (A.4) and show that they behave well under pull-back by analytic maps. Then, in Section A.6, we give two types of p-explicit results about approximation by polynomials in open subsets of $\mathbb{R}^d$ with Lipschitz boundaries. First we estimate the error in approximating high-regularity functions by high-degree polynomials (coming from averaged Taylor expansions), following Brenner and Scott (2008) but keeping careful track of

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

dependence on the polynomial degree and regularity. Second, we estimate the error in approximating fixed-regularity functions by high-degree Legendre polynomials (in one dimension, such results go back at least to Canuto and Quarteroni 1982; see Babuška and Suri 1994, Bernardi and Maday 1997, Schwab 1998 and the references therein). These estimates can then be used to construct good polynomial approximations of a function on a single element, but they cannot be directly used to construct H^1-conforming approximations by piecewise polynomials on all of Ω. In Section A.7 we consider polynomial approximation on simplices and modify the polynomial approximants constructed in Section A.6 so that they can be extended to H^1-conforming piecewise polynomials on the full triangulation. To do this, we use ideas from Melenk and Sauter (2010), who, in turn, use ideas from Muñoz-Sola (1997) and Guo and Zhang (2009).

A.5. H_k^p-norms

To determine the correct choice of H_k^p-norm we return to $\mathbb{R}^d$, where the semiclassical Fourier transform (defined by (2.1)) gives a natural H_k^p-seminorm:

$$|u|^2_{H_k^p(\mathbb{R}^d)} := \frac{1}{(2\pi)^d}\big\||\xi|^p \mathcal{F}_k(u)(\xi)\big\|^2_{L^2(\mathbb{R}^d)}.$$

We now write this in terms of L^2-norms of derivatives:

$$\begin{aligned}
|u|^2_{H_k^p(\mathbb{R}^d)} &= \frac{1}{(2\pi^d)}\int |\xi|^{2p}|\mathcal{F}_k(u)(\xi)|^2\,d\xi \\
&= \frac{1}{(2\pi^d)}\int \sum_{\substack{|\alpha|=p\\ \alpha\in\mathbb{N}^d}} \frac{p!}{\alpha!}\xi_1^{2\alpha_1}\cdots\xi_d^{2\alpha_d}|\mathcal{F}_k(u)(\xi)|^2\,d\xi \\
&= \sum_{\substack{|\alpha|=p\\ \alpha\in\mathbb{N}^d}} \frac{p!}{\alpha!}\|(k^{-1}D_{x_1})^{\alpha_1}(k^{-1}D_{x_2})^{\alpha_2}\cdots(k^{-1}D_{x_d})^{\alpha_d}u\|^2_{L^2(\mathbb{R}^d)}.
\end{aligned}$$

Therefore, on an open set $\Omega\subset\mathbb{R}^d$, it is natural to define $|u|_{H_k^p(\Omega)}$ as in (A.4).

Now, we need to determine the appropriate way to combine the H_k^p-seminorms into a norm. It is natural to consider a norm that contains a multiple of each H_k^ℓ-seminorm, $\ell = 0,\dots,p$, such that for analytic functions (whose αth derivative goes like $C^{|\alpha|}|\alpha|!$), each term in the sum has roughly the same size and whose total size grows at most exponentially in p. The Sobolev norm in (A.4) then achieves these goals.

To make sure our choice of norm is appropriate, we check two important properties. First we estimate the norm of the derivative in H_k^p.

Lemma A.13. Let $\Omega\subset\mathbb{R}^d$. Then, for all $p\ge 0$,

$$\|k^{-1}\nabla u\|_{H_k^p(\Omega)} \le (p+1)\|u\|_{H_k^{p+1}(\Omega)}.$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Proof. By the definition (A.4),

$$\begin{aligned}
&\|k^{-1}\nabla u\|^2_{H^p_k(\Omega)} \\
&\quad= \sum_{\ell=0}^{p} \sum_{\substack{|\alpha|=\ell \\ \alpha\in\mathbb{N}^d}} \sum_{j=1}^{d} \frac{1}{\ell!\alpha!} \|(k^{-1}D_{x_1})^{\alpha_1}(k^{-1}D_{x_2})^{\alpha_2}\cdots(k^{-1}D_{x_d})^{\alpha_d}(k^{-1}D_{x_j})u\|^2_{L^2(\Omega)} \\
&\quad\le \sum_{\ell=0}^{p} \sum_{\substack{|\alpha|=\ell+1 \\ \alpha\in\mathbb{N}^d}} \frac{(\ell+1)^2}{(\ell+1)!\alpha!} \|(k^{-1}D_{x_1})^{\alpha_1}(k^{-1}D_{x_2})^{\alpha_2}\cdots(k^{-1}D_{x_d})^{\alpha_d}u\|^2_{L^2(\Omega)} \\
&\quad\le \sum_{\ell=0}^{p+1} \sum_{\substack{|\alpha|=\ell \\ \alpha\in\mathbb{N}^d}} \frac{\ell^2}{(\ell!)^2} |u|^2_{H^\ell_k} \le (p+1)^2 \|u\|^2_{H^{p+1}_k(\Omega)}.
\end{aligned}$$

□

Next, we show that the H^p-norm defined by (A.4) behaves well under pull-backs with analytic maps.

Lemma A.14. Let $d \ge 1$, $C_0 > 0$. Then there exists $C_1 > 0$ such that for all $t \ge 0$, $U, V \subset \mathbb{R}^d$ open, $\gamma\colon U \to V$ a bijection satisfying

$$\|\partial^\alpha \gamma\|_{L^\infty} \le C_0^{|\alpha|}\alpha! \quad \text{for } \alpha \in \mathbb{N}^d,\ |\alpha| \le \max(p,1) \text{ and } \|(\partial\gamma)^{-1}\|_{L^\infty} \le C_0,$$

and $u \in H^t(V)$,

$$\frac{1}{t!}|\gamma^* u|_{H^t(U)} \le C_1^t \|u\|_{H^t(V)},$$

where $\gamma^* u := u \circ \gamma$ denotes the pull-back operator.

Proof. We start from the Faà di Bruno formula (Constantine and Savits 1996, Theorem 2.1). Let $\mu, \nu \in \mathbb{N}^d$. We say $\mu \prec \nu$ if one of the following holds:

(1) $|\mu| < |\nu|$, or
(2) there exists $1 \le r \le d$ such that $\mu_i = \nu_i$, $i = 1 \ldots, r-1$ and $\mu_r < \nu_r$.

Then, defining

$$p_s(\nu,\lambda) := \left\{ (\mu_1,\ldots,\mu_s;\ell_1,\ldots,\ell_s) \in (\mathbb{N}^d)^{2s} \colon |\mu_i| > 0,\ 0 \prec \ell_1 \prec \cdots \prec \ell_s, \ \sum_{i=1}^{s} \mu_i = \lambda,\ \sum_{i=1}^{s} |\mu_i|\ell_i = \nu \right\},$$

we have

$$\begin{aligned}
&\frac{1}{(t!)^2}|u\circ\gamma|^2_{H^t(U)} \\
&\quad= \sum_{|\nu|=t} \int_U \left(\sum_{m=1}^{t} \sum_{|\lambda|=m} (\partial^\lambda u)\circ\gamma \sum_{s=1}^{t} \sum_{p_s(\nu,\lambda)} \frac{\nu!}{\sqrt{\nu!t!}} \prod_{j=1}^{s} \frac{[\partial^{\ell_j}\gamma]^{\mu_j}}{\mu_j![\ell_j!]^{|\mu_j|}} \right)^2
\end{aligned}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

$$\leq C_0^{2t} \sum_{|\nu|=t} \int_U \left(\sum_{m=1}^{t} \sum_{|\lambda|=m} \frac{1}{m!} (\partial^\lambda u) \circ \gamma \sum_{s=1}^{t} \sum_{p_s(\nu,\lambda)} \frac{\nu! m!}{\sqrt{\nu! t!}} \prod_{j=1}^{s} \frac{1}{\mu_j!} \right)^2$$

$$\leq C_0^{2t} C_1 \|u\|^2_{H^t(V)} \sum_{|\nu|=t} \sum_{m=1}^{t} \sum_{|\lambda|=m} \frac{\lambda!}{m!} \left(\sum_{s=1}^{t} \sum_{p_s(\nu,\lambda)} \frac{\nu! m!}{\sqrt{\nu! t!}} \prod_{j=1}^{s} \frac{1}{\mu_j!} \right)^2.$$

Therefore we claim that there exists $C > 0$ such that

$$b_t := \sum_{|\nu|=t} \sum_{m=1}^{t} \sum_{|\lambda|=m} \frac{\lambda!}{m!} \left(\sum_{s=1}^{t} \sum_{p_s(\nu,\lambda)} \frac{\nu! m!}{\sqrt{\nu! t!}} \prod_{j=1}^{s} \frac{1}{\mu_j!} \right)^2 \leq C^t. \tag{A.5}$$

To do this, we convert the sum over multi-indices μ to a sum over only their lengths. Let $r \in \mathbb{N}^s$ with $|r| = |\lambda|$ and consider the following:

$$a_{r,\lambda} := \sum_{\substack{\mu_1 + \cdots + \mu_s = \lambda \\ |\mu_i| = r_i}} \prod_{j=1}^{s} \frac{1}{\mu_j!}.$$

By the multinomial theorem,

$$\sum_{|\mu|=r} \frac{x^\mu}{\mu!} = \frac{(x_1 + \cdots + x_d)^r}{r!}.$$

Thus

$$\begin{aligned}
\sum_{|\lambda|=|r|} a_{r,\lambda} x^\lambda &= \sum_{\substack{\mu_1,\ldots,\mu_s \\ |\mu_i|=r_i}} \prod_{j=1}^{s} \frac{x^{\mu_j}}{\mu_j!} \\
&= \prod_{j=1}^{s} \sum_{|\mu|=r_j} \frac{x^\mu}{\mu!} \\
&= \prod_{j=1}^{s} \frac{(x_1 + \cdots + x_d)^{r_j}}{r_j!} \\
&= \frac{(x_1 + \cdots + x_d)^{|r|}}{r!} \\
&= \frac{1}{r!} \sum_{|\lambda|=|r|} \frac{|r|! x^\lambda}{\lambda!}.
\end{aligned}$$

Therefore $a_{r,\lambda}$ is given by the coefficient of x^λ on the right-hand side:

$$a_{r,\lambda} = \frac{|\lambda|!}{\lambda! r!}.$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Hence, defining

$$q_s(\nu,\lambda) := \left\{ \begin{array}{r} (r;\ell_1,\dots,\ell_s) \in \mathbb{N}^s \times (\mathbb{N}^d)^s : r_i > 0,\ 0 \prec \ell_1 \prec \cdots \prec \ell_s, \\ |r| = |\lambda|,\ \sum_{i=1}^{s} r_i \ell_i = \nu \end{array} \right\},$$

we have

$$b_t = \sum_{|\nu|=t} \sum_{m=1}^{t} \sum_{|\lambda|=m} \frac{\nu! m!}{t!\lambda!} \left(\sum_{s=1}^{t} \sum_{q_s(\nu,\lambda)} \frac{m!}{r!} \right)^2. \tag{A.6}$$

Observe that for fixed $\{\ell_i\}_{i=1}^s$ and r, $m!/r!$ is the number of ordered m-tuples $(w_1,\dots,w_m) \in (\mathbb{N}^d \setminus \{0\})^m$ in which exactly r_i entries are equal to ℓ_i. Therefore

$$\sum_{s=1}^{t} \sum_{q_s(\nu,\lambda)} \frac{m!}{r!} \le \#\left\{ (w_1,\dots,w_m) \in (\mathbb{N}^d \setminus \{0\})^m : \sum_i w_i = \nu \right\}. \tag{A.7}$$

Now

$$\#\left\{ (w_1,\dots,w_m) \in (\mathbb{N}^d \setminus \{0\})^m : \sum_i w_i = \nu \right\} \le \prod_{j=1}^{d} \frac{(\nu_j + m - 1)!}{(m-1)!\nu_j!}, \tag{A.8}$$

since for the jth coordinate of ν we break ν_j into m non-empty groups. Finally, using the arithmetic mean–geometric mean inequality,

$$\left(\prod_{i=1}^{d} a_i \right)^{1/d} \le \frac{1}{d} \sum_{i=1}^{d} a_i \quad \text{for } a_i \ge 0,$$

we obtain

$$\begin{aligned} \prod_{j=1}^{d} \frac{(\nu_j + m - 1)!}{(m-1)!\nu_j!} &\le \prod_{j=1}^{d} \frac{(\nu_j + m - 1)^{m-1}}{(m-1)!} \\ &\le \frac{(|\nu|/d + m - 1)^{d(m-1)}}{[(m-1)!]^d} \\ &\le C^{d(m-1)} \left(\frac{|\nu|}{d(m-1)} + 1 \right)^{d(m-1)}. \end{aligned}$$

Now

$$\sup_{x \in (0,\infty)} (1+x)^{1/x} < \infty$$

so that

$$\prod_{j=1}^{d} \frac{(\nu_j + m - 1)!}{(m-1)!\nu_j!} \le C^{d(m-1)} C^{|\nu|}. \tag{A.9}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

The combination of (A.6), (A.7), (A.8) and (A.9) implies that

$$b_t \leq \sum_{|\nu|=t} \sum_{m=1}^{t} \sum_{|\lambda|=m} \frac{\nu! m!}{t! \lambda!} C^t.$$

Now

$$\sum_{|\lambda|=m} \frac{m!}{\lambda!} = d^m \quad \text{and} \quad \sum_{|\nu|=t} \frac{\nu!}{t!} \leq \#\{\nu : |\nu| = t\} \leq (t+1)^{d-1}$$

(by, respectively, the multinomial theorem and the fact that $\nu! \leq |\nu|!$). Thus $b_t \leq C^t$, from which (A.5), and hence the result, follows. □

A.6. Polynomial approximation on subsets of $\mathbb{R}^d$

The following result shows that one can approximate high-regularity functions very accurately by piecewise polynomials of high degree.

Proposition A.15. Given $d \in \mathbb{N}$, there exists $C > 0$ such that the following holds. Let $\Omega \Subset \mathbb{R}^d$ with $\gamma(\Omega) < \infty$. Then there are $\mathcal{I}^{p-1} : H^p_k(\Omega) \to \mathbb{P}^{p-1}$, $p \geq 1$ such that for $|\alpha| \leq p$, $D^\alpha \mathcal{I}^{p-1} = \mathcal{I}^{p-1-|\alpha|} D^\alpha$, and for $\ell \in \{0, 1, \ldots, p\}$, $u \in H^p_k$,

$$\|(I - \mathcal{I}^{p-1})u\|^2_{H^\ell_k(\Omega)} \leq C \sum_{j=0}^{\ell} \frac{(k \operatorname{diam}(\Omega))^{2p-2j}}{[(p-j)!j!]^2} (\gamma+1)^{2d} d^{p-j} |u|^2_{H^p_k(\Omega)}. \quad \text{(A.10)}$$

Moreover, if $u \in \mathbb{P}^{p-1}$, then $\mathcal{I}^{p-1} u = u$.

To prove Proposition A.15, we follow Brenner and Scott (2008, Chapter 4).

Lemma A.16 (Brenner and Scott 2008, Corollary 4.2.18 and Propositions 4.1.17 and 4.2.8). Given $d \in \mathbb{N}$, there exists $C > 0$ such that the following holds. For all $\Omega \Subset \mathbb{R}^d$, with $\gamma(\Omega) < \infty$, there is a ball B such that Ω is star-shaped with respect to B, and there are $\mathcal{I}^{p-1} : H^p_k(\Omega) \to \mathbb{P}^{p-1}$, $p \geq 0$ such that for $|\alpha| \leq p-1$, $D^\alpha \mathcal{I}^{p-1} = \mathcal{I}^{p-1-|\alpha|} D^\alpha$ and for $u \in H^p_k(\Omega)$,

$$(u - \mathcal{I}^{p-1} u)(x) = p \sum_{|\alpha|=p} \int_{C_x} k_\alpha(x,z) D^\alpha u(z)\, dz,$$

where $k_\alpha(x,z) = \frac{1}{\alpha!}(x-z)^\alpha k(x,z)$,

$$|k(x,z)| \leq C(\gamma(\Omega)+1)^d |x-z|^{-d},$$

and C_x is the convex hull of $\{x\} \cup B$.

Sketch of proof. The result follows by averaging the Taylor polynomial over a ball and writing the integral form of the remainder. □

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Lemma A.17. Let $\Omega \subset \mathbb{R}^d$, and for $m > 0$ define the operator

$$T_m u(x) := \int_\Omega |x - z|^{m-d} u(z)\, dz.$$

For all $m > 0$,

$$\|T_m u\|_{L^2(\Omega)} \leq \frac{(\operatorname{diam}(\Omega))^m}{m} \|u\|_{L^2(\Omega)}.$$

Proof. The result follows from the Schur test for boundedness (see e.g. Hörmander 1985, Lemma 18.1.12) since

$$\begin{aligned}\sup_{x\in\Omega} \int_\Omega |x - z|^{m-n}\, dz &= \sup_{z\in\Omega} \int_\Omega |x - z|^{m-d}\, dx \\ &\leq \int_0^{\operatorname{diam}(\Omega)} r^{m-d+d-1}\, dr \\ &\leq \frac{(\operatorname{diam}(\Omega))^m}{m}.\end{aligned}$$

□

Proof of Proposition A.15. We start with the case $m = 0$. Then, by Lemma A.16, there exists $\mathcal{I}^{p-1}$ with the required properties such that

$$\begin{aligned}|u - \mathcal{I}^{p-1}u|(x) &\leq Cp(\gamma+1)^d \sum_{|\alpha|=p} \int_\Omega \frac{1}{\alpha!} |(x-z)^\alpha| |x-z|^{-d} |D^\alpha u(z)|\, dz \\ &\leq Cp \frac{1}{p!} (\gamma+1)^d \sum_{|\alpha|=p} \frac{p!}{\alpha!} \int_\Omega |x-z|^{p-d} |D^\alpha u(z)|\, dz.\end{aligned}$$

The fact that for $u \in \mathbb{P}^{p-1}$, $\mathcal{I}^{p-1}u = u$ follows from the equality $D^\alpha \mathcal{I}^{p-1} = \mathcal{I}^{p-1-|\alpha|} D^\alpha$.

By Lemma A.17 and the triangle inequality, the Cauchy–Schwarz inequality, (7.34), and the definition of $|\cdot|_{H_k^p(\Omega)}$ (A.4),

$$\begin{aligned}&\|u - \mathcal{I}^{p-1}u\|_{L^2(\Omega)} \\ &\quad\leq C \frac{(\operatorname{diam}(\Omega))^p}{p!} (\gamma+1)^d \sum_{|\alpha|=p} \frac{p!}{\alpha!} \|D^\alpha u\|_{L^2(\Omega)} \\ &\quad\leq C \frac{(\operatorname{diam}(\Omega))^p}{p!} (\gamma+1)^d \left(\sum_{|\alpha|=p} \frac{p!}{\alpha!} \right)^{1/2} \left(\sum_{|\alpha|=p} \frac{p!}{\alpha!} \|D^\alpha u\|^2_{L^2(\Omega)} \right)^{1/2} \\ &\quad= C \frac{(k \operatorname{diam}(\Omega))^p}{p!} (\gamma+1)^d d^{p/2} |u|_{H_k^p(\Omega)}. \qquad \text{(A.11)}\end{aligned}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

By the definition of $|\cdot|_{H_k^\ell(\Omega)}$ (A.4), the property $D^\alpha \mathcal{I}^{p-1} = \mathcal{I}^{p-1-|\alpha|} D^\alpha$, (A.11), and (A.4) again,

$$
\begin{aligned}
&|u - \mathcal{I}^{p-1}u|^2_{H_k^\ell(\Omega)} \\
&\quad = \sum_{|\alpha|=\ell} \frac{\ell!}{\alpha!} \|k^{-|\alpha|} D^\alpha u - k^{-|\alpha|} D^\alpha \mathcal{I}^{p-1} u\|^2_{L^2(\Omega)} \\
&\quad = \sum_{|\alpha|=\ell} \frac{\ell!}{\alpha!} \|k^{-|\alpha|} D^\alpha u - \mathcal{I}^{p-1-|\alpha|}(k^{-|\alpha|} D^\alpha u)\|^2_{L^2(\Omega)} \\
&\quad \le C \frac{(k \operatorname{diam}(\Omega))^{2p-2\ell}}{[(p-\ell)!]^2} (\gamma+1)^{2d} d^{p-\ell} \sum_{|\alpha|=\ell} \frac{\ell!}{\alpha!} |k^{-|\alpha|} D^\alpha u|^2_{H_k^{p-|\alpha|}(\Omega)} \\
&\quad \le C \frac{(k \operatorname{diam}(\Omega))^{2p-2\ell}}{[(p-\ell)!]^2} (\gamma+1)^{2d} d^{p-\ell} \\
&\qquad \times \sum_{|\alpha|=\ell} \frac{\ell!}{\alpha!} \sum_{\substack{|\beta|=p-\ell \\ \beta\in\mathbb{N}^d}} \frac{(p-\ell)!}{\beta!} \|(k^{-1}D_{x_1})^{\beta_1} \cdots (k^{-1}D_{x_d})^{\beta_d} (k^{-1}D)^\alpha u\|^2_{L^2(\Omega)} \\
&\quad \le C \frac{(k \operatorname{diam}(\Omega))^{2p-2\ell}}{[(p-\ell)!]^2} (\gamma+1)^{2d} d^{p-\ell} |u|^2_{H_k^p(\Omega)}.
\end{aligned}
$$

where in the last line we used

$$
\sup_{\substack{|\alpha|=\ell \\ |\beta|=p-\ell}} \frac{(\alpha+\beta)!}{\alpha!\beta!} = \frac{p!}{\ell!(p-\ell)!}
$$

Interpreting the left-hand side combinatorially, we see that the number on the left is less than the number of ways of choosing a group of ℓ objects from a group of p objects, and this number is obviously achieved when α and β have only one non-zero entry.

Therefore, by (A.4),

$$
\begin{aligned}
\|u - \mathcal{I}^{p-1}u\|^2_{H_k^\ell(\Omega)} &= \sum_{j=0}^{\ell} \frac{1}{(j!)^2} |u - \mathcal{I}^{p-1}u|^2_{H_k^j(\Omega)} \\
&\le \sum_{j=0}^{\ell} C \frac{(k \operatorname{diam}(\Omega))^{2p-2j}}{[(p-j)!j!]^2} (\gamma+1)^{2d} d^{p-j} |u|^2_{H_k^p(\Omega)},
\end{aligned}
$$

which is the result (A.10). □

In Proposition A.15, the regularity of u is tied to the degree of the polynomials, that is, Proposition A.15 applies to $u \in H^m$ with m at least the polynomial degree p. We now show that the error when approximating a *fixed-regularity* function by polynomials decreases with the polynomial degree. The main idea is to use the spectral theory of a self-adjoint elliptic operator whose eigenfunctions are polynomials to

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

approximate a function with fixed regularity. In this case we use the Legendre polynomials: eigenfunctions of the operator $-\partial_x(1-x^2)\partial_x$ on $[-1,1]$.

Proposition A.18. Let $\Omega \Subset \mathbb{R}^d$ open with Lipschitz boundary and $m \geq \ell \geq 0$. Then there exists $C > 0$ such that for all $p \geq m$, there exists $\mathcal{L}^{p-1} \colon H^m_k(\Omega) \to \mathbb{P}^{p-1}$ such that for all $u \in H^m_k(\Omega)$,

$$\|(I - \mathcal{L}^{p-1})u\|_{H^\ell_k(\Omega)} \leq C \frac{(\operatorname{diam}(\Omega)k)^{m-\ell}}{p^{m-\ell}} |u|_{H^m_k(\Omega)}, \tag{A.12}$$

and for $u \in \mathbb{P}^{p-1}$, $\mathcal{L}^{p-1}u = u$.

Proof. We start by considering Ω with $\operatorname{diam}(\Omega) = 1$. Then, without loss of generality, we can assume $\Omega \Subset [-1,1]^d$. By the now-classical result of Calderón (1961), there exists $E \colon H^m(\Omega) \to H^m(\mathbb{R}^d)$ such that

$$\|Eu\|_{H^m(\mathbb{R}^d)} \leq C\|u\|_{H^m(\Omega)}, \quad (Eu)|_\Omega = u.$$

Let $\chi \in C^\infty_c((-1,1)^d)$ with $\operatorname{supp}(1-\chi) \cap \Omega = \emptyset$. Define the sesquilinear form $Q \colon H^1([-1,1]^d) \times H^1([-1,1]^d) \to \mathbb{C}$ by

$$Q(u,v) = \sum_{i=1}^d \langle (1-x_i^2)\partial_{x_i} u, \partial_{x_i} v\rangle.$$

Then, for all $u, v \in H^1$,

$$0 \leq Q(u,u), \quad |Q(u,v)| \leq C\|u\|_{H^1}\|v\|_{H^1},$$

and hence the Friedrichs extension of Q, which we denote by $P \colon L^2([-1,1]^d) \to L^2([-1,1]^d)$, is self-adjoint with domain

$$\mathcal{D}(P) := \{u \in H^1 \colon |Q(u,v)| \leq C_u\|v\|_{L^2} \text{ for all } v \in H^1\}.$$

Observe that $H^2([-1,1]^d) \subset \mathcal{D}(P)$ and, for $u \in H^2$,

$$Pu = -\sum_{i=1}^d \partial_{x_i}(1-x_i^2)\partial_{x_i} u.$$

Let $p_{\lambda_j} \colon [-1,1] \to \mathbb{R}$ be an orthonormal basis of eigenfunctions for P when $d = 1$, i.e. satisfying

$$-\partial_x(1-x^2)\partial_x p_{\lambda_j} - \lambda_j p_{\lambda_j} = 0.$$

We claim that $\lambda_j = j(j+1)$ and p_{λ_j} is a polynomial of degree j. To see this, observe that $\mathbb{P}^\ell$ is invariant under P for every ℓ, and hence P has $\ell+1$ linearly independent eigenfunctions in $\mathbb{P}^\ell$ for every ℓ. This implies that for every n there exists j such that p_{λ_j} is a polynomial of degree exactly n. Since the set of polynomials is dense in $L^2([-1,1])$, we then obtain that for every j there exists n_j such that p_{λ_j} is a polynomial of degree n_j. To compute the eigenvalue, observe that since

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

$p_{\lambda_j}(x) = \sum_{\ell=0}^{n_j} b_\ell x^\ell$ with $b_{n_j} \neq 0$,

$$0 = (-\partial_x(1-x^2)\partial_x p_{\lambda_j} - \lambda_j p_{\lambda_j}) = \sum_{\ell=0}^{n_j} a_\ell x^\ell,$$

where

$$a_{n_j} x^{n_j} = \partial_x x^2 \partial_x b_{n_j} x^{n_j} - \lambda_j b_{n_j} x^{n_j} = (n_j(n_j+1) - \lambda_j) b_{n_j} x^{n_j}.$$

Hence $\lambda_j = n_j(n_j+1)$. We now now relabel these eigenfunctions by their degree, and recall that, in fact, $p_n \colon [-1,1] \to \mathbb{R}$, $n = 0, 1, \ldots$, are the Legendre polynomials.

For $\vec{n} \in \mathbb{N}^d$, define $p_{\vec{n}} \colon [-1,1]^d \to \mathbb{R}$ by

$$p_{\vec{n}}(x) \coloneqq \prod_{j=1}^{d} p_{\vec{n}_j}(x_j).$$

By the definition of the Legendre polynomials,

$$\left(P - \sum_{j=1}^{d} \vec{n}_j(\vec{n}_j+1)\right) p_{\vec{n}} = 0,$$

and hence $\{p_{\vec{n}}\}_{\vec{n}\in\mathbb{N}^d}$ forms an orthonormal basis of eigenfunctions of P for $L^2([-1,1]^d)$.

In particular, there exists $c > 0$ such that with $\Pi_{p-1} \colon L^2([-1,1]^d) \to \mathbb{P}^{p-1}$ the orthogonal projector, $[\Pi_{p-1}, P] = 0$ and, since

$$c_d \left(\sum_j \vec{n}_j\right)^2 \leq \sum_j \vec{n}_j(\vec{n}_j+1)$$

for some $c_d > 0$,

$$1_{(-\infty, cp^2]}(P)\Pi_{p-1} = 1_{(-\infty,cp^2]}(P).$$

Now set $\widetilde{\mathcal{L}}^{p-1} \coloneqq 1_\Omega \Pi_{p-1} \chi E$. Observe that since $1_\Omega E = I$ and $1_\Omega \chi = 1_\Omega$,

$$\begin{aligned}(I - \widetilde{\mathcal{L}}^{p-1}) &= 1_\Omega (I - \Pi_{p-1})\chi E \\ &= 1_\Omega (P+1)^{-m/2} 1_{(cp^2,\infty)}(P)(I-\Pi_{p-1})(P+1)^{m/2}\chi E.\end{aligned}$$

Now, by local elliptic regularity,

$$\|1_\Omega (P+1)^{-t/2}\|_{L^2([-1,1]^d)\to H^t(\Omega)} \leq C_t,$$

so that

$$\begin{aligned}&\|(I - \widetilde{\mathcal{L}}^{p-1})\|_{H^\ell(\Omega)\to H^m(\Omega)} \\ &\quad \leq C_{m-\ell} \|(P+1)^{(m-\ell)/2} 1_{(cp^2,\infty)}(P)\|_{L^2\to L^2} \|(P+1)^{\ell/2}\chi E\|_{H^\ell(\Omega)\to L^2} \\ &\quad \leq C\langle p\rangle^{-\ell+m}.\end{aligned}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

To define $\mathcal{L}^{p-1}$, we need to construct an operator so that $(I-\mathcal{L}^{p-1})u$ is controlled by the H^m-seminorm of u (as in (A.12)). For this, let $\mathcal{Z}\colon H^m(\Omega)\to\mathbb{P}_{m-1}$ be the unique operator such that

$$\int_\Omega \partial^\alpha(u-\mathcal{Z}u)=0,\quad |\alpha|\le m-1, \tag{A.13}$$

and set

$$\mathcal{L}^{p-1}u=\widetilde{\mathcal{L}}^{p-1}(u-\mathcal{Z}u)+\mathcal{Z}u.$$

Then

$$\|u-\mathcal{L}^{p-1}u\|_{H^\ell(\Omega)}=\|u-\mathcal{Z}u-\widetilde{\mathcal{L}}^{p-1}(u-\mathcal{Z}u)\|_{H^\ell}\le C\langle p\rangle^{-\ell+m}\|u-\mathcal{Z}u\|_{H^m}.$$

Using the Poincaré inequality together with (A.13) repeatedly, we obtain

$$\|u-\mathcal{L}^{p-1}u\|_{H^\ell(\Omega)}\le C\langle p\rangle^{-\ell+m}|u-\mathcal{Z}u|_{H^m(\Omega)}=C\langle p\rangle^{-\ell+m}|u|_{H^m(\Omega)},$$

and the result (A.12) then follows by scaling.

The fact that $\mathcal{L}^{p-1}u=u$ for $u\in\mathbb{P}^{p-1}$ follows from the fact that $u=\tilde{\mathcal{L}}u$ if $u\in\mathbb{P}^{p-1}$. □

A.7. Boundary-compatible approximations

We now show that given a polynomial approximant on a simplex with control on a sufficiently high Sobolev norm of the error (more than $d/2$ derivatives in L^2), we can modify the polynomial approximant to one that is *boundary-compatible* in the sense of Definition A.20 below. Boundary-compatible polynomial approximants can then be extended to conforming piecewise polynomial approximants on C^1-triangulations (in the sense of Definition A.5); this is done in Section A.8.

A.7.1. Simplices and boundary compatibility

Definition A.19. Let $K\subset\mathbb{R}^d$ be an open simplex. The boundary of K can be decomposed as

$$\bigcup_{j=1}^{d}\bigcup_{i=1}^{N_{j,d}}F_{j,i},$$

where $F_{j,i}$ is an open simplex in a $(d-j)$-dimensional hyperplane for $j<d$, $F_{d,i}=\{x_i\}$, for some $x_i\in\mathbb{R}^d$, and $F_{j,i}\cap F_{\ell,m}=\emptyset$ for $(j,i)\ne(\ell,m)$. We call $F_{j,i}$ a *face of codimension j for K*. For convenience, we set $F_{0,1}=K$ and $N_{0,1}=1$.

Definition A.20. Let $K\subset\mathbb{R}^d$ be a simplex. An operator $T\colon C^\infty(K)\to C^\infty(K)$ is *boundary-compatible* if, for all faces F,

$$u|_{\overline{F}}=0\quad\Rightarrow\quad (Tu)|_{\overline{F}}=0, \tag{A.14}$$

and for any pair of faces of K, F_1 and F_2, and any affine isomorphism $\kappa\colon\overline{F_1}\to\overline{F_2}$,

$$\left[T(u|_{\overline{F_2}}\circ\kappa)\right]\circ\kappa^{-1}=Tu|_{\overline{F_2}}. \tag{A.15}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Note that (A.15) makes sense since, by (A.14), T defines a map from $C^\infty(F)$ to itself.

The goal of this subsection (Section A.7) is to establish the following proposition, which shows that one can modify polynomial approximants to boundary-compatible polynomial approximants.

Proposition A.21. Let $K \subset \mathbb{R}^d$ be an open d-simplex. Then, for each face of K, $F_{j,i}$, there are operators $Q_{j,i}^{p-1} : H^{1/2}(F_{j,i}) \to \mathbb{P}^{p-1}(F_{j,i})$, $p = 2, 3, \ldots$ such that for all $m > d/2, m \geq 1$ there exists $C > 0$ such that, if $\ell \geq m$, $\mathcal{J}^{p-1} : H_p^\ell(K) \to \mathbb{P}^{p-1}$, and $(I - \mathcal{J}^{p-1}) : H_p^\ell(K) \to H_p^m(K)$, then there is a boundary-compatible operator $\mathcal{C}^{p-1} : H^\ell(K) \to \mathbb{P}^{p-1}$ satisfying

$$\|(I - \mathcal{C}^{p-1})u\|_{H_p^1(K)} \leq C\|(I - \mathcal{J}^{p-1})u\|_{H_p^m(K)}$$

and

$$(\mathcal{C}^{p-1}u)|_{F_{j,i}} = Q_{j,i}^{p-1}(u|_{F_{j,i}}). \tag{A.16}$$

In addition, if, for all $u \in \mathbb{P}^{p-1}$, $\mathcal{J}^{p-1}u = u$, then, for all $u \in \mathbb{P}^{p-1}$, $\mathcal{C}^{p-1}u = u$.

Remark A.22. Proposition A.21 is one important place where our construction of approximants on reference elements differs from that in Melenk and Sauter (2010). Indeed, we construct approximants satisfying (A.15), with this property used in our construction of globally H^1-conforming piecewise polynomial approximants (in Section A.9). In contrast, Melenk and Sauter (2010) construct approximants having an 'element-by-element construction' in the sense of their Definition 5.3. Such approximants do not have the property (A.15) because they involve, for example,

$$\operatorname{argmin}\left\{p\|u - q\|_{L^2(f)} + \|u - q\|_{H^1(f)}\right\} \tag{A.17}$$

(see Melenk and Sauter 2010, equation 5.5). In general,

$$\begin{aligned}&\operatorname{argmin}\left\{p\|u - q\|_{L^2(f)} + \|u - q\|_{H^1(f)} : q \in \mathbb{P}^{p-1}\right\}\\ &\quad \neq (A^{-1})^* \operatorname{argmin}\left\{p\|A^*u - q\|_{L^2(A^{-1}(f))} + \|A^*u - q\|_{H^1(A^{-1}(f))} : q \in \mathbb{P}^{p-1}\right\}\end{aligned}$$

for an affine isomorphism A. (We can see this in one dimension, e.g. from the fact that $\|u\|_{L^2}$ and $|u|_{H^1}$ scale differently.) To deal with this issue, we average minimizations similar to (A.17) over all possible boundary faces (see (A.37) below).

We prove Proposition A.21 in Section A.8, but we discuss the ideas used in the proof here.

Idea of the proof of Proposition A.21. To modify $\mathcal{J}^{p-1}$ into a boundary-compatible operator, we follow the steps from Melenk and Sauter (2010, Appendix B). Let $e_d = (I - \mathcal{J}^{p-1})u$. For $j = d, d-1, \ldots, 1$, let g_j be a polynomial of degree p such that on each codimension j face of K, $F_{j,i}$, $g_j|_{F_{j,i}} =: \tilde{T}_{j,i}(e_j|_{F_{j,i}})$, for some operator $T_{j,i}$, is a good polynomial approximant of $e_j|_{F_{j,i}}$ with zero boundary conditions

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

on $\partial F_{j,i}$ (in the case $j = d$, there are no boundary conditions and the polynomial g_d matches e_d exactly at each vertex), and such that

$$\tilde{T}_{j,i}u = (\kappa^{-1})^*(\tilde{T}_{j,i'}(\kappa^* u)) \tag{A.18}$$

for any affine isomorphism κ that sends $F_{j,i'}$ to $F_{j,i}$ (compare to (A.15)). Set $e_{j-1} \coloneqq e_j - g_j$. Then the desired boundary-compatible operator $\mathcal{C}^{p-1} = u - e_0$.

The key ingredient in this procedure is the construction of g_j. To achieve this, we take the average of the best $H^{1/2}$-approximations of $\kappa^*(e_j|_{F_{j,i}})$ (among polynomials with zero boundary conditions), where κ ranges over all affine isomorphisms mapping $F_{j,i'}$ to $F_{j,i}$ for some i' (strictly speaking, pulled back to $F_{j,i}$ using $(\kappa^{-1})^*$), and extend it to all of K with good enough H^1-estimates. This is done in Section A.7.3 following ideas from Muñoz-Sola (1997) and Bernardi, Dauge and Maday (1995) to find a p-independent polynomial extension operator that is bounded from $H^{1/2}$ of a codimension 1 boundary face to H^1 of a simplex (see Lemma A.26). Using an induction procedure together with some ideas from Melenk and Sauter (2010, Appendix B), we use this extension to construct a p-dependent extension operator from any boundary face into H^1 of all lower-codimension boundary faces having improved estimates as a function of p (see Lemma A.27). One of the key elements of this last step is a polynomial approximation lemma that respects zero boundary conditions (see Lemma A.23).

A.7.2. Polynomial approximation in simplices respecting boundary conditions
Before constructing boundary-compatible polynomial approximation operators (via Proposition A.21), we need one further lemma on polynomial approximation specifically in simplices. This lemma shows that one can both accurately approximate H^2-functions (in H^1) with high-degree polynomials *and* respect boundary conditions. The proof of the lemma is similar to that of Proposition A.18, but we use the associated Legendre polynomials of order two, i.e. eigenfunctions of $-\partial_x(1-x^2)\partial_x + 4(1-x^2)^{-1}$ on $[-1,1]$, when we need to preserve boundary data.

Lemma A.23. Suppose that $K \subset \mathbb{R}^d$ is an open d-simplex and $0 \le m \le d$. Then there exists $C > 0$ such that for all $p \ge 1$, there exists $\mathcal{A}^{p-1}\colon \tilde{H}^1_{m,0}(K) \to \mathbb{P}^{p-1} \cap \tilde{H}^1_{m,0}(K)$ satisfying

$$\|(I - \mathcal{A}^{p-1})u\|_{H^1(K)} \le C p^{-1} \|u\|_{H^2(K)},$$

where

$$\tilde{H}^1_{m,0}(K) \coloneqq \{u \in H^1(K)\colon u|_{F_{1,j}} = 0,\ j = 0, \dots, m\}.$$

Proof. The proof will be similar to that of Proposition A.18 except that we change the operator P to impose boundary conditions. To do this, we use the reflection symmetry of the operator

$$-\partial_x(1-x^2)\partial_x + \frac{4}{1-x^2}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

(which has polynomial eigenfunctions on $[-1,1]$) to find an operator with certain Dirichlet boundary conditions on the faces of a simplex that still has polynomial eigenfunctions.

Without loss of generality, we can assume that

$$K = \{x \in \mathbb{R}^d : \ -1 < x_1 < x_2 < \cdots < x_d < 1\}$$

and

$$F_{1,0} \subset \{x_1 = -1\}, \quad F_{1,j} \subset \{x_{j-1} < x_j\},\ j = 1, \dots, d-1,\ F_{1,d} \subset \{x_d = 1\}.$$

Define also

$$\tilde{K} := \{x \in \mathbb{R}^m : \ -1 < x_1 < x_2 < \cdots < x_m < 1\} \times (-1,1)^{d-m}.$$

Let $\tilde{E} : C^\infty(K) \to C^\infty(\tilde{K})$ such that $1_K \tilde{E} = I$, and $\tilde{E} : H^s(K) \to H^s(\tilde{K})$ is bounded for all $s \geq 0$.

For $\sigma : \{1, \dots, m\} \to \{1, \dots, m\}$ a permutation, define

$$\tilde{K}^\sigma := \{x \in \mathbb{R}^m : \ -1 < x_{\sigma(1)} < x_{\sigma(2)} < \cdots < x_{\sigma(m)} < 1\} \times (-1,1)^{d-m},$$

and

$$\iota_\sigma(x_1, \dots, x_d) = (x_{\sigma(1)}, \dots, x_{\sigma(m)}, x_{m+1}, \dots, x_d).$$

Let $\operatorname{sgn}(\sigma)$ denote the signature of a permutation. Let

$$H^1_m(\tilde{K}) := \left\{u \in H^1(\tilde{K}) : u|_{x_1=-1} = 0,\ u|_{x_{j-1}=x_j} = 0,\ j = 1, \dots, m\right\},$$
$$H^1_m([-1,1]^d) := \left\{u \in H^1([-1,1]^d) : u|_{x_j=\pm 1} = 0,\ j = 1, \dots, m\right\}.$$

We then define an extension operator $E : H^1_{m,0}(K) \to H^1_{m,0}([-1,1]^d)$ by

$$(Eu)(x) = (-1)^{\operatorname{sgn}(\sigma)} \tilde{E} u(\iota_\sigma^{-1}(x)), \quad x \in \overline{\tilde{K}^\sigma}.$$

Note also that

$$E : H^1_{m,0}(K) \cap H^2(K) \to H^1_{m,0}([-1,1]^d) \cap H^2([-1,1]^d).$$

We now define two self-adjoint operators, one on $L^2(\tilde{K})$ and one on $L^2([-1,1]^d)$. Define the sesquilinear forms $Q : H^1_{m,0}([-1,1]^d) \times H^1_{m,0}([-1,1]^d) \to \mathbb{C}$ and $Q^{\tilde{K}} : H^1_{m,0}(\tilde{K}) \times H^1_{m,0}(\tilde{K}) \to \mathbb{C}$ by

$$Q(u,v) := \sum_{i=1}^d \langle (1-x_i^2)\partial_{x_i} u, \partial_{x_i} v\rangle_{L^2([-1,1]^d)} + 4 \sum_{i=1}^m \left\langle \tfrac{1}{(1-x_i^2)} u, v \right\rangle_{L^2([-1,1]^d)},$$
$$Q^{\tilde{K}}(u,v) := \sum_{i=1}^d \langle (1-x_i^2)\partial_{x_i} u, \partial_{x_i} v\rangle_{L^2(\tilde{K})} + 4 \sum_{i=1}^m \left\langle \tfrac{1}{(1-x_i^2)} u, v \right\rangle_{L^2(\tilde{K})}.$$

Then, using the Hardy inequality,

$$\int_0^a \left| \frac{1}{t} \int_x^{x+t} f(s)\, ds \right|^2 dt \leq 4 \int_0^a |f(x+t)|^2\, dt, \tag{A.19}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

for the upper bounds,

$$0 \le Q(u,u), \qquad |Q(u,v)| \le C\|u\|_{H^1([-1,1]^d)}\|v\|_{H^1([-1,1]^d)},$$
$$0 \le Q^{\tilde{K}}(u,u), \quad |Q^{\tilde{K}}(u,v)| \le C\|u\|_{H^1(\tilde{K})}\|v\|_{H^1(\tilde{K})},$$

and hence the Friedrichs extensions of Q and $Q^{\tilde{K}}$, respectively $P\colon L^2([-1,1]^d) \to L^2([-1,1]^d)$ and $P^{\tilde{K}}\colon L^2(\tilde{K}) \to L^2(\tilde{K})$, are self-adjoint with domains

$$\mathcal{D}(P) := \left\{u \in H^1_{m,0}([-1,1]^d)\colon |Q(u,v)| \le C_u\|v\|_{L^2} \text{ for all } v \in H^1_{m,0}([-1,1]^d)\right\},$$
$$\mathcal{D}(P^{\tilde{K}}) := \left\{u \in H^1_{m,0}(K)\colon |Q^{\tilde{K}}(u,v)| \le C_u\|v\|_{L^2} \text{ for all } v \in H^1_{m,0}(\tilde{K})\right\}.$$

Notice that

$$H^2([-1,1]^d) \cap H^1_{m,0}([-1,1]^d) \subset \mathcal{D}(P)$$

and, for $u \in H^2 \cap H^1_{m,0}([-1,1]^d$,

$$Pu = -\sum_{i=1}^{d} \partial_{x_i}(1-x_i^2)\partial_{x_i}u + \sum_{i=1}^{m} \tfrac{4}{(1-x_i^2)}u.$$

Let $p^{(2)}_{n+2}\colon [-1,1] \to \mathbb{R}$, $n = 0,1,\dots$ be the associated Legendre polynomials of order 2 (see e.g. DLMF, §14.2(ii)) normalized such that

$$\left\langle p^{(2)}_{n+2}, p^{(2)}_{l+2}\right\rangle_{L^2([-1,1])} = \delta_{nl},$$

and let $p_n\colon [-1,1] \to \mathbb{R}$, $n = 0,1,\dots$ be the Legendre polynomials normalized such that

$$\langle p_n, p_l\rangle_{L^2([-1,1])} = \delta_{nl}.$$

Then, for $\vec{n} \in \mathbb{N}^d$, define $p_{\vec{n}}\colon [-1,1]^d \to \mathbb{R}$ by

$$p_{\vec{n}}(x) := \prod_{i=m+1}^{d} p_{\vec{n}_i}(x_i) \prod_{j=1}^{m} p^{(2)}_{\vec{n}_j+2}(x_j),$$

so that by the definition of $p^{(2)}_{n+2}$ and p_n,

$$\left(P - \sum_{j=1}^{m}(\vec{n}_j+2)(\vec{n}_j+3) - \sum_{i=1}^{m+1}(\vec{n}_i)(\vec{n}_i+1)\right)p_{\vec{n}} = 0.$$

We claim that the associated Legendre polynomials of order 2 are complete. Indeed, observe that by DLMF, (14.7.8),

$$p^{(2)}_{n+2} = (1-x^2)\partial_x^2 p_{n+2},$$

and hence, if $u \in L^2$ is orthogonal to $p^{(2)}_{n+2}$ for all n, then

$$\langle u, (1-x^2)\partial_x^2 p_{n+2}\rangle_{[-1,1]} = 0 \quad \text{for all } n.$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

But, since p_n is a polynomial of degree n, $\operatorname{span}\{\partial_x^2 p_{n+2}\}_{n=0}^{\infty}$ contains all polynomials and hence is dense in L^2. In particular, this implies that $(1-x^2)u = 0$ and hence $u = 0$.

Since both the Legendre polynomials and the associated Legendre polynomials of order 2 are complete, $\{p_{\vec{n}}\}_{\vec{n}\in\mathbb{N}^d}$ forms an orthonormal basis of eigenfunctions of P for $L^2([-1,1]^d)$.

Now, suppose that $P^{\tilde{K}}u = \lambda u$. By the definitions of E, Q and $Q^{\tilde{K}}$, $PEu = \lambda Eu$ and hence Eu is an eigenfunction of P with eigenvalue λ. This implies that there exists $c > 0$ such that

$$1_{(-\infty,cp^2]}(P^{\tilde{K}})\Pi_{p-1} = 1_{(-\infty,cp^2]}(P^{\tilde{K}}),$$

where $\Pi_{p-1}\colon L^2(\tilde{K}) \to \mathbb{P}^{p-1}\cap H_0^1(\tilde{K})$ is the orthogonal projector, $[\Pi_{p-1}, P^{\tilde{K}}] = 0$ (this is very similar to the proof of Proposition A.18). Let $\mathcal{A}^{p-1} := 1_K\Pi_{p-1}\tilde{E}$. Then

$$\begin{aligned}(I-\mathcal{A}^{p-1}) &= 1_K(I-\Pi_{p-1})\tilde{E}\\ &= 1_K(P^{\tilde{K}}+1)^{-1}1_{(cp^2,\infty)}(P^K)(I-\Pi_{p-1})(P^K+1)\tilde{E}.\end{aligned}$$

Now, $(P^{\tilde{K}}+1)^{-1/2}\colon L^2(\tilde{K}) \to H^1_{m,0}(\tilde{K})$, $(P^{\tilde{K}}+1)\colon H^1_{m,0}(\tilde{K})\cap H^2(\tilde{K}) \to L^2$ and

$$\|(P^{\tilde{K}}+1)^{-1/2}1_{(cp^2,\infty)}\|_{L^2\to L^2} \le c\langle p\rangle^{-1}.$$

Therefore

$$\begin{aligned}&\|(I-\mathcal{A}^{p-1})\|_{H^2(K)\cap H^1_{m,0}(K)\to H^1_{m,0}(K)}\\ &\quad\le \|(P^K+1)^{-1/2}1_{(cp^2,\infty)}(P)\|_{L^2\to L^2}\|(P^K+1)\chi E\|_{H^2(K)\cap H^1_{m,0}(K)\to L^2}\\ &\quad\le C\langle p\rangle^{-1}.\end{aligned}$$

□

A.7.3. Extension operators

To extend polynomials effectively from faces into a simplex K, we need to construct extension operators that are bounded from $H^{1/2}$ to H^1, respect traces, and preserve the class of polynomials of a fixed degree.

We start by adapting the arguments from Muñoz-Sola (1997, Section 2) to construct a p-independent extension operator. Let $Q = (0,1)^d$, $Q_I := (0,1)^d \times (0,\frac{1}{2})$, $\chi \in L^\infty(Q)$ with $\int \chi = 1$ and define the operator $E_\chi\colon C^\infty(Q) \to L^\infty(Q_I)$ by

$$(E_\chi u)(x,t) := \int u((1-t)x+ty)\chi(y)\,dy.$$

Lemma A.24. Let $d \ge 1$. Then there exists $C > 0$ such that for all $\chi \in L^\infty$ with $\int \chi = 1$ and $u \in C^\infty(\overline{Q})$,

$$\|E_\chi u\|_{L^2(Q_I)} \le C\|\chi\|_{L^\infty}\min_i \|x_i^{1/2}u\|_{L^2(Q)}. \tag{A.20}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Proof. To prove (A.20), we change variables so that

$$|E_\chi u(x,t)| = \left| t^{-d} \int u(w)\chi(t^{-1}(w-(1-t)x))\,dw \right|$$
$$\leq \|\chi\|_{L^\infty} t^{-d} \int |u(w)| 1_{tQ}(w-(1-t)x)\,dw.$$

We prove (A.20) with $i = 1$; the other estimates come from a permutation of coordinates. Then, by the Hardy inequality (A.19) in the seventh line and the L^2-boundedness of the Hardy–Littlewood maximal function

$$M_{d-1}f(z) := \sup_{t>0} \left| \frac{1}{t^{d-1}} \int_{tQ_{d-1}} f(z-w)\,dw \right|$$

in the eighth line,

$$\begin{aligned}
&\|F_\Omega u\|^2_{L^2(Q_I)} \\
&\leq \int_0^{1/2} \int_Q \left| t^{-d}\|\chi\|_{L^\infty} \int |u(w)| 1_{tQ}(w-(1-t)x)\,dw \right|^2 dx\,dt \\
&= \int_0^{1/2} \int_{(1-t)Q} (1-t)^{-d} \left| t^{-d}\|\chi\|_{L^\infty} \int |u(w)| 1_{tQ}(w-z)\,dw \right|^2 dz\,dt \\
&= 2^d \|\chi\|^2_{L^\infty} \\
&\quad \times \int_Q \int_0^{\min(\min_i(1-z_i),\frac{1}{2})} \left| t^{-1} \int_{z_1}^{z_1+t} t^{-d+1} \int |u(w)| 1_{tQ_{d-1}}(w'-z')\,dw'\,dw_1 \right|^2 dt\,dz \\
&= 2^d \|\chi\|^2_{L^\infty} \int_Q \int_0^{\min(\min_i(1-z_i),\frac{1}{2})} \left| t^{-1} \int_{z_1}^{z_1+t} \int M_{d-1}|u(w_1,\bullet)|(z')\,dw_1 \right|^2 dt\,dz \\
&= 2^{d+2} \|\chi\|^2_{L^\infty} \int_Q \int_0^{\min(\min_i(1-z_i),\frac{1}{2})} \left| \int M_{d-1}|u(z_1+t,\bullet)|(z') \right|^2 dt\,dz \\
&= 2^{d+2} \|\chi\|^2_{L^\infty} \int_0^1 \int_0^{\min(1-z_1,\frac{1}{2})} \int_{Q_{d-1}} \left| \int M_{d-1}|u(z_1+t,\bullet)|(z') \right|^2 dz'\,dt\,dz_1 \\
&\leq C 2^{d+2} \|\chi\|^2_{L^\infty} \int_0^1 \int_0^{\min(1-z_1,\frac{1}{2})} \|u(z_1+t,\bullet)\|^2_{L^2_{z'}}\,dt\,dz_1 \\
&= C 2^{d+2} \|\chi\|^2_{L^\infty} \int_0^{1/2} \int_{z_1}^{\min(1,\frac{1}{2}+z_1)} \|u(s,\bullet)\|^2_{L^2_{z'}}\,ds\,dz_1 \\
&\leq C 2^{d+2} \|\chi\|^2_{L^\infty} \int_0^1 \int_0^s \|u(s,\bullet)\|^2_{L^2_{z'}}\,dz_1\,ds \leq C 2^{d+2} \|\chi\|^2_{L^\infty} \|x_1^{1/2} u\|^2_{L^2}.
\end{aligned}$$

□

Lemma A.25. Let $\Omega \subset Q$ be convex and $\chi \in C^1_{\text{comp}}(\Omega)$. Then there exists $C > 0$ such that

$$\|E_\chi u\|_{H^1(\Omega\times(0,1/2)} \leq C\|u\|_{H^{1/2}(\Omega)} \tag{A.21}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

and, for any $1 \le j \le d-1$,

$$\left\| \prod_{i=1}^{j} x_i \left[E_\chi u(\bullet) \prod_{i=1}^{j} (\bullet)_i^{-1} \right](x,t) \right\|_{H^1(\Omega\times(0,1/2))} \le C\left(\sum_{i=1}^{j} \|x_i^{-1/2} u\|_{L^2(\Omega)} + \|u\|_{H^{1/2}(\Omega)} \right). \tag{A.22}$$

Proof. We begin by proving (A.21) and first change variables to $z = (1-t)x$. (Since $0 \le t \le \frac{1}{2}$, this is a diffeomorphism.) By (A.20), it remains to estimate $\partial_{z_i} E_\chi u$ and $\partial_t E_\chi u$ in L^2. Changing variables as in Lemma A.24, and setting $z = (1-t)x$, we have

$$\begin{aligned} \partial_{z_i} E_\chi &= t^{-d-1} \int u(w) \partial_{z_i} \chi(t^{-1}(w-z))\, dw \\ &= t^{-d-1} \int (u(w)-u(z))(\partial_{z_i} \chi(t^{-1}(w-z))\, dw \\ &= t^{-\frac{d+1}{2}} \int \frac{u(w)-u(z)}{|z-w|^{\frac{d+1}{2}}} t^{-\frac{d+1}{2}} |w-z|^{\frac{d+1}{2}} \partial_{z_i} \chi(t^{-1}(w-z))\, dw. \end{aligned}$$

Hence, letting $\tilde\chi \in C_c^\infty(\Omega)$ with $\operatorname{supp}(1-\tilde\chi) \cap \operatorname{supp}\chi = \emptyset$ and using the Cauchy–Schwarz inequality,

$$\begin{aligned} &\|\partial_{z_i} E_\chi u\|_{L^2}^2 \\ &\le \int_0^{1/2} \int t^{-d-1} \left| \int \frac{u(w)-u(z)}{|z-w|^{\frac{d+1}{2}}} t^{-\frac{d+1}{2}} |w-z|^{\frac{d+1}{2}} \partial_{z_i} \chi(t^{-1}(w-z))\, dw \right|^2 dz\, dt \\ &\le \int_0^{1/2} \int t^{-d-1} \left(\int \frac{|u(w)-u(z)|^2}{|z-w|^{d+1}} \tilde\chi(t^{-1}(w-z))\, dw \right. \\ &\qquad \left. \times \int t^{-d+1} |w-z|^{d+1} |\partial_{z_i} \chi(t^{-1}(w-z))|^2\, dw \right) dz\, dt \\ &\le C \int_0^{1/2} \int t^{-1} \int \frac{|u(w)-u(z)|^2}{|z-w|^{d+1}} \tilde\chi(t^{-1}(w-z))\, dw\, dz\, dt \\ &\le C \int \int \frac{|u(w)-u(z)|^2}{|z-w|^{d+1}} \int_{c|z-w|}^{C\min(|z-w|,\frac{1}{2})} t^{-1}\, dt\, dw\, dz \\ &\le C \int \int \frac{|u(w)-u(z)|^2}{|z-w|^{d+1}}\, dw\, dz \le C\|u\|_{H^{1/2}}^2 \end{aligned} \tag{A.23}$$

by the definition of the Slobodeckij seminorm (see e.g. McLean 2000, equation

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

3.18). Next, consider

$$\begin{aligned}
&\partial_t E_\chi u(z,t)\\
&\quad = \partial_t (E_\chi u(z,t) - u(z))\\
&\quad = -t^{-d-1}\\
&\qquad \times \int (u(w)-u(z))\left(d\chi(t^{-1}(w-z)) - \sum_i t^{-1}(w_i - z_i)\partial_{z_i}\chi(t^{-1}(w-z))\right) dw.
\end{aligned}$$

Now, since $\|s_i \partial_{z_i}\chi(s)\|_{L^\infty} < \infty$ and $\operatorname{supp}\chi \subset \Omega$, the proof that $\|\partial_t E_\chi u\| \le C\|u\|_{H^{1/2}}$ follows from the same argument leading to (A.23).

We now prove (A.22). Let

$$v(x,t) := \prod_{i=1}^{j} x_i \left[E_\chi u(\bullet) \prod_{i=1}^{j} (\bullet)_i^{-1} \right](x,t)$$

and first observe that for $x \in Q$,

$$|v(x,t)| \le E_\chi(|u|)(x,t).$$

Hence, by (A.20),

$$\|v\|_{L^2(Q_I)} \le \|E_\chi(|u|)\|_{L^2(Q_I)} \le C \min_i \|x_i^{1/2} u\|_{L^2(Q)}.$$

Thus it remains to estimate the derivatives of v.

First we estimate $\partial_{z_\ell} v$ for $1 \le \ell \le j$. In fact, up to a permutation of indices, it is enough to estimate $\partial_{z_1} v$. Then

$$\begin{aligned}
&\partial_{z_1} v\\
&\quad = \frac{1}{[(1-t)t]^d} \int u(w) \prod_{i=2}^{j} \frac{z_i}{w_i} \left(w_1^{-1}\chi(t^{-1}(w-z)) + \frac{z_1}{w_1} t^{-1} \partial_{z_1}\chi(t^{-1}(w-z))\right) dw\\
&\quad =: I + II.
\end{aligned}$$

For I, on $\operatorname{supp}\chi(t^{-1}(w-z))$, $w_i \ge z_i$, and hence

$$|I| \le |E_\chi |u(\bullet_1)^{-1}|(z,t)|.$$

Hence (A.20) implies

$$\|I\|_{L^2} \le C\|x_1^{-1/2} u\|_{L^2}.$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

For II,

$$\begin{aligned} II &= \frac{1}{[(1-t)t]^d} \int \left[(u(w)-u(z)) \prod_{i=1}^{j} \frac{z_i}{w_i} \right. \\ &\quad \left. + u(z) \left(\prod_{i=1}^{j} \frac{z_i}{w_i} - 1 \right) \right] t^{-1} \partial_{z_1} \chi(t^{-1}(w-z))\, dw \\ &=: II_a + II_b. \end{aligned}$$

Since $|z_i/w_i| \le 1$ on the support of the integrand, II_a can be treated as in (A.23) to obtain

$$\|II_a\|_{L^2} \le C \|u\|_{H^{1/2}}.$$

For II_b, observe that the integrand is supported on $z_\ell \le w_\ell \le z_\ell + t$ for $1 \le \ell \le d$, and hence

$$\begin{aligned} \int \left| \frac{z_i}{w_i} - 1 \right| t^{-1} |\partial_{z_1} \chi(t^{-1}(w-z))| &\le C t^{d-2} \int_{z_i}^{z_i+t} 1 - \frac{z_i}{w_i}\, dw_i \\ &= C t^{d-2} \left(t - z_i \log\left(1 + \frac{t}{z_i} \right) \right) \\ &\le C t^{d-2} \min(t^2 z_i^{-1}, t). \end{aligned}$$

Therefore, since

$$\begin{aligned} \left| \prod_{i=1}^{j} \frac{z_i}{w_i} - 1 \right| &\le \sum_{\ell=1}^{j-1} \left| \left(\frac{z_\ell}{w_\ell} - 1 \right) \prod_{i=\ell+1}^{j} \frac{z_i}{w_i} \right| + \left| \frac{z_j}{w_j} - 1 \right| \\ &\le \sum_{i=1}^{j} \left| \frac{z_j}{w_j} - 1 \right|, \end{aligned}$$

we have

$$\|II_b\|_{L^2}^2 \le C \sum_{i=1}^{j} \int |u(z)|^2 \int_0^{1/2} \min(z_i^{-2}, t^{-2})\, dt\, dz \le C \sum_{i=1}^{j} \|z_i^{-1/2} u\|_{L^2}^2.$$

The combination of the bounds on I, II_a and II_b implies that

$$\|\partial_{z_1} v\|_{L^2} \le C \sum_{i=1}^{j} \|z_i^{-1/2} u\|_{L^2} + \|u\|_{H^{1/2}}.$$

Now, for $\ell > j$, we have only terms of the form II and hence the bounds there imply

$$\|\partial_{z_\ell} v\|_{L^2} \le C \sum_{i=1}^{j} \|z_i^{-1/2} u\|_{L^2} + \|u\|_{H^{1/2}}.$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Finally,

$$\begin{aligned}
&\partial_t\left[(1-t)^{-d}v\right]\\
&\quad=\partial_t\left(t^{-d}\int\left(u(w)\prod_{i=1}^{j}\frac{z_i}{w_i}-u(z)\right)\chi(t^{-1}(w-z))\,dw\right.\\
&\quad=\int\left((u(w)-u(z))\prod_{i=1}^{j}\frac{z_i}{w_i}+u(z)\left(\prod_{i=1}^{j}\frac{z_i}{w_i}-1\right)\right)\partial_t(t^{-d}\chi(t^{-1}(w-z)))\,dw,
\end{aligned}$$

and each term can be estimated as above (as in the estimate for $\partial_t E_\chi$, we use that the term arising from differentiating $\chi(t^{-1}(w-z))$ in t has an extra power of $|z-w|$.) □

Lemma A.26 (Extension operators: version 1). Let $K\subset\mathbb{R}^d$ be an open d-simplex. Then there is an extension operator $E\colon H^{1/2}(\partial K)\to H^1(K)$ and $C>0$ such that for all $p\geq 0$, $E\colon \mathbb{P}^p(\partial K)\to\mathbb{P}^p(K)$ and

$$E|_{\partial K}=I,\quad \|Eu\|_{H^1(K)}\leq C\|u\|_{H^{1/2}(\partial K)},$$

where

$$\|u\|_{H^s(\partial K)}:=\inf\left\{\sum_i\|u|_{F_{1,i}}\|_{H^s(F_{1,i})}\colon U\in H^{s+1/2}(K)\colon U|_{\partial K}=u\right\}.$$

Proof. Without loss of generality, we can assume that

$$K=\left\{x\in\mathbb{R}^d\colon 0<x_i,\ \textstyle\sum_{i=1}^d x_i<1\right\}$$

and

$$F_0:=F_{1,1}=\left\{x\in\mathbb{R}^{d-1}\colon x_d=0,\ 0<x_i,\ \textstyle\sum_{i=1}^{d-1}x_i<1\right\}$$

There is an invertible affine map taking $\tilde\Phi\colon\overline{K}\to\overline{K}$ so that

$$\tilde\Phi(F_0)=\left\{x\in\mathbb{R}^{d-1}\colon 0<x_i,\ \textstyle\sum_{i=1}^d x_i=1\right\}=:F.$$

Fix $\chi\in C_c^1(F)$, and for $u\in C^\infty(\overline{F})$ set

$$\tilde Eu(x',x_d):=[E_\chi u]\left(\left(1-\tfrac12 x_d\right)^{-1}x',\tfrac12 x_d\right)$$

and observe that

$$K\subset\left\{(x',x_d)\colon\left(1-\tfrac12 x_d\right)^{-1}x'\in F,\ 0<x_d<1\right\}.$$

Then define

$$E_0:=\tilde\Phi^*\tilde E(\tilde\Phi^{-1})^*,$$

where, for a map $g\colon\overline{K}\to\overline{K}$, we let g^* denote the pull-back, i.e. $g^*u(x):=u(g(x))$. Then, for all $p\geq 0$, $E_0\colon H^{1/2}(F_{1,1})\to H^1(K)$, $E_0\colon\mathbb{P}^p(F_{1,1})\to\mathbb{P}^p(K)$ and

$$\|E_0u\|_{H^1(K)}\leq C\|u\|_{H^{1/2}(F_{1,1})},\quad E_0|_{F_{1,1}}=I.$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Suppose by induction that for some $0 \le j < d$ we have constructed E_j such that

$$E_j \colon H^1(\partial K) \cap C^\infty(\partial K) \to C^\infty(\overline{K}), \tag{A.24}$$

$$E_j \colon \mathbb{P}^p(\partial K) \to \mathbb{P}^p(K) \quad \text{for all } p \ge 0, \tag{A.25}$$

$$\|E_j u\|_{H^1(K)} \le C\|u\|_{H^{1/2}(\partial K)}, \tag{A.26}$$

$$E_j|_{\cup_{1\le i\le j+1} F_{1,i}} = I. \tag{A.27}$$

Here we write $C^\infty(\partial K)$ for the restriction to ∂K of $C^\infty(\overline{K})$.

We now construct E_{j+1} satisfying (A.24) to (A.27) with j replaced by $j+1$.

Let $\Phi_j \colon \overline{K} \to \overline{K}$ be an invertible affine map such that

$$\Phi_j(F_{1,j}) = F_0, \quad \Phi_j(F_{1,i}) \subset \{x_i = 0\}, \quad i = 1, \dots, j+1.$$

Then define

$$E_{j+1}u \coloneqq E_j u - \Phi_j^*\big(x_1 \cdots x_{j+1} E_0\big[\big(((\Phi_j^{-1})^*(E_j u - u))/x_1 \cdots x_{j+1}\big)|_{F_0}\big]\big)$$

We first show (A.24) for E_{j+1}, i.e. $E_{j+1} \colon C^\infty(\partial K) \to C^\infty(\overline{K})$. Indeed, observe that for $u \in C^\infty(\partial K)$, $E_j u \in C^\infty(\overline{K})$ and $(\Phi_j^{-1})^* E_j u|_{x_i=0} = (\Phi_j^{-1})^* u|_{x_i=0}$ for $i = 1, \dots, j+1$. Hence $x_1^{-1} \cdots x_{j+1}^{-1}(\Phi_j^{-1})^*(E_j u - u) \in C^\infty(\partial K)$, which implies $E_{j+1}u \in C^\infty(\overline{K})$.

Next, we show (A.25) for E_{j+1}, i.e. $E_{j+1} \colon \mathbb{P}^p(\partial K) \to \mathbb{P}^p(K)$. Indeed, observe that for $u \in \mathbb{P}^p(\partial K)$, $E_j u \in \mathbb{P}^p(K)$ and $(\Phi_j^{-1})^*(E_j u - u)|_{x_i=0} = 0$ for $i = 1, \dots, j+1$. Hence $x_1^{-1} \cdots x_{j+1}^{-1}(\Phi_j^{-1})^*(E_j u - u)|_{\partial K} \in \mathbb{P}^{p-j-1}$, which implies $E_{j+1}u \in \mathbb{P}^p(K)$ since $E_0 \colon \mathbb{P}^{p-j-1}(F_0) \to \mathbb{P}^{p-j-1}(K)$.

Next, we claim that (A.26) and (A.27) hold, that is,

$$\|E_{j+1}u\|_{H^1(K)} \le C\|u\|_{H^{1/2}(\partial K)}, \tag{A.28}$$

$$E_{j+1}|_{F_{1,i}} = I, \quad 1 \le i \le j+2. \tag{A.29}$$

To see (A.28), first observe that by (A.22),

$$\begin{aligned}&\big\|\Phi_j^*\big(x_1 \cdots x_{j+1} E_0\big[((w)/x_1 \cdots x_{j+1})|_{F_0}\big]\big)\big\|_{H^1(K)}\\ &\quad\le C\Bigg(\sum_{i=0}^{j+1} \|x_i^{-1/2} w\|_{L^2(F_0)} + \|w\|_{H^{1/2}(F_0)}\Bigg).\end{aligned}$$

Therefore, to prove (A.28), we need to check that

$$\sum_{i=0}^{j+1} \|x_i^{-1/2}(\Phi_j^{-1})^*(E_j u - u)\|_{L^2(F_0)} + \|(\Phi_j^{-1})^*(E_j u - u)\|_{H^{1/2}(F_0)} \le C\|u\|_{H^{1/2}(\partial K)}. \tag{A.30}$$

For this, let $u \in H^{1/2}(\partial K)$ and let $u_n \in C^\infty(\partial K)$ with $u_n \overset{H^{1/2}}{\to} u$. The boundedness of E_j implies that $(\Phi_j^{-1})^* E_j u_n \to (\Phi_j^{-1})^* u$ in $H^1(K)$.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Then $(\Phi_j^{-1})^* E_j u_n \in C^\infty(K)$ and $(\Phi_j^{-1})^*(E_j u_n - u_n)|_{\{x_i=0\}\cap \partial K} = 0$ for $1 \le i \le j+1$. Moreover,

$$\|(\Phi_j^{-1})^* E_j (u_n - u)\|_{H^{1/2}(\partial K)} \le C \|(\Phi_j^{-1})^* E_j (u_n - u)\|_{H^1(K)} \le C \|u_n - u\|_{H^{1/2}(\partial K)}.$$

In particular, $[(\Phi_j^{-1})^*(E_j u_n - u_n)]|_{F_0}$ converges to

$$[(\Phi_j^{-1})^*(E_j u - u)]|_{F_0} \text{ in } H^{1/2}(\partial K).$$

Next, observe that for $w \in H^1(\partial K)$ with $w \in H_0^1(F_0)$,

$$\begin{aligned} |w|^2_{H^{1/2}(\partial K)} &\ge c \int_{F_0} \int_{\Phi_j(F_{1,i})} \frac{|w(x)|^2}{|x_i^2 + |x' - y'|^2 + y_d^2|^{(d+1)/2}} \, dy_d \, dy' \, dx' \, dx_i \\ &\ge c \|x_i^{-1/2} w\|^2_{L^2}. \end{aligned}$$

Since $(\Phi_j^{-1})^*(E_j u_n - u_n)|_{\{x_i=0\}} = 0$, $i = 1, \dots, j+1$, and $u_n \in H^1(\partial K)$, this implies

$$\begin{aligned} &\sum_{i=0}^{j+1} \|x_i^{-1/2} (\Phi_j^{-1})^* [E_j(u_n - u_m) - u_n + u_m\|_{L^2(F_0)} \\ &\quad \le C \|(\Phi_j^{-1})^* [E_j(u_n - u_m) - u_n + u_m]\|_{H^{1/2}(F_0)}, \end{aligned}$$

This implies that for $i = 1, \dots, j+1$, $x_i^{-1/2}(\Phi_j^{-1})^* E_j u \in L^2$ and

$$\sum_{i=0}^{j+1} \|x_i^{-1/2} (\Phi_j^{-1})^* E_j u\|_{L^2(F_0)} + \|(\Phi_j^{-1})^* E_j u\|_{H^{1/2}(F_0)} \le C \|u\|_{H^{1/2}(\partial K)},$$

which is (A.30), and hence we have proved (A.28).

The equalities (A.29) follow easily for $u \in C^\infty(\partial K)$ and follow for $u \in H^{1/2}(\partial K)$ by density. The proof is now complete by induction. □

Lemma A.27 (Extension operators: version 2). Let $K \subset \mathbb{R}^d$ be an open d-simplex. Then there exists $C > 0$ such that for all $p \ge 0$, there are extension operators $E_{j,i} \colon \mathbb{P}^p(F_{j,i}) \cap H_0^1(F_{j,i}) \to \mathbb{P}^p(K)$, $1 \le j < d$, and $E_{d,i} \colon \mathbb{C} \to \mathbb{P}^p(K)$ such that

$$\begin{aligned} &E_{j,i}|_{\overline{F_{j,i}}} = I, \quad E_{j,i}|_{\overline{F_{j,m}}} = 0, \ m \ne i, \\ &E_{j,i}|_{\overline{F_{r,m}}} = 0, \quad \overline{F_{r,m}} \cap \overline{F_{j,i}} = \emptyset, \\ &\|E_{j,i}|_{F_{r,l}} q\|_{H_p^1(F_{r,l})} \le C p^{(r-j)/2} \|q\|_{H_p^{1/2}(F_{j,i})}, \quad 0 \le r < j. \end{aligned}$$

where

$$\|u\|_{H_p^1(F)} \coloneqq \|u\|_{L^2} + \langle p \rangle^{-1} |u|_{H^1},$$

Proof. We claim that it is enough to show that for an open $(m+1)$-simplex $K_m \subset \mathbb{R}^m$, there exists $E_m \colon \mathbb{P}^p(\partial K_m) \to \mathbb{P}^p(K_m)$ such that

$$E_m^{K_m}|_{\partial K_m} = I \tag{A.31}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

and

$$\|E_m^{K_m} u\|_{H_p^1(K_m)} \leq C\langle p\rangle^{-1/2} \|u\|_{H_p^{1/2}(\partial K)}. \tag{A.32}$$

Indeed, suppose such an operator exists for any K_m.

When $d = 1$, $E_m^{K_m}$ is the required extension. Therefore we can assume by induction that the statement of the lemma holds for some $d \geq 1$.

Let $K \subset \mathbb{R}^{d+1}$ be an open $(d+1)$-simplex. Then, by the inductive hypothesis, for any $j > 1$, and any i, and any $F_{1,l}$ such that $F_{j,i} \subset \overline{F}_{1,l}$, there is an extension

$$E_{j,i}^l \colon \mathbb{P}^p(F_{j,i}) \cap H_0^1(F_{j,i}) \to \mathbb{P}^p(F_{1,l})$$

such that

$$\begin{aligned} &E_{j,i}^l|_{F_{j,i}} = I, \quad E_{j,i}^l|_{F_{j,m}} = 0, \quad m \neq i, F_{j,m} \subset \overline{F}_{1,l}, \\ &E_{j,i}^l|_{F_{r,m}} = 0, \quad F_{r,m} \subset \overline{F}_{1,l} \quad \text{and} \quad \overline{F_{r,m}} \cap \overline{F_{j,i}} = \emptyset, \end{aligned}$$

and, for $F_{r,m} \subset \overline{F_{1,l}}$,

$$\|E_{j,i}|_{F_{r,m}} u\|_{H_p^1(F_{r,m})} \leq C p^{\frac{r-j}{2}} \|u\|_{H_p^{1/2}(F_{j,i})}.$$

Notice that for $u \in H_0^1(F_{j,i})$, defining v on ∂K by

$$v := \begin{cases} E_{j,i}^l u, & F_{j,i} \subset \overline{F_{1,l}}, \\ 0, & F_{j,i} \cap \overline{F_{1,l}} = \emptyset, \end{cases}$$

we have $v \in H^1(\partial K)$, since it is in H^1 on each boundary face and is continuous across codimension 2 boundary faces.

Fix l such that $F_{j,i} \subset \overline{F_{1,l}}$. Then, by our claims (A.31) and (A.32), there exists $E_{d+1}^K \colon \mathbb{P}^p(\partial K) \to \mathbb{P}^p(K)$ such that

$$E_{d+1}^K|_{\overline{F_{1,l}}} = I, \quad \left(E_{d+1}^K v\right)|_{F_{r,m}} = 0 \quad \text{if } r = j,\ m \neq i \text{ or } \overline{F_{r,m}} \cap \overline{F_{j,i}} = \emptyset,$$

and

$$\|E_{d+1}^K v\|_{H_p^1(K)} \leq C\langle p\rangle^{-1/2} \|v\|_{H_p^{1/2}(\partial K)} \leq C\langle p\rangle^{-1/2} \|v\|_{H_p^1(\partial K)} \leq \langle p\rangle^{-j/2} \|u\|_{H_p^{1/2}},$$

which provides the required extension for $j > 1$.

For $j = 1$ and $u \in H_0^1(F_{1,i})$, we define $\tilde{u} \in H^1(\partial K)$ by

$$\tilde{u}|_{\overline{F_{1,l}}} = \begin{cases} u, & l = i, \\ 0, & \text{else}, \end{cases}$$

Then define

$$E_{1,i} u = E_d^K \tilde{u},$$

and the required estimates follow from (A.31) and (A.32).

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

It therefore remains to prove that $E_m^{K_m}$ exists. Without loss of generality, we can assume that

$$K_m = \left\{ x \in \mathbb{R}^m \colon 0 < x_i,\ \sum_{i=1}^{m} x_i < 1 \right\}.$$

Let $\psi \in C_c^\infty(\mathbb{R}; [0,1])$ with $0 \notin \operatorname{supp}(1-\psi)$, and define

$$\psi_{p,K_m}(x) := \psi(\langle p \rangle d(x, \partial K_m)).$$

Let E be the extension operator constructed in Lemma A.26, and define

$$E_m^{K_m} := E + \Pi^{\mathbb{P}^p}_{H_0^1(K_m)}(\psi_{p,K_m} - 1)E,$$

where $\Pi^{\mathbb{P}^p}_{H_0^1(K_m)}$ is the orthogonal projector with respect to the H^1-norm onto

$$\mathbb{P}^p(K_m) \cap H_0^1(K_m).$$

It is now enough to show that

$$\|E_m u\|_{H_p^1} \leq C\langle p \rangle^{-1/2} \|u\|_{H_p^{1/2}}, \tag{A.33}$$

To see this, we first estimate the H^1-norm

$$\begin{aligned} \langle p \rangle^{-1} \|E_m^{K_m} u\|_{H^1(K_m)} &\leq 2\langle p \rangle^{-1} \|Eu\|_{H^1(K_m)} + \|\psi_{p,K_m} Eu\|_{H_p^1} \\ &\leq C\langle p \rangle^{-1} \|Eu\|_{H^{1/2}} + \langle p \rangle^{-1} \|\psi_{p,K_m} Eu\|_{H^1}. \end{aligned}$$

Thus we need to show that

$$\langle p \rangle^{-1} \|\psi_{p,K_m} Eu\|_{H^1} \leq C\langle p \rangle^{-1/2} \|u\|_{H_p^{1/2}} \tag{A.34}$$

to obtain

$$\langle p \rangle^{-1} \|E_m^{K_m} u\|_{H^1(K_m)} \leq C\langle p \rangle^{-1/2} \|u\|_{H_p^{1/2}}. \tag{A.35}$$

To see (A.34), we make a Lipschitz change of coordinates so that $d(x, \partial K_m)$ becomes x_d (note that if γ is a Lipschitz bijection, then $\|u \circ \gamma\|_{H^1} \leq C\|u\|_{H^1}$, $\|u \circ \gamma\|_{L^2} \leq C\|u\|_{L^2}$ and it is enough to estimate in the new coordinates), and hence

$$\psi_{p,K_m} Eu(x_d, x') = \chi(\langle p \rangle x_d) \frac{1}{x_d} \int_0^{x_d} \partial_{x_d}(Eu)(x', x_d)\, ds + \chi(\langle p \rangle x_d) Eu(x', 0).$$

Therefore, by the Hardy inequality (A.19),

$$\begin{aligned} \|\psi(\langle p \rangle x_d) Eu\|_{L^2} &\leq \|\psi(\langle p \rangle x_d) x_d\|_{L^\infty} \|Eu\|_{H^1} + C\langle p \rangle^{-1/2} \|u\|_{L^2} \\ &\leq C\langle p \rangle^{-1} \|Eu\|_{H^1} + C\langle p \rangle^{-1/2} \|u\|_{L^2} \\ &\leq C\langle p \rangle^{-1/2} \big(\langle p \rangle^{-1/2} \|u\|_{H^{1/2}} + \|u\|_{L^2} \big) \\ &\leq C\langle p \rangle^{-1/2} \|u\|_{H_p^{1/2}}. \end{aligned} \tag{A.36}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Next,

$$\begin{aligned}\langle p\rangle^{-1}\|\nabla\psi(\langle p\rangle x_d)Eu\|_{L^2} &\le \|\psi'(\langle p\rangle x_d)Eu\|_{L^2} + \langle p\rangle^{-1}\|Eu\|_{H^1}\\ &\le C\langle p-1\rangle^{-1/2}\|u\|_{H^{1/2}_p} + C\langle p\rangle^{-1}\|u\|_{H^{1/2}}\\ &\le C\langle p\rangle^{-1/2}\|u\|_{H^{1/2}_p},\end{aligned}$$

and hence we have proved (A.34) and, in particular, (A.35).

For the L^2-bound, by (A.35), (A.36) and the self-adjointness of $\big(I-\Pi^{\mathbb{P}^p}_{\tilde H^1_0}\big)$ on H^1,

$$\begin{aligned}&\|E^{K_m}_m u\|_{L^2}\\ &\quad\le \|\psi_{p,K_m}Eu\|_{L^2} + \big\|\big(I-\Pi^{\mathbb{P}^p}_{H^1_0}\big)(\psi_{p,K_m}-1)Eu\big\|_{L^2}\\ &\quad\le C\langle p\rangle^{-1/2}\|u\|_{H^{1/2}_p} + \big\|\big(I-\Pi^{\mathbb{P}^p}_{\tilde H^1_0}\big)\big\|_{H^1_{1,0}\to L^2}\|(\psi_{p,K_m}-1)Eu\|_{H^1}\\ &\quad\le C\langle p\rangle^{-1/2}\|u\|_{H^{1/2}_p} + C\big\|\big(I-\Pi^{\mathbb{P}^p}_{H^1_0}\big)\big\|_{H^2\cap H^1_0\to H^1_0}\big(\langle p\rangle^{1/2}\|u\|_{H^{1/2}_p} + \|u\|_{H^{1/2}}\big).\end{aligned}$$

Now, Lemma A.23 implies that

$$\big\|\big(I-\Pi^{\mathbb{P}^p}_{\tilde H^1_0}\big)\big\|_{H^2\cap H^1_{1,0}\to H^1_{1,0}} \le C\langle p\rangle^{-1},$$

and hence

$$\begin{aligned}\|E^{K_m}_m u\|_{L^2} &\le C\langle p\rangle^{-1/2}\|u\|_{H^{1/2}_p} + C\langle p\rangle^{-1/2}\big(\|u\|_{H^{1/2}_p} + \langle p\rangle^{-1/2}\|u\|_{H^{1/2}}\big)\\ &\le C\|u\|_{H^{1/2}_p}.\end{aligned}$$

Combining this with (A.35) implies (A.33) and hence completes the proof. □

A.8. *Proof of Proposition A.21*

Proof of Proposition A.21. We follow the steps outlined in the subsection 'Idea of the proof of Proposition A.21' (page 137).

First observe that it is enough to consider

$$K = \Big\{x\in\mathbb{R}^d \;:\; 0<\sum_{i=1}^d x_i < 1,\; x_i>0\Big\}.$$

For $0<r<d$, define

$$\mathcal{A}_{r,i} := \{A\colon F_{r,i}\to F_{r,i'}\colon A \text{ is an affine isomorphism, } 1\le i'\le N_{r,d}\},$$

and let

$$\begin{aligned}&(\tilde T_{r,i}u)(x) := q(x) \quad \text{for } x\in F_{r,i},\\ &q := \frac{1}{|\mathcal{A}_{r,i}|}\sum_{A\in\mathcal{A}_{r,i}} (A^{-1})^*\mathcal{M}_{A^{-1}(F_{r,i})}\big(A^*u, H^1_0(A^{-1}(F_{r,i}))\big),\end{aligned} \tag{A.37}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

where

$$\mathcal{M}_F(u,V) \coloneqq \operatorname{argmin}\left\{\|u-q\|_{H_p^{1/2}(F)} \colon q \in V \cap \mathbb{P}^{p-1}\right\},$$

and set $(\tilde{T}_{d,i}u) \coloneqq u|_{F_{d,i}} \in \mathbb{C}$. Then define

$$T_{r,i} \coloneqq E_{r,i}\tilde{T}_{r,i},$$

where $E_{r,i}$ is the extension from Lemma A.27. Let

$$e_d \coloneqq (I-\mathcal{J}^{p-1})u, \quad e_j \coloneqq e_{j+1} - \sum_{i=1}^{N_{j+1,d}} T_{j+1,i}e_{j+1}, \quad j=0,\dots,d-1 \tag{A.38}$$

(i.e. $\sum_{i=1}^{N_{j+1,d}} T_{j+1,i}e_{j+1}$ is the function g_{j+1} in the discussion of the ideas in the proof above). We then let

$$\mathcal{C}^{p-1}u \coloneqq u - e_0.$$

Note first that $\mathcal{C}^{p-1}u \in \mathbb{P}^{p-1}$, since the image of all the $T_{r,i}$ maps is contained in $\mathbb{P}^{p-1}$. We now define recursively $Q_{d,i}(u|_{F_{d,i}}) \coloneqq u|_{F_{d,i}}$, and for $1 \le j < d$,

$$Q_{j,i}v \coloneqq \frac{1}{|\mathcal{A}_{j,i}|}\sum_{A\in\mathcal{A}_{j,i}} (A^{-1})^*\mathcal{M}_{A^{-1}(F_{j,i})}\big(A^*v, \{q\in C^\infty \colon q|_{A^{-1}(F_{j+1,\ell})} = A^*Q_{j+1,\ell}v \text{ if } F_{j+1,\ell}\subset\partial F_{j,i}\}\big). \tag{A.39}$$

We claim that (A.16) holds with this $Q_{j,i}$, so that, in particular, (A.14) holds.

To prove this, we claim that for all $1 \le j \le d$ and all i,

$$e_{j-1}|_{F_{j,i}} = u|_{F_{j,i}} - Q_{j,i}\big(u|_{\overline{F_{j,i}}}\big). \tag{A.40}$$

Once this is proved,

$$\mathcal{C}^{p-1}u|_{F_{j,i}} = (u-e_0)|_{F_{j,i}} = Q_{j,i}\big(u|_{\overline{F_{j,i}}}\big),$$

and hence (A.16) holds.

To prove (A.40), first observe that by the definition of e_{d-1} and the fact $T_{d,i}e_{d,i}|_{F_{d,j}} = e_{d,i}\delta_{ij}$ for all i and j,

$$e_{d-1}|_{F_{d,i}} = 0 = u|_{F_{d,i}} - Q_{d,i}(u|_{F_{d,i}}).$$

Suppose by induction that for some $0 < j \le d-1$, all i, and all $r = d, d-1, \dots, j+1$, $e_j|_{F_{r,i}} = u - Q_{r,i}u$. Then, by (A.38),

$$e_{j-1} = e_j - \sum_{i=1}^{N_{j,d}} T_{j,i}e_j \tag{A.41}$$

and $T_{j,i}e_j|_{F_{r,i}} = 0$ for $j < r \le d$ (since $T_{j,i}e_j|_{F_{j,i}}$ has zero boundary data by the

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

definition of $\tilde{T}_{j,i}$ and $T_{j,i}|_{\overline{F_{j,i'}}} = 0$ for $i \neq i'$ by the definition of $E_{j,i}$), and hence

$$e_{j-1}|_{F_{r,i}} = e_j|_{F_{r,i}}, \quad j < r \leq d.$$

Therefore, to prove (A.40), it remains to check that $e_{j-1}|_{F_{j,i}} = u - Q_{j,i}u$. By (A.41) and (A.37),

$$\begin{aligned} e_{j-1}|_{F_{j,i}} &= e_j|_{F_{j,i}} - \frac{1}{|\mathcal{A}_{j,i}|} \sum_{A \in \mathcal{A}_{j,i}} (A^{-1})^* \mathcal{M}_{A^{-1}(F_{j,i})}\big(A^* e_j, H_0^1(A^{-1}(F_{j,i}))\big) \\ &= e_j|_{F_{j,i}} - \frac{1}{|\mathcal{A}_{j,i}|} \sum_{A \in \mathcal{A}_{j,i}} (A^{-1})^* \\ &\quad \times \operatorname{argmin}\big\{\|A^*(e_j - u) + A^* u - q\|_{H_p^{1/2}(A^{-1}(F_{j,i}))} : q \in \mathbb{P}^{p-1} \cap H_0^1(A^{-1}(F_{j,i}))\big\}. \end{aligned}$$

Since $e_j - u \in \mathbb{P}^{p-1}$, we have

$$\begin{aligned} e_{j-1}|_{F_{j,i}} &= e_j|_{F_{j,i}} + u - e_j|_{F_{j,i}} - \frac{1}{|\mathcal{A}_{j,i}|} \\ &\quad \times \sum_{A \in \mathcal{A}_{j,i}} (A^{-1})^* \mathcal{M}_{A^{-1}(F_{j,i})}\big(A^* u, \{q \in C^\infty : (q + A^*(e_j - u))|_{\partial A^{-1}(F_{j,i})} = 0\}\big) \\ &= u - Q_{j,i}\big(u|_{\overline{F_{j,i}}}\big), \end{aligned}$$

where we have used (A.40) in the last step.

Now, to check (A.15), observe that for $\kappa : F_{r,i_2} \to F_{r,i_1}$, an affine isomorphism,

$$\mathcal{A}_{r,i_2} = \{A \circ \kappa : A \in \mathcal{A}_{r,i_1}\}.$$

Hence (A.15) with T replaced by $Q_{j,i}$ follows directly from the definition of $Q_{j,i}$.

Notice that if $\mathcal{J}^{p-1}u = u$, then $e_d = 0$ and hence $e_0 = 0$. Thus $\mathcal{C}^{p-1}u = u$. Therefore, if $\mathcal{J}^{p-1}u = u$ for all $u \in \mathbb{P}^{p-1}$, then $\mathcal{C}^{p-1}u = u$ for all $u \in \mathbb{P}^{p-1}$.

Finally, to prove estimates on

$$\|e_0\|_{H_p^j(k)} = \|(I - \mathcal{C}^{p-1})u\|_{H_p^j(K)},$$

we start by observing that

$$\|e_d\|_{H_p^j} = \|(I - \mathcal{J}^{p-1})u\|_{H_p^j(K)} \tag{A.42}$$

by the definition of e_d (A.38). First note that since $j > d/2$, for $0 \leq l \leq d$, and all i,

$$p^{-\frac{l}{2}} \|e_d|_{F_{l,i}}\|_{H_p^{j-\frac{l}{2}}} \leq C \|e_d\|_{L^2}^{1-\frac{l}{2j}} \|e_d\|_{H_p^j}^{\frac{l}{2j}}.$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

We now proceed by induction. Suppose for some $0 < n \le d$

$$\|e_n\|_{H^1_p(K)} + \sum_{\substack{0<l\le n\\ i}} p^{\frac{-l}{2}} \|e_n|_{F_{l,i}}\|_{H^{1/2}_p} \le C\|e_d\|_{L^2}^{1-\frac{d}{2j}} \|e_d\|_{H^j_p}^{\frac{d}{2j}}. \tag{A.43}$$

By Lemma A.27, for each i,

$$\begin{aligned}\sum_{l<n} p^{\frac{-l}{2}} \|(T_{n,i}e_n)|_{F_{l,i}}\|_{H^1_p(F_{l,i})} &\le Cp^{-\frac{n}{2}} \|\tilde{T}_{n,i}e_n\|_{H^{1/2}_{00,p}(F_{n,i})}\\ &\le Cp^{-\frac{n}{2}} \|e_n|_{F_{n,i}}\|_{H^{1/2}_p}\\ &\le C\|e_d\|_{L^2}^{1-d/(2j)} \|e_d\|_{H^j_p}^{\frac{d}{2j}}.\end{aligned}$$

Therefore, by (A.38) and (A.43),

$$\|e_{n-1}\|_{H^1_p(K)} \le \|e_n\|_{H^1_p(K)} + C\|e_d\|_{L^2}^{1-\frac{d}{2j}} \|e_d\|_{H^j_p}^{\frac{d}{2j}} \le C\|e_d\|_{L^2}^{1-\frac{d}{2j}} \|e_d\|_{H^j_p}^{\frac{d}{2j}},$$

and for $1 < l \le n-1$,

$$\begin{aligned}p^{-\frac{l}{2}} \|e_{n-1}\|_{H^{1/2}_p(F_{l,i})} &\le p^{-\frac{l}{2}} \|e_n\|_{H^{1/2}_p(F_{l,i})} + C\|e_d\|_{L^2}^{1-\frac{d}{2j}} \|e_d\|_{H^j_p}^{\frac{d}{2j}}\\ &\le C\|e_d\|_{L^2}^{1-\frac{d}{2j}} \|e_d\|_{H^j_p}^{\frac{d}{2j}}.\end{aligned}$$

Hence (A.43) holds with n replaced by $n-1$ and the induction is complete. Since

$$\|e_d\|_{L^2}^{1-\frac{d}{2j}} \|e_d\|_{H^j_p}^{\frac{d}{2j}} \le \|e_d\|_{H^j_p},$$

the bound on $\|e_0\|_{H^1_p(K)} = \|(I - \mathcal{C}^{p-1})u\|_{H^1_p(K)}$ follows from (A.42). □

A.9. Proof of Theorems A.11 and A.12

We now combine Proposition A.21 with Propositions A.18 and A.15 to prove Theorem A.11 (with Proposition A.18 used for the fixed-regularity approximation, and Proposition A.15 used for the high-regularity approximation).

Proof of Theorem A.11. Let $\widehat{T}$ be the reference element for $\mathcal{T}$, $p \ge \max(t-1, 1)$, let $\mathcal{L}^p : H^t(\widehat{T}) \to \mathbb{P}^p$ be the operator constructed in Proposition A.18, and let $\mathcal{I}^p$ be the operator constructed in Proposition A.15 with $\Omega = \widehat{T}$.

Then, by Proposition A.21, for each $T \in \mathcal{T}$, there are boundary-compatible operators $\mathcal{C}_i^p : H^\ell(\widehat{T}) \to \mathbb{P}^p$, $i = 1, 2$ such that by (A.12) (with $k = 1$ and $\ell = 0, 1, \dots, j$),

$$\left\|\left(I - \mathcal{C}_1^p\right)v\right\|_{H^1_p(\widehat{T})} \le C\|(I - \mathcal{L}^p)v\|_{H^j_p(\widehat{T})} \le \frac{C}{p^t}|v|_{H^t(\widehat{T})}, \tag{A.44}$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

by (A.10) (with $k = 1$ and $\ell = 0, 1, \dots, j$),

$$\left\|\left(I - \mathcal{C}_2^p\right) v\right\|_{H_p^1(\widehat{T})} \le C \|(I - \mathcal{I}^p) v\|_{H_p^j(\widehat{T})} \le \frac{C^{p+1}}{p^{p+1}} |v|_{H^{p+1}(\widehat{T})}, \tag{A.45}$$

and

$$\mathcal{C}_1^p v|_{\partial\widehat{T}} = \mathcal{C}_2^p v|_{\partial\widehat{T}}. \tag{A.46}$$

Then define $\mathcal{C}_{\mathcal{T}}^p \colon C^0(\Omega) \cap \oplus_{T\in\mathcal{T}} H_k^t(T) \to \mathcal{P}_{\mathcal{T}}^p$ by

$$\mathscr{F}_T^*\left(\mathcal{C}_{\mathcal{T}}^p u\right)|_T \coloneqq \begin{cases} \mathcal{C}_1^p(\mathscr{F}_T^* u|_T), & T \in \mathcal{T}_1 \\ \mathcal{C}_2^p(\mathscr{F}_T^* u|_T), & T \in \mathcal{T}_2. \end{cases} \tag{A.47}$$

Observe that $\mathcal{C}_{\mathcal{T}}^p$ is triangulation-preserving (in the sense of Definition A.10) since $\mathcal{C}_j^p$, $j = 1, 2$, are boundary-compatible. Furthermore, since $\mathcal{T}$ is a C^{p+1}-simplicial triangulation, by (A.44) and Lemma A.14, for $\ell = 0, 1$,

$$\begin{aligned}
\sum_{T\in\mathcal{T}_1} k^{-2\ell} \left|\left(I - \mathcal{C}_{\mathcal{T}}^p\right) u\right|_{H_p^\ell(T)}^2 &\le C \sum_{T\in\mathcal{T}_1} h_T^{d-2\ell} k^{-2\ell} \left\|\left(I - \mathcal{C}_1^p\right) \mathscr{F}_T^* u\right\|_{H_p^\ell(\widehat{T})}^2 \\
&\le \sum_{T\in\mathcal{T}_1} \frac{C}{p^{t-\ell}} k^{-2\ell} h_T^{d-2\ell} |A_T^* R_T^* u|_{H^t(\widehat{T})}^2 \\
&\le \frac{C}{p^{2(t-\ell)}} \sum_{T\in\mathcal{T}_1} h_T^{2(t-\ell)} k^{-2\ell} |R_T^* u|_{H^t(A_T^{-1}(\widehat{T}))}^2 \\
&\le \frac{C}{p^{2(t-\ell)}} \sum_{T\in\mathcal{T}_1} h_T^{2(t-\ell)} k^{-2\ell} \|u\|_{H^t(T)}^2 \\
&\le \frac{C}{p^{2(t-\ell)}} \sum_{T\in\mathcal{T}_1} (h_T k)^{2(t-\ell)} k^{-t} \|u\|_{H^t(T)}^2.
\end{aligned}$$

and, by (A.45) and Lemma A.14 (applied with $t = p + 1$), for $\ell = 0, 1$

$$\begin{aligned}
\sum_{T\in\mathcal{T}_2} k^{-2\ell} \left|\left(I - \mathcal{C}_{\mathcal{T}}^p\right) u\right|_{H_p^\ell(T)}^2 &\le C \sum_{T\in\mathcal{T}_2} h_T^{d-2\ell} k^{-2\ell} \left|\left(I - \mathcal{C}_2^p\right) \mathscr{F}_T^* u\right|_{H_p^{p+1}(\widehat{T})}^2 \\
&\le \sum_{T\in\mathcal{T}_2} \frac{C^{2(p+1)}}{p^{2(p+1)}} k^{-2\ell} h_T^{d-2\ell} |A_T^* R_T^* u|_{H^{p+1}(\widehat{T})}^2 \\
&\le \frac{C^{2(p+1)}}{p^{2(p+1)}} \sum_{T\in\mathcal{T}_2} h_T^{2(p+1-\ell)} k^{-2\ell} |R_T^* u|_{H^{p+1}(A_T^{-1}(\widehat{T}))}^2 \\
&\le C^{2(p+1)} \sum_{T\in\mathcal{T}_2} (h_T k)^{2(p+1-\ell)} k^{-2(p+1)} \|u\|_{H^{p+1}(T)}^2.
\end{aligned}$$

Therefore, to establish the result we only need to show that $\mathcal{C}_{\mathcal{T}}^p u \in H^1(\Omega)$.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

It is enough to check that $\mathcal{C}^p_{\mathcal{T}} u \in C^0(\Omega)$. It is clear that $\mathcal{C}^p_{\mathcal{T}} \in \oplus_{T\in\mathcal{T}} \overline{C^0}(T)$. Therefore we need only check that if $x \in T_1 \cap T_2$, then

$$\left[\mathcal{C}^p_{\ell_1}(\mathscr{F}^*_{T_1} u)\right](\mathscr{F}^{-1}_{T_1}(x)) = \left[\mathcal{C}^p_{\ell_2}(\mathscr{F}^*_{T_2} u)\right](\mathscr{F}^{-1}_{T_2}(x)), \tag{A.48}$$

where $T_1 \in \mathcal{T}_{\ell_1}, T_2 \in \mathcal{T}_{\ell_2}$. By the definition of a triangulation, if $x \in T_1 \cap T_2$, then $x \in \partial T_1 \cap \partial T_2$. In particular, since $\mathcal{T}$ is affine-conforming (in the sense of Definition A.8), there are (j_1, i_1) and (j_2, i_2) such that $x \in F_{j_1,i_1}(T_1) \cap F_{j_2,i_2}(T_2)$ and hence $F_{j_1,i_1}(T_1) = F_{j_2,i_2}(T_2)$ and there is an affine isomorphism

$$\kappa\colon \overline{\mathscr{F}^{-1}_{T_1}(F_{j_1,i_1}(T_1))} \to \overline{\mathscr{F}^{-1}_{T_2}(F_{j_2,i_2}(T_2))}$$

such that

$$\mathscr{F}_{T_1}|_{\overline{\mathscr{F}^{-1}_{T_1}(F_{j_1,i_1}(T_1))}} = \mathscr{F}_{T_2}|_{\overline{\mathscr{F}_{T_2^{-1}(F_{j_2,i_2}(T_2)}}} \circ \kappa. \tag{A.49}$$

Since $u \in C^0$,

$$(\mathscr{F}^*_{T_1} u)|_{\overline{\mathscr{F}^{-1}_{T_1}(F_{j_1,i_1})}} = \left((\mathscr{F}^*_{T_2} u)|_{\overline{\mathscr{F}^{-1}_{T_2}(F_{j_2,i_2})}}\right) \circ \kappa. \tag{A.50}$$

By (A.50), (A.49), (A.15) with T replaced by $\mathcal{C}^p_{\ell_1}$ (which holds since $\mathcal{C}^p_{\ell}, \ell = 1, 2$, are boundary-compatible) and (A.46), for $T_1 \in \mathcal{T}_{\ell_1}$ and $T_2 \in \mathcal{T}_{\ell_2}$,

$$\begin{aligned}
\mathcal{C}^p_{\ell_1}(\mathscr{F}^*_{T_1} u)(\mathscr{F}^{-1}_{T_1}(x)) &= \mathcal{C}^p_{\ell_1}(\kappa^* \mathscr{F}^*_{T_2} u)(\mathscr{F}^{-1}_{T_1}(x)) \\
&= \mathcal{C}^p_{\ell_1}(\kappa^* \mathscr{F}^*_{T_2} u)(\kappa^{-1} \circ \mathscr{F}^{-1}_{T_2}(x)) \\
&= \mathcal{C}^p_{\ell_1}(\mathscr{F}^*_{T_2} u)(\mathscr{F}^{-1}_{T_2}(x)) \\
&= \mathcal{C}^p_{\ell_2}(\mathscr{F}^*_{T_2} u)(\mathscr{F}^{-1}_{T_2}(x)),
\end{aligned}$$

as claimed in (A.48). □

The proof of Theorem A.12 is simpler than that of Theorem A.11; indeed, one can directly use $\mathcal{L}^p$ or $\mathcal{I}^p$ on each element $\mathcal{T}$ and therefore we do not require the triangulation to be affine-conforming.

B. Definition of the standard boundary integral operators

In this section we show how scattering of a plane wave by an obstacle with zero Dirichlet or Neumann boundary conditions can be reformulated as a BIE involving the operators A_k, A'_k (3.8) (for the Dirichlet problem) and $B_{\rm reg}, B'_{\rm reg}$ (3.9) (for the Neumann problem). The proof that the general Dirichlet and Neumann problems (i.e. with arbitrary boundary data) can be reformulated via these BIEs is very similar; see e.g. Chandler-Wilde *et al.* (2012, §2.6).

Let $\Omega_- \subset \mathbb{R}^d$, $d \geq 2$ be a bounded open set such that its open complement $\Omega_+ := \mathbb{R}^d \setminus \overline{\Omega_-}$ is connected. Let $\Gamma_- := \partial\Omega_-$. The results in Section 8 require that Γ is C^∞, but the results in this appendix hold when Γ is Lipschitz. Let ν be the

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

outward-pointing unit normal vector to Ω_-, and let $\gamma^\pm$ and $\partial_\nu^\pm$ denote the Dirichlet and Neumann traces, respectively, on Γ_- from $\Omega^\pm$.

Definition B.1 (Plane-wave sound-soft/-hard scattering problems). Given $k > 0$ and the incident plane wave $u^I(x) := \exp(ikx \cdot a)$ for $a \in \mathbb{R}^d$ with $|a|_2 = 1$, find the total field $u \in H^1_{\mathrm{loc}}(\Omega_+)$ satisfying $\Delta u + k^2 u = 0$ in Ω_+,

$$\text{either } \gamma^+ u = 0 \text{ (sound-soft) or } \partial_\nu^+ u = 0 \text{ (sound-hard) on } \Gamma,$$

and, where $u^S := u - u^I$ is the scattered field,

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

The results proving existence and uniqueness of the Helmholtz scattering problem (4.1) also show that the solutions of the sound-soft and sound-hard plane-wave scattering problems exist and are unique; see e.g. Chandler-Wilde *et al.* (2012, Theorem 2.12 and Corollary 2.13).

Let $\Phi_k(x, y)$ be the fundamental solution of the Helmholtz equation defined by

$$\Phi_k(x,y) := \frac{i}{4}\left(\frac{k}{2\pi|x-y|}\right)^{(d-2)/2} H^{(1)}_{(d-2)/2}(k|x-y|) = \begin{cases} \dfrac{i}{4}H_0^{(1)}(k|x-y|), & d=2, \\ \dfrac{e^{ik|x-y|}}{4\pi|x-y|}, & d=3, \end{cases} \tag{B.2}$$

where $H_\nu^{(1)}$ denotes the Hankel function of the first kind of order ν. The single- and double-layer potentials, $\mathcal{S}\ell$ and $\mathcal{D}\ell$ respectively, are defined for $k \in \mathbb{C}$, $\phi \in L^2(\Gamma)$ and $x \in \mathbb{R}^d \setminus \Gamma$ by

$$\mathcal{S}\ell\varphi(x) = \int_\Gamma \Phi_k(x,y)\varphi(y)\,ds(y) \quad \text{and} \quad \mathcal{D}\ell\varphi(x) = \int_\Gamma \frac{\partial \Phi_k(x,y)}{\partial \nu(y)}\varphi(y)\,ds(y). \tag{B.3}$$

The standard single-layer, adjoint-double-layer, double-layer and hypersingular operators are defined for $k \in \mathbb{C}$, $\phi \in L^2(\Gamma_-)$, $\psi \in H^1(\Gamma_-)$ and $x \in \Gamma_-$ by

$$V_k\phi(x) := \int_{\Gamma_-} \Phi_k(x,y)\phi(y)\,ds(y), \qquad K_k'\phi(x) := \int_{\Gamma_-} \frac{\partial \Phi_k(x,y)}{\partial \nu(x)}\phi(y)\,ds(y), \tag{B.4}$$

$$K_k\phi(x) := \int_{\Gamma_-} \frac{\partial \Phi_k(x,y)}{\partial \nu(y)}\phi(y)\,ds(y), \quad H_k\psi(x) := \frac{\partial}{\partial_{\nu(x)}}\int_{\Gamma_-} \frac{\partial \Phi_k(x,y)}{\partial \nu(y)}\psi(y)\,ds(y). \tag{B.5}$$

We use the notation V_k for the single-layer operator instead of S_k (used, in particular, in Galkowski *et al.* 2025), to avoid a notational clash with the smoothing operator in Section 6. Similarly, we use the notation K_k, K_k' for the double-layer operator and its adjoint, instead of D_k, D_k', to avoid a notational clash with the operator $D := -i\partial$ used in the rest of the paper.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Theorem B.2.

(i) If u is a solution of the sound-soft scattering problem of Definition B.1, then

$$A'_k \partial_\nu^+ u = \partial_\nu^+ u^I - iu^I \quad \text{and} \quad u = u^I - \mathcal{S}\ell(\partial_\nu^+ u). \tag{B.6}$$

(ii) If $v \in L^2(\Gamma_-)$ is the solution to

$$A_k v = -\gamma^+ u^I, \quad \text{then} \quad u := u^I + (\mathcal{D}\ell - i\mathcal{S}\ell)v \tag{B.7}$$

is the solution of the sound-soft scattering problem of Definition B.1.

(iii) If u is a solution of the sound-hard scattering problem of Definition B.1, then

$$B_{\mathrm{reg}} \gamma^+ u = i\gamma^+ u^I - S_{ik} \partial_\nu^+ u^I \quad \text{and} \quad u = u^I + \mathcal{D}\ell(\gamma^+ u). \tag{B.8}$$

(iv) If $v \in L^2(\Gamma_-)$ is the solution to

$$B'_{\mathrm{reg}} v = -\partial_\nu^+ u^I, \quad \text{then} \quad u := u^I + (\mathcal{D}\ell S_{ik} - i\mathcal{S}\ell)v$$

is the solution of the sound-hard scattering problem of Definition B.1.

References for the proof. Part (i) is proved in Chandler-Wilde *et al.* (2012, Theorem 2.46). Part (ii) is proved in Chandler-Wilde *et al.* (2012, equations 2.70–2.72). Part (iii) is proved in Galkowski *et al.* (2022, equation 1.6). Part (iv) is proved in Galkowski *et al.* (2022, equation 1.8). □

Lemma B.3. If Γ is C^1, then

$$\|A'_k\|_{L^2(\Gamma_-)\to L^2(\Gamma_-)} \geq 1/2, \qquad \|(A'_k)^{-1}\|_{L^2(\Gamma_-)\to L^2(\Gamma_-)} \geq 2,$$

$$\|B_{\mathrm{reg}}\|_{L^2(\Gamma_-)\to L^2(\Gamma_-)} \geq \left(\frac{1}{2}+\frac{1}{4}\right), \quad \|(B_{\mathrm{reg}})^{-1}\|_{L^2(\Gamma_-)\to L^2(\Gamma_-)} \geq \left(\frac{1}{2}+\frac{1}{4}\right)^{-1}.$$

Proof. The results for A'_k are proved in Chandler-Wilde, Graham, Langdon and Lindner (2009, Lemma 4.1) using that $A'_k - (1/2)I$ is compact when Γ is C^1. The results for B_{reg} are proved in an analogous way, using that $B_{\mathrm{reg}} - (i/2 - 1/4)I$ is compact by Galkowski *et al.* (2022, Proof of Theorem 2.2). □

C. Precise statement of Informal Theorem 5.5

Theorem C.1 (Properties of semiclassical pseudodifferential operators).

(i) *Quantization.* The *quantization* map

$$\mathrm{Op}_\hbar \colon S^m(\mathbb{R}^{2d}) \to \Psi_\hbar^m(\mathbb{R}^d)$$

defined by (5.3) is such that $\mathrm{Op}_\hbar(a)$ is properly supported, $\mathrm{Op}_\hbar(1) = I$, and $[\mathrm{Op}_\hbar(a)u](x) = a(x)u(x)$ when a is independent of ξ.

(ii) *Principal symbol.* There is a *principal symbol map*

$$\sigma_\hbar \colon \Psi_\hbar^m(\mathbb{R}^d) \to S^m(\mathbb{R}^d)/\hbar S^{m-1}(\mathbb{R}^d)$$

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

such that

$$\sigma_\hbar \circ \mathrm{Op}_\hbar = I \mod \hbar S^{m-1}(\mathbb{R}^d).$$

If $A \in \Psi_\hbar^m(\mathbb{R}^d)$ and $\hbar^{-1}\sigma_\hbar(A) \in S^{m-1}$, then $\hbar^{-1}A \in \Psi_\hbar^{m-1}(\mathbb{R}^d)$.

(iii) *Boundedness in Sobolev norms.* Given $a \in S^m(\mathbb{R}^{2d})$, for all s, there exists $C_s > 0$ such that

$$\|\mathrm{Op}_\hbar(a)\|_{H_k^s(\mathbb{R}^d)\to H_k^{s-m}(\mathbb{R}^d)} \le C_s \tag{C.1}$$

(i.e. the quantization of a symbol of order m decreases Sobolev regularity by m). Furthermore, for $a \in S^0(\mathbb{R}^{2d})$, there exists $C > 0$ such that

$$\|\mathrm{Op}_\hbar(a)\|_{L^2(\mathbb{R}^d)\to L^2(\mathbb{R}^d)} \le \sup_{(x,\xi)\in\mathbb{R}^{2d}} |a(x,\xi)| + C\hbar. \tag{C.2}$$

(iv) *Composition.* For $a \in S^{m_1}(\mathbb{R}^{2d})$, $b \in S^{m_2}(\mathbb{R}^{2d})$,

$$\mathrm{Op}_\hbar(a)\,\mathrm{Op}_\hbar(b) \in \Psi_\hbar^{m_1+m_2}(\mathbb{R}^d)$$

and

$$\sigma_\hbar(\mathrm{Op}_\hbar(a)\,\mathrm{Op}_\hbar(b)) - ab \in \hbar S^{m_1+m_2-1}(\mathbb{R}^{2d}).$$

Furthermore, there exists $c \in S^{m_1+m_2}(\mathbb{R}^{2d})$ with $\operatorname{supp} c \subset \operatorname{supp} a \cap \operatorname{supp} b$ such that

$$\mathrm{Op}_\hbar(a)\,\mathrm{Op}_\hbar(b) = \mathrm{Op}_\hbar(c) + O(\hbar^\infty)_{\Psi_\hbar^{-\infty}}.$$

(v) *Adjoints.* For $a \in S^m(\mathbb{R}^d)$, $(\mathrm{Op}_\hbar(a))^* \in \Psi_\hbar^m(\mathbb{R}^d)$ and

$$\sigma((\mathrm{Op}_\hbar(a))^*) - \bar{a} \in \hbar S^{m-1}(\mathbb{R}^{2d}).$$

(vi) *Commutators.* For $a \in S^{m_1}(\mathbb{R}^{2d})$ and $b \in S^{m_2}(\mathbb{R}^{2d})$,

$$i\hbar^{-1}[\mathrm{Op}_\hbar(a), \mathrm{Op}_\hbar(b)] \in \Psi_\hbar^{m_1+m_2-1}(\mathbb{R}^d)$$

and

$$\sigma_\hbar(i\hbar^{-1}[\mathrm{Op}_\hbar(a), \mathrm{Op}_\hbar(b)]) - \{a,b\} \in \hbar S^{m_1+m_2-2},$$

where $\{a,b\} = \langle \partial_\xi a, \partial_x b\rangle - \langle \partial_\xi b, \partial_x a\rangle$.

(vii) *Sharp Gårding inequality.* If $a \in S^m(\mathbb{R}^{2d})$, $w \in S^{m/2}(\mathbb{R}^{2d};\mathbb{R})$ and $\operatorname{Re} a \ge w^2$, then

$$\operatorname{Re}\langle \mathrm{Op}_\hbar(a)u, u\rangle \ge \|\mathrm{Op}_\hbar(w)u\|^2_{L^2(\mathbb{R}^d)} - C\hbar\|u\|^2_{H_k^{(m-1)/2}(\mathbb{R}^d)}.$$

References for the proof. Part (i) follows from the definition (5.3).

Part (ii) is proved in Dyatlov and Zworski (2019, Proposition E.16) and Zworski (2012, §9.3.3). Note that Dyatlov and Zworski (2019, Appendix E) consider a restricted class of symbols, so-called *polyhomogeneous symbols*; this ensures that the principal symbol is a function, rather than an equivalence class, but this distinction is not important for the proof of Theorem C.1.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

For part (iii), the bound (C.1) is reduced to the proof of $L^2 \to L^2$ boundedness for elements of $\Psi^0_\hbar$ in Dyatlov and Zworski (2019, Proposition E.19). The $L^2 \to L^2$ bound (C.2) is proved in Zworski (2012, Theorem 13.13), with a weaker bound (still sufficient to imply (C.1)) proved in Hörmander (1985, Theorem 18.1.11).

Parts (iv) and (v) are proved in Zworski (2012, Theorems 9.5 and 4.14) (see also Dyatlov and Zworski 2019, Proposition E.8).

Part (vi) is Dyatlov and Zworski (2019, E.1.24), and follows as a consequence of the composition result in Zworski (2012, Theorem 9.5) and Dyatlov and Zworski (2019, Proposition E.8).

Part (vii) is proved in Zworski (2012, Theorem 9.11). □

Acknowledgements

This article collects research done in collaboration with Martin Averseng (CNRS, Angers), Dean Baskin (Texas A&M), Théophile Chaumont-Frelet (INRIA, Lille), Shihua Gong (CUHK, Shenzhen), Ivan Graham (University of Bath), David Lafontaine (CNRS, Toulouse), Manas Rachh (ITT Bombay) and Jared Wunsch (Northwestern University) over the last ten years, and it is a pleasure to acknowledge these collaborations here.

Furthermore, the authors thank Martin Averseng, Théophile Chaumont-Frelet, Simon Chandler-Wilde (University of Reading), David Lafontaine, Markus Melenk (TU Wien), Mostafa Meliani (University of Bath), María Ignacia Fierro-Piccardo (University of Bath), Siavash Sadeghi (University of Reading), Alastair Spence (University of Bath) and Maciej Zworski (UC Berkeley) for discussions about drafts of this article.

JG was supported by EPSRC grants EP/V001760/1 and EP/V051636/1 and Leverhulme Research Project Grant RPG-2023-325, and JG and ES were both supported by the ERC synergy grant 'PSINumScat' 101167139.

References

M. Ainsworth (2004), Discrete dispersion relation for hp-version finite element approximation at high wave number, *SIAM J. Numer. Anal.* **42**, 553–575.

G. Alessandrini (2012), Strong unique continuation for general elliptic equations in 2D, *J. Math. Anal. Appl.* **386**, 669–676.

S. Amini and P. J. Harris (1990), A comparison between various boundary integral formulations of the exterior acoustic problem, *Comput. Methods Appl. Mech. Engrg* **84**, 59–75.

T. Apel and J. M. Melenk (2015), Interpolation and quasi-interpolation in h- and hp-version finite element spaces (extended version). ASC Report No. 39/2015. Available at https://www.tuwien.at/index.php?eID=dumpFile&t=f&f=190654&token=cf4edf3c7e8601fea9d11b7d3c9cc471e42cc6fa.

J. P. Aubin (1967), Behavior of the error of the approximate solutions of boundary value problems for linear elliptic operators by Galerkin's and finite difference methods, *Ann. Sc. Norm. Super. Pisa Cl. Sci.* **21**, 599–637.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

M. Averseng, J. Galkowski and E. A. Spence (2024), Helmholtz FEM solutions are locally quasi-optimal modulo low frequencies, *Adv. Comput. Math.* **50**, art. 112.

M. Averseng, J. Galkowski and E. A. Spence (2025), Non-uniform finite-element meshes defined by ray dynamics for Helmholtz problems. Available at arXiv:2506.15630.

A. K. Aziz, R. B. Kellogg and A. B. Stephens (1988), A two point boundary value problem with a rapidly oscillating solution, *Numer. Math.* **53**, 107–121.

I. Babuška and S. A. Sauter (1997), Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers?, *SIAM J. Numer. Anal.* **34**, 2392–2423.

I. Babuška and M. Suri (1987), The optimal convergence rate of the p-version of the finite element method, *SIAM J. Numer. Anal.* **24**, 750–776.

I. Babuška and M. Suri (1994), The p and h-p versions of the finite element method, basic principles and properties, *SIAM Rev.* **36**, 578–632.

I. Babuška, A. Craig, J. Mandel and J. Pitkäranta (1991), Efficient preconditioning for the p-version finite element method in two dimensions, *SIAM J. Numer. Anal.* **28**, 624–661.

G. Bachman, L. Narici and E. Beckenstein (2000), *Fourier and Wavelet Analysis*, Springer.

J. M. Ball, Y. Capdeboscq and B. Tsering-Xiao (2012), On uniqueness for time harmonic anisotropic Maxwell's equations with piecewise regular coefficients, *Math. Models Methods Appl. Sci.* **22**, art. 1250036.

L. Banjai and S. Sauter (2007), A refined Galerkin error and stability analysis for highly indefinite variational problems, *SIAM J. Numer. Anal.* **45**, 37–53.

C. Bardos, G. Lebeau and J. Rauch (1992), Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary, *SIAM J. Control Optim.* **30**, 1024–1065.

H. Barucq, T. Chaumont-Frelet and C. Gout (2017), Stability analysis of heterogeneous Helmholtz problems and finite element solution based on propagation media approximation, *Math. Comp.* **86**, 2129–2157.

D. Baskin, E. A. Spence and J. Wunsch (2016), Sharp high-frequency estimates for the Helmholtz equation and applications to boundary integral equations, *SIAM J. Math. Anal.* **48**, 229–267.

M. Bellassoued (2003), Carleman estimates and distribution of resonances for the transparent obstacle and application to the stabilization, *Asymptot. Anal.* **35**, 257–279.

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

C. Bernardi (1989), Optimal finite-element interpolation on curved domains, *SIAM J. Numer. Anal.* **26**, 1212–1240.

C. Bernardi and Y. Maday (1997), Spectral methods, in *Techniques of Scientific Computing 2* (P. G. Ciarlet and J. L. Lions, eds), Vol. 5 of Handbook of Numerical Analysis, Elsevier, pp. 209–485.

C. Bernardi, M. Dauge and Y. Maday (1995), Interpolation of nullspaces for polynomial approximation of divergence-free functions in a cube, in *Boundary Value Problems and Integral Equations in Nonsmooth Domains*, Vol. 167 of Lecture Notes in Pure and Applied Mathematics, Dekker, pp. 27–46.

M. Bernkopf, T. Chaumont-Frelet and J. M. Melenk (2025), Wavenumber-explicit stability and convergence analysis of hp finite element discretizations of Helmholtz problems in piecewise smooth media, *Math. Comp.* **94**, 73–122.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

C. O. Bloom (1973), Estimates for solutions of reduced hyperbolic equations of the second order with a large parameter, *J. Math. Anal. Appl.* **44**, 310–332.

C. O. Bloom and N. D. Kazarinoff (1974), Local energy decay for a class of nonstar-shaped bodies, *Arch. Ration. Mech. Anal.* **55**, 73–85.

C. O. Bloom and N. D. Kazarinoff (1977), *A priori* bounds for solutions of the Dirichlet problem for $[\Delta + \lambda^2 n(x)]u = f(x, \lambda)$ on an exterior domain, *J. Differential Equations* **24**, 437–465.

J.-F. Bony, N. Burq and T. Ramond (2010), Minoration de la résolvante dans le cas captif, *C.R. Math.* **348**(23–24), 1279–1282.

Y. Boubendir and C. Turc (2013), Wave-number estimates for regularized combined field boundary integral operators in acoustic scattering problems with Neumann boundary conditions, *IMA J. Numer. Anal.* **33**, 1176–1225.

N. Bouziani, F. Nataf and P.-H. Tournier (2023), A unified framework for double sweep methods for the Helmholtz equation, *J. Comput. Phys.* **490**, art. 112305.

H. Brakhage and P. Werner (1965), Über das Dirichletsche Aussenraumproblem für die Helmholtzsche Schwingungsgleichung, *Archiv Math.* **16**, 325–329.

J. H. Bramble and J. Pasciak (2007), Analysis of a finite PML approximation for the three dimensional time-harmonic Maxwell and acoustic scattering problems, *Math. Comp.* **76**, 597–614.

J. H. Bramble and J. E. Pasciak (2013), Analysis of a Cartesian PML approximation to acoustic scattering problems in $\mathbb{R}^2$ and $\mathbb{R}^3$, *J. Comput. Appl. Math.* **247**, 209–230.

S. C. Brenner and L. R. Scott (2008), *The Mathematical Theory of Finite Element Methods*, Vol. 15 of Texts in Applied Mathematics, third edition, Springer.

D. L. Brown, D. Gallistl and D. Peterseim (2017), Multiscale Petrov–Galerkin method for high-frequency heterogeneous Helmholtz equations, in *Meshfree Methods for Partial Differential Equations VIII*, Springer, pp. 85–115.

O. Bruno, T. Elling and C. Turc (2012), Regularized integral equations and fast high-order solvers for sound-hard acoustic scattering problems, *Internat. J. Numer. Methods Engrg* **91**, 1045–1072.

N. Burq (1998), Décroissance de l'énergie locale de l'équation des ondes pour le problème extérieur et absence de résonance au voisinage du réel, *Acta Math.* **180**, 1–29.

N. Burq (2002), Semi-classical estimates for the resolvent in nontrapping geometries, *Int. Math. Res. Not.* **2002**, 221–241.

A. J. Burton (1976), *Numerical Solution of Acoustic Radiation Problems*, National Physical Laboratory, Division of Numerical Analysis and Computing.

A. J. Burton and G. F. Miller (1971), The application of integral equation methods to the numerical solution of some exterior boundary-value problems, *Proc. R. Soc. A* **323**, 201–210.

A. P. Calderón (1961), Lebesgue spaces of differentiable functions and distributions, in *Partial Differential Equations* (C. B. Morrey Jr, ed.), Vol. 4 of Proceedings of Symposia in Pure Mathematics, American Mathematical Society, pp. 33–49.

C. Canuto and A. Quarteroni (1982), Approximation results for orthogonal polynomials in sobolev spaces, *Math. Comp.* **38**, 67–86.

F. Cardoso and G. Vodev (2002), Uniform estimates of the resolvent of the Laplace–Beltrami operator on infinite volume Riemannian manifolds II, *Ann. Henri Poincaré* **3**, 673–691.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

T. Carleman (1939), Sur un problème d'unicité pour les systèmes d'équations aux dérivées partielles à deux variables indépendantes, *Ark. Mat., Astr. Fys.* **26**, issue 17.
J. Céa (1964), Approximation variationnelle des problèmes aux limites, *Ann. Inst. Fourier* **14**, 345–444.
S. N. Chandler-Wilde and S. Langdon (2007), A Galerkin boundary element method for high frequency scattering by convex polygons, *SIAM J. Numer. Anal.* **45**, 610–640.
S. N. Chandler-Wilde and P. Monk (2008), Wave-number-explicit bounds in time-harmonic scattering, *SIAM J. Math. Anal.* **39**, 1428–1455.
S. N. Chandler-Wilde and S. Sadeghi (2025), The cut-off resolvent can grow arbitrarily fast in obstacle scattering. Available at arXiv:2508.16958.
S. N. Chandler-Wilde, I. G. Graham, S. Langdon and M. Lindner (2009), Condition number estimates for combined potential boundary integral operators in acoustic scattering, *J. Integral Equ. Appl.* **21**, 229–279.
S. N. Chandler-Wilde, I. G. Graham, S. Langdon and E. A. Spence (2012), Numerical-asymptotic boundary integral methods in high-frequency acoustic scattering, *Acta Numer.* **21**, 89–305.
S. N. Chandler-Wilde, D. P. Hewett, S. Langdon and A. Twigger (2015), A high frequency boundary element method for scattering by a class of nonconvex obstacles, *Numer. Math.* **129**, 647–689.
S. N. Chandler-Wilde, E. A. Spence, A. Gibbs and V. P. Smyshlyaev (2020), High-frequency bounds for the Helmholtz equation under parabolic trapping and applications in numerical analysis, *SIAM J. Math. Anal.* **52**, 845–893.
T. Chaumont Frelet (2015), Approximation par éléments finis de problèmes d'Helmholtz pour la propagation d'ondes sismiques. PhD thesis, Rouen, INSA.
T. Chaumont-Frelet (2025), A ill-posed scattering problem saturating Weyl's law. Available at arXiv:2505.10346.
T. Chaumont-Frelet and S. Nicaise (2020), Wavenumber explicit convergence analysis for finite element discretizations of general wave propagation problem, *IMA J. Numer. Anal.* **40**, 1503–1543.
T. Chaumont-Frelet and E. A. Spence (2023), Scattering by finely layered obstacles: frequency-explicit bounds and homogenization, *SIAM J. Math. Anal.* **55**, 1319–1363.
T. Chaumont-Frelet and E. A. Spence (2025), The geometric error is less than the pollution error when solving the high-frequency Helmholtz equation with high-order FEM on curved domains, *IMA J. Numer. Anal.* Available at doi:10.1093/imanum/draf035.
T. Chaumont-Frelet, V. Dolean and M. Ingremeau (2024*a*), Efficient approximation of high-frequency Helmholtz solutions by Gaussian coherent states, *Numer. Math.* **156**, 1385–1426.
T. Chaumont-Frelet, J. Galkowski and E. A. Spence (2024*b*), Sharp error bounds for edge-element discretisations of the high-frequency Maxwell equations. Available at arXiv:2408.04507. To appear in *Math. Comp.*
T. Chaumont-Frelet, S. Gallistl, D.and Nicaise and J. Tomezyk (2022), Wavenumber explicit convergence analysis for finite element discretizations of time-harmonic wave propagation problems with perfectly matched layers, *Commun. Math. Sci.* **20**, 1–52.
J. Cheeger and M. Taylor (1982*a*), Diffraction by conical singularities II, *Commun. Pure Appl. Math.* **35**, 487–529.
J. Cheeger and M. Taylor (1982*b*), On the diffraction of waves by conical singularities I, *Commun. Pure Appl. Math.* **35**, 275–331.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Z. Chen and X. Xiang (2013), A source transfer domain decomposition method for Helmholtz equations in unbounded domain, *SIAM J. Numer. Anal.* **51**, 2331–2356.

H. Christianson (2018), High-frequency resolvent estimates on asymptotically Euclidean warped products, *Commun. Partial Differential Equations* **43**, 1306–1362.

H. Christianson and J. Wunsch (2013), Local smoothing for the Schrödinger equation with a prescribed loss, *Amer. J. Math.* **135**, 1601–1632.

P. G. Ciarlet (2002), *The Finite Element Method For Elliptic Problems*, Classics in Applied Mathematics, SIAM.

D. Colton and R. Kress (1983), *Integral Equation Methods in Scattering Theory*, Wiley.

G. M. Constantine and T. H. Savits (1996), A multivariate Faà di Bruno formula with applications, *Trans. Amer. Math. Soc.* **348**, 503–520.

M. Costabel, M. Dauge and S. Nicaise (2010), Corner singularities and analytic regularity for linear elliptic systems, I: Smooth domains. Available at https://hal.archives-ouvertes.fr/file/index/docid/453934/filename/CoDaNi_Analytic_Part_I.pdf.

P. Cummings and X. Feng (2006), Sharp regularity coefficient estimates for complex-valued acoustic and elastic Helmholtz equations, *Math. Models Methods Appl. Sci.* **16**, 139–160.

K. Datchev and A. Vasy (2012*a*), Propagation through trapped sets and semiclassical resolvent estimates, *Ann. Inst. Fourier* **62**, 2347–2377.

K. Datchev and A. Vasy (2012*b*), Semiclassical resolvent estimates at trapped sets, *Ann. Inst. Fourier* **62**, 2379–2384.

A. Demlow, J. Guzmán and A. H. Schatz (2011), Local energy estimates for the finite element method on sharply varying grids, *Math. Comp.* **80**, 1–9.

A. Deraemaeker, I. Babuška and P. Bouillard (1999), Dispersion and pollution of the FEM solution for the Helmholtz equation in one, two and three dimensions, *Internat. J. Numer. Methods Engrg* **46**, 471–499.

J. Descloux (1975), Interior regularity and local convergence of Galerkin finite element approximations for elliptic equations, in *Topics in Numerical Analysis II* (J. J. H. Miller, ed.), Academic Press, pp. 27–41.

M. Dimassi and J. Sjöstrand (1999), *Spectral Asymptotics in the Semi-Classical Limit*, Vol. 268 of London Mathematical Society Lecture Note Series, Cambridge University Press.

DLMF (2025), NIST Digital Library of Mathematical Functions (F. W. J. Olver *et al.*, eds). Release 1.2.4 of 2025-03-15. Available at https://dlmf.nist.gov/.

V. Dolean, P. Jolivet and F. Nataf (2015), *An Introduction to Domain Decomposition Methods: Algorithms, Theory, and Parallel Implementation*, SIAM.

J. Douglas Jr, J. E. Santos, D. Sheen and L. S. Bennethum (1993), Frequency domain treatment of one-dimensional scalar waves, *Math. Models Methods Appl. Sci.* **3**, 171–194.

Y. Du and H. Wu (2015), Preasymptotic error analysis of higher order FEM and CIP-FEM for Helmholtz equation with high wave number, *SIAM J. Numer. Anal.* **53**, 782–804.

Y. Du and H. Wu (2020), A pure source transfer domain decomposition method for Helmholtz equations in unbounded domain, *J. Sci. Comput.* **83**, art. 68.

S. Dyatlov and M. Zworski (2019), *Mathematical Theory of Scattering Resonances*, Vol. 200 of Graduate Studies in Mathematics, American Mathematical Society.

B. Engquist and A. Majda (1977*a*), Absorbing boundary conditions for numerical simulation of waves, *Proc. Nat. Acad. Sci. USA* **74**, 1765–1766.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

B. Engquist and A. Majda (1977*b*), Absorbing boundary conditions for the numerical simulation of waves, *Math. Comp.* **31**, 629–651.
B. Engquist and O. Runborg (2003), Computational high frequency wave propagation, *Acta Numer.* **12**, 181–266.
B. Engquist and L. Ying (2011), Sweeping preconditioner for the Helmholtz equation: moving perfectly matched layers, *Multiscale. Model. Simul.* **9**, 686–710.
B. Engquist and H.-K. Zhao (1998), Absorbing boundary conditions for domain decomposition, *Appl. Numer. Math.* **27**, 341–365.
Y. A. Erlangga (2008), Advances in iterative methods and preconditioners for the Helmholtz equation, *Arch. Comput. Methods Engrg* **15**, 37–66.
A. Ern and J.-L. Guermond (2021), *Finite Elements I: Approximation and Interpolation*, Vol. 72 of Texts in Applied Mathematics, Springer Nature.
O. G. Ernst and M. J. Gander (2012), Why it is difficult to solve Helmholtz problems with classical iterative methods, in *Numerical Analysis of Multiscale Problems* (I. G. Graham *et al.*, eds), Vol. 83 of Lecture Notes in Computational Science and Engineering, Springer, pp. 325–363.
L. Escauriaza, S. Montaner and C. Zhang (2017), Analyticity of solutions to parabolic evolutions and applications, *SIAM J. Math. Anal.* **49**, 4064–4092.
S. Esterhazy and J. M. Melenk (2012), On stability of discretizations of the Helmholtz equation, in *Numerical Analysis of Multiscale Problems* (I. G. Graham *et al.*, eds), Springer, pp. 285–324.
X. Feng and H. Wu (2009), Discontinuous Galerkin methods for the Helmholtz equation with large wave number, *SIAM J. Numer. Anal.* **47**, 2872–2896.
X. Feng and H. Wu (2011), hp-discontinuous Galerkin methods for the Helmholtz equation with large wave number, *Math. Comp.* **80**, 1997–2024.
N. Filonov (2001), Second-order elliptic equation of divergence form having a compactly supported solution, *J. Math. Sci.* **106**, 3078–3086.
K. Friedrichs (1934), Spektraltheorie halbbeschränkter Operatoren und Anwendung auf die Spektralzerlegung von Differentialoperatoren, *Math. Ann.* **109**, 465–487.
J. Galkowski (2019), *Distribution of Resonances in Scattering by Thin Barriers*, Vol. 259 of Memoirs of the American Mathematical Society, American Mathematical Society.
J. Galkowski (2025), Lower bounds for piecewise polynomial approximations of oscillatory functions, *J. Approx. Theory* **305**, art. 106100.
J. Galkowski and E. A. Spence (2023), Does the Helmholtz boundary element method suffer from the pollution effect?, *SIAM Rev.* **65**, 806–828.
J. Galkowski and E. A. Spence (2025), Sharp preasymptotic error bounds for the Helmholtz h-FEM, *SIAM J. Numer. Anal.* **63**, 1–23.
J. Galkowski and E. A. Spence (2026), Convergence theory for two-level hybrid Schwarz preconditioners for high-frequency Helmholtz problems, *SIAM J. Numer. Anal.* **64**, 29–54.
J. Galkowski, S. Gong, I. G. Graham, D. Lafontaine and E. A. Spence (2024*a*), Convergence of overlapping domain decomposition methods with PML transmission conditions applied to nontrapping Helmholtz problems. Available at arXiv:2404.02156.
J. Galkowski, D. Lafontaine and E. A. Spence (2023*a*), Perfectly-matched-layer truncation is exponentially accurate at high frequency, *SIAM J. Math. Anal.* **55**, 3344–3394.
J. Galkowski, D. Lafontaine and E. A. Spence (2024*b*), Local absorbing boundary conditions on fixed domains give order-one errors for high-frequency waves, *IMA J. Numer. Anal.* **44**, 1946–2069.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

J. Galkowski, D. Lafontaine, E. A. Spence and J. Wunsch (2023*b*), Decompositions of high-frequency Helmholtz solutions via functional calculus, and application to the finite element method, *SIAM J. Math. Anal.* **55**, 3903–3958.
J. Galkowski, D. Lafontaine, E. A. Spence and J. Wunsch (2024*c*), The hp-FEM applied to the Helmholtz equation with PML truncation does not suffer from the pollution effect, *Commun. Math. Sci.* **22**, 1761–1816.
J. Galkowski, P. Marchand and E. A. Spence (2022), High-frequency estimates on boundary integral operators for the Helmholtz exterior Neumann problem, *Integral Equations Operator Theory* **94**, art. 36.
J. Galkowski, E. H. Müller and E. A. Spence (2019), Wavenumber-explicit analysis for the Helmholtz h-BEM: Error estimates and iteration counts for the Dirichlet problem, *Numer. Math.* **142**, 329–357.
J. Galkowski, M. Rachh and E. A. Spence (2025), Helmholtz boundary integral methods and the pollution effect. Available at arXiv:2507.22797.
J. Galkowski, E. A. Spence and J. Wunsch (2020), Optimal constants in nontrapping resolvent estimates, *Pure Appl. Anal.* **2**, 157–202.
M. J. Gander (2006), Optimized Schwarz methods, *SIAM J. Numer. Anal.* **44**, 699–731.
M. J. Gander and H. Zhang (2019), A class of iterative solvers for the Helmholtz equation: Factorizations, sweeping preconditioners, source transfer, single layer potentials, polarized traces, and optimized Schwarz methods, *SIAM Rev.* **61**, 3–76.
M. J. Gander and H. Zhang (2022), Schwarz methods by domain truncation, *Acta Numer.* **31**, 1–134.
M. Ganesh and C. Morgenstern (2020), A coercive heterogeneous media Helmholtz model: Formulation, wavenumber-explicit analysis, and preconditioned high-order FEM, *Numer. Algorithms* **83**, 1441–1487.
I. G. Graham and S. A. Sauter (2020), Stability and finite element error analysis for the Helmholtz equation with variable coefficients, *Math. Comp.* **89**, 105–138.
I. G. Graham, M. Löhndorf, J. M. Melenk and E. A. Spence (2015), When is the error in the h-BEM for solving the Helmholtz equation bounded independently of k?, *BIT Numer. Math.* **55**, 171–214.
I. G. Graham, O. R. Pembery and E. A. Spence (2019), The Helmholtz equation in heterogeneous media: *A priori* bounds, well-posedness, and resonances, *J. Differential Equations* **266**, 2869–2923.
G. Grubb (2008), *Distributions and Operators*, Vol. 252 of Graduate Texts in Mathematics, Springer.
B. Guo and J. Zhang (2009), Stable and compatible polynomial extensions in three dimensions and applications to the p and h-p finite element method, *SIAM J. Numer. Anal.* **47**, 1195–1225.
I. Harari and T. J. R. Hughes (1991), Finite element methods for the Helmholtz equation in an exterior domain: Model problems, *Comput. Methods Appl. Mech. Engrg* **87**, 59–96.
U. Hetmaniuk (2007), Stability estimates for a class of Helmholtz problems, *Commun. Math. Sci.* **5**, 665–678.
R. Hiptmair, A. Moiola and I. Perugia (2016), A survey of Trefftz methods for the Helmholtz equation, in *Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations*, Springer, pp. 237–279.
T. Hohage, F. Schmidt and L. Zschiedrich (2003), Solving time-harmonic scattering problems based on the pole condition II: Convergence of the PML method, *SIAM J. Math. Anal.* **35**, 547–560.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

L. Hörmander (1959), On the uniqueness of the Cauchy problem II, *Math. Scand.* **7**, 177–190.

L. Hörmander (1985), *The Analysis of Linear Partial Differential Operators III: Pseudo-Differential Operators*, Springer.

F. Ihlenburg (1998), *Finite Element Analysis of Acoustic Scattering*, Springer.

F. Ihlenburg and I. Babuška (1995*a*), Dispersion analysis and error estimation of Galerkin finite element methods for the Helmholtz equation, *Internat. J. Numer. Methods Engrg* **38**, 3745–3774.

F. Ihlenburg and I. Babuška (1995*b*), Finite element solution of the Helmholtz equation with high wave number I: The h-version of the FEM, *Comput. Math. Appl.* **30**, 9–37.

F. Ihlenburg and I. Babuška (1997), Finite element solution of the Helmholtz equation with high wave number II: The hp version of the FEM, *SIAM J. Numer. Anal.* **34**, 315–358.

D. Jerison and C. E. Kenig (1985), Unique continuation and absence of positive eigenvalues for Schrödinger operators, *Ann. of Math.* **121**, 463–488.

S. Kim and J. E. Pasciak (2010), Analysis of a Cartesian PML approximation to acoustic scattering problems in $\mathbb{R}^2$, *J. Math. Anal. Appl.* **370**, 168–186.

J. J. Kohn and L. Nirenberg (1965), An algebra of pseudo-differential operators, *Commun. Pure Appl. Math.* **18**, 269–305.

D. Lafontaine, E. A. Spence and J. Wunsch (2021), For most frequencies, strong trapping has a weak effect in frequency-domain scattering, *Commun. Pure Appl. Math.* **74**, 2025–2063.

D. Lafontaine, E. A. Spence and J. Wunsch (2022*a*), A sharp relative-error bound for the Helmholtz h-FEM at high frequency, *Numer. Math.* **150**, 137–178.

D. Lafontaine, E. A. Spence and J. Wunsch (2022*b*), Wavenumber-explicit convergence of the hp-FEM for the full-space heterogeneous Helmholtz equation with smooth coefficients, *Comput. Math. Appl.* **113**, 59–69.

D. Lahaye, J. Tang and K. Vuik, eds (2017), *Modern Solvers for Helmholtz Problems*, Geosystems Mathematics, Birkhäuser.

M. Lassas and E. Somersalo (1998), On the existence and convergence of the solution of PML equations, *Computing* **60**, 229–241.

M. Lassas and E. Somersalo (2001), Analysis of the PML equations in general convex geometry, *Proc. Roy. Soc. Edinburgh Sect. A* **131**, 1183–1207.

P. D. Lax and R. S. Phillips (1989), *Scattering Theory*, second edition, Academic Press.

R. Leis (1965), Zur Dirichletschen Randwertaufgabe des Aussenraumes der Schwingungsgleichung, *Math. Z.* **90**, 205–211.

W. Leng and L. Ju (2019), An additive overlapping domain decomposition method for the Helmholtz equation, *SIAM J. Sci. Comput.* **41**, A1252–A1277.

W. Leng and L. Ju (2021), A diagonal sweeping domain decomposition method with source transfer for the Helmholtz equation, *Commun. Comput. Phys.* **29**, 357–398.

W. Leng and L. Ju (2022), Trace transfer-based diagonal sweeping domain decomposition method for the Helmholtz equation: Algorithms and convergence analysis, *J. Comput. Phys.* **455**, art. 110980.

M. Lenoir (1986), Optimal isoparametric finite elements and error estimates for domains involving curved boundaries, *SIAM J. Numer. Anal.* **23**, 562–580.

N. Lerner (2019), *Carleman Inequalities: An Introduction and More*, Vol. 353 of Grundlehren der mathematischen Wissenschaften, Springer.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

Y. Li and H. Wu (2019), FEM and CIP-FEM for Helmholtz equation with high wave number and perfectly matched layer truncation, *SIAM J. Numer. Anal.* **57**, 96–126.
M. Löhndorf and J. M. Melenk (2011), Wavenumber-explicit hp-BEM for high frequency scattering, *SIAM J. Numer. Anal.* **49**, 2340–2363.
C. H. Makridakis, F. Ihlenburg and I. Babuška (1996), Analysis and finite element methods for a fluid–solid interaction problem in one dimension, *Math. Models Methods Appl. Sci.* **6**, 1119–1141.
S. Marburg (2002), Six boundary elements per wavelength: Is that enough?, *J. Comput. Acous.* **10**, 25–51.
S. Marburg (2016), Numerical damping in the acoustic boundary element method, *Acta Acust.* **102**, 415–418.
S. Marburg (2017), Benchmark problem identifying a pollution effect in boundary element method, *J. Acoust. Soc. Am.* **141**, 3975–3975.
P. Marchand, J. Galkowski, A. Spence and E. A. Spence (2022), Applying GMRES to the Helmholtz equation with strong trapping: How does the number of iterations depend on the frequency?, *Adv. Comput. Math.* **48**, 1–63.
W. McLean (2000), *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press.
J. M. Melenk (1995), On generalized finite element methods. PhD thesis, The University of Maryland.
J. M. Melenk (2004), *hp-Finite Element Methods for Singular Perturbations*, Springer.
J. M. Melenk and S. Sauter (2010), Convergence analysis for finite element discretizations of the Helmholtz equation with Dirichlet-to-Neumann boundary conditions, *Math. Comp* **79**(272), 1871–1914.
J. M. Melenk and S. Sauter (2011), Wavenumber explicit convergence analysis for Galerkin discretizations of the Helmholtz equation, *SIAM J. Numer. Anal.* **49**, 1210–1243.
J. M. Melenk and S. A. Sauter (2021), Wavenumber-explicit hp-FEM analysis for Maxwell's equations with transparent boundary conditions, *Found. Comput. Math.* **21**, 125–241.
J. M. Melenk, A. Parsania and S. Sauter (2013), General DG-methods for highly indefinite Helmholtz problems, *J. Sci. Comput.* **57**, 536–581.
R. B. Melrose (1979), Singularities and energy decay in acoustical scattering, *Duke Math. J.* **46**, 43–59.
R. B. Melrose and J. Sjöstrand (1978), Singularities of boundary value problems I, *Commun. Pure Appl. Math.* **31**, 593–617.
R. B. Melrose and J. Sjöstrand (1982), Singularities of boundary value problems II, *Commun. Pure Appl. Math.* **35**, 129–168.
R. B. Melrose and M. E. Taylor (1985), Near peak scattering and the corrected Kirchhoff approximation for a convex obstacle, *Adv. Math.* **55**, 242–315.
K. Miller (1974), Nonunique continuation for uniformly parabolic and elliptic equations in self-adjoint divergence form with Hölder continuous coefficients, *Arch. Ration. Mech. Anal.* **54**, 105–117.
A. Moiola and E. A. Spence (2014), Is the Helmholtz equation really sign-indefinite?, *SIAM Rev.* **56**, 274–312.
A. Moiola and E. A. Spence (2019), Acoustic transmission problems: Wavenumber-explicit bounds and resonance-free regions, *Math. Models Methods Appl. Sci.* **29**, 317–354.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

C. S. Morawetz (1975), Decay for solutions of the exterior problem for the wave equation, *Commun. Pure Appl. Math.* **28**, 229–264.
C. S. Morawetz and D. Ludwig (1968), An inequality for the reduced wave operator and the justification of geometrical optics, *Commun. Pure Appl. Math.* **21**, 187–203.
C. S. Morawetz, J. V. Ralston and W. A. Strauss (1977), Decay of solutions of the wave equation outside nontrapping obstacles, *Commun. Pure Appl. Math.* **30**, 447–508.
R. Muñoz-Sola (1997), Polynomial liftings on a tetrahedron and applications to the hp version of the finite element method in three dimensions, *SIAM J. Numer. Anal.* **34**, 282–314.
F. Nataf, F. Rogier and E. de Sturler (1994), Optimal interface conditions for domain decomposition methods. Technical Report 301, CMAP Ecole Polytechnique. Available at hal-02194208.
J. Nitsche (1968), Ein Kriterium für die Quasi-Optimalität des ritzschen Verfahrens, *Numer. Math.* **11**, 346–348.
J. A. Nitsche and A. H. Schatz (1974), Interior estimates for Ritz–Galerkin methods, *Math. Comp.* **28**, 937–958.
M. Ohlberger and B. Verfurth (2018), A new heterogeneous multiscale method for the Helmholtz equation with high contrast, *Multiscale Model. Simul.* **16**, 385–411.
P. Olver (1993), *Applications of Lie Groups to Differential Equations*, second edition, Springer.
O. I. Panich (1965), On the question of the solvability of exterior boundary-value problems for the wave equation and for a system of Maxwell's equations (in Russian), *Uspekhi Mat. Nauk* **20**, 221–226.
O. R. Pembery (2020), The Helmholtz equation in heterogeneous and random media: Analysis and numerics. PhD thesis, University of Bath. Available at https://research-portal.bath.ac.uk/en/studentTheses/the-helmholtz-equation-in-heterogeneous-and-random-media-analysis.
B. Perthame and L. Vega (1999), Morrey–Campanato estimates for Helmholtz equations, *J. Funct. Anal.* **164**, 340–355.
A. Plis (1963), On non-uniqueness in Cauchy problem for an elliptic second order differential equation, *Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys* **11**, 95–100.
M. Reed and B. Simon (1980), *Methods of Modern Mathematical Physics I: Functional Analysis*, second edition, Academic Press.
F. Rellich (1940), Darstellung der Eigenwerte von $\Delta u + \lambda u = 0$ durch ein Randintegral, *Math. Z.* **46**, 635–636.
F. Rellich (1943), Über das asymptotische Verhalten der Lösungen von $\Delta u + \lambda u = 0$ in unendlichen Gebieten, *Jahresber. Dtsch. Math.-Ver.* **53**, 57–65.
A. Royer, C. Geuzaine, E. Béchet and A. Modave (2022), A non-overlapping domain decomposition method with perfectly matched layer transmission conditions for the Helmholtz equation, *Comput. Methods Appl. Mech. Engrg* **395**, art. 115006.
Y. Saad and M. H. Schultz (1986), GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems, *SIAM J. Sci. Statist. Comput.* **7**, 856–869.
S. A. Sauter (2006), A refined finite element convergence theory for highly indefinite Helmholtz problems, *Computing* **78**, 101–115.
S. A. Sauter and C. Schwab (2011), *Boundary Element Methods*, Springer.
A. H. Schatz (1974), An observation concerning Ritz–Galerkin methods with indefinite bilinear forms, *Math. Comp.* **28**, 959–962.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

A. H. Schatz and L. B. Wahlbin (1977), Interior maximum norm estimates for finite element methods, *Math. Comp.* **31**, 414–442.

A. H. Schatz and L. B. Wahlbin (1982), On the quasi-optimality in L_∞ of the $\overset{\circ}{H}_1$-projection into finite element spaces, *Math. Comp.* **38**, 1–22.

A. H. Schatz and J. Wang (1996), Some new error estimates for Ritz–Galerkin methods with minimal regularity assumptions, *Math. Comp.* **65**, 19–27.

C. Schwab (1998), *p- and hp- Finite Element Methods: Theory and Applications in Solid and Fluid Mechanics*, Numerical Mathematics and Scientific Computation, Clarendon Press.

L. R. Scott (1973), Finite element techniques for curved boundaries. PhD thesis, Massachusetts Institute of Technology.

C. E. Shannon (1949), Communication in the presence of noise, *Proc. IRE* **37**, 10–21.

J. Sjöstrand and M. Zworski (1991), Complex scaling and the distribution of scattering poles, *J. Amer. Math. Soc.* **4**, 729–769.

E. A. Spence (2014), Wavenumber-explicit bounds in time-harmonic acoustic scattering, *SIAM J. Math. Anal.* **46**, 2987–3024.

E. A. Spence (2023), A simple proof that the hp-FEM does not suffer from the pollution effect for the constant-coefficient full-space Helmholtz equation, *Adv. Comput. Math.* **49**, art. 27.

E. A. Spence, I. V. Kamotski and V. P. Smyshlyaev (2015), Coercivity of combined boundary integral equations in high-frequency scattering, *Commun. Pure Appl. Math.* **68**, 1587–1639.

C. C. Stolk (2013), A rapidly converging domain decomposition method for the Helmholtz equation, *J. Comput. Phys.* **241**, 240–252.

M. Taus, L. Zepeda-Núñez, R. J. Hewett and L. Demanet (2019), L-sweeps: A scalable, parallel preconditioner for the high-frequency Helmholtz equation, *J. Comput. Phys.* **420**, art. 109706.

M. E. Taylor (1996), *Partial Differential Equations II: Qualitative Studies of Linear Equations*, Vol. 116 of Applied Mathematical Sciences, Springer.

M. E. Taylor (2023), *Partial Differential Equations III: Nonlinear Equations*, Vol. 117 of Applied Mathematical Sciences, third edition, Springer.

A. Toselli (1998), Overlapping method with perfectly matched layers for the solution of the Helmholtz equation, in *Proceedings of the 11th International Conference on Domain Decomposition Methods*, pp. 551–557. Available at http://www.ddm.org/DD11/DD11proc.pdf.

B. R. Vainberg (1975), 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 Math. Surveys* **30**, 1–58.

F. Vico, L. Greengard and Z. Gimbutas (2014), Boundary integral equation analysis on the sphere, *Numer. Math.* **128**, 463–487.

G. Vodev (2000), On the exponential bound of the cutoff resolvent, *Serdica Math. J.* **26**, 49–58.

L. B. Wahlbin (1991), Local behavior in finite element methods, in *Finite Element Methods 1* (P. G. Ciarlet and J. L. Lions, eds), Vol. II of Handbook of Numerical Analysis, Elsevier, pp. 353–522.

E. T. Whittaker (1915), On the functions which are represented by the expansions of the interpolation-theory, *Proc. Roy. Soc. Edinburgh* **35**, 181–194.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press

T. Wolff (1993), Recent work on sharp estimates in second-order elliptic unique continuation problems, *J. Geom. Anal.* **3**, 621–650.

T. H. Wolff (1992), A property of measures in $\mathbb{R}^N$ and an application to unique continuation, *Geom. Funct. Anal.* **2**, 225–284.

H. Wu (2014), Pre-asymptotic error analysis of CIP-FEM and FEM for the Helmholtz equation with high wave number I: Linear version, *IMA J. Numer. Anal.* **34**, 1266–1288.

L. Zepeda-Núñez and L. Demanet (2016), The method of polarized traces for the 2D Helmholtz equation, *J. Comput. Phys.* **308**, 347–388.

M. Zlámal (1973), Curved elements in the finite element method I, *SIAM J. Numer. Anal.* **10**, 229–240.

M. Zworski (2012), *Semiclassical Analysis*, Vol. 138 of Graduate Studies in Mathematics, American Mathematical Society.

https://doi.org/10.1017/S0962492925100275 Published online by Cambridge University Press