Computers and Mathematics with Applications 113 (2022) 59–69

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

www.elsevier.com/locate/camwa

Wavenumber-explicit convergence of the *hp*-FEM for the full-space heterogeneous Helmholtz equation with smooth coefficients

D. Lafontaine [a], E.A. Spence [a,*], J. Wunsch [b]

[a] Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK
[b] Department of Mathematics, Northwestern University, 2033 Sheridan Road, Evanston IL 60208-2730, USA

ARTICLE INFO

Keywords:
Helmholtz equation
hp-FEM
High frequency
Pollution effect

ABSTRACT

A convergence theory for the *hp*-FEM applied to a variety of constant-coefficient Helmholtz problems was pioneered in the papers [35], [36], [15], [34]. This theory shows that, if the solution operator is bounded polynomially in the wavenumber k, then the Galerkin method is quasioptimal provided that $hk/p \leq C_1$ and $p \geq C_2 \log k$, where C_1 is sufficiently small, C_2 is sufficiently large, and both are independent of k, h, and p. The significance of this result is that if $hk/p = C_1$ and $p = C_2 \log k$, then quasioptimality is achieved with the total number of degrees of freedom proportional to k^d; i.e., the *hp*-FEM does not suffer from the pollution effect.

This paper proves the analogous quasioptimality result for the heterogeneous (i.e. variable-coefficient) Helmholtz equation, posed in $\mathbb{R}^d$, $d = 2, 3$, with the Sommerfeld radiation condition at infinity, and C^∞ coefficients. We also prove a bound on the relative error of the Galerkin solution in the particular case of the plane-wave scattering problem. These are the first ever results on the wavenumber-explicit convergence of the *hp*-FEM for the Helmholtz equation with variable coefficients.

1. Introduction

1.1. Context

Over the last 10 years, a wavenumber-explicit convergence theory for the *hp*-FEM applied to the Helmholtz equation

$$\Delta u + k^2 u = -f \tag{1.1}$$

was established in the papers [35], [36], [15], [34]. This theory is based on decomposing solutions of the Helmholtz equation into two components:

(i) an analytic component, satisfying bounds with the same k-dependence as those satisfied by the full Helmholtz solution, and
(ii) a component with finite regularity, satisfying bounds with improved k-dependence compared to those satisfied by the full Helmholtz solution.

Such a decomposition was obtained for

- the Helmholtz equation (1.1) posed in $\mathbb{R}^d$, $d = 2, 3$, with compactly-supported f, and with the Sommerfeld radiation condition

$$\frac{\partial u}{\partial r}(x) - \mathrm{i}ku(x) = o\left(\frac{1}{r^{(d-1)/2}}\right) \tag{1.2}$$

  as $r := |x| \to \infty$, uniformly in $\widehat{x} := x/r$ [35, Lemma 3.5],
- the Helmholtz exterior Dirichlet problem where the obstacle has analytic boundary [36, Theorem 4.20],
- the Helmholtz interior impedance problem where the domain is either smooth ($d = 2, 3$) [36, Theorem 4.10], [34, Theorem 4.5], or polygonal [36, Theorem 4.10], [15, Theorem 3.2].

This decomposition was then used to prove quasioptimality of the *hp*-FEM applied to the standard Helmholtz variational formulation in [35], [36], [15], and applied to a discontinuous Galerkin formulation in [34]. Indeed, for the standard variational formulation (defined for the full-space problem in Definition 2.2 below) applied to the boundary value problems above, if the solution operator of the problem is bounded polynomially in k (see Definition 2.6 below), then there exist C_1, C_2, and C_{qo} (independent of k, h, and p) such that if

$$\frac{hk}{p} \leq C_1 \quad \text{and} \quad p \geq C_2 \log k \tag{1.3}$$

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

* Corresponding author.
E-mail addresses: D.Lafontaine@bath.ac.uk (D. Lafontaine), E.A.Spence@bath.ac.uk (E.A. Spence), jwunsch@math.northwestern.edu (J. Wunsch).

https://doi.org/10.1016/j.camwa.2022.03.007
Received 20 September 2021; Received in revised form 31 January 2022; Accepted 5 March 2022

$$\|u-u_N\|_{H^1_k} \le C_{\rm qo} \min_{v_N\in V_N} \|u-v_N\|_{H^1_k},$$

where V_N is the hp approximation space and the norm $\|\cdot\|_{H^1_k}$ is the standard weighted H^1 norm (defined by (2.7) below). Since the total number of degrees of freedom of the approximation space is proportional to $(p/h)^d$, the significance of this result is that it shows there is a choice of h and p such that the Galerkin solution is quasioptimal, with quasioptimality constant (i.e. $C_{\rm qo}$) independent of k, and with the total number of degrees of freedom proportional to k^d; thus, with these choices of k and p, the hp-FEM does not suffer from the pollution effect [2].

Over the last few years, there has been increasing interest in the numerical analysis of the heterogeneous Helmholtz equation, i.e. the Helmholtz equation with variable coefficients

$$\nabla\cdot(\mathsf{A}\nabla u)+k^2 nu=-f; \tag{1.4}$$

see, e.g., [8], [3], [10], [18], [38], [21], [16], [29], [19]. However there do not yet exist in the literature analogous results to those in [35], [36], [15], [34] for the variable-coefficient Helmholtz equation.

1.2. Informal statement and discussion of the main results

The main results This paper considers the variable-coefficient Helmholtz equation (1.4) with C^∞ coefficients posed in $\mathbb{R}^d$, $d=2,3$, with the Sommerfeld radiation condition at infinity. We obtain analogous results to those obtained in [35] for this scenario with constant coefficients. That is, we prove quasioptimality of the hp-FEM under the conditions (1.3) and provided that the solution operator is polynomially bounded in k; see Theorem 3.4 below.

We obtain this result by decomposing the solution u to (1.4) into two components:

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

where $u_{H^2}\in H^2(B_R)$ and $u_{\mathcal{A}}$ is analytic in B_R, where B_R denotes the ball of radius R centred at the origin (and R is arbitrary); see Theorem 3.1 below. This is exactly analogous to the decomposition obtained in [35], except that now u satisfies the variable-coefficient equation (1.4) instead of (1.1).

Overview of the ideas behind the decomposition and subsequent bounds The idea in [35] was to decompose the data f in (1.1) into “low-” and “high-” frequency components, with $u_{\mathcal{A}}$ the Helmholtz solution for the low-frequency component of f and u_{H^2} the Helmholtz solution for the high-frequency component of f. The frequency cut-offs were defining using the indicator function

$$1_{B_{\lambda k}}(\zeta):=\begin{cases}1 & \text{for } |\zeta|\le\lambda k,\\ 0 & \text{for } |\zeta|\ge \lambda k,\end{cases} \tag{1.5}$$

with λ a free parameter (see [35, Equation 3.31] and the surrounding text). In [35] the frequency cut-off (1.5) was then used with (a) the expression for u as a convolution of the fundamental solution and the data f, and (b) the fact that the fundamental solution is known explicitly when $\mathsf{A}=\mathsf{I}$ and $n=1$, to obtain the appropriate bounds on $u_{\mathcal{A}}$ and u_{H^2} using explicit calculation.

In this paper we use the same idea as in [35] of decomposing into low- and high-frequency components, but apply frequency cut-offs to the solution u as opposed to the data f. Then, given any cut-off function that is zero for $|\zeta|\ge Ck$, bounding the corresponding low-frequency component $u_{\mathcal{A}}$ is relatively straightforward using basic properties of the Fourier-transform (namely the expression for the Fourier transform of a derivative and Parseval's theorem). Indeed, in Fourier space each derivative corresponds to a power of the Fourier variable ζ, and the frequency cut-off means that $|\zeta|\le Ck$ for $u_{\mathcal{A}}$; i.e. every derivative of $u_{\mathcal{A}}$ brings down a power of k compared to $u_{\mathcal{A}}$ (see §5.3 below). The main difficulty therefore is showing that the high-frequency component u_{H^2} satisfies a bound with one power of k improvement over the bound satisfied by u.

The main idea of the present paper is that the high-frequency cut-off can be chosen so that the (scaled) Helmholtz operator

$$P_k:=-\big(k^{-2}\nabla\cdot(\mathsf{A}\nabla\cdot)+n\big) \tag{1.6}$$

is *semiclassically elliptic* on the support of the high-frequency cut-off. Furthermore, choosing the cut-off function to be smooth (as opposed to discontinuous, as in (1.5)) then allows us to use basic facts about the “nice” behaviour of elliptic semiclassical pseudodifferential operators (namely, they are invertible up to a small error) to prove the required bound on u_{H^2}. (Recall that semiclassical pseudodifferential operators are just pseudodifferential operators with a large/small parameter; in this case the large parameter is k.)

We now discuss further the frequency cut-offs and the bound on u_{H^2} via ellipticity.

The frequency cut-offs In contrast to (1.5), we choose $\chi_\mu\in C^\infty_{\rm comp}(\mathbb{R}^d)$ such that

$$\chi_\mu\big(k^{-2}|\zeta|^2\big)=\begin{cases}1 & \text{for } |\zeta|\le\sqrt{\mu}\,k,\\ 0 & \text{for } |\zeta|\ge\sqrt{2\mu}\,k,\end{cases} \tag{1.7}$$

where the parameter μ is chosen later in the argument. With the Fourier transform and its inverse defined by

$$\mathcal{F}\varphi(\zeta):=\int_{\mathbb{R}^d}\exp\big(-{\rm i}x\cdot\zeta\big)\varphi(x)\,{\rm d}x \quad\text{and}$$

$$\mathcal{F}^{-1}\psi(x):=(2\pi)^{-d}\int_{\mathbb{R}^d}\exp\big({\rm i}x\cdot\zeta\big)\psi(\zeta)\,{\rm d}\zeta, \tag{1.8}$$

we define the low-frequency cut-off Π_L by

$$\Pi_L v(x):=\mathcal{F}^{-1}\Big(\chi_\mu\big(k^{-2}|\zeta|^2\big)\,\mathcal{F}v(\zeta)\Big), \tag{1.9}$$

and the high-frequency cut-off Π_H by

$$\Pi_H v(x):=\mathcal{F}^{-1}\Big(\big(1-\chi_\mu\big(k^{-2}|\zeta|^2\big)\big)\,\mathcal{F}v(\zeta)\Big), \tag{1.10}$$

so that $\Pi_L+\Pi_H=I$. We let $\varphi\in C^\infty_c$ be equal to one on B_{R+1} and vanish outside B_{R+2}, and then

$$u_{\mathcal{A}}:=\Pi_L(\varphi u)\Big|_{B_R}\quad\text{and}\quad u_{H^2}:=\Pi_H(\varphi u)\Big|_{B_R}. \tag{1.11}$$

The bound on the high-frequency component u_{H^2} via ellipticity Recall that a PDE is elliptic if its principal symbol is non-zero. The concept of ellipticity for semiclassical differential operators (or, more generally, semiclassical pseudodifferential operators) is analogous, except that it now involves the *semiclassical principal symbol* (see (4.17) below). The semiclassical principal symbol of P_k (1.6) is

$$\langle\mathsf{A}\xi,\xi\rangle-n, \tag{1.12}$$

where $\langle\cdot,\cdot\rangle$ denotes the ℓ^2 inner product and $\xi=k^{-1}\zeta$ (see (4.12) below and the surrounding text).

If the parameter μ in the cut-off function χ_μ (1.7) is chosen to be a certain function of A and n (see (5.7) below), then the symbol (1.12) is bounded away from zero when $k^{-2}|\zeta|^2\ge\mu$, i.e. in the region of Fourier space where Π_H is non-zero; one therefore describes P_k as “microlocally elliptic”, where the adjective “microlocal” indicates that we have ellipticity on just a region of phase space (rather than on all of phase space in the more familiar global ellipticity).

These ellipticity properties are then used with the standard microlocal elliptic estimate for pseudodifferential operators, appearing in the semiclassical setting in, e.g., [14, Appendix E], and stated in this setting as Theorem 4.3 below. The whole point is that a semiclassical pseudodifferential operator that is elliptic in some region of phase space can be inverted (up to some small error) in that region, and the norm of the inverse is bounded uniformly in the large parameter (here k) as long as one uses weighted norms (analogous to the familiar H^1_k norm (2.7)).

The result is that u_{H^2} satisfies a bound with one power of k improvement over the bound satisfied by u (compare (3.1) and (2.12)). To give a simple illustration of how ellipticity can give this improved k-dependence, we contrast the solutions of

$$P_k u := -(\Delta + k^2)u = f \quad \text{and} \quad \widetilde{P}_k v := -(\Delta - k^2)v = f,$$

with both equations posed in $\mathbb{R}^d$ with compactly-supported f, and with u satisfying the Sommerfeld radiation condition (1.2) and v satisfying boundedness at infinity. The $L^2 \to L^2$ bounds that are sharp in terms of k-dependence are

$$\|u\|_{L^2(B_R)} \lesssim k^{-1}\|f\|_{L^2(\mathbb{R}^d)} \quad \text{and} \quad \|v\|_{L^2(\mathbb{R}^d)} \lesssim k^{-2}\|f\|_{L^2(\mathbb{R}^d)},$$

with the former given by Part (i) of Theorem 2.7, and the latter following from the Lax-Milgram theorem. The operator P_k is not semiclassically elliptic on all of phase space (its semiclassical principal symbol is $|\xi|^2 - 1$), whereas $\widetilde{P}_k$ is semiclassically elliptic on all of phase space (its semiclassical principal symbol is $|\xi|^2 + 1$); we therefore see that ellipticity has resulted in the solution operator having improved k-dependence. The proof of the bound on u_{H^2} is more technical, but the idea – that the improvement in k-dependence comes from ellipticity – is the same.

The assumption that the solution operator is polynomially bounded in k We need to assume that the solution operator is polynomially bounded in k (in sense of Definition 2.6 below), *both* in proving the bound on u_{H^2}, *and* in proving quasi-optimality of the hp-FEM.

The k-dependence of the Helmholtz solution operator depends on whether the problem is *trapping* or *nontrapping*. For the heterogeneous Helmholtz equation (1.4) posed in $\mathbb{R}^d$ (i.e. with no obstacle), trapping can be created by the coefficients A and n; see, e.g., [39]. If the problem is nontrapping, then the Helmholtz solution operator (measured in the natural norms) is bounded in k. However, under the strongest form of trapping, the Helmholtz solution operator can grow exponentially in k [39]. Nevertheless, it has recently been proved that, if a set of frequencies of arbitrarily small measure is excluded, then the solution operator is polynomially bounded under any type of trapping [28]. Therefore, the result that the hp-FEM is quasi-optimal holds for a wide class of Helmholtz problems; see Corollary 3.5 below.

Why do we need C^∞ coefficients? As highlighted above, our proof of the decomposition relies on standard results about semiclassical pseudodifferential operators (recapped in §4). These results are usually stated for C^∞ symbols, and thus to fit into this framework A and n must be C^∞. However, examining the results we use, we see that we only need the symbol of the PDE to be in C^L where L depends only on the dimension d and on the exponent M appearing in the assumption that the solution operator is polynomially bounded (see Definitions 2.5 and 2.6 below). Therefore, while we consider $\mathsf{A}, n \in C^\infty$ to easily use results about semiclassical pseudodifferential operators from [52], [14, Appendix E], our results hold for $\mathsf{A} \in C^L$ and $n \in C^L$, where $L = L(d, M)$.

Extending the decomposition result to the solution of other PDEs Our proof of the decomposition result only relies on the principal symbol of the differential operator being bounded below at infinity (in the sense of (3.8) below). Therefore, the decomposition result Theorem 3.1 is valid for a much larger class of PDEs (and indeed pseudodifferential operators) than (1.4); see Remark 3.7 below for more details.

In the follow-up paper [27], we use the ideas of the present paper combined with much more sophisticated tools of semiclassical and microlocal analysis (namely the black-box scattering framework of Sjöstrand–Zworski [45], the Helffer–Sjöstrand functional calculus [23], and associated results by Helffer, Robert, and Sjöstrand [22], [40], [44]) to prove analogous decompositions for a wide variety of scattering problems (albeit with slightly weaker estimates on $u_{\mathcal{A}}$). In particular, the main result of the present paper, Theorem 3.1, is rederived in this more general context as [27, Theorem 1.16].

We also note that, as announced in the abstract [4], Bernkopf, Chaumont–Frelet, and Melenk are also studying the question of k-explicit convergence of the hp-FEM for the Helmholtz equation with variable coefficients.

Outline of the paper §2 gives the definitions of the boundary-value problem and the finite-element method. §3 states the main results. §4 recaps results about semiclassical pseudodifferential operators, with [52] and [14, Appendix E] as the main references. §5 proves the result about the decomposition $u|_{B_R} = u_{H^2} + u_{\mathcal{A}}$ (Theorem 3.1). §6 proves the result about quasioptimality of the hp-FEM (Theorem 3.4).

2. Formulation of the problem

2.1. The boundary value problem

Assumption 2.1 *(Assumptions on the coefficients).* $\mathsf{A} \in C^\infty(\mathbb{R}^d, \mathsf{SPD})$ (where SPD is the set of $d \times d$ real, symmetric, positive-definite matrices) is such that $\operatorname{supp}(\mathsf{I} - \mathsf{A})$ is compact in $\mathbb{R}^d$ and there exist $0 < A_{\min} \le A_{\max} < \infty$ such that, in the sense of quadratic forms,

$$A_{\min} \le \mathsf{A}(x) \le A_{\max} \quad \text{for all } x \in \mathbb{R}^d. \tag{2.1}$$

$n \in C^\infty(\mathbb{R}^d, \mathbb{R})$ is such that $\operatorname{supp}(1-n)$ is compact in $\mathbb{R}^d$ and there exist $0 < n_{\min} \le n_{\max} < \infty$ such that

$$n_{\min} \le n(x) \le n_{\max} \quad \text{for all } x \in \mathbb{R}^d. \tag{2.2}$$

Let $R > 0$ be such that $\operatorname{supp}(\mathsf{I} - \mathsf{A}) \cup \operatorname{supp}(1-n) \Subset B_R$, where B_R denotes the ball of radius R about the origin and $\Subset$ denotes compact containment. Let γ and ∂_n denote the Dirichlet and Neumann traces, respectively, on ∂B_R, where the normal vector for the Neumann trace points out of B_R.

Define $\mathrm{DtN}_k : H^{1/2}(\partial B_R) \to H^{-1/2}(\partial B_R)$ to be the Dirichlet-to-Neumann map for the equation $\Delta u + k^2 u = 0$ posed in the exterior of B_R with the Sommerfeld radiation condition (1.2). The definition of DtN_k in terms of Hankel functions and polar coordinates (when $d = 2$)/spherical polar coordinates (when $d = 3$) is given in, e.g., [35, Equations 3.7 and 3.10].

Definition 2.2 *(Heterogeneous Helmholtz Problem on $\mathbb{R}^d$).* Given A and n satisfying Assumption 2.1, $R > 0$ such that $\operatorname{supp}(\mathsf{I} - \mathsf{A}) \cup \operatorname{supp}(1-n) \Subset B_R$, $k > 0$, and $F \in (H^1(B_R))^*$, $u \in H^1(B_R)$ satisfies the *Heterogeneous Helmholtz Problem on $\mathbb{R}^d$* if u satisfies the variational problem

$$\text{find } u \in H^1(B_R) \text{ such that} \quad a(u, v) = F(v) \quad \text{for all } v \in H^1(B_R), \tag{2.3}$$

where

$$a(u, v) := \int_{B_R} \Big((\mathsf{A}\nabla u) \cdot \overline{\nabla v} - k^2 n u \overline{v} \Big) - \big\langle \mathrm{DtN}_k(\gamma u), \gamma v \big\rangle_{\partial B_R}, \tag{2.4}$$

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

Lemma 2.3 *(Helmholtz boundary value problems included in Definition 2.2).*

(i) *If*

$$F(v) := \int_{B_R} f\, \overline{v} \tag{2.5}$$

with $f \in L^2(B_R)$, then the solution u to (2.3) equals $\widetilde{u}|_{B_R}$, where $\widetilde{u} \in H^1_{\rm loc}(\mathbb{R}^d)$ is the solution to

$$\nabla \cdot (\mathsf{A}\nabla\widetilde{u}) + k^2 n\widetilde{u} = -f \quad \text{in } \mathbb{R}^d,$$

and $\widetilde{u}$ satisfies the Sommerfeld radiation condition (1.2).

(ii) *If*

$$F(v) := \int_{\partial B_R} \left(\partial_n u^I - \mathrm{DtN}_k(\gamma u^I)\right)\overline{\gamma v} \quad \text{with} \quad u^I(x) := \exp(\mathrm{i}kx\cdot a), \tag{2.6}$$

where $a \in \mathbb{R}^d$ with $|a| = 1$, then the solution u to (2.3) equals $\widetilde{u}|_{B_R}$, where $\widetilde{u} \in H^1_{\rm loc}(\mathbb{R}^d)$ is the solution of the Helmholtz plane-wave scattering problem*; i.e.*

$$\nabla \cdot (\mathsf{A}\nabla\widetilde{u}) + k^2 n\widetilde{u} = 0 \quad \text{in } \mathbb{R}^d,$$

and $\widetilde{u}^S := \widetilde{u} - u^I$ satisfies the Sommerfeld radiation condition (1.2).

Part (i) of Lemma 2.3 is proved in, e.g., [20, Lemma 3.3]; the proof of Part (ii) is similar.

Let the weighted H^1 norm, $\|\cdot\|_{H^1_k(B_R)}$, be defined by

$$\|u\|^2_{H^1_k(B_R)} := \|\nabla u\|^2_{L^2(B_R)} + k^2\|u\|^2_{L^2(B_R)}. \tag{2.7}$$

Lemma 2.4. *The solution of the Heterogeneous Helmholtz Problem on $\mathbb{R}^d$ (defined in Definition 2.2) exists, is unique, and there exists $C(k, \mathsf{A}, n, R) > 0$ such that*

$$\|u\|_{H^1_k(B_R)} \le C\,\|F\|_{(H^1_k(B_R))^*} \quad \text{for all } k > 0. \tag{2.8}$$

Proof. Uniqueness follows from the unique continuation principle; see [20, §1], [21, §2] and the references therein. Since $a(\cdot,\cdot)$ satisfies a Gårding inequality (see (6.4) below), Fredholm theory then gives existence and the bound (2.8). ∎

Properties of DtN_k *and* $a(\cdot,\cdot)$ We use later the following two properties of DtN_k: given $k_0, R_0 > 0$, there exists $C_{\rm DtN} = C_{\rm DtN}(k_0R_0)$ such that, for all $k \ge k_0$ and $R \ge R_0$,

$$\left|\left\langle \mathrm{DtN}_k(\gamma u), \gamma v\right\rangle_{\partial B_R}\right| \le C_{\rm DtN1}\,\|u\|_{H^1_k(B_R)}\,\|v\|_{H^1_k(B_R)} \tag{2.9}$$

for all $u, v \in H^1(B_R)$, and

$$-\Re\left\langle \mathrm{DtN}_k\phi, \phi\right\rangle_{\partial B_R} \ge 0 \quad \text{for all } \phi \in H^{1/2}(\partial B_R). \tag{2.10}$$

For a proof of (2.9), see [35, Lemma 3.3]. For a proof of (2.10), see [37, Theorem 2.6.4] (for $d = 3$) and [7, Corollary 3.1] or [35, Lemma 3.10] (for $d = 2, 3$).

Let $C_{\rm cont} = C_{\rm cont}(\mathsf{A}, n, R, k_0)$ be the *continuity constant* of the sesquilinear form $a(\cdot,\cdot)$ (defined in (2.4)) in the norm $\|\cdot\|_{H^1_k(B_R)}$; i.e.

$$a(u, v) \le C_{\rm cont}\,\|u\|_{H^1_k(B_R)}\,\|v\|_{H^1_k(B_R)} \quad \text{for all } u, v \in H^1(B_R) \text{ and } k \ge k_0.$$

By the Cauchy-Schwarz inequality and (2.9),

$$C_{\rm cont} \le \max\{A_{\max}, n_{\max}\} + C_{\rm DtN1}. \tag{2.11}$$

2.2. The behaviour of the solution operator for large k

Definition 2.5 *($C_{\rm sol}$).* Given $f \in L^2(B_R)$, let u be the solution of the heterogeneous Helmholtz equation (1.4) with the Sommerfeld radiation condition (1.2) (i.e. u is the solution of the variational problem (2.3) with $F(v)$ given by (2.5)). Given $k_0 > 0$, let $C_{\rm sol} = C_{\rm sol}(k, \mathsf{A}, n, R, k_0) > 0$ be such that

$$\|u\|_{H^1_k(B_R)} \le C_{\rm sol}\,\|f\|_{L^2(B_R)} \quad \text{for all } k > 0. \tag{2.12}$$

$C_{\rm sol}$ exists by Lemma 2.4; indeed, with C given by (2.8), $C_{\rm sol} := C/k$.

How $C_{\rm sol}$ depends on k is crucial to the analysis below, and to emphasise this we write $C_{\rm sol} = C_{\rm sol}(k)$. Below we consider $C_{\rm sol}$ with different values of R, and we then write, e.g., $C_{\rm sol}(k; R)$ (as in the bound (3.2) below).

A key assumption in the analysis of the Helmholtz hp-FEM is that $C_{\rm sol}(k)$ is polynomially bounded in k in the following sense.

Definition 2.6 *($C_{\rm sol}$ is polynomially bounded in k).* Given k_0 and $K \subset [k_0, \infty)$, $C_{\rm sol}(k)$ is polynomially bounded for $k \in K$ if there exists $C > 0$ and $M > 0$ such that

$$C_{\rm sol}(k) \le Ck^M \text{ for all } k \in K, \tag{2.13}$$

where C and M are independent of k (but depend on k_0 and possibly also on K, A, n, d, R).

There exist C^∞ coefficients A and n such that $C_{\rm sol}(k_j) \ge c_1\exp(c_2k_j)$ for $0 < k_1 < k_2 < \ldots$ with $k_j \to \infty$ as $j \to \infty$, see [39], but this exponential growth is the worst-possible, since $C_{\rm sol}(k) \le c_3\exp(c_4k)$ for all $k \ge k_0$ by [5, Theorem 2]. We now recall results on when $C_{\rm sol}(k)$ is polynomially bounded in k.

Theorem 2.7 *(Conditions under which $C_{\rm sol}(k)$ is polynomially bounded in k).*

(i) *A and n are C^∞ and nontrapping (i.e. all the trajectories of the Hamiltonian flow defined by the symbol of (1.4) starting in B_R leave B_R after a uniform time), then $C_{\rm sol}(k)$ is independent of k for all k, i.e., (2.13) holds for all k with $M = 0$.*

(ii) *If $n = 1$ and A is $C^{0,1}$ then, given $k_0 > 0$ and $\delta > 0$ there exists a set $J \subset [k_0, \infty)$ with $|J| \le \delta$ such that*

$$C_{\rm sol}(k) \le Ck^{5d/2+1+\varepsilon} \quad \text{for all } k \in [k_0, \infty)\setminus J, \tag{2.14}$$

for any $\varepsilon > 0$, where C depends on $\delta, \varepsilon, d, k_0$, and A. If A is $C^{1,\sigma}$ for some $\sigma > 0$ then the exponent is reduced to $5d/2 + \varepsilon$.

References for the proof. (i) is proved using *either* (a) the propagation of singularities results of [13] combined with either the parametrix argument of [48, Theorem 3]/[49, Chapter 10, Theorem 2] or Lax–Phillips theory [30], *or* (b) the defect-measure argument of [6, Theorem 1.3 and §3]. It has recently been proved that, for this situation, $C_{\rm sol}$ is proportional to the length of the longest trajectory in B_R; see [16, Theorems 1 and 2, and Equation 6.32].

(ii) is proved in [28, Theorem 1.1 and Corollary 3.6]. ∎

2.3. The finite-element method

Let $(V_N)_{N=0}^\infty$ be a sequence of finite-dimensional subspaces of $H^1(B_R)$ that converge to $H^1(B_R)$ in the sense that, for all $v \in H^1(B_R)$,

$$\lim_{N\to\infty}\left(\min_{v_N\in V_N}\|v - v_N\|_{H^1(B_R)}\right) = 0.$$

Later we specialise to the triangulations described in [35, §5], which allow curved elements and thus fit ∂B_R exactly.

The finite-element method for the variational problem (2.3) is the Galerkin method applied to the variational problem (2.3), i.e.

$$\text{find } u_N \in V_N \text{ such that } \; a(u_N, v_N) = F(v_N) \; \text{ for all } v_N \in V_N. \tag{2.15}$$

3. Statement of the main results

Theorem 3.1 *(Decomposition of the solution). Let A and n satisfy Assumption 2.1 and let $R > 0$ be such that $\mathrm{supp}(\mathsf{I} - \mathsf{A}) \cup \mathrm{supp}(1 - n) \Subset B_R$. Given $f \in L^2(B_R)$, let u satisfy $\nabla\cdot(\mathsf{A}\nabla u) + k^2nu = -f$ in $\mathbb{R}^d$ and the Sommerfeld radiation condition (1.2).*

If $C_{\rm sol}(k)$ is polynomially bounded (in the sense of Definition 2.6) for $k \in K \subset [k_0,\infty)$, then there exist $C_3, C_4, C_5 > 0$ such that

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

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

$$\|\partial^\alpha u_{H^2}\|_{L^2(B_R)} \le C_3 k^{|\alpha|-2} \|f\|_{L^2(B_R)}$$
$$\text{for all } |\alpha| \le 2 \text{ and for all } k \in K \subset [k_0,\infty), \tag{3.1}$$

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

$$\left\|\partial^\beta u_{\mathcal{A}}\right\|_{L^2(B_R)} \le C_{\rm sol}(k; R+2)\, C_4 \left(C_5 k\right)^{|\beta|-1} \|f\|_{L^2(B_R)}$$
$$\text{for all } \beta \text{ and for all } k \in K \subset [k_0,\infty), \tag{3.2}$$

where C_3, C_4, and C_5 depend on A, n, d, and k_0, but are independent of k, f, α, and β.

Remark 3.2 *($u_{\mathcal{A}}$ is analytic).* Since C_4 and C_5 are independent of β, the bound (3.2) implies that $u_{\mathcal{A}}$ is in the class of analytic functions on B_R, $\mathcal{A}(B_R)$, defined by

$$\mathcal{A}(B_R) := \Big\{ v \in \bigcap_{n\in\mathbb{N}} H^n(B_R) : \exists\, c_0, c_1 > 0, \text{ independent of } n,$$
$$\text{such that } |u|_{H^n(B_R)} \le c_1 c_0^n n! \Big\},$$

where $|u|^2_{H^n} := \sum_{|\alpha|=n} \|\partial^\alpha u\|^2_{L^2}$. See, e.g., [11, §1.1.b], both for this definition, and for how the definition implies convergence of the Taylor series of elements of $\mathcal{A}(B_R)$ at every point in $\overline{B_R}$.

Remark 3.3 *(The bounds of Theorem 3.1 written with the notation ∇^n).* The analogous bounds to (3.1) and (3.2) in [35], [36] are written using the notation

$$\left|\nabla^n u(x)\right|^2 := \sum_{|\alpha|=n} \frac{n!}{\alpha!}\left|\partial^\alpha u(x)\right|^2.$$

Since $\sum_{|\alpha|=n}(n!/\alpha!) = d^n$,

$$\text{if } \|\partial^\alpha u\|_{L^2(B_R)} \le C_1 \left(C_2\right)^{|\alpha|} \text{ for all } \alpha \text{ with } |\alpha| = n,$$
$$\text{then } \|\nabla^n u\|_{L^2(B_R)} \le C_1 \left(C_2\sqrt{d}\right)^n,$$

and so the bounds (3.1) and (3.2) can also be written as bounds on $\|\nabla^n u_{H^2}\|_{L^2(B_R)}$ and $\|\nabla^n u_{\mathcal{A}}\|_{L^2(B_R)}$ respectively.

The following result about quasioptimality of the hp-FEM is then obtained by combining Theorem 3.1, well-known results about the convergence of the Galerkin method based on duality arguments (recapped in Lemma 6.4 below), and results about the hp approximation spaces in [35, §5] (used in Lemma 6.5 below).

Theorem 3.4 *(Quasioptimality of the hp-FEM if $C_{\rm sol}(k)$ is polynomially bounded). Let $d = 2$ or 3, and let $k_0 > 0$. Let $(V_N)_{N=0}^\infty$ be the piecewise-polynomial approximation spaces described in [35, §5] (where, in particular, the triangulations are quasi-uniform), and let u_N be the Galerkin solution defined by (2.15).*

If $C_{\rm sol}(k)$ is polynomially bounded (in the sense of Definition 2.6) for $k \in K \subset [k_0,\infty)$ 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 if (1.3) holds, then, for all $k \in K$, the Galerkin solution exists, is unique, and satisfies the quasi-optimal error bound

$$\|u - u_N\|_{H^1_k(B_R)} \le C_{\rm qo} \min_{v_N \in V_N} \|u - v_N\|_{H^1_k(B_R)}, \tag{3.3}$$

with

$$C_{\rm qo} := \frac{2\big(\max\{A_{\max}, n_{\max}\} + C_{\rm DtN1}\big)}{A_{\min}} \tag{3.4}$$

Combining Theorem 3.4 with the results on $C_{\rm sol}(k)$ recapped in Theorem 2.7, we obtain the following specific examples of coefficients A and n when quasioptimality holds.

Corollary 3.5 *(Quasioptimality under specific conditions on A and n). Let $d = 2$ or 3, and let $k_0 > 0$.*

(i) *If A and n are nontrapping, 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 if (1.3) holds then, for all $k \ge k_0$, the Galerkin solution exists, is unique, and satisfies the quasi-optimal error bound (3.3) with $C_{\rm qo}$ given by (3.4).*
(ii) *If A is C^∞ and $n = 1$ then, given $\delta > 0$, there exist a set J with $|J| \le \delta$ and constants $\widetilde{C}_1, \widetilde{C}_2 > 0$, with all three depending on A, n, R, d, and k_0, but independent of k, and $\widetilde{C}_2$ additionally depending on δ and k_0 such that, for all $k \in [k_0,\infty)\setminus J$, if (1.3) holds (with C_1, C_2 replaced by $\widetilde{C}_1, \widetilde{C}_2$) then the Galerkin solution exists, is unique, and satisfies (3.3) with $C_{\rm qo}$ given by (3.4).*

For the plane-wave scattering problem (i.e. for $F(v)$ given by (2.6)), the regularity result

$$|u|_{H^2(B_R)} \le C_{\rm osc} k \|u\|_{H^1_k(B_R)} \tag{3.5}$$

was recently proved in [29, Theorem 9.1 and Remark 9.10], where $C_{\rm osc}$ depends on A, n, d, and R, but is independent of k. The polynomial approximation bounds in [35, §B] imply that, for the sequence of approximation spaces $(V_N)_{N=0}^\infty$ described in [35, §5],

$$\min_{v_N \in V_N} \|u - v_N\|_{H^1_k(B_R)} \le C_6 \frac{h}{p}\left(1 + \frac{kh}{p}\right)|u|_{H^2(B_R)} \tag{3.6}$$

where C_6 only depends on the constants in [35, Assumption 5.2] (which depend on the element maps from the reference element). Using (3.6) and (3.5) to bound the right-hand side of (3.3), we obtain the following bound on the relative error of the Galerkin solution.

Corollary 3.6 *(Bound on the relative error of the Galerkin solution). Let the assumptions of Theorem 3.4 hold and, furthermore, let $F(v)$ be given by (2.6) (so that u is the solution of the plane-wave scattering problem). If $C_{\rm sol}(k)$ is polynomially bounded (in the sense of Definition 2.6) for $k \in K \subset [k_0,\infty)$, then there exists $C_6 > 0$, independent of k, h, and p, such that if (1.3) holds, then, for all $k \in K$,*

$$\frac{\|u - u_N\|_{H^1_k(B_R)}}{\|u\|_{H^1_k(B_R)}} \le C_{\rm qo} C_6 C_{\rm osc} C_1 \big(1 + C_1\big), \tag{3.7}$$

with $C_{\rm qo}$ given by (3.4); i.e. the relative error can be made arbitrarily small by making C_1 smaller.

Remark 3.7 *(Theorem 3.1 is valid for solutions of a much larger class of PDEs).* Inspecting the proof of Theorem 3.1 below, we see that the conclusion, i.e. the decomposition $u = u_{H^2} + u_{\mathcal{A}}$ with u_{H^2} and $u_{\mathcal{A}}$ satisfying the bounds (3.1) and (3.2) respectively, holds under much weaker assumptions. Indeed, the conclusion still holds under the following three assumptions only.

(i) P_k is a family of properly-supported second-order pseudo-differential operators, with principal symbol $p_k(x,\zeta)$,
(ii) $p_k(x,\zeta)$ is coercive at infinity in the sense that

$$\liminf_{|\xi|\to\infty,\, x\in\mathbb{R}^d} \langle k\xi\rangle^{-2} p_k(x, k\xi) \ge c > 0, \tag{3.8}$$

where $c > 0$ does not depend on k, and
(iii) the solution to $P_k u = -f$, posed in $\mathbb{R}^d$ with $\operatorname{supp} f \subset B_R$ and $f \in L^2(B_R)$, satisfies the bound

$$\|u\|_{L^2(B_{R+2})} \le Ck^M \|f\|_{L^2(B_R)},$$

with C and M independent of k, u, and f. (In fact, the 2 in the $R+2$ on the left-hand side of the bound can be replaced by any number >0.)

In particular, no assumption is made about lower-order terms of P_k, or the behaviour of u at infinity (such as a radiation condition).

4. Recap of relevant results about semiclassical pseudodifferential operators

The proof of Theorem 3.1 relies on standard results about semiclassical pseudodifferential operators. We review these here, with our default references being [52] and [14, Appendix E]. Homogeneous – as opposed to semiclassical – versions of the results in this section can be found in, e.g., [47, Chapter 7], [41, Chapter 7], [25, Chapter 6].[1]

While the use of homogeneous pseudodifferential operators in numerical analysis is well established, see, e.g., [41], [25], there has been less use of semiclassical pseudodifferential operators. However, these are ideally-suited for studying the high-frequency behaviour of Helmholtz solutions. Indeed, semiclassical pseudodifferential operators are just pseudodifferential operators with a large/small parameter, and behaviour with respect to this parameter is then explicitly kept track of in the associated calculus.

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., [52, §1.2], [14, Page 82], [32, Chapter 1]). In this notation, the Helmholtz equation $\nabla\cdot(\mathsf{A}\nabla u) + k^2 nu = -f$ becomes

$$P_\hbar u = \hbar^2 f, \quad \text{where} \quad P_\hbar := -\hbar^2\nabla\cdot(\mathsf{A}\nabla\cdot) - n. \tag{4.1}$$

While some results in semiclassical analysis are valid in the limit $\hbar$ small, the results we recap in this section are valid for all $0<\hbar\le\hbar_0$ with $\hbar_0<\infty$ arbitrary.

The semiclassical Fourier transform $\mathcal{F}_\hbar$ The semiclassical Fourier transform is defined for $\hbar>0$ by

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

and its inverse by

$$\mathcal{F}_\hbar^{-1}\psi(x) := (2\pi\hbar)^{-d}\int_{\mathbb{R}^d} \exp\big(\mathrm{i}x\cdot\xi/\hbar\big)\psi(\xi)\,\mathrm{d}\xi; \tag{4.2}$$

see [52, §3.3]. Then

$$\mathcal{F}_\hbar\Big(\big(-\mathrm{i}\hbar\partial\big)^\alpha\phi\Big) = \xi^\alpha\,\mathcal{F}_\hbar\phi \tag{4.3}$$

and

$$\|\phi\|_{L^2(\mathbb{R}^d)} = \frac{1}{(2\pi\hbar)^{d/2}}\,\|\mathcal{F}_\hbar\phi\|_{L^2(\mathbb{R}^d)}. \tag{4.4}$$

[1] The counterpart of "semiclassical" involving differential/pseudodifferential operators without a small parameter is usually called "homogeneous" (owing to the homogeneity of the principal symbol) rather than "classical." "Classical" describes the behaviour in either calculus in the small-$\hbar$ or high-frequency limit respectively, where commutators of operators become Poisson brackets of symbols, hence classical particle dynamics replaces wave motion.

Semiclassical Sobolev spaces In the same way that it is convenient to work with the weighted H^1 norm (2.7) when studying the Helmholtz equation with parameter k, it is convenient to use norms weighted with $\hbar$ when studying (4.1). Therefore on the space

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

we use the norm

$$\|u\|^2_{H_\hbar^s(\mathbb{R}^d)} := (2\pi\hbar)^{-d}\int_{\mathbb{R}^d}\langle\xi\rangle^{2s}|\mathcal{F}_\hbar u(\xi)|^2\,\mathrm{d}\xi; \tag{4.5}$$

see [52, §8.3], [14, §E.1.8]. We abbreviate $H_\hbar^s(\mathbb{R}^d)$ to $H_\hbar^s$ and $L^2(\mathbb{R}^d)$ to L^2.

We record for later the fact that, by (4.3) and (4.4), for multiindices α,

$$\begin{aligned}\hbar^{|\alpha|}\,\|\partial^\alpha\phi\|_{L^2} &= \Big\|\big(-\mathrm{i}\hbar\partial\big)^\alpha\phi\Big\|_{L^2} = \frac{1}{(2\pi\hbar)^{d/2}}\,\|\xi^\alpha\,\mathcal{F}_\hbar\phi\|_{L^2}\\ &\le \frac{1}{(2\pi\hbar)^{d/2}}\Big\|\langle\xi\rangle^{|\alpha|}\,\mathcal{F}_\hbar\phi\Big\|_{L^2} = \|\phi\|_{H_\hbar^{|\alpha|}}.\end{aligned} \tag{4.6}$$

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 $\{(x,\xi): x\in\mathbb{R}^d, \xi\in\mathbb{R}^d\}$.

To deal with the behaviour of functions on phase space uniformly near $\xi=\infty$ (so-called *fiber infinity*), we consider the *radial compactification* in the ξ variable of $T^*\mathbb{R}^d$. This is defined by

$$\overline{T}^*\mathbb{R}^d := \mathbb{R}^d\times B^d,$$

where B^d denotes the closed unit ball, considered as the closure of the image of $\mathbb{R}^d$ under the radial compactification map

$$\mathsf{RC}: \xi\mapsto \xi/(1+\langle\xi\rangle);$$

see [14, §E.1.3]. Near the boundary of the ball, $|\xi|^{-1}\circ\mathsf{RC}^{-1}$ is a smooth function, vanishing to first order at the boundary, with $(|\xi|^{-1}\circ\mathsf{RC}^{-1}, \widehat{\xi}\circ\mathsf{RC}^{-1})$ thus giving local coordinates on the ball near its boundary. The boundary of the ball should be considered as a sphere at infinity consisting of all possible *directions* of the momentum variable. More generally, we denote $\overline{T}^*X := X\times B^d$ for $X\subset\mathbb{R}^d$, and where appropriate (e.g., in dealing with finite values of ξ only), we abuse notation by dropping the composition with RC from our notation and simply identifying $\mathbb{R}^d$ with the interior of B^d.

Symbols, quantisation, and semiclassical pseudodifferential operators A symbol is a function on $T^*\mathbb{R}^d$ that is also allowed to depend on $\hbar$, and thus can be considered as an $\hbar$-dependent family of functions. Such a family $a=(a_\hbar)_{0<\hbar\le\hbar_0}$, 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

$$|\partial_x^\alpha\partial_\xi^\beta a(x,\xi)| \le C_{\alpha,\beta}\langle\xi\rangle^{m-|\beta|} \quad \text{for all } (x,\xi)\in T^*\mathbb{R}^d \text{ and for all } 0<\hbar\le\hbar_0, \tag{4.7}$$

where $C_{\alpha,\beta}$ does not depend on $\hbar$, x, or ξ; see [52, p. 207], [14, §E.1.2]. In this paper, we only consider these symbol classes on $\mathbb{R}^d$, and so we abbreviate $S^m(\mathbb{R}^d)$ to S^m.

For $a\in S^m$, we define the *semiclassical quantisation* of a, $\mathrm{Op}_\hbar(a): \mathcal{S}(\mathbb{R}^d)\to\mathcal{S}(\mathbb{R}^d)$, by

$$\big(\mathrm{Op}_\hbar(a)v\big)(x) := (2\pi\hbar)^{-d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\exp\big(\mathrm{i}(x-y)\cdot\xi/\hbar\big)\,a(x,\xi)v(y)\,\mathrm{d}y\mathrm{d}\xi \tag{4.8}$$

for $v\in\mathcal{S}(\mathbb{R}^d)$; [52, §4.1] [14, Page 543]. The integral in (4.8) need not converge, and can be understood *either* as an oscillatory integral in the sense of [52, §3.6], [24, §7.8], *or* as an iterated integral, with the y integration performed first; see [14, Page 543].

Conversely, if A can be written in the form above, i.e. $A = \mathrm{Op}_\hbar(a)$ with $a \in S^m$, we say that A is a *semiclassical pseudo-differential operator of order* m and we write $A \in \Psi^m_\hbar$. We use the notation $a \in \hbar^l S^m$ if $\hbar^{-l} a \in S^m$; similarly $A \in \hbar^l \Psi^m_\hbar$ if $\hbar^{-l} A \in \Psi^m_\hbar$.

Theorem 4.1 *(Composition and mapping properties of semiclassical pseudo-differential operators [52, Theorem 8.10], [14, Proposition E.17 and Proposition E.19]). If $A \in \Psi^{m_1}_\hbar$ and $B \in \Psi^{m_2}_\hbar$, then*

(i) $AB \in \Psi^{m_1+m_2}_\hbar$,
(ii) $[A,B] := AB - BA \in \hbar \Psi^{m_1+m_2-1}_\hbar$,
(iii) *For any $s \in \mathbb{R}$, A is bounded uniformly in $\hbar$ as an operator from $H^s_\hbar$ to $H^{s-m_1}_\hbar$.*

Residual class We say that $A = O(\hbar^\infty)_{\Psi^{-\infty}}$ if, for any $s > 0$ and $N \geq 1$, there exists $C_{s,N} > 0$ so that

$$\|A\|_{H^{-s}_\hbar \to H^s_\hbar} \leq C_{N,s} \hbar^N; \tag{4.9}$$

i.e. $A \in \Psi^{-\infty}_\hbar$ and furthermore all of its operator norms are bounded by any algebraic power of $\hbar$.

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 $\hbar S^{m-1}$. For any m, there is a linear, surjective map

$$\sigma^m_\hbar : \Psi^m_\hbar \to S^m/\hbar S^{m-1},$$

called the *principal symbol map*, such that, for $a \in S^m$,

$$\sigma^m_\hbar\big(\mathrm{Op}_\hbar(a)\big) = a \quad \mod \hbar S^{m-1}; \tag{4.10}$$

see [52, Page 213], [14, Proposition E.14] (observe that (4.10) implies that $\ker(\sigma^m_\hbar) = \hbar \Psi^{m-1}_\hbar$).

When applying the map $\sigma^m_\hbar$ to elements of $\Psi^m_\hbar$, 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). Key properties of the principal symbol that we use below are that

$$\sigma_\hbar(AB) = \sigma_\hbar(A)\sigma_\hbar(B), \tag{4.11}$$

$$\sigma_\hbar(P_\hbar) = \langle \mathsf{A}\xi, \xi\rangle - n, \tag{4.12}$$

where $\langle \cdot,\cdot\rangle$ denotes the ℓ^2 inner product on $\mathbb{R}^d$. The property (4.11) is proved in [14, Proposition E.17], (4.12) follows from (4.10) since $P_\hbar = \mathrm{Op}_\hbar\big(\langle \mathsf{A}\xi,\xi\rangle - n - i\hbar\xi_\ell \partial_j A_{j\ell}\big)$ (where we sum over the indices j and ℓ).

Operator wavefront set $\mathrm{WF}_\hbar$ We say that $(x_0,\xi_0) \in \overline{T}^*\mathbb{R}^d$ is *not* in the *semiclassical operator wavefront set* of $A = \mathrm{Op}_\hbar(a) \in \Psi^m_\hbar$, 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 $C_{\alpha,\beta,U,N} > 0$ (independent of $\hbar$) so that, for all $0 < \hbar \leq \hbar_0$,

$$|\partial^\alpha_x \partial^\beta_\xi a(x,\xi)| \leq C_{\alpha,\beta,U,N} \hbar^N \langle\xi\rangle^{-N} \quad \text{for all } (x, \mathrm{RC}(\xi)) \in U; \tag{4.13}$$

i.e. outside its semiclassical operator wavefront set an operator vanishes faster than any algebraic power of both $\hbar$ and $\langle\xi\rangle^{-1}$; see [52, Page 194], [14, Definition E.27]. Three properties of the semiclassical operator wavefront set that we use below are

$$\mathrm{WF}_\hbar(AB) \subset \mathrm{WF}_\hbar A \cap \mathrm{WF}_\hbar B \tag{4.14}$$

(see [52, §8.4], [14, E.2.5]),

$$\mathrm{WF}_\hbar\big(\mathrm{Op}_\hbar(a)\big) \subset \operatorname{supp} a \tag{4.15}$$

(since $(\operatorname{supp} a)^c \subset (\mathrm{WF}_\hbar(\mathrm{Op}_\hbar(a)))^c$ by (4.13)), and

$$\mathrm{WF}_\hbar A = \emptyset \iff A = O(\hbar^\infty)_{\Psi^{-\infty}} \tag{4.16}$$

(see [14, E.2.2]).

Compactly-supported operators We say that A is *compactly supported* if its Schwartz kernel is compactly supported in some set $K \Subset \mathbb{R}^d \times \mathbb{R}^d$, for all $0 < \hbar \leq \hbar_0$. We recall that if $\mathcal{D}(\mathbb{R}^d) := C^\infty_{\mathrm{comp}}(\mathbb{R}^d)$ (i.e. the set of test functions) and $\mathcal{D}'(\mathbb{R}^d)$ denote the set of linear functionals on $\mathcal{D}(\mathbb{R}^d)$ (i.e. the set of distributions), given a bounded, sequentially-continuous operator $A : \mathcal{D} \to \mathcal{D}'$ there exists a *Schwartz kernel* $\mathcal{K}_A \in \mathcal{D}'(\mathbb{R}^d \times \mathbb{R}^d)$ such that

$$Av(x) = \int_{\mathbb{R}^d} \mathcal{K}_A(x,y) v(y)\, \mathrm{d}y,$$

in the sense of distributions; see, e.g., [24, Theorem 5.2.1], [14, §A.7]. We use below the facts that

- A is compactly supported iff there exist $\chi_1, \chi_2 \in \mathcal{D}$ such that $A = \chi_1 A \chi_2$, thus
- if $\chi_1, \chi_2 \in \mathcal{D}$ are compactly supported functions, then $\chi_1 A \chi_2$ is compactly supported, and
- if P is a differential operator and $\chi \in \mathcal{D}$, then both χP and $P\chi$ are compactly supported.

Ellipticity We say that $B \in \Psi^m_\hbar$ is *elliptic* on $X \subset \overline{T}^*\mathbb{R}^d$ if there exists $c > 0$, independent of $\hbar$, such that

$$\langle\xi\rangle^{-m} \Big|\sigma_\hbar(B)(x,\xi)\Big| \geq c, \quad \text{for all } (x, \mathrm{RC}(\xi)) \in X \text{ and for all } 0 < \hbar \leq \hbar_0. \tag{4.17}$$

A key feature of elliptic operators is that they are microlocally invertible; this is reflected in the following result.

Proposition 4.2 *(Elliptic parametrix [14, Proposition E.32]).* [2]*Let $A \in \Psi^m_\hbar$ and $B \in \Psi^\ell_\hbar$ be such that B is elliptic on $\mathrm{WF}_\hbar(A)$. Then there exist $Q, Q' \in \Psi^{m-\ell}_\hbar$ such that*

$$A = BQ + O(\hbar^\infty)_{\Psi^{-\infty}} = Q'B + O(\hbar^\infty)_{\Psi^{-\infty}}.$$

Theorem 4.3 *(Elliptic estimate [14, Theorem E.33]).* [2]*Let $A \in \Psi^{m_1}_\hbar$, $B_1 \in \Psi^{m_2}_\hbar$, and $P \in \Psi^\ell_\hbar$ be so that $B_1 P$ is elliptic on $\mathrm{WF}_\hbar(A)$.*

(i) *Given $s, N > 0$, and $M > 0$, if $v \in \mathcal{D}'$ and $B_1 P v \in H^{s-m_2-\ell}$ then $Av \in H^{s-m_1}$ and there exists $C_s > 0$, $C_{N,M,s} > 0$ (independent of v and $\hbar$) such that*

$$\|Av\|_{H^{s-m_1}_\hbar} \leq C_s \left\|B_1 P v\right\|_{H^{s-m_2-\ell}_\hbar} + C_{N,M,s} \hbar^M \|v\|_{H^{-N}_\hbar}. \tag{4.18}$$

(ii) *If, in addition, A and $B_1 P$ are compactly supported, then there exists $\widetilde{\chi} \in C^\infty_{\mathrm{comp}}$ so that*

$$\|Av\|_{H^{s-m_1}_\hbar} \leq C_s \left\|B_1 P v\right\|_{H^{s-m_2-\ell}_\hbar} + C_{N,M,s} \hbar^M \left\|\widetilde{\chi} v\right\|_{H^{-N}_\hbar}. \tag{4.19}$$

Part (i) of Theorem 4.3 is proved by using Proposition 4.2 with $B = B_1 P \in \Psi^{m_2+\ell}_\hbar$, applying the resulting operator equation to v, and taking norms. The operator $Q' \in \Psi^{m_1-m_2-\ell}_\hbar$ and the constant C_s is then $\|Q'\|_{H^{s-m_2-\ell}_\hbar \to H^{s-m_1}_\hbar}$. The proof of Part (ii) is similar, using that, since A and $B_1 P$ are both compactly supported, there exists $\widetilde{\chi} \in C^\infty_{\mathrm{comp}}$ such that $(A - B_1 P)v = (A - B_1 P)\widetilde{\chi} v$.

[2] We highlight that working in $\mathbb{R}^d$ (as opposed to on a general manifold defined by coordinate charts) allows us to remove the proper-support assumption appearing in [14, Proposition E.32, Theorem E.33].

5. Proof of Theorem 3.1

In the notation introduced in §4, Theorem 3.1 becomes the following.

Theorem 5.1. *Let* A *and* n *satisfy Assumption 2.1 and let* $R > 0$ *be such that* $\operatorname{supp}(\mathsf{I} - \mathsf{A}) \cup \operatorname{supp}(1 - n) \Subset B_R$. *Given* $f \in L^2(B_R)$, *let* u *satisfy* $P_\hbar u = \hbar^2 f$ *in* $\mathbb{R}^d$ *and the Sommerfeld radiation condition (1.2). Assume that, given* $k_0 > 0$, $C_{\rm sol}(k)$ *is polynomially bounded (in the sense of Definition 2.6) for* $k \in K \subset [k_0, \infty)$. *Given* $k_0 > 0$, *let* $\hbar_0 := k_0^{-1}$, *and let* $H := \{k^{-1} : k \in K\} \subset (0, \hbar_0]$.

Then there exist $C_3, C_4, C_5 > 0$ *such that*

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

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

$$\|\partial^\alpha u_{H^2}\|_{L^2(B_R)} \le C_3 \hbar^{2-|\alpha|} \|f\|_{L^2(B_R)}$$

$$\text{for all } |\alpha| \le 2 \text{ and for all } \hbar \in H \subset (0, \hbar_0], \tag{5.1}$$

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

$$\left\|\partial^\beta u_{\mathcal{A}}\right\|_{L^2(B_R)} \le C_{\rm sol}\left(\hbar^{-1}; R+2\right) C_4 \left(\frac{\hbar}{C_5}\right)^{1-|\beta|} \|f\|_{L^2(B_R)}$$

$$\text{for all } \beta \text{ and for all } \hbar \in H \subset (0, \hbar_0], \tag{5.2}$$

where C_3, C_4, *and* C_5 *depend on* A, n, d, *and* $\hbar_0$, *but are independent of* $\hbar$, f, α, *and* β.

5.1. Step 0: restatement of bounds on the solution operator in semiclassical notation

The definition of $C_{\rm sol}$ (Definition 2.5) implies that, in semiclassical notation,

$$\|u\|_{H^1_\hbar(B_R)} \le \hbar\, C_{\rm sol}(\hbar^{-1}) \|f\|_{L^2(B_R)} \qquad \text{for all } \hbar > 0. \tag{5.3}$$

It is convenient to record here in semiclassical notation the bound on the solution operator when $C_{\rm sol}$ is polynomially bounded.

Lemma 5.2 *(Polynomial boundedness rewritten in terms of* $\hbar$*). Given* $f \in L^2_{\rm comp}(\mathbb{R}^d)$, *let* $u \in H^1_{\rm loc}(\mathbb{R}^d)$ *be the solution to*

$$P_\hbar u = \hbar^2 f$$

satisfying the Sommerfeld radiation condition (1.2) (with $k = \hbar^{-1}$*).*

If $C_{\rm sol}(k)$ *is polynomially bounded for* $k \in K \subset [k_0, \infty)$ *(in the sense of Definition 2.6), then there exists* $M > 0$ *(independent of* $\hbar$*) such that, given* $\chi \in C^\infty_{\rm comp}(\mathbb{R}^d)$, *there exists* $C > 0$ *(independent of* $\hbar$ *but dependent on* χ*) such that*

$$\|\chi u\|_{L^2} \le C \hbar^{1-M} \|f\|_{L^2} \qquad \text{for } \hbar \in H \subset (0, \hbar_0], \tag{5.4}$$

where $\hbar_0 := k_0^{-1}$ *and* $H := \{k^{-1} : k \in K\}$.

The bound (5.4) also holds with $\|\chi u\|_{L^2}$ replaced by $\|\chi u\|_{H^1_\hbar}$, but we only need it in the form (5.4) for what follows.

5.2. Step 1: the definitions of $u_{\mathcal{A}}$ and u_{H^2}

The cut-off functions χ *and* χ_μ Let $\chi \in C^\infty_{\rm comp}(\mathbb{R}^d; [0,1])$ be such that

$$\chi = \begin{cases} 1 & \text{in } B_1 \\ 0 & \text{outside } B_2. \end{cases} \tag{5.5}$$

For $\mu > 0$, let

$$\chi_\mu(\cdot) := \chi\left(\frac{\cdot}{\mu}\right). \tag{5.6}$$

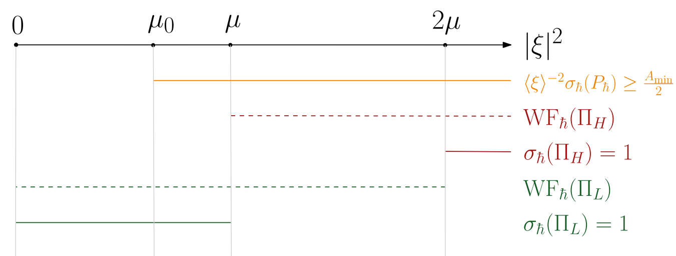

Fig. 5.1. The locations of $\mathrm{WF}_\hbar(\Pi_H)$ and $\mathrm{WF}_\hbar(\Pi_L)$, the regions where the principal symbols of Π_H and Π_L equal one, and the region where $P_\hbar$ is elliptic.

We define $\mu_0 = \mu_0(\mathsf{A}, n)$ by

$$\mu_0(\mathsf{A}, n) := \left(1 + \frac{2 n_{\max}}{A_{\min}}\right). \tag{5.7}$$

The reason for this definition is that it implies that

$$\text{if} \quad |\xi|^2 \ge \mu_0 \quad \text{then} \quad \langle\xi\rangle^{-2}\sigma_\hbar(P) \ge \frac{A_{\min}}{2} > 0. \tag{5.8}$$

Indeed, by (4.12),

$$\langle\xi\rangle^{-2}\sigma_\hbar(P) \ge \frac{A_{\min}|\xi|^2 - n_{\max}}{1+|\xi|^2} = \frac{A_{\min}}{2} + \left(\frac{A_{\min}}{2}\right)\left(\frac{|\xi|^2 - 1 - 2n_{\max}/A_{\min}}{1+|\xi|^2}\right),$$

and (5.8) follows. The importance of the property (5.8) is explained at the end of this subsection.

The frequency cut-offs Π_L *and* Π_H We define Π_L and Π_H, the projections on low and high frequencies respectively, by (1.9) and (1.10). The definition of the quantisation $\mathrm{Op}_\hbar$ (4.8) and the change of variable $\zeta = \xi/\hbar$ imply that

$$\Pi_L = \mathrm{Op}_\hbar\left(\chi_\mu(|\xi|^2)\right) \tag{5.9}$$

and

$$\Pi_H = I - \Pi_L. \tag{5.10}$$

These definitions and the definition of $\Psi^m_\hbar(\mathbb{R}^d)$ in §4 imply that $\Pi_L \in \Psi^{-\infty}_\hbar(\mathbb{R}^d)$ and $\Pi_H \in \Psi^0_\hbar(\mathbb{R}^d)$.

The locations of the wavefront sets of the frequency cut-offs, and the regions where their symbols equal one In Fig. 5.1 we show, as functions of $|\xi|^2$, the locations of $\mathrm{WF}_\hbar(\Pi_H)$ and $\mathrm{WF}_\hbar(\Pi_L)$, and the regions where $\sigma_\hbar(\Pi_H)$, and $\sigma_\hbar(\Pi_L)$ equal one. These locations/regions are obtained using (4.15) and (4.10) respectively. For example, since $1 - \chi_\mu(|\xi|^2) = 1$ for $|\xi|^2 \ge 2\mu$ and $= 0$ for $|\xi|^2 \le \mu$, (4.10) and (4.15) imply that

$$\sigma_\hbar(\Pi_H) = 1 \text{ on } \left\{\xi : |\xi|^2 \ge 2\mu\right\} \quad \text{and} \quad \mathrm{WF}_\hbar(\Pi_H) \subset \left\{\xi : |\xi|^2 \ge \mu\right\}. \tag{5.11}$$

We also record the following key consequence of the results summarised in Fig. 5.1.

Lemma 5.3. *If* $\mu \ge \mu_0$, *then* $P_\hbar$ *is elliptic on* $\mathrm{WF}_\hbar(\Pi_H)$.

This property is central to our proof of the bound (5.1) on u_{H^2}, i.e., the high-frequency component. It is a consequence of (5.8), and the reason why we choose μ_0 as in (5.7) is for this ellipticity result to hold.

The definitions of $u_{\mathcal{A}}$ *and* u_{H^2} As described in §1.2, we choose $\varphi \in C^\infty_{\rm comp}(\mathbb{R}^d)$ be equal to one on B_{R+1} and vanish outside B_{R+2}. We then let

$$w := \varphi u$$

and we define

$$u_{\mathcal{A}} := (\Pi_L w)\Big|_{B_R} \quad \text{and} \quad u_{H^2} := (\Pi_H w)\Big|_{B_R}.$$

5.3. Step 2: proof of the bound (5.2) on u_A (the low-frequency component)

Since $\Pi_L \in \Psi_\hbar^{-\infty}$, Part (iii) of Theorem 4.1, together with Sobolev embedding, gives $\Pi_L w \in C^\infty$.

The definition of Π_L (1.9) and Plancherel's identity (4.4) for the standard (i.e. non semiclassical) Fourier transform imply that

$$\left\|\partial^\beta\left(\Pi_L w\right)\right\|_{L^2} = \frac{1}{(2\pi)^{d/2}}\left\|(\cdot)^\beta \mathcal{F}(\Pi_L w)(\cdot)\right\|_{L^2} = \frac{1}{(2\pi)^{d/2}}\left\|(\cdot)^\beta \chi_\mu\left(\hbar^2|\cdot|^2\right)\mathcal{F}w(\cdot)\right\|_{L^2}. \tag{5.12}$$

The definitions of χ (5.5) and χ_μ (5.6) imply that $\chi_\mu(\xi)=0$ for $|\xi|\geq 2\mu$, so

$$\chi_\mu(\hbar^2|\zeta|^2)=0 \quad \text{for } |\zeta|\geq \sqrt{2\mu}\,\hbar^{-1}.$$

Using this fact, and then (in this order) the fact that $|\chi_\mu|\leq 1$, Plancherel's identity for the standard Fourier transform, the fact that $\varphi=0$ outside B_{R+2}, and the definition of $C_{\rm sol}$ (2.12), we find from (5.12) that

$$\begin{aligned}\left\|\partial^\beta\left(\Pi_L \varphi u\right)\right\|_{L^2} &\leq \frac{(2\mu)^{|\beta|/2}}{(2\pi)^{d/2}}\hbar^{-|\beta|}\left\|\chi_\mu\left(\hbar^2|\cdot|^2\right)\mathcal{F}(\varphi u)(\cdot)\right\|_{L^2}\\ &\leq \frac{(2\mu)^{|\beta|/2}}{(2\pi)^{d/2}}\hbar^{-|\beta|}\left\|\mathcal{F}(\varphi u)\right\|_{L^2}\\ &\leq (2\mu)^{|\beta|/2}\hbar^{-|\beta|}\left\|\varphi u\right\|_{L^2}\\ &\leq (2\mu)^{|\beta|/2}\hbar^{-|\beta|}\hbar\, C_{\rm sol}(\hbar^{-1};R+2)\left\|f\right\|_{L^2(B_R)}.\end{aligned}$$

Since

$$\left\|\partial^\beta u_A\right\|_{L^2(B_R)} = \left\|\partial^\beta(\Pi_L w)\right\|_{L^2(B_R)} \leq \left\|\partial^\beta(\Pi_L w)\right\|_{L^2},$$

the bound (5.2) then follows with $C_4 := \sqrt{2\mu}$ and $C_5 := \sqrt{2\mu}$.

5.4. Step 3: proof of the bound (5.1) on u_{H^2} (the high-frequency component)

By the inequality (4.6), it is sufficient to prove that

$$\left\|\Pi_H w\right\|_{H_\hbar^2} \leq C_3\hbar^2\left\|f\right\|_{L^2(B_R)} \qquad \text{for all } \hbar\in H\subset(0,\hbar_0). \tag{5.13}$$

It is instructive to first prove (5.13) under the assumption that $C_{\rm sol}(k)\lesssim 1$ (which, by Theorem 2.7 is ensured if A and n are nontrapping). Indeed, as discussed in §1.2, this proof only requires that $P_\hbar$ is elliptic on $\mathrm{WF}_\hbar(\Pi_H)$; i.e., Lemma 5.3. Throughout the rest of this section, therefore, we assume that $\mu\geq\mu_0$, so that the result of Lemma 5.3 holds.

5.4.1. Proof of (5.13) under the assumption that $C_{\rm sol}(k)\lesssim 1$

We seek to apply Part (i) of Theorem 4.3 with $A=\Pi_H$ (so $m_1=0$), $B_1=1$ (so $m_2=0$), and $P=P_\hbar$ (so $\ell=2$). By Lemma 5.3, B_1P is elliptic on $\mathrm{WF}_\hbar(A)$. We can therefore apply Theorem 4.3 and obtain that, given $N,N'>0$,

$$\left\|\Pi_H w\right\|_{H_\hbar^2} \lesssim \left\|P_\hbar w\right\|_{L^2} + \hbar^{N'}\left\|w\right\|_{H_\hbar^{-N}}, \tag{5.14}$$

where the omitted constant in $\lesssim$ depends on N and N'. Since $P_\hbar u=\hbar^2 f$,

$$P_\hbar w = [P_\hbar,\varphi]u + \hbar^2\varphi f,$$

where $[\cdot,\cdot]$ is the standard commutator defined by $[A_1,A_2]:=A_1A_2-A_2A_1$, so that (5.14) becomes

$$\left\|\Pi_H w\right\|_{H_\hbar^2} \lesssim \left\|[P_\hbar,\varphi]u\right\|_{L^2} + \hbar^2\left\|f\right\|_{L^2} + \hbar^{N'}\left\|w\right\|_{H_\hbar^{-N}}. \tag{5.15}$$

Direct calculation, using the fact that $\operatorname{supp}\varphi\subset B_{R+2}$, implies that

$$\left\|[P_\hbar,\varphi]u\right\|_{L^2} \lesssim \hbar\left\|u\right\|_{H_\hbar^1(B_{R+2})}, \tag{5.16}$$

where the omitted constant depends on φ, and hence on R.

Combining (5.15) and (5.16), and recalling that $\operatorname{supp}\varphi\subset B_{R+2}$, we have

$$\left\|\Pi_H w\right\|_{H_\hbar^2} \lesssim \hbar\left\|u\right\|_{H_\hbar^1(B_{R+2})} + \hbar^2\left\|f\right\|_{L^2(B_R)} + \hbar^{N'}\left\|u\right\|_{H_\hbar^{-N}(B_{R+2})}.$$

Choosing $N=0$ and $N'=1$, and then using (5.3), we obtain

$$\left\|\Pi_H w\right\|_{H_\hbar^2} \lesssim \hbar^2\left(1+C_{\rm sol}(\hbar^{-1})\right)\left\|f\right\|_{L^2(B_R)}. \tag{5.17}$$

If $C_{\rm sol}(\hbar^{-1})\lesssim 1$, then this implies (5.13). However, if $C_{\rm sol}(\hbar^{-1})\gg 1$ (as occurs when $C_{\rm sol}$ is polynomially bounded in the sense of Definition 2.6 with $M>0$) then (5.17) is a weaker bound than (5.13).

5.4.2. Proof of (5.13) under the assumption that $C_{\rm sol}(k)$ is polynomially bounded

Inspecting the argument in §5.4.1, we see that the assumption that $C_{\rm sol}(k)\lesssim 1$ is needed to get a good bound on the commutator term $[P_\hbar,\varphi]u$. To remove this commutator term, one idea is to use the elliptic estimate in Part (i) of Theorem 4.3, using the fact that $P_\hbar$ is elliptic on $\mathrm{WF}_\hbar(\Pi_H\varphi)$, and apply the estimate with $v:=u$. However, the error term would not be compactly supported and we would be unable to control it using the polynomial bound on the solution operator (5.4). We therefore introduce additional spatial cut-offs on the left of $\Pi_H\varphi$ and $P_\hbar$ to create compactly-supported operators and have a compactly-supported error term thanks to Part (ii) of Theorem 4.3.

To this end, let φ_1, $\varphi_2\in C^\infty_{\rm comp}(\mathbb{R}^d)$ be such that $\varphi_1=1$ on $\operatorname{supp}\varphi$ and $\varphi_2=1$ on $\operatorname{supp}\varphi_1$; we then write

$$\Pi_H\varphi u = (1-\varphi_1)\Pi_H\varphi u + \varphi_1\Pi_H\varphi u. \tag{5.18}$$

Since $1-\varphi_1=0$ on $\operatorname{supp}\varphi$, using (4.14) and (4.15), we obtain that

$$\mathrm{WF}_\hbar\left((1-\varphi_1)\Pi_H\varphi\right)\subset \overline{T}^*(\operatorname{supp}(1-\varphi_1))\cap\overline{T}^*(\operatorname{supp}\varphi)=\emptyset.$$

Hence, by (4.16), $(1-\varphi_1)\Pi_H\varphi = O(\hbar^\infty)_{\Psi^{-\infty}}$, and, by the definition of the residual class (4.9), for any $N\geq 1$ there exists $C_N>0$ so that

$$\|(1-\varphi_1)\Pi_H\varphi u\|_{H_\hbar^2} = \|(1-\varphi_1)\Pi_H\varphi\varphi_1 u\|_{H_\hbar^2} \leq C_N\hbar^N\|\varphi_1 u\|_{L^2}, \tag{5.19}$$

were we used the fact that $\varphi_1=1$ on $\operatorname{supp}\varphi$ in the first equality.

It therefore remains to control $\varphi_1\Pi_H\varphi u$; to do this, we use the elliptic estimate of Theorem 4.3.

Lemma 5.4. *$\varphi_2P_\hbar$ is elliptic on* $\mathrm{WF}_\hbar(\varphi_1\Pi_H\varphi)$.

Proof. By (4.14) and (4.15), $\mathrm{WF}_\hbar(\varphi_1\Pi_H\varphi)\subset\overline{T}^*(\operatorname{supp}\varphi_1)\cap\mathrm{WF}_\hbar\Pi_H$. Since $\varphi_2=1$ on $\operatorname{supp}\varphi_1$, the result is a direct consequence of Lemma 5.3. ■

By the facts about compactly-supported operators recalled in §4, $\varphi_1\Pi_H\varphi$ and $\varphi_2P_\hbar$ are compactly supported. Therefore, by Lemma 5.4, we can apply Part (ii) of Theorem 4.3 with $A=\varphi_1\Pi_H\varphi$, $B_1=\varphi_2$, $P=P_\hbar$, $m_1=0$, $m_2=0$, $\ell=2$. This result implies that there exists $\widetilde{\chi}\in C^\infty_{\rm comp}$, and, for any $N'\geq 1$, there exists $C_{N'}>0$ such that

$$\begin{aligned}\|\varphi_1\Pi_H\varphi u\|_{H_\hbar^2} &\lesssim \|\varphi_2P_\hbar u\|_{L^2} + C_{N'}\hbar^{N'}\|\widetilde{\chi}u\|_{L^2}\\ &= \hbar^2\|\varphi_2 f\|_{L^2} + C_{N'}\hbar^{N'}\|\widetilde{\chi}u\|_{L^2}.\end{aligned} \tag{5.20}$$

Collecting (5.18), (5.19), (5.20), using (5.4), and choosing $N=N'=M+1$, we obtain (5.13).

6. Proof of Theorem 3.4

The two ingredients for the proof of Theorem 3.4 are

- Lemma 6.4, which is the standard duality argument giving a condition for quasi-optimality to hold in terms of how well the solution of the adjoint problem is approximated by the finite-element space (measured by the quantity $\eta(V_N)$ defined by (6.3)), and
- Lemma 6.5 that bounds $\eta(V_N)$ using the decomposition from Theorem 3.1.

Regarding Lemma 6.4: we recall that this argument came out of ideas introduced in [43], was then formalised in [42], and has been used extensively in the analysis of the Helmholtz FEM; see, e.g., [1,26,33,42,35,36,51,50,12,9,31,10,17,21,16].

Before stating Lemma 6.4 we need to introduce some notation.

Definition 6.1 *(The adjoint sesquilinear form $a^*(\cdot,\cdot)$).* The adjoint sesquilinear form, $a^*(u,v)$, to the sesquilinear form $a(\cdot,\cdot)$ defined in (2.4) is given by

$$a^*(u,v) := \overline{a(v,u)} = \int_{B_R} \left((\mathsf{A}\nabla u)\cdot \overline{\nabla v} - k^2 n u \overline{v} \right) - \left\langle \gamma u, \mathrm{DtN}_k(\gamma v)\right\rangle_{\partial B_R}.$$

A key role is played by the solution operator of the adjoint variational problem with data in $L^2(B_R)$; we therefore introduce the following notation.

Definition 6.2 *(Adjoint solution operator $\mathcal{S}^*$).* Given $f\in L^2(B_R)$, let $\mathcal{S}^* f$ be defined as the solution of the variational problem

$$\text{find } \mathcal{S}^* f \in H^1(B_R) \quad \text{such that}$$
$$a^*(\mathcal{S}^* f, v) = \int_{B_R} f\,\overline{v} \quad \text{for all } v\in H^1(B_R). \tag{6.1}$$

Green's second identity applied to solutions of the Helmholtz equation satisfying the Sommerfeld radiation condition (1.2) implies that $\left\langle \mathrm{DtN}_k \psi, \overline{\phi}\right\rangle_{\partial B_R} = \left\langle \mathrm{DtN}_k \phi, \overline{\psi}\right\rangle_{\partial B_R}$ (see, e.g., [46, Lemma 6.13]); thus $a(\overline{v},u) = a(\overline{u},v)$ and so the definition (6.1) implies that

$$a(\overline{\mathcal{S}^* f}, v) = (\overline{f}, v)_{L^2(B_R)} \quad \text{for all } v\in H^1(B_R). \tag{6.2}$$

Definition 6.3 *($\eta(V_N)$).* Given a sequence of finite-dimensional spaces $(V_N)_{N=0}^\infty$ (as described in §2.3), let

$$\eta(V_N) := \sup_{0\neq f\in L^2(B_R)} \min_{v_N\in V_N} \frac{\left\| \mathcal{S}^* f - v_N\right\|_{H^1_k(B_R)}}{\left\| f\right\|_{L^2(B_R)}}. \tag{6.3}$$

Lemma 6.4 *(Conditions for quasi-optimality). If*

$$k\,\eta(V_N) \le \frac{1}{C_{\rm cont}}\sqrt{\frac{A_{\min}}{2\left(n_{\max}+A_{\min}\right)}},$$

then the Galerkin equations (2.15) have a unique solution which satisfies

$$\left\| u-u_h\right\|_{H^1_k(B_R)} \le \frac{2C_{\rm cont}}{A_{\min}}\left(\min_{v_N\in V_N} \left\| u-v_N\right\|_{H^1_k(B_R)}\right).$$

Proof. Using the inequality (2.10), we see that $a(\cdot,\cdot)$ satisfies the Gårding inequality

$$\Re\left(a(v,v)\right) \ge A_{\min}\left\| v\right\|^2_{H^1_k(B_R)} - 2k^2\left(n_{\max}+A_{\min}\right)\left\| v\right\|^2_{L^2(B_R)} \tag{6.4}$$

and the result follows from, e.g., the account [46, Theorem 6.32] of the standard duality argument with (in the notation of [46]) $\alpha = A_{\min}$ and $C_{\mathcal{V}} = 2k^2\left(n_{\max}+A_{\min}\right)$. ■

Lemma 6.5 *(Bound on $\eta(V_N)$ using the decomposition from Theorem 3.1). Let A and n satisfy Assumption 2.1 and let $R>0$ be such that* $\operatorname{supp}(\mathsf{I}-\mathsf{A})\cup\operatorname{supp}(1-n)\Subset B_R$. *Let $(V_N)_{N=0}^\infty$ be the piecewise-polynomial approximation spaces described in [35, §5]. There exists $C_6, C_7, \sigma>0$, all independent of k, h, and p, such that*

$$k\,\eta(V_N) \le C_6 C_3 \frac{hk}{p}\left(1+\frac{kh}{p}\right) + C_7 C_{\rm sol}(k)\left[\left(\frac{h}{h+\sigma}\right)^p\left(1+\frac{hk}{h+\sigma}\right) + k\left(\frac{kh}{\sigma p}\right)^p\left(\frac{1}{p}+\frac{kh}{\sigma p}\right)\right]. \tag{6.5}$$

The constants C_6 and σ only depend on the constants in [35, Assumption 5.2] defining the element maps from the reference element; C_7 depends on these constants, and additionally on C_5.

Proof. This proof is very similar to the proof of [35, Theorem 5.5]. Indeed, [35, Theorem 5.5] proves a bound very similar to (6.5) starting from bounds almost identical to the bounds (3.1) and (3.2) (recalling Remark 3.3 about notation). The only difference is that the bound (3.2) contains $C_{\rm sol}$, which depends on k (whereas in [35] $C_{\rm sol}\sim 1$), and so we now need to keep track of how $C_{\rm sol}$ enters the proof of [35, Theorem 5.5].

From the definition (6.3), it is sufficient to show that, given $f\in L^2(B_R)$, there exists $w_N\in V_N$ such that

$$\left\| \mathcal{S}^* f - w_N\right\|_{H^1_k(B_R)} \le C\left\| f\right\|_{L^2(B_R)}, \tag{6.6}$$

where C is the right-hand side of (6.5) divided by k. Let $v := \mathcal{S}^* f$; by (6.2) and Part (i) of Lemma 2.3, $\overline{v}$ satisfies the assumptions of Theorem 3.1 with f replaced by $\overline{f}$, and so the bounds (3.1) and (3.2) hold with u replaced by v.

By [35, First equation on Page 1896] (which uses [35, Theorem B.4]), the bound (3.6) holds, and thus there exists $w_N^{(1)}\in V_N$ such that

$$\left\| v_{H^2} - w_N^{(1)}\right\|_{H^1_k(B_R)} \le C_6\frac{h}{p}\left(1+\frac{kh}{p}\right)|v|_{H^2(B_R)}$$

and so

$$\left\| v_{H^2} - w_N^{(1)}\right\|_{H^1_k(B_R)} \le C_6\frac{h}{p}\left(1+\frac{kh}{p}\right)C_3\left\| f\right\|_{L^2(B_R)} \tag{6.7}$$

by (3.1).

For the approximation of $v_{\mathcal{A}}$, the only change to the argument in [35] is that a multiplicative factor of $(C_{\rm sol})^2$ must be included on the right-hand side of [35, Equation 5.8]. Then [35, Equations 5.8 and 5.9] implies that there exists C_7 and $w_N^{(2)}\in V_N$ such that

$$k\left\| v_{\mathcal{A}} - w_N^{(2)}\right\|_{H^1_k(B_R)} \le C_7 C_{\rm sol}(k)\left[\left(\frac{h}{h+\sigma}\right)^p\left(1+\frac{hk}{h+\sigma}\right) + k\left(\frac{kh}{\sigma p}\right)^p\left(\frac{1}{p}+\frac{kh}{\sigma p}\right)\right]\left\| f\right\|_{L^2(B_R)} \tag{6.8}$$

(observe that this equation is identical to [35, Last equation on Page 1896] except for the factor $C_{\rm sol}$ on the right-hand side).

Let $w_N := w_N^{(1)} + w_N^{(2)}$. By the triangle inequality, the decomposition $v = v_{H^2} + v_{\mathcal{A}}$ on B_R, and the inequalities (6.7) and (6.8), the inequality (6.6) holds with C the right-hand side of (6.5) and the proof is complete. ■

Corollary 6.6 *(Conditions under which $k\,\eta(V_N)$ is arbitrarily small). Let the assumptions of Lemma 6.5 hold. Given $\varepsilon>0$ and $k_0>0$, there exists $C_1, C_2>0$, depending only on $\varepsilon, C_3, C_6, C_7, \sigma$, and k_0, such that if*

$$\frac{hk}{p}\le C_1 \quad \text{and} \quad p \ge C_2\left(1+\log k + \log\left(C_{\rm sol}(k)\right)\right),$$

then

$$k\,\eta(V_N)\le \varepsilon \quad \text{for all } k\ge k_0.$$

Proof. This proof is essentially identical to the proofs of [35, Corollary 5.6] and [36, Theorem 5.8]. First choose C_1 sufficiently small such that $C_1<\sigma$ and

$$C_6\, C_3\, C_1\left(1+C_1\right) \le \frac{\varepsilon}{2}$$

From the bound on $k\eta(V_N)$ (6.5), it is then sufficient to show that

$$C_7\, C_{\rm sol}(k)\left[\left(\frac{h}{h+\sigma}\right)^p\left(1+\frac{hk}{h+\sigma}\right) + k\left(\frac{kh}{\sigma p}\right)^p\left(\frac{1}{p}+\frac{kh}{\sigma p}\right)\right] \tag{6.9}$$

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

$$\theta_1 := \frac{h}{h+\sigma} \quad \text{and} \quad \theta_2 := \frac{C_1}{\sigma},$$

so that (6.9) is bounded by

$$C_7\, C_{\mathrm{sol}}(k)\left[(\theta_1)^p\left(1+\frac{C_1 p}{\sigma}\right)+k(\theta_2)^p\left(\frac{1}{p}+\frac{C_1}{\sigma}\right)\right];$$

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

Proof of Theorem 3.4. This follows by combining Lemma 6.4 and Corollary 6.6. ■

Acknowledgements

The authors thank Martin Averseng (ETH Zürich) and an anonymous referee for highlighting simplifications of the arguments in a earlier version of the paper. We also thank Théophile Chaumont-Frelet (INRIA, Nice) for useful discussions about the results of [35], [36]. DL and EAS acknowledge support from EPSRC grant EP/1025995/1. JW was partly supported by Simons Foundation grant 631302.

References

[1] A.K. Aziz, R.B. Kellogg, A.B. Stephens, A two point boundary value problem with a rapidly oscillating solution, Numer. Math. 53 (1988) 107–121.
[2] I.M. Babuška, S.A. Sauter, Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers?, SIAM Rev. (2000) 451–484.
[3] H. Barucq, T. Chaumont-Frelet, C. Gout, Stability analysis of heterogeneous Helmholtz problems and finite element solution based on propagation media approximation, Math. Comput. 86 (2017) 2129–2157.
[4] M. Bernkopf, T. Chaumont-Frelet, J.M. Melenk, Stability and convergence of Galerkin discretizations of the Helmholtz equation in piecewise smooth media, https://numericalwaves.sciencesconf.org/data/program/abstract_melenk.pdf, 2020.
[5] N. Burq, Décroissance de l'énergie locale de l'équation des ondes pour le problème extérieur et absence de resonance au voisinage du réel, Acta Math. 180 (1998) 1–29.
[6] N. Burq, Semi-classical estimates for the resolvent in nontrapping geometries, Int. Math. Res. Not. 2002 (2002) 221–241.
[7] S.N. Chandler-Wilde, P. Monk, Wave-number-explicit bounds in time-harmonic scattering, SIAM J. Math. Anal. 39 (2008) 1428–1455.
[8] T. Chaumont-Frelet, On high order methods for the heterogeneous Helmholtz equation, Comput. Math. Appl. 72 (2016) 2203–2225.
[9] T. Chaumont-Frelet, S. Nicaise, High-frequency behaviour of corner singularities in Helmholtz problems, ESAIM: Math. Model. Numer. Anal. 52 (2018) 1803–1845.
[10] T. Chaumont-Frelet, S. Nicaise, Wavenumber explicit convergence analysis for finite element discretizations of general wave propagation problem, IMA J. Numer. Anal. 40 (2020) 1503–1543.
[11] M. Costabel, M. Dauge, S. Nicaise, Corner singularities and analytic regularity for linear elliptic systems. Part I: smooth domains, https://hal.archives-ouvertes.fr/file/index/docid/453934/filename/CoDaNi_Analytic_Part_I.pdf, 2010.
[12] Y. Du, H. Wu, Preasymptotic error analysis of higher order FEM and CIP-FEM for Helmholtz equation with high wave number, SIAM J. Numer. Anal. 53 (2015) 782–804.
[13] J.J. Duistermaat, L. Hörmander, Fourier integral operators. II, Acta Math. 128 (1972) 183–269.
[14] S. Dyatlov, M. Zworski, Mathematical Theory of Scattering Resonances, Graduate Studies in Mathematics, vol. 200, American Mathematical Society, 2019.
[15] S. Esterhazy, J.M. Melenk, On stability of discretizations of the Helmholtz equation, in: I.G. Graham, T.Y. Hou, O. Lakkis, R. Scheichl (Eds.), Numerical Analysis of Multiscale Problems, Springer, 2012, pp. 285–324.
[16] J. Galkowski, E.A. Spence, J. Wunsch, Optimal constants in nontrapping resolvent estimates, Pure Appl. Anal. 2 (2020) 157–202.
[17] D. Gallistl, T. Chaumont-Frelet, S. Nicaise, J. Tomezyk, Wavenumber explicit convergence analysis for finite element discretizations of time-harmonic wave propagation problems with perfectly matched layers, hal preprint 01887267, 2018.
[18] M. Ganesh, C. Morgenstern, A coercive heterogeneous media Helmholtz model: formulation, wavenumber-explicit analysis, and preconditioned high-order FEM, Numer. Algorithms (2019) 1–47.
[19] S. Gong, I.G. Graham, E.A. Spence, Domain decomposition preconditioners for high-order discretisations of the heterogeneous Helmholtz equation, IMA J. Numer. Anal. 41 (2021) 2139–2185.
[20] I.G. Graham, O.R. Pembery, E.A. Spence, The Helmholtz equation in heterogeneous media: a priori bounds, well-posedness, and resonances, J. Differ. Equ. 266 (2019) 2869–2923.
[21] I.G. Graham, S.A. Sauter, Stability and finite element error analysis for the Helmholtz equation with variable coefficients, Math. Comput. 89 (2020) 105–138.
[22] B. Helffer, D. Robert, Calcul fonctionnel par la transformation de Mellin et opérateurs admissibles, J. Funct. Anal. 53 (1983) 246–268.
[23] B. Helffer, J. Sjöstrand, Équation de Schrödinger avec champ magnétique et équation de Harper, in: Schrödinger Operators, Sønderborg, 1988, in: Lecture Notes in Phys., vol. 345, Springer, Berlin, 1989, pp. 118–197.
[24] L. Hörmander, The Analysis of Linear Differential Operators. I, Distribution Theory and Fourier Analysis, Springer-Verlag, Berlin, 1983.
[25] G.C. Hsiao, W.L. Wendland, Boundary Integral Equations, Applied Mathematical Sciences, vol. 164, Springer, 2008.
[26] F. Ihlenburg, I. Babuška, Finite element solution of the Helmholtz equation with high wave number Part I: the h-version of the FEM, Comput. Math. Appl. 30 (1995) 9–37.
[27] D. Lafontaine, E.A. Spence, J. Wunsch, Decompositions of high-frequency Helmholtz solutions via functional calculus, and application to the finite element method, arXiv preprint, arXiv:2102.13081, 2021.
[28] D. Lafontaine, E.A. Spence, J. Wunsch, For most frequencies, strong trapping has a weak effect in frequency-domain scattering, Commun. Pure Appl. Math. 74 (2021) 2025–2063.
[29] D. Lafontaine, E.A. Spence, J. Wunsch, A sharp relative-error bound for the Helmholtz h-FEM at high frequency, Numer. Math. 150 (2022) 137–178.
[30] P.D. Lax, R.S. Phillips, Scattering Theory, revised ed., Academic Press, 1989.
[31] Y. Li, H. Wu, FEM and CIP-FEM for Helmholtz equation with high wave number and perfectly matched layer truncation, SIAM J. Numer. Anal. 57 (2019) 96–126.
[32] A. Martinez, An Introduction to Semiclassical and Microlocal Analysis, vol. 994, Springer, 2002.
[33] J.M. Melenk, On generalized finite element methods, PhD thesis, The University of Maryland, 1995.
[34] J.M. Melenk, A. Parsania, S. Sauter, General DG-methods for highly indefinite Helmholtz problems, J. Sci. Comput. 57 (2013) 536–581.
[35] J.M. Melenk, S. Sauter, Convergence analysis for finite element discretizations of the Helmholtz equation with Dirichlet-to-Neumann boundary conditions, Math. Comput. 79 (2010) 1871–1914.
[36] J.M. Melenk, S. Sauter, Wavenumber explicit convergence analysis for Galerkin discretizations of the Helmholtz equation, SIAM J. Numer. Anal. 49 (2011) 1210–1243.
[37] J.C. Nédélec, Acoustic and Electromagnetic Equations: Integral Representations for Harmonic Problems, Springer Verlag, 2001.
[38] O.R. Pembery, The Helmholtz Equation in Heterogeneous and Random Media: Analysis and Numerics, PhD thesis, University of Bath, 2020.
[39] J.V. Ralston, Trapped rays in spherically symmetric media and poles of the scattering matrix, Commun. Pure Appl. Math. 24 (1971) 571–582.
[40] D. Robert, Autour de l'approximation semi-classique, Progress in Mathematics, vol. 68, Birkhäuser Boston, Inc., Boston, MA, 1987.
[41] J. Saranen, G. Vainikko, Periodic Integral and Pseudodifferential Equations with Numerical Approximation, Springer, 2002.
[42] S.A. Sauter, A refined finite element convergence theory for highly indefinite Helmholtz problems, Computing 78 (2006) 101–115.
[43] A.H. Schatz, An observation concerning Ritz-Galerkin methods with indefinite bilinear forms, Math. Comput. 28 (1974) 959–962.
[44] J. Sjöstrand, A trace formula and review of some estimates for resonances, in: Microlocal Analysis and Spectral Theory, Lucca, 1996, in: NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci., vol. 490, Kluwer Acad. Publ., Dordrecht, 1997, pp. 377–437.
[45] J. Sjöstrand, M. Zworski, Complex scaling and the distribution of scattering poles, J. Am. Math. Soc. 4 (1991) 729–769.
[46] E.A. Spence, Overview of variational formulations for linear elliptic PDEs, in: A.S. Fokas, B. Pelloni (Eds.), Unified Transform Method for Boundary Value Problems: Applications and Advances, SIAM, 2015, pp. 93–159.
[47] M.E. Taylor, Partial Differential Equations II, Qualitative Studies of Linear Equations, Applied Mathematical Sciences, vol. 116, Springer-Verlag, New York, 1996.
[48] 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, Russ. Math. Surv. 30 (1975) 1–58.
[49] B.R. Vainberg, Asymptotic Methods in Equations of Mathematical Physics, Gordon & Breach Science Publishers, New York, 1989, Translated from the Russian by E. Primrose.
[50] 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 (2014) 1266–1288.
[51] L. Zhu, 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 (2013) 1828–1852.
[52] M. Zworski, Semiclassical Analysis, Graduate Studies in Mathematics, vol. 138, American Mathematical Society, Providence, RI, 2012.