At the interface between semiclassical analysis and numerical analysis of wave-scattering problems

E. A. Spence*

December 17, 2021

Abstract

These are the lecture notes for the course https://math.ethz.ch/fim/activities/nachdiplom-lectures/euan-spence.html. The abstract for the course is as follows.

Semiclassical analysis (SCA) is a branch of microlocal analysis concerned with rigorously analysing PDEs with large (or small) parameters. On the other hand, numerical analysis (NA) seeks to design numerical methods that are accurate, efficient, and robust, with theorems guaranteeing these properties.

In the context of high-frequency wave scattering, both SCA and NA share the same goal – that of understanding the behaviour of the scattered wave – but these two fields have operated largely in isolation, mainly because the tools and techniques of the two fields are somewhat disjoint.

This by-and-large self-contained course focuses on the Helmholtz equation, which is arguably the simplest possible model of wave propagation. Our first goal will be to show how even relatively-simple tools from semiclassical analysis can be used to prove fundamental results about the numerical analysis of finite-element method applied to the high-frequency Helmholtz equation.

The course will aim at being accessible both to students coming from a numerical-analysis/applied-maths background and to students coming from an analysis background.

Contents

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

0 Introduction

0.1 Where does the Helmholtz equation come from? (Short version)

The Helmholtz equation is the simplest model of wave propagation; indeed, seeking solutions of the wave equation

$$\Delta U - \frac{1}{c^2}\frac{\partial^2 U}{\partial t^2} = 0$$

in the form $U(x,t) = \exp(-\mathrm{i}\omega t)u(x)$ (so-called "time-harmonic solutions"), we find that $u(x)$ satisfies

$$\Delta u + k^2 u = 0 \tag{0.1}$$

with $k = \omega/c$; since c is a speed (with dimension (length)(time)$^{-1}$), and ω has dimension (time)$^{-1}$, k has dimension (length)$^{-1}$; k is called the *wavenumber*, and ω is called the *angular frequency*.

In this course, we are interested in the Helmholtz equation (0.1) (and its generalisation with variable coefficients) when k is large, and we approach this question mathematically by studying the behaviour of Helmholtz solutions as $k \to \infty$. We often refer to this limit as the "high-frequency" limit, rather than the "large-wavenumber" limit, because the former terminology is more familiar to most people than the latter.

0.2 Why is the high-frequency Helmholtz equation difficult?

Dividing the Helmholtz equation (0.1) by k^2, we obtain

$$k^{-2}\Delta u + u = 0. \tag{0.2}$$

Recall that a *singular limit* of a PDE with a parameter is one in which the coefficient of the highest-order term vanishes with the parameter is formally set to the limit; a *regular limit* is one in which this coefficient does not vanish. The limit $k \to \infty$ is therefore a singular limit for the Helmholtz equation, and the limit $k \to 0$ a regular limit. This explains why the Helmholtz equation with k large is harder to solve than the Helmholtz equation with k small, or the Laplace equation. Furthermore, one can show that the equations governing the $k \to \infty$ limit of solutions of the Helmholtz equation are *nonlinear*.

The high-frequency Helmholtz equation $\Delta u + k^2 u = 0$ is difficult to solve numerically for the following three reasons:

1. The solutions of the homogeneous Helmholtz equation oscillate on a scale of $1/k$, and so to approximate them accurately with piecewise polynomial functions (e.g. using the finite element method) one needs the total number of degrees of freedom, N, to be proportional to k^d as k increases, where d is the spatial dimension.

2. The *pollution effect* means that for fixed-order finite-element methods with $N \sim k^d$, even though the best-approximation error is bounded independently of k, the relative error grows with k. The fact that $N \gg k^d$ is required for the relative error to be bounded independently of k leads to very large matrices, and hence to large (and sometimes intractable) computational costs.

3. The standard variational formulation of the Helmholtz equation is not coercive (i.e. it is sign-indefinite) when k is sufficiently large; in other words, zero is in the *numerical range* or *field of values* of the operator. This indefiniteness is inherited by the Galerkin linear

system; therefore even when the linear system has a unique solution (which depends on the discretisation and on k), one expects iterative methods to behave badly if the system is not preconditioned.

0.3 Where does the Helmholtz equation come from? (Longer version)

Answer 1: from the acoustic approximation of elastic waves. The acoustic approximation of the elastic wave equation removes the (longitudinal) shear waves and keeps the (transverse) compressional waves to obtain the wave equation

$$\frac{1}{\kappa}\frac{\partial^2 p}{\partial t^2} - \nabla\cdot\left(\frac{1}{\rho}\nabla p\right) = \nabla\cdot\left(\frac{1}{\rho}\mathbf{f}\right),$$

where p is the pressure, ρ is the density, κ is given in terms of the Lamé parameters and has the same dimension as pressure, and $\mathbf{f}$ is a force term; see, e.g., [35, §1.2.6]. Assuming $p(x,t) = u(x)\exp(-\mathrm{i}\omega t)$, we obtain

$$\nabla\cdot\left(\frac{1}{\rho}\nabla u\right) + \frac{\omega^2}{\kappa}u = -\nabla\cdot\left(\frac{1}{\rho}\mathbf{f}\right). \tag{0.3}$$

Let ρ_0 and κ_0 be reference values of ρ and κ. Let $A := \rho_0/\rho$ and $n := \kappa_0/\kappa$ be the (dimensionless) relative variations of ρ and κ, respectively (observe that matrix-valued A then corresponds to anisotropic density), and let $c_0 = \sqrt{\rho_0/\kappa_0}$ (which one can check has the dimensions of speed). Multiplying (0.3) by ρ_0 and using these definitions, we find that u satisfies

$$\nabla\cdot(A\nabla u) + k^2 n u = -\rho_0\nabla\cdot\left(\frac{1}{\rho}\mathbf{f}\right).$$

Answer 2: from transverse electric (TE) or transverse magnetic (TM) modes of the time-harmonic Maxwell equations. The time-harmonic Maxwell equations are

$$\nabla\times\mathbf{H} + \mathrm{i}\omega\varepsilon\mathbf{E} = (\mathrm{i}\omega)^{-1}\mathbf{J}, \qquad \nabla\times\mathbf{E} - \mathrm{i}\omega\mu\mathbf{H} = \mathbf{0} \qquad \text{in } \mathbb{R}^3, \tag{0.4}$$

where ε is the electric permittivity and μ is the magnetic permeability. When all fields and parameters involved depend only on two Cartesian space variables, say x and y, the equations (0.4) reduce to the Helmholtz equation in $\mathbb{R}^2$.

In the transverse-magnetic (TM) mode, J and E are given by $J = (0,0,J_z(x,y))$ and $E = (0,0,E_z(x,y))$ so, when additionally the permittivity ε is a scalar and the permeability μ satisfies

$$\mu = \begin{pmatrix} \widetilde{\mu} & 0 \\ 0 & 1 \end{pmatrix} \tag{0.5}$$

for $\widetilde{\mu}$ a 2×2 symmetric positive-definite matrix, E_z satisfies

$$\nabla\cdot(A\nabla E_z) + \omega^2 n E_z = -f \tag{0.6}$$

with $n = \varepsilon$,

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}^T (\widetilde{\mu})^{-1} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \tag{0.7}$$

and $f = J_z$. Similarly, in the case of the transverse-electric (TE) mode, J and H are given by $J = (J_x(x,y), J_y(x,y), 0)$ and $H = (0,0,H_z(x,y))$, so that when μ is a scalar and ε satisfies an equation analogous to (0.5), the PDE (0.6) holds with E_z replaced by H_z, A given by (0.7) with $\widetilde{\mu}$ replaced by $\widetilde{\varepsilon}$, $n = \mu$, and

$$f = -\frac{1}{\mathrm{i}\omega}\nabla\cdot\left[\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}(\widetilde{\varepsilon})^{-1}\begin{pmatrix} J_x \\ J_y \end{pmatrix}\right].$$

1 Wellposedness of the Helmholtz equation

1.1 The Helmholtz exterior Dirichlet problem

Notation: $L^p(\Omega)$ denotes complex-valued L^p functions on a Lipschitz open set Ω. When the range of the functions is not $\mathbb{C}$, it will be given in the second argument; e.g. $L^\infty(\Omega, \mathbb{R}^{d\times d})$ denotes the space of $d \times d$ matrices with each entry a real-valued L^∞ function on Ω. We use γ to denote the trace operator $H^1(\Omega) \to H^{1/2}(\partial\Omega)$ and ∂_ν to denote the normal derivative trace operator $H^1(\Omega, \Delta) \to H^{-1/2}(\partial\Omega)$, where $H^1(\Omega, \Delta) := \{v \in H^1(\Omega) : \Delta v \in L^2(\Omega)\}$.

Assumption 1.1. (Assumptions on the domain and coefficients.)
(i) $\Omega_- \subset \mathbb{R}^d, d = 2, 3$, *is a bounded open Lipschitz set such that its open complement* $\Omega_+ := \mathbb{R}^d \setminus \overline{\Omega_-}$ *is connected.*

(ii) $A \in C^{0,1}(\Omega_+, \mathrm{SPD})$ *(where* SPD *is the set of* $d \times d$ *real, symmetric, positive-definite matrices) is such that* $\mathrm{supp}(I - A)$ *is bounded and there exist* $0 < A_{\min} \le A_{\max} < \infty$ *such that, for all* $\xi \in \mathbb{R}^d$,

$$A_{\min}|\xi|^2 \le \big(A(x)\xi\big) \cdot \xi \le A_{\max}|\xi|^2 \quad \text{for every } x \in \Omega_+. \tag{1.1}$$

(iii) $n \in L^\infty(\Omega_+, \mathbb{R})$ *is such that* $\mathrm{supp}(1 - n)$ *is bounded and there exist* $0 < n_{\min} \le n_{\max} < \infty$ *such that*

$$n_{\min} \le n(x) \le n_{\max} \quad \text{for almost every } x \in \Omega_+. \tag{1.2}$$

Definition 1.2. (Exterior Dirichlet Problem (EDP).) *Given* Ω_-, A, *and* n *satisfying Assumption 1.1,* $k > 0$, *and*

- $f \in L^2(\Omega_+)$ *with* $\mathrm{supp} f$ *bounded,*
- $g_D \in H^{1/2}(\Gamma_D)$, *where* $\Gamma_D := \partial\Omega_-$,

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

$$k^{-2}\nabla \cdot (A\nabla u) + nu = -f \quad \text{in } \Omega_+, \qquad \gamma u = g_D \quad \text{on } \Gamma_D, \tag{1.3}$$

and u *satisfies the Sommerfeld radiation condition*

$$k^{-1}\frac{\partial u}{\partial r}(x) - \mathrm{i}u(x) = o\left(\frac{1}{r^{(d-1)/2}}\right) \tag{1.4}$$

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

Some remarks:

(i) The PDE in (1.3) is understood in the following weak sense:

$$\int_{\Omega_+} k^{-2}u\nabla \cdot (A\overline{\nabla\phi}) + nu\overline{\phi} = -\int_{\Omega_+} f\,\overline{\phi} \quad \text{for all } \phi \in C^\infty_{\mathrm{comp}}(\Omega_+), \tag{1.5}$$

where $C^\infty_{\mathrm{comp}}(\Omega_+) := \{\phi \in C^\infty(\Omega_+) : \mathrm{supp}\,\phi \text{ is a compact subset of } \Omega_+\}$.

(ii) We can legitimately impose the radiation condition (1.4) on the function $u \in H^1_{\mathrm{loc}}(\Omega_+)$ since u satisfies the equation $k^{-2}\Delta u + u = 0$ outside a ball of finite radius, and then u is C^∞ outside this ball by elliptic regularity (see §2.3 below).

Definition 1.3. (Helmholtz plane-wave sound-soft scattering problem.) *Given* $k > 0$ *and* $a \in \mathbb{R}^d$ *with* $|a| = 1$, *let* $u^I(x) := \exp(\mathrm{i}kx \cdot a)$. *Given* Ω_-, A, *and* n *satisfying Assumption 1.1, we say* $u \in H^1_{\mathrm{loc}}(\Omega_+)$ *satisfies the* Helmholtz plane-wave scattering problem *if*

$$k^{-2}\nabla \cdot (A\nabla u) + nu = 0 \quad \text{in } \Omega_+, \qquad \gamma u = 0 \quad \text{on } \Gamma_D, \tag{1.6}$$

and $u^S := u - u^I$ *satisfies the Sommerfeld radiation condition* (1.4) *(with* u *replaced by* u^S*) as* $r := |x| \to \infty$, *uniformly in* $\widehat{x} := x/r$.

(If the zero Neumann boundary condition $\partial_n u = 0$ is prescribed instead of the zero Dirichlet condition, then the problem is the sound-hard scattering problem.)

Lemma 1.4. (Atkinson-Wilcox expansion.) *(i) If $u \in C^2(\mathbb{R}^d \setminus \overline{B_R})$ (for some $R > 0$) is a solution of the Helmholtz equation satisfying the Sommerfeld radiation condition, then*

$$u(x) = \frac{\mathrm{e}^{\mathrm{i}kr}}{r^{(d-1)/2}} \sum_{n=0}^{\infty} \frac{f_n(\widehat{x})}{r^n} \qquad \text{for all } r > R,$$

where both the series and all its term-by-term derivatives converge absolutely and uniformly with respect to r and $\widehat{x}$.

(ii) If $f_0 = 0$, then $u = 0$ in $\mathbb{R}^d \setminus \overline{B_R}$.

Proof. (i) can be proved using the explicit solution (obtained by separation of variables) for the solution of the Helmholtz equation outside a ball (see (1.10) below), or integral equations and properties of the fundamental solution; for the former approach, see [118, §2.6.3], for the latter approach, see [41, Theorem 3.6].

(ii) is proved using (i) in, e.g., [41, Corollary 3.8]. □

Remark 1.5. (An "outgoing solution" of the Helmholtz equation.) *A solution of the Helmholtz equation satisfying the Sommerfeld radiation condition (1.4) is often called an* outgoing *solution. The reason for this is that, if u satisfies the Sommerfeld radiation condition and $U(x,t) := u(x)\exp(-\mathrm{i}\omega t)$, then, with $c = \omega/k$,*

$$U(x,t) = \frac{\mathrm{e}^{\mathrm{i}k(r-ct)}}{r^{(d-1)/2}} \sum_{n=0}^{\infty} \frac{f_n(\widehat{x})}{r^n} \qquad \text{for all } r \geq R;$$

recall that a function of the form $(r,t) \mapsto w(r-ct)$ can be seen as a wave moving in the positive r direction with speed c as t increases (i.e., "outgoing").

1.2 Weighted norms

For Ω a bounded Lipschitz open set, let

$$\|v\|^2_{H^m_k(\Omega)} := \sum_{0 \leq |\alpha| \leq m} k^{-2|\alpha|} \|D^\alpha v\|^2_{L^2(\Omega)}, \tag{1.7}$$

so that, in particular,

$$\|v\|^2_{H^1_k(\Omega)} := k^{-2} \|\nabla v\|^2_{L^2(\Omega)} + \|v\|^2_{L^2(\Omega)}. \tag{1.8}$$

The rationale for using these norms is that if a function v oscillates with frequency k, then we (roughly) expect $|\nabla v| \sim k|v|$; e.g., if $v(x) = \exp(\mathrm{i}kx \cdot a)$, then $\nabla v(x) = \mathrm{i}ka \exp(\mathrm{i}kx \cdot a)$.

1.3 The Dirichlet-to-Neumann map in the exterior of a ball

Let $R > 0$ be such that $\overline{\Omega_-} \cup \mathrm{supp}(I - A) \cup \mathrm{supp}(1 - n) \Subset B_R$, where B_R denotes the ball of radius R about the origin; see Figure 1.1. Let $\Omega_R := \Omega_+ \cap B_R$, and let $\Gamma_R := \partial B_R$.

Definition 1.6. (Dirichlet-to-Neumann map in the exterior of B_R.) *Given $g \in H^{1/2}(\Gamma_R)$, let u be the outgoing solution of*

$$(-k^{-2}\Delta - 1)u = 0 \quad \text{in } \mathbb{R}^d \setminus \overline{B_R} \quad \text{and} \quad \gamma u = g \text{ on } \Gamma_R. \tag{1.9}$$

Define the map $\mathrm{DtN}_k : H^{1/2}(\Gamma_R) \to H^{-1/2}(\Gamma_R)$ by

$$\mathrm{DtN}_k g := k^{-1} \partial_\nu u,$$

where $\boldsymbol{\nu} := x/R = \widehat{x}$ (i.e., $\boldsymbol{\nu}$ is the outward-pointing unit normal vector to B_R).

When $d = 2$, the outgoing solution to (1.9) is given in polar coordinates by

$$u(r,\theta) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} \frac{H^{(1)}_n(kr)}{H^{(1)}_n(kR)} \exp(\mathrm{i}n\theta)\widehat{g}(n), \quad \text{where} \quad \widehat{g}(n) := \int_0^{2\pi} \exp(-\mathrm{i}n\theta) g(R,\theta)\,\mathrm{d}\theta, \tag{1.10}$$

so that

$$\mathrm{DtN}_k g = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} \frac{H_n^{(1)'}(kR)}{H_n^{(1)}(kR)} \exp(\mathrm{i}n\theta)\widehat{g}(n).$$

An analogous expression is available for $d = 3$; see, e.g., [33, Equation 3.6] [118, §2.6.3], [108, Equation 3.10].

Lemma 1.7. (Key properties of DtN_k.)

(i) Given $k_0, R_0 > 0$ there exists $C_{\mathrm{DtN}1} = C_{\mathrm{DtN}1}(k_0 R_0)$ such that for all $k \geq k_0$ and $R \geq R_0$,

$$\left|\left\langle \mathrm{DtN}_k(\gamma u), \gamma v \right\rangle_{\Gamma_R}\right| \leq k\, C_{\mathrm{DtN}1}\, \|u\|_{H^1_k(\Omega_R)}\, \|v\|_{H^1_k(\Omega_R)} \qquad \text{for all } u, v \in H^1(\Omega_R). \tag{1.11}$$

where $\langle \cdot, \cdot \rangle_{\Gamma_R}$ denotes the duality pairing on Γ_R that is linear in the first argument and antilinear in the second argument.

(ii)

$$\Im\left\langle \mathrm{DtN}_k \phi, \phi \right\rangle_{\Gamma_R} > 0 \qquad \text{for all } \phi \in H^{1/2}(\Gamma_R) \setminus \{0\}. \tag{1.12}$$

(iii) Given $k_0, R_0 > 0$ there exists $C_{\mathrm{DtN}2} = C_{\mathrm{DtN}2}(k_0 R_0)$ such that for all $k \geq k_0$ and $R \geq R_0$,

$$-\Re\left\langle \mathrm{DtN}_k \phi, \phi \right\rangle_{\Gamma_R} \geq C_{\mathrm{DtN}2}(kR)^{-1}\, \|\phi\|^2_{L^2(\Gamma_R)} \qquad \text{for all } \phi \in H^{1/2}(\Gamma_R). \tag{1.13}$$

Proof. The proof of Part (ii) is Exercise 1 in §1.6. For the proofs of Parts (i) and (iii), see [108, Lemma 3.3]. □

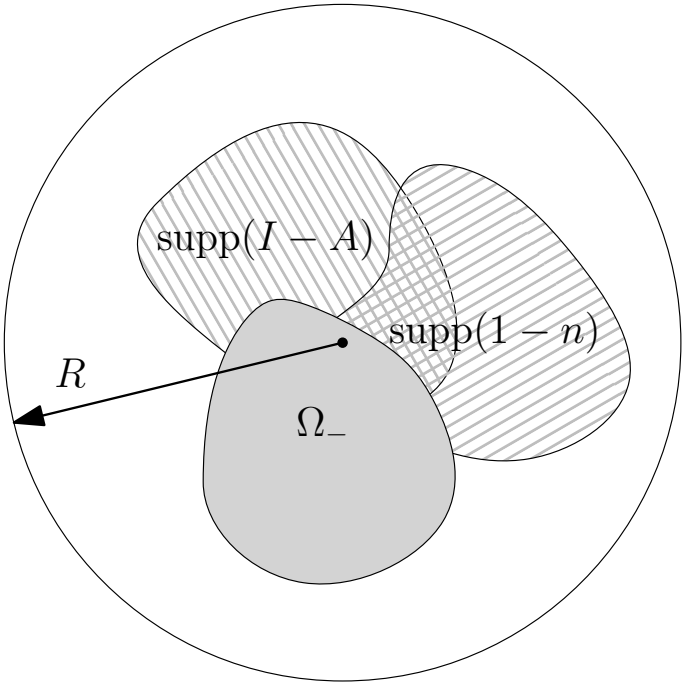

Figure 1.1: A schematic of Ω_-, the supports of $I - A$ and $1 - n$, and B_R.

Remark 1.8. (Energy.) *If $u(x)$ is a solution of the Helmholtz equation in Ω_+, then $\Im\langle \partial_\nu u, \gamma u \rangle_{\Gamma_R}$ (where $\boldsymbol{\nu} = \widehat{x}$) is proportional to the flux in the $\widehat{x}$ direction across Γ_R of the energy of the wave equation solution $u(x)\exp(-\mathrm{i}\omega t)$; see, e.g., [136, Page 262]. The sign property (1.12) is therefore another illustration (on top of that in Remark 1.5) that the Sommerfeld radiation condition implies that the associated solution $u(x)\exp(-\mathrm{i}\omega t)$ of the wave equation corresponds to a wave moving away from Ω_- towards infinity; indeed, if $\Im\langle \partial_\nu u, \gamma u \rangle_{\Gamma_R} > 0$, then energy is moving across Γ_R in the positive r direction.*

1.4 Variational formulation of the EDP with zero Dirichlet data

This formulation is based on Green's identity.

Lemma 1.9. (Green's identity (Version 1).) *Let Ω be a bounded Lipschitz open set with outward-pointing unit normal vector field $\boldsymbol{\nu}$. If $A \in C^{0,1}(\Omega, \mathbb{R}^{d\times d})$, $u \in H^2(\Omega)$, and $v \in H^1(\Omega)$, then*

$$\int_{\partial\Omega} \boldsymbol{\nu} \cdot \gamma(A\nabla u)\,\overline{\gamma v} = \int_\Omega (A\nabla u)\cdot\overline{\nabla v} + \overline{v}\nabla\cdot(A\nabla u). \tag{1.14}$$

Proof. First assume that $v \in C^\infty(\overline{\Omega}) := \{w|_{\overline{\Omega}} : w \in C^\infty(\mathbb{R}^d)\}$. Then (1.14) follows from applying the divergence theorem

$$\int_\Omega \nabla\cdot\mathbf{F} = \int_{\partial\Omega} \boldsymbol{\nu}\cdot\gamma\mathbf{F}, \qquad \text{for all } \mathbf{F} \in H^1(\Omega, \mathbb{C}^{d\times d}), \tag{1.15}$$

with $\mathbf{F} = vA\nabla u$, which is indeed in H^1 since the product of a Lipschitz function and an H^1 function is H^1. Since $C^\infty(\overline{\Omega})$ is dense in $H^1(\Omega)$ and (1.14) is continuous in v with respect to the $H^1(D)$ norm, (1.14) holds for all $v \in H^1(\Omega)$. □

Lemma 1.10. (Green's identity (Version 2) and conormal derivative.) *Let Ω be a bounded Lipschitz open set with outward-pointing unit normal vector field $\boldsymbol{\nu}$. If $A \in C^{0,1}(\Omega, \mathbb{R}^{d\times d})$, $u \in H^1(\Omega)$, and $\nabla\cdot(A\nabla u) \in L^2(\Omega)$ (understood as in (1.5)[1]) then there exists a uniquely defined $\varphi \in H^{-1/2}(\partial\Omega)$ such that*

$$\langle\varphi, \gamma v\rangle_{\partial\Omega} = \int_\Omega (A\nabla u)\cdot\overline{\nabla v} + \overline{v}\,\nabla\cdot(A\nabla u) \qquad \textit{for all } v \in H^1(\Omega), \tag{1.16}$$

where $\langle\cdot,\cdot\rangle_{\partial\Omega}$ denotes the duality pairing on $\partial\Omega$ that is linear in the first argument and antilinear in the second argument. Furthermore, if $u \in H^2(\Omega)$ then $\varphi = \boldsymbol{\nu}\cdot\gamma(A\nabla u)$ and thus we denote φ by $\partial_{\nu,A}u$.

Proof. Define $\varphi \in H^{-1/2}(\partial\Omega)$ by

$$\langle\varphi,\psi\rangle_{\partial\Omega} := \int_\Omega (A\nabla u)\cdot\overline{\nabla(\eta\psi)} + \overline{\eta\psi}\nabla\cdot(A\nabla u) \qquad \text{for all } \psi \in H^{1/2}(\Gamma), \tag{1.17}$$

where $\eta : H^{1/2}(\partial\Omega) \to H^1(\Omega)$ is a continuous right inverse to the trace operator $\gamma : H^1(\Omega) \to H^{1/2}(\partial\Omega)$; i.e. $\gamma\eta\psi = \psi$ for all $\psi \in H^{1/2}(\partial\Omega)$. We now need to show that (1.16) holds.

We first prove that if $u \in H^1(\Omega)$ with $\nabla\cdot(A\nabla u) \in L^2(\Omega)$ and $w \in H^1_0(\Omega)$,

$$\int_\Omega A\nabla u\cdot\overline{\nabla w} = -\int_\Omega \overline{w}\nabla\cdot(A\nabla u). \tag{1.18}$$

By the definition of the weak derivative, for all $\phi \in C^\infty_{\text{comp}}(\Omega)$,

$$\int_\Omega \overline{\phi}\nabla\cdot(A\nabla u) = \int_\Omega u\nabla\cdot(A\overline{\nabla\phi}) = -\int_\Omega (A\nabla u)\cdot\overline{\nabla\phi}, \tag{1.19}$$

where for the last equality we have used the divergence theorem (1.15) with $\mathbf{F} = uA\nabla\phi$, which is in H^1 (again since the product of a Lipschitz function and an H^1 function is H^1). Since $C^\infty_{\text{comp}}(\Omega)$ is dense in $H^1_0(\Omega)$ and (1.19) is continuous in v with respect to the $H^1(\Omega)$ norm, (1.18) holds.

To prove (1.16), we apply (1.17) with $\psi = \gamma v$ to obtain that

$$\langle\varphi,\gamma v\rangle_\Gamma = \int_\Omega (A\nabla u)\cdot\overline{\nabla(\eta\gamma v)} + \overline{\eta\gamma v}\nabla\cdot(A\nabla u). \tag{1.20}$$

We then apply (1.18) with $w = v - \eta\gamma v$, which is in $H^1_0(\Omega)$ because $\gamma\eta = I$, to obtain that

$$\int_\Omega (A\nabla u)\cdot\overline{\nabla\big(v-\eta\gamma v\big)} = -\int_\Omega \overline{\big(v-\eta\gamma v\big)}\nabla\cdot(A\nabla u). \tag{1.21}$$

The result (1.16) then follows from adding (1.20) and (1.21). The result that $\varphi = \boldsymbol{\nu}\cdot\gamma(A\nabla u)$ when $u \in H^2(\Omega)$ follows from comparing (1.16) and (1.14). □

[1] That is, there exists $g \in L^2(\Omega)$ such that $\int_\Omega u\nabla\cdot(A\nabla\overline{\phi}) = \int_\Omega g\overline{\phi}$ for all $\phi \in C^\infty_{\text{comp}}(\Omega)$.

Definition 1.11. (Variational formulation of EDP with $g_D = 0$.) *Given Ω_-, A, and n satisfying Assumption 1.1 and $f \in L^2(\Omega_+)$ with $\operatorname{supp} f$ bounded, choose $R > 0$ such that $\overline{\Omega_-} \cup \operatorname{supp}(I - A) \cup \operatorname{supp}(1 - n) \cup \operatorname{supp} f \Subset B_R$. Let*

$$H^1_{0,D}(\Omega_R) := \big\{ v \in H^1(\Omega_R) : \gamma v = 0 \text{ on } \Gamma_D \big\}. \tag{1.22}$$

The variational formulation of the EDP of Definition 1.2 with $g_D = 0$ is

$$\text{find } \widetilde{u} \in H^1_{0,D}(\Omega_R) \text{ such that} \quad a(\widetilde{u}, v) = F(v) \quad \text{for all } v \in H^1_{0,D}(\Omega_R), \tag{1.23}$$

where

$$a(\widetilde{u}, v) := \int_{\Omega_R} \Big(k^{-2} (A \nabla \widetilde{u}) \cdot \overline{\nabla v} - n u \overline{v} \Big) - k^{-1} \langle \mathrm{DtN}_k \gamma \widetilde{u}, \gamma v \rangle_{\Gamma_R} \quad \text{and} \quad F(v) := \int_{\Omega_R} f \, \overline{v}. \tag{1.24}$$

Lemma 1.12. (Equivalence of the formulations.)

(i) If u is a solution of the Helmholtz EDP of Definition 1.2 with $g_D = 0$, then $u|_{\Omega_R}$ is a solution of the variational problem (1.23) with $a(\cdot,\cdot)$ and $F(\cdot)$ as in (1.24). Conversely, if $\widetilde{u}$ is a solution of this variational problem, then there exists a solution u of the Helmholtz EDP of Definition 1.2 with $g_D = 0$ such that $u|_{\Omega_R} = \widetilde{u}$.

(ii) If u is a solution of the plane-wave sound-soft scattering problem of Definition 1.3, then $u|_{\Omega_R}$ is a solution of the variational problem (1.23) with $a(\cdot,\cdot)$ as in (1.24) and

$$F(v) := k^{-1} \int_{\Gamma_R} \left(k^{-1} \frac{\partial u^I}{\partial r} - \mathrm{DtN}_k(\gamma u^I) \right) \overline{\gamma v}. \tag{1.25}$$

Conversely, if $\widetilde{u}$ is a solution of this variational problem, then there exists a solution of the plane-wave sound-soft scattering problem of Definition 1.3 such that $u|_{\Omega_R} = \widetilde{u}$.

Proof. The proof of Part (i) is Exercise 2 in §1.6; the proof of Part (ii) is similar. □

Lemma 1.13. (Continuity of the sesquilinear form.) *Given $k_0, R_0 > 0$, for all $k \geq k_0$ and $R \geq R_0$,*

$$|a(u, v)| \leq C_{\mathrm{cont}} \, \|u\|_{H^1_k(B_R)} \, \|v\|_{H^1_k(B_R)} \quad \text{for all } u, v \in H^1(B_R), \tag{1.26}$$

where

$$C_{\mathrm{cont}} := \max\{A_{\max}, n_{\max}\} + C_{\mathrm{DtN}1}. \tag{1.27}$$

Proof of Lemma 1.13. This follows from the Cauchy-Schwarz inequality, the definition of $\|\cdot\|_{H^1_k(\Omega_R)}$ (1.8), and the inequalites (1.1), (1.2), and (1.11). □

1.5 Wellposedness of the EDP

Theorem 1.14. (Unique continuation principle (UCP).) *Suppse that Ω is Lipschitz, $A \in C^{0,1}(\Omega, \mathrm{SPD})$, $n \in L^\infty(\Omega, \mathbb{R})$, and u satisfies $k^{-2} \nabla \cdot (A \nabla u) + n u = 0$ (in the sense of (1.5)). If $u = 0$ on $B_r(x_0)$ for some $r > 0$ and $x_0 \in \Omega$ such that $B_r(x_0) \Subset \Omega_R$, then $u = 0$ in Ω.*

References for the proof. This follows from the unique continuation results of [66, 91] (for Lipschitz A) and [87, 150] (for $n \in L^{3/2}$); see, e.g., [71, Theorem 2.1]. □

Remark 1.15. *Actually, in 2-d the UCP holds when A is L^∞ and $n \in L^p$ for some $p > 1$ [3]. An example of an $A \in C^{0,\alpha}$ for all $\alpha < 1$ for which the UCP fails in 3-d is given in [57]. Nevertheless, the UCP can be extended from Lipschitz A to piecewise-Lipschitz A by the Baire-category argument in [13] (see also [101, Proposition 2.11]).*

Lemma 1.16. *If Ω_-, A, and n satisfy Assumption 1.1, then the solution to the variational problem (1.23) is unique.*

Proof. This is Exercise 3 in §1.6. □

Theorem 1.17. (Wellposedness of the EDP.) *The EDP of Definition 1.2 has a unique solution which depends continuously on the data.*

Proof. We first consider the case when $g_D = 0$. By Lemma 1.12, it is sufficient to prove that the variational problem (1.23) has a unique solution which depends continuously on the data. By the inequalities (1.13), (1.1), and (1.2), and the definition of $\|\cdot\|_{H^1_k(\Omega_R)}$ (1.8), for any $v \in H^1_{0,D}(\Omega_R)$,

$$\Re a(v,v) \geq A_{\min} k^{-2} \|\nabla v\|^2_{L^2(\Omega_R)} - n_{\max} \|v\|^2_{L^2(\Omega_R)} = A_{\min} \|v\|^2_{H^1_k(\Omega_R)} - (n_{\max} + A_{\min}) \|v\|^2_{L^2(\Omega_R)}; \tag{1.28}$$

i.e., $a(\cdot,\cdot)$ satisfies a Gårding inequality. Since the sesquilinear form is continuous (by Lemma 1.13) and satisfies a Gårding inequality, Fredholm theory implies that existence of a solution to the variational problem and continuous dependence of the solution on the data both follow from uniqueness; see, e.g., [106, Theorem 2.34], [53, §6.2.8], [135, Theorem 6.31].

When $g_D \neq 0$, let $\chi \in C^\infty_{\rm comp}(B_R)$ with $\chi = 1$ on $B_{R'}$ for some R' satisfying $\operatorname{diam}(\Omega_-) < R' < R$. Then, given u satisfying the EDP with $g_D \not\equiv 0$, $u - \chi\eta g_D$ satisfies the EDP with zero Dirichlet data and with suitably modified f and g (the bounded support of $\chi\eta g_D$ ensures that $u - \chi\eta g_D$ satisfies the radiation condition). Existence of the solution of the EDP then follows from the case $g_D = 0$. □

Remark 1.18. (Transmission problems.) *Definition 1.2 allows for discontinuous n, but not discontinuous A. The only change needed to allow $A \in L^\infty(\Omega_+, \mathrm{SPD})$ in Assumption 1.1 (instead of $A \in C^{0,1}(\Omega_+, \mathrm{SPD})$) is for the PDE in (1.3) to be understood in the sense that*

$$\int_{\Omega_+} \Big(-k^{-2} \nabla u \cdot (A \overline{\nabla \phi}) + n u \overline{\phi} \Big) = -\int_{\Omega_+} f\, \overline{\phi} \quad \text{for all } \phi \in C^\infty_{\rm comp}(\Omega_+). \tag{1.29}$$

Lemma 1.10 then holds with $\nabla \cdot (A \nabla u) \in L^2(\Omega)$ understood as in (1.29). Wellposedness then follows as above, using the fact that a UCP holds for piecewise Lipschitz A – see the references in Remark 1.15.

1.6 Exercises

1. Prove Part (ii) of Lemma 1.7. Hint: use Lemma 1.4 and Green's identity.
2. Prove Part (i) of Lemma 1.12. Hint: use Green's identity and the fact that a piecewise H^1 function is globally H^1 if it is continuous.
3. Prove Lemma 1.16. Hint: use Part (ii) of Lemma 1.7.

2 The k-dependence of the Helmholtz solution operator

2.1 The operator norm of the Helmholtz solution operator

Definition 2.1. *Let $C_{\rm sol}(k, A, n, \Omega_-, R)$ be the operator norm of the map $L^2(\Omega_R) \ni f \mapsto u \in H^1(\Omega_R)$, where u is the solution of the variational problem (1.23) (i.e., the EDP with $g_D = 0$); i.e.,*

$$C_{\rm sol}(k, A, n, \Omega_-, R) := \sup_{\substack{f \in L^2(B_R) \\ \|f\|_{L^2(\Omega_R)} = 1}} \|u\|_{H^1_k(\Omega_R)}.$$

This definition implies that the solution of the EDP with $g_D = 0$ satisfies

$$\|u\|_{H^1_k(\Omega_R)} \leq C_{\rm sol} \|f\|_{L^2(\Omega_R)}. \tag{2.1}$$

Theorem 1.17 implies that $C_{\rm sol} < \infty$, and, more generally, that the map $(H^1_{0,D}(\Omega_R))' \ni f \mapsto u \in H^1_{0,D}(\Omega_R)$ is bounded. The following lemma gives a bound on this latter map in terms of $C_{\rm sol}$.

Lemma 2.2. *Given $F \in (H^1_{0,D}(\Omega_R))'$, let u be the solution of the variational problem $a(u,v) = F(v)$ for all $v \in H^1_{0,D}(\Omega_R)$, where $a(\cdot,\cdot)$ is given by (1.24). Then u satisfies*

$$\|u\|_{H^1_k(\Omega_R)} \leq \frac{(1 + 2C_{\rm sol} n_{\max})}{\min\{A_{\min}, n_{\min}\}} \|F\|_{(H^1_k(\Omega_R))'},$$

where

$$\|F\|_{(H^1_k(\Omega_R))'} := \sup_{v\in H^1_{0,D}(\Omega_R)\setminus\{0\}} \frac{|F(v)|}{\|v\|_{H^1_k(\Omega_R)}}.$$

Proof. This is Exercise 1 in §2.5 (with this result going back to at least [33, Text between Lemmas 3.3 and 3.4]). □

2.2 The k-dependence of C_{sol}

There could be a whole lecture course studying the k-dependence of C_{sol}; it would be a great course, but it is not this course. The goal of this section is to summarise the results about how C_{sol} behaves when k is large that we need later. The summary is that

- C_{sol} grows at least linearly in k (Lemma 2.4) and at most exponentially in k (Theorem 2.5).
- Linear growth is attained when Ω_-, A, and n are *nontrapping* (Theorem 2.7).
- Exponential growth is attained through a sequence of ks when one of Ω_-, A, and n is such that the problem has the strongest form of trapping (Theorem 2.9); however, for most frequencies C_{sol} is polynomially bounded in k (Theorem 2.11).

Although this section is focused on recapping rather than proving, I believe that everyone studying the Helmholtz equation should have proved at least one k-explicit upper bound on C_{sol} in their lives! Theorem 2.14 therefore records the bound on C_{sol} when $A = I, n = 1$, and Ω_- is star-shaped proved by Morawetz (using only integration by parts), and the proof of this is Exercise 3 in §2.5.

Remark 2.3. **($k\to\infty$ vs $kR\to\infty$.)** *Since k has units (length)$^{-1}$ but kR is nondimensional, it makes more sense to talk about the limit $kR\to\infty$ than the limit $k\to\infty$, and to write bounds in terms of kR (or k multiplied by some other parameter with dimension length) instead of k. However, since we are focused on the case when R is fixed and k can be arbitrarily large, we usually work with k instead of kR (the next lemma and Theorem 2.14 being the only exceptions).*

Lemma 2.4. **(C_{sol} grows at least linearly in k with no scatterer.)** *If $A = I, n = 1, \Omega_- = \emptyset$, then, given $k_0, R_0 > 0$, there exists $C > 0$ such that $C_{\mathrm{sol}} \geq CkR$ for all $k \geq k_0$ and $R \geq R_0$.*

Proof. This is Exercise 2 in §2.5. □

Theorem 2.5. (Exponential upper bound on C_{sol}.) *If Ω_-, A, and n are all C^∞, then given $k_0 > 0$ there exists $C, \alpha > 0$ such that, for all $k \geq k_0$,*

$$C_{\mathrm{sol}} \leq C\exp(\alpha k).$$

References for the proof. This was first proved in [25, Theorem 2] using Carleman estimates; for a proof of the analogue of this result for scattering by a potential (involving the operator $-k^{-2}\Delta + V - z$)), see [52, Theorem 6.25]. □

The key geometric conditions that govern the k-dependence of C_{sol} are those of *trapping* and *nontrapping*. We now define nontrapping for the case when $\Omega_- = \emptyset$; the definition for general smooth Ω_- is much more technical (and requires the notion of the Melrose-Sjöstrand generalised bicharacteristic flow; see [110, 111], [79, §24.3]), and when Ω_- has corners more technical still (see [16] for the notion of a *nontrapping polygon*).

Definition 2.6. (Nontrapping A and n.) *Suppose A and n satisfy Assumption 1.1 and are additionally both $C^{1,1}$. Let $R > 0$ be such that $\mathrm{supp}(I - A) \cup \mathrm{supp}(1 - n) \Subset B_R$. Consider the solutions $(x(s), \xi(s)) \in \mathbb{R}^d \times \mathbb{R}^d$ of the Hamiltonian system*

$$\frac{\mathrm{d}x_i}{\mathrm{d}s}(s) = \frac{\partial}{\partial\xi_i}H\big(x(s),\xi(s)\big), \qquad \frac{\mathrm{d}\xi_i}{\mathrm{d}s}(s) = -\frac{\partial}{\partial x_i}H\big(x(s),\xi(s)\big), \tag{2.2}$$

satisfying $H(x(s),\xi(s))=0$, where the Hamiltonian $H(x,\xi)$ is given by

$$H(x,\xi) := \sum_{i=1}^{d}\sum_{j=1}^{d} A_{ij}(x)\xi_i\xi_j - n(x). \tag{2.3}$$

We say that A and n are nontrapping *if there exists $S(R)>0$ such that all solutions of (2.2) with $|x(0)|<R$ satisfy $|x(s)|>R$ for all $s\geq S(R)$.*

Four remarks: (i) if $A=I$ and $n=1$, then $H=|\xi|^2-1$ and (2.2) becomes $\dot{x}_i=2\xi_i$ and $\dot{\xi}_i=0$, with solution $x=x_0+2s\xi_0$, $\xi=\xi_0$ i.e., straight-line motion with speed 2, (ii) the projections in x of the solutions of (2.2) are the *rays* of the Helmholtz equation, (iii) we see later that the significance of the Hamiltonian (2.3) is that it is the *semiclassical principal symbol* of the Helmholtz equation, and (iv) the requirement that A and n are both $C^{1,1}$ means that the coefficients of the ODE system (2.2) are Lipschitz, and then the solutions of (2.2) exist by the Picard–Lindelöf/Cauchy–Lipschitz theorem (see, e.g., [8, §31]).

Theorem 2.7. (Nontrapping bound on C_{sol}.) *If $\Omega_-=\emptyset$ and A and n are both $C^{1,1}$ and nontrapping in the sense of Definition 2.6, then given $k_0>0$ there exists $C>0$ such that, for all $k\geq k_0$,*

$$C_{\mathrm{sol}} \leq Ck. \tag{2.4}$$

References for the proof. This is proved in [64] using the defect-measure argument of [26, Theorem 1.3 and §3]. When A and n are both C^∞ and nontrapping, the bound (2.4) follows from the results about propagation of singularities of the wave equation in [51, §VI] combined by [147, Theorem 3]/ [148, Chapter 10, Theorem 2] or [97]. (The argument in [147] takes results about propagation of singularities for the wave equation, and outputs a bound on C_{sol}; see, e.g., the account of this argument in [52, Theorem 4.43]). □

Remark 2.8. (The constant in the nontrapping bound on C_{sol}.) *In fact, [64] shows that, if k_0 is sufficiently large and one works in an H^1_k norm weighted with A and n, then the constant C in (2.4) is a multiple of the length of the longest ray in B_R; see [64, Theorem 1 and Equation 6.32].*

We write $a\lesssim b$ if there exists $C>0$, independent of k, such that $a\leq Cb$. We write $a\gtrsim b$ if $b\lesssim a$, and $a\sim b$ if $a\lesssim b$ and $a\gtrsim b$.

In the rest of the course, we will informally describe the case when $C_{\mathrm{sol}}\lesssim k$ as "nontrapping". We emphasise however, that there exist A,n, and Ω_- for which $C_{\mathrm{sol}}\lesssim k$, but for which the concept of nontrapping is not well defined because either the coefficients A,n or the domain Ω_- are too rough; an example of the former situation is given in [70, Theorem 2.7], an example of the latter situation is given in Theorem 2.14 below.

Theorem 2.9. (Exponential blow up of C_{sol} through a sequence of ks for trapping Ω_-.) *Suppose $d=2$, $A=I$ and $n=1$. Given $a_1>a_2>0$, let*

$$E := \left\{(x_1,x_2) \,:\, \left(\frac{x_1}{a_1}\right)^2+\left(\frac{x_2}{a_2}\right)^2<1\right\}. \tag{2.5}$$

Assume that Γ_D coincides with the boundary of E in the neighbourhoods of the points $(0,\pm a_2)$, and that $\overline{\Omega_+}$ contains the convex hull of the union of these neighbourhoods (see, e.g., Figure 2.1).

Then there exists $C_1,C_2>0$ and $\{k_j\}_{j=1}^\infty$ with $k_j\to\infty$ as $k\to\infty$ such that, for all j,

$$C_{\mathrm{sol}}(k_j)\geq C_1\exp(C_2k_j).$$

References for the proof. This is proved in [19, Equation A.16]; the idea is that there exists families of Laplace eigenfunctions of the ellipse that localise exponentially around the minor axis. These eigenfunctions, chopped off with suitable cut-off functions, form functions through which the exponential growth of $C_{\mathrm{sol}}(k)$ is attained (with k_j such that k_j^2 is the appropriate eigenvalue). This exponential localisation of Laplace eigenfunctions is also proved in [119, Theorem 3.1]. □

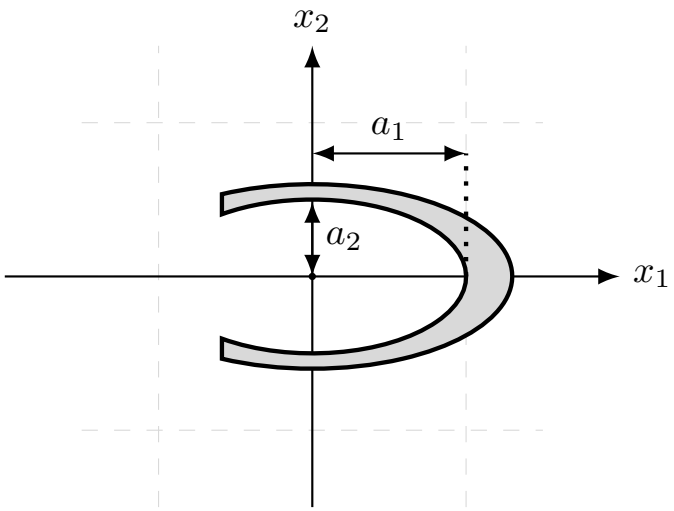

Figure 2.1: An example of an Ω_- satisfying the conditions of Theorem 2.9.

Remark 2.10. (Other results on the blow up of C_{sol}.) *Superalgebraic blow up of C_{sol} through a sequence of ks for more general Ω_- than in Theorem 2.9 is proved in [30, Theorem 1].*

An example where the trapping is created by smooth coefficients (as opposed to the obstacle) is given in [125]: here C_{sol} grows exponentially through a sequences of ks, $\Omega_- = \emptyset$, $A = I$, and n is C^∞ and spherically symmetric. For an example where trapping is caused by discontinuous coefficients, see [124, 28, 29, 2] (with these results summarised for an "applied" audience in [113, §6]).

The frequencies $\{k_j\}_{j=1}^\infty$, together with the (compactly-supported) functions through which the blow-up occurs, are called quasimodes. *The existence of quasimodes is linked to the existence of* resonances, *i.e., poles of the meromorphic continuation of the solution operator of (0.1) from $\Im k \geq 0$ to $\Im k < 0$. The relationship between trapping, resonances, and quasimodes is a classic topic in scattering theory; see [140, 141, 144, 137, 138] and [52, Chapter 7].*

Theorem 2.11. (C_{sol} is polynomially bounded for most frequencies.) *If Ω_-, A, and n satisfy Assumption 1.1 then given $k_0 > 0$ and $\delta > 0$ there exists a set $J \subset [k_0, \infty)$ with $|J| \leq \delta$ such that, for any $\varepsilon > 0$, there exists $C > 0$ such that*

$$C_{\mathrm{sol}}(k) \leq C k^{5d/2+1+\varepsilon} \qquad \text{for all } k \in [k_0, \infty) \setminus J.$$

If both Ω_- and A are $C^{1,\sigma}$ for some $\sigma > 0$ then the exponent is reduced to $5d/2 + \varepsilon$.

The result of Theorem 2.11 holds for a much wider range of scattering problems, namely those fitting in the framework of *black-box scattering* introduced in [133] (see [52, Chapter 4]); see [95, Theorem 1.1].

References for the proof of Theorem 2.11. This is proved for $n = 1$ in [95, Theorem 1.1 and Corollary 3.6]; the proof for more-general n follows from [94, Lemma 2.3] (which shows that the EDP of Definition 1.2 fits into the framework of black-box scattering).

Under an additional assumption about the location of resonances, a similar result to Theorem 2.11 with a larger exponent can also be extracted from [139, Proposition 3] by using the Markov inequality. □

As mentioned above, the definition of nontrapping Ω_- is technical because one needs to define reflection of the rays from Γ_D. Nevertheless, a geometric condition that implies nontrapping (at least for smooth domains) is that of star-shapedness.

Definition 2.12. (Star-shaped and star-shaped with respect to a ball.)
(i) Ω *is* star-shaped with respect to the point x_0 *if the segment* $[x_0, x] \subset \Omega$ *for all* $x \in \Omega$.
(ii) Ω *is* star-shaped with respect to the ball $B_a(x_0)$ *if it is star-shaped with respect to every point in* $B_a(x_0)$.

These definitions make sense even for non-Lipschitz Ω, but when Ω is Lipschitz one can characterise star-shapedness with respect to a point or ball in terms of $(x - x_0) \cdot \boldsymbol{\nu}(x)$ for $x \in \partial\Omega$, where $\boldsymbol{\nu}(x)$ is the outward-pointing unit normal vector at $x \in \partial\Omega$.

Lemma 2.13. ([112, Lemma 5.4.1]) *(i) If* Ω *is Lipschitz (with outward-pointing unit normal vector field* $\boldsymbol{\nu}(x)$*), then it is star-shaped with respect to* x_0 *if and only if* $(x - x_0) \cdot \boldsymbol{\nu}(x) \geq 0$ *for all* $x \in \partial\Omega$ *for which* $\boldsymbol{\nu}(x)$ *is defined.*
(ii) If Ω *is Lipschitz, then* Ω *is star-shaped with respect to* $B_a(x_0)$ *if and only if* $(x - x_0) \cdot \boldsymbol{\nu}(x) \geq a$ *for all* $x \in \partial\Omega$ *for which* $\boldsymbol{\nu}(x)$ *is defined.*

Theorem 2.14. (Morawetz bound on C_{sol} for star-shaped Ω_-.) *If* $A = I$, $n = 1$, *and* Ω_- *is Lipschitz and star-shaped with respect to the origin, then, for all* $k > 0$ *and* $R > \operatorname{diam}(\Omega_-)$,

$$C_{\mathrm{sol}} \leq 2kR\sqrt{1 + \left(\frac{d-1}{2kR}\right)^2}. \tag{2.6}$$

Proof. This is Exercise 3 in §2.5. □

Remark 2.15 (The idea behind the proof of Theorem 2.14)**.** *The proof of Theorem 2.14 is based on the following identity: if*

$$\mathcal{L}v := k^{-2}\Delta v + v \quad \text{and} \quad \mathcal{M}_{\beta,\alpha}v := x \cdot \nabla v - \mathrm{i}k\beta v + \alpha v,$$

with β *and* α *real-valued functions, then*

$$\begin{aligned} 2\Re\big(\overline{\mathcal{M}_{\beta,\alpha}v}\,\mathcal{L}v\big) = \nabla \cdot \Big[2k^{-1}\Re\big(\overline{\mathcal{M}_{\beta,\alpha}v}\,k^{-1}\nabla v\big) + \big(|v|^2 - k^{-2}|\nabla v|^2\big)x\Big] \\ - 2\Re\big(\overline{v}\,(\mathrm{i}\nabla\beta + k^{-1}\nabla\alpha)\cdot k^{-1}\nabla v\big) - \big(d - 2\alpha\big)|v|^2 - \big(2\alpha - d + 2\big)k^{-2}|\nabla v|^2. \end{aligned} \tag{2.7}$$

The idea of multiplying second-order PDEs with first-order expressions has been used by many authors; multiplying Δv *by a derivative of* v *goes back to Rellich [126, 127], and multiplying* $\nabla \cdot (A\nabla v)$ *by a derivative of* v *goes back to Hörmander [77] and Payne and Weinberger [121] (e.g., the identity (2.7) with* α *and* β *equal zero appears as [121, Equation 2.4]). These identities have been independently discovered by, e.g., Jerison and Kenig [88, 90, 89] and Pohozaev [123].*

In the context of the Helmholtz equation, the identity (2.7) with x *replaced by a general vector field, and* α *and* β *replaced by general scalar fields was the heart of Morawetz's paper [115], following both the earlier work by Morawetz and Ludwig [116] using a special case of (2.7) (see (2.18) below) and Morawetz's earlier work on the wave equation [114].*

Similar identities are available with A *and* n *variable; the analogous identity to (2.7) with* A *variable and* $n \equiv 1$ *was used by Bloom in [20], and the one with* $A \equiv I$ *and variable* n *was used in Bloom and Kazarinoff in [21]. Although these identities have a long history, the recent papers [113], [70], [39] contain new results about the variable-coefficient Helmholtz equation obtained with these identities, and [34] contains new results about the constant-coefficient Helmholtz equation.*

The reason why *the identity (2.7) can be used to prove a bound on* C_{sol} *can be understood using semiclassical analysis; this is not a focus of this lecture course, but we refer the interested reader to [70, Section 7] for an introduction to this.*

2.3 Bounding the H^2 norm of the solution of the EDP

Given Ω_- and A satisfying Assumption 1.1, with Ω_- in addition $C^{1,1}$, let $R > 0$ such that $\overline{\Omega_-} \cup \operatorname{supp}(I - A) \Subset B_R$. By elliptic regularity (see, e.g., [106, Theorem 4.18], [72, Theorem 2.4.2.5], [53, §6.3.2]), there exists $C > 0$ (depending on the $W^{1,\infty}$ norm of A) such that

$$|v|_{H^2(\Omega_R)} \leq C\Big(\|\nabla \cdot (A\nabla v)\|_{L^2(\Omega_{R+1})} + R^{-1}\|\nabla v\|_{L^2(\Omega_{R+1})} + R^{-2}\|v\|_{L^2(\Omega_{R+1})} \Big)$$

so that

$$k^{-2}|u|_{H^2(\Omega_R)} \le C\Big[C_{\rm sol}(k,A,n,\Omega_-,R+1)\Big(n_{\max}+(kR)^{-1}+(kR)^{-2}\Big)+1\Big]\,\|f\|_{L^2(\Omega_R)}\,.$$

Combining this last inequality with (2.1), we can bound $\|u\|_{H^2_k(\Omega_R)}$ in terms of $\|f\|_{L^2(\Omega_R)}$ with a constant involving both $C_{\rm sol}(k,A,n,\Omega_-,R)$ and $C_{\rm sol}(k,A,n,\Omega_-,R+1)$. We could certainly live with this; however, the FEM theory in §3 requires the following Poisson regularity result, and using this result we can bound $k^{-2}|u|_{H^2(\Omega_R)}$ in terms of $\|f\|_{L^2(\Omega_R)}$ with a constant involving only $C_{\rm sol}(k,A,n,R)$ (see (2.10) below).

Lemma 2.16. (H^2 regularity of Poisson's equation with DtN_k boundary condition.) *Given Ω_- and A satisfying Assumption 1.1, with Ω_- in addition $C^{1,1}$, let $R>0$ be such that $\overline{\Omega_-}\cup\operatorname{supp}(I-A)\Subset B_R$. Then there exists C_{H^2} such that given $f\in L^2(\Omega_R)$ there exists $v\in H^1_{0,D}(\Omega_R)$ satisfying*

$$\nabla\cdot(A\nabla v)=-\widetilde{f}\ \text{ in }\Omega_R,\qquad \text{and}\qquad k^{-1}\partial_\nu v=\mathrm{DtN}_k(\gamma v)\ \text{ on }\Gamma_R,\tag{2.8}$$

and, for all $k>0$,

$$|v|_{H^2(\Omega_R)}\le C_{H^2}\left(\|\widetilde{f}\|_{L^2(\Omega_R)}+R^{-1}\,\|\nabla v\|_{L^2(\Omega_R)}+R^{-2}\,\|v\|_{L^2(\Omega_R)}\right).\tag{2.9}$$

The key point in (2.9) is that, although v in (2.8) depends on k via the boundary condition on Γ_R, C_{H^2} is independent of k.

References for the proof of Lemma 2.16. This is proved in [96, Theorem 6.1] following the proof of the analogous result [38, Theorem 3.1] when DtN_k is replaced by $\mathrm{i}k$, $A=I$, and $\Omega_-=\emptyset$. Both these proofs rely on ideas from [72, §3] used to prove H^2 regularity results on C^2 convex domains; see [72, Theorem 3.1.1.1 and proof of Theorem 3.1.2.3] and Exercise 4 in §2.5 below. □

Corollary 2.17. *Suppose that Ω_-, A, and n satisfy Assumption 1.1 and, in addition, Ω_- is $C^{1,1}$. The solution u of the EDP of Definition 1.2 with $g_D=0$ is in $H^2(\Omega_R)$ and, if $kR\ge 1$, then*

$$|u|_{H^2_k(\Omega_R)}\le C_{H^2}\Big[C_{\rm sol}(k,A,n,\Omega_-,R)\Big(n_{\max}+\sqrt{2}\Big)+1\Big]\,\|f\|_{L^2(\Omega_R)}\,,\tag{2.10}$$

where the weighted semi-norm $|u|_{H^2_k(\Omega_R)}:=k^{-2}|u|_{H^2(\Omega_R)}$ (in analogy with (1.7)).

2.4 Using Green's identity to bound the L^2 norm of ∇u in terms of the L^2 norm of u (and vice versa) plus norms of the data and traces

Using Green's identity in this way is well-known, see, e.g., [115, Theorem I.1] and [134, Lemma 2.2], and the idea is similar to that of Caccioppoli inequalities in the Calculus of Variations.

Lemma 2.18. (Bounding the H^1 semi-norm of u via the L^2 norm and L^2 norm of f.) *Assume there exists a solution to the EDP of Definition 1.2.*

(i) Let $R>0$ be such that $\overline{\Omega_-}\cup\operatorname{supp}(I-A)\cup\operatorname{supp}(1-n)\cup\operatorname{supp}f\Subset\Omega_R$. Then, for all $k>0$,

$$k^{-2}A_{\min}\,\|\nabla u\|^2_{L^2(\Omega_R)}\le\frac{3}{2}n_{\max}\,\|u\|^2_{L^2(\Omega_R)}+\frac{1}{2n_{\max}}\,\|f\|^2_{L^2(\Omega_+)}+k^{-2}\,\|\gamma u\|_{L^2(\Gamma_D)}\,\|\partial_{\nu,A}u\|_{L^2(\Gamma_D)}\,.\tag{2.11}$$

(ii) Let $R>0$ be such that $\overline{\Omega_-}\subset\subset\Omega_R$. Then, for all $k>0$,

$$\begin{aligned}n_{\min}\,\|u\|^2_{L^2(\Omega_R)}\le\frac{4}{k^2}\left(A_{\max}+\frac{6(A_{\max})^2}{n_{\min}k^2}\right)\|\nabla u\|^2_{L^2(\Omega_{R+1})}+\frac{2}{n_{\min}}\,\|f\|^2_{L^2(\Omega_+)}\\ +4k^{-2}\,\|\gamma u\|_{L^2(\Gamma_D)}\,\|\partial_{\nu,A}u\|_{L^2(\Gamma_D)}\,.\end{aligned}\tag{2.12}$$

Proof. (i) Applying Green's identity (1.16) with $\Omega = \Omega_R$, with u the solution of the EDP, and with $v = u$, we obtain

$$-k^{-2}\langle \partial_{\nu,A}u, \gamma u\rangle_{\Gamma_D} + k^{-2}\int_{\Gamma_R} \overline{u}\frac{\partial u}{\partial r} = \int_{\Omega_R} k^{-2}(A\nabla u)\cdot\overline{\nabla u} - n|u|^2 - \overline{u}f, \tag{2.13}$$

where we have used the fact that $u \in C^\infty$ in a neighbourhood of Γ_R (by elliptic regularity) to write the duality pairing on Γ_R as an integral. The key point now is that the inequality (1.13) involving the term on Γ_R in (2.13) allows us to obtain an upper bound on $\int_{\Omega_R}(A\nabla u)\cdot\overline{\nabla u}$. Indeed, taking the real part of (2.13), using the inequality (1.13), and then the Cauchy-Schwarz inequality and the inequalities on A and n (1.1) and (1.2), we obtain that

$$k^{-2}A_{\min}\|\nabla u\|^2_{L^2(\Omega_R)} \le n_{\max}\|u\|^2_{L^2(\Omega_R)} + \|u\|_{L^2(\Omega_R)}\|f\|_{L^2(\Omega_R)} + k^{-2}\|\gamma u\|_{L^2(\Gamma_D)}\|\partial_{\nu,A}u\|_{L^2(\Gamma_D)}.$$

The result (2.11) then follows from using the inequality

$$2\alpha\beta \le \varepsilon\alpha^2 + \varepsilon^{-1}\beta^2 \quad \text{for all } \alpha,\beta,\varepsilon > 0, \tag{2.14}$$

on $\|u\|_{L^2(\Omega_R)}\|f\|_{L^2(\Omega_R)}$.

(ii) The sign property of the inequality (1.13) does not allow us to obtain an upper bound on $\int_{\Omega_R} n|u|^2$ via the argument in Part (i). Instead we apply Green's identity in Ω_{R+1} with u the solution of the EDP and $v = \chi u$, where $\chi(x) := \chi(r)$ is such that $\chi \equiv 1$ on $[0,R]$, $\chi(R+1) = 0$, and $\chi(r) = F(R+1-r)$ for $r \in [R, R+1]$, where $F(t) := t^2(3-2t)$. Observe that F increases from 0 to 1 as t increases from 0 to 1, and thus χ decreases from 1 to 0 as r increases from R to $R+1$. This particular choice of F is motivated by the fact that there exists an $M > 0$ such that

$$\frac{(F'(t))^2}{F(t)} \le M \quad \text{for all } 0 \le t \le 1; \tag{2.15}$$

in fact, one can easily verify that this last inequality holds with $M = 12$.

Applying Green's identity (1.16) as described above we obtain

$$-k^{-2}\langle \partial_{\nu,A}u, \gamma u\rangle_{\Gamma_D} = \int_{\Omega_{R+1}} k^{-2}\chi(A\nabla u)\cdot\overline{\nabla u} + k^{-2}(A\nabla u)\cdot(\overline{u}\,\nabla\chi) - n\chi|u|^2 - \chi\overline{u}f \tag{2.16}$$

(where we use the convention on Γ_D that the normal points *out* of Ω_- and thus *into* Ω_{R+1}); observe that, since $\chi(R+1) = 0$, there is no contribution from Γ_R, and thus we have avoided the issue with the sign in the inequality (1.13). Now, by the Cauchy-Schwarz inequality and (2.14),

$$\left|\int_{\Omega_{R+1}}(A\nabla u)\cdot(\overline{u}\,\nabla\chi)\right| \le \frac{\varepsilon}{2}\int_{\Omega_{R+1}}\chi|u|^2 + \frac{1}{2\varepsilon}\int_{\Omega_{R+1}}\frac{|A\nabla u|^2|\nabla\chi|^2}{\chi}. \tag{2.17}$$

Then, using the second inequality in (1.1), the inequality (2.15) with $M = 12$, and choosing $\varepsilon = n_{\min}k^2$ we obtain that

$$\left|\int_{\Omega_{R+1}}(A\nabla u)\cdot(\overline{u}\,\nabla\chi)\right| \le \frac{n_{\min}k^2}{2}\int_{\Omega_{R+1}}\chi|u|^2 + \frac{6(A_{\max})^2}{n_{\min}k^2}\int_{\Omega_{R+1}}|\nabla u|^2.$$

Using this last inequality in (2.16), we find that

$$\frac{n_{\min}}{2}\int_{\Omega_{R+1}}\chi|u|^2 \le \frac{1}{k^2}\left(A_{\max} + \frac{6(A_{\max})^2}{n_{\min}k^2}\right)\|\nabla u\|^2_{L^2(\Omega_{R+1})} + \int_{\Omega_{R+1}}\chi\overline{u}f + k^{-2}\|\gamma u\|_{L^2(\Gamma_D)}\|\partial_{\nu,A}u\|_{L^2(\Gamma_D)}.$$

Using the Cauchy inequality (2.14) again on the term $\int_{\Omega_{R+1}}\chi\overline{u}f$ with weight $\varepsilon = n_{\min}$, we obtain (2.12). □

Remark. (Dimensions of the factors in (2.11) and (2.12).) *The dimensions of the factors in front of the norms in (2.11) and (2.12) are as expected apart from*

$$\left(A_{\max} + \frac{6(A_{\max})^2}{n_{\min}k^2}\right);$$

this expression should be non-dimensional, but instead the second term has dimension (length)^2. This discrepancy is because there is the factor $1 = ((R+1)-R)^2$ (the distance between B_{R+1} and B_R squared) multiplying the k^2s, providing the missing (length)$^{-2}$.

2.5 Exercises

1. Prove Lemma 2.2. Hint: let $u = u_+ + w$ where u_+ is the solution of the variational problem $a_+(u_+, v) = F(v)$ for all $v \in H^1_{0,D}(\Omega_R)$, where

$$a_+(u,v) := \int_{\Omega_R} \left(k^{-2}(A\nabla u)\cdot\overline{\nabla v} + nu\overline{v}\right) - k^{-1}\langle \mathrm{DtN}_k\gamma u, \gamma v\rangle_{\Gamma_R}.$$

2. Prove Lemma 2.4. Hint: consider $u(x) = \exp(\mathrm{i}kx_1)\chi(r/R)$ for $\chi \in C^\infty_{\mathrm{comp}}(\mathbb{R};\mathbb{R})$ with $\mathrm{supp}\chi \Subset (0,1)$.

3. Via the following steps, prove Theorem 2.14 under the simplifying assumption that Ω_- is $C^{1,1}$ so that $u \in H^2(\Omega_R)$ (for a proof when Ω_- is only C^0, see [33, Lemma 3.8], [70, Remark 2.13]).

   (a) Prove the identity (2.7). Hint: split $\mathcal{M}_{\beta,\alpha}v$ up into its component parts and prove these three identities separately by expanding the divergences on the right-hand sides, and using that

$$2\Re\{\nabla v\cdot(x\cdot\nabla)\,\overline{\nabla v}\} = \nabla\cdot\left[|\nabla v|^2 x\right] - d|\nabla v|^2$$

   and

$$2\Re\{v\,x\cdot\overline{\nabla v}\} = \nabla\cdot\left[|v|^2 x\right] - d|v|^2.$$

   (b) With $\mathcal{L}v$ as above, show that if $\alpha\in\mathbb{R}$, then

$$\begin{aligned}2\Re\{\overline{\mathcal{M}_{r,\alpha}v}\mathcal{L}v\} = \nabla\cdot\left[2k^{-1}\Re\{\overline{\mathcal{M}_{r,\alpha}v}\,k^{-1}\nabla v\} + \left(|v|^2 - k^{-2}|\nabla v|^2\right)x\right] - \left|k^{-1}v_r - \mathrm{i}v\right|^2 \\ + \left(2\alpha-(d-1)\right)\left(|v|^2 - k^{-2}|\nabla v|^2\right) - k^{-2}\left(|\nabla v|^2 - |v_r|^2\right),\end{aligned} \tag{2.18}$$

   where $v_r = x\cdot\nabla v/r$ (this identity first appeared as [116, Equation 1.2]).

   (c) With the identity (2.18) written as $\nabla\cdot\mathbf{Q}_{r,\alpha}(v) = P_{r,\alpha}(v)$, show that if u is an outgoing solution of $\mathcal{L}u = 0$ in $\mathbb{R}^d\setminus\overline{B_{R_0}}$, then, for all $\alpha\in\mathbb{R}$,

$$\int_{\Gamma_{R_1}} \mathbf{Q}_{R_1,\alpha}(u)\cdot\widehat{x} \to 0 \quad \text{as } R_1\to\infty. \tag{2.19}$$

   (d) Show that if u is an outgoing solution of $\mathcal{L}u = 0$ in $\mathbb{R}^d\setminus\overline{B_{R_0}}$, for some $R_0 > 0$, then, for $R > R_0$,

$$\int_{\Gamma_R} \mathbf{Q}_{R,(d-1)/2}(u)\cdot\widehat{x} \le 0. \tag{2.20}$$

   Hint: integrate the identity (2.18) over $B_{R_1}\setminus B_R$. [2]

[2] An analogous inequality to (2.20) holds if Γ_R is replaced by the boundary of a Lipschitz domain that is star-shaped with respect to a ball, and u satisfies the impedance boundary condition $\partial_n u - \mathrm{i}ku = g$ on this boundary; see [70, Lemma A.11]. This is essentially the reason why [11, Page 109], [107, Prop. 8.1.4], [103, Proposition 2.1], [46], [76] were able to independently discover the multiplier $x\cdot\nabla u$ and use it to prove bounds on Helmholtz problems with an impedance boundary condition.

(e) With the identity (2.7) written as $\nabla \cdot \mathbf{Q}_{\beta,\alpha}(v) = P_{\beta,\alpha}(v)$, show that if $u \in H^2(\Omega_R)$ with $\gamma u = 0$ on Γ_D, then

$$\int_{\Gamma_D} \mathbf{Q}_{\beta,\alpha}(u) \cdot \mathbf{n} = \int_{\Gamma_D} \big(x \cdot \mathbf{n}(x)\big) k^{-2} \left| \frac{\partial u}{\partial n} \right|^2 .$$

(f) By using Parts (a), (d), and (e), prove the bound (2.6); i.e., that if Ω_- is star-shaped with respect to the origin and u is the solution of the EDP with $g_D = 0$, $A = I$, and $n = 1$, then

$$\|u\|_{H^1_k(\Omega_R)} \leq 2kR \sqrt{1 + \left(\frac{d-1}{2kR} \right)^2} \, \|f\|_{L^2(\Omega_R)} .$$

4. The purpose of this exercise is to show how properties of DtN_k enter the proof of Lemma 2.16 (for the full proof, see [96, Proof of Theorem 6.1]). Assume the following two results.

- If D is a bounded, convex, open set of $\mathbb{R}^d$ with C^2 boundary and $\mathbf{v} \in H^1(D; \mathbb{C}^d)$, then

$$\int_D \left(|\nabla \cdot \mathbf{v}|^2 - \sum_{i,j=1}^{n} \int_D \frac{\partial v_i}{\partial x_j} \overline{\frac{\partial v_j}{\partial x_i}} \right) \geq -2\Re \big\langle (\gamma \mathbf{v})_T, \nabla_T (\gamma \mathbf{v} \cdot \mathbf{n}) \big\rangle_{\partial D}, \tag{2.21}$$

  where ∇_T is the surface gradient on ∂D and $(\gamma \mathbf{v})_T := \gamma \mathbf{v} - \mathbf{n}(\gamma \mathbf{v} \cdot \mathbf{n})$ is the tangential component of $\gamma \mathbf{v}$; this follows from [72, Theorem 3.1.1.1] and the fact that the second fundamental form of ∂D (defined in, e.g., [72, §3.1.1]) is non-positive (see [72, Proof of Theorem 3.1.2.3]).

- The solution of (2.8) is in $H^2(\Omega_R)$; this follows from results about the regularity of transmission problems (see [42, Theorem 5.2.1 and §5.4b]) since Γ_R is C^2 and $A = I$ in a neighbourhood of Γ_R.

Use these two results along with Lemma 1.7 to prove the bound (2.9) when $A = I$ and $\Omega_- = \emptyset$.

3 Convergence of the h-FEM: sharp k-explicit results for $p = 1$

3.1 Definition of the Galerkin method

Let $\mathcal{H} := H^1_{0,D}(\Omega_R)$, and let $\mathcal{H}_N$ be a finite-dimensional subspace of $\mathcal{H}$. We restate the variational formulation (1.23) of the EDP with $g_D = 0$ as

$$\text{find } u \in \mathcal{H} \text{ such that } \; a(u, v) = F(v) \; \text{ for all } v \in \mathcal{H}. \tag{3.1}$$

The Galerkin method applied to the variational problem (3.1) is

$$\text{find } u_N \in \mathcal{H}_N \text{ such that } \; a(u_N, v_N) = F(v_N) \; \text{ for all } v_N \in \mathcal{H}_N. \tag{3.2}$$

Observe that setting $v = v_N$ in (3.1) and combining this with (3.2) we obtain the *Galerkin orthogonality* that

$$a(u - u_N, v_N) = 0 \quad \text{ for all } v_N \in \mathcal{H}_N. \tag{3.3}$$

These definitions of course make sense for (3.1) with $a(\cdot, \cdot)$ a general sesquilinear form and $F(\cdot)$ a general antilinear functional (i.e., not just with $a(\cdot, \cdot)$ and $F(v)$ defined by (1.24)).

3.2 Abstract results about convergence of the Galerkin method and their applicability (or not) to the Helmholtz equation

Let $(\mathcal{H}_N)_{N=1}^{\infty}$ be a sequence of finite-dimensional subspaces of $\mathcal{H}$ that is *asymptotically dense* in $\mathcal{H}$, meaning that, for all $v \in \mathcal{H}$,

$$\min_{w_N \in \mathcal{H}_N} \|v - w_N\|_{\mathcal{H}} \to 0 \quad \text{as} \quad N \to \infty.$$

We say that the Galerkin method (3.2) converges if there exists $N_0 \in \mathbb{N}$ such that for all $N \geq N_0$ and for all $F(\cdot) \in \mathcal{H}'$, u_N defined by (3.2) exists and is unique, and $\|u_N - u\|_{\mathcal{H}} \to 0$ as $N \to \infty$.

In this section, we focus on proving *either* a bound on the relative error of the Galerkin solution, i.e., a bound on

$$\frac{\|u - u_N\|_{\mathcal{H}}}{\|u\|_{\mathcal{H}}},$$

or that the Galerkin method is *quasioptimal*, i.e., there exists a $C_{\text{qo}} > 0$ and $N_0 \in \mathbb{N}$ such that, for $N \geq N_0$,

$$\|u - u_N\|_{\mathcal{H}} \leq C_{\text{qo}} \min_{v_N \in \mathcal{H}_N} \|u - v_N\|_{\mathcal{H}}. \tag{3.4}$$

(If (3.4) holds with $C_{\text{qo}} = 1$ then the method is *optimal*.)

To state the following theorem we recall that given a sesquilinear form $a : \mathcal{H} \times \mathcal{H} \to \mathbb{C}$, there exists a unique linear operator $\mathcal{A} : \mathcal{H} \to \mathcal{H}$ such that $a(u, v) = (\mathcal{A}u, v)_{\mathcal{H}}$ for all $u, v \in \mathcal{H}$. (see, e.g., [130, Lemma 2.1.38]).

Theorem 3.1. (The main abstract theorem on convergence of the Galerkin method.) *Let $\mathcal{A} : \mathcal{H} \to \mathcal{H}$ be a bounded linear operator.*

(a) If $\mathcal{A}$ is invertible then there exists a sequence $(\mathcal{H}_N)_{N=1}^{\infty}$ for which the Galerkin method (3.2) converges.

(b) If $\mathcal{A}$ is coercive, i.e., there exists $\alpha > 0$ such that

$$\left|(\mathcal{A}v, v)_{\mathcal{H}}\right| \geq \alpha \|v\|_{\mathcal{H}}^2 \quad \text{for all } v \in \mathcal{H}, \tag{3.5}$$

then, for every sequence $(\mathcal{H}_N)_{N=1}^{\infty}$, the Galerkin equations (3.2) have a unique solution u_N for every N and

$$\left\|u - u_N\right\|_{\mathcal{H}} \leq \frac{\|\mathcal{A}\|_{\mathcal{H}\to\mathcal{H}}}{\alpha} \min_{v_N \in \mathcal{H}_N} \left\|u - v_N\right\|_{\mathcal{H}}.$$

(c) If $\mathcal{A}$ is invertible then the following are equivalent:

(i) The Galerkin method (3.2) converges for every sequence $(\mathcal{H}_N)_{N=1}^{\infty}$ that is asymptotically dense in $\mathcal{H}$.

(ii) $\mathcal{A} = \mathcal{A}_0 + \mathcal{K}$ where $\mathcal{A}_0$ is coercive and $\mathcal{K}$ is compact.

References for the proof. Part (a) was first proved in [104, Theorem 1]; see also [68, Chapter II, Theorem 4.1]. Part (b) is Céa's Lemma; see [31]. Part (c) was first proved in [104, Theorem 2], with this result building on results in [149]; see also [68, Chapter II, Lemma 5.1 and Theorem 5.1] □

Remark 3.2. *The fact that Point (ii) in Part (c) implies Point (i) is very well known in the numerical-analysis community (see, e.g., [130, Theorem 4.2.9], [142, Theorem 8.11]). However, the fact that Point (i) implies Point (ii) appears not to be well known (e.g., it was recently independently rederived in [7, Theorem 5.2]); this implication is nevertheless is quoted in, e.g., [80, Page 303].*

The following two results show that Part (b) of Theorem 3.1 is *not* applicable to the Helmholtz equation, but Part (c) *is*.

Lemma 3.3. (The Helmholtz sesquilinear form is not coercive for k sufficiently large.) *Suppose that A, n, and Ω_- satisfy Assumption 1.1 and let $a(\cdot, \cdot)$ be defined by (1.24). Let $\lambda_1 > 0$ be the first Dirichlet eigenvalue of $n^{-1}\nabla \cdot (A\nabla \cdot)$ in Ω_R. If $k^2 \geq \lambda_1$, then there exists $v \in H^1_{0,D}(\Omega_R)$ such that $a(v, v) = 0$.*

Proof. If λ_j is a Dirichlet eigenvalue of $n^{-1}\nabla\cdot(A\nabla\cdot)$ in Ω_R with corresponding eigenfunction $u_j \in H^1_0(\Omega_R) := \{v \in H^1_{0,D}(\Omega_R) : \gamma v = 0 \text{ on } \Gamma_R\}$, then

$$\big(A\nabla u_j, \nabla v\big)_{L^2(\Omega_R)} - \lambda_j \big(n u_j, v\big)_{L^2(\Omega_R)} = 0 \quad \text{for all } v \in H^1_{0,D}(\Omega_R),$$

and thus

$$\big(A\nabla u_j, \nabla u_j\big)_{L^2(\Omega_R)} - \lambda_j \big(n u_j, u_j\big)_{L^2(\Omega_R)} = 0.$$

Therefore, by the definition of $a(\cdot,\cdot)$ (1.24),

$$a(u_j, u_j) = k^{-2}(\lambda_j - k^2)\big(n u_j, u_j\big)_{L^2(\Omega_R)},$$

and thus $a(u_j, u_j) = 0$ if $k^2 = \lambda_j$. Furthermore, if $\lambda_1 < k^2 < \lambda_j$ (for some $j > 1$) then

$$a(u_1, u_1) < 0 < a(u_j, u_j). \tag{3.6}$$

The *numerical range* of a general sesquilinear form $a : \mathcal{H} \times \mathcal{H} \to \mathbb{C}$ is defined to equal

$$\left\{ \frac{a(v,v)}{\|v\|^2_{\mathcal{H}}} : v \in \mathcal{H} \setminus \{0\} \right\} \subset \mathbb{C}; \tag{3.7}$$

furthermore, the numerical range is convex by, e.g., [74, Theorem 1.1-2]. Therefore (3.6) implies that there exists $v \in H^1_{0,D}(\Omega_R)$ such that $a(v,v) = 0$. □

Lemma 3.4. (Helmholtz sesquilinear form = coercive + compact.) *Suppose that A, n, and Ω_- satisfy Assumption 1.1. Let $\mathcal{A}$ be the operator associated with the sesquilinear form $a(\cdot,\cdot)$ defined by (1.24) and let $\mathcal{H} := H^1_{0,D}(\Omega_R)$. Then $\mathcal{A} = \mathcal{A}_0 + \mathcal{K}$ with $\mathcal{A}_0$ bounded and coercive on $\mathcal{H}$ and $\mathcal{K}$ compact on $\mathcal{H}$.*

Proof. This follows from the Gårding inequality (1.28); see Exercise 3 in §3.12. □

Although Part (c) of Theorem 3.1 is applicable to the Helmholtz equation, this abstract result gives no information about the k-dependence of either the threshold N_0 for the Galerkin solution to exist or the Galerkin error. The main results of this section (in §3.4) are all about this k-dependence.

3.3 Definition of the FEM and assumptions for the results in this section

The finite-element method (FEM) is the Galerkin method applied with finite-dimensional subspaces consisting of piecewise polynomials. The h-version of the FEM is where accuracy is increased by decreasing the meshwidth h, the p-version of the FEM is where accuracy is increased by increasing the polynomial degree p, and the hp-version of the FEM is where accuracy is increased by *both* decreasing h *and* increasing p.

The main results in this section (Theorems 3.11 and 3.15) are proved for a sequence of subspaces $(\mathcal{H}_h)_{0<h\le h_0}$ satisfying the following assumption.

Assumption 3.5. *There exists $C_{\rm int} > 0$ such that for all $0 < h \le h_0$ there exists $I_h : H^2(\Omega_R) \to \mathcal{H}_h$ such that*

$$\|v - I_h v\|_{L^2(\Omega_R)} + h\,\|\nabla(v - I_h v)\|_{L^2(\Omega_R)} \le C_{\rm int} h^2 |v|_{H^2(\Omega_R)} \quad \text{for all } v \in H^2(\Omega_R). \tag{3.8}$$

We now show that the standard piecewise-polynomial subspaces of the h-version of the FEM satisfy Assumption 3.5 (with I_h the so-called nodal interpolant) – see Lemma 3.8 below – hence our choice of the notation $(\mathcal{H}_h)_{0<h\le h_0}$.

Definition 3.6. (Triangulation [40, Page 61].) *A finite collection of sets $\mathcal{T}$ is a* triangulation *of Ω if the following properties hold.*

1. $\overline{\Omega} = \bigcup_{K\in\mathcal{T}} K$

2. *Each $K \in \mathcal{T}$ is closed and its interior, $\mathring{K}$, is non-empty and connected.*

3. *If $K_1, K_2 \in \mathcal{T}$ and $K_1 \neq K_2$ then $\mathring{K}_1 \cap \mathring{K}_2 = \emptyset$.*

4. *Each $K \in \mathcal{T}$ is Lipschitz.*

For $K \in \mathcal{T}$, let $h_K := \operatorname{diam}(K) = \max_{x,\mathbf{y}\in K} |x - \mathbf{y}|$ and let the *mesh width* $h := \max_{K\in\mathcal{T}} h_K$. We denote a triangulation with mesh width h by $\mathcal{T}_h$. We now consider a family of triangulations $(\mathcal{T}_h)_{0<h\leq h_0}$ for some h_0. Let ρ_K be the diameter of the largest ball contained in K (so $\rho_K \leq h_K$).

Definition 3.7. (Shape-regular and quasi-uniform [23, Definition 4.4.13], [40, Pages 128 and 135].)
(i) The family $(\mathcal{T}_h)_{0<h\leq h_0}$ is shape-regular *(or* non-degenerate*) if there exists $C > 0$ such that $h_K \leq C\rho_K$ for all $K \in \mathcal{T}_h$ and for all $0 < h \leq h_0$.*
(ii) The family $(\mathcal{T}_h)_{0<h\leq h_0}$ is quasi-uniform *if there exists $C > 0$ such that $h \leq Ch_K$ for all $K \in \mathcal{T}_h$ and for all $0 < h \leq h_0$; i.e.,*

$$\max_{K\in\mathcal{T}_h} \operatorname{diam}(K) \leq C \min_{K\in\mathcal{T}_h} \operatorname{diam}(K) \quad \text{for all } 0 < h \leq h_0.$$

Given a triangulation $\mathcal{T}_h$, let

$$\mathcal{H}^{p,1}(\mathcal{T}_h) := \Big\{ v \in H^1(\Omega_R) : v|_K \text{ is a polynomial of degree } p \text{ for each } K \in \mathcal{T}_h \text{ and } v = 0 \text{ on } \Gamma_D \Big\}. \tag{3.9}$$

If $(\mathcal{T}_h)_{0<h\leq h_0}$ is quasi-uniform, then the dimension of $\mathcal{H}^{p,1}(\mathcal{T}_h)$ is proportional to $(p/h)^d$. From here on, we abbreviate the family $(\mathcal{H}^{p,1}(\mathcal{T}_h))_{0<h\leq h_0}$ (with p fixed) by $(\mathcal{H}_h)_{0<h\leq h_0}$.

Lemma 3.8. (Conditions under which Assumption 3.5 holds.) *Let $(\mathcal{T}_h)_{0<h\leq h_0}$ be a sequence of shape-regular triangulations, satisfying the addition conditions that (i) each $K \in \mathcal{T}_h$ is a simplex, and (ii) any face of any simplex K_1 in the triangulation is either a subset of the boundary of the domain, or a face of another simplex K_2 in the triangulation.*

Then Assumption 3.5 holds for $(\mathcal{H})_{0<h\leq h_0}$ with I_h the nodal interpolant *and C_{int} depending only on p, d, and the shape-regularity constant of $(\mathcal{T}_h)_{0<h\leq h_0}$.*

References for the proof. This follows from, e.g., [40, Theorem 17.1], [23, Proposition 3.3.17 and §4.4] (with (3.8) following from [23, Equation 4.4.28]). □

The main results in this section (Theorems 3.11 and 3.15) require Ω_R to be at least $C^{1,1}$ (so that the solution of the EDP with $g_D = 0$ is in $H^2(\Omega_R)$). For such Ω_R it is not possible to fit $\partial\Omega_R$ exactly with simplicial elements (i.e. when each element of $\mathcal{T}_h$ is a simplex), and fitting $\partial\Omega_R$ with isoparametric elements (see, e.g, [40, Chapter VI], [23, §4.7]) or curved elements (see, e.g., [18]) is often impractical. Some analysis of non-conforming error is therefore required, but since this is standard (see, e.g., [23, Chapter 10]), we ignore this issue here.

Finally, we make some simplifying assumptions on the parameters k, R, and h.

Assumption 3.9. $R \geq R_0 > 0,\ k \geq k_0 > 0,\ k_0R_0 \geq 1,\quad \text{and}\quad hk \leq 1.$

Observe that (3.8), the definition of $\|\cdot\|_{H^1_k(\Omega_R)}$ (1.8), and the assumption that $hk \leq 1$ imply that

$$\|v - I_h v\|_{H^1_k(\Omega_R)} \leq \sqrt{2} C_{\text{int}} hk |v|_{H^2_k(\Omega_R)}. \tag{3.10}$$

Remark 3.10. (Approximating DtN_k.) *Implementing the operator DtN_k is computationally expensive, and so in practice one seeks to approximate this operator by* either *imposing an absorbing boundary condition on Γ_R,* or *using a perfectly-matched layer (PML),* or *using boundary integral equations (so-called "FEM-BEM coupling"). For simplicity, in this section we analyse the FEM assuming that DtN_k is realised exactly. Recent k-explicit results on the error incurred by approximating DtN_k by absorbing boundary conditions or PML can be found in [60] and [61], respectively.*

3.4 Statement of the main results of this section

Theorem 3.11 gives sufficient conditions for the Galerkin method to be quasioptimal (with constant of quasioptimality independent of k), and Theorem 3.15 gives sufficient conditions for the relative error of the Galerkin solution to be controllably small, independent of k. Both these results are sharp for $p = 1$ when $C_{\text{sol}} \sim k$.

Theorem 3.11. (Quasioptimality.) *Let u be the solution of the EDP. Suppose that Assumptions 1.1, 3.5, and 3.9 hold, and Ω_- is $C^{1,1}$. Let C_{sol} be defined by (2.1), C_{cont} by (1.27), C_{H^2} by (2.9), and C_{int} by (3.8).*

If

$$hkC_{\text{sol}} \leq C_0 \tag{3.11}$$

where

$$C_0 := \frac{1}{C_{\text{cont}}} \sqrt{\frac{A_{\min}}{2\big(n_{\max} + A_{\min}\big)}} \left[\sqrt{2} C_{\text{int}} C_{H^2} \left(n_{\max} + \frac{1}{C_{\text{sol}}} + \sqrt{2}\right)\right]^{-1},$$

then the Galerkin solution u_N to the variational problem (3.2) exists, is unique, and satisfies the bound

$$\|u - u_N\|_{H^1_k(\Omega_R)} \leq \frac{2C_{\text{cont}}}{A_{\min}} \left(\min_{v_N \in \mathcal{H}_h} \|u - v_N\|_{H^1_k(\Omega_R)}\right). \tag{3.12}$$

The quantity C_0 depends on k via C_{sol}. However, under the assumption that $C_{\text{sol}} \gtrsim 1$, $C_0 \sim 1$.

When $C_{\text{sol}} \lesssim k$ (informally, the problem is nontrapping), the condition (3.11) is that hk^2 is sufficiently small; this condition is sharp – see Figure 3.1.

Our bound on the relative error requires the following assumption about the oscillatory character of u.

Assumption 3.12. (Oscillatory behaviour.) *Given $k_0 > 0$, there exists $C_{\text{osc}} = C_{\text{osc}}(A, n, \Omega_-, R, k_0)$ ('osc' standing for 'oscillation') such that, for all $k \geq k_0$, the solution of the EDP of Definition 1.2 satisfies*

$$|u|_{H^2_k(\Omega_R)} \leq C_{\text{osc}} \|u\|_{H^1_k(\Omega_R)}. \tag{3.13}$$

Theorem 3.13. *Assumption 3.12 is satisfied when u is the solution of the plane-wave sound-soft scattering problem, Ω_- is $C^{1,1}$, and $C_{\text{sol}} \lesssim k$.*

Reference for the proof. See [96, Theorem 9.1] (note that [96, Remark 9.10] outlines how the assumption that $C_{\text{sol}} \lesssim k$ can be removed). □

Remark 3.14. (Discussion of Assumption 3.12.) *Assumption 3.12 is* not *satisfied for the solution of the EDP for general $f \in L^2(\Omega_R)$. For example, consider the 1-d problem*

$$k^{-2}u'' + u = -f \quad \text{in } (0,1), \quad u(0) = 0, \quad \text{and} \quad k^{-1}u'(1) - \mathrm{i}u(1) = 0 \tag{3.14}$$

with

$$f(x) := -k^{-2}\big[\exp(\mathrm{i}k^n x)\chi(x)\big]'' - \big[\exp(\mathrm{i}k^n x)\chi(x)\big], \tag{3.15}$$

where χ has compact support in $(0,1)$. The solution to (3.14) is then $u(x) = \exp(\mathrm{i}k^n x)\chi(x)$, which oscillates on a scale of k^{-n}, i.e., a smaller scale than k^{-1} when $n > 1$.

If Assumption 3.12 holds then, using (3.10) and (3.13) in (3.12), we obtain that if $hkC_{\text{sol}} \leq C_0$, then

$$\frac{\|u - u_N\|_{H^1_k(\Omega_R)}}{\|u\|_{H^1_k(\Omega_R)}} \lesssim hk \leq \frac{C_0}{C_{\text{sol}}};$$

i.e., a bound on the relative error. When $C_{\text{sol}} \sim k$, this bound says that the relative error decreases like k^{-1} when hk^2 is sufficiently small. This leaves open the possibility that the relative error is bounded in k when hk^a is sufficiently small, for some $1 < a < 2$; the following theorem shows that (when $C_{\text{sol}} \sim k$) this is true when $a = 3/2$.

This next result involves the following constant; by, e.g., [23, §5.3], [146, Corollary A.15], there exists $C_{\rm PF} = C_{\rm PF}(\Omega_-)$ ('PF' standing for 'Poincaré–Friedrichs') such that

$$R^{-2}\,\|v\|^2_{L^2(\Omega_R)} \le C_{\rm PF}\left(R^{-1}\,\|\gamma v\|^2_{L^2(\Gamma_R)} + \|\nabla v\|^2_{L^2(\Omega_R)}\right) \tag{3.16}$$

for all $v \in H^1(\Omega_R)$.

Theorem 3.15. (Relative-error bound.) *Let u be the solution of the EDP. Suppose that Assumptions 1.1, 3.5, 3.9, and 3.12 hold, and Ω_- is $C^{1,1}$. Let $C_{\rm DtN1}$ be defined by (1.11), $C_{\rm DtN2}$ by (1.13), $C_{\rm sol}$ by (2.1), $C_{\rm cont}$ by (1.27), C_{H^2} by (2.9), $C_{\rm int}$ by (3.8), $C_{\rm osc}$ by (3.13), and $C_{\rm PF}$ by (3.16).*

If

$$(hk)^2 C_{\rm sol} \le C_1, \tag{3.17}$$

then the Galerkin solution u_N to the variational problem (3.2) exists, is unique, and satisfies the bound

$$\frac{\|u - u_N\|_{H^1_k(\Omega_R)}}{\|u\|_{H^1_k(\Omega_R)}} \le C_2 hk + C_3 (hk)^2 C_{\rm sol}, \tag{3.18}$$

where

$$C_1 := \frac{1}{4(A_{\max} + C_{\rm DtN1}) n_{\max} (C_{H^2})^2 (C_{\rm int})^2} \left(1 + \frac{\sqrt{2}}{\min\left\{A_{\min}(1+C_{\rm PF})^{-1}, C_{\rm DtN2}(C_{\rm PF})^{-1}\right\}}\right)^{-1} \times \left(n_{\max} + \frac{1}{C_{\rm sol}} + \sqrt{2}\right)^{-1},$$

$$C_2 := \frac{\sqrt{2} C_{\rm int} C_{\rm osc}}{A_{\min}} \left(\max\left\{A_{\max}, n_{\max}\right\} + C_{\rm DtN1}\right),$$

and

$$C_3 := \frac{4\sqrt{2}}{\sqrt{A_{\min}}} (A_{\max} + C_{\rm DtN1}) (C_{\rm int})^2 C_{H^2} C_{\rm osc} \sqrt{n_{\max} + A_{\min}} \left(n_{\max} + \frac{1}{C_{\rm sol}} + \sqrt{2}\right).$$

When $C_{\rm sol} \sim k$ (informally, the problem is nontrapping), the condition (3.17) is that h^2k^3 is sufficiently small. The quantity C_2 is independent of k and h. The quantities C_0, C_1, and C_3 are independent of h, but depend on k via $C_{\rm sol}$. Under the assumption that $C_{\rm sol} \gtrsim 1$ however, C_0, C_1, and C_3 all ~ 1. The requirement that h^2k^3 be sufficiently small for the relative-error to be controllably small (independent of k) is sharp – see Figure 3.1.

The history of the results in Theorems 3.11 and 3.15 and the techniques used to obtain them are discussed in §3.11 below.

Remark 3.16. (The pollution effect.) *The pollution effect for the h-FEM can be defined by* either *saying that the h-FEM suffers the pollution effect if the condition "hk sufficiently small (independent of k)" is not enough to ensure that the relative error is controllably small, independently of k – see, e.g., [83, §4.6.1] –* or *by saying that the h-FEM suffers the pollution effect if the condition "hk sufficiently small (independent of k)" is not enough to ensure quasi-optimality with constant independent of k – see, e.g. [12, Definition 2.1].*

With either definition, the sharpness of the conditions on h and k in Theorems 3.11 and 3.15 shows that the h-FEM with $p = 1$ suffers from the pollution effect.

More generally, the pollution effect can be defined by replacing "hk sufficiently small" in the above definitions by "sufficiently large (but k-independent) number of degrees of freedom per wavelength", since the number of degrees of freedom per wavelength for the h-FEM is proportional to $(p/h)(2\pi/k)$.

3.5 Numerical experiments illustrating Theorems 3.11 and 3.15

Figure 3.1.

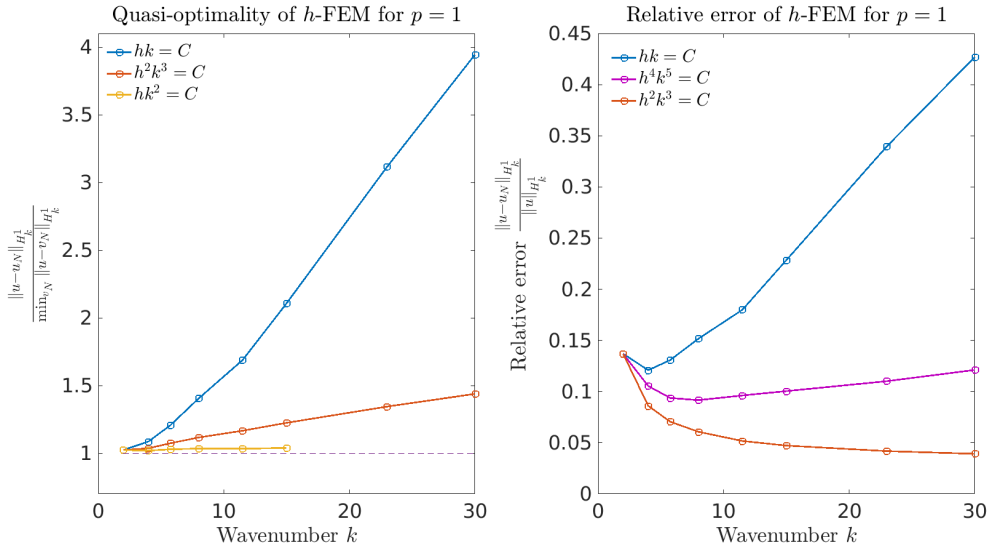

Figure 3.1: The ratio of the Galerkin error and the best-approximation error (left) and the relative Galerkin error (right) for the h-FEM with $p = 1$

3.6 The adjoint solution operator with L^2 data

The proofs of Theorems 3.11 and 3.15 crucially rely on properties of the adjoint solution operator.

Definition 3.17. (Adjoint solution operator $\mathcal{S}^*$.) *Given $f \in L^2(\Omega_R)$, let $\mathcal{S}^* f \in H^1_{0,D}(\Omega_R)$ be defined as the solution of the variational problem*

$$a(v, \mathcal{S}^* f) = (v, f)_{L^2(\Omega_R)} \quad \text{for all } v \in H^1_{0,D}(\Omega_R). \tag{3.19}$$

$\mathcal{S}^*$ is therefore the solution operator of the adjoint problem to the variational problem (3.1) with data in $L^2(\Omega_R)$.

Lemma 3.18. *If $\mathcal{S}^*$ is defined as in (3.19) then*

$$a(\overline{\mathcal{S}^* f}, v) = (\overline{f}, v)_{L^2(\Omega_R)} \quad \text{for all } v \in H^1_{0,D}(\Omega_R); \tag{3.20}$$

i.e., $\mathcal{S}^* f$ is the complex-conjugate of an outgoing Helmholtz solution.

Proof. This is Exercise 2 in §3.12. □

Following [129], let

$$\eta(\mathcal{H}_N) := \sup_{f \in L^2(\Omega_R)} \min_{v_N \in \mathcal{H}_N} \frac{\|\mathcal{S}^* f - v_N\|_{H^1_k(\Omega_R)}}{\|f\|_{L^2(\Omega_R)}}; \tag{3.21}$$

observe that this definition implies that, given $f \in L^2(\Omega_R)$,

$$\text{there exists } w_N \in \mathcal{H}_N \text{ such that } \|\mathcal{S}^* f - w_N\|_{H^1_k(\Omega_R)} \leq \eta(\mathcal{H}_N) \|f\|_{L^2(\Omega_R)}. \tag{3.22}$$

Lemma 3.19. (Bound on $\eta(\mathcal{H}_N)$.) *Suppose that Ω_-, A, and n satisfy Assumption 1.1 and, in addition, Ω_- is $C^{1,1}$. If Assumption 3.5 holds and $kR \geq 1$, then*

$$\eta(\mathcal{H}_N) \leq hk\, C_{\text{sol}} \left[\sqrt{2} C_{\text{int}} C_{H^2} \left(n_{\max} + \frac{1}{C_{\text{sol}}} + \sqrt{2} \right) \right]. \tag{3.23}$$

Proof. By the consequence (3.10) of (3.8), there exists $v_N \in \mathcal{H}_N$ such that

$$\|\mathcal{S}^* f - v_N\|_{H^1_k(\Omega_R)} \leq \sqrt{2} C_{\text{int}} hk |\mathcal{S}^* f|_{H^2_k(\Omega_R)}$$

(indeed, we can take $v_N = I_h(\mathcal{S}^* f)$). The result then follows from Lemma 3.18, the bound (2.10), and the fact that $kR \geq 1$. □

3.7 Sufficient conditions for quasioptimality/relative error controllably small in terms of $\eta(\mathcal{H}_N)$

Lemma 3.20. (Sufficient conditions for quasi-optimality.) *If*

$$\eta(\mathcal{H}_N) \leq \frac{1}{C_{\text{cont}}} \sqrt{\frac{A_{\min}}{2\big(n_{\max} + A_{\min}\big)}}, \tag{3.24}$$

then the Galerkin solution u_N to the variational problem (3.2) exists, is unique, and satisfies the bound

$$\|u - u_N\|_{H^1_k(\Omega_R)} \leq \frac{2C_{\text{cont}}}{A_{\min}} \left(\min_{v_N \in V_N} \|u - v_N\|_{H^1_k(\Omega_R)} \right). \tag{3.25}$$

Lemma 3.21. (Sufficient conditions for relative error controllably small.) *Suppose that Assumptions 1.1, 3.5, 3.9, and 3.12 hold, and Ω_- is $C^{1,1}$. Let $C_{\text{cont}\star}$ and $C_{H^2\star}$ be defined by (3.42) and (3.46) below. If*

$$hk\,\eta(\mathcal{H}_N) \leq \mathcal{C}_1, \qquad \text{where} \qquad \mathcal{C}_1 := \frac{1}{2\sqrt{2}C_{\text{cont}\star}C_{H^2\star}C_{\text{int}}n_{\max}}, \tag{3.26}$$

then the Galerkin solution u_h to the variational problem (3.2) exists, is unique, and satisfies the bound

$$\frac{\|u - u_h\|_{H^1_k(\Omega_R)}}{\|u\|_{H^1_k(\Omega_R)}} \leq \mathcal{C}_2 hk + \mathcal{C}_3 hk\,\eta(\mathcal{H}_N), \tag{3.27}$$

where

$$\mathcal{C}_2 := \frac{\sqrt{2}C_{\text{cont}}C_{\text{int}}C_{\text{osc}}}{A_{\min}} \qquad \text{and} \qquad \mathcal{C}_3 := \frac{4C_{\text{cont}\star}C_{\text{int}}C_{\text{osc}}\sqrt{n_{\max} + A_{\min}}}{\sqrt{A_{\min}}}. \tag{3.28}$$

Lemma 3.20 holds for any finite-dimensional subspace $\mathcal{H}_N$, whereas Lemma 3.21 is geared to be used for a member of $(\mathcal{H}_h)_{0<h\leq h_0}$ (because of the presence of h in (3.26), coming from a use of (3.8)). Nevertheless, to avoid swapping between the notations $\mathcal{H}_N$ and $\mathcal{H}_h$, from now on we denote the finite dimensional subspaces as $(\mathcal{H}_N)_{N=0}^{\infty}$.

Theorems 3.11 and 3.15 follow immediately from Lemmas 3.20 and 3.21, respectively, by using the bound (3.23) on $\eta(\mathcal{H}_N)$.

Observe that Lemma 3.20 requires no assumptions on A, n, and Ω_-, but Lemma 3.21 does. The reason is that the proof of Lemma 3.21 requires H^2 regularity of both u – hence the presence of C_{osc} in $\mathcal{C}_2$ and $\mathcal{C}_3$ – and the solution of a different elliptic boundary-value problem on Ω_- with coefficient A – hence the presence of $C_{H^2\star}$ in $\mathcal{C}_1$ (see (3.36) and Lemma 3.24 below for the definition of this boundary-value problem).

We now prove Lemma 3.20 (§3.8); with this proof in hand, we then describe the ideas behind Lemma 3.21 (§3.9).

3.8 Proof of Lemma 3.20

We first prove the result under the assumption that the Galerkin solution u_N exists.

Step 1: Use the Gårding inequality and Galerkin orthogonality. Using (in this order) the Gårding inequality (1.28), Galerkin orthogonality (3.3) and continuity of $a(\cdot,\cdot)$ (1.26), we have that, for any $v_N \in \mathcal{H}_N$,

$$\begin{aligned} A_{\min}\|u - u_N\|^2_{H^1_k(\Omega_R)} &\leq \Re a(u - u_N, u - u_N) + \big(n_{\max} + A_{\min}\big)\|u - u_N\|^2_{L^2(\Omega_R)} \\ &= \Re a(u - u_N, u - v_N) + \big(n_{\max} + A_{\min}\big)\|u - u_N\|^2_{L^2(\Omega_R)} \\ &\leq C_{\text{cont}}\|u - u_N\|_{H^1_k(\Omega_R)}\|u - v_N\|_{H^1_k(\Omega_R)} + \big(n_{\max} + A_{\min}\big)\|u - u_N\|^2_{L^2(\Omega_R)}. \end{aligned} \tag{3.29}$$

Step 2: Prove an Aubin-Nitsche type bound using the definition of $\eta(\mathcal{H}_N)$. The quasioptimal error bound (3.25) under the condition (3.24) follows from (3.29) if we can prove that

$$\|u - u_N\|_{L^2(\Omega_R)} \leq C_{\text{cont}} \eta(\mathcal{H}_N) \|u - u_N\|_{H^1_k(\Omega_R)}. \tag{3.30}$$

By the definition of S^* (3.19), Galerkin orthogonality (3.3), and continuity (1.26),

$$\|u - u_N\|^2_{L^2(\Omega_R)} = a\big(u - u_N, S^*(u - u_N)\big) = a\big(u - u_N, S^*(u - u_N) - v_N\big) \tag{3.31}$$

$$\leq C_c \|u - u_N\|_{H^1_k(\Omega_R)} \|S^*(u - u_N) - v_N\|_{H^1_k(\Omega_R)} \tag{3.32}$$

for any $v_N \in \mathcal{H}_N$. The definition of $\eta(\mathcal{H}_N)$ (3.21) implies that there exists a $w_N \in \mathcal{H}_N$ such that

$$\|S^*(u - u_N) - w_N\|_{H^1_k(\Omega_R)} \leq \eta(\mathcal{H}_N) \|u - u_N\|_{L^2(\Omega_R)}$$

(see (3.22)). Using this last inequality in (3.32), we obtain (3.30).

Step 3: Prove that u_N exists. We have so far assumed that u_N exists. Recall that an $N \times N$ matrix B is invertible if and only if B has full rank, which is the case if and only if the only solution of $B\mathbf{v} = 0$ is $\mathbf{v} = 0$. Therefore, to show that u_N exists, we only need to show that u_N is unique. Seeking a contradiction, suppose that there exists a $\widetilde{u}_N \in \mathcal{H}_N$ such that

$$a(\widetilde{u}_N, v_N) = 0 \quad \text{for all } v_N \in \mathcal{H}_N.$$

Let $\widetilde{u}$ be such that

$$a(\widetilde{u}, v) = 0 \quad \text{for all } v \in \mathcal{H}; \tag{3.33}$$

thus $\widetilde{u}_N$ is the Galerkin approximation to $\widetilde{u}$. Repeating the argument in the first part of the proof we see that if (3.24) holds then the quasi-optimal error bound (3.25) holds (with u replaced by $\widetilde{u}$ and u_N replaced by $\widetilde{u}_N$). By assumption, the only solution to the variational problem (3.33) is $\widetilde{u} = 0$, and then (3.25) implies that $\widetilde{u}_N = 0$. We have therefore shown that the solution u_N exists under the condition (3.24) and the proof is complete.

Remark 3.22. (Aubin-Nitsche-type bound.) *We describe (3.30) as an "Aubin-Nitsche-type bound", since the argument that obtains (3.30) was first introduced in the coercive case by Aubin [9, Theorem 3.1] and Nitsche [120] (see, e.g., [40, Theorem 19.1] for (3.30) exactly as stated); the history of these arguments is discussed further in §3.11.*

3.9 The ideas behind the proof of Lemma 3.21

The start of the proof of Lemma 3.21 is the same as Step 1 in the proof of Lemma 3.20; i.e., one arrives at (3.29). The end of the proof – justifying that u_N exists – uses the same arguments as in Step 3 of the proof of Lemma 3.20.

We now claim that Lemma 3.21 follows from using in (3.29) the result that if $hk\,\eta(\mathcal{H}_N)$ is sufficiently small then

$$\|u - u_N\|_{L^2(\Omega_R)} \lesssim \eta(\mathcal{H}_N) \|u - w_N\|_{H^1_k(\Omega_R)} \quad \text{for all } w_N \in \mathcal{H}_N. \tag{3.34}$$

Observe that (3.34) is a stronger bound than (3.30), since w_N on the right-hand side of (3.34) is arbitrary.

To justify the claim, observe that inputting (3.34) into (3.29), choosing $w_N = v_N$, and using the inequality (2.14) on the first term on the right-hand side of (3.29), we obtain that, if $hk\,\eta(\mathcal{H}_N)$ is sufficiently small, then

$$\|u - u_N\|_{H^1_k(\Omega_R)} \lesssim \big(1 + \eta(\mathcal{H}_N)\big) \|u - v_N\|_{H^1_k(\Omega_R)} \quad \text{for all } v_N \in \mathcal{H}_N.$$

At first it might look like we have just rederived the condition that the Galerkin solution is quasi-optimal (with constant independent of k) if $\eta(\mathcal{H}_N)$ is sufficiently small. However, the key point is

that "$hk\,\eta(\mathcal{H}_N)$ sufficiently small" is a weaker condition that "$\eta(\mathcal{H}_N)$ sufficiently small". Assuming H^2 regularity of the solution, and using (3.10), we obtain that, if $hk\eta(\mathcal{H}_N)$ is sufficiently small, then

$$\|u-u_N\|_{H^1_k(\Omega_R)} \lesssim \big(1+\eta(\mathcal{H}_N)\big)hk|u|_{H^2_k(\Omega_R)}, \tag{3.35}$$

and this becomes (3.27) after using the oscillatory-behaviour bound (3.13).

We now describe how to prove (3.34). Define the sesquilinear form $a_\star(\cdot,\cdot)$ by

$$a_\star(u,v) := \int_{\Omega_R} k^{-2}(A\nabla u)\cdot\overline{\nabla v} - k^{-1}\big\langle \mathrm{DtN}_k\gamma u, \gamma v\big\rangle_{\Gamma_R}; \tag{3.36}$$

one can show that $a_\star(\cdot,\cdot)$ is continuous and coercive in $H^1_{0,D}(\Omega_R)$ (see Lemma 3.23 below). We argue as in Step 2 of the proof of Lemma 3.20 to obtain (3.31), but then introduce $a_\star(\cdot,\cdot)$, where for brevity we use the notation that $\xi = S^*(u-u_N)$,

$$\begin{aligned}\|u-u_N\|^2_{L^2(\Omega_R)} = a(u-u_N,\xi) &= a(u-u_N,\xi-v_N),\\ &= a_\star(u-u_N,\xi-v_N) - \big(n(u-u_N),\xi-v_N\big)_{L^2(\Omega_R)}.\end{aligned} \tag{3.37}$$

Given $z\in H^1_{0,D}(\Omega_R)$, define $\mathcal{P}_N z\in\mathcal{H}_N$ by

$$a_\star(w_N,\mathcal{P}_N z) = a_\star(w_N,z) \quad \text{for all } w_N\in\mathcal{H}_N;$$

since $a_\star(\cdot,\cdot)$ is continuous and coercive in $H^1_{0,D}(\Omega_R)$, the Lax–Milgram theorem implies that $\mathcal{P}_N$ is well defined. The definition of $\mathcal{P}_N$ implies the Galerkin-orthogonality property that

$$a_\star(w_N, z-\mathcal{P}_N z) = 0 \quad \text{for all } w_N\in\mathcal{H}_N. \tag{3.38}$$

Choosing $v_N = P_N\xi$ in (3.37) and then using (3.38), we obtain that, for all $w_N\in\mathcal{H}_N$,

$$\begin{aligned}\|u-u_N\|^2_{L^2(\Omega_R)} &= a_\star(u-w_N,\xi-\mathcal{P}_N\xi) - \big(n(u-u_N),\xi-\mathcal{P}_N\xi\big)_{L^2(\Omega_R)},\\ &\lesssim \|u-w_N\|_*\,\|\xi-\mathcal{P}_N\xi\|_* + \|u-u_N\|_{L^2(\Omega_R)}\,\|\xi-\mathcal{P}_N\xi\|_{L^2(\Omega_R)},\end{aligned} \tag{3.39}$$

where

$$\|v\|_\star := \sqrt{a_\star(v,v)} \lesssim \|v\|_{H^1_k(\Omega_R)}. \tag{3.40}$$

Comparing (3.32) and (3.39), and using this last norm inequality, we see that Galerkin orthogonality for $a_\star(\cdot,\cdot)$ has allowed us to obtain $\|u-w_N\|_{H^1_k}$ (with w_N arbitrary) as opposed to $\|u-u_N\|_{H^1_k}$ on the right-hand side – this is ultimately what leads to the bound (3.34) instead of (3.30). To get there, we need to

(i) use Part (b) of Theorem 3.1, (3.40), the fact that $\xi = \mathcal{S}^*(u-u_N)$, and the definition of $\eta(\mathcal{H}_N)$ to get

$$\|\xi-\mathcal{P}_N\xi\|_* \lesssim \min_{v_N\in\mathcal{H}_N}\|\xi-v_N\|_* \lesssim \min_{v_N\in\mathcal{H}_N}\|\xi-v_N\|_{H^1_k(\Omega_R)} \lesssim \eta(\mathcal{H}_N)\|u-u_N\|_{L^2(\Omega_R)}$$

(see (3.48) below), and

(ii) use the argument in the proof of Lemma 3.20 to show that $\|\xi-\mathcal{P}_N\xi\|_{L^2(\mathbb{R}^d)} \lesssim hk\|\xi-\mathcal{P}_N\xi\|_*$ (see (3.49) below), and thus controlling the last term on the right-hand side of (3.39) leads to the condition that $hk\,\eta(\mathcal{H}_N)$ is sufficiently small.

The arguments in (ii) crucially use the H^2 regularity result of Lemma 2.16 – we see now that the non-standard boundary-value problem (2.8) comes from the sesquilinear form $a_\star(\cdot,\cdot)$.

3.10 Proof of Lemma 3.21

Lemma 3.23. (Continuity and coercivity of $a_\star(\cdot,\cdot)$.) *With $a_\star(\cdot,\cdot)$ defined by (3.36), for all $u, v \in H^1_{0,D}(\Omega_R)$,*

$$\left|a_\star(u,v)\right| \le C_{\mathrm{cont}\star}\,\|u\|_{H^1_k(\Omega_R)}\,\|v\|_{H^1_k(\Omega_R)} \quad \textit{and} \quad \Re a_\star(v,v) \ge C_{\mathrm{coer}\star}\,\|v\|^2_{H^1_{k,R}(\Omega_R)}\,, \tag{3.41}$$

where

$$C_{\mathrm{cont}\star} := A_{\max} + C_{\mathrm{DtN}1}, \qquad C_{\mathrm{coer}\star} := \min\left\{A_{\min}(1+C_{\mathrm{PF}})^{-1}, C_{\mathrm{DtN}2}(C_{\mathrm{PF}})^{-1}\right\}, \tag{3.42}$$

and

$$\|v\|^2_{H^1_{k,R}(\Omega_R)} := k^{-2}\,\|\nabla u\|^2_{L^2(\Omega_R)} + (kR)^{-2}\,\|v\|^2_{L^2(\Omega_R)}\,. \tag{3.43}$$

Proof. The first inequality in (3.41) follows from the inequality (1.11) and the Cauchy–Schwarz inequality. The second inequality in (3.41) follows from (1.13) and (3.16); indeed, for any $C > 0$,

$$\begin{aligned}\Re a_\star(v,v) &\ge \big(A_{\min} - C\big)k^{-2}\,\|\nabla v\|^2_{L^2(\Omega_R)} + \big(C_{\mathrm{DtN}2} - C\big)k^{-2}R^{-1}\,\|\gamma v\|^2_{L^2(\Gamma_R)} \\ &\qquad + Ck^{-2}\left(\|\nabla v\|^2_{L^2(\Omega_R)} + R^{-1}\,\|\gamma v\|^2_{L^2(\Gamma_R)}\right) \\ &\ge \big(A_{\min} - C\big)k^{-2}\,\|\nabla v\|^2_{L^2(\Omega_R)} + \big(C_{\mathrm{DtN}2} - C\big)k^{-2}R^{-1}\,\|\gamma v\|^2_{L^2(\Gamma_R)} + \frac{C}{C_{\mathrm{PF}}}(kR)^{-2}\,\|v\|^2_{L^2(\Omega_R)}\,.\end{aligned}$$

The value of C that produces the largest multiple of $\|v\|^2_{H^1_{k,R}(\Omega_R)}$ (3.43) on the right-hand side of the last inequality is $C = C' := A_{\min}C_{\mathrm{PF}}(1+C_{\mathrm{PF}})^{-1}$. If $C_{\mathrm{DtN}2} \ge C'$, then coercivity holds with $C_{\mathrm{coer}*} = A_{\min}(1+C_{\mathrm{PF}})^{-1}$. If $C_{\mathrm{DtN}2} < C'$, then we take $C = C_{\mathrm{DtN}2}$, and obtain coercivity with $C_{\mathrm{coer}*} = C_{\mathrm{DtN}2}/C_{\mathrm{PF}}$; i.e., coercivity holds with $C_{\mathrm{coer}*}$ given in (3.42). □

By Lemma 3.23,

$$C_{\mathrm{coer}\star}\,\|v\|^2_{H^1_{k,R}(\Omega_R)} \le \left|a_\star(v,v)\right| \le C_{\mathrm{cont}\star}\,\|v\|^2_{H^1_k(\Omega_R)} \qquad \text{for all } v \in H^1_{0,D}(\Omega_R); \tag{3.44}$$

we then define the new norm on $H^1_{0,D}(\Omega_R)$,

$$\|v\|_\star := \sqrt{a_\star(v,v)}.$$

Lemma 3.24. (Bounds on the solution of the variational problem associated with $a_\star(\cdot,\cdot)$.) *The solution of the variational problem*

$$\textit{find } u \in H^1_{0,D}(\Omega_R) \textit{ such that } \; a_\star(u,v) = (f,v)_{L^2(\Omega_R)} \; \textit{ for all } v \in H^1_{0,D}(\Omega_R)$$

satisfies

$$\|u\|_{H^1_{k,R}(\Omega_R)} \le \frac{kR}{C_{\mathrm{coer}\star}}\,\|f\|_{L^2(\Omega_R)} \qquad \textit{and} \qquad |u|_{H^2_k(\Omega_R)} \le C_{H^2\star}\,\|f\|_{L^2(\Omega_R)}\,, \tag{3.45}$$

where

$$C_{H^2\star} := C_{H^2}\left(1 + \sqrt{2}(C_{\mathrm{coer}\star})^{-1}\right). \tag{3.46}$$

Proof. Since $a_\star(\cdot,\cdot)$ is continuous and coercive in $H^1_{0,D}(\Omega_R)$, the first bound in (3.45) follows from the Lax–Milgram theorem and the fact that

$$\sup_{v\in\mathcal{H}} \frac{\left|(f,v)_{L^2(\Omega_R)}\right|}{\|v\|_{H^1_{k,R}(\Omega_R)}} \le kR\,\|f\|_{L^2(\Omega_R)}\,,$$

by the definition of $\|\cdot\|_{H^1_{k,R}(\Omega_R)}$ (3.43). The second bound in (3.45) follows from combining the first bound in (3.45) and the bound (2.9). □

We now define the particular Galerkin projection known in the literature as the "elliptic projection" (see the discussion in §3.11).

Definition 3.25. (Elliptic projection $\mathcal{P}_N$.) *Given $u \in H^1_{0,D}(\Omega_R)$, define $\mathcal{P}_N u \in \mathcal{H}_N$ by*

$$a_\star(v_N, \mathcal{P}_N u) = a_\star(v_N, u) \quad \text{for all } v_N \in \mathcal{H}_N.$$

Since $a_\star(\cdot,\cdot)$ is continuous and coercive in $H^1_{0,D}(\Omega_R)$ by Lemma 3.23, the Lax–Milgram theorem implies that $\mathcal{P}_N$ is well defined. The definition of $\mathcal{P}_N$ then immediately implies the Galerkin-orthogonality property that

$$a_\star(v_N, u - \mathcal{P}_N u) = 0 \quad \text{for all } v_N \in \mathcal{H}_N. \tag{3.47}$$

Lemma 3.26. (Approximation properties of $\mathcal{P}_N$.) *For all $u \in H^1_{0,D}(\Omega_R)$,*

$$\|u - \mathcal{P}_N u\|_\star \le \sqrt{C_{\mathrm{cont}\star}} \min_{v_N \in \mathcal{H}_N} \|u - v_N\|_{H^1_k(\Omega_R)} \qquad \text{and} \tag{3.48}$$

$$\|u - \mathcal{P}_N u\|_{L^2(\Omega_R)} \le hk\sqrt{2} C_{\mathrm{int}} C_{H^2\star} \sqrt{C_{\mathrm{cont}\star}} \, \|u - \mathcal{P}_N u\|_\star \,. \tag{3.49}$$

Proof. By the Cauchy–Schwarz inequality $a_\star(\cdot,\cdot)$ is continuous in the $\|\cdot\|_\star$ norm with continuity constant equal to one, and, by definition, $a_\star(\cdot,\cdot)$ is coercive in this norm. Therefore Céa's lemma implies that

$$\|u - \mathcal{P}_N u\|_\star \le \min_{v_N \in \mathcal{H}_N} \|u - v_N\|_\star \,,$$

and (3.48) follows from the norm equivalence (3.44).

To prove (3.49) we argue as in the proof of Lemma 3.20. Given $u \in H^1_{0,D}(\Omega_R)$, let ξ be the solution of the variational problem

$$\text{find } \xi \in H^1_{0,D}(\Omega_R) \text{ such that } \; a_\star(\xi, v) = (u - \mathcal{P}_N u, v)_{L^2(\Omega_R)} \quad \text{for all } v \in H^1_{0,D}(\Omega_R). \tag{3.50}$$

Then, by Galerkin orthogonality (3.47) and continuity of $a_\star(\cdot,\cdot)$, for all $v_N \in \mathcal{H}_N$,

$$\|u - \mathcal{P}_N u\|^2_{L^2(\Omega_R)} = a_\star(\xi, u - \mathcal{P}_N u) = a_\star(\xi - v_N, u - \mathcal{P}_N u) \le \|\xi - v_N\|_\star \, \|u - \mathcal{P}_N u\|_\star \tag{3.51}$$

By the norm equivalence (3.44), the consequence (3.10) of the definition of C_{int}, the definition of ξ (3.50), and the second bound in (3.45),

$$\begin{aligned} \|\xi - I_h\xi\|_\star \le \sqrt{C_{\mathrm{cont}\star}} \, \|\xi - I_h\xi\|_{H^1_k(\Omega_R)} &\le \sqrt{C_{\mathrm{cont}\star}}\sqrt{2} C_{\mathrm{int}} hk |\xi|_{H^2_k(\Omega_R)}, \\ &\le \sqrt{C_{\mathrm{cont}\star}}\sqrt{2} C_{\mathrm{int}} hk C_{H^2\star} \, \|u - \mathcal{P}_N u\|_{L^2(\Omega_R)} \,, \end{aligned}$$

and the result (3.49) follows from combining this last inequality with (3.51). □

Lemma 3.27. (Aubin-Nitsche analogue via elliptic projection.) *Assuming that the Galerkin solution u_N to the variational problem (3.2) exists, if (3.26) holds, then*

$$\|u - u_N\|_{L^2(\Omega_R)} \le 2 C_{\mathrm{cont}\star} \eta(\mathcal{H}_N) \, \|u - w_N\|_{H^1_k(\Omega_R)} \quad \text{for all } w_N \in \mathcal{H}_N.$$

Proof. Let $\xi = \mathcal{S}^*(u - u_N)$; i.e. ξ is the solution of variational problem

$$\text{find } \xi \in H^1_{0,D}(\Omega_R) \text{ such that } a(v, \xi) = (v, u - u_N)_{L^2(\Omega_R)} \quad \text{for all } v \in H^1_{0,D}(\Omega_R).$$

Then, by Galerkin orthogonality (3.47) and the definition of $a_\star(\cdot,\cdot)$ (3.36), for all $v_N \in \mathcal{H}_N$,

$$\begin{aligned} \|u - u_N\|^2_{L^2(\Omega_R)} = a(u - u_N, \xi) &= a(u - u_N, \xi - v_N), \\ &= a_\star(u - u_N, \xi - v_N)_{L^2(\Omega_R)} - \big(n(u - u_N), \xi - v_N\big)_{L^2(\Omega_R)}. \end{aligned}$$

We choose $v_N = \mathcal{P}_N \xi$, and then use (in the following order) (i) the Galerkin orthogonality (3.47), (ii) continuity of $a_\star(\cdot,\cdot)$, (iii) the bound (3.49), (iv) the upper bound in the norm equivalence (3.44)

and the bound (3.48), and (v) the consequence (3.22) of the definition of η to obtain that, for all $w_N \in \mathcal{H}_N$,

$$\begin{aligned}
\|u-u_N\|^2_{L^2(\Omega_R)} &= a_\star(u-w_N,\xi-\mathcal{P}_N\xi)_{L^2(\Omega_R)} - \big(n(u-u_N),\xi-\mathcal{P}_N\xi\big)_{L^2(\Omega_R)} \\
&\le \|u-w_N\|_\star \|\xi-\mathcal{P}_N\xi\|_\star + n_{\max}\|u-u_N\|_{L^2(\Omega_R)}\|\xi-\mathcal{P}_N\xi\|_{L^2(\Omega_R)} \\
&\le \Big(\|u-w_N\|_\star + hk\sqrt{2}C_{\rm int}C_{H^2\star}\sqrt{C_{{\rm cont}\star}}n_{\max}\|u-u_N\|_{L^2(\Omega_R)}\Big)\|\xi-\mathcal{P}_N\xi\|_\star \\
&\le \Big(\sqrt{C_{{\rm cont}\star}}\|u-w_N\|_{H^1_k(\Omega_R)} + hk\sqrt{2}C_{\rm int}C_{H^2\star}\sqrt{C_{{\rm cont}\star}}n_{\max}\|u-u_N\|_{L^2(\Omega_R)}\Big) \\
&\qquad\qquad \times \sqrt{C_{{\rm cont}\star}}\min_{v_N\in\mathcal{H}_N}\|\xi-v_N\|_{H^1_k(\Omega_R)} \\
&\le \Big(\sqrt{C_{{\rm cont}\star}}\|u-w_N\|_{H^1_k(\Omega_R)} + hk\sqrt{2}C_{\rm int}C_{H^2\star}\sqrt{C_{{\rm cont}\star}}n_{\max}\|u-u_N\|_{L^2(\Omega_R)}\Big) \\
&\qquad\qquad \times \sqrt{C_{{\rm cont}\star}}\eta(\mathcal{H}_N)\|u-u_N\|_{L^2(\Omega_R)};
\end{aligned}$$

the result then follows. □

Proof of Lemma 3.21. We first assume that the Galerkin solution u_N exists. Having proved the result under the assumption of existence, the fact that the condition (3.26) implies existence follows by arguing exactly as at the end of the proof of Lemma 3.20

Using the Gårding inequality (1.28), Galerkin orthogonality (3.3) and continuity of $a(\cdot,\cdot)$ (1.26), we find that that (3.29) holds for any $v_N \in \mathcal{H}_N$. Using first the inequality (2.14) with $\alpha = \|u-u_N\|_{H^1_k(\Omega_R)}$, $\beta = C_{\rm cont}\|u-v_N\|_{H^1_k(\Omega_R)}$, $\varepsilon = A_{\min}$, and then Lemma 3.27, we find that if (3.26) holds, then, for any $v_N \in \mathcal{H}_N$,

$$\begin{aligned}
\frac{A_{\min}}{2}\|u-u_N\|^2_{H^1_k(\Omega_R)} &\le \frac{(C_{\rm cont})^2}{2A_{\min}}\|u-v_N\|^2_{H^1_k(\Omega_R)} + \big(n_{\max}+A_{\min}\big)\|u-u_N\|^2_{L^2(\Omega_R)} \\
&\le \left[\frac{(C_{\rm cont})^2}{2A_{\min}} + 4\big(n_{\max}+A_{\min}\big)(C_{{\rm cont}\star})^2\big(\eta(\mathcal{H}_N)\big)^2\right]\|u-v_N\|^2_{H^1_k(\Omega_R)}.
\end{aligned} \tag{3.52}$$

By the consequence (3.10) of the definition of $C_{\rm int}$ and the bound (3.13),

$$\|u-I_hu\|_{H^1_k(\Omega_R)} \le \sqrt{2}hkC_{\rm int}|u|_{H^2_k(\Omega_R)} \le \sqrt{2}hkC_{\rm int}C_{\rm osc}\|u\|_{H^1_k(\Omega_R)}. \tag{3.53}$$

Choosing $v_N = I_hu$ in (3.52), using (3.53), taking the square root and using the inequality $\sqrt{a^2+b^2} \le a+b$ for all $a,b>0$, we find the result (3.27). □

3.11 History of the results in this section and the arguments used in the proofs

The arguments used to prove Theorem 3.11. As mentioned in §3.9, the argument that obtains (3.30) from (3.32) was first introduced in the coercive case by Aubin [9, Theorem 3.1] and Nitsche [120], with (3.30) appearing exactly as stated in, e.g., [40, Theorem 19.1]. This argument is often called a "duality argument" because it uses the adjoint sesquilinear form (in our case, $\eta(\mathcal{H}_N)$ involves the adjoint solution operator).

Schatz [131] considered second-order linear elliptic PDEs satisfying a Gårding inequality (such as (1.28)) proving existence and uniqueness of the Galerkin solution for h sufficiently small. The fact that these arguments also give quasioptimality was recognised in [11, Theorem 3.1], with this result proving the analogue of Theorem 3.11 (i.e., quasioptimality if hk^2 is sufficiently small) for the 1-d problem (3.14). Other uses of this argument on 1-d problems included [49, Lemma 2.6], [86, Theorem 3], [103, Theorem 3.2]. The analogue of Theorem 3.11 for the Helmholtz interior impedance problem

$$k^{-2}\Delta u + u = -f \text{ in } \Omega, \qquad k^{-1}\partial_n u - {\rm i}u = g \text{ on } \partial\Omega,$$

with $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, was then proved by Melenk [107, Proposition 8.2.7]; this advance was thanks to the availability of a k-explicit bound on the analogue of $C_{\rm sol}$ for this problem using essentially the identity (2.7) – see the discussion in Exercise 3 in §2.5.

Sauter [129] introduced the notation $\eta(\mathcal{H}_N)$ and emphasised the role of "adjoint approximability", with then [14] formulating more abstractly this idea of obtaining sufficient conditions for quasioptimality in terms of approximability of solutions of the adjoint equation (see Exercises 4 and 5 in §3.12 below). The argument was then used in the framework of [129] for Helmholtz problems with variable coefficients in [71, 64] and in domains with corners [36].

The arguments used to prove Theorem 3.15. The initial ideas behind the elliptic-projection argument were introduced in the Helmholtz context in [55, 56] for interior-penalty discontinuous Galerkin methods, and then further developed for the standard FEM and continuous interior-penalty methods in [151, 154]. The argument has been subsequently used by, e.g., [50, 15, 152, 36, 65, 100, 96] with all these papers considering the Helmholtz interior impedance problem apart from [100], which considered PML truncation, and [96] which considered the sesquilinear form $a(\cdot,\cdot)$ (1.24) (i.e., involving DtN_k).

The elliptic-projection argument for the Helmholtz equation with impedance boundary conditions requires the analogous result to Lemma 2.16 for Poisson's equation with the impedance boundary condition $k^{-1}\partial_{\mathbf{n}} v = \mathrm{i}\gamma v$. This result was explicitly assumed in [56, Lemma 4.3], implicitly assumed in [151, 154, 15, 36], and recently proved in [38]. Lemma 2.16 was proved in [96, Theorem 6.1] using arguments from [38], which in turn use results from [72] (see Exercise 4 in §2.5).

In all of the works referenced above except [96], the elliptic-projection argument is used to obtain (3.35) and then the analogue of the H^2 bound (2.10) is used to obtain a bound on the Galerkin error in terms of the data. The ingredient need to prove a bound on the relative error – the measure of the error most used in practical applications [3] – is the oscillatory behaviour bound (3.13), proved for the Helmholtz EDP in [96, Theorem 9.1]. We note that oscillatory behaviour similar to (3.13) of Helmholtz solutions has been an assumption in many analyses of finite- and boundary-element methods; see, e.g., [84, First equation in §3.4], [85, Definition 3.2], [24, Definition 4.6], [14, Definition 3.5], [48, Assumption 3.4]. Rigorous results other than [96, Theorem 9.1] proving such behaviour are [69, Theorems 1.1 and 1.2] and [63, Theorem 1.11(c)]. These results concern the Neumann trace of the solution of the Helmholtz plane-wave scattering problem with $A = I$ and $n = 1$, and are then used in [69] and [63] to analyse boundary-element methods applied to this problem; in common with the proof of (3.13) in [96, Theorem 9.1], these results are obtained using semiclassical-analysis techniques.

3.12 Exercises for Section 3

1. The goal of this exercise is to show how the quantity h^2k^3 (appearing in Theorem 3.15 under the assumption that $C_{\rm sol} \sim k$) arises from analysing solutions of the Galerkin linear system in 1-d. This material, and significant extensions of it, appear in [75, 84, 86, 83, 1].

   (a) Consider the finite-element discretisation of the 1-d model problem (3.14) on a uniform grid with meshwidth h, nodes x_j, and with piecewise-linear hat functions ϕ_j such that $\phi_j(x_i) = \delta_{ij}$. If x_j and x_{j+1} are both away from the boundary, show that

$$a(\phi_j, \phi_j) = \frac{2}{k^2 h}\left(1 - \frac{(hk)^2}{3}\right) =: \frac{2}{k^2 h} S(hk)$$

   and

$$a(\phi_j, \phi_{j+1}) = \frac{1}{k^2 h}\left(-1 - \frac{(hk)^2}{6}\right) =: \frac{1}{k^2 h} R(hk)$$

   so that, at least in the interior of the domain, the nodal values of Galerkin solution u_N satisfy

$$R(hk)u_N(x_j - h) + 2S(hk)u(x_j) + R(hk)u_N(x_j + h) = 0. \tag{3.54}$$

[3] For example, according to Google Scholar out of the articles published between 2010 and 2021 in the journal "Computer Methods in Applied Mechanics and Engineering" (a top, mathematically-inclined engineering journal), 21 contain the phrase "quasi optimality", 111 contain "quasi optimal", and 1020 contain "relative error".

(b) Seeking a solution of (3.54) of the form $u_N(x_j) = \exp(\mathrm{i}\widetilde{k}x_j)$, show that the constraint that $\widetilde{k}$ is real implies that $hk < \sqrt{12}$. Under this constraint, show that

$$\widetilde{k} = \frac{1}{h}\cos^{-1}\left(-\frac{S(hk)}{R(hk)}\right) = k - \frac{k^3h^2}{24} + \mathcal{O}(k^5h^4); \tag{3.55}$$

i.e., if the Galerkin solution is a propagating wave, then its "discrete wavenumber" $\widetilde{k}$ differs from the true wavenumber k by (to leading order) a constant times h^2k^3.

(This type of analysis is often called "dispersion analysis" for the following reason. Recall that a wave of the form $f(kx - \omega t)$ has phase velocity ω/k; when this phase velocity is independent of k, the wave is *non-dispersive*, and when the phase velocity depends on k, the wave is *dispersive*. The solution of the wave equation $\exp(\mathrm{i}\widetilde{k}x - \mathrm{i}\omega t)$ with $k = \omega/c$ has phase velocity $\omega/\widetilde{k} = (k/\widetilde{k})c$ (as in §0.1), which depends on k when $\widetilde{k}$ is given by (3.55).)

2. Prove Lemma 3.3. Hint: let $\mathcal{B} : L^2(\Omega_R) \to H^1_{0,D}(\Omega_R)$ be defined by

$$(\mathcal{B}u, v)_{H^1_k(\Omega_R)} = (nu, \imath v)_{L^2(\Omega_R)} \quad \text{for all } u \in L^2(\Omega_R),\ v \in H^1_{0,D}(\Omega_R),$$

where $\imath$ is the inclusion map $H^1_{0,D}(\Omega_R) \to L^2(\Omega_R)$, and recall that $\imath$ is compact by a result of Rellich; see, e.g., [106, Theorem 3.27].

3. Prove Lemma 3.18. Hint: using Green's identity and the radiation condition, show that $\langle \mathrm{DtN}_k\psi, \overline{\phi}\rangle_{\Gamma_R} = \langle \mathrm{DtN}_k\phi, \overline{\psi}\rangle_{\Gamma_R}$ for all $\phi, \psi \in H^{1/2}(\Gamma_R)$.

4. The goal of this exercise is to show how the conditions for quasioptimality in Lemma 3.20 can be formulated more abstractly (with this done in [14, Theorem 2.1]).

   As in §3.2, let $\mathcal{A} : \mathcal{H} \to \mathcal{H}$ be the linear operator such that $a(u, v) = (\mathcal{A}u, v)_{\mathcal{H}}$ for all $u, v \in \mathcal{H}$. Given $\mathcal{H}_N$ closed in $\mathcal{H}$, let P_N be the orthogonal projection onto $\mathcal{H}_N$ so that, in particular, $\|(I - P_N)u\|_{\mathcal{H}} = \min_{v_N \in \mathcal{H}_N} \|u - v_N\|_{\mathcal{H}}$. Suppose that $\mathcal{A}_0 : \mathcal{H} \to \mathcal{H}$ is a bounded linear operator that is coercive (i.e., (3.5) holds with $\mathcal{A}$ replaced by $\mathcal{A}_0$).

   Let u be the solution of the variational problem (3.1), and let u_N be the Galerkin solution defined by (3.2). Show that if

$$\left\|(I - P_N)(\mathcal{A}^*)^{-1}(\mathcal{A}^* - \mathcal{A}_0^*)\right\|_{\mathcal{H}\to\mathcal{H}} \le \frac{\alpha}{2\,\|\mathcal{A}\|_{\mathcal{H}\to\mathcal{H}}}, \tag{3.56}$$

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

$$\|u - u_N\|_{\mathcal{H}} \le \frac{2\,\|\mathcal{A}\|_{\mathcal{H}\to\mathcal{H}}}{\alpha}\,\|(I - P_N)u\|_{\mathcal{H}}\,. \tag{3.57}$$

   Hint: define $\mathcal{T} : \mathcal{H} \to \mathcal{H}$ by

$$a(w, \mathcal{T}v) = -(a - a_0)(w, v) \quad \text{for all } w \in \mathcal{H},$$

   where $a_0(w, v) = (\mathcal{A}_0 w, v)_{\mathcal{H}}$, let $\eta(\mathcal{H}_N) := \|(I - P_N)\mathcal{T}\|_{\mathcal{H}\to\mathcal{H}}$, and use the ideas in the proof of Lemma 3.20.

   (This result is useful when $(\mathcal{A}^*)^{-1}(\mathcal{A}^* - A_0^*)$ is smoothing; recall from Exercise 2 that the sesquilinear form of the EDP (1.24) fits into this framework – with also $\mathcal{A}_0$ coercive.)

5. The goal of this exercise is to show how the $\|\mathcal{A}\|_{\mathcal{H}\to\mathcal{H}}$ in the quasi-optimality constant in (3.57) can be replaced by $\|\mathcal{A}_0\|_{\mathcal{H}\to\mathcal{H}}$ – this is useful for proving quasioptimality of the Galerkin method applied to Helmholtz boundary integral equations where the norms grow with k; see [102], [62].

   Assume that $\mathcal{A}$ and $\mathcal{A}_0$ are as in Exercise 2.

(a) Show that

$$\alpha \|u-u_N\|_{\mathcal{H}}^2 \leq \|\mathcal{A}_0\|_{\mathcal{H}\to\mathcal{H}} \|u-u_N\|_{\mathcal{H}} \|u-P_N u\|_{\mathcal{H}} + \left|\left((\mathcal{A}-\mathcal{A}_0)(u-u_N), u_N - P_N u\right)_{\mathcal{H}}\right|.$$

(Note that $\|u-P_N u\|_{\mathcal{H}}$ on the right-hand side is multiplied by $\|\mathcal{A}_0\|_{\mathcal{H}\to\mathcal{H}}$, instead of by $\|A\|_{\mathcal{H}\to\mathcal{H}}$ as in the argument leading to (3.57).)

(b) Show that, for all $w_N \in \mathcal{H}_N$,

$$\left|\left((\mathcal{A}-\mathcal{A}_0)(u-u_N), w_N\right)_{\mathcal{H}}\right| \leq \left(\|\mathcal{A}_0\|_{\mathcal{H}\to\mathcal{H}} + \|(I-P_N)(A-\mathcal{A}_0)\|_{\mathcal{H}\to\mathcal{H}} \right) \times \left\|(I-P_N)(\mathcal{A}^*)^{-1}(\mathcal{A}^*-\mathcal{A}_0^*)\right\|_{\mathcal{H}\to\mathcal{H}} \|u-u_N\|_{\mathcal{H}} \|w_N\|_{\mathcal{H}}.$$

(c) By writing

$$\|u-u_N\|_{\mathcal{H}} \leq \|u-P_N u\|_{\mathcal{H}} + \|u_N - P_N u\|_{\mathcal{H}}$$

and using Parts (a) and (b), show that if

$$\left\|(I-P_N)(\mathcal{A}^*)^{-1}(\mathcal{A}^*-\mathcal{A}_0^*)\right\|_{\mathcal{H}\to\mathcal{H}} \leq \frac{\alpha}{4\|\mathcal{A}_0\|_{\mathcal{H}\to\mathcal{H}}} \tag{3.58}$$

and

$$\|(I-P_N)(\mathcal{A}-\mathcal{A}_0)\|_{\mathcal{H}\to\mathcal{H}} \leq \|\mathcal{A}_0\|_{\mathcal{H}\to\mathcal{H}}. \tag{3.59}$$

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

$$\|u-u_N\|_V \leq \left(1+\frac{2\|\mathcal{A}_0\|_{\mathcal{H}\to\mathcal{H}}}{\alpha}\right) \|(I-P_N)u\|_{\mathcal{H}}. \tag{3.60}$$

(This result is similar to that in [102, Theorem 3.8]; in this latter result, instead of $\mathcal{A}_0$ being coercive, $\mathcal{A}_0$ satisfies a discrete inf-sup condition in $\mathcal{H}_N$ with constant α, and then quasioptimality holds with constant $2(1+\|\mathcal{A}_0\|/\alpha)$. The proof is very similar to above, but starts by writing $\|u-u_N\|_{\mathcal{H}} \leq \|u-P_N u\|_{\mathcal{H}} + \|u_N - P_N u\|_{\mathcal{H}}$ and then bounding $\|u_N - P_N u\|_{\mathcal{H}}$ using steps similar to those above.)

4 Sharp k-explicit convergence results about the h-FEM (with $p>1$) and hp-FEM via frequency-splitting of high-frequency Helmholtz solutions

4.1 Assumptions on the finite-dimensional subspaces

The main results of this section are proved under one or both of the following two assumptions on the finite-dimensional subspaces. We highlight immediately that both these assumptions are satisfied by the standard hp-finite-element space $\mathcal{H}^{p,1}(\mathcal{T}_h)$ (3.9), provided that the triangulations are constructed by refining a fixed triangulation that has analytic element maps; see Theorems 4.2 and 4.5 below.

The results in this section hold for $\Omega_- = \emptyset$, in which case $\Omega_R = B_R$, and thus we work on B_R for the whole of the section.

Assumption 4.1. (Approximation of functions with finite regularity.) *Given s, d with $s > d/2$, there exists $C_{\text{approx}_1} > 0$ such that if $v \in H^s(B_R)$ and $p \geq s-1$, then*

$$\min_{w_N \in \mathcal{H}_N} \|v-w_N\|_{H_k^1(B_R)} \leq C_{\text{approx}_1} \left(\frac{hk}{p}\right)^{s-1} \left(1+\frac{hk}{p}\right) |v|_{H_k^s(B_R)}, \tag{4.1}$$

where $|v|_{H_k^s(B_R)} := k^{-s}|v|_{H^s(B_R)}$.

Assumption 4.1 is a generalisation of Assumption 3.5 (although Assumption 4.1 is stated in k-weighted norms, whereas Assumption 3.5 is stated in unweighted norms). Indeed, if Assumption 4.1 holds, then the consequence (3.10) of Assumption 3.5 follows from (4.1) with $p = 1$ and $s = 2$ and $\sqrt{2}C_{\text{int}}$ replaced by $C_{\text{approx}_1}(1 + hk)$.

Theorem 4.2. (Conditions under which Assumption 4.1 holds.) *Assume that $d = 2, 3$, and $(\mathcal{T}_h)_{0<h\leq h_0}$ is a family of quasi-uniform triangulations (in the sense of Definition 3.7). Assume further that each $K \in \mathcal{T}$ is the image of a triangle/tetrahedron under the image of a bi-Lipschitz map (i.e., both the map and its inverse are Lipschitz). Then $\mathcal{H}^{p,1}(\mathcal{T}_h)$ defined by (3.9) satisfies Assumption 4.1.*

References for the proof. This follows from [108, Theorem B.4] (a result about approximation on the reference element) and a scaling argument (see [108, Bottom of Page 1895]). The result [108, Theorem B.4] builds on the results of [117, Theorem 4.1], [153], and [73]; see the discussion at the start of [108, Appendix B]. □

Assumption 4.1 is about approximating in k-weighted norms an arbitrary function in H^s for some $s > 0$. In contrast, Assumption 4.3 is about approximating in k-weighted norms a function that depends on k, in the sense that its derivatives satisfy certain k-dependent bounds.

Assumption 4.3. (Approximation of certain analytic functions.) *Suppose $v \in C^\infty(B_R)$ is such that, given $k_0 > 0$ there exists $C_1, C_2 > 0$ such that*

$$\left\|(k^{-1}\partial)^\alpha v\right\|_{L^2(B_R)} \leq C_1(C_2)^{|\alpha|} \quad \text{for all } k \geq k_0. \tag{4.2}$$

Given $\widetilde{C}$, there exist $\sigma, C_{\text{approx}_2}$, depending on C_2 and $\widetilde{C}$ (but not C_1), such that, if $k \geq k_0$ and k, h, and p satisfy

$$h + \frac{hk}{p} \leq \widetilde{C}, \tag{4.3}$$

then

$$\min_{w_N \in \mathcal{H}_N} \|v - w_N\|_{H^1_k(B_R)} \leq C_1 C_{\text{approx}_2} \left[k^{-1}\left(\frac{h}{\sigma + h}\right)^p \left(\frac{1 + hk}{h + \sigma}\right) + \left(\frac{hk}{\sigma p}\right)^p \frac{1}{\sigma}\left(\frac{1}{p} + \frac{hk}{p}\right)\right]. \tag{4.4}$$

We make the following three remarks about Assumption 4.3.

(i) The term in square brackets on the right-hand side of (4.4) is small if both $h/(\sigma + h)$ and hk/p are small.

(ii) In Assumption 4.1 one cannot take $s \to \infty$ (and hence also $p \to \infty$, since $p \geq s - 1$), since the constant C_{approx_1} depends in an unspecified way on s. In contrast, Assumption 4.3 is valid for arbitrarily large p, with the right-hand side of (4.4) explicit in p; this feature is achieved by restricting attention to functions satisfying (4.2).

(iii) The term in square brackets on the right-hand side of (4.4) should be dimensionless. Strictly speaking, the factor of $h/(\sigma + h)$ should be $(h/L)(h/L + \sigma)$ and the factor $1/(\sigma + h)$ should be $(1/L)/(h/L + \sigma)$ where L is some parameter with dimension length. In writing (4.4) we have absorbed this parameter L into σ and C_{approx_2}. Similarly, the h in (4.3) should really be h/L.

The title of Assumption 4.3 is explained by the following result.

Lemma 4.4. (Analyticity from derivative bounds.) *If $u \in C^\infty(D)$ and there exist $C_1, C_2 > 0$ such that*

$$\|\partial^\alpha u\|_{L^2(D)} \leq C_1(C_2)^{|\alpha|}|\alpha|! \quad \text{for all } \alpha, \tag{4.5}$$

then u is real analytic in D.

Proof. This is Exercise 1 in §4.9. □

We see that v in (4.2) is better than analytic, in the sense that there is no $|\alpha|!$ on the right-hand side of (4.2).

Theorem 4.5. (Conditions under which Assumption 4.3 holds.) *If $(\mathcal{T}_h)_{0<h\leq h_0}$ satisfies [108, Assumption 5.2] (informally, $(\mathcal{T}_h)_{0<h\leq h_0}$ is quasi-uniform with the maps from the reference element analytic), then $\mathcal{H}^{p,1}(\mathcal{T}_h)$ defined by (3.9) satisfies Assumption 4.3.*

Proof. This is proved in [108, Proof of Theorem 5.5], using the results about *hp*-approximation of analytic functions in [108, Appendix C] – see the last equation on [108, Page 1896]. Note that (i) the weighted H^1 norm in [108] is k times that in (1.8), and (ii) the term in square brackets on the right-hand side of (4.4) is not exactly the same as the analogous term in the last equation on [108, Page 1896], since, in obtaining the latter, [108] estimate $1/(\sigma+h) \lesssim 1$ and also absorb an instance of σ into C_{approx_2}, whereas we do not. □

Note that triangulations satisfying [108, Assumption 5.2] can be constructed by refining a fixed triangulation that has analytic element maps; see [109, Remark 5.2].

4.2 Statement of the main results of this section

4.2.1 Results about the h-FEM with $p > 1$

Theorem 4.6. (Quasioptimality of the h-FEM for $p \geq 1$.) *Suppose that Assumption 1.1 holds, $\Omega_- = \emptyset$, both A and n are C^∞. Suppose that $(\mathcal{H}_N)_{N=0}^\infty$ satisfy Assumption 4.1. Then, given $p > 0$, $k_0 > 0$, there exists C_p (depending on p, A, n, k_0, and R, but independent of h and k) such that the following holds. Given $F \in (H^1(B_R))'$, let u be the solution of the variational problem (3.1) with $a(\cdot,\cdot)$ defined by (1.24) and $\mathcal{H} = H^1(B_R)$. Then if $k \geq k_0$ and h and k satisfy*

$$(hk)^p C_{\text{sol}}(k, R+2) \leq C_p, \tag{4.6}$$

then the Galerkin solution u_N exists, is unique, and satisfies the quasi-optimal error bound (3.12).

If $C_{\text{sol}} \sim k$, then the condition (4.6) is that $h^p k^{p+1}$ is sufficiently small; this implies that the pollution effect is less pronounced for p large. Numerical experiments indicate that the condition "$h^p k^{p+1}$ sufficiently small" for quasioptimality is sharp [4].

Theorem 4.7. (Non-sharp bound on the relative-error of the h-FEM for $p \geq 1$.) *Suppose that Assumptions 1.1 and 3.9 hold, with additionally $hk/p \leq 1$. Suppose that $\Omega_- = \emptyset$, both A and n are C^∞, and $(\mathcal{H}_N)_{N=0}^\infty$ satisfy Assumption 4.1. Then there exist $C_{\text{split},H^2}, C_{\text{split},\mathcal{A}} > 0$ (depending on A, n, R, and k_0) such that the following holds. Let u be the solution of the variational problem (3.1) and suppose that Assumption 3.12 holds. If $k \geq k_0$ and h and k satisfy*

$$\left(\frac{hk}{p}\right)^2 C_{\text{split},H^2} + \left(\frac{hk}{p}\right)^{p+1} C_{\text{sol}}(k, R+2)\left(C_{\text{split},\mathcal{A}}\right)^{p+1} \leq \mathcal{C}_1 \tag{4.7}$$

then the Galerkin solution u_N to the variational problem (3.2) exists, is unique, and satisfies the bound

$$\frac{\|u-u_N\|_{H^1_k(\Omega_R)}}{\|u\|_{H^1_k(\Omega_R)}} \leq \mathcal{C}_2 \frac{hk}{p} + 2\mathcal{C}_3 C_{\text{approx}_1}\left[C_{\text{split},H^2}\left(\frac{hk}{p}\right)^2 + (C_{\text{split},\mathcal{A}})^{p+1} C_{\text{sol}}(k, R+2)\left(\frac{hk}{p}\right)^{p+1}\right], \tag{4.8}$$

where

$$\mathcal{C}_1 := \frac{1}{8(C_{\text{approx}_1})^2 C_{\text{cont}\star} C_{H^2\star} n_{\max}},$$

$$\mathcal{C}_2 := \frac{2C_{\text{cont}} C_{\text{approx}_1} C_{\text{osc}}}{A_{\min}}, \quad \text{and} \quad \mathcal{C}_3 := \frac{4\sqrt{2} C_{\text{cont}\star} C_{\text{approx}_1} C_{\text{osc}} \sqrt{n_{\max} + A_{\min}}}{\sqrt{A_{\min}}}. \tag{4.9}$$

[4] experiments to come

Recall that by Assumption 4.1, C_{approx_1} in Theorem 4.7 depends on p. If $C_{\text{sol}} \sim k$, then the condition (4.7) is that $h^{p+1}k^{p+2}$ is sufficiently small. Numerical experiments indicate that the sharp condition for the relative error of the h-FEM with $p > 1$ to be controllably small when $C_{\text{sol}} \sim k$ is actually "$h^{2p}k^{2p+1}$ sufficiently small"; this has not yet been proved for the Helmholtz EDP – see the discussion in §4.8.

4.2.2 Results about the *hp*-FEM

The results in §4.2.1 imply that the larger p is, the less pronounced the pollution effect. A natural question is then, can the pollution effect be eliminated with a choice of p that $\to \infty$ as $k \to \infty$? The results in this section show that the answer is yes.

Definition 4.8. (C_{sol} is polynomially bounded in k.) *Given k_0 and $K \subset [k_0, \infty)$, $C_{\text{sol}}(k)$ is polynomially bounded for $k \in K$ if there exists $C, M > 0$ (independent of k, but depending on k_0, K, A, n, d, and R) such that*

$$C_{\text{sol}}(k) \leq Ck^M \text{ for all } k \in K. \tag{4.10}$$

Recalling Theorems 2.7 and 2.11, we see that (i) $C_{\text{sol}}(k)$ is polynomially bounded with $K = [k_0, \infty)$ when $\Omega_- = \emptyset$ and A and n are nontrapping in the sense of Definition 2.6, and (ii) given $\delta > 0$, $C_{\text{sol}}(k)$ is polynomially bounded with $K = [k_0, \infty) \setminus J$ with $|J| \leq \delta$ by Theorem 2.11 for Lipschitz Ω_- and A and bounded n.

Theorem 4.9. (Quasioptimality of the *hp*-FEM if $C_{\text{sol}}(k)$ is polynomially bounded.) *Suppose that Assumption 1.1 holds, $\Omega_- = \emptyset$, both A and n are C^∞ and $C_{\text{sol}}(k)$ is polynomially bounded (in the sense of Definition 4.8) for $k \in K \subset [k_0, \infty)$. Suppose that $(\mathcal{H}_N)_{N=0}^\infty$ satisfy Assumptions 4.1 and 4.3. Then there exist $C_1, C_2 > 0$, depending on A, n, R, and d, and k_0, but independent of k, h, and p, such that the following holds. Let u be the solution of the variational problem (3.1). If*

$$\frac{hk}{p} \leq C_1 \quad \text{and} \quad p \geq C_2 \log k, \tag{4.11}$$

then, for all $k \in K$, the Galerkin solution u_N exists, is unique, and satisfies the quasi-optimal error bound (3.12).

If h and p are chosen so that $hk/p = C_1$, then the number of degrees of freedom per wavelength is proportional to $1/C_1$ (recall the end of Remark 3.16). Theorem 4.9 therefore says that if $hk/p = C_1$ and $p \geq C_2 \log k$ then the *hp*-FEM does *not* suffer from the pollution effect.

Using in the quasi-optimal bound (3.12) the polynomial-approximation result (4.1) with $s = 2$ and then the oscillatory-behaviour bound (3.13), we obtain the following corollary of Theorem 4.9.

Corollary 4.10. (Bound on the relative error of the *hp*-FEM solution.) *Let the assumptions of Theorem 4.9 hold and, furthermore, suppose that Assumption 3.12 holds. If (4.11) holds, then, for all $k \in K$,*

$$\frac{\|u - u_N\|_{H_k^1(B_R)}}{\|u\|_{H_k^1(B_R)}} \leq \frac{2C_{\text{cont}}}{A_{\min}} C_{\text{approx}_1} C_{\text{osc}} \frac{hk}{p} \left(1 + \frac{hk}{p}\right) \leq \frac{2C_{\text{cont}}}{A_{\min}} C_{\text{approx}_1} C_{\text{osc}} C_1 \big(1 + C_1\big); \tag{4.12}$$

i.e. the relative error can be made arbitrarily small by making hk/p small.

4.3 Frequency-splitting of high-frequency Helmholtz solutions

Theorem 4.11. (Frequency-splitting of the Helmholtz solution.) *Let $\Omega_- = \emptyset$, let A and n satisfy Assumption 1.1, let $R > 0$ be such that $\operatorname{supp}(I - A) \cup \operatorname{supp}(1 - n) \Subset B_R$, and assume further that both A and n are C^∞. Given $k_0 > 0$, if $C_{\text{sol}}(k)$ is polynomially bounded (in the sense of Definition 4.8) for $k \in K \subset [k_0, \infty)$, then there exist $C_{\text{split},H^2}, C_{\text{split},\mathcal{A}} > 0$ such that the following holds. Given $f \in L^2(B_R)$, let u satisfy $k^{-2}\nabla \cdot (A\nabla u) + nu = -f$ in $\mathbb{R}^d$ and the Sommerfeld radiation condition (1.4). Then*

$$u|_{B_R} = u_{H^2} + u_{\mathcal{A}}$$

where $u_{H^2} \in H^2(B_R)$ *with*

$$\|u_{H^2}\|_{H^2_k(B_R)} \le C_{\mathrm{split},H^2} \, \|f\|_{L^2(B_R)} \qquad \text{for all } k \in K \subset [k_0, \infty), \tag{4.13}$$

and $u_{\mathcal{A}} \in C^\infty(B_R)$ *with*

$$\left\|(k^{-1}\partial)^\alpha u_{\mathcal{A}}\right\|_{L^2(B_R)} \le C_{\mathrm{sol}}(k, R+2) \left(C_{\mathrm{split},\mathcal{A}}\right)^{|\alpha|} \|f\|_{L^2(B_R)} \text{ for all } \alpha \text{ and for all } k \in K \subset [k_0, \infty), \tag{4.14}$$

The key points about this decomposition are that

(i) the bound on u_{H^2} is one power of k better than the bound on u when A and n are nontrapping – compare (4.13) to (2.4), and

(ii) the bound on $u_{\mathcal{A}}$ has the same k dependence as the bound on u – both are governed by C_{sol} – although $u_{\mathcal{A}}$ is C^∞ (and indeed analytic by Lemma 4.4).

4.4 Proofs of Theorems 4.6, 4.7, and 4.9 using Theorem 4.11

Lemma 4.12. (Bound on $\eta(\mathcal{H}_N)$ using the splitting of Theorem 4.11 and Assumption 4.1.) *Suppose Assumption 4.1 and the assumptions of Theorem 4.11 hold. Then, for all* $k \in K$,

$$\eta(\mathcal{H}_N) \le C_{\mathrm{approx}_1} \left(1 + \frac{hk}{p}\right) \left[\left(\frac{hk}{p}\right) C_{\mathrm{split},H^2} + \left(\frac{hk}{p}\right)^p C_{\mathrm{sol}}(k, R+2) \big(C_{\mathrm{split},\mathcal{A}}\big)^{p+1}\right] \tag{4.15}$$

Proof. From the definition of $\eta(\mathcal{H}_N)$ (3.21) and Lemma 3.18, it is sufficient to show the following: given $f \in L^2(B_R)$, there exists $w_N \in \mathcal{H}_N$ such that, if u is the solution of $k^{-2}\nabla\cdot(A\nabla u) + nu = -f$ in $\mathbb{R}^d$ satisfying the Sommerfeld radiation condition (1.4) (i.e., u is as in Theorem 4.11), then

$$\|u - w_N\|_{H^1_k(B_R)} \le C \, \|f\|_{L^2(B_R)}\,, \tag{4.16}$$

where C is the right-hand side of (4.15).

This follows by (i) applying (4.1) with $s = 2$ to u_{H^2} and using the bound (4.13), (ii) applying (4.1) with $s = p+1$ to $u_{\mathcal{A}}$ and using the bound (4.14), (iii) using the triangle inequality and the decomposition $u = u_{H^2} + u_{\mathcal{A}}$ on B_R. □

Theorem 4.6 follows immediately from using the bound on $\eta(\mathcal{H}_N)$ from Lemma 4.12 in Lemma 3.20. To prove Theorem 4.6 we use this bound on $\eta(\mathcal{H}_N)$ in Lemma 3.21, but we modify Lemma 3.21 to take advantage of the polynomial approximation result (4.1).

Proof of Theorem 4.7. Repeating the argument leading to Lemma 3.21 with the polynomial approximation result (3.8) replaced by (4.1), and assuming that $hk/p \le 1$, we find that

$$\frac{hk}{p}\,\eta(\mathcal{H}_N) \le \mathcal{C}_1, \qquad \text{where} \qquad \mathcal{C}_1 := \frac{1}{4C_{\mathrm{cont}\star}C_{H^2\star}C_{\mathrm{approx}_1}n_{\max}},$$

then the Galerkin solution u_h to the variational problem (3.2) exists, is unique, and satisfies the bound

$$\frac{\|u - u_h\|_{H^1_k(\Omega_R)}}{\|u\|_{H^1_k(\Omega_R)}} \le \mathcal{C}_2 \frac{hk}{p} + \mathcal{C}_3 \frac{hk}{p}\,\eta(\mathcal{H}_N),$$

where $\mathcal{C}_1, \mathcal{C}_2$ and $\mathcal{C}_3$ are as in (4.9); the result then follows from using the bound on $\eta(\mathcal{H}_N)$ (4.15). □

Theorem 4.9 is proved by using the following bound on $\eta(\mathcal{H}_N)$ in Lemma 3.20.

Lemma 4.13. (Bound on $\eta(\mathcal{H}_N)$ using the splitting from Theorem 4.11 and Assumptions 4.1 and 4.3.) *Suppose that Assumptions 4.1 and 4.3 the assumptions of Theorem 4.11 hold. Given $\widetilde{C}$, there exist $\sigma, C_{\text{approx}_2}$, depending on $C_{\text{split},\mathcal{A}}$ and $\widetilde{C}$, such that, if $k \geq k_0$ and k, h, and p satisfy*

$$h + \frac{hk}{p} \leq \widetilde{C},$$

then, for all $k \in K$,

$$\begin{aligned}\eta(\mathcal{H}_N) \leq\ & C_{\text{approx}_1} C_{\text{split},H^2} \frac{hk}{p}\left(1 + \frac{hk}{p}\right) \\ & + C_{\text{approx}_2} C_{\text{sol}}(k, R+2)\left[k^{-1}\left(\frac{h}{\sigma+h}\right)^p \left(\frac{1+hk}{\sigma+h}\right) + \left(\frac{hk}{\sigma p}\right)^p \frac{1}{\sigma}\left(\frac{1}{p} + \frac{hk}{p}\right)\right]. \quad (4.17)\end{aligned}$$

Proof. As in the proof of Lemma 4.12, it is sufficient to show that there exists $w_N \in \mathcal{H}_N$ such that (4.16) holds, where u is as in Theorem 4.11. and C is the right-hand side of (4.17).

We approximate u_{H^2} using Assumption 4.1 and $u_{\mathcal{A}}$ by Assumption 4.3. By (4.1) and (4.13), there exists $w_N^{(1)} \in \mathcal{H}_N$ such that

$$\begin{aligned}\left\| v_{H^2} - w_N^{(1)} \right\|_{H_k^1(B_R)} &\leq C_{\text{approx}_1}\left(1 + \frac{hk}{p}\right)\left(\frac{hk}{p}\right) |u|_{H_k^2(B_R)} \\ &\leq C_{\text{approx}_1}\left(1 + \frac{hk}{p}\right)\left(\frac{hk}{p}\right) C_{\text{split},H^2} \|f\|_{L^2(B_R)}. \quad (4.18)\end{aligned}$$

By (4.4) and (4.14), there exists $w_N^{(2)} \in V_N$ such that

$$\left\| v_{\mathcal{A}} - w_N^{(2)} \right\|_{H_k^1(B_R)} \leq C_{\text{approx}_2} C_{\text{sol}}(k, R+2)\left[k^{-1}\left(\frac{h}{\sigma+h}\right)^p \left(\frac{1+hk}{\sigma+h}\right) + \left(\frac{hk}{\sigma p}\right)^p \frac{1}{\sigma}\left(\frac{1}{p} + \frac{hk}{p}\right)\right]. \quad (4.19)$$

Let $w_N := w_N^{(1)} + w_N^{(2)}$. By the triangle inequality, the decomposition $u = u_{H^2} + u_{\mathcal{A}}$ on B_R, and the inequalities (4.18) and (4.19), the inequality (4.16) holds with C the right-hand side of (4.17) and the proof is complete. □

Proof of Theorem 4.9. The result follows if we can show that given $\varepsilon > 0$ and $k_0 > 0$, there exists $\mathcal{C}_1, \mathcal{C}_2 > 0$, depending only on $\varepsilon, C_{\text{approx}_1}, C_{\text{approx}_2}, C_{\text{split},H^2}, \sigma$, and k_0, such that if

$$\frac{hk}{p} \leq \mathcal{C}_1 \quad \text{and} \quad p \geq \mathcal{C}_2\Big(1 + \log\big(C_{\text{sol}}(k, R+2)\big)\Big),$$

then

$$\eta(\mathcal{H}_N) \leq \varepsilon \quad \text{for all } k \in K.$$

First choose $\mathcal{C}_1$ sufficiently small such that $\mathcal{C}_1 < \sigma$ and

$$C_{\text{approx}_1} C_{\text{split},H^2} \mathcal{C}_1 (1 + \mathcal{C}_1) \leq \frac{\varepsilon}{2}.$$

From the bound on $\eta(\mathcal{H}_N)$ (4.17), it is then sufficient to show that

$$C_{\text{approx}_2} C_{\text{sol}}(k, R+2)\left[k^{-1}\left(\frac{h}{\sigma+h}\right)^p \left(\frac{1+hk}{\sigma+h}\right) + \left(\frac{kh}{\sigma p}\right)^p \frac{1}{\sigma}\left(\frac{1}{p} + \frac{kh}{p}\right)\right] \quad (4.20)$$

can be made $\leq \varepsilon/2$. Let

$$\theta_1 := \frac{h}{\sigma + h} \quad \text{and} \quad \theta_2 := \frac{\mathcal{C}_1}{\sigma},$$

so that (4.20) is bounded by

$$C_{\text{approx}_2} C_{\text{sol}}(k, R+2)\left[k^{-1}(\theta_1)^p \left(\frac{1 + \mathcal{C}_1 p}{\sigma}\right) + (\theta_2)^p \frac{1}{\sigma}\left(\frac{1}{p} + \mathcal{C}_1\right)\right];$$

the result then follows since $\theta_1, \theta_2 < 1$. □

4.5 Definition of u_{H^2} and $u_{\mathcal{A}}$

Let $\chi \in C^\infty_{\rm comp}(\mathbb{R}^d, [0,1])$ be such that

$$\chi = \begin{cases} 1 & \text{in } B_1 \\ 0 & \text{outside } B_2. \end{cases} \tag{4.21}$$

For $\mu > 0$, let

$$\chi_\mu(\cdot) := \chi\left(\frac{\cdot}{\mu}\right), \tag{4.22}$$

and observe then that

$$\chi_\mu\big(k^{-2}|\zeta|^2\big) = \begin{cases} 1 & \text{for } |\zeta| \le \sqrt{\mu}\, k, \\ 0 & \text{for } |\zeta| \ge \sqrt{2\mu}\, k. \end{cases} \tag{4.23}$$

With the Fourier transform and its inverse defined by

$$\mathcal{F}\varphi(\zeta) := \int_{\mathbb{R}^d} \exp\big(-\mathrm{i}x\cdot\zeta\big)\varphi(x)\,\mathrm{d}x \quad \text{and} \quad \mathcal{F}^{-1}\psi(x) := (2\pi)^{-d}\int_{\mathbb{R}^d} \exp\big(\mathrm{i}x\cdot\zeta\big)\psi(\zeta)\,\mathrm{d}\zeta, \tag{4.24}$$

we define the low-frequency cut-off Π_L by

$$\Pi_L v(x) := \mathcal{F}^{-1}\Big(\chi_\mu\big(k^{-2}|\zeta|^2\big)\,\mathcal{F}v(\zeta)\Big), \tag{4.25}$$

and the high-frequency cut-off Π_H by

$$\Pi_H v(x) := \mathcal{F}^{-1}\Big(\big(1-\chi_\mu\big(k^{-2}|\zeta|^2\big)\big)\,\mathcal{F}v(\zeta)\Big), \tag{4.26}$$

so that $\Pi_L + \Pi_H = I$. This splitting contains the arbitrary parameter μ; we fix this when proving the bound (4.13) on u_{H^2}.

We let $\varphi \in C^\infty_{\rm comp}(\mathbb{R}^d, [0,1])$ be equal to one on B_{R+1} and vanish outside B_{R+2}, and then set

$$u_{\mathcal{A}} := \big(\Pi_L(\varphi u)\big)\big|_{B_R} \quad \text{and} \quad u_{H^2} := \big(\Pi_H(\varphi u)\big)\big|_{B_R}. \tag{4.27}$$

4.6 Proof of the bound (4.14) on $u_{\mathcal{A}}$

Recall that the Fourier transform satisfies

$$\mathcal{F}\big(\big(-\mathrm{i}\partial\big)^\alpha\phi\big)(\zeta) = \zeta^\alpha(\mathcal{F}\phi)(\zeta) \tag{4.28}$$

and

$$\|\phi\|_{L^2(\mathbb{R}^d)} = \frac{1}{(2\pi)^{d/2}}\,\|\mathcal{F}\phi\|_{L^2(\mathbb{R}^d)}\,. \tag{4.29}$$

The properties (4.28) and (4.29) and the definition of Π_L (4.25) imply that

$$\begin{aligned}\big\|\partial^\alpha\big(\Pi_L\varphi u\big)\big\|_{L^2(\mathbb{R}^d)} &= \frac{1}{(2\pi)^{d/2}}\,\big\|(\cdot)^\alpha\mathcal{F}\big(\Pi_L\varphi u\big)(\cdot)\big\|_{L^2(\mathbb{R}^d)} \\ &= \frac{1}{(2\pi)^{d/2}}\,\big\|(\cdot)^\alpha\chi_\mu\big(k^{-2}|\cdot|^2\big)\mathcal{F}(\varphi u)(\cdot)\big\|_{L^2(\mathbb{R}^d)}\,.\end{aligned}$$

The definitions of χ (4.21) and χ_μ (4.22) imply that $\chi_\mu(\xi) = 0$ for $|\xi| \ge 2\mu$, so

$$\chi_\mu\big(k^{-2}|\zeta|^2\big) = 0 \quad \text{for } |\zeta| \ge \sqrt{2\mu}\, k.$$

Using this fact, and then (in this order) the fact that $|\chi_\mu| \le 1$, the property (4.29), the fact that $\varphi = 0$ outside B_{R+2}, and the definition of $C_{\rm sol}$ (2.1), we find that

$$\big\|\partial^\alpha\big(\Pi_L\varphi u\big)\big\|_{L^2(\mathbb{R}^d)} \le \frac{(2\mu)^{|\alpha|/2}}{(2\pi)^{d/2}}k^{|\alpha|}\,\big\|\chi_\mu\big(k^{-2}|\cdot|^2\big)\mathcal{F}(\varphi u)(\cdot)\big\|_{L^2(\mathbb{R}^d)}$$

$$
\begin{aligned}
&\leq \frac{(2\mu)^{|\alpha|/2}}{(2\pi)^{d/2}} k^{|\alpha|} \left\| \mathcal{F}(\varphi u) \right\|_{L^2(\mathbb{R}^d)} \\
&\leq (2\mu)^{|\alpha|/2} k^{|\alpha|} \left\| \varphi u \right\|_{L^2(\mathbb{R}^d)} \\
&\leq (2\mu)^{|\alpha|/2} k^{|\alpha|} C_{\mathrm{sol}}(k, R+2) \left\| f \right\|_{L^2(B_R)} .
\end{aligned}
$$

Since

$$
\left\| \partial^\alpha u_{\mathcal{A}} \right\|_{L^2(B_R)} = \left\| \partial^\alpha \big(\Pi_L \varphi u \big) \right\|_{L^2(B_R)} \leq \left\| \partial^\alpha \big(\Pi_L \varphi u \big) \right\|_{L^2(\mathbb{R}^d)} ,
$$

the bound (4.14) then follows with $C_{\mathrm{split},\mathcal{A}} := \sqrt{2\mu}$.

4.7 Informal explanation of why the bound (4.13) on u_{H^2} holds

To complete the proof of Theorem 4.11, we need to prove the bound (4.13) on u_{H^2}. We will do this in §8, using basic results about semiclassical pseudodifferential operators from §7. However, here we give an informal explanation as to why (4.13) holds.

	Helmholtz equation	"modified Helmholtz equation"
PDE $Pu = f$	$-k^{-2}\nabla \cdot (A\nabla u) - nu = f$	$-k^{-2}\nabla \cdot (A\nabla u) + nu = f$
Symbol	$k^{-2}A_{j\ell}\zeta_j\zeta_\ell - \mathrm{i}k^{-2}\zeta_\ell \partial_j A_{j\ell} - n$	$k^{-2}A_{j\ell}\zeta_j\zeta_\ell - \mathrm{i}k^{-2}\zeta_\ell \partial_j A_{j\ell} + n$
Principal symbol	$k^{-2}A_{j\ell}\zeta_j\zeta_\ell$	$k^{-2}A_{j\ell}\zeta_j\zeta_\ell$
Elliptic?	Yes	Yes
SC symbol	$A_{j\ell}\xi_j\xi_\ell - \mathrm{i}k^{-1}\xi_\ell \partial_j A_{j\ell} - n$	$A_{j\ell}\xi_j\xi_\ell - \mathrm{i}k^{-1}\xi_\ell \partial_j A_{j\ell} + n$
SC principal symbol	$A_{j\ell}\xi_j\xi_\ell - n$	$A_{j\ell}\xi_j\xi_\ell + n$
SC elliptic?	No	Yes
Bound on $Pu = f$ in $\mathbb{R}^d$ with $\mathrm{supp} f \subset B_R$	$\|u\|_{H^1_k(B_R)} \lesssim k\|f\|_{L^2(B_R)}$ (if u satisfies (1.4))	$\|u\|_{H^1_k(\mathbb{R}^d)} \lesssim \|f\|_{L^2(B_R)}$

Table 4.1: The Helmholtz and "modified Helmholtz equation"s, their symbols, principal symbols, semiclassical (SC) symbols, and semiclassical principal symbols.

Table 4.1 compares the Helmholtz equation and the "modified Helmholtz" equation (where the sign of the term containing n is swapped).

The row labelled "Symbol" gives the symbols of these operators as pseudodifferential operators; these symbols are obtained by replacing derivatives ∂_j that act on u in the PDE by $\mathrm{i}\zeta_j$. Recall that a pseudodifferential operator is called *elliptic* if its principal symbol is never zero; we see in the table that both the Helmholtz and modified Helmholtz operators are elliptic.

In §7, we recall the concept of *semiclassical pseudodifferential operators*; we'll see that these are just standard pseudodifferential operators with a small parameter – in our case k^{-1}– where behaviour with respect to this parameter is explicitly kept track of in the associated calculus. To obtain the semiclassical (SC) symbol of a PDE operator, one replaces derivatives ∂_j that act on u by $\mathrm{i}k\xi_j$ (so that the SC symbol is just the standard symbol under a change of variables).

The Helmholtz operator is *not* SC elliptic, since its principal symbol vanishes at points (x, ξ) when $(A(x)\xi, \xi)_2 = n(x)$, whilst the modified Helmholtz operator *is* SC elliptic when A and n satisfy Assumption 1.1, since then $(A(x)\xi, \xi)_2 + n(x) \geq A_{\min}|\xi|^2 + n_{\min} > 0$.

This difference means that, at least when the equations are posed in $\mathbb{R}^d$, the norm of the modified-Helmholtz solution operator is one power of k better than the Helmholtz solution operator (see Theorem 2.7 and Exercise 2 in §4.9).

Nevertheless, the Helmholtz operator *is* SC elliptic when $|\xi| \geq \sqrt{\mu}$ for μ sufficiently large (depending on A and n), i.e., when $|\zeta| \geq \sqrt{\mu}k$, i.e., the support (in Fourier space) of the high-frequency cut-off Π_H (4.26). It is this fact that leads to $u_{H^2} := \Pi_H(\varphi u)\big|_{B_R}$ satisfying a bound that is one power of k better than the bound u itself satisfies when A and n are nontrapping.

4.8 History and context of the results in this section

Quasioptimality of the h- and hp-FEM [108, 109] [93, 94], [37].

The error of the h-FEM for $p > 1$ [151, 154, 50], [122]

4.9 Exercises for Section 4

1. Prove Lemma 4.4 via the following steps.

   (a) Show that the result follows if there exists $n_0 \in \mathbb{Z}^+$ such that

   $$\|\partial^\alpha u\|_{L^\infty(D)} \leq \widetilde{C}_1(\widetilde{C}_2)^{|\alpha|}(|\alpha| + n_0)!. \tag{4.30}$$

   Hint: bound the Lagrange form of the remainder in the Taylor-series up to $n-1$ terms, i.e.,

   $$\sum_{|\alpha|=n} \frac{(x-x')^\alpha}{\alpha!}\big(\partial^\alpha u\big(x' + c(x-x')\big)\big),$$

   for some $c \in (0,1)$, and use the consequence of the binomial theorem that

   $$\sum_{|\alpha|=n} \frac{n!}{\alpha!} = d^n. \tag{4.31}$$

   (b) Prove (4.30) using the Sobolev embedding theorem (see, e.g., [106, Theorem 3.26]).

2. (Proof of the bound on the solution of the "modified Helmholtz equation" in Table 4.1.) Given $f \in L^2(\mathbb{R}^d)$, and A and n satisfying Assumption 1.1 with $\Omega_- = \emptyset$, let $u \in H^1(\mathbb{R}^d)$ be the solution of $-k^{-2}\nabla \cdot (A\nabla u) + nu = f$ in $\mathbb{R}^d$. Prove that u exists, is unique, and satisfies the bound

   $$\|u\|_{H^1_k(\mathbb{R}^d)} \leq \frac{1}{\min\{A_{\min}, n_{\min}\}}\|f\|_{L^2(\mathbb{R}^d)}$$

   for all $k > 0$. Hint: consider the variational problem satisfied by u.

5 Semiclassical Fourier multipliers

5.1 Plan for the next few sections

To complete the proof of Theorem 4.11, we need to prove the bound (4.13) on u_{H^2}. As discussed in §4.7, our proof of the bound (4.13) on u_{H^2} uses basic results about semiclassical pseudodifferential operators; these results are recapped in §7, with the bound on u_{H^2} proved in §8.

As a warm-up, in this section we recap results about (semiclassical) Fourier multipliers, i.e., operators defined by multiplication in Fourier space. In §6 we use these results to prove the bound (4.13) on u_{H^2} when $A = I$ and $n = 1$. Importantly, the results in this section and the corresponding proof in §6 only involve basic facts about the Fourier transform (namely, the rule for the Fourier transform of a derivative, and Plancherel's theorem).

Since pseudodifferential operators are a generalisation of Fourier multipliers, the goal of this section and §6 is to provide a "bridge" into the theory and use of pseudodifferential operators, and to illustrate to an audience unfamiliar with the theory of pseudodifferential operators how this theory is the natural generalisation of Fourier analysis to study linear PDEs with variable coefficients.

5.2 The semiclassical parameter $\hbar = k^{-1}$

Instead of working with the parameter k and being interested in the large-k limit, the semiclassical literature usually works with a parameter $h := k^{-1}$ and is interested in the small-h limit. So that we can easily recall results from this literature, we also work with the small parameter k^{-1}, but to avoid a notational clash with the meshwidth of the FEM, we let $\hbar := k^{-1}$ (the notation $\hbar$ comes from the fact that the semiclassical parameter is related to Planck's constant, which is written as $2\pi\hbar$; see, e.g., [155, §1.2], [52, Page 82], [105, Chapter 1]).

5.3 The semiclassical Fourier transform $\mathcal{F}_\hbar$

The semiclassical Fourier transform is defined for $\hbar > 0$ by

$$\mathcal{F}_\hbar \phi(\xi) := \int_{\mathbb{R}^d} \exp\big(-\mathrm{i}x\cdot\xi/\hbar\big)\phi(x)\,\mathrm{d}x;$$

i.e.,

$$\mathcal{F}_\hbar \phi(\xi) = \mathcal{F}\phi(\xi/\hbar), \tag{5.1}$$

where $\mathcal{F}$ is defined in (4.24). The inverse of $\mathcal{F}$ is given by

$$\mathcal{F}_\hbar^{-1}\psi(x) := (2\pi\hbar)^{-d}\int_{\mathbb{R}^d} \exp\big(\mathrm{i}x\cdot\xi/\hbar\big)\psi(\xi)\,\mathrm{d}\xi; \tag{5.2}$$

see [155, §3.3]. Let

$$\mathscr{S}(\mathbb{R}^d) := \Big\{\phi\in C^\infty(\mathbb{R}^d) : \sup_{x\in\mathbb{R}^d}\big|x^\alpha\partial^\beta\phi(x)\big| < \infty \text{ for all multiindices } \alpha \text{ and } \beta\Big\}; \tag{5.3}$$

i.e., $\mathscr{S}(\mathbb{R}^d)$ is the Schwartz space of rapidly decreasing, C^∞ functions. Let $\mathscr{S}^*(\mathbb{R}^d)$ be the space of continuous linear functions on $\mathscr{S}(\mathbb{R}^d)$; recall that $\mathcal{F}_\hbar : \mathscr{S}(\mathbb{R}^d)\to\mathscr{S}(\mathbb{R}^d)$ and then, by duality, $\mathcal{F}_\hbar : \mathscr{S}^*(\mathbb{R}^d)\to\mathscr{S}^*(\mathbb{R}^d)$ (see, e.g., [106, Page 72]). Recall also the property

$$\mathcal{F}_\hbar\big(\big(-\mathrm{i}\hbar\partial\big)^\alpha\phi\big)(\xi) = \xi^\alpha\,\mathcal{F}_\hbar\phi(\xi) \tag{5.4}$$

and Plancherel's theorem (sometimes known as Parseval's theorem)

$$\|\phi\|_{L^2(\mathbb{R}^d)} = \frac{1}{(2\pi\hbar)^{d/2}}\,\|\mathcal{F}_\hbar\phi\|_{L^2(\mathbb{R}^d)}\,. \tag{5.5}$$

5.4 Semiclassical (i.e., weighted) Sobolev spaces

For $s\in\mathbb{R}$, let

$$H^s_\hbar(\mathbb{R}^d) := \Big\{u\in\mathscr{S}^*(\mathbb{R}^d),\ \langle\xi\rangle^s\mathcal{F}_\hbar u\in L^2(\mathbb{R}^d)\Big\},\qquad \text{where } \langle\xi\rangle := (1+|\xi|^2)^{1/2},$$

and let

$$\|u\|^2_{H^s_\hbar(\mathbb{R}^d)} := (2\pi\hbar)^{-d}\int_{\mathbb{R}^d}\langle\xi\rangle^{2s}|\mathcal{F}_\hbar u(\xi)|^2\,\mathrm{d}\xi; \tag{5.6}$$

we abbreviate $H^s_\hbar(\mathbb{R}^d)$ to $H^s_\hbar$ and $L^2(\mathbb{R}^d)$ to L^2. Thanks to (5.4), up to dimension-dependent constants, $\|u\|_{H^s_\hbar(\mathbb{R}^d)}$ defined by (5.6) is equivalent to $\|u\|_{H^s_k(\mathbb{R}^d)}$ defined by (1.7); we use this clashing notation to avoid writing $H^s_{\hbar^{-1}}(\mathbb{R}^d)$ and $\|\cdot\|_{H^s_{\hbar^{-1}}(\mathbb{R}^d)}$.

To be more precise about the norm equivalence, let $C_j = C_j(s,d) > 0$, $j=1,2$, be such that

$$C_1\sum_{|\alpha|\le s}\xi^{2\alpha} \le (1+|\xi|^2)^s \le C_2\sum_{|\alpha|\le s}\xi^{2\alpha},$$

then

$$\sqrt{C_1}\,\|u\|_{H^s_k(\mathbb{R}^d)} \le \|u\|_{H^s_\hbar(\mathbb{R}^d)} \le \sqrt{C_2}\,\|u\|_{H^s_k(\mathbb{R}^d)}\,.$$

Finally, recall that, for $s\in\mathbb{R}$, $H^{-s}_\hbar(\mathbb{R}^d)$ is an isometric realisation of the dual space of $H^s_\hbar(\mathbb{R}^d)$; i.e.,

$$H^{-s}_\hbar(\mathbb{R}^d) = (H^s_\hbar(\mathbb{R}^d))^*; \tag{5.7}$$

see, e.g., [106, Page 76].

5.5 Definition of Fourier multipliers

We say that $a \in C^\infty(\mathbb{R}^d)$ is a *Fourier symbol* if there exists $m \in \mathbb{R}$ such that for any multiindex β there exists C_β such that

$$|\partial_\xi^\beta a(\xi)| \le C_\beta \langle \xi \rangle^{m-|\beta|} \quad \text{for all } \xi \in \mathbb{R}^d. \tag{5.8}$$

We say that m is the *order* of the Fourier symbol and use the (non-standard) notation that $a \in (FS)^m$.

Example 5.1. (Examples of Fourier symbols.)
(i) $a(x,\xi) := \sum_{|\alpha|\le m} a_\alpha \xi^\alpha$, *where* a_α *are constants, is in* $(FS)^m$.
(ii) $a(\xi) := |\xi|^2 - 1 \in (FS)^2$.
(iii) $\langle \xi \rangle^{-m} := (1+|\xi|^2)^{-m/2} \in (FS)^{-m}$.
(iv) If $\chi \in C^\infty_{\rm comp}(\mathbb{R}^d)$, *then* $\chi \in (FS)^{-N}$ *for all* $N \ge 1$.

Given a Fourier symbol a, the *Fourier multiplier* defined by a is given by

$$\big(a(\hbar D)v\big)(x) = \mathcal{F}_\hbar^{-1}\big(a(\cdot)(\mathcal{F}_\hbar v)(\cdot)\big)(x). \tag{5.9}$$

The rationale for this notation is that, by (5.4), the semiclassical Fourier transform $\mathcal{F}_\hbar$ converts $\hbar D$, where $D := -i\partial$, into ξ.

Lemma 5.2. $a(\hbar D) : \mathscr{S}(\mathbb{R}^d) \to \mathscr{S}(\mathbb{R}^d)$ *and* $\mathscr{S}^*(\mathbb{R}^d) \to \mathscr{S}^*(\mathbb{R}^d)$.

Proof. This is Exercise 1 in §5.6. Note that this proof is the only place in this section where (5.8) with $|\beta| > 0$ (i.e., the differentiability property of a) is used. □

Example 5.3. (Examples of Fourier multipliers.)
(i) If $a(\xi) = 1$, *then* $a(\hbar D)v(x) = v(x)$; *i.e.,* $1(\hbar D) = I$.
(ii) If $a(\xi) := \sum_{|\alpha|\le m} a_\alpha \xi^\alpha$, *then* $a(\hbar D)v(x) = \sum_{|\alpha|\le m} a_\alpha (\hbar D)^\alpha v(x)$.
(iii) If $a(\xi) := |\xi|^2 - 1$, *then* $a(\hbar D)v(x) = (-\hbar^2\Delta - 1)v$.

Theorem 5.4. (Composition and mapping properties of Fourier multipliers.) *If* $a \in (FS)^{m_1}$ *and* $b \in (FS)^{m_2}$ *then the following hold.*
(i) $ab \in (FS)^{m_1+m_2}$.
(ii) $a(\hbar D)b(\hbar D) = (ab)(\hbar D) = b(\hbar D)a(\hbar D)$.
(iii) $a(\hbar D) : H^s_\hbar \to H^{s-m_1}_\hbar$ *and there exists* $C > 0$ *such that, for all* $s \in \mathbb{R}$ *and* $\hbar > 0 m$

$$\|a(\hbar D)\|_{H^s_\hbar \to H^{s-m_1}_\hbar} \le C.$$

i.e., $a(\hbar D)$ *is bounded uniformly in both* $\hbar$ *and* s *as an operator from* $H^s_\hbar$ *to* $H^{s-m_1}_\hbar$.

Proof. This is Exercise 2 in §5.6. □

5.6 Exercises for §5

1. Prove Lemma 5.2.
2. Prove Theorem 5.4.

6 Proof of the bound on u_{H^2} when $A = I$ and $n = 1$

Recap of the definition and properties of u when $A = I$ and $n = 1$. Given $f \in L^2(B_R)$, let u be the solution of $P_\hbar u := (-\hbar^2\Delta - 1)u = f$ in $\mathbb{R}^d$ satisfying the Sommerfeld radiation condition (1.4) (with $k = \hbar^{-1}$). Recall that

$$\|u\|_{H^1_\hbar(B_R)} \le 2\hbar^{-1} R \sqrt{1 + \left(\frac{d-1}{2kR}\right)^2}\, \|f\|_{L^2(B_R)} \quad \text{for all } \hbar > 0 \tag{6.1}$$

by Theorem 2.14 (taking $\Omega_- = \emptyset$).

Recap of the definition of u_{H^2} and the bound (4.13) we need to prove. Let $\varphi \in C^\infty_{\mathrm{comp}}(\mathbb{R}^d, [0,1])$ be equal to one on B_{R+1} and vanish outside B_{R+2}, and set

$$u_{H^2} := \big(\Pi_H(\varphi u)\big)\big|_{B_R}, \tag{6.2}$$

where Π_H is defined in terms of the Fourier transform by (4.26). More precisely, Π_H contains the arbitrary parameter μ, and then u_{H^2} is defined by (6.2) with μ chosen sufficiently large (as prescribed below).

Using the equivalence of $\|\cdot\|_{H^2_k}$ and $\|\cdot\|_{H^2_\hbar}$, the bound (4.13) is equivalent to

$$\|u_{H^2}\|_{H^2_\hbar(B_R)} \lesssim \|f\|_{L^2(B_R)} \qquad \text{for all } 0 < \hbar \le \hbar_0. \tag{6.3}$$

Note that, by (6.1), in the set up we are considering ($A = I$, $n = 1$, and $\Omega_- = \emptyset$) C_{sol} is polynomially bounded for all k (in fact $C_{\mathrm{sol}} \lesssim k$), and thus we prove the bound (4.13) with $K = [k_0, \infty)$.

The Helmholtz operator with $A = I$ and $n = 1$ and Π_H as Fourier multipliers. By Example 5.3 (iii), $P_\hbar := (-\hbar^2\Delta - 1) = p(\hbar D)$ with $p(\xi) := |\xi|^2 - 1$. By the relationship (5.1) between $\mathcal{F}$ and $\mathcal{F}_\hbar$ and the definition (5.9) of a Fourier multiplier,

$$\Pi_H = \mathcal{F}_\hbar^{-1}\big(1 - \chi_\mu(|\cdot|^2)\big)\mathcal{F}_\hbar = \big(1 - \chi_\mu(|\cdot|^2)\big)(\hbar D). \tag{6.4}$$

Proof of the bound (4.13) on u_{H^2} when $A = I$ and $n = 1$. By (6.4) and Part (ii) of Theorem 5.4,

$$\|\Pi_H(\varphi u)\|_{H^2_\hbar(\mathbb{R}^d)} = \big\|\big(1 - \chi_\mu(|\cdot|^2)\big)(\hbar D)(\varphi u)\big\|_{H^2_\hbar(\mathbb{R}^d)} = \left\|\left(\frac{1 - \chi_\mu(|\cdot|^2)}{p(\cdot)}\right)(\hbar D)\, p(\hbar D)(\varphi u)\right\|_{H^2_\hbar(\mathbb{R}^d)}. \tag{6.5}$$

Lemma 6.1. *If $\mu \ge 4$, then $(1 - \chi_\mu(|\xi|^2)/p(\xi) \in (FS)^{-2}$.*

Proof. By the definitions of χ_μ (4.23) (recalling that $\xi = \zeta/k$) and $p(\xi) := |\xi|^2 - 1$,

$$\frac{1 - \chi_\mu(|\xi|^2)}{p(\xi)} = \begin{cases} 0 & \text{for } |\xi| \le \sqrt{\mu}, \\ \big(1 - \chi_\mu(|\xi|^2)\big)(|\xi|^2 - 1)^{-1} & \text{for } \sqrt{\mu} \le |\xi| \le \sqrt{2\mu}, \\ (|\xi|^2 - 1)^{-1} & \text{for } |\xi| \ge \sqrt{2\mu}. \end{cases}$$

Therefore, if $\mu \ge 4$, then $(1 - \chi_\mu(|\xi|^2)/p(\xi)$ is bounded for all $\xi \in \mathbb{R}$. The derivative bounds in (5.8) with $m = -2$ follow by differentiating, essentially using the fact that the symbol equals $1/p$ outside a compact set. □

By using Lemma 6.1 and Part (iii) of Theorem 5.4 in (6.5), we have

$$\|\Pi_H(\varphi u)\|_{H^2_\hbar(\mathbb{R}^d)} \lesssim \|p(\hbar D)(\varphi u)\|_{L^2} = \|P_\hbar(\varphi u)\|_{L^2} = \|\varphi P_\hbar u + [P_\hbar, \varphi]u\|_{L^2} \lesssim \|f\|_{L^2} + \|[P_\hbar, \varphi]u\|_{L^2}, \tag{6.6}$$

where we have used the fact that $\varphi \equiv 1$ on $\operatorname{supp} f$, and where the commutator $[A, B]$ is defined as $AB - BA$. By direct calculation,

$$[P, \varphi]u = -\hbar^2\big(u\Delta\varphi + 2\nabla\varphi \cdot \nabla u\big),$$

so that

$$\big\|[P_\hbar, \varphi]u\big\|_{L^2} \lesssim \hbar\, \|u\|_{H^1_\hbar(B_{R+2})}, \tag{6.7}$$

where the omitted constant depends on φ. Therefore, by combining (6.6) and (6.7), and using (6.1) (with R replaced by $R + 2$), we have

$$\|\Pi_H(\varphi u)\|_{H^2_\hbar(\mathbb{R}^d)} \lesssim \|f\|_{L^2} + \hbar\, \|u\|_{H^1_\hbar(B_{R+2})} \lesssim \|f\|_{L^2};$$

since $\|u_{H^2}\|_{H^2_\hbar(B_R)} \le \|\Pi_H(\varphi u)\|_{H^2_\hbar(\mathbb{R}^d)}$, the result (6.3) follows.

Remark 6.2. (The analogue of the elliptic parametrix for Fourier multipliers.) *The above proof uses the fact that if $a \in (FS)^{m_A}$, $b \in (FS)^{m_B}$, and there exists $c > 0$ such that $b(\xi) \geq c$ for $\xi \in \operatorname{supp} a$ then*

$$a(\hbar D) = q(\hbar D)b(\hbar D)$$

where $q \in (FS)^{m_A - m_B}$ is defined by $q(\xi) := a(\xi)/b(\xi)$. The generalisation of this result to pseudodifferential operators is the so-called elliptic parametrix in Theorem 7.24.

7 Semiclassical pseudodifferential operators

This section states results about semiclassical pseudodifferential operators. When the proofs are straightforward, we give them here, but we postpone the more-involved proofs to §9. The intermediate section (§8) then uses these results to prove the bound (4.13) on u_{H^2}. The reason for separating the more-involved proofs of the pseudodifferential-operator results from their statements is that I want you to see as soon as possible these results "in action" in the proof of (4.13). I hope that seeing how these pseudodifferential-operator results naturally lead to the proof of (4.13) will then provide motivation to tackle the more-involved proofs; we highlight at this stage that our presentation of the pseudodifferential-operator material is based on those in [155], [52, Appendix E], and [59].

7.1 Semiclassical pseudodifferential operators vs "standard" pseudodifferential operators

As mentioned in §4, semiclassical pseudodifferential operators are pseudodifferential operators with a small parameter – in our case k^{-1}– where behaviour with respect to this parameter is explicitly kept track of in the associated calculus. Semiclassical pseudodifferential operators are effective for problems where oscillations happen at frequency k (which is assumed to be large); and are thus tailor-made to study high-frequency Helmholtz problems.

The standard (a.k.a. *homogeneous*) – as opposed to semiclassical – versions of the results in this section can be found in, e.g., [145, Chapter 7], [128, Chapter 7], [81, Chapter 6]. The use of homogeneous pseudodifferential operators in numerical analysis is well established, particular in the field of boundary integral equations; see, e.g., [44, 143, 45, 43, 132, 98, 128, 99, 5, 81] (for a selection of well-established results) and, e.g., [47, 22, 10, 4, 6, 67, 32] (for a selection of recent results). However, perhaps surprisingly, the pseudodifferential operators tailor-made for studying the Helmholtz equation, namely semiclassical pseudodifferential operators, have been little used in the numerical analysis of the Helmholtz equation.

7.2 Phase space

The set of all possible positions x and momenta (i.e. Fourier variables) ξ is denoted by $T^*\mathbb{R}^d$; this is known informally as "phase space". Strictly, $T^*\mathbb{R}^d := \mathbb{R}^d \times (\mathbb{R}^d)^*$, but for our purposes, we can consider $T^*\mathbb{R}^d$ as $\{(x,\xi) : x \in \mathbb{R}^d, \xi \in \mathbb{R}^d\}$. (This notation comes from the fact that pseudodifferential operators on a general manifold M are defined using the notion of the *cotangent bundle* T^*M; see [155, Chapter 14].)

7.3 Symbols, quantisation, and semiclassical pseudodifferential operators

A symbol is a function on $T^*\mathbb{R}^d$ that is also allowed to depend on $\hbar$ (i.e., it is an $\hbar$-dependent family of functions) Such a family $a = (a_\hbar)_{0<\hbar\leq\hbar_0}$, with $a_\hbar \in C^\infty(T^*\mathbb{R}^d)$, is a *symbol of order m*, written as $a \in S^m(\mathbb{R}^d)$, if, for any multiindices α, β, there exists $C_{\alpha,\beta}$ such that

$$|\partial_x^\alpha \partial_\xi^\beta a_\hbar(x,\xi)| \leq C_{\alpha,\beta}\langle\xi\rangle^{m-|\beta|} \quad \text{for all } (x,\xi) \in T^*\mathbb{R}^d \text{ and for all } 0 < \hbar \leq \hbar_0 \tag{7.1}$$

(i.e., $C_{\alpha,\beta}$ does not depend on $\hbar$, x, or ξ); see [155, p. 207], [52, §E.1.2]. In these notes, we only consider these symbol classes on $\mathbb{R}^d$, and so we abbreviate $S^m(\mathbb{R}^d)$ to S^m. In the literature, one

usually omits the $\hbar$ dependence of a in the notation, writing $a(x,\xi)$ instead of $a_\hbar(x,\xi)$, and we do the same here.

Example 7.1. (Examples of symbols.) *Exercise 1 in §7.11 asks you to show that*
(i) $a(x,\xi) := \sum_{|\alpha|\le m} a_\alpha(x)\xi^\alpha$, *where* $a_\alpha \in C^\infty$ *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 $\chi \in C^\infty_{\mathrm{comp}}(T^*\mathbb{R}^d)$, *then* $\chi \in S^{-N}$ *for every* $N \ge 1$.

Remark 7.2. (Kohn–Nirenberg symbols.) *The symbol class* S^m *defined above, where one gains one power of* $\langle\xi\rangle$ *on differentiation with respect to* ξ, *and no powers of* $\langle\xi\rangle$ *on differentiation with respect to* x, *is known as the Kohn–Nirenberg symbol class [92]; much wider classes of symbols exist – see, e.g., [145, Chapter 7, §1], [155, §4.4.1], [52, Equation E.1.48].*

For $a \in S^m$, we define the *semiclassical quantisation* of a, $\mathrm{Op}_\hbar(a)$ by (see, e.g., [155, §4.1] [52, Page 543])

$$\big(\mathrm{Op}_\hbar(a)v\big)(x) := (2\pi\hbar)^{-d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \exp\big(\mathrm{i}(x-y)\cdot\xi/\hbar\big)\, a(x,\xi)v(y)\,\mathrm{d}y\mathrm{d}\xi \tag{7.2}$$

for $v \in \mathscr{S}(\mathbb{R}^d)$, where the integral is understood as an iterated integral, with the y integration performed first, i.e.,

$$\big(\mathrm{Op}_\hbar(a)v\big)(x) = (2\pi\hbar)^{-d}\int_{\mathbb{R}^d} \exp\big(\mathrm{i}x\cdot\xi/\hbar\big)\, a(x,\xi)\mathcal{F}_\hbar v(\xi)\,\mathrm{d}\xi. \tag{7.3}$$

In analogy with the notation $a(\hbar D)$ for Fourier multipliers (5.9), the semiclassical quantisation $\mathrm{Op}_\hbar(a)$ is often denoted by $a(x,\hbar D)$.

Lemma 7.3. $\mathrm{Op}_\hbar(a) : \mathscr{S}(\mathbb{R}^d) \to \mathscr{S}(\mathbb{R}^d)$.

Proof. This is Exercise 2 in §7.11. □

Conversely, if A can be written in the form above, i. e. $A = \mathrm{Op}_\hbar(a)$ with $a \in S^m$, then A is a *semiclassical pseudodifferential operator of order* m and we write $A \in \Psi^m_\hbar(\mathbb{R}^d)$, which we then abbreviate to $A \in \Psi^m_\hbar$.

We say $a \in S^{-\infty}$ if $a \in S^{-N}$ for all $N \ge 1$. We say $a \in \hbar^l S^m$ if $\hbar^{-l}a \in S^m$; similarly $A \in \hbar^l\Psi^m_\hbar$ if $\hbar^{-l}A \in \Psi^m_\hbar$.

Example 7.4. (Examples of quantisations.) *(Compare to Example 5.3.)*
(i) If $a(x,\xi) = 1$, *then* $(\mathrm{Op}_\hbar(a)v)(x) = v(x)$, *i.e.,* $\mathrm{Op}_\hbar(1) = I$.
(ii) If $a(x,\xi) = a(x)$, *then* $(\mathrm{Op}_\hbar(a)v)(x) = a(x)v(x)$.
(iii) If $a(x,\xi) = a(\xi)$, *then* $(\mathrm{Op}_\hbar(a)v)(x) = \mathcal{F}_\hbar^{-1}\big(a(\cdot)(\mathcal{F}_\hbar v)(\cdot)\big)(x)$, *i.e.,* $\mathrm{Op}_\hbar(a)$ *is a Fourier multiplier (5.9).*
(iv) If $a(x,\xi) := \sum_{|\alpha|\le m} a_\alpha(x)\xi^\alpha$, *where* $a_\alpha \in C^\infty$, *then* $(\mathrm{Op}_\hbar(a)v)(x) = \sum_{|\alpha|\le m} a_\alpha(x)(\hbar D)^\alpha$ *(recall that* $D := -\mathrm{i}\partial$*).*

In the following result, given $A : \mathscr{S}(\mathbb{R}^d) \to \mathscr{S}(\mathbb{R}^d)$, its formal adjoint $A^* : \mathscr{S}^*(\mathbb{R}^d) \to \mathscr{S}^*(\mathbb{R}^d)$ is defined by $\langle A^*u, v\rangle_{\mathbb{R}^d} = \langle u, Av\rangle_{\mathbb{R}^d}$.

Theorem 7.5. (Composition and mapping properties of semiclassical pseudodifferential operators.) *If* $A \in \Psi^{m_A}_\hbar$ *and* $B \in \Psi^{m_B}_\hbar$, *then*

(i) $A^* : \mathscr{S}(\mathbb{R}^d) \to \mathscr{S}(\mathbb{R}^d)$ *and* $A^* \in \Psi^{m_A}_\hbar$,

(ii) $AB \in \Psi^{m_A+m_B}_\hbar$,

(iii) $[A,B] := AB - BA \in \hbar\Psi^{m_A+m_B-1}_\hbar$,

(iv) Given $s \in \mathbb{R}$ *and* $\hbar_0 > 0$, *there exists* $C > 0$ *such that*

$$\|A\|_{H^s_\hbar \to H^{s-m_A}_\hbar} \le C \quad \text{for all } 0 < \hbar \le \hbar_0;$$

i.e., A is bounded uniformly in $\hbar$ *as an operator from* $H^s_\hbar$ *to* $H^{s-m_A}_\hbar$.

Theorem 5.4 and Exercise 3 in §7.11 show that Theorem 7.5 is straightforward to prove if the pseudodifferential operators are Fourier multipliers. In §9 we see that the proof in the general case is much more involved.

7.4 Residual class

At the level of operators, we say that $A = O(\hbar^\infty)_{\Psi^{-\infty}}$ if, for any $s > 0$ and $N \geq 1$, there exists $C_{s,N} > 0$ such that

$$\|A\|_{H_\hbar^{-s} \to H_\hbar^s} \leq C_{s,N} \hbar^N; \tag{7.4}$$

i.e., all of the operator norms are bounded by any algebraic power of $\hbar$.

At the level of symbols, we say that

$$a \in \hbar^\infty S^{-\infty} \quad \text{if} \quad a \in \bigcap_{N=1}^{\infty} \hbar^N S^{-N},$$

that is, for any multiindices α, β and any $N \geq 1$, there exists $C_{\alpha,\beta,N} > 0$ so that

$$|\partial_x^\alpha \partial_\xi^\beta a(x,\xi)| \leq C_{\alpha,\beta,N} \hbar^N \langle \xi \rangle^{-N} \quad \text{for all } (x,\xi) \in T^*\mathbb{R}^d \tag{7.5}$$

(see [52, E.1.10]).

Lemma 7.6. *If $a \in \hbar^\infty S^{-\infty}$, then $\operatorname{Op}_\hbar(a) = O(\hbar^\infty)_{\Psi^{-\infty}}$.*

Proof. This is Exercise 4 in §7.11. □

7.5 The principal symbol $\sigma_\hbar$

Let the quotient space $S^m/\hbar S^{m-1}$ be defined by identifying elements of S^m that differ only by an element of $\hbar S^{m-1}$.

Definition 7.7. (Principal symbol.) *For any m, let the* principal symbol map

$$\sigma_\hbar^m : \Psi_\hbar^m \to S^m/\hbar S^{m-1},$$

be defined for $a \in S^m$ by

$$\sigma_\hbar^m\big(\operatorname{Op}_\hbar(a)\big) = a \mod \hbar S^{m-1}. \tag{7.6}$$

Observe that $\sigma_\hbar^m$ is linear and surjective, and that $\ker(\sigma_\hbar^m) = \hbar\Psi_\hbar^{m-1}$. When applying the map $\sigma_\hbar^m$ to elements of $\Psi_\hbar^m$, we denote it by $\sigma_\hbar$ (i.e. we omit the m dependence) and we use $\sigma_\hbar(A)$ to denote one of the representatives in S^m (with the results we use then independent of the choice of representative).

Exercise 5 in §7.11 asks you to show that if $P_\hbar u := -\hbar^2 \nabla \cdot (A\nabla u) - nu$, then $\sigma_\hbar(P_\hbar) = (A\xi)\cdot\xi - n$, as claimed in Table 4.1.

The following result involves the Poisson bracket $\{\cdot,\cdot\}$ defined by

$$\{a,b\} := \sum_j \Big((\partial_{\xi_j} a)(\partial_{x_j} b) - (\partial_{x_j} a)(\partial_{\xi_j} b) \Big) = \langle \partial_{\xi_j} a, \overline{\partial_{x_j} b} \rangle - \langle \partial_{x_j} a, \overline{\partial_{\xi_j} b} \rangle. \tag{7.7}$$

Lemma 7.8. (Key properties of the principal symbol.)

$$\sigma_\hbar(A^*) = \overline{\sigma_\hbar(A)}, \tag{7.8}$$

$$\sigma_\hbar(AB) = \sigma_\hbar(A)\sigma_\hbar(B), \tag{7.9}$$

$$\sigma_\hbar\left(\frac{\mathrm{i}}{\hbar}[A,B]\right) = \big\{\sigma_\hbar(A), \sigma_\hbar(B)\big\}, \tag{7.10}$$

and

$$\text{if } e \in \hbar^\infty S^{-\infty} \text{ then } \sigma_\hbar\big(\operatorname{Op}_\hbar(a+e)\big) = \sigma_\hbar\big(\operatorname{Op}_\hbar(a)\big). \tag{7.11}$$

Proof. (7.8), (7.9), and (7.10) are proved in §9. (7.11) follows from (7.6), since if $a \in S^m$, then $e \in \hbar S^{m-1}$. □

7.6 Operator wavefront set $\mathrm{WF}_\hbar$

Definition 7.9. (Operator wavefront set.) $(x_0,\xi_0)\in T^*\mathbb{R}^d$ *is* not *in the* semiclassical operator wavefront set *of* $A=\mathrm{Op}_\hbar(a)\in\Psi_\hbar^m$*, denoted by* $\mathrm{WF}_\hbar A$*, if there exists a neighbourhood* U *of* (x_0,ξ_0) *such that for all multiindices* α,β *and all* $N\geq 1$ *there exists* $C_{\alpha,\beta,N,U}>0$ *such that*

$$|\partial_x^\alpha\partial_\xi^\beta a(x,\xi)|\leq C_{\alpha,\beta,N,U}\hbar^N \quad \text{for all } (x,\xi)\in U \text{ and } 0<\hbar\leq\hbar_0; \tag{7.12}$$

That is, outside the semiclassical wavefront set of the operator, the symbol vanishes faster than any algebraic power of $\hbar$.

Lemma 7.10. *If* a *is independent of* $\hbar$*, then* $\mathrm{WF}_\hbar(\mathrm{Op}_\hbar(a))=\operatorname{supp} a$.

Proof. This is Exercise 6 in §7.11. □

Lemma 7.11. (Key properties of the semiclassical operator wavefront set.) *If* $A\in\Psi_\hbar^{m_A}$ *and* $B\in\Psi_\hbar^{m_B}$*, then*

$$\mathrm{WF}_\hbar(A+B)\subset\mathrm{WF}_\hbar A\cup\mathrm{WF}_\hbar B, \tag{7.13}$$

$$\mathrm{WF}_\hbar(AB)\subset\mathrm{WF}_\hbar A\cap\mathrm{WF}_\hbar B, \tag{7.14}$$

$$\mathrm{WF}_\hbar\big(\mathrm{Op}_\hbar(a)\big)\subset\operatorname{supp} a, \tag{7.15}$$

and

$$\text{if } e\in\hbar^\infty S^{-\infty} \text{ then } \mathrm{WF}_\hbar\big(\mathrm{Op}_\hbar(a+e)\big)=\mathrm{WF}_\hbar\big(\mathrm{Op}_\hbar(a)\big). \tag{7.16}$$

Proof. By De Morgan's laws, (7.13) is equivalent to $(\mathrm{WF}_\hbar A)^c\cap(\mathrm{WF}_\hbar B)^c\subset(\mathrm{WF}_\hbar(A+B))^c$, and this follows from the definition of $\mathrm{WF}_\hbar$ (7.12). (7.14) is proved in §9. (7.15) holds because since $(\operatorname{supp} a)^c\subset(\mathrm{WF}_\hbar(\mathrm{Op}_\hbar(a)))^c$ by (7.12), and (7.16) follows from the definitions of $\hbar^\infty S^{-\infty}$ (7.5) and $\mathrm{WF}_\hbar$ (7.12). □

Recall that the informal explanation in §4.7 of why the bound (4.13) on u_{H^2} holds talked about "the support (in Fourier space) of the high-frequency cut-off Π_H (4.26)". The notion of $\mathrm{WF}_\hbar$ allows us to formulate this notion of support, and we see in (8.6) below that (as a consequence of Lemma 7.10) $\mathrm{WF}_\hbar(\Pi_H)=\big\{\xi\,:\,|\xi|^2\geq\mu\big\}$.

7.7 Schwartz kernel

Let $\mathcal{D}(\mathbb{R}^d):=C^\infty_{\mathrm{comp}}(\mathbb{R}^d)$ (i.e. the set of test functions) and let $\mathcal{D}'(\mathbb{R}^d)$ denote the set of linear functionals on $\mathcal{D}(\mathbb{R}^d)$ (i.e. the set of distributions).

Theorem 7.12. (Schwartz kernel.) *Given a bounded, sequentially-continuous operator* $A:\mathcal{D}(\mathbb{R}^d)\to\mathcal{D}'(\mathbb{R}^d)$ *there exists a* Schwartz kernel $K_A\in\mathcal{D}'(\mathbb{R}^d\times\mathbb{R}^d)$ *such that*

$$Av(x)=\int_{\mathbb{R}^d}K_A(x,y)v(y)\,\mathrm{d}y,$$

in the sense of distributions.

References for the proof. See, e.g., [78, Theorem 5.2.1]. □

7.8 Compactly supported and properly supported operators

Definition 7.13. (Compactly supported and properly supported operators.)
A is compactly supported *if its Schwartz kernel* K_A *is compactly supported.*
A is properly supported *if, for any compact* $X,Y\subset\mathbb{R}^d$*,*

$$\big\{(x,y)\in\operatorname{supp}K_A\,:\,x\in X\big\} \text{ is compact} \tag{7.17a}$$

and

$$\big\{(x,y)\in\operatorname{supp}K_A\,:\,y\in Y\big\} \text{ is compact.} \tag{7.17b}$$

The name "properly supported" comes from the fact that the conditions (7.17) can be written as the x and y projections being proper maps from $\operatorname{supp} K_A$ to X and Y, where a *proper map* is one such that the preimage of any compact set is itself compact; see [52, Page 482].

Lemma 7.14. (Definitions of compactly and properly supported in terms of cut-off functions.)
(i) A is compactly supported iff there exist $\chi_1, \chi_2 \in \mathcal{D}$ such that $A = \chi_1 A \chi_2$.
(ii) A is properly supported iff for any $\chi \in \mathcal{D}$ there exist $\chi_1, \chi_2 \in \mathcal{D}$ such that

$$\chi A = \chi A \chi_1, \qquad A\chi = \chi_2 A \chi.$$

Proof. This is Exercise 6 in §7.11. □

The following lemma collects properties of compactly- and properly-supported operators that can all be proved in a straightforward way using either the Definition 7.13 or Lemma 7.14.

Lemma 7.15. (Key properties of compactly- and properly-supported operators.)
(i) If $K_A(x,y) = K(x-y)L(x,y)$, then A is properly supported if K has compact support.
(ii) Any differential operator is properly supported.
(iii) The operation of multiplication by $\phi \in \mathcal{D}(\mathbb{R}^d)$ is compactly supported (since $K_A(x,y) = \phi(x)\delta(x-y)$).
(iv) The composition of two properly supported operators is properly supported.
(v) The composition of a compactly supported operator with a properly supported operator is compactly supported.

Lemma 7.16. *If $A \in \Psi^m_\hbar$, then there exists a properly-supported $\widetilde{A} \in \Psi^m_\hbar$ and $e \in \hbar^\infty S^{-\infty}$ such that $A = \widetilde{A} + \operatorname{Op}_\hbar(e) = \widetilde{A} + O(\hbar^\infty)_{\Psi^{-\infty}_\hbar}$. Furthermore*

$$\sigma_\hbar(\widetilde{A}) = \sigma_\hbar(A) \quad \text{and} \quad \operatorname{WF}_\hbar(\widetilde{A}) = \operatorname{WF}_\hbar(A). \tag{7.18}$$

Observe that, once we establish that $A = \widetilde{A} + \operatorname{Op}_\hbar(e)$, then the properties in (7.18) follow immediately from (7.11) and (7.16) respectively.

7.9 A restricted class of symbols

Definition 7.17. (Symbol class S^m_{phg}.) *$a \in S^m_{\mathrm{phg}}$ if $a \in S^m$ and there exist $a_j \in S^{m-j}$, independent of $\hbar$, such that, for all $N \in \mathbb{Z}^+$,*

$$a - \sum_{j=0}^{N-1} \hbar^j a_j \in \hbar^N S^{m-N}. \tag{7.19}$$

If $A = \operatorname{Op}_\hbar(a)$ for $a \in S^m_{\mathrm{phg}}$, we write $A \in \Psi^m_{\mathrm{phg}}$.

If $a \in S^m$ satisfies (7.19) for $a_j \in S^{m-j}$, we write $a \sim \sum_{j=0}^\infty \hbar^j a_j$.

Remark 7.18. (Why "phg"?) *The subscript "phg" stands for "polyhomogeneous", since the symbols in Definition 7.17 are similar to the class of* semiclassical polyhomogeneous symbols *defined in [52, Definition E.3]; our class is actually slightly simpler than than in [52, Definition E.3], since this latter class imposes conditions on the behaviour of the a_js as $|\xi| \to \infty$, and we don't need these conditions for the results in these notes.*

Example 7.19. *(i) if $a \in S^m$ is independent of $\hbar$, then $a \in S^m_{\mathrm{phg}}$ with $a_0 = a$. Therefore, all the symbols in Example 7.1 are in S^m_{phg}.*
(ii) If $a_{\alpha j} \in C^\infty$ and $\partial^\gamma a_{\alpha j} \in L^\infty$ for all γ, α, and j, then

$$a(x,\xi) = \sum_{|\alpha| \le m} \left(\sum_{j=0}^{m-|\alpha|} \hbar^j a_{\alpha j}(x) \right) \xi^\alpha \in S^m_{\mathrm{phg}}.$$

When $d = 1$, this symbol is

$$a_{20}\xi^2 + \big(a_{10} + a_{11}\hbar)\xi + (a_{00} + a_{01}\hbar + a_{02}\hbar^2)$$

so

$$a_0 = a_{20}\xi^2 + a_{10}\xi + a_{00}, \quad a_1 = a_{11}\xi + a_{01}, \quad \text{and} \quad a_2 = a_{02}.$$

(iii) $a = \hbar\xi^2 \notin S^2_{\mathrm{phg}}$.

The big advantage of working in the class S^m_{phg} is that the principal symbols are *both* independent of $\hbar$ *and* can be understood as actual functions (as opposed to equivalence classes).

Corollary 7.20. *If $a \in S^m_{\mathrm{phg}}$ then*

$$\sigma_\hbar\big(\operatorname{Op}_\hbar(a)\big) = a_0. \tag{7.20}$$

Furthermore, if $A \in \Psi^m_{\mathrm{phg}}$ is such that $\sigma_\hbar(A) = 0$, then $A \in \hbar\Psi^{m-1}_{\mathrm{phg}}$.

Proof. This follows from Definitions 7.7 and Definition 7.17. □

Lemma 7.21. *Parts (i)-(iii) of Theorem 7.5 hold with $\Psi^m_\hbar$ replaced by Ψ^m_{phg}.*

Proof. This is proved in §9. □

Theorem 7.22. (Borel's theorem.) *Given $a_j \in S^{m-j}$, $j = 0, 1, \ldots$, there exists $a \in S^m$ such that $a \sim \sum_{j=0}^\infty \hbar^j a_j$ (in the sense of (7.19)).*

Proof. This is Exercise 8 in §7.11. □

Lemma 7.23. *Suppose $a \in S^m$ and $a_j \in S^{m-j}$ $j = 0, 1, \ldots$ are such that $a \sim \sum_{j=0}^\infty \hbar^j a_j$. If $a_j \in S^{m-j}_{\mathrm{phg}}$, then $a \in S^m_{\mathrm{phg}}$.*

Proof. This is Exercise 9 in §7.11. □

7.10 Ellipticity

We say that $B \in \Psi^m_\hbar$ is *elliptic* on $X \subset T^*\mathbb{R}^d$ if there exists $c > 0$ such that

$$\langle\xi\rangle^{-m}\big|\sigma_\hbar(B)(x,\xi)\big| \geq c \quad \text{for all } (x,\xi) \in X \text{ and for all } 0 < \hbar \leq \hbar_0. \tag{7.21}$$

A key feature of elliptic operators is that, up to a term in the residual class, they are invertible.

Theorem 7.24. (Elliptic parametrix.) *Let $A \in \Psi^{m_A}_{\mathrm{phg}}$ and $B \in \Psi^{m_B}_{\mathrm{phg}}$ be such that B is elliptic on $\mathrm{WF}_\hbar(A)$. Then there exist $Q_R, Q_L \in \Psi^{m_A - m_B}_{\mathrm{phg}}$ such that*

$$A = BQ_R + O(\hbar^\infty)_{\Psi^{-\infty}} = Q_L B + O(\hbar^\infty)_{\Psi^{-\infty}}. \tag{7.22}$$

(Compare to Remark 6.2.)

We have avoided defining the elliptic set of B, i.e., the points in phase space where $\langle\xi\rangle^{-m}|\sigma_\hbar(B)|$ is positive, to avoid dealing with the issues of uniformity of this property as either $|x|$ or $|\xi| \to \infty$; this issues can be dealt with by compactifying (i.e., introducing "the point at infinity"); see [52, §E.1.3] for this done in the ξ variable (so-called *fibre-radial compactification*) and [59] for this done in both x and ξ variables (so-called *fiber-radial and radial compactification*).

Theorem 7.25. (Elliptic estimate.) *Let $A \in \Psi^{m_A}_{\mathrm{phg}}$, $B_1 \in \Psi^{m_B}_{\mathrm{phg}}$, and $P \in \Psi^{m_P}_{\mathrm{phg}}$ be such that $B_1 P$ is elliptic on $\mathrm{WF}_\hbar(A)$.*

(i) Given $s, N, M > 0$, if $v \in \mathcal{D}'$ and $B_1 P v \in H^{s-m_B-m_P}$ then $Av \in H^{s-m_A}$ and there exists $C_s > 0$, $C_{N,M,s} > 0$ (independent of v and $\hbar$) such that, for all $0 < \hbar \leq \hbar_0$,

$$\|Av\|_{H^{s-m_A}_\hbar} \leq C_s \|B_1 P v\|_{H^{s-m_B-m_P}_\hbar} + C_{N,M,s}\,\hbar^M \|v\|_{H^{-N}_\hbar}. \tag{7.23}$$

(ii) If, in addition, A and B_1 are compactly supported and P is properly supported, then there exists $\widetilde{\chi} \in C^\infty_{\mathrm{comp}}$ (independent of v, M, N, and s) such that, for all $0 < \hbar \leq \hbar_0$,

$$\|Av\|_{H^{s-m_A}_\hbar} \leq C_s \|B_1 P v\|_{H^{s-m_B-m_P}_\hbar} + C_{N,M,s}\,\hbar^M \|\widetilde{\chi} v\|_{H^{-N}_\hbar}. \tag{7.24}$$

Idea of the proof. For Part (i), use the elliptic parametrix. For Part (ii), follow the proof of Part (i), and then use the properties of compactly and properly supported operators to show that the $O(\hbar^\infty)_{\Psi^{-\infty}}$ remainder is compactly supported.

Proof. (i) Applying Theorem 7.24 with $B = B_1 P \in \Psi_\hbar^{m_B+m_P}$ and using the definition of the residual class (7.4), we have that there exists $Q_L \in \Psi_\hbar^{m_A-(m_B+m_P)}$ such that

$$\begin{aligned}\|Av\|_{H_\hbar^{s-m_A}} &\leq \|Q_L B_1 P v\|_{H_\hbar^{s-m_A}} + C_{N,M,s}\,\hbar^M \|v\|_{H_\hbar^{-N}} \\ &\leq \|Q_L\|_{H_\hbar^{s-m_B-m_P}\to H_\hbar^{s-m_A}} \|B_1 P v\|_{H_\hbar^{s-m_B-m_P}} + C_{N,M,s}\,\hbar^M \|v\|_{H_\hbar^{-N}}\,.\end{aligned}$$

(ii) By Lemma 7.16 and (7.22), we can assume that Q' is properly supported. By Theorem 7.24, $A - Q_L B_1 P = E$ with $E = O(\hbar^\infty)_{\Psi^{-\infty}}$. Then, since P and Q_L are properly supported, and A and B_1 are compactly supported, Parts (iv) and (v) of Lemma 7.15 imply that $A - Q_L B_1 P = E$ is compactly supported; therefore there exists χ such that $E = E\chi$. The proof of (7.24) then follows in a similar way to the proof of (7.23) in Part (i). □

Remark 7.26. (Why did we work in $S^m_{\rm phg}$ for the elliptic parametrix?) *We see below that the proof of Theorem 7.24 requires that* $\operatorname{supp}\sigma_\hbar(A) \subset \mathrm{WF}_\hbar(A)$. *This is true when $A \in \Psi^{m_A}_{\rm phg}$, since then $\sigma_\hbar(A) = a_0$ is independent of $\hbar$, but this inclusion need not be the case in general (with our definition of* $\mathrm{WF}_\hbar$*).*

For example, if $a = \exp(-1/\hbar)$, then $a \in S^0$, and $\sigma_\hbar(\mathrm{Op}_\hbar(a)) = \exp(-1/\hbar)$ mod $\hbar S^{-1}$. However, $\mathrm{WF}_\hbar(\mathrm{Op}_\hbar(a)) = \emptyset$. *Let $B = 0$; then B is elliptic on* $\mathrm{WF}_\hbar(\mathrm{Op}_\hbar(a)) = \emptyset$, *but the result of Theorem 7.24 does not hold since* $\exp(-1/\hbar) \neq O(\hbar^\infty)_{\Psi_\hbar^{-\infty}}$. *Our way of ensuring that* $\operatorname{supp}\sigma_\hbar(A) \subset \mathrm{WF}_\hbar(A)$ *(and thus excluding this example) is to work in $S^m_{\rm phg}$.*

Remark 7.27. (Summary of results to be proved in §9.)

- *Theorem 7.5,*
- *the composition properties* (7.9) *and* (7.14) *of, respectively, the principal symbol and operator wavefront set,*
- *the relation* (7.8) *between the principal symbol of the adjoint and the principal symbol of the operator,*
- *Lemma 7.21,*
- *Lemma 7.16, and*
- *Theorem 7.24.*

7.11 Exercises for §7

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

   (ii) Show that $\langle\xi\rangle^{-m} \in S^{-m}$ for $m \in \mathbb{Z}^+$.

   (iii) Show that if $\chi \in C^\infty_{\rm comp}(T^*\mathbb{R}^d)$, then $\chi \in S^{-N}$ for every $N \geq 1$.

2. Prove Lemma 7.3. Hint: start with the expression (7.3) and then use the definition of $\mathcal{S}(\mathbb{R}^d)$ (5.3).

3. Prove Lemma 7.6. Hint: given $s > 0, N \geq 1$, choose an appropriate $M \geq 1$, and use that $a \in \hbar^M S^{-M}$.

4. If $P_\hbar u := -\hbar^2 \nabla \cdot (A\nabla u) - nu$, show that

   (i) $P_\hbar$ is the quantisation of a symbol in $S^2_{\rm phg}$, and

   (ii) $\sigma_\hbar(P_\hbar) = (A\xi)\cdot\xi - n \in S^2$.

5. Prove that if $a(x,\xi)$ is independent of $\hbar$, then $\mathrm{WF}_\hbar(\mathrm{Op}_\hbar(a)) = \operatorname{supp} a$ (i.e., Lemma 7.10).

6. Prove Lemma 7.14.

7. Prove Theorem 7.22 via the following steps.

   (a) Let $\chi \in C^\infty_{\mathrm{comp}}(\mathbb{R})$ with $\chi \equiv 1$ on $[-1,1]$. Show that if $\{\lambda_j\}_{j=0}^\infty \subset \mathbb{R}$ with $\lambda_j \to \infty$, the sum
   $$a(x,\xi) := \sum_{j=0}^\infty \chi\left(\frac{\lambda_j \hbar}{\langle \xi\rangle}\right)\hbar^j a_j(x,\xi)$$
   converges.

   (b) Show that, given β and $\chi \in C^\infty_{\mathrm{comp}}(\mathbb{R})$, there exists $C_{\beta,\chi}$ such that
   $$\partial_\xi^\beta\left(\chi\left(\frac{\lambda_j\hbar}{\langle\xi\rangle}\right)\right) \le \frac{C_{\beta,\chi}}{\lambda_j\hbar}\langle\xi\rangle^{1-|\beta|}. \tag{7.25}$$

   (c) Show that there is an increasing sequence $\{\lambda_j\}_{j=0}^\infty$ with $\lambda_j \to \infty$ such that for any multiindices $\alpha,\beta \in \mathbb{N}^d$ with $|\alpha|+|\beta| \le j$,
   $$\left|\partial_x^\alpha\partial_\xi^\beta\left(\chi\left(\frac{\lambda_j\hbar}{\langle\xi\rangle}\right)a_j\right)\right| \le 2^{-j}\hbar^{-1}\langle\xi\rangle^{m-j-|\beta|+1}.$$

   (d) With the choice of λ_j from (c), show that for any $\alpha,\beta\in\mathbb{N}^d$ with $|\alpha|+|\beta|\le N$,
   $$\left|\partial_x^\alpha\partial_\xi^\beta\left(a(x,\xi) - \sum_{j=0}^N a_j(x,\xi)\right)\right| \le C_{\alpha\beta N}\hbar^N\langle\xi\rangle^{m-|\beta|-N}, \tag{7.26}$$
   and conclude that $a \sim \sum_j h^j a_j$.

8. Prove Lemma 7.23.

8 Proof of the bound on u_{H^2} (i.e., the end of the proof of Theorem 4.11)

8.1 Restatement of bounds on the solution operator in semiclassical notation

Let $P_\hbar u := -\hbar^2\nabla\cdot(A\nabla u) - nu$, so that the Helmholtz equation $k^{-2}\nabla\cdot(A\nabla u) + nu = -f$ is $P_\hbar u = f$.

Given $f \in L^2(\mathbb{R}^d)$ with $\operatorname{supp} f \subset B_R$, let $u \in H^1_{\mathrm{loc}}(\mathbb{R}^d)$ be the solution to $P_\hbar u = f$ satisfying the Sommerfeld radiation condition (1.4) (with $k = \hbar^{-1}$). The definition of C_{sol} (Definition 2.1) and (2.1) imply that

$$\|u\|_{H^1_\hbar(B_R)} \lesssim C_{\mathrm{sol}}(\hbar^{-1}, R)\,\|f\|_{L^2(B_R)} \qquad \text{for all } \hbar > 0; \tag{8.1}$$

the reason we have $\lesssim$ and not $\le$ is that $\|\cdot\|_{H^1_\hbar(B_R)}$ is not equal to $\|\cdot\|_{H^1_k(B_R)}$, only equivalent (with constant only depending on d).

The bound (4.13) on u_{H^2} is proved under the assumption that $C_{\mathrm{sol}}(k)$ is polynomially bounded for $k \in K \subset [k_0,\infty)$ (in the sense of Definition 4.8); see Theorem 4.11. This implies that there exists $M > 0$ such that, given $\psi \in C^\infty_{\mathrm{comp}}(\mathbb{R}^d)$, there exists $C > 0$ such that

$$\|\psi u\|_{L^2(\mathbb{R}^d)} \le C\hbar^{-M}\,\|f\|_{L^2(B_R)} \qquad \text{for all } \hbar \in H \subset (0,\hbar_0], \tag{8.2}$$

where $\hbar_0 := k_0^{-1}$ and $H := \{k^{-1} : k \in K\}$. The bound (8.2) also holds with $\|\psi u\|_{L^2}$ replaced by $\|\psi u\|_{H^1_\hbar}$, but we only need it in the form (8.2) for what follows.

By the definitions of u_{H^2} (4.27) and $\|\cdot\|_{H^2_\hbar}$ (5.6), it is sufficient to prove that, given $\hbar_0 > 0$, there exists $C'_{\text{split},H^2} > 0$ such that

$$\|\Pi_H w\|_{H^2_\hbar(\mathbb{R}^d)} \le C'_{\text{split},H^2} \|f\|_{L^2(B_R)} \qquad \text{for all } \hbar \in H \subset (0,\hbar_0], \tag{8.3}$$

where $w := \varphi u$, where $\varphi \in C^\infty_{\text{comp}}(\mathbb{R}^d,[0,1])$ is defined in §4.5 to be equal to one on B_{R+1} and vanish outside B_{R+2}. (The constant C_{split,H^2} in (4.13) will then equal C'_{split,H^2} multiplied by the d-dependent constant coming from the equivalence of $\|\cdot\|_{H^2_\hbar}$ and $\|\cdot\|_{H^2_k}$.)

8.2 The frequency cut-offs as semiclassical pseudodifferential opeators and choosing the parameter μ

8.2.1 The frequency cut-offs as semiclassical pseudodifferential operators.

With Π_L and Π_H defined by (4.25) and (4.26), the definition of $\text{Op}_\hbar$ (7.2), the change of variable $\zeta = \xi/\hbar$, and the relationship (5.1) between $\mathcal{F}$ and $\mathcal{F}_\hbar$ imply that

$$\Pi_L = \text{Op}_\hbar\left(\chi_\mu(|\cdot|^2)\right). \tag{8.4}$$

By Part (iii) of Exercise 1 in §7.11, $\Pi_L \in \Psi^{-\infty}_\hbar(\mathbb{R}^d) := \cup_{N\ge 0}\Psi^{-N}_\hbar(\mathbb{R}^d)$. Since

$$\Pi_H = I - \Pi_L = \text{Op}_\hbar\left(1-\chi(|\cdot|^2)\right) \tag{8.5}$$

and $I \in \Psi^0_\hbar(\mathbb{R}^d)$, $\Pi_H \in \Psi^0_\hbar(\mathbb{R}^d)$.

8.2.2 The operator wavefront set of Π_H

Since $1-\chi_\mu(|\xi|^2) = 0$ for $|\xi|^2 \le \mu$, Lemma 7.10 implies that

$$\text{WF}_\hbar(\Pi_H) = \left\{(x,\xi) \,:\, |\xi|^2 \ge \mu\right\}. \tag{8.6}$$

8.2.3 Choosing the parameter μ

Recall from Exercise 5 in §7.11 that

$$\sigma_\hbar(P_\hbar) = (A\xi)\cdot\xi - n. \tag{8.7}$$

Let $\mu_0 = \mu_0(A,n)$ be defined by

$$\mu_0(A,n) := \left(1 + \frac{2n_{\max}}{A_{\min}}\right). \tag{8.8}$$

Lemma 8.1. *If $\mu \ge \mu_0$, then $P_\hbar$ is elliptic on $\text{WF}_\hbar(\Pi_H)$.*

Proof. It is sufficient to prove that

$$\text{if} \quad |\xi|^2 \ge \mu_0 \quad \text{then} \quad \langle\xi\rangle^{-2}\sigma_\hbar(P_\hbar) \ge \frac{A_{\min}}{2} > 0; \tag{8.9}$$

i.e., $P_\hbar$ is elliptic on $\{|\xi|^2 \ge \mu_0\}$.

By (8.7),

$$\begin{aligned}\langle\xi\rangle^{-2}\sigma_\hbar(P_\hbar) \ge \frac{A_{\min}|\xi|^2 - n_{\max}}{1+|\xi|^2} &= A_{\min}\left(\frac{(1+|\xi|^2)/2 + (|\xi|^2-1)/2 - n_{\max}/A_{\min}}{1+|\xi|^2}\right)\\ &= \frac{A_{\min}}{2} + \left(\frac{A_{\min}}{2}\right)\left(\frac{|\xi|^2 - 1 - 2n_{\max}/A_{\min}}{1+|\xi|^2}\right),\end{aligned}$$

and (8.9) follows. □

8.3 Proof of (8.17) under the assumption that $C_{\mathrm{sol}}(k) \lesssim k$

This proof follows very closely the proof in §6 for the case when $A = I$ and $n = 1$, with the elliptic parametrix replacing the argument in Remark 6.2.

As highlighted in §4.7, the main idea behind the bound on u_{H^2} is that $P_\hbar$ is (semiclassically) elliptic on the "support" of Π_H (now understood as $\mathrm{WF}_\hbar(\Pi_H)$), provided that μ is sufficiently large. Throughout the rest of this section, we therefore assume that $\mu \geq \mu_0$, so that the result of Part (i) of Lemma 8.1 holds.

We seek to apply Part (i) of Theorem 7.25 with $A = \Pi_H$ (so $m_A = 0$), $B_1 = 1$ (so $m_B = 0$), and $P = P_\hbar$ (so $m_P = 2$); observe that these are quantisations of elements of $S^0_{\mathrm{phg}}, S^0_{\mathrm{phg}}$, and S^2_{phg}, respectively (for $P_\hbar$ this is shown in Exercise 5 in §7.11). By Part (i) of Lemma 8.1, $B_1 P$ is elliptic on $\mathrm{WF}_\hbar(A)$. We can therefore apply Part (i) of Theorem 7.25 with $s = 2$ and obtain that, given $N, N' > 0$,

$$\left\|\Pi_H w\right\|_{H^2_\hbar(\mathbb{R}^d)} \lesssim \|P_\hbar w\|_{L^2(\mathbb{R}^d)} + \hbar^{N'} \|w\|_{H^{-N}_\hbar(\mathbb{R}^d)}\,, \tag{8.10}$$

where the omitted constant in $\lesssim$ depends on N and N'. Since $P_\hbar u = f$,

$$P_\hbar w = [P_\hbar, \varphi]u + \varphi f,$$

where $[\cdot,\cdot]$ is the standard commutator defined by $[A_1, A_2] := A_1A_2 - A_2A_1$, so that (8.10) becomes

$$\left\|\Pi_H w\right\|_{H^2_\hbar(\mathbb{R}^d)} \lesssim \|[P_\hbar, \varphi]u\|_{L^2(\mathbb{R}^d)} + \|f\|_{L^2(\mathbb{R}^d)} + \hbar^{N'} \|w\|_{H^{-N}_\hbar(\mathbb{R}^d)}\,. \tag{8.11}$$

Direct calculation, using the fact that $\operatorname{supp}\varphi \subset B_{R+2}$, implies that

$$\|\nabla\cdot(A\nabla(\varphi u)) - \varphi\nabla\cdot(A\nabla u)\|_{L^2(\mathbb{R}^d)} \lesssim \|\nabla u\|_{L^2(B_{R+2})} + \|u\|_{L^2(B_{R+2})}\,,$$

where the omitted constant depends on A and φ; therefore,

$$\left\|[P_\hbar, \varphi]u\right\|_{L^2(\mathbb{R}^d)} \lesssim \hbar\,\|u\|_{H^1_\hbar(B_{R+2})}\,. \tag{8.12}$$

Combining (8.11) and (8.12), and recalling that $\operatorname{supp}\varphi \subset B_{R+2}$, we have

$$\left\|\Pi_H w\right\|_{H^2_\hbar(\mathbb{R}^d)} \lesssim \hbar\,\|u\|_{H^1_\hbar(B_{R+2})} + \|f\|_{L^2(B_R)} + \hbar^{N'}\,\|u\|_{H^{-N}_\hbar(B_{R+2})}\,.$$

Choosing $N = 0$ and $N' = 1$, and then using (8.1), we obtain

$$\left\|\Pi_H w\right\|_{H^2_\hbar(\mathbb{R}^d)} \lesssim \Big(1 + \hbar C_{\mathrm{sol}}(\hbar^{-1}, R+2)\Big)\,\|f\|_{L^2(B_R)}\,. \tag{8.13}$$

If $C_{\mathrm{sol}}(\hbar^{-1}) \lesssim \hbar^{-1}$, i.e., $C_{\mathrm{sol}}(k) \lesssim k$, then (8.13) implies (8.3). However, if $C_{\mathrm{sol}}(\hbar^{-1}) \gg 1$ (as occurs when C_{sol} is polynomially bounded in the sense of Definition 4.8 with $M > 1$) then (8.13) is a weaker bound than (8.17).

8.4 Proof of (4.13) under the assumption that $C_{\mathrm{sol}}(k)$ is polynomially bounded (i.e., the end of the proof of Theorem 4.11)

8.5 The ideas of the proof.

Inspecting the argument in §8.3, we see that the assumption that $C_{\mathrm{sol}}(k) \lesssim 1$ is needed to get a good bound on the commutator term $[P_\hbar, \varphi]u$.

Idea 1: remove the commutator term by, instead of using that

$$P_\hbar \text{ is elliptic on } \mathrm{WF}_\hbar(\Pi_H),$$

use that

$$P_\hbar \text{ is elliptic on } \mathrm{WF}_\hbar(\Pi_H\varphi) \subset \mathrm{WF}_\hbar(\Pi_H) \cap \mathrm{WF}_\hbar(\varphi)$$

(by (7.14)); i.e., we apply the elliptic estimate (7.23) to u and not $w := \varphi u$. This gives

$$\|\Pi_H\varphi u\|_{H^2_\hbar(\mathbb{R}^d)} \lesssim \|P_\hbar u\|_{L^2(\mathbb{R}^d)} + \hbar^{N'}\,\|u\|_{H^{-N}_\hbar(\mathbb{R}^d)}\,. \tag{8.14}$$

However, the issue now is that we only have control of χu (via (8.2)), but the remainder term in (8.14) is not compactly supported.

Idea 2: aiming to use the elliptic estimate (7.24), we introduce a compactly supported B_1 in front of $P_\hbar$: let $\psi \in \mathcal{D}$ be such that $\psi = 1$ on $\operatorname{supp}\varphi$ and use that

$$\psi P_\hbar \text{ is elliptic on } \mathrm{WF}_\hbar(\Pi_H\varphi) \subset \mathrm{WF}_\hbar(\Pi_H) \cap \mathrm{WF}_\hbar(\varphi).$$

Indeed, this follows since (by the definition of ψ) $\psi P_\hbar = P_\hbar$ on $\operatorname{supp}\varphi = \mathrm{WF}_\hbar(\varphi)$ (by Lemma 7.10) and $P_\hbar$ is elliptic on $\mathrm{WF}_\hbar(\Pi_H)$ (by Lemma 8.1).

We now want to apply (7.24) with $A = \Pi_H\varphi$, $B_1 = \psi$, and $P = P_\hbar$; observe that these are quantisations of elements of $S^0_{\mathrm{phg}}, S^0_{\mathrm{phg}}$, and S^2_{phg}, respectively. By Lemma 7.15, B_1 is compactly supported, and P is properly supported; however, for A to be compactly supported we need Π_H to be properly supported, which it isn't. We therefore use Lemma 7.16 to replace Π_H by a properly-supported operator, up to an $O(\hbar^\infty)_{\Psi_\hbar^{-\infty}}$ remainder, which is controlled by the assumption that C_{sol} is polynomially bounded; for simplicity, we work with this properly-supported frequency cut-off from the beginning. Finally, the (now compactly-supported) $O(\hbar^\infty)_{\Psi_\hbar^{-\infty}}$ remainder coming from the elliptic estimate is controlled using the assumption that C_{sol} is polynomially bounded.

8.5.1 The details of the proof

The following result is a direct corollary of Lemma 7.16

Corollary 8.2. (Properly-supported frequency cut-offs.) *There exist a properly-supported operator $\widetilde{\Pi}_H$ such that*

$$\Pi_H = \widetilde{\Pi}_H + E, \text{ with } E = O(\hbar^\infty)_{\Psi^{-\infty}}. \tag{8.15}$$

Furthemore,

$$\mathrm{WF}_\hbar(\Pi_H) = \mathrm{WF}_\hbar(\widetilde{\Pi}_H), \tag{8.16}$$

To prove (8.3) it is now sufficient to prove that there exists $C''_{\mathrm{split},H^2} > 0$ such that

$$\left\|\widetilde{\Pi}_H\varphi u\right\|_{H^2_\hbar(\mathbb{R}^d)} \leq C''_{\mathrm{split},H^2}\|f\|_{L^2(B_R)} \qquad \text{for all } \hbar \in H \subset (0,\hbar_0]. \tag{8.17}$$

Indeed, given (8.17), for any $N > 0$ there exists C_N (by the definition of $O(\hbar^\infty)_{\Psi^{-\infty}}$ (7.4)) so that

$$\|\Pi_H\varphi u\|_{H^2_\hbar(\mathbb{R}^d)} \leq \left\|\widetilde{\Pi}_H\varphi u\right\|_{H^2_\hbar(\mathbb{R}^d)} + \left\|E\varphi u\right\|_{H^2_\hbar(\mathbb{R}^d)} \leq C''_{\mathrm{split},H^2}\|f\|_{L^2(\mathbb{R}^d)} + C_N\hbar^N\|\varphi u\|_{L^2(\mathbb{R}^d)},$$

Using the bound on the solution operator (8.2) and taking $N = M + 1$, we obtain (8.3).

Given φ (defined in §4.5), let $\psi \in \mathcal{D}$ be such that $\psi \equiv 1$ on $\operatorname{supp}\varphi$.

Lemma 8.3. *If $\mu \geq \mu_0$, then $\psi P_\hbar$ is elliptic on $\mathrm{WF}_\hbar(\widetilde{\Pi}_H\varphi)$.*

Proof. By (7.14), Lemma 7.10, and (8.16),

$$\mathrm{WF}_\hbar(\widetilde{\Pi}_H\varphi) \subset \mathrm{WF}_\hbar(\widetilde{\Pi}_H) \cap \mathrm{WF}_\hbar(\varphi) = \mathrm{WF}_\hbar(\widetilde{\Pi}_H) \cap \operatorname{supp}\varphi = \mathrm{WF}_\hbar(\Pi_H) \cap \operatorname{supp}\varphi.$$

The result then follows by recalling that, by the definition of ψ, $\psi P_\hbar = P_\hbar$ on $\operatorname{supp}\varphi$ and, by Lemma 8.1, $P_\hbar$ is elliptic on $\mathrm{WF}_\hbar(\Pi_H)$. □

Let $A = \widetilde{\Pi}_H\varphi$, $B_1 = \psi$, and $P = P_\hbar$ (so $m_A = 0$, $m_B = 0$, and $m_P = 2$); observe that these are quantisations of elements of $S^0_{\mathrm{phg}}, S^0_{\mathrm{phg}}$, and S^2_{phg}, respectively. Lemma 8.3 implies that B_1P is elliptic on $\mathrm{WF}_\hbar(A)$. Furthermore, A and B_1 are compactly supported (by Parts (iii) and (v) of Lemma 7.15, and P is properly supported (by Part (ii) of Lemma 7.15)

Therefore, by Part (ii) of Theorem 7.25, there exists $\chi \in \mathcal{D}$ such that, given $N, N' > 0$,

$$\left\|\widetilde{\Pi}_H\varphi u\right\|_{H^2_\hbar(\mathbb{R}^d)} \lesssim \|\psi P_\hbar u\|_{L^2(\mathbb{R}^d)} + \hbar^{N'}\|\chi u\|_{H^{-N}_\hbar(\mathbb{R}^d)} = \|f\|_{L^2(B_R)} + \hbar^{N'}\|\chi u\|_{H^{-N}_\hbar(\mathbb{R}^d)},$$

since $P_\hbar u = f$ and $\operatorname{supp} f \subset B_R \subset \{\varphi \equiv 1\} \subset \{\psi \equiv 1\}$.

Choosing $N = 0$ and $N' = M + 1$, and then using (8.2), we obtain (8.17) and the proof is complete.

8.6 Bibliographical remarks

9 The remaining proofs of the results in §7

Recall from Remark 7.27 that we need to prove the following.

- Theorem 7.5 and Lemma 7.21 (with the latter the specialisation of the former from S^m to S^m_{phg}),
- the composition properties (7.9) and (7.14) of, respectively, the principal symbol and operator wavefront set,
- the relation (7.8) between the principal symbol of the adjoint and the principal symbol of the operator,
- Lemma 7.16, and
- Theorem 7.24.

We prove these roughly in reverse order, since the proof of Theorem 7.5 is by far the most technical. (Note that we avoid circular reasoning – whereas the proof of, e.g., Theorem 7.24 uses most of the earlier results in §7, the proofs of these earlier results do not use Theorem 7.24.)

When the proofs can be summarised in a short "idea", we do so between the statement of the result and the proof. For longer proofs (where such "idea" summaries can become unwieldy) we split the proofs into steps (with the "idea" then formed by combining the instructions for each step).

9.1 Proof of Theorem 7.24 (the elliptic parametrix)

We give the proof of the existence of Q_L; the proof of existence of Q_R is very similar.

Step 1: Define the principal symbol of Q_L by "dividing" A by B at the level of principal symbols. Let

$$q_0 := \sigma_\hbar(A)/\sigma_\hbar(B). \tag{9.1}$$

Observe that $\operatorname{supp}\sigma(A) \subset \mathrm{WF}_\hbar(A)$ (since the former is independent of $\hbar$). Since B is elliptic on $\mathrm{WF}_\hbar(A)$, $|\sigma(B)| \geq c > 0$ on $\operatorname{supp}\sigma(A)$. Then, by Exercise 1 in §9.6, $q_0 \in S^{m_A - m_B}$. Since $\sigma_\hbar(A)$ and $\sigma_\hbar(B)$ are independent of $\hbar$, so is q_0, and thus $q_0 \in S^{m_A-m_B}_{\mathrm{phg}}$. Observe further that $\operatorname{supp}(q_0) \subset \operatorname{supp}\sigma(A) \subset \mathrm{WF}_\hbar(A)$. Let $Q_0 := \mathrm{Op}_\hbar(q_0)$; by (7.9),

$$\sigma^{m_A}_\hbar(Q_0 B - A) = q_0 \sigma^{m_B}_\hbar(B) - \sigma^{m_A}_\hbar(A) = 0$$

Therefore, by the fact that $Q_0 B - A \in \Psi^{m_A}_{\mathrm{phg}}$ and Corollary 7.20, there exists $R_1 \in \Psi^{m_A-1}_{\mathrm{phg}}$ such that $Q_0 B - A = \hbar R_1$.

Step 2: "Divide" the remainder by B. Suppose we have found $q_i \in S^{m_A-m_B-i}_{\mathrm{phg}}$, $i = 0, 1, \ldots, N-1$, such that $\operatorname{supp} q_i \in \mathrm{WF}_\hbar(A)$ and

$$Q_{N-1} := \sum_{j=0}^{N-1} \hbar^j \,\mathrm{Op}_\hbar(q_j) \in \Psi^{m_A-m_B}_{\mathrm{phg}} \tag{9.2}$$

is such that there exists $R_N \in \Psi^{m_A-N}_{\mathrm{phg}}$ such that

$$Q_{N-1} B - A = \hbar^N R_N. \tag{9.3}$$

We now construct a $q_N \in S^{m_A-m_B-N}_{\mathrm{phg}}$ with $\operatorname{supp} q_i \in \mathrm{WF}_\hbar(A)$ such that (9.3) holds with N replaced by $N+1$. By (9.3),

$$\mathrm{WF}_\hbar(R_N) = \mathrm{WF}_\hbar\left(\hbar^{-N}(Q_N B - A)\right) \subset \mathrm{WF}_\hbar(A),$$

where we have used (i) the fact that $\mathrm{WF}_\hbar(\hbar^{-N}C) = \mathrm{WF}_\hbar(C)$ by (7.12), (ii) the union and composition and properties (7.13) and (7.14), (iii) the support property (7.15), and (iv) the fact that $\operatorname{supp} q_i \subset \mathrm{WF}_\hbar(A)$. Since B is elliptic on $\mathrm{WF}_\hbar(A)$, B is elliptic on $\mathrm{WF}_\hbar(R_N)$ and thus, similar to above (i.e., using Exercise 2 in §9.6 and the fact that principal symbols in S^m_{phg} are independent of $\hbar$)

$$q_N := -\frac{\sigma_\hbar(R_N)}{\sigma_\hbar(B)} \in S^{m_A-N-m_B}_{\mathrm{phg}}. \tag{9.4}$$

Therefore,

$$\big(Q_{N-1} + \hbar^N \operatorname{Op}_\hbar(q_N)\big)B - A = \hbar^N\big(R_N + \operatorname{Op}_\hbar(q_N)B\big) \tag{9.5}$$

but

$$\sigma_\hbar^{m_A-N}(R_N + \operatorname{Op}_\hbar(q_N)B) = 0$$

by the definition of q_N, so that (by Corollary 7.20 again) there exists $R_{N+1} \in \Psi^{m_A-N-1}_{\mathrm{phg}}$ such that $R_N + \operatorname{Op}_\hbar(q_N)B = \hbar R_{N+1}$. Using this in (9.5), we have proved that (9.3) holds with N replaced by $N+1$.

Step 3: Define Q_L via Borel's theorem. We have therefore shown that there exist $q_i \in S^{m_A-m_B-i}_{\mathrm{phg}}$, $i = 0, 1, \ldots$, such that, with Q_{N-1} defined by (9.2), (9.3) holds for any N. By Theorem 7.22 there exists $q \in S^{m_A-m_B}$ such that $q \sim \sum_{j=0}^\infty \hbar^j q_j$; since $q_i \in S^{m_A-m_B-i}_{\mathrm{phg}}$, $q \in S^{m_A-m_B}_{\mathrm{phg}}$ by Lemma 7.23. The proof is then completed by setting $Q_L := \operatorname{Op}_\hbar(q)$.

9.1.1 Two natural questions regarding the elliptic parametrix

Can one easily write down q_L (i.e., the symbol of Q_L) explicitly (or at least an expansion of it)? Unfortunately no, and this is shown by the formula for symbol of the composition of two pseudodifferential operators in Theorem 9.4 below. Indeed, in the elliptic parametrix proof, we are essentially seeking $q_L \in S^{m_A-m_B}$ such that

$$q_L \# b - a \in \hbar^\infty S^{-\infty}, \tag{9.6}$$

where $q_L \# b$ is defined by (9.32). The expansion (9.34) shows that, given a and b, it is difficult to write down an explicit expansion for q_L satisfying (9.6).

Does anything simplify in our use of the elliptic parametrix in §8.4 to prove (4.13)? Unfortunately also no. Indeed, here $A = \operatorname{Op}_\hbar(a)$ and $B = \operatorname{Op}_\hbar(b)$ with

$$a(x,\xi) := \big(1 - \chi_\mu(|\xi|^2)\varphi(x) \quad \text{and} \quad b(x,\xi) := \psi(x)\Big((A(x)\xi)\cdot\xi - n(x) - \mathrm{i}\hbar\xi_\ell\partial_j A_{j\ell}(x)\Big).$$

The symbol q_0 is defined by (9.1), and the symbol q_1 is defined in terms of R_1 by (9.4). Suppose we want to write down an expansion of $R_1 := \hbar^{-1}\operatorname{Op}_\hbar(q_0\#b - a)$. In our case,

$$q_0\#b(x,\xi) = \left(\frac{(1-\chi_\mu(|\xi|^2)\varphi(x)}{\psi(x)\Big((A(x)\xi)\cdot\xi - n(x)\Big)}\right)\#\Big(\psi(x)\Big((A(x)\xi)\cdot\xi - n(x) - \mathrm{i}\hbar\xi_\ell\partial_j A_{j\ell}(x)\Big)\Big);$$

the expansion of this symbol has infinitely-many terms, since (without further assumptions on χ_μ, ψ, and n) all derivatives of the first argument in ξ are non-zero and all derivatives of the second argument in x are non-zero.

9.2 Proof of Lemma 7.16 (any pseudo is the sum of a properly supported pseudo and an $O(\hbar^\infty)_{\Psi^{-\infty}_\hbar}$ operator)

Idea of the proof. Write the Schwartz kernel $K_A(x,y)$ as

$$(1-\psi_0)(x-y)K_A(x,y) + \psi_0(x-y)K_A(x,y),$$

where $\psi \in C^\infty(\mathbb{R}^d)$ with $\psi_0 = 0$ on B_1 and $\psi_0 = 1$ outside B_2. By Part (i) of Lemma 7.15, the operator with Schwartz kernel $(1-\psi_0)(x-y)K_A(x,y)$ is properly supported. We then show, by repeated integration by parts, that the symbol of the operator with Schwartz kernel $\psi_0(x-y)K_A(x,y)$ is in $\hbar^\infty S^{-\infty}$, and thus the operator is in $O(\hbar^\infty)_{\Psi_\hbar^{-\infty}}$. [5]

Details of the proof. By the definition (7.2) of $\mathrm{Op}_\hbar$,

$$Av(x) = \int_{\mathbb{R}^d} K_A(x,y)v(y)\,\mathrm{d}y \tag{9.7}$$

where

$$K_A(x,y) := \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d} \exp\big(\mathrm{i}(x-y)\cdot\xi/\hbar\big)a(x,\xi)\,\mathrm{d}\xi. \tag{9.8}$$

Let

$$\widetilde{A}v(x) := \int_{\mathbb{R}^d} (1-\psi_0)(x-y)K_A(x,y)v(y)\,\mathrm{d}y, \tag{9.9}$$

where $\psi_0 \in C^\infty(\mathbb{R}^d)$ with $\psi_0 = 0$ on B_1 and $\psi_0 = 1$ outside B_2; by Part (i) of Lemma 7.15, $\widetilde{A}$ has proper support. Then $A = \widetilde{A} + E$ where

$$Ev(x) := \int_{\mathbb{R}^d} \psi_0(x-y)K_A(x,y)v(y)\,\mathrm{d}y. \tag{9.10}$$

We now claim that $E = \mathrm{Op}_\hbar(e)$ with

$$e(x,\xi) = \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \exp\big(\mathrm{i}(x-y)\cdot(\zeta-\xi)/\hbar\big)\,a(x,\zeta)\psi_0(x-y)\,\mathrm{d}y\,\mathrm{d}\zeta \tag{9.11}$$

(note that if $a(x,\zeta)$ does not decay sufficiently fast as $|\zeta|\to\infty$, then the ζ integral does not converge, and this oscillatory integral is understood in a distributional sense; see, e.g., [155, §3.6]). Indeed, by the definition (7.2) of $\mathrm{Op}_\hbar$, $e(x,\xi)$ is such that

$$\begin{aligned}\frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d} \exp\big(\mathrm{i}(x-y)\cdot\eta/\hbar\big)\,e(x,\eta)\,\mathrm{d}\eta &= \psi_0(x-y)K_A(x,y)\\ &= \psi_0(x-y)\frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d} \exp\big(\mathrm{i}(x-y)\cdot\zeta/\hbar\big)\,a(x,\zeta)\,\mathrm{d}\zeta;\end{aligned}$$

multiplying both sides by $\exp(-\mathrm{i}(x-y)\cdot\xi/\hbar)$ and then integrating in y (using that $\mathcal{F}_\hbar^{-1}(1) = \delta$) gives (9.11).

We need to show that $e \in \hbar^\infty S^{-\infty}$, i.e., for any multiindices α,β and for any N there exists $C_{\alpha,\beta,N} > 0$ so that

$$|\partial_x^\alpha\partial_\xi^\beta e(x,\xi)| \le C_{\alpha,\beta,N}\,\hbar^N\,\langle\xi\rangle^{-N}; \tag{9.12}$$

observe that it is sufficient to prove this for N sufficiently large (i.e., that there exists $N_0 > 0$ such that (9.12) holds for all $N \ge N_0$); we show that (9.12) holds for $N > d$.

From (9.11)

$$\partial_x^\alpha\partial_\xi^\beta e(x,\xi) = \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}(x-y)\cdot(\zeta-\xi)/\hbar}\left(\frac{\mathrm{i}(\zeta-\xi)}{\hbar}\right)^\alpha\left(\frac{\mathrm{i}(y-x)}{\hbar}\right)^\beta \psi_0(x-y)\,a(x,\zeta)\,\mathrm{d}\zeta\,\mathrm{d}y \tag{9.13}$$

(where again the integral is understood in a distributional sense).

[5] This latter property is well known; see, e.g., [155, Page 204, Step 2], [155, Theorem 9.6(iii)], and, e.g., [145, Chapter 7, Proposition 2.1], [81, Theorem 6.1.2] for related homogeneous – as opposed to semiclassical – results proved in the same way.

Overview of the rest of the proof. The plan is to now integrate by parts in (9.11), using the multidimensional version of $\exp(\mathrm{i}\lambda t) = (\mathrm{i}\lambda)^{-1}\partial_t(\exp(\mathrm{i}\lambda t))$, with t "equal" to either y or ζ. We first integrate by parts in ζ; this obtains

- positive powers of $\hbar$,
- negative powers of $|x-y|$, which make the y integral converge (note that there is not a problem at $x = y$ since $|x-y|$ is bounded away from zero on the support of ψ_0), and
- negative powers of $\langle\zeta\rangle$ (via the differentiation of $a(x,\zeta) \in S^m$), which make the ζ integral converge.

We then integrate by parts in y to obtain negative powers of $\langle\zeta-\xi\rangle$. Peetre's inequality (proved below) implies that

$$\frac{1}{\langle\zeta-\xi\rangle\langle\zeta\rangle} \le \frac{2}{\langle\xi\rangle}, \tag{9.14}$$

and thus negative powers of $\langle\zeta\rangle$ and $\langle\zeta-\xi\rangle$ can be converted into negative powers of $\langle\xi\rangle$ to show that (9.12) is satisfied.

Integration by parts in ζ. With $D = -\mathrm{i}\partial$, let

$$L := \hbar\frac{(x-y)}{|x-y|^2}\cdot D_\zeta \quad \text{so that} \quad L\big(\mathrm{e}^{\mathrm{i}(x-y)\cdot(\zeta-\xi)/\hbar}\big) = \mathrm{e}^{\mathrm{i}(x-y)\cdot(\zeta-\xi)/\hbar}.$$

Note that L is well defined on the support of the integrand of (9.13) since ψ_0 is supported away from zero, and thus $|x-y|$ is bounded below on the support of the integrand. Then, for any $M \in \mathbb{Z}^+$,

$$\begin{aligned}
&\partial_x^\alpha\partial_\xi^\beta e(x,\xi)\\
&= \frac{1}{(2\pi\hbar)^d}\left(\frac{\mathrm{i}}{\hbar}\right)^{|\alpha|+|\beta|}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}(x-y)\cdot(\zeta-\xi)/\hbar}\,\big(-L\big)^M\Big((\zeta-\xi)^\alpha(y-x)^\beta\psi_0(x-y)\,a(x,\zeta)\Big)\,\mathrm{d}\zeta\,\mathrm{d}y\\
&= \frac{\hbar^{M-d-|\alpha|-|\beta|}\,(\mathrm{i})^{|\alpha|+|\beta|}}{(2\pi)^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}(x-y)\cdot(\zeta-\xi)/\hbar}\,(y-x)^\beta\psi_0(x-y)\\
&\qquad\qquad\times\left(-\frac{(x-y)}{|x-y|^2}\cdot D_\zeta\right)^M\Big((\zeta-\xi)^\alpha a(x,\zeta)\Big)\,\mathrm{d}\zeta\,\mathrm{d}y.
\end{aligned}\tag{9.15}$$

Now

$$\left(-\frac{(x-y)}{|x-y|^2}\cdot D_\zeta\right)^M\Big((\zeta-\xi)^\alpha\,a(x,\zeta)\Big) = f(x-y)\,\widetilde{a}(x,\zeta), \tag{9.16}$$

where

$$|f(x-y)| \lesssim |x-y|^{-M} \quad \text{and} \quad |\partial_y^\gamma f(x-y)| \lesssim |x-y|^{-M-\gamma} \tag{9.17}$$

and, using the fact that $a \in S^m$,

$$\big|\widetilde{a}(x,\zeta)\big| \lesssim c_0|\zeta-\xi|^{|\alpha|}\langle\zeta\rangle^{m-M} + \ldots + c_{|\alpha|}\langle\zeta\rangle^{m-M+|\alpha|}, \tag{9.18}$$

where $c_j = c_j(m,\alpha)$ can be expressed in terms of multinomial coefficients

Integration by parts in y. Now let

$$\widetilde{L} := \frac{1-\hbar(\zeta-\xi)\cdot D_y}{1+|\zeta-\xi|^2} \quad \text{so that} \quad \widetilde{L}\big(\mathrm{e}^{\mathrm{i}(x-y)\cdot(\zeta-\xi)/\hbar}\big) = \mathrm{e}^{\mathrm{i}(x-y)\cdot(\zeta-\xi)/\hbar}.$$

Therefore, using (9.15) and (9.16), we find that, for any $M' \in \mathbb{Z}^+$,

$$\partial_x^\alpha\partial_\xi^\beta e(x,\xi) = \frac{\hbar^{M-d-|\alpha|-|\beta|}\,(\mathrm{i})^{|\alpha|+|\beta|}}{(2\pi)^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}(x-y)\cdot(\zeta-\xi)/\hbar}\,(y-x)^\beta\psi_0(x-y)f(x-y)\widetilde{a}(x,\zeta)\,\mathrm{d}\zeta\,\mathrm{d}y$$

$$= \frac{\hbar^{M-d-|\alpha|-|\beta|}\,(\mathrm{i})^{|\alpha|+|\beta|}}{(2\pi)^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}(x-y)\cdot(\zeta-\xi)/\hbar}\,\widetilde{a}(x,\zeta) \left(\frac{1+\hbar(\zeta-\xi)\cdot D_y}{1+|\zeta-\xi|^2}\right)^{M'} \Big((y-x)^\beta \psi_0(x-y) f(x-y)\Big)\,\mathrm{d}\zeta\,\mathrm{d}y. \tag{9.19}$$

Therefore, using in this last expression the bounds (9.17) and (9.18), we obtain that

$$\begin{aligned}\left|\partial_x^\alpha \partial_\xi^\beta e(x,\xi)\right| &\lesssim \hbar^{M-d-|\alpha|-|\beta|}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \frac{c_0|\zeta-\xi|^{|\alpha|}\langle\zeta\rangle^{m-M}+\ldots+c_\alpha\langle\zeta\rangle^{m-M+|\alpha|}}{\langle\zeta-\xi\rangle^{2M'}\,|x-y|^{M-|\beta|}}\,\mathrm{d}\zeta\,\mathrm{d}y \\ &\lesssim \hbar^{M-d-|\alpha|-|\beta|}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \frac{1}{\langle\zeta\rangle^{M-m-|\alpha|}\,\langle\zeta-\xi\rangle^{2M'-|\alpha|}\,|x-y|^{M-|\beta|}}\,\mathrm{d}\zeta\,\mathrm{d}y, \end{aligned} \tag{9.20}$$

where we have used (i) that $|\zeta-\xi|\le\langle\zeta-\xi\rangle$, and (ii) the fact that the least decay in $|x-y|$ occurs when all the derivatives in y in (9.19) fall on ψ_0).

We now choose M so that

$$M \ge N+d+|\alpha|+|\beta|; \tag{9.21}$$

this ensures that we obtain $\hbar^N$ in (9.12); in addition, since $M>d+|\beta|$, the integral in y converges. Therefore, to prove (9.12), we only need to bound the terms in the integrand in (9.20) that depend on ζ and ξ by

$$\frac{1}{\langle\xi\rangle^N\langle\zeta\rangle^N}; \tag{9.22}$$

indeed, this bound obtains the $\langle\xi\rangle^{-N}$ in (9.12), and, when $N>d$, this bound ensures convergence of the ζ integral.

Peetre's inequality is usually stated in the form that, for all $x,y\in\mathbb{R}^d$,

$$\frac{\langle x\rangle}{\langle y\rangle}\le\sqrt{2}\langle x-y\rangle; \tag{9.23}$$

[6] letting $x=\xi$ and $y=\zeta$ we see that (9.14) holds. Thus, to bound the terms in the integrand in (9.20) that depend on ζ and ξ by (9.22), it is sufficient to bound them by

$$\frac{1}{\langle\zeta-\xi\rangle^N\langle\zeta\rangle^{2N}}; \tag{9.24}$$

we therefore choose

$$M'\ge N/2+|\alpha|/2 \quad\text{and}\quad M\ge 2N+m+|\alpha|. \tag{9.25}$$

Therefore, to ensure (9.21) and (9.25), we choose

$$M' := N/2+|\alpha|/2 \quad\text{and}\quad M:=\max\Big\{2N+m+|\alpha|\,,\,N+d+|\alpha|+|\beta|\Big\},$$

and the proof of (9.12) is complete.

9.3 Proof of Part (iv) of Theorem 7.5 ($H^s_\hbar\to H^{s-m}_\hbar$ boundedness of elements of $\Psi^m_\hbar$)

Lemma 9.1. (The Schur test for boundedness.) *If A has Schwartz kernel K_A,*

$$\sup_{x\in\mathbb{R}^d}\int_{\mathbb{R}^d}|K_A(x,y)|\,\mathrm{d}y\le C_1, \qquad \textit{and} \qquad \sup_{y\in\mathbb{R}^d}\int_{\mathbb{R}^d}|K_A(x,y)|\,\mathrm{d}x\le C_2,$$

then

$$\|A\|_{L^2(\mathbb{R}^d)\to L^2(\mathbb{R}^d)}\le\sqrt{C_1C_2}.$$

[6] This follows from

$$1+|x|^2=1+|(x-y)+y|^2=1+|x-y|^2+|y|^2+2(x-y)\cdot y\le 1+2|x-y|^2+2|y|^2\le 2(1+|x-y|^2)(1+|y|^2).$$

Idea of the proof. Starting from the definition of $|Au(x)|^2$, use the Cauchy–Schwarz inequality.

Proof. By the Cauchy–Schwarz inequality,

$$\begin{aligned}|Au(x)|^2 \le \Big(\int_{\mathbb{R}^d} |K_A(x,y)u(y)|\mathrm{d}y\Big)^2 &= \Big(\int_{\mathbb{R}^d} \big|K_A(x,y)\big|^{1/2}|u(y)|\,\big|K_A(x,y)\big|^{1/2}\mathrm{d}y\Big)^2 \\ &\le \Big(\int_{\mathbb{R}^d} \big|K_A(x,y)\big||u(y)|^2\mathrm{d}y\Big)\Big(\int_{\mathbb{R}^d} |K_A(x,y)|\mathrm{d}y\Big) \\ &\le C_1 \int_{\mathbb{R}^d} \big|K_A(x,y)\big||u(y)|^2\mathrm{d}y.\end{aligned}$$

Therefore,

$$\|Au\|^2_{L^2(\mathbb{R}^d)} \le C_1 \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \big|K_A(x,y)\big||u(y)|^2\,\mathrm{d}y\right)\mathrm{d}x.$$

By Tonelli's theorem,

$$\|Au\|^2_{L^2(\mathbb{R}^d)} \le C_1 \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \big|K_A(x,y)\big|\,\mathrm{d}x\right)|u(y)|^2\,\mathrm{d}y \le C_1C_2\|u\|^2_{L^2(\mathbb{R}^d)}.$$

□

We first prove L^2 boundness for $A \in \Psi^m_\hbar$ with $m < -d$, and then use "Hörmander's square root trick" (see [79, Proof of Theorem 18.1.11]) to convert this to L^2 boundedness for $A \in \Psi^0_\hbar$.

Lemma 9.2. *If $A \in \Psi^m_\hbar$ with $m < -d$, then $A : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ is uniformly bounded for $0 < \hbar \le \hbar_0$.*

Idea of the proof. By the Schur test, it is sufficient to bound both $\sup_x \int |K_A(x,y)|\,\mathrm{d}y$ and $\sup_y \int |K_A(x,y)|\,\mathrm{d}x$ independently of $\hbar$, where

$$K_A(x,y) = \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}(x-y)\cdot\xi/\hbar}a(x,\xi)\,\mathrm{d}\xi. \tag{9.26}$$

Since $m < -d$ and $a \in S^m$, the ξ integral converges absolutely ($\int_{\mathbb{R}^d}\langle\xi\rangle^m\mathrm{d}\xi < \infty$ when $m < -d$). Integrate by parts in ξ to bring down enough powers of $(1+\hbar^{-1}|x-y|)^{-1}$ to make the x and y integrals converge and get rid of the $\hbar^{-d}$.

Proof. We show that $\sup_x \int |K_A(x,y)|\,\mathrm{d}y$ is bounded independently of $\hbar$; the proof that $\sup_y \int |K_A(x,y)|\,\mathrm{d}x$ is bounded independently of $\hbar$ is very similar, and then the result follows from Lemma 9.1.

We now integrate by parts the expression (9.26) for the Schwartz kernel; a convenient way to do this is to let

$$L := \frac{1 + \frac{(x-y)}{|x-y|}\cdot D_\xi}{1+\hbar^{-1}|x-y|} \quad \text{so that} \quad L\big(\mathrm{e}^{\mathrm{i}(x-y)\cdot\xi/\hbar}\big) = \mathrm{e}^{\mathrm{i}(x-y)\cdot\xi/\hbar}. \tag{9.27}$$

Therefore, for any $N > 0$,

$$K_A(x,y) = \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d} L^N\big(\mathrm{e}^{\mathrm{i}(x-y)\cdot\xi/\hbar}\big)a(x,\xi)\,\mathrm{d}\xi = \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}(x-y)\cdot\xi/\hbar}(L^*)^N a(x,\xi)\,\mathrm{d}\xi.$$

By the definition of L and the fact that $a \in S^m$, given $N \in \mathbb{Z}^+$ there exists $C_N > 0$ such that

$$|(L^*)^N a(x,\xi)| \le C_N \frac{\langle\xi\rangle^m}{(1+\hbar^{-1}|x-y|)^N}.$$

Therefore, since $m < -d$,

$$\big|K_A(x,y)\big| \le \frac{C_N}{(2\pi\hbar)^d}\frac{1}{(1+\hbar^{-1}|x-y|)^N}\int_{\mathbb{R}^d}\langle\xi\rangle^m\,\mathrm{d}\xi \le \frac{C'_N}{(2\pi\hbar)^d}\frac{1}{(1+\hbar^{-1}|x-y|)^N}.$$

for some $C'_N > 0$. Therefore,

$$\int_{\mathbb{R}^d} \big|K_A(x,y)\big|\mathrm{d}y \leq \frac{C'_N}{(2\pi)^d}\hbar^{-d}\int_{\mathbb{R}^d}\frac{1}{(1+\hbar^{-1}|x-y|)^N}\mathrm{d}y = \frac{C'_N}{(2\pi)^d}\hbar^{-d}\hbar^d\int_{\mathbb{R}^d}\frac{1}{(1+|w|)^N}\mathrm{d}w,$$

where we have used the change of variables $y = x + \hbar w$. Choosing $N > d$ (so that the last integral is finite), we obtain that $\sup_x \int |K_A(x,y)|\,\mathrm{d}y \lesssim 1$ and the proof is complete. □

Remark. *Exercise 4 in §9.6 asks you to prove Lemma 9.2 using a less-sophisticated integration-by-parts operator L, albeit then splitting the integral to deal separately with the regions $|x-y| \leq \hbar$ and $|x-y| \geq \hbar$. The point of this exercise is to demonstrate that, even if one can't come up with a clever choice of L to deal with everything (i.e., $\hbar$ dependence and convergence) together, the results can still be obtained with simpler L.*

Lemma 9.3. *If $A \in \Psi^0_\hbar$, then $A : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ is uniformly bounded for $0 < \hbar \leq \hbar_0$. Moreover, given $\delta > 0$ and $\hbar_0 > 0$ there exists $C > 0$ such that, for all $0 < \hbar \leq \hbar_0$,*

$$\|A\|_{L^2(\mathbb{R}^d)\to L^2(\mathbb{R}^d)} \leq (1+\delta)\sup_{T^*\mathbb{R}^d}|\sigma_\hbar(A)| + C\hbar^{1/2}. \tag{9.28}$$

In fact, an improved bound holds with $\hbar^{1/2}$ replaced by $\hbar$; see [155, Theorem 13.13].

Proof.

Step 1: Use Parts (i) and (ii) of Theorem 7.5 and that $\|A\|^2 = \|A^*A\|$ to show that the result of Lemma 9.2 holds in fact for all $m < 0$.

We first prove that for all $m < 0$

$$\text{any } A \in \Psi^m_\hbar \text{ is bounded on } L^2 \text{ uniformly for } 0 < \hbar \leq \hbar_0. \tag{9.29}$$

By Lemma 9.2, (9.29) is true for $m < -d$. We now show that if (9.29) holds for $m = m_1$, then (9.29) holds for $m = m_1/2$; the fact that (9.29) then holds for all $m < 0$ follows. Let $m_1 < 0$ and suppose that (9.29) holds for $m = m_1$. Let $A \in \Psi^{m_1/2}_\hbar$; then, by Part (i) of Theorem 7.5, $A^* \in \Psi^{m_1}_\hbar$ and, by Part (ii) of Theorem 7.5, $A^*A \in \Psi^{m_1}_\hbar$. By assumption therefore, A^*A is bounded on L^2 uniformly for $0 < \hbar \leq \hbar_0$. Since $\|A\|^2 = \|A^*A\|$, $A : L^2 \to L^2$ is also bounded uniformly for $0 < \hbar \leq \hbar_0$.

Step 2: Let $M := (1+\delta)\sup_{T^*\mathbb{R}^d}|\sigma_\hbar(A)|$, let $B = \mathrm{Op}_\hbar(b)$ where

$$b(x,\xi) := \big(M^2 - |\sigma_\hbar(A)(x,\xi)|^2\big)^{1/2}, \tag{9.30}$$

and consider $\sigma_\hbar(B^*B - (M^2 - A^*A))$. We first claim that we can assume, without loss of generality, that $\sup_{T^*\mathbb{R}^d}|\sigma_\hbar(A)| > 0$. Indeed, if $\sup_{T^*\mathbb{R}^d}|\sigma_\hbar(A)| = 0$, then $\sigma_\hbar(A) = 0$ and thus $A \in \hbar\Psi^{-1}_\hbar$ (by (7.6)). Then, by Step 1, $\|A\|_{L^2\to L^2} \leq C'\hbar \leq C\hbar^{1/2}$ for $0 < \hbar \leq \hbar_0$.

Therefore,

$$M^2 - |\sigma_\hbar(A)|^2(x,\xi) \geq (2\delta+\delta^2)\sup_{T^*\mathbb{R}^d}|\sigma_\hbar(A)| > 0$$

for all (x,ξ); by Exercise 2 in §9.6, $b \in S^0$. Let $B = \mathrm{Op}_\hbar(b)$; by (7.8) and (7.9),

$$\sigma_\hbar\big(B^*B - (M^2 - A^*A)\big)\big) = b^2 - \big(M^2 - |\sigma_\hbar(A)|^2\big) = 0.$$

Therefore, by the definition of $\sigma_\hbar$ (Definition 7.7),

$$B^*B = M^2 - A^*A + \hbar R$$

for some $R \in \Psi^{-1}_\hbar$. Acting this last equality on u, and then pairing with u, we find that

$$\|Bu\|^2_{L^2(\mathbb{R}^d)} + \|Au\|^2_{L^2(\mathbb{R}^d)} = M^2\|u\|^2_{L^2(\mathbb{R}^d)} + \hbar\langle Ru, u\rangle;$$

and thus

$$\|Au\|^2_{L^2(\mathbb{R}^d)} \leq M^2\|u\|^2_{L^2(\mathbb{R}^d)} + \hbar\langle Ru, u\rangle.$$

Step 3: By Step 1, $R : L^2 \to L^2$ is uniformly bounded; hence get result.

Since $R \in \Psi_\hbar^{-1}$, by the fact that (9.29) holds for all $m < 0$, $R : L^2 \to L^2$ is uniformly bounded for $0 < \hbar \leq \hbar_0$; thus, given $\hbar_0 > 0$ there exists $C' > 0$ such that, for all $0 < \hbar \leq \hbar_0$,

$$\|Au\|^2_{L^2(\mathbb{R}^d)} \leq (M^2 + C'\hbar)\, \|u\|^2_{L^2(\mathbb{R}^d)}\,;$$

the result (9.28) follows with $C = \sqrt{C'}$ by recalling that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. □

Remark. *Bounds of the type (9.28), i.e., of the L^2 operator norm in terms of properties of the principal symbol, go back to [27]; see, e.g., the discussion in [82].*

We now prove Part (iv) of Theorem 7.5.

Idea of the proof. Prove that

$$\|u\|^2_{H^s_\hbar(\mathbb{R}^d)} = \|\mathrm{Op}_\hbar(\langle\xi\rangle^s)u\|^2_{L^2(\mathbb{R}^d)} \tag{9.31}$$

and then use the $L^2 \to L^2$ boundedness of elements of $\Psi^0_\hbar$ (proved in Lemma 9.3).

Proof of Part (iv) of Theorem 7.5. By (5.6), (5.5), and (7.3),

$$\|u\|^2_{H^s_\hbar(\mathbb{R}^d)} := (2\pi\hbar)^{-d}\, \|\langle\xi\rangle^s \mathcal{F}_\hbar u\|^2_{L^2(\mathbb{R}^d)} = \left\|\mathcal{F}^{-1}_\hbar\left(\langle\xi\rangle^s \mathcal{F}_\hbar u\right)\right\|^2_{L^2(\mathbb{R}^d)} = \|\mathrm{Op}_\hbar(\langle\xi\rangle^s)u\|^2_{L^2(\mathbb{R}^d)}\,,$$

and (9.31) follows. By (9.31),

$$\|Au\|_{H^{s-m}_\hbar} = \left\|\mathrm{Op}_\hbar(\langle\xi\rangle^{s-m})Au\right\|_{L^2(\mathbb{R}^d)} = \left\|\mathrm{Op}_\hbar(\langle\xi\rangle^{s-m})A\,\mathrm{Op}_\hbar(\langle\xi\rangle^{-s})\,\mathrm{Op}_\hbar(\langle\xi\rangle^s)u\right\|_{L^2(\mathbb{R}^d)}\,,$$

where we have used that

$$\mathrm{Op}_\hbar(\langle\xi\rangle^{-s})\,\mathrm{Op}_\hbar(\langle\xi\rangle^s) = \mathrm{Op}_\hbar(1) = I,$$

by Part (ii) of Theorem 5.4. Since $\langle\xi\rangle^{-m} \in S^{-m}$ (by Exercise 1 in §7.11), $\mathrm{Op}_\hbar(\langle\xi\rangle^{s-m})A\,\mathrm{Op}_\hbar(\langle\xi\rangle^{-s}) \in \Psi^0_\hbar$ by Part (ii) of Theorem 7.5. Then, by Lemma 9.3,

$$\|Au\|_{H^{s-m}_\hbar} \leq C\, \|\mathrm{Op}_\hbar(\langle\xi\rangle^s)u\|_{L^2(\mathbb{R}^d)} = C\, \|u\|_{H^s_\hbar(\mathbb{R}^d)}\,.$$

□

9.4 The proofs of Parts (ii) and (iii) of Theorem 7.5 and the composition properties

The basis of all these results is the following composition property for symbols.

Theorem 9.4. (Composition property for symbols.) *Given $a \in S^{m_A}$ and $b \in S^{m_B}$, let*

$$a\#b(x,\xi) := \frac{1}{(2\pi\hbar)^d} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \exp\left(-\mathrm{i}\widetilde{x}\cdot\widetilde{\xi}/\hbar\right) a(x,\xi+\widetilde{\xi})\, b(x+\widetilde{x},\xi)\, \mathrm{d}\widetilde{\xi}\, \mathrm{d}\widetilde{x}. \tag{9.32}$$

Then $a\#b \in S^{m_A+m_B}$,

$$\mathrm{Op}_\hbar(a)\,\mathrm{Op}_\hbar(b) = \mathrm{Op}_\hbar(a\#b), \tag{9.33}$$

and, for all $N \in \mathbb{Z}^+$,

$$a\#b(x,\xi) - \sum_{|\alpha|\leq N-1} \frac{(\mathrm{i}\hbar)^{|\alpha|}}{\alpha!}\left(D^\alpha_\xi a(x,\xi)\right)\left(D^\alpha_x b(x,\xi)\right) \in \hbar^N S^{m_A+m_B-N}. \tag{9.34}$$

Proofs of the composition properties (7.9) and (7.14) using Theorem 9.4. The composition property (7.9) of the principal symbol follows immediately from (9.34).

$$\sigma_\hbar(AB) = \sigma_\hbar(\mathrm{Op}_\hbar(a\#b)) = a(x,\xi)b(x,\xi) = \sigma_\hbar(A)\sigma_\hbar(B).$$

For the composition property (7.14) of the operator wavefront set, observe that, by (9.34), if *either* $\partial^\alpha_x\partial^\beta_\xi a(x,\xi)$ *or* $\partial^\alpha_x\partial^\beta_\xi b(x,\xi)$ is superalgebraically small in $\hbar$ for all α,β, then $\partial^\alpha_x\partial^\beta_\xi a\#b(x,\xi)$ is superalgebraically small in $\hbar$ for all α, β. The result (7.14) is equivalent to

$$(\mathrm{WF}_\hbar\, A)^c \cup (\mathrm{WF}_\hbar\, B)^c \subset (\mathrm{WF}_\hbar(AB))^c,$$

and thus follows from the definition of $\mathrm{WF}_\hbar$ (7.12). □

Proofs of Parts (ii) and (iii) of Theorem 7.5 and (7.10) using Theorem 9.4. Part (ii) follows immediately from the fact that $a\#b \in S^{m_A+m_B}$. Part (iii) follows since (9.34) implies that

$$\begin{aligned}\operatorname{Op}_\hbar(a)\operatorname{Op}_\hbar(b) - \operatorname{Op}_\hbar(b)\operatorname{Op}_\hbar(a) &= \operatorname{Op}_\hbar(a\#b) - \operatorname{Op}_\hbar(b\#a)\\ &= \operatorname{Op}_\hbar(ab) - \operatorname{Op}_\hbar(ba) + \hbar S^{m_A+m_B-1} = \hbar S^{m_A+m_B-1}.\end{aligned}$$

Furthermore,

$$\begin{aligned}&\operatorname{Op}_\hbar(a\#b) - \operatorname{Op}_\hbar(b\#a)\\ &\quad= \sum_{|\alpha|=1} \mathrm{i}\hbar\big(D_\xi^\alpha a(x,\xi)\big)\big(D_x^\alpha b(x,\xi)\big) - \sum_{|\alpha|=1} \mathrm{i}\hbar\big(D_x^\alpha a(x,\xi)\big)\big(D_\xi^\alpha b(x,\xi)\big) + \hbar^2 S^{m_A+m_B-2}\\ &\quad= -\sum_{|\alpha|=1} \mathrm{i}\hbar\big(\partial_\xi^\alpha a(x,\xi)\big)\big(\partial_x^\alpha b(x,\xi)\big) + \sum_{|\alpha|=1} \mathrm{i}\hbar\big(\partial_x^\alpha a(x,\xi)\big)\big(\partial_\xi^\alpha b(x,\xi)\big) + \hbar^2 S^{m_A+m_B-2}\\ &\quad= -\mathrm{i}\hbar\{a,b\} + \hbar^2 S^{m_A+m_B-2},\end{aligned}$$

by the definition of the Poisson bracket $\{\cdot,\cdot\}$ (7.7), so that (7.10) follows. □

The idea of the proof is to

- first consider the case when $a, b \in \mathscr{S}$,
- prove (9.33) by manipulating the definition of the left-hand side (with these manipulations straightforward to justify since $a, b \in \mathscr{S}$),
- observe that the integral defining $a\#b$ in (9.32) is an oscillatory integral (with a large parameter $\hbar^{-1}$), and use results about such integrals to prove that the appropriate analogue of (9.34) (see (9.43) below), and
- use the density of $\mathscr{S}$ in the symbol class $S(m)$ (defined in Definition 9.11 below) to prove the result for $a, b \in S(m_A m_B)$, and
- prove the result for $a \in S^{m_A}$ and $b \in S^{m_B}$ via the relationship between the classes S^m and $S(\langle\xi\rangle^m)$ encapsulated by Lemma 9.12 below.

We highlight here that a similar strategy is used in the proof of the results about the adjoint (Part (i) of Theorem 7.5 and (7.8)); see §9.5.

Since the plan above involves using results about asymptotics of oscillatory integrals, we collect these results first.

9.4.1 Asymptotics of oscillatory integrals and related results

Theorem 9.5. (Asymptotics of oscillatory integral with quadratic phase.) *Let Q be a non-singular, symmetric, real matrix. Let $\operatorname{sgn} Q$ be the number of positive eigenvalues of Q minus the number of negative eigenvalues. For $a \in \mathscr{S}$, let*

$$I(\hbar, a) := \int_{\mathbb{R}^d} \exp\left(\frac{\mathrm{i}\langle Qx, x\rangle}{2\hbar}\right) a(x)\,\mathrm{d}x.$$

Then, for all $N \in \mathbb{Z}^+$,

$$I(\hbar,a) = (2\pi\hbar)^{d/2}\frac{\exp\left(\mathrm{i}(\pi/4)\operatorname{sgn} Q\right)}{|\det Q|^{1/2}}\left[\sum_{j=0}^{N-1}\frac{\hbar^j}{j!}\left(\left(\frac{\langle Q^{-1}D, D\rangle}{2\mathrm{i}}\right)^j a\right)(0) + \frac{\hbar^N}{N!}R_N(\hbar,a)\right], \quad (9.35)$$

where

$$R_N(\hbar,a) := \frac{N}{(2\pi)^d}\int_0^1 (1-t)^{N-1}\Big(\exp\Big(-\mathrm{i}t\hbar\langle Q^{-1}D,D\rangle/2\Big)\left(-\frac{\mathrm{i}}{2}\langle Q^{-1}D,D\rangle\right)^N a\Big)(0)\mathrm{d}t \quad (9.36)$$

We make two remarks: (i) the operator $\langle Q^{-1}D, D\rangle^j$ in (9.35) and analogous operators in (9.36) are understood as Fourier multipliers; i.e.

$$\big(\langle Q^{-1}D, D\rangle^j a\big)(x) = \mathcal{F}^{-1}_{\xi\to x}\big(\langle Q^{-1}\xi,\xi\rangle^j \mathcal{F}_{x\to\xi} a\big)(x)$$

(compare to (5.9)), and (ii) compare the asymptotics (9.35) to the well-known stationary phase asymptotics in the 1-d case: if φ is real-valued, the only zero of φ' is at zero, and $\varphi''(0)\neq 0$, then

$$\int_{-\infty}^{\infty} \exp\Big(\mathrm{i}\varphi(x)/\hbar\Big)a(x)\,\mathrm{d}x = \exp\big(\mathrm{i}\varphi(0)\hbar\big)\exp\Big(\mathrm{i}(\pi/4)\operatorname{sgn}(Q''(0))\Big)\sqrt{\frac{2\pi\hbar}{|\varphi''(0)|}}\Big(a(0)+O(\hbar)\Big); \tag{9.37}$$

see, e.g., [17, §6.5].

To prove Theorem 9.5 we recall that, for Q as in the theorem and $\mathcal{F}$ the (non-semiclassical) Fourier transform defined by (4.24),

$$\mathcal{F}\Big(\exp\big(\mathrm{i}\langle Qx,x\rangle/2\big)\Big) = \frac{(2\pi)^{d/2}\exp\big(\mathrm{i}(\pi/4)\operatorname{sgn}Q\big)}{|\det Q|^{1/2}}\exp\big(-\mathrm{i}\langle Q^{-1}\xi,\xi\rangle/2\big) \tag{9.38}$$

(see Exercise [7]) and, for $u, v\in\mathscr{S}$,

$$\int_{\mathbb{R}^d} u\,\overline{v} = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \mathcal{F}u\,\overline{\mathcal{F}v}; \tag{9.39}$$

i.e., Plancherel's theorem (cf. the consequence (5.5), written in terms of $\mathcal{F}_\hbar$).

Proof of Theorem 9.5. Using (9.39) (with $v=\overline{a}$) and (9.38) (the latter letting $Q\mapsto Q/\hbar$), we find

$$I(\hbar,a) = \left(\frac{\hbar}{2\pi}\right)^{d/2}\frac{\exp\big(\mathrm{i}(\pi/4)\operatorname{sgn}Q\big)}{|\det Q|^{1/2}}\int_{\mathbb{R}^d}\exp\big(-\mathrm{i}\hbar\langle Q^{-1}\xi,\xi\rangle/2\big)\,\overline{\mathcal{F}\overline{a}(\xi)}\,\mathrm{d}\xi.$$

Recalling that $\overline{\mathcal{F}\overline{a}(\xi)} = \mathcal{F}a(-\xi)$, and then making the change of variables $\xi\mapsto-\xi$, we obtain that

$$I(\hbar,a) = \left(\frac{\hbar}{2\pi}\right)^{d/2}\frac{\exp\big(\mathrm{i}(\pi/4)\operatorname{sgn}Q\big)}{|\det Q|^{1/2}}\int_{\mathbb{R}^d}\exp\big(-\mathrm{i}\hbar\langle Q^{-1}\xi,\xi\rangle/2\big)\,\mathcal{F}a(\xi)\,\mathrm{d}\xi.$$

Taylor's theorem implies that

$$f(\hbar) - \sum_{j=0}^{N-1}\frac{\hbar^j}{j!}f^{(j)}(0) = \frac{1}{(N-1)!}\int_0^\hbar(\hbar-s)^{N-1}f^{(N)}(s)\,\mathrm{d}s = \frac{\hbar^N}{(N-1)!}\int_0^1(1-t)^{N-1}f^{(N)}(\hbar t)\,\mathrm{d}t,$$

so that

$$\begin{aligned}\exp\big(-\mathrm{i}\hbar\langle Q^{-1}\xi,\xi\rangle/2\big) = {}& \sum_{j=0}^{N-1}\frac{\hbar^j}{j!}\left(\frac{-\mathrm{i}\langle Q^{-1}\xi,\xi\rangle}{2}\right)^j \\ & + \frac{\hbar^N}{(N-1)!}\int_0^1(1-t)^{N-1}\exp\big(-\mathrm{i}t\hbar\langle Q^{-1}\xi,\xi\rangle/2\big)\left(\frac{-\mathrm{i}\langle Q^{-1}\xi,\xi\rangle}{2}\right)^N\mathrm{d}t.\end{aligned}$$

Therefore,

$$I(\hbar,a) = \left(\frac{\hbar}{2\pi}\right)^{d/2}\frac{\exp\big(\mathrm{i}(\pi/4)\operatorname{sgn}Q\big)}{|\det Q|^{1/2}}\left(\sum_{j=0}^{N-1}\frac{\hbar^j}{j!}\int_{\mathbb{R}^d}\left(\frac{-\mathrm{i}\langle Q^{-1}\xi,\xi\rangle}{2}\right)^j\mathcal{F}a(\xi)\,\mathrm{d}\xi + \frac{\hbar^N}{N!}(2\pi)^d R_N(\hbar,a)\right),$$

with

$$R_N(\hbar,a) := \frac{N}{(2\pi)^d}\int_0^1(1-t)^{N-1}\left(\int_{\mathbb{R}^d}\exp\big(-\mathrm{i}t\hbar\langle Q^{-1}\xi,\xi\rangle/2\big)\left(-\frac{\mathrm{i}}{2}\langle Q^{-1}\xi,\xi\rangle\right)^N\mathcal{F}a(\xi)\,\mathrm{d}\xi\right).$$

[7] to be created

By the Fourier inversion theorem $u(0) = (2\pi)^{-d} \int_{\mathbb{R}^d} \mathcal{F}u(\xi)\,\mathrm{d}\xi$, and thus

$$\int_{\mathbb{R}^d} \left(\frac{-\mathrm{i}\langle Q^{-1}\xi,\xi\rangle}{2}\right)^j \mathcal{F}a(\xi)\,\mathrm{d}\xi = \left(\left(\frac{-\mathrm{i}\langle Q^{-1}D,D\rangle}{2}\right)^j a\right)(0),$$

and similarly,

$$\begin{aligned}\int_{\mathbb{R}^d} \exp\Big(-\mathrm{i}t\hbar\langle Q^{-1}\xi,\xi\rangle/2\Big)\left(-\frac{\mathrm{i}}{2}\langle Q^{-1}\xi,\xi\rangle\right)^N \mathcal{F}a(\xi)\,\mathrm{d}\xi \\ = \left(\exp\Big(-\mathrm{i}t\hbar\langle Q^{-1}D,D\rangle/2\Big)\left(-\frac{\mathrm{i}}{2}\langle Q^{-1}D,D\rangle\right)^N a\right)(0),\end{aligned}$$

and the result (9.35) follows. □

We now use Theorem 9.5 to obtain the asymptotics of an integral of the form (9.32).

Corollary 9.6. *If $a \in \mathscr{S}(\mathbb{R}^{2d})$ then, for all $N \in \mathbb{Z}^+$,*

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \exp\big(\mathrm{i}\langle x,y\rangle/\hbar\big)\,a(x,y)\,\mathrm{d}x\,\mathrm{d}y = (2\pi\hbar)^d\left(\sum_{j=0}^{N-1}\frac{\hbar^j}{j!}\Big(\big(-\mathrm{i}\langle D_x,D_y\rangle\big)^j a\Big)(0,0) + \frac{\hbar^N}{N!}R_N(\hbar,a)\right), \tag{9.40}$$

where

$$R_N(\hbar,a) := \frac{N}{(2\pi)^d}\int_0^1 (1-t)^{N-1}\Big(\exp\big(-\mathrm{i}t\hbar\langle D_x,D_y\rangle\big)\big(-\mathrm{i}\langle D_x,D_y\rangle\big)^N a\Big)(0,0)\,\mathrm{d}t.$$

Proof. We apply Theorem 9.5 with $d \mapsto 2d$ and (x,y) denoting a point in $\mathbb{R}^{2d}$. Let

$$Q := \begin{pmatrix} 0 & I \\ I & 0\end{pmatrix} \in \mathbb{R}^{2d\times 2d},$$

and observe that $Q = Q^T = Q^{-1}$, $|\det Q| = 1$, $\operatorname{sgn}(Q) = 0$, $Q(x,y) = (y,x)$, and thus $\langle Q(x,y),(x,y)\rangle/2 = \langle x,y\rangle$. Since $D = (D_x,D_y)$, $\langle Q^{-1}D,D\rangle/2 = \langle D_x,D_y\rangle$. The result then follows from Theorem 9.5. □

Lemma 9.7.

$$\langle D_x,D_y\rangle^j a(x,y) = \sum_{|\alpha|=j}\frac{j!}{\alpha!}D_x^\alpha D_y^\alpha a(x,y). \tag{9.41}$$

Proof. The multinomial theorem states that

$$(x_1+\ldots+x_d)^j = \sum_{|\alpha|=j}\binom{j}{\alpha}x^\alpha, \quad \text{where } \binom{j}{\alpha} := \frac{j!}{\alpha_1!\ldots\alpha_d!} = \frac{j!}{\alpha!}. \tag{9.42}$$

Therefore, by its definition as a Fourier multiplier,

$$\langle D_x,D_y\rangle^j a(x,y) = \mathcal{F}^{-1}_{(\zeta_x,\zeta_y)\to(x,y)}\Big(\langle\zeta_x,\zeta_y\rangle^j \mathcal{F}_{(x,y)\to(\zeta_x,\zeta_y)}a\Big)(x,y).$$

Now, by (9.42) and the definition of multiindices,

$$\langle\zeta_x,\zeta_y\rangle^j = \big(\zeta_{x_1}\zeta_{y_1}+\cdots+\zeta_{x_d}\zeta_{y_d}\big)^j = \sum_{|\alpha|=j}\frac{j!}{\alpha!}\zeta_x^\alpha\zeta_y^\alpha.$$

Therefore

$$\langle D_x,D_y\rangle^j a(x,y) = \sum_{|\alpha|=j}\frac{j!}{\alpha!}\mathcal{F}^{-1}_{(\zeta_x,\zeta_y)\to(x,y)}\Big(\zeta_x^\alpha\,\zeta_y^\alpha\,\mathcal{F}_{(x,y)\to(\zeta_x,\zeta_y)}a\Big)(x,y) = \sum_{|\alpha|=j}\frac{j!}{\alpha!}D_x^\alpha D_y^\alpha a(x,y).$$

□

9.4.2 Proof of the analogue of Theorem 9.4 when $a, b \in \mathscr{S}$

Lemma 9.8. (Composition property for Schwartz symbols.) *Given $a, b \in \mathscr{S}$, let $a\#b$ be defined by (9.32). Then $a\#b \in \mathscr{S}$, $\operatorname{Op}_\hbar(a)\operatorname{Op}_\hbar(b) = \operatorname{Op}_\hbar(a\#b)$ (i.e., (9.33) holds), and, for all $N \in \mathbb{Z}^+$,*

$$a\#b(x,\xi) = \sum_{|\alpha|\le N-1} \frac{(\mathrm{i}\hbar)^{|\alpha|}}{\alpha!}\big(D_\xi^\alpha a(x,\xi)\big)\big(D_x^\alpha b(x,\xi)\big) + \frac{\hbar^N}{N!}R_N, \tag{9.43}$$

where

$$R_N := \frac{N}{(2\pi)^d}\int_0^1 (1-t)^{N-1}\exp\big(\mathrm{i}t\hbar\langle D_{\widetilde{x}}, D_{\widetilde{\xi}}\rangle\big)\big(\mathrm{i}\langle D_{\widetilde{x}}, D_{\widetilde{\xi}}\rangle\big)^N a(x,\xi+\widetilde{\xi})\, b(x+\widetilde{x},\xi)\Big|_{\widetilde{x}=0,\widetilde{\xi}=0}\mathrm{d}t. \tag{9.44}$$

Proof. By the definitions (7.2) and (7.3),

$$\begin{aligned}
\operatorname{Op}_\hbar(a)\operatorname{Op}_\hbar(b)v(x) &= \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\exp(\mathrm{i}\langle x-y,\eta\rangle/\hbar)a(x,\eta)\big(\operatorname{Op}_\hbar(b)v\big)(y)\,\mathrm{d}y\,\mathrm{d}\eta\\
&= \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\exp(\mathrm{i}\langle x-y,\eta\rangle/\hbar)a(x,\eta)\exp(\mathrm{i}y\cdot\xi/\hbar)b(y,\xi)(\mathcal{F}_\hbar v)(\xi)\,\mathrm{d}y\,\mathrm{d}\eta\,\mathrm{d}\xi\\
&= \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d}\exp(\mathrm{i}\langle x,\xi\rangle/\hbar)c(x,\xi)(\mathcal{F}_\hbar v)(\xi)\,\mathrm{d}\xi\\
&= \operatorname{Op}_\hbar(c)v(x),
\end{aligned}$$

where

$$c(x,\xi) := \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\exp\big(-\mathrm{i}\langle x-y,\xi-\eta\rangle/\hbar\big)a(x,\eta)b(y,\xi)\,\mathrm{d}y\,\mathrm{d}\eta.$$

Using the change of variables $y = x+\widetilde{x}$ and $\eta = \xi+\widetilde{\xi}$, we find that $c(x,\xi) = (a\#b)(x,\xi)$ as claimed.

Letting $\widetilde{\xi} \mapsto -\widetilde{\xi}$ in (9.32), we obtain that

$$a\#b(x,\xi) := \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\exp\big(\mathrm{i}\widetilde{x}\cdot\widetilde{\xi}/\hbar\big)\,a(x,\xi-\widetilde{\xi})b(x+\widetilde{x},\xi)\,\mathrm{d}\widetilde{\xi}\,\mathrm{d}\widetilde{x},$$

which has the same form as the left-hand side of (9.40). Observe that (9.41) implies that

$$\big(-\mathrm{i}\langle D_{\widetilde{x}}, D_{\widetilde{\xi}}\rangle\big)^j a(x,\widetilde{\xi}-\xi)b(x+\widetilde{x},\xi)\Big|_{\widetilde{x}=0,\widetilde{\xi}=0} = \sum_{|\alpha|=j}\frac{j!(-1)^j(-\mathrm{i})^j}{\alpha!}\big(D_\xi^\alpha a(x,\xi)\big)\big(D_x^\alpha b(x,\xi)\big).$$

Therefore, by the asymptotics in (9.40),

$$a\#b(x,\xi) = \sum_{|\alpha|\le N-1}\frac{(\mathrm{i}\hbar)^{|\alpha|}}{\alpha!}\big(D_\xi^\alpha a(x,\xi)\big)\big(D_x^\alpha b(x,\xi)\big) + \frac{\hbar^N}{N!}R_N,$$

where

$$R_N := \frac{N}{(2\pi)^d}\int_0^1 (1-t)^{N-1}\exp\big(-\mathrm{i}t\hbar\langle D_{\widetilde{x}}, D_{\widetilde{\xi}}\rangle\big)\big(-\mathrm{i}\langle D_{\widetilde{x}}, D_{\widetilde{\xi}}\rangle\big)^N a(x,\xi-\widetilde{\xi})\, b(x+\widetilde{x},\xi)\Big|_{\widetilde{x}=0,\widetilde{\xi}=0}\mathrm{d}t.$$

Making the change of variables $\widetilde{\xi} \mapsto -\widetilde{\xi}$, we find (9.44). □

9.4.3 Lemmas about converting between symbol classes

Definition 9.9. (Order function.) *m is an* order function *if there exist $C, N > 0$ such that*

$$m(w) \le C\langle z-w\rangle^N m(z) \quad \text{for all } w, z \in \mathbb{R}^{2d}.$$

Example 9.10. *The following are order functions:*

(i) $m(w) = 1$.

(ii) $m(w) = \langle w\rangle$ (by Peetre's inequality (9.23)),

(iii) $m((x,\xi)) = \langle x\rangle^a\langle\xi\rangle^b$ for any $a, b \in \mathbb{R}$.

Furthermore, if m_1 and m_2 are order functions, then so is m_1m_2.

Definition 9.11. (The symbol class $S(m)$.) *Given an order function m on $\mathbb{R}^{2d}$, let*

$$S(m) := \Big\{ a \in C^\infty(\mathbb{R}^{2d}) : \textit{for all } \alpha \textit{ there exists } C_\alpha \textit{ such that, for all } 0 < \hbar \le \hbar_0, \\ |\partial^\alpha m(w)| \le C_\alpha m(w) \Big\}.$$

Lemma 9.12. (Converting between S^m and $S(\langle\xi\rangle^m)$.)
(a) For all $m \in \mathbb{Z}$,

$$S^m \subset S(\langle\xi\rangle^m)$$

(b) Let c be a symbol and assume that there exists a sequence $(\sigma_j)_{j\in\mathbb{N}}$ of symbols such that

(i) $\sigma_j \in S^{m-j}$ for all $j \in \mathbb{Z}^+$,

(ii) For each $N \in \mathbb{Z}^+$,

$$c - \sum_{j=0}^{N-1} \hbar^j \sigma_j \in \hbar^N S(\langle\xi\rangle)^{m-N}.$$

Then $c \in S^m$ and, for all $N \in \mathbb{Z}^+$,

$$c - \sum_{j=0}^{N-1} \hbar^j \sigma_j \in \hbar^N S^{m-N}.$$

Proof. Part (a) follows immediately from the definitions of $S(m)$ (Definition 9.11) and S^m (7.1).

For Part (b), given α, β, let $N = |\beta|$. By Property (ii), for all $0 < \hbar \le \hbar_0$,

$$\big|\partial_x^\alpha \partial_\xi^\beta c\big| \le \sum_{j=0}^{N-1} \hbar_0^j \big|\partial_x^\alpha \partial_\xi^\beta \sigma_j\big| + \hbar_0^N \big|\partial_x^\alpha \partial_\xi^\beta R_N\big|, \tag{9.45}$$

where

$$R_N := \hbar^{-N}\Big(c - \sum_{j=0}^{N-1} \hbar^j \sigma_j\Big).$$

By assumption

$$R_N \in S\big(\langle\xi\rangle^{m-N}\big) = S\big(\langle\xi\rangle^{m-|\beta|}\big).$$

Therefore, using this last inclusion and the fact that $\sigma_j \in S^{m-j}$ in (9.45), we see that there exists $C_{\alpha\beta} > 0$ such that, for all $0 < \hbar \le \hbar_0$,

$$|\partial_x^\alpha \partial_\xi^\beta c_\hbar(x,\xi)| \le C_{\alpha\beta}\langle\xi\rangle^{m-|\beta|} \quad \text{for all } (x,\xi);$$

i.e., $c \in S^m$. We now argue similarly to show that $R_N \in S^{m-N}$. By definition, for any $M \in \mathbb{Z}^+$,

$$\hbar^N R_N = c - \sum_{j=0}^{N-1} \hbar^j \sigma_j = \sum_{j=N}^{N+M-1} \hbar^j \sigma_j + \hbar^{N+M} R_{N+M},$$

for some $R_{M+N} \in S(\langle\xi\rangle^{m-N-M})$. Rearranging, we find that

$$R_N = \sum_{\ell=0}^{M-1} \hbar^\ell \sigma_{N+\ell} + \hbar^M R_{M+N},$$

Differentiating and arguing similar to above (using that $\sigma_{N+\ell} \in S^{m-N-\ell}$), we find that $R_N \in S^{m-N}$. □

9.4.4 Proof of Theorem 9.4

Lemma 9.13. *Let Q be a symmetric non-singular matrix and let m be an order function. Then* $\exp(\mathrm{i}\hbar\langle QD, D\rangle/2) : \mathscr{S} \to \mathscr{S}$ *has a unique extension to an operator* $\exp(\mathrm{i}\hbar\langle QD, D\rangle/2) : S(m) \to S(m)$.

Proof of Theorem 9.4 using Lemma 9.13. Since $a \in S^{m_A}$ and $b \in S^{m_B}$, Part (a) of Lemma 9.12 implies that $a \in S(\langle\xi\rangle^{m_A})$ and $b \in S(\langle\xi\rangle^{m_B})$. Using this, the density of $\mathscr{S}$ in $S(m)$, and Lemma 9.13, we find that (9.43) holds with R_N given by (9.44).

Our plan now is to show that the assumptions of Part (b) of Lemma 9.12 are satisfied with $m := m_A + m_B$ and

$$\sigma_j := \sum_{|\alpha|=j} \frac{(\mathrm{i}\hbar)^{|\alpha|}}{\alpha!} \big(D_\xi^\alpha a(x,\xi)\big)\big(D_x^\alpha b(x,\xi)\big).$$

Since $a \in S^{m_A}$ and $b \in S^{m_B}$, one can show using the Leibnitz rule that $\sigma_j \in S^{m_A+m_B-j}$.

To apply Part (b) of Lemma 9.12, from which the result follows, we therefore only need to show that R_N defined by (9.44) is in $S(\langle\xi\rangle^{m_A+m_B-N})$. Using Lemma 9.7 and the assumptions that $a \in S^{m_A}$ and $b \in S^{m_B}$, we have

$$\big(\mathrm{i}\langle D_{\widetilde{x}}, D_{\widetilde{\xi}}\rangle\big)^N a(x,\xi+\widetilde{\xi})\, b(x+\widetilde{x},\xi)\Big|_{\widetilde{x}=0,\widetilde{\xi}=0} \in S\big(\langle\xi\rangle^{m_A+m_B-N}\big).$$

By Lemma 9.13,

$$\exp\big(\mathrm{i}t\hbar\langle D_{\widetilde{x}}, D_{\widetilde{\xi}}\rangle\big)\big(\mathrm{i}\langle D_{\widetilde{x}}, D_{\widetilde{\xi}}\rangle\big)^N a(x,\xi+\widetilde{\xi})\, b(x+\widetilde{x},\xi)\Big|_{\widetilde{x}=0,\widetilde{\xi}=0} \in S\big(\langle\xi\rangle^{m_A+m_B-N}\big);$$

thus R_N defined by (9.44) is in $S(\langle\xi\rangle^{m_A+m_B-N})$, Part (b) of Lemma 9.12 applies, and the result follows. □

Proof of Lemma 9.13. [8] □

9.5 Proof of Part (i) of Theorem 7.5 and (7.8)

Part (i) of Theorem 7.5 and (9.46) follow from the following result.

Theorem 9.14. *If $A = \mathrm{Op}_\hbar(a) \in \Psi_\hbar^m$, then $A^* \in \Psi_\hbar^m$ with $A^* = \mathrm{Op}_\hbar(a^*)$, where*

$$a^*(x,\xi) = \frac{1}{(2\pi\hbar)^d} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \exp\big(-\mathrm{i}\langle\widetilde{x},\widetilde{\xi}\rangle/\hbar\big)\overline{a(x+\widetilde{x},\xi+\widetilde{\xi})}\,\mathrm{d}\widetilde{x}\,\mathrm{d}\widetilde{\xi}. \tag{9.46}$$

Furthermore, for every $N \in \mathbb{Z}^+$,

$$a^*(x,\xi) - \sum_{|\alpha|\le N-1} \frac{(\mathrm{i}\hbar)^{|\alpha|}}{\alpha!} D_x^\alpha D_\xi^\alpha \overline{a(x,\xi)} \in \hbar^N S^{m-N}. \tag{9.47}$$

Observe the similarity of Theorem 9.14 to Theorem 9.4, and of (9.46) to (9.32). Similar to the proof of Theorem 9.4, we first prove the result when $a \in \mathscr{S}$.

Lemma 9.15. *If $a \in \mathscr{S}$ and $A = \mathrm{Op}_\hbar(a)$, then $A^* = \mathrm{Op}_\hbar(a^*)$ where a^* is defined by (9.46), and*

$$a^*(x,\xi) = \sum_{|\alpha|\le N-1} \frac{\hbar^{|\alpha|}(-1)^{|\alpha|}}{\alpha!} D_x^\alpha D_\xi^\alpha \overline{a(x,\xi)} + \frac{\hbar^N}{N!} R_N \tag{9.48}$$

where

$$R_N := \frac{N}{(2\pi)^d} \int_0^1 (1-t)^{N-1} \exp\big(-\mathrm{i}t\hbar\langle D_{\widetilde{x}}, D_{\widetilde{\xi}}\rangle\big)\big(-\mathrm{i}\langle D_{\widetilde{x}}, D_{\widetilde{\xi}}\rangle\big)^N \overline{a(x+\widetilde{x},\xi-\widetilde{\xi})}\Big|_{\widetilde{x}=0,\widetilde{\xi}=0} \mathrm{d}t. \tag{9.49}$$

[8] to come

Proof. Recall that, given $A : \mathscr{S}(\mathbb{R}^d) \to \mathscr{S}(\mathbb{R}^d)$, its formal adjoint $A^* : \mathscr{S}^*(\mathbb{R}^d) \to \mathscr{S}^*(\mathbb{R}^d)$ is defined by $\langle A^* u, v\rangle_{\mathbb{R}^d} = \langle u, Av\rangle_{\mathbb{R}^d}$.

By the definition of $\mathrm{Op}_\hbar$ (7.2),

$$\begin{aligned}\langle Au, v\rangle_{\mathbb{R}^d} &= \int_{\mathbb{R}^d} Au(x)\,\overline{v(x)}\,\mathrm{d}x \\ &= \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \exp\Big(\mathrm{i}\langle x-y,\xi\rangle/\hbar\Big) a(x,\xi)\,u(y)\,\mathrm{d}\xi\,\mathrm{d}y\right)\overline{v(x)}\,\mathrm{d}x \\ &= \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d} u(y)\left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \exp\Big(\mathrm{i}\langle x-y,\xi\rangle/\hbar\Big) a(x,\xi)\,\overline{v(x)}\,\mathrm{d}\xi\,\mathrm{d}x\right)\mathrm{d}y.\end{aligned}$$

Since we want to obtain $\overline{\mathrm{Op}_\hbar(c)v(y)}$ under the y integral, and since this can be written in terms of $\mathcal{F}_\hbar v$ by (7.3), we use the Fourier inversion formula (5.2) to write

$$\begin{aligned}\langle Au, v\rangle_{\mathbb{R}^d} &= \frac{1}{(2\pi\hbar)^{2d}}\int_{\mathbb{R}^d} u(y)\left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \exp\Big(\mathrm{i}\langle x-y,\xi\rangle/\hbar - \mathrm{i}\langle x,\eta\rangle/\hbar\Big)\overline{\mathcal{F}_\hbar v(\eta)}\,a(x,\xi)\mathrm{d}x\,\mathrm{d}\xi\,\mathrm{d}\eta\right)\mathrm{d}y \\ &= \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d} u(y)\,\overline{\left(\int_{\mathbb{R}^d}\exp\Big(\mathrm{i}\langle y,\eta\rangle/\hbar\Big)c(y,\eta)\,\mathcal{F}_\hbar v(\eta)\,\mathrm{d}\eta\right)}\mathrm{d}y,\end{aligned}$$

where

$$c(y,\eta) = \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\exp\Big(-\mathrm{i}\langle x-y,\xi-\eta\rangle/\hbar\Big)\overline{a(x,\xi)}\,\mathrm{d}x\,\mathrm{d}\xi.$$

Relabelling $x \to \widetilde{x}, \xi \to \widetilde{\xi}, y \to x$, and $\eta \to \xi$, we obtain

$$c(x,\xi) = \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\exp\Big(-\mathrm{i}\langle \widetilde{x}-x,\widetilde{\xi}-\xi\rangle/\hbar\Big)\overline{a(\widetilde{x},\widetilde{\xi})}\,\mathrm{d}\widetilde{x}\,\mathrm{d}\widetilde{\xi}.$$

The expression (9.46) then follows by changing variables $\widetilde{x} \mapsto \widetilde{x} + x$ and $\widetilde{\xi} \mapsto \widetilde{\xi} + \xi$.

Changing variables $\widetilde{\xi} \mapsto -\widetilde{\xi}$, we find that

$$a^*(x,\xi) = \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\exp\big(\mathrm{i}\langle\widetilde{x},\widetilde{\xi}\rangle/\hbar\big)\overline{a(x+\widetilde{x},\xi-\widetilde{\xi})}\,\mathrm{d}\widetilde{x}\,\mathrm{d}\widetilde{\xi}.$$

The expansion (9.48) with remainder (9.49) then follow from the asymptotics (9.40) and Lemma 9.7. □

The proof of Theorem 9.14 using Lemma 9.15 follows the same steps as the proof of Theorem 9.4 in §9.4.4.

9.6 Exercises for §9

1. Show that if $a \in S^{m_A}$ and $b \in S^{m_B}$ with $|b(x,\xi)| \geq c\langle\xi\rangle^{m_B}$ on $\operatorname{supp} a$ for some $c > 0$, then $a/b \in S^{m_A - m_B}$.

   Solution: We prove by induction on $N = |\alpha| + |\beta|$ that

   $$\big|\partial_x^\alpha \partial_\xi^\beta (a/b)\big| \leq C_{\alpha\beta}\langle\xi\rangle^{m_A - m_B - |\beta|}. \tag{9.50}$$

   For $N = 0$ this bound holds by the lower bound on b and the fact that $a \in S^{m_A}$.

   Assume that (9.50) holds for some N. By the Leibniz rule,

   $$\partial_{x_i}\partial_x^\alpha\partial_\xi^\beta\Big(a/b\Big) = \partial_x^\alpha\partial_\xi^\beta\Big((\partial_{x_i}a)/b - a(\partial_{x_i}b)/b^2\Big).$$

   Since $\partial_{x_i}a \in S^{m_A}$, (9.50) with $a \mapsto \partial_{x_i}a$ and $b \mapsto b$ implies that

   $$\big|\partial_x^\alpha\partial_\xi^\beta\big((\partial_{x_i}a)/b\big)\big| \leq C_{\alpha\beta}\langle\xi\rangle^{m_A - m_B - |\beta|}. \tag{9.51}$$

Now $a\partial_{x_i} b \in S^{m_A+m_B}$ and $b^2 \in S^{2m_B}$ with $|b(x,\xi)^2| \geq c^2\langle\xi\rangle^{2m_B}$ on $\operatorname{supp}(a\partial_{x_i} b) \subset \operatorname{supp} a$. Therefore, (9.50) with $a \mapsto a\partial_{x_i} b$ and $b \mapsto b^2$ implies that

$$\left|\partial_x^\alpha \partial_\xi^\beta \left(a(\partial_{x_i} b)/b^2\right)\right| \leq C_{\alpha\beta}\langle\xi\rangle^{m_A+m_B-2m_B} = C_{\alpha\beta}\langle\xi\rangle^{m_A-m_B}. \tag{9.52}$$

The combination of (9.51) and (9.52) implies that

$$\left|\partial_{x_i}\partial_x^\alpha \partial_\xi^\beta (a/b)\right| \leq C_{\alpha\beta}\langle\xi\rangle^{m_A-m_B-|\beta|}.$$

The bound

$$\left|\partial_{\xi_i}\partial_x^\alpha \partial_\xi^\beta (a/b)\right| \leq C_{\alpha\beta}\langle\xi\rangle^{m_A-m_B-|\beta|-1}$$

can be proved similarly, and this completes the proof.

2. Show that if $F \in C^\infty(\mathbb{R})$ and $a \in S^0$ is real valued, then $F(a) \in S^0$.

   Solution: Similar to the proof in Question 1, we prove by induction on $N = |\alpha| + |\beta|$ that

   $$\left|\partial_x^\alpha \partial_\xi^\beta F(a)\right| \leq C_{\alpha\beta}\langle\xi\rangle^{-|\beta|}. \tag{9.53}$$

   For $N = 0$ this bound holds since $|a| \leq C$ (since $a \in S^0$).

   Assume that (9.50) holds for some N. By the Leibniz rule,

   $$\begin{aligned}
   |\partial_{x_i}\partial_x^\alpha \partial_\xi^\beta F(a)| &= \left|\partial_x^\alpha \partial_\xi^\beta \left(F'(a)(\partial_{x_i} a)\right)\right| \\
   &= \left|\sum_{(\alpha',\beta')\leq(\alpha,\beta)} \binom{(\alpha,\beta)}{(\alpha',\beta')} \partial_x^{\alpha'}\partial_\xi^{\beta'}\left(F'(a)\right)\partial_x^{\alpha-\alpha'}\partial_\xi^{\beta-\beta'}(\partial_{x_i} a_k)\right| \\
   &\leq C_{\alpha\beta}\langle\xi\rangle^{-|\beta'|}\langle\xi\rangle^{-|\beta|+|\beta'|} \leq C_{\alpha\beta}\langle\xi\rangle^{-|\beta|},
   \end{aligned}$$

   where we have used (9.53) applied with F replaced by F' and the fact that $\partial_{x_i} a \in S^0$.

3. Show that if $a \in S^0$ and a is real valued with $a(x,\xi) \geq c > 0$ for all (x,ξ), then $\sqrt{a} \in S^0$. Hint: use the result of Question 2.

   Solution: We cannot apply the result of Question 2 directly since $\sqrt{\cdot}$ is not C^∞. However, if $I := a(T^*\mathbb{R}^d)$, then $I \subset \mathbb{R} \setminus (-c,c)$ and $\sqrt{\cdot}$ is C^∞ on I. We now construct a C^∞ extension F of $\sqrt{\cdot}$. Let

   $$F(x) := \phi(x)\sqrt{x},$$

   where $\phi = 1$ on $\mathbb{R}^d \setminus (-c,c)$ and $\phi = 0$ on $(-c/2, c/2)$. Then $F(a) = \sqrt{a}$ and $F(a) \in S^0$ by the result of Question 2.

4. Prove Lemma 9.2 by considering the cases $|x-y| \leq \hbar$ and $|x-y| \geq \hbar$ separately, and, in the latter case, integrating by parts using

   $$L := \frac{\hbar(x-y)\cdot D_\xi}{|x-y|^2}. \tag{9.54}$$

   Solution: As in the proof above, we show that $\sup_x \int |K_A(x,y)|\,\mathrm{d}y$ is bounded independently of $\hbar$; the proof that $\sup_y \int |K_A(x,y)|\,\mathrm{d}x$ is bounded independently of $\hbar$ is very similar, and then the result follows from Lemma 9.1.

   If $|x-y| \leq \hbar$, we estimate

   $$|K_A(x,y)| \leq \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d} |a(x,\xi)|\mathrm{d}\xi \lesssim \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d}\langle\xi\rangle^m \mathrm{d}\xi \lesssim \hbar^{-d},$$

   since $m < -d$. Then

   $$\int_{|x-y|\leq\hbar} |K_A(x,y)|\,\mathrm{d}y \lesssim \hbar^d\hbar^{-d} \lesssim 1.$$

If $|x-y| \geq \hbar$, we integrate by parts the expression (9.26) for the Schwartz kernel. With L defined by (9.54)

$$L\big(\mathrm{e}^{\mathrm{i}(x-y)\cdot\xi/\hbar}\big) = \mathrm{e}^{\mathrm{i}(x-y)\cdot\xi/\hbar};$$

therefore, for any $N > 0$,

$$K_A(x,y) = \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d} L^N\big(\mathrm{e}^{\mathrm{i}(x-y)\cdot\xi/\hbar}\big)a(x,\xi)\,\mathrm{d}\xi = \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}(x-y)\cdot\xi/\hbar}(L^*)^N a(x,\xi)\,\mathrm{d}\xi.$$

By the definition of L and the fact that $a \in S^m$, given $N \in \mathbb{Z}^+$ there exists $C_N > 0$ such that

$$|(L^*)^N a(x,\xi)| \leq C_N \frac{\hbar^N}{|x-y|^N}\langle\xi\rangle^{m-N}.$$

Therefore, since $m < -d$,

$$\big|K_A(x,y)\big| \leq \frac{C_N}{(2\pi\hbar)^d}\frac{\hbar^N}{|x-y|^N}\int_{\mathbb{R}^d}\langle\xi\rangle^{m-N}\,\mathrm{d}\xi \leq C'_N\frac{\hbar^{N-d}}{|x-y|^N},$$

for some $C'_N > 0$. Therefore,

$$\int_{|x-y|\geq\hbar}\big|K_A(x,y)\big|\mathrm{d}y \leq C'_N\hbar^{N-d}\int_{|x-y|\geq\hbar}\frac{1}{|x-y|^N}\mathrm{d}y. \tag{9.55}$$

Now

$$\int_{|x-y|\geq\hbar}\frac{1}{|x-y|^N}\mathrm{d}y = C_d\int_\hbar^\infty \frac{1}{r^N}r^{d-1}\,\mathrm{d}r = C_d\hbar^{d-N}\int_1^\infty\frac{1}{t^N}t^{d-1}\mathrm{d}t, \tag{9.56}$$

where C_d is the surface area of the unit sphere in d dimensions (i.e., $C_d = 2\pi^{d/2}/\Gamma(d/2)$). Choosing $N > d$ (so that the last integral is finite) and combining (9.55) and (9.56), we obtain that $\sup_x \int |K_A(x,y)|\,\mathrm{d}y \lesssim 1$ and the proof is complete.

10 The Hamiltonian flow defined by the principal symbol of the Helmholtz equation

10.1 Recap of Hamilton's equations

Given a function $H(x,\xi)$ (the *Hamiltonian*), Hamilton's equations are

$$\frac{\mathrm{d}x_i}{\mathrm{d}t}(t) = \frac{\partial}{\partial\xi_i}H\big(x(t),\xi(t)\big), \qquad \frac{\mathrm{d}\xi_i}{\mathrm{d}t}(t) = -\frac{\partial}{\partial x_i}H\big(x(t),\xi(t)\big). \tag{10.1}$$

One usually thinks of the variable x as corresponding to position, and the variable ξ corresponding to momentum.

Lemma 10.1. (Evolution of quantities along the flow.) *For a function $f(x,\xi;t)$, if $(x(t),\xi(t))$ satisfies Hamilton's equations (10.1), then*

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(f(x(t),\xi(t);t)\big) = \left(\{H,f\} + \frac{\partial f}{\partial t}\right)\big(x(t),\xi(t);t\big), \tag{10.2}$$

where the Poisson bracket $\{\cdot,\cdot\}$ is defined by (7.7).

We write (10.2) more succinctly as

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \left(\{H,f\} + \frac{\partial f}{\partial t}\right). \tag{10.3}$$

Proof of Lemma 10.1. By the chain rule and the definition (7.7) of $\{\cdot,\cdot\}$,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\big(f(x(t),\xi(t);t)\big) &= \sum_j \left(\frac{\partial f}{\partial x_j}\frac{\mathrm{d}x_j}{\mathrm{d}t} + \frac{\partial f}{\partial \xi_j}\frac{\mathrm{d}\xi_j}{\mathrm{d}t}\right) + \frac{\partial f}{\partial t} \\ &= \sum_j \left(\frac{\partial f}{\partial x_j}\frac{\partial H}{\partial \xi_j} - \frac{\partial f}{\partial \xi_j}\frac{\partial H}{\partial x_j}\right) + \frac{\partial f}{\partial t} = \{H,f\} + \frac{\partial f}{\partial t}.\end{aligned}$$

□

Since $\{H,H\}=0$, Lemma 10.1 has the following corollary.

Corollary 10.2. *$H(x(t),\xi(t))$ is constant as a function of t.*

Notation 10.3. (The flow $\varphi_t(\rho)$.) *Given $\rho=(x_0,\xi_0)$,*

$$\varphi_t(\rho) := (x(t),\xi(t)),$$

where $(x(t),\xi(t))$ is the solution of (10.1) with initial condition $(x(0),\xi(0))=(x_0,\xi_0)$.

Given $a(x,\xi)$, (10.2)/(10.3) can therefore be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}t}(a\circ\varphi_t) = \{H,a\}. \tag{10.4}$$

10.2 The case when the Hamiltonian is the semiclassical principal symbol of the Helmholtz equation

Let $P_\hbar := -\hbar^2\nabla\cdot(A\nabla\cdot) - n$, so that $\sigma_\hbar(P_\hbar) = \langle A\xi,\xi\rangle - n$ (as in (8.7)). Hamilton's equations (10.1) with $H=\sigma_\hbar(P_\hbar)$ are then

$$\frac{\mathrm{d}x_i}{\mathrm{d}t}(t) = 2(A(x)\xi)_i, \qquad \frac{\mathrm{d}\xi_i}{\mathrm{d}t}(t) = -\left\langle \frac{\partial A}{\partial x_i}(x)\xi,\xi\right\rangle + \frac{\partial n}{\partial x_i}(x). \tag{10.5}$$

As in Definition 2.6, we assume that A and n are both $C^{1,1}$, which implies that, given an initial condition, the solution of (10.5) is unique (by the Picard–Lindelöf theorem).

As noted below Definition 2.6, if $A=I$ and $n=1$, then $\sigma_\hbar(P_\hbar)=|\xi|^2-1$ and (10.5) become $\dot{x}_i = 2\xi_i$ and $\dot{\xi}_i = 0$, with solution

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

i.e., straight-line motion with speed $2|\xi_0|$. Special importance is played by the flow with $\sigma_\hbar(P_\hbar)=0$, which in this case implies that $|\xi_0|=1$; i.e., the flow has speed 2.

Let π_x denote projection in the x variables; i.e. $\pi_x((x,\xi))=x$.

Lemma 10.4. ("Going backwards with reversed speed = going forwards".)
(i) If $(x(t),\xi(t))$ is a solution to (10.5) then so is $(\widetilde{x}(t),\widetilde{\xi}(t)) := (x(-t),-\xi(-t))$.
(ii) $\pi_x(\varphi_t(x_0,\xi_0)) = \pi_x(\varphi_{-t}(x_0,-\xi_0))$.

Proof. (i) This proof relies on the fact that $\sigma_\hbar(P_\hbar)$ is even in ξ (and its ξ derivative is therefore odd in ξ). therefore Since $\widetilde{x}(t)=x(-t)$,

$$\frac{\mathrm{d}\widetilde{x}_i}{\mathrm{d}t}(t) = -\frac{\mathrm{d}x_i}{\mathrm{d}t}(-t) = -2\big(A(x(-t))\xi(-t)\big)_i = 2\big(A\big(\widetilde{x}(t)\big)\widetilde{\xi}(t)\big)_i$$

and since $\widetilde{\xi}(t) = -\xi(-t)$,

$$\begin{aligned}\frac{\mathrm{d}\widetilde{\xi}_i}{\mathrm{d}t}(t) = \frac{\mathrm{d}\xi_i}{\mathrm{d}t}(-t) &= -\left\langle \frac{\partial A}{\partial x_i}\big(x(-t)\big)\xi(-t),\xi(-t)\right\rangle + \frac{\partial n}{\partial x_i}\big(x(-t)\big) \\ &= -\left\langle \frac{\partial A}{\partial x_i}\big(\widetilde{x}(t)\big)\widetilde{\xi}(t),\widetilde{\xi}(t)\right\rangle + \frac{\partial n}{\partial x_i}\big(\widetilde{x}(t)\big);\end{aligned}$$

i.e. $(\widetilde{x}(t), \widetilde{\xi}(t))$ solve (10.5).

(ii) By definition $\varphi_t(x_0, \xi_0)$ is $(x(t), \xi(t))$ satisfying (10.5) with initial condition (x_0, ξ_0). By uniqueness of the solution and Part (i), $\varphi_{-t}(x_0, -\xi_0)$ is then $(\widetilde{x}(t), \widetilde{\xi}(t)) := (x(-t), -\xi(-t))$. Since $\{x(t)\}_{t\geq 0} = \{x(-t)\}_{t\leq 0}$, the result follows. □

Definition 10.5. (The forward and backward trapped sets.) *Let*

$$\Gamma_{\mathrm{fw}} := \left\{(x,\xi) : \left|\left(\pi_x(\varphi_t(x,\xi))\right)\right| \nrightarrow \infty \quad \textit{as} \quad t \to \infty\right\} \tag{10.6}$$

i.e., Γ_{bw} *is the* forward trapped set. *Let*

$$\Gamma_{\mathrm{bw}} := \left\{(x,\xi) : \left|\left(\pi_x(\varphi_t(x,\xi))\right)\right| \nrightarrow \infty \quad \textit{as} \quad t \to -\infty\right\} \tag{10.7}$$

i.e., Γ_{bw} *is the* backward trapped set.

Using the notation Γ for the trapped sets is common (see, e.g. [52, Definition 6.1]) and so we use it here, despite the slight notational clash of Γ_D for the Dirichlet boundary and $\Gamma_R := \partial B_R$.

Lemma 10.6. (Forward trapping $\iff$ backward trapping.) $\Gamma_{\mathrm{fw}} \neq \emptyset$ *iff* $\Gamma_{\mathrm{bw}} \neq \emptyset$.

Proof. We prove the forward implication; the proof of the reverse implication is very similar. Since $\Gamma_{\mathrm{fw}} \neq \emptyset$, by (10.6) there exists (x, ξ) such that

$$\left|\left(\pi_x(\varphi_t(x,\xi))\right)\right| \nrightarrow \infty \quad \text{as} \quad t \to \infty.$$

By Part (ii) of Lemma 10.4,

$$\left|\left(\pi_x(\varphi_{-t}(x,-\xi))\right)\right| \nrightarrow \infty \quad \text{as} \quad t \to \infty;$$

thus, letting $t \mapsto -t$,

$$\left|\left(\pi_x(\varphi_t(x,-\xi))\right)\right| \nrightarrow \infty \quad \text{as} \quad t \to -\infty,$$

so that $\Gamma_{\mathrm{bw}} \neq \emptyset$ by (10.7). □

Corollary 10.7. (Forward nontrapping $\iff$ backward nontrapping.) $\Gamma_{\mathrm{fw}} = \emptyset$ *iff* $\Gamma_{\mathrm{bw}} = \emptyset$.

The definition of nontrapping in Definition 2.6 says that coefficients A and n are *nontrapping* if $\Gamma_{\mathrm{fw}} = \emptyset$ *and* all trajectories escape uniformly; we now show that the latter condition is ensured by A and n satisfying Assumption 1.1.

Lemma 10.8. (Definition 2.6 $\equiv (\Gamma_{\mathrm{fw}} = \emptyset) \equiv (\Gamma_{\mathrm{bw}} = \emptyset)$.) *Suppose* $\Omega_- = \emptyset$ *and* A *and* n *satisfy Assumption 1.1 and consider the flow with* $\sigma_\hbar(P) = 0$. *Then* A, n *are nontrapping in the sense of Definition 2.6 iff* $\Gamma_{\mathrm{fw}} = \emptyset$ *iff* $\Gamma_{\mathrm{bw}} = \emptyset$.

Proof. By Corollary 10.7, the only thing we need to prove is that if $\Gamma_{\mathrm{fw}} = \emptyset$, then all trajectories starting in B_R leave B_R in a uniform time. This follows if we can show that $\|\mathrm{d}x_i/\mathrm{d}t\| \geq C$ for all i, with $C > 0$ independent of x. The first equation in (10.5), the fact that $A(x)$ is symmetric positive definite for each x, and the lower bound in (1.1) imply that

$$\left\|\frac{\mathrm{d}x_i}{\mathrm{d}t}\right\| \geq 2A_{\min}\|\xi\|.$$

However, since $\sigma_\hbar(P) = \langle A\xi, \xi\rangle - n = 0$, the bounds in (1.1) and (1.2) imply that

$$n_{\min} \leq \left\|A^{1/2}\xi\right\| \leq (A_{\max})^{1/2}\|\xi\|, \quad \text{and thus} \quad \left\|\frac{\mathrm{d}x_i}{\mathrm{d}t}\right\| \geq \frac{2A_{\min}n_{\min}}{(A_{\max})^{1/2}},$$

which is the desired lower bound on $\|\mathrm{d}x_i/\mathrm{d}t\|$. □

11 Defect measures

11.1 Functions that are locally bounded, uniformly in $\hbar$

Definition 11.1. $\{u(\hbar)\}_{0<\hbar\leq\hbar_0}$ *is* uniformly locally bounded *if given* $\chi \in C^\infty_{\mathrm{comp}}(\mathbb{R}^d)$ *there exists* $C > 0$ *such that*

$$\|\chi u(\hbar)\|_{L^2(\mathbb{R}^d)} \leq C \quad \text{for all } 0 < \hbar \leq \hbar_0.$$

The theory of defect measures is simpler for functions that are uniformly bounded in $L^2(\mathbb{R}^d)$, instead of in $L^2_{\mathrm{loc}}(\mathbb{R}^d)$. The reason we consider the latter is that the outgoing solution of $(\hbar^2\Delta + 1)u = 0$ is not in $L^2(\mathbb{R}^d)$ but is in $L^2_{\mathrm{loc}}(\mathbb{R}^d)$.

Definition 11.2.

$$S^{\mathrm{comp}} := \Big\{ a \in C^\infty_{\mathrm{comp}}(T^*\mathbb{R}^d) \,:\, \operatorname{supp} a \subset K \text{ for } K \text{ an } \hbar\text{-independent compact set} \Big\} \subset S^{-\infty}.$$

11.2 An important technicality

For the results in this section, we use the quantisation

$$\big(\operatorname{Op}^{\mathrm{ps}}_\hbar(a)v\big)(x) := (2\pi\hbar)^{-d} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \exp\big(\mathrm{i}(x-y)\cdot\xi/\hbar\big)\, a(x,\xi)\, v(y)\, \chi_0\big(|x-y|\big)\, \mathrm{d}y\, \mathrm{d}\xi \tag{11.1}$$

where χ_0 is a fixed function in $C^\infty_{\mathrm{comp}}(\mathbb{R})$ that is equal to one in a neighbourhood of zero; without loss of generality, we assume that $\chi_0(t) = 0$ for $|t| \geq 1/2$.

The advantage of using this quantisation is that $\operatorname{Op}^{\mathrm{ps}}_\hbar(a)$ is properly-supported (hence the "ps" superscript) for all symbols a by Part (i) of Lemma 7.15. In fact, in the proof of Lemma 7.16, we saw that

$$\operatorname{Op}_\hbar(a) = \operatorname{Op}^{\mathrm{ps}}_\hbar(a) + O(\hbar^\infty)_{\Psi^{-\infty}_\hbar}; \tag{11.2}$$

(observe that $\operatorname{Op}^{\mathrm{ps}}_\hbar(a)$ is essentially $\widetilde{A}$ defined by (9.9) with $\chi_0 = 1 - \psi_0$; the only difference is that the cut-off χ_0 is a function of $|x-y|$, whereas in (9.9) ψ_0 was a function of $x-y$).

The reason having properly-support operators is useful is given by the following lemma.

Lemma 11.3. *If* $a \in S^{\mathrm{comp}}$ *then* $\operatorname{Op}^{\mathrm{ps}}_\hbar(a)$ *is compactly supported with*

$$\operatorname{Op}^{\mathrm{ps}}_\hbar(a) = \chi_1 \operatorname{Op}^{\mathrm{ps}}_\hbar(a) = \chi_1 \operatorname{Op}^{\mathrm{ps}}_\hbar(a)\chi_2 \tag{11.3}$$

for any real-valued $\chi_1 \in \mathcal{D}$ *with* $\chi_1 \equiv 1$ *on* $\pi_x(\operatorname{supp} a)$*, and any real-valued* $\chi_2 \in \mathcal{D}$ *with* $\chi_2(y) = 1$ *for all* y *such that* $\operatorname{dist}(y, \operatorname{supp}\chi_1) \leq 1/2$.

Proof. Since $\chi_0(t) = 0$ for $|t| \geq 1/2$,

$$a(x,\xi)\, v(y)\, \chi_0\big(|x-y|\big) = \chi_1(x)\, a(x,\xi)\Big(\chi_2(y)\, v(y)\Big)\chi_0\big(|x-y|\big),$$

and the result follows from the definition (11.1) of $\operatorname{Op}^{\mathrm{ps}}_\hbar$. □

In the theory of defect measures, we consider pairings of the form

$$\big\langle \operatorname{Op}^{\mathrm{ps}}_\hbar(a)u, u\big\rangle$$

for $a \in S^{\mathrm{comp}}$ and $u \in L^2_{\mathrm{loc}}(\mathbb{R}^d)$. The property (11.3) ensures such pairings make sense; indeed,

$$\big\langle \operatorname{Op}^{\mathrm{ps}}_\hbar(a)u, u\big\rangle = \big\langle \chi_1 \operatorname{Op}^{\mathrm{ps}}_\hbar(a)\chi_2 u, u\big\rangle = \big\langle \operatorname{Op}^{\mathrm{ps}}_\hbar(a)\chi_2 u, \chi_1 u\big\rangle. \tag{11.4}$$

By Lemma 9.3, for $a \in S^{\mathrm{comp}}$, $\|\operatorname{Op}^{\mathrm{ps}}_\hbar(a)\|_{L^2(\mathbb{R}^d)\to L^2(\mathbb{R}^d)}$ is bounded independently of $\hbar$. Using this, and the fact that $\chi_j u \in L^2(\mathbb{R}^d)$, $j = 1, 2$, we see that the pairing $\langle \operatorname{Op}^{\mathrm{ps}}_\hbar(a)u, u\rangle$ is well defined; i.e.,

$$\big|\big\langle \operatorname{Op}^{\mathrm{ps}}_\hbar(a)u, u\big\rangle\big| < \infty.$$

11.3 Definition, existence, and positivity

11.3.1 Statement of results

Defect measures give a precise notion of where a sequence of functions $\{u(\hbar)\}_{0<\hbar\leq\hbar_0}$ lives in phase space in the limit $\hbar \to 0$.

Definition 11.4. (Defect measure.) *Given $\{u(\hbar)\}_{0<\hbar\leq\hbar_0}$, uniformly locally bounded, and a sequence $\hbar_n \to 0$, $\{u(\hbar)\}_{0<\hbar\leq\hbar_0}$ has defect measure μ if, for all $a \in S^{\mathrm{comp}}$,*

$$\lim_{n\to\infty} \left\langle \mathrm{Op}_\hbar^{\mathrm{ps}}(a)u(\hbar_n), u(\hbar_n)\right\rangle = \int_{T^*\mathbb{R}^d} a \,\mathrm{d}\mu. \tag{11.5}$$

Observe that (11.5) implies that if A is the quantisation of a symbol $a \in S^{\mathrm{comp}}$, then

$$\lim_{n\to\infty} \left\langle Au(\hbar_n), u(\hbar_n)\right\rangle = \int_{T^*\mathbb{R}^d} \sigma_\hbar(A) \,\mathrm{d}\mu;$$

indeed, this follows since $A - \mathrm{Op}_\hbar(\sigma_\hbar(A)) \in \hbar\Psi_\hbar^{-\infty}$, by the definition of the principal symbol and the fact that $A \in \Psi_\hbar^{-\infty}$.

Example 11.5. (Defect measure of plane wave.) *Let $u^I(x) := \exp(\mathrm{i}\langle x, \widehat{a}\rangle)/\hbar$ with $|\widehat{a}| = 1$ (note this in not in $L^2(\mathbb{R}^d)$, but is in $L^2_{\mathrm{loc}}(\mathbb{R}^d)$). Then*

$$\begin{aligned}
&\langle \mathrm{Op}_\hbar(b)u^I, u^I\rangle \\
&= \frac{1}{(2\pi\hbar)^d} \int_{\mathbb{R}^d} \exp\big(-\mathrm{i}\langle x, \widehat{a}\rangle)/\hbar\big) \left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \exp\big(\mathrm{i}\langle x-y, \xi\rangle)/\hbar\big) \exp\big(\mathrm{i}\langle y, \widehat{a}\rangle)/\hbar\big)\, b(x,\xi)\,\mathrm{d}y\,\mathrm{d}\xi\right)\mathrm{d}x \\
&= \int_{\mathbb{R}^d} \exp\big(-\mathrm{i}\langle x, \widehat{a}\rangle\big)/\hbar\Big(\int_{\mathbb{R}^d} \delta(\xi-\widehat{a})\, \exp\big(\mathrm{i}\langle y, \widehat{a}\rangle)/\hbar\big)\, b(x,\xi)\,\mathrm{d}\xi\Big)\mathrm{d}x \\
&= \int_{\mathbb{R}^d} b(x,\widehat{a})\,\mathrm{d}x \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} b(x,\xi)\,\delta(\xi-\widehat{a})\,\mathrm{d}x\,\mathrm{d}\xi,
\end{aligned}$$

where we have first performed the y integral to obtain $\delta(\xi - \widehat{a})$. This calculation shows that, for any $\hbar_n \to 0$, u^I has defect measure equal to the product of $\delta(\xi - \widehat{a})$ and Lebesgue measure in x.

In the above calculation we used $\mathrm{Op}_\hbar$ and not $\mathrm{Op}_\hbar^{\mathrm{ps}}$. Unlike above, $\langle \mathrm{Op}_\hbar(b)u^I, u^I\rangle$ cannot be calculated exactly (because of the $\chi_0(|x-y|)$ in the integrand), but its asymptotics as $\hbar \to 0$ can be obtained by using stationary phase. Indeed, after changing variables $\xi \mapsto \xi - \widehat{a}$ and $y \mapsto y - x$, the integral in y and ξ is of the form considered in Corollary 9.6. The asymptotics (9.40) then imply that

$$\left\langle \mathrm{Op}_\hbar^{\mathrm{ps}}(b)u^I, u^I\right\rangle = \int_{\mathbb{R}^d} b(x,\widehat{a})\,\mathrm{d}x + O(\hbar) \qquad \text{as } \hbar \to 0.$$

Theorem 11.6. (Existence of defect measures.) *Suppose that $\{u(\hbar)\}_{0<\hbar\leq\hbar_0}$ is uniformly locally bounded and $\hbar_n \to 0$. Then there exists a subsequence $\{\hbar_{n_\ell}\}_{\ell=1}^\infty$ and a Radon measure μ on $T^*\mathbb{R}^d$ such that $\{u(\hbar_{n_\ell})\}_{\ell=1}^\infty$ has defect measure μ.*

Remark 11.7. (Our use of Radon measures.) *For the precise definition of a Radon measure, see, e.g., [54, Chapter 1], [58, Chapter 7]. The only facts we use about such measures on $\mathbb{R}^d$ in what follows are the following.*

(i) Given a bounded, positive functional on $C_{\mathrm{comp}}(\mathbb{R}^d)$ there exists a Radon measure such that the action of the functional is integration against the measure – this is a version of the Riesz representation theorem (see, e.g., [54, §1.8], [58, §7.2]), and is used in the proof of Theorem 11.6.

(ii) If μ is a Radon measure on $\mathbb{R}^d$, then $C_{\mathrm{comp}}(\mathbb{R}^d)$ is dense in $L^1(\mathbb{R}^d;\mu)$ [58, Prop. 7.9].

Lemma 11.8. (Positivity of defect measure.) *Suppose that $\hbar_n \to 0$ and $\{u(\hbar_n)\}_{n=1}^\infty$ is uniformly locally bounded with defect measure μ. Then μ is positive; i.e., if $a \geq 0$, then $\int a\,\mathrm{d}\mu \geq 0$.*

11.3.2 Proofs of Theorem 11.6 and Lemma 11.8

11.4 Defect measures and PDEs

Lemma 11.9. (If $Pu = 0$, then defect measure supported in $\{\sigma_\hbar(P) = 0\}$.) *Let $P \in \Psi^m_\hbar$ be properly supported. Suppose that $\{u(\hbar_n)\}$ has defect measure μ, and satisfies*

$$\|Pu(\hbar_n)\|_{L^2(\mathbb{R}^d)} \to 0 \quad \text{as } n \to \infty.$$

Then $\mu(\{\sigma_\hbar(P) \neq 0\}) = 0$; i.e., if $\operatorname{supp} a \subset \{\sigma_\hbar(P) \neq 0\}$, then $\int a \, \mathrm{d}\mu = 0$.

The interpretation of this result is that solutions of $Pu = 0$ "live" in $\{\sigma_\hbar(P) = 0\}$. Two examples:

(i) we saw in Example 11.5 that the defect measure of a plane wave was supported in $\{|\xi|^2 - 1 = 0\}$, i.e., $\{\sigma_\hbar(-\hbar^2\Delta - 1) = 0\}$, and

(ii) if $p(\hbar D)u = 0$, where $p(\hbar D)$ is a semiclassical Fourier multiplier (as in §5.5), then $p(\xi)(\mathcal{F}_\hbar u)(\xi) = 0$, and thus $\operatorname{supp}(\mathcal{F}_\hbar u) \subset \{p(\xi) = 0\}$.

Recall that in §10.2 we stated that the flow with $\sigma_\hbar(P) = 0$ plays a special role; Lemma 11.9 is the reason for this.

Proof. If we can show that

$$\int b \, \sigma_\hbar(P) \, \mathrm{d}\mu = 0 \quad \text{for all } b \in S^{\mathrm{comp}}, \tag{11.6}$$

then the result follows. Indeed, given $a \in S^{\mathrm{comp}}$ with $\operatorname{supp} a \subset \{\sigma_\hbar(P) \neq 0\}$, let $b := a/\sigma_\hbar(P)$ (which is in S^{comp}) and apply (11.6).

Idea of the rest of the proof: first show that $\langle \operatorname{Op}^{\mathrm{ps}}_\hbar(b)\, Pu, u\rangle \to 0$ using that $Pu \to 0$ and $\operatorname{Op}^{\mathrm{ps}}_\hbar(b)$ is bounded, then use (11.5) to show that $\langle \operatorname{Op}^{\mathrm{ps}}_\hbar(b)\, Pu, u\rangle \to \int b\, \sigma_\hbar(P)\, \mathrm{d}\mu$. The difficulty comes in dealing with the cut-off functions needed to deal with $u \in L^2_{\mathrm{loc}}(\mathbb{R}^d)$ (if $u \in L^2(\mathbb{R}^d)$ then the proof is simpler; see [59, Lemma 4.4]).

By Lemma 9.3, if $a \in S^{\mathrm{comp}}$ then $\operatorname{Op}_\hbar(a)$ is $L^2 \to L^2$ bounded, uniformly in $\hbar$. Using this and Lemma 11.3, we have that, on the one hand,

$$\left|\langle \operatorname{Op}^{\mathrm{ps}}_\hbar(b)\, Pu, u\rangle\right| = \left|\langle \chi_1 \operatorname{Op}^{\mathrm{ps}}_\hbar(b)\chi_2\, Pu, u\rangle\right| \leq C \|Pu\|_{L^2(\mathbb{R}^d)} \|\chi_1 u\|_{L^2(\mathbb{R}^d)} \to 0 \quad \text{as } n \to \infty. \tag{11.7}$$

On the other hand, by the composition formula (Theorem 9.4) and the definition of $\sigma_\hbar$ (Definition 7.7), $\operatorname{Op}_\hbar(b)P = \operatorname{Op}_\hbar(b\,\sigma_\hbar(P)) + \hbar E_1$, where $E_1 \in \Psi^{-\infty}_\hbar$. Therefore, by (11.2),

$$\operatorname{Op}^{\mathrm{ps}}_\hbar(b)P = \operatorname{Op}^{\mathrm{ps}}_\hbar(b\,\sigma_\hbar(P)) + \hbar E_2, \quad \text{where } E_2 \in \Psi^{-\infty}_\hbar.$$

Since $\operatorname{Op}^{\mathrm{ps}}_\hbar(b)$ and P are properly supported and $b \in S^{\mathrm{comp}}$, Part (iv) of Lemma 7.15 and Lemma 11.3 imply that there exist $\chi_1 \in \mathcal{D}$, with $\chi_1 \equiv 1$ on $\pi_x(\operatorname{supp} b)$, and $\chi_2 \in \mathcal{D}$ such that $\operatorname{Op}^{\mathrm{ps}}_\hbar(b)P = \chi_1 \operatorname{Op}^{\mathrm{ps}}_\hbar(b)\, P\chi_2$. Inspecting the proof of Lemma 11.3, we see that

$$\chi_2(y) = 1 \quad \text{on} \quad \big\{y : \operatorname{dist}(y, \pi_x(\operatorname{supp} b)) \leq 1/2\big\} \tag{11.8}$$

(in fact χ_2 will be equal to one on a larger set). Therefore,

$$\begin{aligned}
\langle \operatorname{Op}^{\mathrm{ps}}_\hbar(b)\, Pu, u\rangle = \langle \chi_1 \operatorname{Op}^{\mathrm{ps}}_\hbar(b)\, P\chi_2 u, u\rangle &= \Big\langle \chi_1 \Big(\operatorname{Op}^{\mathrm{ps}}_\hbar\big(b\,\sigma_\hbar(P)\big) + \hbar E_2\Big)\chi_2 u, u\Big\rangle \\
&= \Big\langle \chi_1 \operatorname{Op}^{\mathrm{ps}}_\hbar\big(b\,\sigma_\hbar(P)\big)\chi_2 u, u\Big\rangle + O(\hbar_n), \\
&= \Big\langle \operatorname{Op}^{\mathrm{ps}}_\hbar\big(b\,\sigma_\hbar(P)\big)u, u\Big\rangle + O(\hbar_n),
\end{aligned} \tag{11.9}$$

where we have used that $E_2 : L^2 \to L^2$ is bounded uniformly in $\hbar$, that u is uniformly locally bounded, that $\chi_1 \equiv 1$ on $\pi_x(\operatorname{supp} b)$ (and hence on $\pi_x(\operatorname{supp}(b\,\sigma_\hbar(P)))$), and (11.8). The equation (11.6) then follows by combining (11.5), (11.7), and (11.9), and the proof is complete. □

Theorem 11.10. (Invariance of defect measure under the flow.) *Suppose that $P \in \Psi^m_\hbar$ is properly supported and formally self adjoint, $\{u(\hbar_n)\}$ has defect measure μ, and*

$$\|Pu(\hbar_n)\|_{L^2(\mathbb{R}^d)} = o(h_n) \quad \text{as } n \to \infty. \tag{11.10}$$

Then

$$\int \{\sigma_\hbar(P), a\}\, \mathrm{d}\mu = 0 \quad \text{for all } a \in S^{\mathrm{comp}}. \tag{11.11}$$

Interpretation of (11.11): By (10.4),

$$\int \{\sigma_\hbar(P), a\}\, \mathrm{d}\mu = \int \frac{\mathrm{d}}{\mathrm{d}t}(a \circ \varphi_t)\, \mathrm{d}\mu = \frac{\mathrm{d}}{\mathrm{d}t} \int (a \circ \varphi_t)\, \mathrm{d}\mu, \tag{11.12}$$

so that (11.11) is the statement that $\int (a \circ \varphi_t)\, \mathrm{d}\mu$ is constant as a function of t; i.e., the defect measure is invariant under the flow.

Corollary 11.11. (Invariance under the flow written in terms of sets.) *Given a Borel set $B \subset T^*\mathbb{R}^d$,*

$$\mu\big(\varphi_t(B)\big) = \mu(B) \quad \text{for all } t, \qquad \text{i.e.,} \qquad \int 1_{\varphi_t(B)}\, \mathrm{d}\mu = \int 1_B\, \mathrm{d}\mu \quad \text{for all } t \tag{11.13}$$

Idea of the proof of Theorem 11.10. The idea is that, by (7.10),

$$\{\sigma_\hbar(P), a\} = \sigma_\hbar\left(\frac{\mathrm{i}}{\hbar}[P, \mathrm{Op}^{\mathrm{ps}}_\hbar(a)]\right).$$

If $u \in L^2(\mathbb{R}^d)$, then, by the formal self-adjointness of P, $L^2 \to L^2$ boundedness of $\mathrm{Op}^{\mathrm{ps}}_\hbar(a)$, and (11.10),

$$\begin{aligned} \big\langle [P, \mathrm{Op}^{\mathrm{ps}}_\hbar(a)]u, u\big\rangle &= \big\langle P\, \mathrm{Op}^{\mathrm{ps}}_\hbar(a)u, u\big\rangle - \big\langle\, \mathrm{Op}^{\mathrm{ps}}_\hbar(a)Pu, u\big\rangle \\ &= \big\langle\, \mathrm{Op}^{\mathrm{ps}}_\hbar(a)u, Pu\big\rangle - \big\langle\, \mathrm{Op}^{\mathrm{ps}}_\hbar(a)Pu, u\big\rangle = o(\hbar); \end{aligned}$$

the result then follows from (11.5). The difficultly in the proof comes from dealing with the contributions from the cut-offs we need to insert to deal with the fact that u is only in $L^2_{\mathrm{loc}}(\mathbb{R}^d)$ (and not $L^2(\mathbb{R}^d)$).

Proof of Theorem 11.10. Since $a \in S^{\mathrm{comp}}$ and $P \in \Psi^m$, $b := \{\sigma_\hbar(P), a\} \in S^{\mathrm{comp}}$. By (11.5),

$$\int \{\sigma_\hbar(P), a\}\, \mathrm{d}\mu = \lim_{n\to\infty} \big\langle\, \mathrm{Op}^{\mathrm{ps}}_\hbar(b)u, u\big\rangle; \tag{11.14}$$

our goal is to prove that the limit on the right-hand side of (11.14) is zero.

By (7.10) and the definition of the principal symbol,

$$\mathrm{Op}_\hbar(b) - \frac{\mathrm{i}}{\hbar}[P, \mathrm{Op}_\hbar(a)] \in \hbar\Psi^{-\infty}_\hbar,$$

so that, by (11.2),

$$\mathrm{Op}^{\mathrm{ps}}_\hbar(b) = \frac{\mathrm{i}}{\hbar}[P, \mathrm{Op}^{\mathrm{ps}}_\hbar(a)] + \hbar E_1, \text{ with } E_1 \in \Psi^{-\infty}_\hbar. \tag{11.15}$$

Then, by Lemma 11.3 there exist $\chi_1, \chi_2 \in \mathcal{D}$ with $\chi_1 \equiv \chi_2 \equiv 1$ on $\operatorname{supp} a$ such that

$$\big\langle\, \mathrm{Op}^{\mathrm{ps}}_\hbar(b)u, u\big\rangle = \big\langle \chi_1\, \mathrm{Op}^{\mathrm{ps}}_\hbar(b)\chi_2 u, u\big\rangle = \frac{\mathrm{i}}{\hbar}\big\langle \chi_1[P, \mathrm{Op}^{\mathrm{ps}}_\hbar(a)]\chi_2 u, u\big\rangle + \hbar\big\langle \chi_1 E_1 \chi_2 u, u\big\rangle.$$

Since E_1 is uniformly bounded $L^2 \to L^2$, to prove that the limit on the right-hand side of (11.14) is zero, it is sufficient to prove that

$$\big|\big\langle \chi_1[P, \mathrm{Op}^{\mathrm{ps}}_\hbar(a)]\chi_2 u, u\big\rangle\big| = o(\hbar_n) \quad \text{as } n \to \infty.$$

Now

$$\begin{aligned}\big\langle \chi_1[P,\operatorname{Op}_\hbar^{\mathrm{ps}}(a)]\chi_2 u,u\big\rangle &= \big\langle \chi_1 P\operatorname{Op}_\hbar^{\mathrm{ps}}(a)\chi_2 u,u\big\rangle - \big\langle \chi_1\operatorname{Op}_\hbar^{\mathrm{ps}}(a)P\chi_2 u,u\big\rangle\\ &= \big\langle \operatorname{Op}_\hbar^{\mathrm{ps}}(a)\chi_2 u,[P,\chi_1]u+\chi_1 Pu\big\rangle - \big\langle \chi_1\operatorname{Op}_\hbar^{\mathrm{ps}}(a)\big([P,\chi_2]u+\chi_2 Pu\big),u\big\rangle.\end{aligned}$$

Therefore, since u is uniformly locally bounded, $\operatorname{Op}_\hbar(a)$ is uniformly bounded $L^2\to L^2$, and (11.10) holds, it is sufficient to prove that

$$\big|\big\langle \operatorname{Op}_\hbar^{\mathrm{ps}}(a)\chi_2 u,[P,\chi_1]u\big\rangle\big| + \big|\big\langle \chi_1\operatorname{Op}_\hbar^{\mathrm{ps}}(a)[P,\chi_2]u,u\big\rangle\big| = o(\hbar_n)\quad \text{as } n\to\infty,$$

and this follows if

$$\big\|(\operatorname{Op}_\hbar^{\mathrm{ps}}(a))^*[P,\chi_1]\big\|_{L^2(\mathbb{R}^d)\to L^2(\mathbb{R}^d)} = o(\hbar)\quad\text{and}\quad \big\|\operatorname{Op}_\hbar^{\mathrm{ps}}(a)[P,\chi_2]\big\|_{L^2(\mathbb{R}^d)\to L^2(\mathbb{R}^d)} = o(\hbar). \tag{11.16}$$

By Part (iii) of Theorem 7.5, a commutator is $O(\hbar)$; this is not quite enough, but we show below that the fact that $\pi_x(\operatorname{supp} a)$ and $\operatorname{supp}\nabla\chi_j$ are disjoint gives faster decay.

By the definition of the principal symbol and the fact that $\chi_2\in\Psi_\hbar^{-\infty}$ (since it has compact support),

$$\frac{\mathrm{i}}{\hbar}[P,\chi_2] = \operatorname{Op}_\hbar\big(\{\sigma_\hbar(P),\chi_2\}\big)+\hbar E_2\quad\text{for } E_2\in\Psi_\hbar^{-\infty}. \tag{11.17}$$

Therefore, by the composition formula (9.33),

$$\operatorname{Op}_\hbar(a)\frac{\mathrm{i}}{\hbar}[P,\chi_2] = \operatorname{Op}_\hbar\big(a\#\{\sigma_\hbar(P),\chi_2\}\big)+\hbar\operatorname{Op}_\hbar(a)E_2.$$

Since $\operatorname{supp} a\cap\operatorname{supp}\nabla\chi_2=\emptyset$, the definition of $\#$ (9.34) implies that

$$a\#\{\sigma_\hbar(P),\chi_2\}\in\hbar^\infty S^{-\infty}; \tag{11.18}$$

the second bound in (11.16) then follows from combining the last two displayed equations and using the composition and mapping properties of Theorem 7.5 (to show that $\operatorname{Op}_\hbar(a)E_2$ is uniformly bounded $L^2\to L^2$).

The first bound in (11.16) follows similarly, except that we use the adjoint expansion (9.47) to deal with $(\operatorname{Op}_\hbar(a))^*$. Indeed, by (9.47),

$$(\operatorname{Op}_\hbar(a))^* = \operatorname{Op}_\hbar(\overline{a})+\hbar\operatorname{Op}_\hbar(\overline{a_1})+\hbar^2E_3,$$

where $\operatorname{supp} a_1\subset\operatorname{supp} a$ and $E_3\in\Psi_\hbar^{-\infty}$. Therefore, by similar arguments to those in (11.17) and (11.18),

$$\begin{aligned}(\operatorname{Op}_\hbar(a))^*\frac{\mathrm{i}}{\hbar}[P,\chi_1] &= \Big(\operatorname{Op}_\hbar(\overline{a})+\hbar\operatorname{Op}_\hbar(\overline{a_1})+\hbar^2E_3\Big)\Big(\operatorname{Op}_\hbar\big(\{\sigma_\hbar(P),\chi_2\}\big)+\hbar E_4\Big)\\ &= \hbar\Big(\operatorname{Op}_\hbar(\overline{a})+\hbar\operatorname{Op}_\hbar(\overline{a_1})\Big)E_4+\hbar^2E_5,\end{aligned}$$

where $E_4,E_5\in\Psi_\hbar^{-\infty}$; the first bound in (11.16) therefore follows from the composition and mapping properties of Theorem 7.5 (similar to in the proof of the second bound in (11.16)). □

Proof of Corollary 11.11. By (11.11) and (11.12),

$$\int (b\circ\varphi_s)(\rho)\,\mathrm{d}\mu = \int b(\rho)\,\mathrm{d}\mu\quad\text{for all } s.$$

By approximating 1_B by smooth symbols and using that $C_{\mathrm{comp}}(\mathbb{R}^d)$ is dense in $L^1(\mathbb{R}^d;\mu)$ when μ is a Radon measure (see, e.g., [58, Prop. 7.9], we have that

$$\int 1_B\big(\varphi_s(\rho)\big)\,\mathrm{d}\mu = \int 1_B(\rho)\,\mathrm{d}\mu\quad\text{for all } s.$$

Now $\varphi_s(\rho)\in B$ iff $\rho\in\varphi_{-s}(B)$, so that $1_B(\varphi_s(\rho)) = 1_{\varphi_{-s}(B)}(\rho)$. The result (11.13) then follows by letting $s=-t$. □

11.5 Defect measures of outgoing Helmholtz solutions

We first recap results about the outgoing solution of $(-\hbar^2\Delta - 1)u = f$ in $\mathbb{R}^d$ (where f has compact support); i.e., the so-called *free resolvent*. Let

$$\big(R_0(\hbar)f\big)(x) := \int_{\mathbb{R}^d} \Phi_\hbar(x,y)\, f(y)\, \mathrm{d}y, \tag{11.19}$$

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

$$\Phi_\hbar(x,y) := \hbar^{-2}\frac{\mathrm{i}}{4}\left(\frac{\hbar^{-1}}{2\pi|x-y|}\right)^{(d-2)/2} H^{(1)}_{d/2-1}\big(\hbar^{-1}|x-y|\big); \tag{11.20}$$

see, e.g., [106, Theorem 9.4] (note that (11.20) is $\hbar^{-2}$ multiplied by the outgoing fundamental solution of $(-\Delta - k^2)u = f$). The definition of the Hankel function $H^{(1)}_{d/2-1}$ implies that

$$\Phi_\hbar(x,y) = \begin{cases} \hbar^{-2}\dfrac{\mathrm{i}}{4}H^{(1)}_0\big(\hbar^{-1}|x-y|\big), & d=2,\\[2ex] \hbar^{-2}\dfrac{\exp\big(\mathrm{i}\hbar^{-1}|x-y|\big)}{4\pi|x-y|}, & d=3. \end{cases}$$

Lemma 11.12. (Properties of the free resolvent.) *For all $\hbar > 0$,*

$$R_0(\hbar) : L^2_{\mathrm{comp}}(\mathbb{R}^d) \to H^2_{\mathrm{loc}}(\mathbb{R}^d).$$

Furthermore, given $\chi \in C^\infty_{\mathrm{comp}}(\mathbb{R}^d)$ and $\hbar_0 > 0$, there exists $C > 0$ such that

$$\|\chi R_0(\hbar)\chi\|_{L^2(\mathbb{R}^d)\to L^2(\mathbb{R}^d)} \leq \frac{C}{\hbar} \quad \text{for all } 0 < \hbar \leq \hbar_0. \tag{11.21}$$

Proof. Boundedness from $L^2_{\mathrm{comp}}(\mathbb{R}^d) \to H^1_{\mathrm{loc}}(\mathbb{R}^d)$ and the bound (11.21) both follow from Theorem 2.14; recall that this theorem is proved by integration by parts using the Morawetz multiplier (see Exercise 3 in §2.5). Boundedness from $L^2_{\mathrm{comp}}(\mathbb{R}^d) \to H^2_{\mathrm{loc}}(\mathbb{R}^d)$ can then be obtained by elliptic regularity (see §2.3). Alternatively, boundedness and (11.21) can (at least in odd dimensions) be proved by expressing R_0 in terms of the wave propagator; see [52, Theorem 3.1]. □

Remark 11.13. (Why is the Helmholtz solution operator called the "resolvent"?) *Recall that the* resolvent *of an operator P is the operator $(P-z)^{-1}$, and the set of z such that this inverse is exists is the* resolvent set. *If $P = -\Delta$ and $z = k^2$, then $(P-z)^{-1} = (-\Delta - k^2)^{-1}$; the Helmholtz solution operator is therefore often called the resolvent (of the Laplacian).*

Theorem 11.14. (The "outgoing" condition at the level of defect measures.) *Given $\hbar_n \to 0$ and $\{g(\hbar_n)\}_{n=1}^\infty$ with $\operatorname{supp} g(\hbar) \subset B_R$, let $u(\hbar_n) := R_0(\hbar_n)g(\hbar_n)$. If $u(\hbar_n)$ has defect measure μ and*

$$\mathcal{I} := \Big\{(x,\xi) : |x| > R+1,\ \langle x,\xi\rangle < 0\Big\},$$

then $\mu(\mathcal{I}) = 0$.

Observe that $\mathcal{I}$ consists of "incoming" directions to the ball B_R (and thus $\mathcal{I}$ is sometimes called the *directly-incoming set* [64, Lemma 3.4], [60, §3.1]). Therefore, Theorem 11.14 (first given as [26, Prop. 3.5]) expresses the fact that the mass in phase space of Helmholtz equations satisfying the Sommerfeld radiation condition (1.4) is concentrated in "outgoing"' directions.

To prove Theorem 11.14 we need the following lemma, which can be proved by integration by parts (see Exercise 1 in §11.6).

Lemma 11.15. (Asymptotics of oscillatory integral with no stationary points.) *Given $\varphi \in C^\infty(\mathbb{R}^n)$, $a \in \mathscr{S}(\mathbb{R}^n)$, let*

$$I(\hbar) := \int_{\mathbb{R}^n} \exp\big(\mathrm{i}\varphi(x)/\hbar\big)\, a(x)\, \mathrm{d}x. \tag{11.22}$$

If $\nabla\phi \neq 0$ on $\operatorname{supp} a$, then

$$I(\hbar) = O(\hbar^\infty).$$

(Compare to Theorem 9.5.)

Proof of Theorem 11.14. We'll prove this result for $d = 3$. The proof for general $d \geq 2$ is similar, using asymptotics of Hankel functions to write the fundamental solution $\Phi_\hbar(x, y)$ as $\exp(\mathrm{i}|x-y|/\hbar)$ multiplied by a (weakly-singular) function.

Step 1: consider $\mathrm{Op}_\hbar^{\mathrm{ps}}(a)R_0 g$ *with the symbol* a *concentrated near* $(x_0, \xi_0) \in \mathcal{I}$.
By approximating indicator functions by smooth functions (as in the proof of Corollary 11.11), it is sufficient to prove that, given $(x_0, \xi_0) \in \mathcal{I}$, $\int a \, \mathrm{d}\mu = 0$ for $a \in S^{\mathrm{comp}}$ such that $a \geq 0$, $a(x, \xi) = 1$ in an open neighbourhood of (x_0, ξ_0), and $\operatorname{supp} a \subset \mathcal{I}$.

By (11.5), the result that $\int a \, \mathrm{d}\mu = 0$ follows if we can show that $(\mathrm{Op}_\hbar^{\mathrm{ps}}(a))R_0(\hbar_n)g(\hbar_n) = o(1)$ as $n \to \infty$; in fact we show that

$$\mathrm{Op}_\hbar^{\mathrm{ps}}(a)R_0(\hbar_n)g(\hbar_n) = O(\hbar_n^\infty) \quad \text{as } n \to \infty. \tag{11.23}$$

By the definitions of $\mathrm{Op}_\hbar^{\mathrm{ps}}$ (11.1) and $R_0(\hbar)$ (11.19),

$$\begin{aligned}
&\big(\mathrm{Op}_\hbar^{\mathrm{ps}}(a)R_0(\hbar_n)g(\hbar_n)\big)(x) \\
&\quad = \frac{1}{(2\pi\hbar)^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \exp\left(\frac{\mathrm{i}\langle x-y, \xi\rangle}{\hbar}\right) a(x,\xi) \big(R_0(\hbar)g(\hbar)\big)(y) \, \chi_0\big(|x-y|\big) \, \mathrm{d}y \, \mathrm{d}\xi \\
&\quad = \frac{\hbar^{-2}}{(2\pi\hbar)^d} \int_{\operatorname{supp} f} \left(\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \exp\left(\frac{\mathrm{i}\big(\langle x-y, \xi\rangle + |y-z|\big)}{\hbar}\right) a(x,\xi) \frac{g(\hbar)(z)}{4\pi|y-z|} \, \chi_0\big(|x-y|\big) \, \mathrm{d}y \, \mathrm{d}\xi \right) \mathrm{d}z,
\end{aligned} \tag{11.24}$$

We now show that $|y - z|$ is bounded below on the support of the integrand, and thus, since, in addition, a, f, and χ_0 all have compact support, the integral in (11.24) exists as a standard integral (i.e., it does not need to be understood as an oscillatory integral in the sense of TK). Without loss of generality, we can assume that the cut-off function χ_0 in the definition of $\mathrm{Op}_\hbar^{\mathrm{ps}}$ (11.1) satisfies $\chi_0(t) = 0$ for $t \geq 1/2$. Therefore $|x - y| \leq 1/2$ in the integrand of (11.24). Since $z \in \operatorname{supp} f$, $|z| \geq R$, and by assumption $|x| > R + 1$ on $\operatorname{supp} a$; therefore $|y - z| \geq 1/2$.

Step 2: show the phase has no stationary points, and integrate by parts using Lemma 11.15.
The integral (11.24) is of the form (11.22) with

$$\varphi(y, \xi) := \langle x - y, \xi\rangle + |y - z|.$$

Then

$$\partial_{y_j}\varphi = -\xi_j + \frac{y_j - z_j}{|y-z|} \quad \text{and} \quad \partial_{\xi_j}\varphi = x_j - y_j,$$

and $\nabla\varphi = 0$ iff $\xi = (x - z)/|x - z|$. This cannot hold, however, since $(x - z)/|x - z|$ is pointing "outwards" from B_R and ξ is pointing "inwards". Therefore, Lemma 11.15 implies the asymptotics (11.23), and the proof is complete. □

Remark 11.16. (Wavefront set of a family of functions.) *The operator wavefront set of Definition 7.9 measures where in phase space a semiclassical pseudodifferential operator is non-negligible. A related concept is the semiclassical wavefront set of a family of functions, which measures where in phase space a family of functions is supported in the $\hbar \to 0$ limit (see [155, §8.4.2], [52, Def. E.36]); defect measures then measure "how concentrated" the family is at these locations.*

The reason we mention this concept here is that the asymptotics (11.23) actually show that $\mathcal{I} \not\subset \mathrm{WF}_\hbar(R_0 f)$; furthermore, the ideas behind the proof of (11.23) can be used to calculate the wavefront set of a general oscillatory integral (i.e., not just $R_0 f$); see [52, Prop. E.37].

11.6 Exercises for §11

1. Prove Lemma 11.15. (Hint: construct L such that $L(\exp(\mathrm{i}\varphi(x)/\hbar)) = \exp(\mathrm{i}\varphi(x)/\hbar)$.)

12 Proof of Theorem 2.7 (bound on solution operator under nontrapping) using defect measures

We now prove Theorem 2.7 using defect measures, with this proof due to Burq [26, Theorem 2]. We assume throughout this section that $\Omega_- = \emptyset$; the same ideas can be used to prove the result for nontrapping Ω_-, although the proof is more technical due to the presence of a boundary – see [26, §4], [64, §4].

Let $P(\hbar) := -\hbar^2\nabla\cdot(A\nabla\cdot) - n$, $P_0(\hbar) := -\hbar^2\Delta - 1$. As in §11.5, let $R_0(\hbar)$ denote the free resolvent.

Lemma 12.1. (Outgoing Helmholtz solutions and the free resolvent.) *Let $R > 0$ be such that $\operatorname{supp}(I-A)\cup\operatorname{supp}(1-n) \Subset B_R$. Given $\{f(\hbar)\}_{0<\hbar\leq\hbar_0}$ with $\operatorname{supp} f(\hbar)$ contained inside an $\hbar$-independent compact set, let $u(\hbar)$ be the outgoing solution to $P(\hbar)u(\hbar) = \hbar f(\hbar)$.*

(i) If $\chi_1 \in C^\infty_{\mathrm{comp}}(\mathbb{R}^d)$ with $B_R \Subset \{\chi_1 \equiv 1\}$, then

$$(1-\chi_1)u = R_0(\hbar)\Big((1-\chi_1)\hbar f - [P_0(\hbar),\chi_1]u\Big) \tag{12.1}$$

(i.e., outside the support of the scatterer, u can be written as a free resolvent).

(ii) Suppose further that $\|f(\hbar)\|_{L^2} \lesssim 1$ and there exists $\chi \in C^\infty_{\mathrm{comp}}(\mathbb{R}^d)$ with $B_R \Subset \{\chi \equiv 1\}$ and $C > 0$ such that

$$\|\chi u\|_{L^2(\mathbb{R}^d)} \leq C \quad \text{for all } 0 < \hbar \leq \hbar_0.$$

Then, given $\widetilde{\chi} \in C^\infty_{\mathrm{comp}}(\mathbb{R}^d)$ there exists $\widetilde{C} > 0$ such that

$$\|\widetilde{\chi}u\|_{L^2(\mathbb{R}^d)} \leq \widetilde{C} \quad \text{for all } 0 < \hbar \leq \hbar_0$$

(i.e., if u is locally bounded for one cut-off function, then it is locally bounded for all cut-off functions).

Remark 12.2. (Alternative definition of outgoing via free resolvent.) *Part (i) of Lemma 12.1 shows that, at least outside the support of the scatterer, an outgoing Helmholtz solution can be written as a free resolvent. Being expressible as a free resolvent is sometimes used as the* definition *of outgoing (instead of the Sommerfeld radiation condition (1.4)); see, e.g., [52, Definition 3.32], [97, Equation 4.19].*

Proof of Lemma 12.1. (i) Since $B_R \Subset \{\chi_1 \equiv 1\}$,

$$P_0\big((1-\chi_1)u\big) = P\big((1-\chi_1)u\big) = (1-\chi_1)\hbar f + [P, 1-\chi_1]u = (1-\chi_1)\hbar f + [P_0, 1-\chi_1]u$$

(where we have suppressed the dependence of P and P_0 on $\hbar$ for brevity). Since the right-hand side of this last equation has compact support, the result (12.1) then follows by applying R_0.

(ii) Given χ as in the statement of the result, let $\chi_1 \in C^\infty_{\mathrm{comp}}(\mathbb{R}^d)$ be such that $\operatorname{supp}\chi_1 \Subset \{\chi \equiv 1\}$ and $B_R \Subset \{\chi_1 \equiv 1\}$ (i.e., the support of the derivatives of χ_1 is between B_R and $\{\chi \equiv 1\}$).

Given $\widetilde{\chi}$, since

$$\widetilde{\chi}u = \widetilde{\chi}(1-\chi)u + \widetilde{\chi}\chi u \quad \text{and} \quad \big\|\widetilde{\chi}(1-\chi)u\big\|_{L^2} \leq \big\|\widetilde{\chi}(1-\chi_1)u\big\|_{L^2},$$

it is sufficient to prove that $\|\widetilde{\chi}(1-\chi)1)u\|_{L^2} \lesssim 1$.

By (12.1) (i.e., the result of Part (i)),

$$\widetilde{\chi}(1-\chi_1)u = \widetilde{\chi}R_0(\hbar)\Big((1-\chi_1)\hbar f + [\hbar^2\Delta,\chi_1]u\Big) = \widetilde{\chi}R_0(\hbar)g$$

where

$$g := (1-\chi_1)\hbar f + 2\hbar^2\nabla u\cdot\nabla\chi_1 + \hbar^2(\Delta\chi_1)u. \tag{12.2}$$

By the bound (11.21) on $R_0(\hbar)$, it is sufficient to prove that $\|g\|_{L^2} \lesssim \hbar$. Since $\|f\|_{L^2} \lesssim 1$, it is sufficient to prove that $\|(\Delta\chi_1)u\|_{L^2} \lesssim \hbar^{-1}$ and $\|\hbar\nabla u\cdot\nabla\chi_1\|_{L^2} \lesssim 1$. Since $\operatorname{supp}(\Delta\chi_1) \subset \{\chi \equiv 1\}$, $\|(\Delta\chi_1)u\|_{L^2} \leq \|\chi u\|_{L^2} \leq C$. Now, since $\|Pu\|_{L^2} \lesssim \hbar$ and $\|\chi u\|_{L^2} \leq C$, $\|\hbar\nabla u\cdot\nabla\chi_1\|_{L^2} \lesssim 1$ by Part (i) of Lemma 2.18. □

Proof of Theorem 2.7. Let $R(\hbar)$ be the resolvent, i.e.,

$$R(\hbar) : L^2_{\text{comp}}(\mathbb{R}^d) \to H^2_{\text{loc}}(\mathbb{R}^d)$$

with both $P(\hbar)R(\hbar)f = f$ and $R(\hbar)f$ outgoing for $f \in L^2_{\text{comp}}(\mathbb{R}^d)$.

The bound (2.4) is therefore equivalent to: given $\chi \in C^\infty_{\text{comp}}(\mathbb{R}^d)$ there exists $\hbar_0 > 0$ and $C > 0$ such that for all $0 < \hbar \le \hbar_0$ and all $f \in L^2(\mathbb{R}^d)$

$$\|\chi R(\hbar)\chi f\|_{L^2(\mathbb{R}^d)} \le \frac{C}{\hbar}\|f\|_{L^2(\mathbb{R}^d)}.$$

Let $R > 0$ be such that $\operatorname{supp}(I-A)\cup\operatorname{supp}(1-n) \Subset B_R$. Without loss of generality we can assume that $\{\chi \equiv 1\} \Supset B_R$. (If we prove the bound for such a χ, then the bound for a smaller χ follows easily.)

Seeking a contradiction, we assume that the bound does *not* hold; i.e., for all $\hbar_0$ and C there exists an $f \in L^2(\mathbb{R}^d)$ and $0 < \hbar \le \hbar_0$ such that

$$\|\chi R(\hbar)\chi f\|_{L^2(\mathbb{R}^d)} > \frac{C}{\hbar}\|f\|_{L^2(\mathbb{R}^d)}.$$

Choosing $C = n$, we see that for all n there exists $\hbar_n \to 0$ and $f_n \in L^2(\mathbb{R}^d)$ such that

$$\|\chi R(\hbar_n)\chi f_n\|_{L^2(\mathbb{R}^d)} > \frac{n}{\hbar_n}\|f_n\|_{L^2(\mathbb{R}^d)}.$$

We now divide f_n by a constant so that

$$\|\chi R(\hbar_n)\chi f_n\|_{L^2(\mathbb{R}^d)} = \frac{1}{\hbar_n} \quad \text{and} \quad \|f_n\|_{L^2(\mathbb{R}^d)} < \frac{1}{n}.$$

Let

$$u_n := R(\hbar_n)\chi f_n \hbar_n.$$

Then

$$\|\chi u_n\|_{L^2(\mathbb{R}^d)} = 1, \qquad P(\hbar_n)u_n = \chi f_n \hbar_n, \quad \text{and} \quad \|P(\hbar_n)u_n\|_{L^2(\mathbb{R}^d)} < \frac{\hbar_n}{n}. \tag{12.3}$$

Observe that $\{u_n\}_{n=0}^\infty$ satisfies the assumptions of Part (ii) of Lemma 12.1, and thus $\{u_n\}_{n=0}^\infty$ is uniformly locally bounded (in the sense of Definition 11.1). Theorem 11.6 therefore implies that there exists a subsequence $\{\hbar_{n_\ell}\}_{\ell=0}^\infty$ such that u_{n_ℓ} has defect measure μ. For brevity, we denote the subsequence by $\{\hbar_n\}_{n=0}^\infty$.

By the last equation in (12.3), $\|P(\hbar_n)u_n\|_{L^2(\mathbb{R}^d)} = o(\hbar_n)$. Therefore Lemma 11.9 implies that μ is supported on $\{\sigma_\hbar(P) = 0\}$ and Theorem 11.10 and Corollary 11.11 imply that the measure is invariant under the flow.

We now show that $\mu \neq 0$. Using (11.5) and Lemma 11.3, there exist $\chi_1, \chi_2 \in \mathcal{D}$ with $\chi_1 \equiv \chi_2 = 1$ on $\operatorname{supp}\chi$ such that

$$\int \chi^2 \,\mathrm{d}\mu = \lim_{n\to\infty} \left\langle \operatorname{Op}^{\text{ps}}_\hbar(\chi^2)u_n, u_n\right\rangle = \lim_{n\to\infty} \left\langle \chi_1 \operatorname{Op}^{\text{ps}}_\hbar(\chi^2)\chi_2 u_n, u_n\right\rangle = \lim_{n\to\infty} \left\langle \operatorname{Op}^{\text{ps}}_\hbar(\chi^2)\chi_2 u_n, \chi_1 u_n\right\rangle. \tag{12.4}$$

Now, by (11.2),

$$\operatorname{Op}^{\text{ps}}_\hbar(\chi^2) = \operatorname{Op}_\hbar(\chi^2) + O(\hbar^\infty)_{\Psi^{-\infty}_\hbar} = \chi^2 + O(\hbar^\infty)_{\Psi^{-\infty}_\hbar}. \tag{12.5}$$

Combining the fact that $\chi_1 = \chi_2 = 1$ on $\operatorname{supp}\chi$, (12.4), (12.5), and the first equation in (12.3), we see that

$$\int \chi^2 \,\mathrm{d}\mu = \lim_{n\to\infty} \left\langle \chi^2 u_n, u_n\right\rangle = \lim_{n\to\infty} \left\|\chi u_n\right\|^2_{L^2(\mathbb{R}^d)} = 1;$$

i.e., $\mu \neq 0$ as claimed.

We now show that u_n can be written as a free resolvent away from $\operatorname{supp}\chi$ using Lemma 12.1, and then apply Theorem 11.14 to show that the measure of the incoming set is zero. Let

$\chi_1 \in C^\infty_{\text{comp}}(\mathbb{R}^d)$ be such that $\operatorname{supp}\chi_1 \Subset \{\chi \equiv 1\}$ and $\{\chi_1 \equiv 1\} \Supset B_R$, i.e., the support of the derivatives of χ_1 is between B_R and $\{\chi \equiv 1\}$ (just as in the Proof of Part (ii) of Lemma 12.1). Let $g(\hbar_n)$ be defined by (12.2) with u replaced by u_n and f replaced by χf_n. By Part (ii) of Lemma 12.1,

$$(1-\chi_1)u_n = R_0(\hbar_n)g(\hbar_n),$$

and $\operatorname{supp} g(\hbar_n) \subset \operatorname{supp}\chi$. By assumption, there exists $\widetilde{R}$ such that $\operatorname{supp}\chi \Subset B_{\widetilde{R}}$. The assumptions of Theorem 11.14 therefore hold with R replaced by $\widetilde{R}$. Therefore, $\mu(\mathcal{I}) = 0$ with

$$\mathcal{I} := \Big\{(x,\xi) : |x| > \widetilde{R}+1,\ \langle x,\xi\rangle < 0\Big\}.$$

We now show that $\mu = 0$, which is the desired contradiction. By Corollary 10.7, the assumption that the forward trapped set Γ_{fw} is empty implies that the backward trapped set Γ_{bw} is empty. This combined with the fact that the flow is just straight-line motion outside $\operatorname{supp}(I-A)\cup\operatorname{supp}(1-n)$ imply that, given any bounded Borel set $B \subset T^*\mathbb{R}^d$, there exists $t_0 > 0$ such that if $t \le -t_0$ then $\varphi_t(B) \subset \mathcal{I}$, and thus $\mu(\varphi_t(B)) = 0$. By Corollary 11.11 (i.e., invariance of μ under the flow), $\mu(B) = 0$. Since B was arbitrary, $\mu = 0$ and the proof is complete. □

References

[1] M. Ainsworth. Discrete dispersion relation for *hp*-version finite element approximation at high wave number. *SIAM Journal on Numerical Analysis*, 42(2):553–575, 2004.

[2] G. S. Alberti and Y. Capdeboscq. *Lectures on elliptic methods for hybrid inverse problems*. Société Mathématique de France, 2018.

[3] G. Alessandrini. Strong unique continuation for general elliptic equations in 2D. *Journal of Mathematical Analysis and Applications*, 386(2):669–676, 2012.

[4] F. Alouges and M. Averseng. New preconditioners for the Laplace and Helmholtz integral equations on open curves: analytical framework and numerical results. *Num. Math.*, 148:255–292, 2021.

[5] X. Antoine and M. Darbas. Generalized combined field integral equations for the iterative solution of the three-dimensional Helmholtz equation. *ESAIM: Mathematical Modelling and Numerical Analysis (M2AN)*, 41(1):147, 2007.

[6] X. Antoine and M. Darbas. An introduction to operator preconditioning for the fast iterative integral equation solution of time-harmonic scattering problems. *Multiscale Science and Engineering*, 3(1):1–35, feb 2021.

[7] W. Arendt, I. Chalendar, and R. Eymard. Galerkin approximation of linear problems in Banach and Hilbert spaces. *IMA J. Num. Anal.*, 2020. https://doi.org/10.1093/imanum/draa067.

[8] V. Arnold. *Ordinary Differential Equations*. MIT Press, 1978.

[9] J. P. Aubin. Behavior of the error of the approximate solutions of boundary value problems for linear elliptic operators by Galerkin's and finite difference methods. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 21(4):599–637, 1967.

[10] M. Averseng. Pseudo-differential analysis of the Helmholtz layer potentials on open curves. *arXiv preprint arXiv:1905.13604*, 2019.

[11] A. K. Aziz, R. B. Kellogg, and A. B. Stephens. A two point boundary value problem with a rapidly oscillating solution. *Numerische Mathematik*, 53(1):107–121, 1988.

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

[13] J. M. Ball, Y. Capdeboscq, and B. Tsering-Xiao. On uniqueness for time harmonic anisotropic Maxwell's equations with piecewise regular coefficients. *Mathematical Models and Methods in Applied Sciences*, 22(11):1250036, 2012.

[14] L. Banjai and S. Sauter. A refined Galerkin error and stability analysis for highly indefinite variational problems. *SIAM Journal on Numerical Analysis*, 45(1):37–53, 2007.

[15] H. Barucq, T. Chaumont-Frelet, and C. Gout. Stability analysis of heterogeneous Helmholtz problems and finite element solution based on propagation media approximation. *Math. Comp.*, 86(307):2129–2157, 2017.

[16] D. Baskin and J. Wunsch. Resolvent estimates and local decay of waves on conic manifolds. *Journal of Differential Geometry*, 95(2):183–214, 2013.

[17] C. M. Bender and S. A. Orszag. *Advanced Mathematical Methods for Scientists and Engineers*. Mcgraw-Hill, New York, 1978.

[18] C. Bernardi. Optimal finite-element interpolation on curved domains. *SIAM J. Numer. Anal.*, 26(5):1212–1240, 1989.

[19] T. Betcke, S. N. Chandler-Wilde, I. G. Graham, S. Langdon, and M. Lindner. Condition number estimates for combined potential boundary integral operators in acoustics and their boundary element discretisation. *Numerical Methods for Partial Differential Equations*, 27(1):31–69, 2011.

[20] C. O. Bloom. Estimates for solutions of reduced hyperbolic equations of the second order with a large parameter. *Journal of Mathematical Analysis and Applications*, 44(2):310–332, 1973.

[21] C. O. Bloom and N. D. Kazarinoff. A priori bounds for solutions of the Dirichlet problem for $[\Delta + \lambda^2 n(x)]u = f(x, \lambda)$ on an exterior domain. *Journal of Differential Equations*, 24(3):437–465, 1977.

[22] Y. Boubendir, V. Dominguez, D. Levadoux, and C. Turc. Regularized combined field integral equations for acoustic transmission problems. *SIAM J. Appl. Math.*, 75(3):929–952, 2015.

[23] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics*. Springer, 3rd edition, 2008.

[24] A. Buffa and S. Sauter. On the acoustic single layer potential: stabilization and Fourier analysis. *SIAM Journal on Scientific Computing*, 28(5):1974–1999, 2006.

[25] N. Burq. Décroissance de l'énergie locale de l'équation des ondes pour le problème extérieur et absence de résonance au voisinage du réel. *Acta Mathematica*, 180(1):1–29, 1998.

[26] N. Burq. Semi-classical estimates for the resolvent in nontrapping geometries. *International Mathematics Research Notices*, 2002(5):221–241, 2002.

[27] A. P. Calderon and R. Vaillancourt. On the boundedness of pseudo-differential operators. *Journal of the Mathematical Society of Japan*, 23(2):374–378, 1971.

[28] Y. Capdeboscq. On the scattered field generated by a ball inhomogeneity of constant index. *Asymptot. Anal.*, 77(3-4):197–246, 2012.

[29] Y. Capdeboscq, G. Leadbetter, and A. Parker. On the scattered field generated by a ball inhomogeneity of constant index in dimension three. In *Multi-scale and high-contrast PDE: from modelling, to mathematical analysis, to inversion*, volume 577 of *Contemp. Math.*, pages 61–80. Amer. Math. Soc., Providence, RI, 2012.

[30] F. Cardoso and G. Popov. Quasimodes with exponentially small errors associated with elliptic periodic rays. *Asymptotic Analysis*, 30(3, 4):217–247, 2002.

[31] J. Céa. Approximation variationnelle des problèmes aux limites. *Ann. Inst. Fourier (Grenoble)*, 14(2):345–444, 1964.

[32] S. Chaillat, M. Darbas, and F. Le Louër. Analytical preconditioners for Neumann elastodynamic boundary element methods. *SN Partial Differential Equations and Applications*, 2(2):1–26, 2021.

[33] S. N. Chandler-Wilde and P. Monk. Wave-number-explicit bounds in time-harmonic scattering. *SIAM Journal on Mathematical Analysis*, 39(5):1428–1455, 2008.

[34] S. N. Chandler-Wilde, E. A. Spence, A. Gibbs, and V. P. Smyshlyaev. High-frequency bounds for the helmholtz equation under parabolic trapping and applications in numerical analysis. *SIAM Journal on Mathematical Analysis*, 52(1):845–893, 2020.

[35] T. Chaumont Frelet. *Approximation par éléments finis de problèmes d'Helmholtz pour la propagation d'ondes sismiques*. PhD thesis, Rouen, INSA, 2015.

[36] T. Chaumont-Frelet and S. Nicaise. High-frequency behaviour of corner singularities in Helmholtz problems. *ESAIM: Math. Model. Numer. Anal.*, 52(5):1803–1845, 2018.

[37] T. Chaumont-Frelet and S. Nicaise. Wavenumber explicit convergence analysis for finite element discretizations of general wave propagation problem. *IMA J. Numer. Anal.*, 40(2):1503–1543, 2020.

[38] T. Chaumont-Frelet, S. Nicaise, and J. Tomezyk. Uniform a priori estimates for elliptic problems with impedance boundary conditions. *Communications on Pure & Applied Analysis*, 19(5):2445, 2020.

[39] T. Chaumont-Frelet and E. A. Spence. Scattering by finely-layered obstacles: frequency-explicit bounds and homogenization. *arXiv preprint arXiv:2109.11267*, 2021.

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

[41] D. L. Colton and R. Kress. *Integral Equation Methods in Scattering Theory*. John Wiley & Sons Inc., New York, 1983.

[42] M. Costabel, M. Dauge, and S. Nicaise. Corner Singularities and Analytic Regularity for Linear Elliptic Systems. Part I: Smooth domains. 2010. https://hal.archives-ouvertes.fr/file/index/docid/453934/filename/CoDaNi_Analytic_Part_I.pdf.

[43] M. Costabel and W. McLean. Spline collocation for strongly elliptic equations on the torus. *Numerische Mathematik*, 62(1):511–538, 1992.

[44] M. Costabel and E. Stephan. A direct boundary integral equation method for transmission problems. *Journal of mathematical analysis and applications*, 106(2):367–413, 1985.

[45] M. Costabel and E. P. Stephan. Strongly elliptic boundary integral equations for electromagnetic transmission problems. *Proc. Roy. Soc. Edinb. A.*, 109(3-4):271–296, 1988.

[46] P. Cummings and X. Feng. Sharp regularity coefficient estimates for complex-valued acoustic and elastic Helmholtz equations. *Mathematical Models and Methods in Applied Sciences*, 16(1):139–160, 2006.

[47] M. Darbas and F. Le Louër. Well-conditioned boundary integral formulations for high-frequency elastic scattering problems in three dimensions. *Mathematical Methods in the Applied Sciences*, 38(9):1705–1733, 2015.

[48] G. C. Diwan, A. Moiola, and E. A. Spence. Can coercive formulations lead to fast and accurate solution of the Helmholtz equation? *J. Comp. Appl. Math.*, 352:110–131, 2019.

[49] J. Douglas Jr., J. E. Santos, D. Sheen, and L. S. Bennethum. Frequency domain treatment of one-dimensional scalar waves. *Mathematical Models and Methods in Applied Sciences*, 3(2):171–194, 1993.

[50] Y. Du and H. Wu. Preasymptotic error analysis of higher order FEM and CIP-FEM for Helmholtz equation with high wave number. *SIAM J. Numer. Anal.*, 53(2):782–804, 2015.

[51] J. J. Duistermaat and L. Hörmander. Fourier integral operators. II. *Acta mathematica*, 128(1):183–269, 1972.

[52] S. Dyatlov and M. Zworski. *Mathematical theory of scattering resonances*. AMS, 2019.

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

[54] L. C. Evans and R. F. Gariepy. *Measure theory and fine properties of functions*. CRC, 1992.

[55] X. Feng and H. Wu. Discontinuous Galerkin methods for the Helmholtz equation with large wave number. *SIAM Journal on Numerical Analysis*, 47(4):2872–2896, 2009.

[56] X. Feng and H. Wu. *hp*-Discontinuous Galerkin methods for the Helmholtz equation with large wave number. *Mathematics of computation*, 80(276):1997–2024, 2011.

[57] N Filonov. Second-order elliptic equation of divergence form having a compactly supported solution. *Journal of Mathematical Sciences*, 106(3):3078–3086, 2001.

[58] G. B. Folland. *Real analysis: modern techniques and their applications*. Wiley New York, 2nd edition, 1999.

[59] J. Galkowski. Semiclassical analysis. 2021. http://www.homepages.ucl.ac.uk/~ucahalk/SemiclassicalAnalysisNotes.pdf.

[60] J. Galkowski, D. Lafontaine, and E. A. Spence. Local absorbing boundary conditions on fixed domains give order-one errors for high-frequency waves. *arXiv preprint*, 2021. https://arxiv.org/abs/2101.02154.

[61] J. Galkowski, D. Lafontaine, and E. A. Spence. Perfectly-matched-layer truncation is exponentially accurate at high frequency. *arXiv preprint*, 2021. https://arxiv.org/abs/2105.07737.

[62] J. Galkowski, P. Marchand, and E. A. Spence. Frequency-explicit error bounds on the h-BEM for high-frequency Helmholtz exterior Dirichlet and Neumann problems. *in preparation*, 2021.

[63] J. Galkowski, E. H. Müller, and E. A. Spence. Wavenumber-explicit analysis for the Helmholtz h-BEM: error estimates and iteration counts for the Dirichlet problem. *Numer. Math.*, 142(2):329–357, 2019.

[64] J. Galkowski, E. A. Spence, and J. Wunsch. Optimal constants in nontrapping resolvent estimates. *Pure and Applied Analysis*, 2(1):157–202, 2020.

[65] D. Gallistl, T. Chaumont-Frelet, S. Nicaise, and J. Tomezyk. Wavenumber explicit convergence analysis for finite element discretizations of time-harmonic wave propagation problems with perfectly matched layers. *hal preprint 01887267*, 2018.

[66] N. Garofalo and F.-H. Lin. Unique continuation for elliptic operators: A geometric-variational approach. *Communications on Pure and Applied Mathematics*, 40(3):347–366, 1987.

[67] H. Gimperlein, J. Stocek, and C. Urzúa-Torres. Optimal operator preconditioning for pseudodifferential boundary problems. *Numerische Mathematik*, 148(1):1–41, 2021.

[68] I. Gohberg and I. A. Fel´dman. *Convolution Equations and Projection Methods for their Solution*. American Mathematical Society, 1974.

[69] I. G. Graham, M. Löhndorf, J. M. Melenk, and E. A. Spence. When is the error in the h-BEM for solving the Helmholtz equation bounded independently of k? *BIT Numer. Math.*, 55(1):171–214, 2015.

[70] I. G. Graham, O. R. Pembery, and E. A. Spence. The Helmholtz equation in heterogeneous media: a priori bounds, well-posedness, and resonances. *Journal of Differential Equations*, 266(6):2869–2923, 2019.

[71] I. G. Graham and S. Sauter. Stability and finite element error analysis for the Helmholtz equation with variable coefficients. *Math. Comp.*, 89(321):105–138, 2020.

[72] P. Grisvard. *Elliptic problems in nonsmooth domains*. Pitman, Boston, 1985.

[73] B. Guo and J. Zhang. Stable and compatible polynomial extensions in three dimensions and applications to the p and $h-p$ finite element method. *SIAM journal on numerical analysis*, 47(2):1195–1225, 2009.

[74] K. E. Gustafson and D. K. M. Rao. *Numerical range; The field of values of linear operators and matrices*. Universitext. Springer-Verlag, New York, 1997.

[75] I. Harari and T. J. R. Hughes. Finite element methods for the Helmholtz equation in an exterior domain: model problems. *Computer methods in applied mechanics and engineering*, 87(1):59–96, 1991.

[76] U. Hetmaniuk. Stability estimates for a class of Helmholtz problems. *Commun. Math. Sci*, 5(3):665–678, 2007.

[77] L. Hörmander. Uniqueness theorems and estimates for normally hyperbolic partial differential equations of the second order. *CR du douzième congrès des mathématiciens scandinaves*, pages 105–115, 1953.

[78] L. Hörmander. *The analysis of linear differential operators. I: Distribution theory and Fourier analysis*. Springer-Verlag, Berlin, 1983.

[79] L. Hörmander. *The Analysis of linear partial differential operators III: pseudo-differential operators*. Springer, 1985.

[80] G. C. Hsiao and W. L. Wendland. The Aubin–Nitsche lemma for integral equations. *The Journal of Integral Equations*, pages 299–315, 1981.

[81] G. C. Hsiao and W. L. Wendland. *Boundary integral equations*, volume 164 of *Applied Mathematical Sciences*. Springer, 2008.

[82] I. Hwang. The L^2-boundedness of pseudodifferential operators. *Transactions of the American Mathematical Society*, 302(1):55–76, 1987.

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

[84] F. Ihlenburg and I. Babuška. Finite element solution of the Helmholtz equation with high wave number Part I: The h-version of the FEM. *Computers & Mathematics with Applications*, 30(9):9–37, 1995.

[85] F. Ihlenburg and I. Babuska. Finite element solution of the Helmholtz equation with high wave number part II: the *hp* version of the FEM. *SIAM J. Numer. Anal.*, 34(1):315–358, 1997.

[86] F. Ihlenburg and I. Babuška. Dispersion analysis and error estimation of Galerkin finite element methods for the Helmholtz equation. *International Journal for Numerical Methods in Engineering*, 38, Issue 22:3745–3774, 1995.

[87] D. Jerison and C. E. Kenig. Unique continuation and absence of positive eigenvalues for schrodinger operators. *Annals of Mathematics*, 121(3):463–488, 1985.

[88] D. S. Jerison and C. E. Kenig. An identity with applications to harmonic measure. *Bulletin of the American Mathematical Society*, 2(3):447–451, 1980.

[89] D. S. Jerison and C. E. Kenig. The Neumann problem on Lipschitz domains. *Bulletin of the American Mathematical Society*, 4(2):203–207, 1981.

[90] D. S. Jerison and C. E. Kenig. The Dirichlet problem in non-smooth domains. *Annals of mathematics*, 113(2):367–382, 1981.

[91] J. L. Kazdan. Unique continuation in geometry. *Comm. Pure Appl. Math*, 41(5):667–681, 1988.

[92] J. J. Kohn and L. Nirenberg. An algebra of pseudo-differential operators. *Communications on Pure and Applied Mathematics*, 18(1-2):269–305, 1965.

[93] D. Lafontaine, E. A. Spence, and J. Wunsch. Wavenumber-explicit convergence of the hp-FEM for the full-space heterogeneous Helmholtz equation with smooth coefficients. *arxiv preprint*, 2020. https://arxiv.org/abs/2010.00585.

[94] D. Lafontaine, E. A. Spence, and J. Wunsch. Decompositions of high-frequency Helmholtz solutions via functional calculus, and application to the finite element method. *arXiv preprint arXiv:2102.13081*, 2021.

[95] D. Lafontaine, E. A. Spence, and J. Wunsch. For most frequencies, strong trapping has a weak effect in frequency-domain scattering. *Communications on Pure and Applied Mathematics*, 74(10):2025–2063, 2021.

[96] D. Lafontaine, E. A. Spence, and J. Wunsch. A sharp relative-error bound for the Helmholtz h-FEM at high frequency. *Numer. Math.*, to appear, 2022. https://arxiv.org/abs/1911.11093.

[97] P. D. Lax and R. S. Phillips. *Scattering theory*, volume 26 of *Pure and Applied Mathematics*. Academic Press Inc., Boston, MA, second edition, 1989. With appendices by Cathleen S. Morawetz and Georg Schmidt.

[98] D. Levadoux. *Etude d'une équation intégrale adaptée à la résolution hautes fréquences de l'équation d'Helmholtz*. PhD thesis, l'Université Paris VI, 2001.

[99] D. P. Levadoux and B. L. Michielsen. Nouvelles formulations intégrales pour les problèmes de diffraction d'ondes. *ESAIM-Math. Model. Num.*, 38(1):157–175, 2004.

[100] Y. Li and H. Wu. FEM and CIP-FEM for Helmholtz Equation with High Wave Number and Perfectly Matched Layer Truncation. *SIAM J. Numer. Anal.*, 57(1):96–126, 2019.

[101] H. Liu, L. Rondi, and J. Xiao. Mosco convergence for $H(\mathrm{curl})$ spaces, higher integrability for Maxwell's equations, and stability in direct and inverse EM scattering problems. *Journal of the European Mathematical Society*, 21(10):2945–2993, 2019.

[102] M. Löhndorf and J. M. Melenk. Wavenumber-Explicit *hp*-BEM for High Frequency Scattering. *SIAM Journal on Numerical Analysis*, 49(6):2340–2363, 2011.

[103] C. H. Makridakis, F. Ihlenburg, and I. Babuška. Analysis and finite element methods for a fluid-solid interaction problem in one dimension. *Mathematical Models and Methods in Applied Sciences*, 6(8):1119–1141, 1996.

[104] A. S. Markus. The reduction method for operators in Hilbert space. In *Nine Papers in Analysis*, volume 103 of *American Mathematical Society Translations: Series 2*, pages 194–200. American Mathematical Society, 1974.

[105] A. Martinez. *An introduction to semiclassical and microlocal analysis*, volume 994. Springer, 2002.

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

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

[108] J. M. Melenk and S. Sauter. Convergence analysis for finite element discretizations of the Helmholtz equation with Dirichlet-to-Neumann boundary conditions. *Math. Comp*, 79(272):1871–1914, 2010.

[109] J. M. Melenk and S. Sauter. Wavenumber explicit convergence analysis for Galerkin discretizations of the Helmholtz equation. *SIAM J. Numer. Anal.*, 49:1210–1243, 2011.

[110] R. B. Melrose and J. Sjöstrand. Singularities of boundary value problems. I. *Communications on Pure and Applied Mathematics*, 31(5):593–617, 1978.

[111] R. B. Melrose and J. Sjöstrand. Singularities of boundary value problems. II. *Communications on Pure and Applied Mathematics*, 35(2):129–168, 1982.

[112] A. Moiola. *Trefftz-discontinuous Galerkin methods for time-harmonic wave problems*. PhD thesis, Seminar for applied mathematics, ETH Zürich, 2011. Available at http://e-collection.library.ethz.ch/view/eth:4515.

[113] A. Moiola and E. A. Spence. Acoustic transmission problems: wavenumber-explicit bounds and resonance-free regions. *Math. Mod. Meth. App. S.*, 29(2):317–354, 2019.

[114] C. S. Morawetz. The decay of solutions of the exterior initial-boundary value problem for the wave equation. *Communications on Pure and Applied Mathematics*, 14(3):561–568, 1961.

[115] C. S. Morawetz. Decay for solutions of the exterior problem for the wave equation. *Communications on Pure and Applied Mathematics*, 28(2):229–264, 1975.

[116] C. S. Morawetz and D. Ludwig. An inequality for the reduced wave operator and the justification of geometrical optics. *Communications on Pure and Applied Mathematics*, 21:187–203, 1968.

[117] R. Muñoz-Sola. Polynomial liftings on a tetrahedron and applications to the *hp* version of the finite element method in three dimensions. *SIAM Journal on Numerical Analysis*, 34(1):282–314, 1997.

[118] J. C. Nédélec. *Acoustic and electromagnetic equations: integral representations for harmonic problems*. Springer Verlag, 2001.

[119] B.-T. Nguyen and D. S. Grebenkov. Localization of Laplacian Eigenfunctions in Circular, Spherical, and Elliptical Domains. *SIAM Journal on Applied Mathematics*, 73(2):780–803, jan 2013.

[120] J. Nitsche. Ein kriterium für die quasi-optimalität des ritzschen verfahrens. *Numerische Mathematik*, 11(4):346–348, 1968.

[121] L. E. Payne and H. F. Weinberger. New bounds for solutions of second order elliptic partial differential equations. *Pacific Journal of Mathematics*, 8(3):551–573, 1958.

[122] O. R. Pembery. *The Helmholtz Equation in Heterogeneous and Random Media: Analysis and Numerics*. PhD thesis, University of Bath, 2020. https://researchportal.bath.ac.uk/en/studentTheses/the-helmholtz-equation-in-heterogeneous-and-random-media-analysis.

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

[124] G. Popov and G. Vodev. Resonances near the real axis for transparent obstacles. *Communications in Mathematical Physics*, 207(2):411–438, 1999.

[125] J. V. Ralston. Trapped rays in spherically symmetric media and poles of the scattering matrix. *Communications on Pure and Applied Mathematics*, 24(4):571–582, 1971.

[126] F. Rellich. Darstellung der Eigenwerte von $\Delta u + \lambda u = 0$ durch ein Randintegral. *Mathematische Zeitschrift*, 46(1):635–636, 1940.

[127] F. Rellich. Über das asymptotische Verhalten der Lösungen von $\Delta u + \lambda u = 0$ in unendlichen Gebieten. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 53:57–65, 1943.

[128] J. Saranen and G. Vainikko. *Periodic integral and pseudodifferential equations with numerical approximation*. Springer, 2002.

[129] S. A. Sauter. A refined finite element convergence theory for highly indefinite Helmholtz problems. *Computing*, 78(2):101–115, 2006.

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

[131] A. H. Schatz. An observation concerning Ritz-Galerkin methods with indefinite bilinear forms. *Mathematics of Computation*, 28(128):959–962, 1974.

[132] C. Schwab and W. L. Wendland. Kernel properties and representations of boundary integral operators. *Mathematische Nachrichten*, 156(1):187–218, 1992.

[133] J. Sjostrand and M. Zworski. Complex scaling and the distribution of scattering poles. *Journal of the American Mathematical Society*, 4(4):729–769, 1991.

[134] E. A. Spence. Wavenumber-explicit bounds in time-harmonic acoustic scattering. *SIAM J. Math. Anal.*, 46(4):2987–3024, 2014.

[135] E. A. Spence. Overview of Variational Formulations for Linear Elliptic PDEs. In A. S. Fokas and B. Pelloni, editors, *Unified transform method for boundary value problems: applications and advances*, pages 93–159. SIAM, 2015.

[136] I. Stakgold. *Boundary value problems of mathematical physics, Volume II*. New York: The Macmillan Company; London: Collier-Macmillan Ltd, 1968.

[137] P. Stefanov. Quasimodes and resonances: sharp lower bounds. *Duke mathematical journal*, 99(1):75–92, 1999.

[138] P. Stefanov. Resonances near the real axis imply existence of quasimodes. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 330(2):105–108, 2000.

[139] P. Stefanov. Resonance expansions and Rayleigh waves. *Mathematical Research Letters*, 8(2):107–124, 2001.

[140] P. Stefanov and G. Vodev. Distribution of resonances for the neumann problem in linear elasticity outside a strictly convex body. *Duke Mathematical Journal*, 78(3):677–714, 1995.

[141] P. Stefanov and G. Vodev. Neumann resonances in linear elasticity for an arbitrary body. *Communications in mathematical physics*, 176(3):645–659, 1996.

[142] O. Steinbach. *Numerical Approximation Methods for Elliptic Boundary Value Problems: Finite and Boundary Elements*. Springer, New York, 2008.

[143] E. P. Stephan. Boundary integral equations for screen problems in $\mathbb{R}^3$. *Integral Equations and Operator Theory*, 10(2):236–257, 1987.

[144] S. H. Tang and M. Zworski. From quasimodes to resonances. *Mathematical Research Letters*, 5:261–272, 1998.

[145] M. Taylor. *Partial differential equations II, Qualitative studies of linear equations, volume 116 of Applied Mathematical Sciences*. Springer-Verlag, New York, 1996.

[146] A. Toselli and O. Widlund. *Domain Decomposition Methods: Algorithms and Theory*. Springer, 2005.

[147] B. R. Vainberg. On the short wave asymptotic behaviour of solutions of stationary problems and the asymptotic behaviour as $t \to \infty$ of solutions of non-stationary problems. *Russian Mathematical Surveys*, 30(2):1–58, 1975.

[148] B. R. Vainberg. *Asymptotic methods in equations of mathematical physics*. Gordon & Breach Science Publishers, New York, 1989. Translated from the Russian by E. Primrose.

[149] G. Vainikko. On the question of convergence of Galerkin's method. *Tartu Rükl. Ul. Toim*, 177:148–152, 1965.

[150] T. H. Wolff. A property of measures in $\mathbb{R}^N$ and an application to unique continuation. *Geometric & Functional Analysis*, 2(2):225–284, 1992.

[151] H. Wu. Pre-asymptotic error analysis of CIP-FEM and FEM for the Helmholtz equation with high wave number. Part I: linear version. *IMA J. Numer. Anal.*, 34(3):1266–1288, 2014.

[152] H. Wu and J. Zou. Finite element method and its analysis for a nonlinear Helmholtz equation with high wave numbers. *SIAM J. Numer. Anal.*, 56(3):1338–1359, 2018.

[153] J. Zhang. *The hp version of the finite element method in three dimensions*. PhD thesis, University of Manitoba, 2008.

[154] L. Zhu and H. Wu. Preasymptotic error analysis of CIP-FEM and FEM for Helmholtz equation with high wave number. Part II: *hp* version. *SIAM J. Numer. Anal.*, 51(3):1828–1852, 2013.

[155] M. Zworski. *Semiclassical analysis*. American Mathematical Society Providence, RI, 2012.