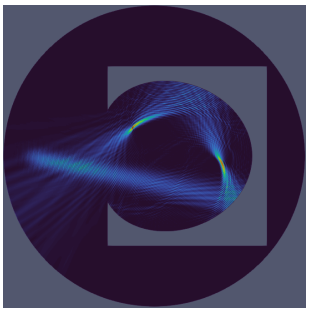
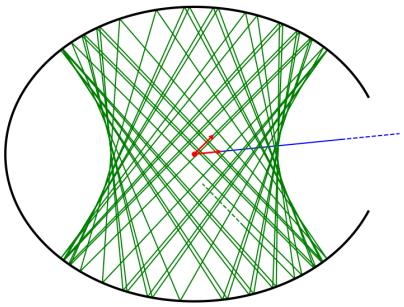
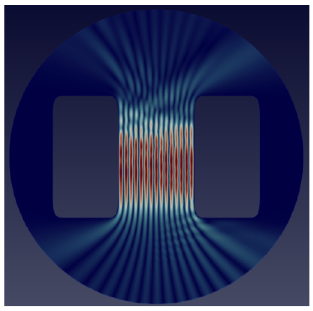
CENTRE
HENRI LEBESGUE
CENTRE DE MATHÉMATIQUES

université
angers

Semiclassical analysis of the high-frequency Helmholtz equation and its numerical approximation

Martin Averseng

Masterclass Angers 2025 – Centre Henri-Lebesgue

Contents

1 Helmholtz problems and their Galerkin approximation **5**
- 1.1 A model Helmholtz problem . . . 5
- 1.2 Abstract Helmholtz problems . . . 10
- 1.3 Galerkin approximation . . . 16
- 1.4 Why should the Galerkin error be small? . . . 17

2 The finite-element method and the pollution effect **21**
- 2.1 Finite-element spaces . . . 21
- 2.2 Elliptic regularity . . . 23
- 2.3 Frequency splitting of the resolvent . . . 26
- 2.4 Pre-asymptotic regime: the elliptic projection argument . . . 27
- 2.5 Construction of $S(k)$ . . . 30

3 Non-uniform meshes defined by billiard trajectories **36**
- 3.1 The Helmholtz resolvent and billiard trajectories . . . 36
- 3.2 Propagation of errors in the finite-element method . . . 41
- 3.3 Sketch of a localized argument . . . 42
- 3.4 Local error estimate . . . 48

4 Pseudo-locality results **51**
- 4.1 Order notation and interior cutoffs . . . 51
- 4.2 Pseudolocality of $S(k)$ and $R^\sharp(k)$ . . . 53

Introduction

The goal of these lectures is to review an are of research spanning from the 1990s to now, with the aim to present, and prove to some extent, the results obtained with J. Galkowski and E. A. Spence in [4]. The presentation heavily draws from the recent review paper [23].[1]

The broad context of this research is that of *high-frequency* wave propagation, described through the Helmholtz equation, which reads

$$-\mathrm{div}(A(x)\nabla u(x)) - k^2 n(x) u(x) = f(x)\,, \quad u, n, f : U \to \mathbb{C}, \quad A : U \to \mathbb{C}^{d\times d}\,, \quad k \in \mathbb{R}_+$$

posed in some possibly unbounded open set $U \subset \mathbb{R}^d$, subject to boundary conditions on ∂U with prescribed behaviour at infinity. Except in very special cases, solutions to this PDE cannot be written down explicitly. However, arbitrarily accurate approximations can be computed by standard computer procedures, given sufficient time and memory. Our main concern here is to establish mathematical results concerning the accuracy of these approximations, as a function of the computational work, in the regime where the *wavenumber* k becomes large – i.e., for high-frequency waves. It is well-known that this is a difficult problem, in the sense that high-frequency problems require very high computing times for a given accuracy. This course is centered around one manifestation of this difficulty, the so-called *pollution effect* (see Chapter 2), and how one can try to mitigate it.

In many numerical methods (in particular in the *finite element method*, which will be the one under study here), the main difficulty is that we need an approximation of u but we only know the result of applying the operator $P(k) = -\mathrm{div}(A(x)\nabla) - k^2 n(x)$ to u. What we *can* do, is find a function $\hat{u}$ which almost solves the same equation as u, i.e., for which $P(k)\hat{u} \approx P(k)u$. This means that the *residual error* $P(k)u - P(k)\hat{u}$ should be small, but at this point, there is no guarantee that the *approximation error* $u - \hat{u}$ (the error that we actually care about) is also small.

To answer this question, it can be seen at this point that the *resolvent* $P(k)^{-1}$ will play a fundamental role in this theory. Indeed, one has

$$u - \hat{u} = P(k)^{-1}(P(k)\hat{u} - P(k)u)$$

so the approximation error will be the residual error, "amplified" by the resolvent $P(k)^{-1}$. Observe that $P(k)^{-1}$ has nothing to do with the numerical approximation method in the first place: it is a genuine, continuous mathematical object that can be studied in its own right. As it turns out, its description for $k \to \infty$ falls under the scope of *semiclassical analysis*. Originally, semiclassical analysis was developed in the context of the theory of quantum dynamics. One of its achievements

[1] Though I have tried my best making these notes as correct and clear as possible, I am certain that they contain many mistakes – sorry about this! I will be grateful if you let me know of any of them; do not hesitate to send me an e-mail at martin.averseng@univ-angers.fr

is the mathematical justification of the so-called "correspondence principle", according to which the predictions of quantum dynamics should agree with those of classical mechanics in the limit of very large systems (hence the name "semiclassical"). Here we will be using another, perhaps more familiar instance of such a correspondence: the fact that high-frequency waves are described by the laws of geometric optics at high-frequency. One of the central aspect of these lectures will be to demonstrate how one can take these informations into account for the analysis of the approximation error $u - \hat{u}$.

Chapter 1

Helmholtz problems and their Galerkin approximation

1.1 A model Helmholtz problem

Throughout the lectures, we will work with an "abstract" formulation of the Helmholtz problem, focusing more on its mathematical structure than on its underlying physical meaning. In this paragraph, we show where such an abstract formulation comes from by considering a typical concrete example from acoustics. Let us point out that many other physical phenomena (for instance quantum dynamics, elastic or electromagnetic waves) lead to a similar mathematical model.

We consider an acoustic wave propagating in an unbounded medium $\Omega_+ := \mathbb{R}^d \setminus \overline{\Omega}_-$ (with $d = 2$ or 3), the open complement of an impenetrable obstacle $\Omega_- \subset \mathbb{R}^d$ (open and bounded). The material properties of the medium are described through smooth bounded functions $\rho : \Omega_+ \to \mathbb{R}$ and $c : \Omega_+ \to \mathbb{R}$ such that ρ and c are constant outside a compact set. Physically, $c(x)$ represents the speed of sound in the medium and $\rho(x)$ describes its density. Assuming that the medium is at rest for all times $t < 0$, and then subjected to some smooth, compactly supported in space, time-harmonic excitation $f(x,t) = f(x)e^{-i\omega t}$ (e.g., by a sinusoidal movement of the membrane of a loudspeaker), the pressure departure from its value at rest, $\delta p(x,t) = p(x,t) - p_0$, settles as $t \to \infty$ to a time-harmonic dependence $\delta p(x,t) = u(x)e^{-i\omega t}$ where the complex *amplitude* $u : \Omega_+ \to \mathbb{C}$ obeys the Helmholtz equation

$$-\operatorname{div}(\rho(x)\nabla u) - k^2 n(x) u = f \quad \text{on } \Omega_+$$

Here, $k = \omega/c_0$ is called the **wavenumber** and $n(x) = c_0^2/c(x)^2$ is (the square of) the refraction index. In many applications, f is a plane wave $\chi(x)e^{i\boldsymbol{k}\cdot x}$ with $|\boldsymbol{k}| = k$ and χ a cutoff function, as depicted in Figure 1.1.

Depending on the obstacle material, some boundary condition is satisfied by u on $\partial\Omega_-$. Here we shall assume that δp, and thus also its amplitude u, vanishes on $\partial\Omega_-$ (so-called "sound-soft" condition). It can be shown that in addition, the property that $\delta p(x,t) = 0$ for $t < 0$, translates into an important asymptotic behaviour of $u(x)$ as $r := \|x\| \to \infty$, the so-called *Sommerfeld radiation*

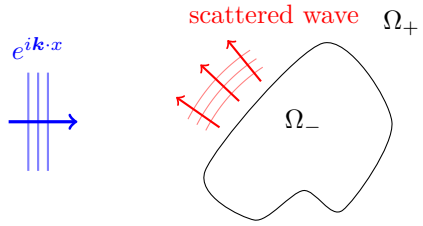


Figure 1.1: Scattering of a plane-wave by an obstacle Ω_-

condition,[1] which reads

$$\frac{\partial u}{\partial r} - iku = o(r^{-(d-1)/2}). \tag{1.1}$$

This condition "selects" solutions of the Helmholtz equation that behave like $O(e^{ikr}/r)$ for large r, and "filters out" those that behave like $O(e^{-ikr}/r)$ (see Exercise 1.1). If we return to the time-dependent problem, this means that we will only retain spherical waves of the form $e^{i(kr-\omega t)}/r$, which propagate towards infinity, and not the "unphysical" waves $e^{i(kr+\omega t)}/r$ which propagate from infinity to the origin. The fact that the time-dependent problem and the time-harmonic problem are related through the Sommerfeld condition is known as the *limiting amplitude principle* (see, e.g., [35, 20])[2]. As a result, the boundary value problem

$$\left\{\begin{array}{rcll} -\operatorname{div}(\rho(x)\nabla u) - k^2 n(x) u &=& f & \text{in } \Omega_+, \\ u &=& 0 & \text{on } \partial\Omega_-, \\ \partial_r u - iku &=& o(r^{-(d-1)/2}) & \text{as } r \to \infty. \end{array}\right. \tag{1.2}$$

admits a unique solution $u \in C^2(\Omega_+) \cap C(\overline{\Omega_+})$.[3] Figure 1.2 below illustrates this by showing a function f and (a numerical approximation of) the solution u to (1.2) truncated to a bounded domain.

With numerical approximation in mind, we would like to reduce the computation to a bounded domain. This can be achieved by replacing the above problem by the *truncated Helmholtz problem*

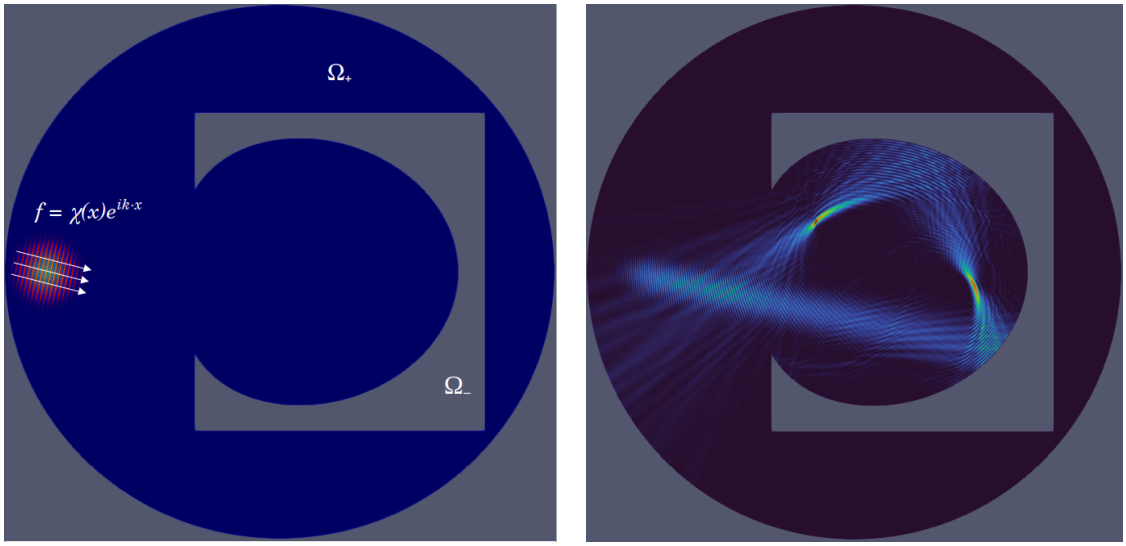
$$\left\{\begin{array}{rcll} -\operatorname{div}(\widetilde{\rho}(x)\nabla \widetilde{u}) - k^2 \widetilde{n}(x) \widetilde{u} &=& f & \text{in } \Omega_+ \cap B_R, \\ \widetilde{u} &=& 0 & \text{on } \partial\Omega_-, \\ \widetilde{u} &=& 0 & \text{on } \partial B_R. \end{array}\right. \tag{1.3}$$

where B_R is a sufficiently large ball and where $\widetilde{\rho} : \Omega_+ \cap B_R \to \mathbb{C}^{d\times d}$ and $\widetilde{n} : \Omega_+ \cap B_R \to \mathbb{C}$ are well-chosen functions. Roughly speaking, $\widetilde{\rho}$ (resp. $\widetilde{n}$) take the same values as ρI_d (resp. n) in some ball B_r with $r < R$, and then are chosen to mimic the properties of a fictitious absorbing material

[1] derived by Arnold Sommerfeld in 1912 [46].

[2] The limit amplitude principle is a cousin of the *limit absorption* principle, which states that the unique solution of (1.2) is the limit as $\varepsilon \to 0^+$ of the unique L^2 solution of the same problem but where k^2 is replaced by $k^2 + i\varepsilon$. This corresponds to adding a small damping in the corresponding time-dependent problem.

[3] uniqueness outside a large ball follows from the Rellich's lemma [39], [13, Lemma 3.11], and on Ω_+ by a unique continuation principle. Existence then follows by a Fredholm argument. References for these proofs are surveyed in [23, Section 4.2].



$\mathrm{Re}(f)$ $\qquad$ $|u|^2$

Figure 1.2: An example of Helmholtz problem in $\mathbb{R}^2$, with a right-hand side $f(x) = \chi(x)e^{i\boldsymbol{k}\cdot x}$, with $\chi(x) = e^{-50|x-x_0|^2}$, $x_0 = (-0.85, 0)$ and $|\boldsymbol{k}| = k = 400$. The visualization is restricted to a disk of radius 1.
Left panel: plot of the real part of f. The arrows indicate the direction of the vector $\boldsymbol{k}$.
Right panel: plot of the magnitude $|u|^2$ of (a numerical approximation of) the unique solution u to (1.2), with $\rho = n = 1$. Brighter areas indicate regions where $|u|^2$ is larger.

in the layer $r \le |x| \le R$. The absorption of this layer is designed to prevent unphysical reflections back towards the obstacle.[4] This truncation procedure is known as *perfectly matched layer*[5]; it can be shown that the solution $\widetilde{u}$ coincides with u on B_ρ up to an error of size $O(e^{-\mu k})$, for some $\mu > 0$ [22]. Roughly speaking, the PML plays the role of an "approximate Sommerfeld condition".

A more pleasant mathematical formulation of (1.3) is obtained by multiplying the first equation by a "test function" $v \in C_c^\infty(\Omega_+ \cap B_R)$, using Green's theorem, noticing that boundary terms vanish due to the boundary conditions. This leads to

$$\int_\Omega \left(k^{-2}\widetilde{\rho}(x)\nabla\widetilde{u}(x)\cdot\overline{\nabla v(x)} - \widetilde{n}(x)\widetilde{u}(x)\overline{v(x)} \right) dx = \int_\Omega k^{-2} f(x)\overline{v(x)}\, dx, \qquad \text{for all } v \in C_c^\infty(\Omega), \tag{1.4}$$

where $\Omega := B_R \cap \Omega_+$ is the *computational domain*. The problem (1.4) is called a *variational formulation* of (1.3). Its mathematical structure is more easily seen by writing

$$a_k(u,v) := \int_\Omega \left(k^{-2}\widetilde{\rho}(x)\nabla u(x)\cdot\overline{\nabla v(x)} - \tilde{n}(x)u(x)\overline{v(x)} \right) dx\,, \quad L(v) := \int_\Omega k^{-2} f(x)\overline{v(x)}\, dx.$$

We are then seeking a function $\tilde{u}$ such that the antilinear forms $a_k(\tilde{u}, \cdot)$ and $L(\cdot)$ agree on $C_c^\infty(\Omega)$. We

[4] Such a principle is in fact used in real-life acoustic experiments. These are often conducted in so-called "anechoic chambers": rooms surrounded by a foam coating playing the role of the absorbing layer.
[5] It was derived by Jean-Pierre Bérenger in [10] (1994).

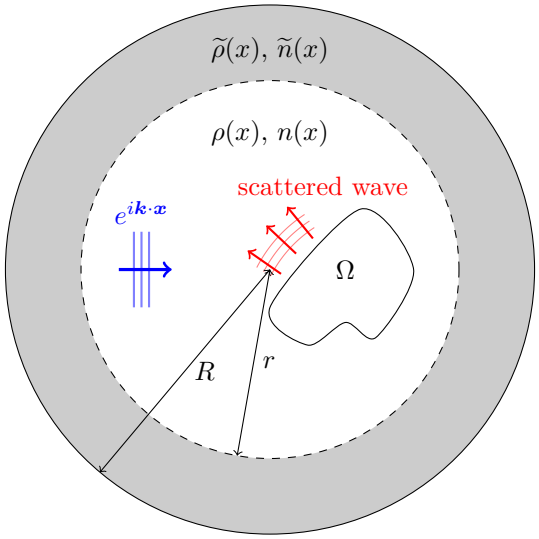


Figure 1.3: The truncated Helmholtz problem with a fictitious absorbing layer (gray shaded region). The larger ball is the *computational domain*. The solution of the truncated problem is a good approximation of the solution of the unbounded problem in B_ρ (white region).

can almost "forget" the specific definition of a_k and L, what matters is that they are a sesquilinear form[6] and a linear form, respectively.

The reason why it is advantageous to reformulate (1.3) in this way is that a systematic existence and uniqueness result is available for variational problems (the Lax-Milgram theorem, see below), provided that we set everything in a suitable **Hilbert space**. This will come at a cost though: we will have to accept that the unique solution provided by this theory may not be a genuine twice-differentiable function (if we wanted to remain in $C^2(\overline{\Omega})$, we could not apply the powerful theory because $C^2(\overline{\Omega})$ is not a Hilbert space), and thus, perhaps, have little to do with the original problem (1.3).

For this reason, it is natural to try to introduce the "smallest" possible Hilbert space. Namely, we consider the adherence of $C^1_c(\Omega)$ (C^1 functions with compact support) in $L^2(\Omega)$ with respect to the "energy norm"

$$\|u\|^2_{H^1} := \int_\Omega |u(x)|^2 + |\nabla u(x)|^2 \, dx. \tag{1.5}$$

This space is known as a *Sobolev space*[7] and is usually denoted by $H^1_0(\Omega)$.[8] It is a (strict) subspace

[6] i.e., linear in the left argument, anti-linear in the right argument.

[7] From Sergĭ L'vovich Sobolev, who studied a family of such spaces systematically between the years 1930-1950, see e.g. [45] (the English translation of his famous 1950 monograph). A standard modern reference is [1].

[8] The subscript 0 comes from the fact that elements of $H^1_0(\Omega)$ satisfy the Dirichlet boundary condition on $\partial\Omega$. More precisely, they satisfy

$$\int_\Omega \nabla u \cdot \mathbf{v} = -\int_\Omega u \operatorname{div} \mathbf{v} \, dx,$$

for all $\mathbf{v} \in C^\infty(\overline{\Omega})^3$, i.e., the expected boundary term vanishes, see Exercise 1.2. Note that one cannot simply say that elements of $H^1_0(\Omega)$ "vanish on $\partial\Omega$", since this set is of Lebesgue measure 0, so the restriction to $\partial\Omega$ of a function

of $L^2(\Omega)$, and, indeed, a Hilbert space for the same norm (see Exercise 1.2). The maps a_k and L admit unique continuous extensions to $H_0^1(\Omega)$ (see Exercise 1.2 below) and the Lax-Milgram theorem gives the existence of a unique function $u_{\mathrm{w}} \in H_0^1(\Omega)$ such that

$$a_k(u_{\mathrm{w}}, v) = L(v)\,, \quad \forall v \in H_0^1(\Omega).$$

Such a solution is known as a *weak solution* of (1.3). Fortunately, if the boundary of Ω_- is a smooth $(d-1)$-submanifold of $\mathbb{R}^d$, it can be shown that weak solutions are in fact (or more exactly, admit a representative that is) infinitely differentiable in Ω, continuous on $\overline{\Omega}$ and which vanishes on $\partial\Omega$; this result is known as *elliptic regularity*[9]. Weak solutions then solve the problem (1.3) ("in the strong sense"); see Exercise 1.3.

Exercise 1.1. (The Sommerfeld condition).
Let Ω_- be the unit ball $B(0,1)$ in $\mathbb{R}^3$, and let $u_\pm : \Omega_+ \to \mathbb{C}$ be given by

$$u_\pm(x) := \frac{e^{\pm ik\|x\|}}{\|x\|} \quad \text{for all } x \in \Omega_+.$$

Show that

$$-\Delta u_\pm - k^2 u_\pm = 0 \quad \text{on } \Omega_+.$$

Let $\varphi \in C_c^\infty(\mathbb{R}^3)$ be such that $\varphi \equiv 1$ on $B(0,2)$ and let $f = (-\Delta - k^2)\varphi$. Deduce that for any $a, b \in \mathbb{C}$ such that $e^{ik}a + e^{-ik}b = -1$, the function

$$u_{a,b} := \varphi + a u_+ + b u_-$$

satisfies the boundary value problem

$$\begin{cases} (-\Delta - k^2)u_{a,b} &= f & \text{in } \Omega_+, \\ u &= 0 & \text{on } \partial\Omega_-. \end{cases}$$

Check that $u_{a,b}$ satisfies the Sommerfeld condition if and only if $b = 0$.

Exercise 1.2. (The space $H_0^1(\Omega)$).
Let $\Omega \subset \mathbb{R}^d$ be a non-empty open set.

1. *Show that $H_0^1(\Omega)$ is a dense subspace of $L^2(\Omega)$ (i.e., given $u \in L^2(\Omega)$ and $\varepsilon > 0$, there exists $u_\varepsilon \in H_0^1(\Omega)$ such that $\|u - u_\varepsilon\|_{L^2(\Omega)} \leq \varepsilon$).*
2. *Show that $u \mapsto \|u\|_{H^1}$ defined on $C_c^1(\Omega)$ admits a unique continuous extension to $H_0^1(\Omega)$, and that this is extension is again a norm on $H_0^1(\Omega)$. Check that $H_0^1(\Omega)$ is a Hilbert space for the norm thus defined.*
3. *Show that the gradient operator $\nabla : C_c^1(\Omega) \to (L^2(\Omega))^d$ admits a unique linear continuous extension (denoted by the same symbol) $\nabla : H_0^1(\Omega) \to L^2(\Omega)$. For $u \in H_0^1(\Omega)$, the function ∇u is called the* weak gradient *of u.*

in $L^2(\Omega)$ is not well-defined.

[9] Elliptic regularity is an important field of PDE theory; in its general form, it is related to the 19th of Hilbert's 23 problems formulated 1900, which was solved in landmark papers by Ennio De Giorgi [17] in 1956 and John Nash [36] in 1958.

4. *Deduce that a_k (resp. L) admits a unique sesquilinear continuous extension (resp. linear continuous) to $H^1_0(\Omega)\times H^1_0(\Omega)$ (resp. $H^1_0(\Omega)$).*

5. *Let $u\in H^1_0(\Omega)$. Prove that for any $\boldsymbol{\varphi}\in C^\infty(\overline{\Omega})^d$,*

$$\int_\Omega \nabla u\cdot\boldsymbol{\varphi}\,dx = -\int_\Omega u\,\mathrm{div}\,\boldsymbol{\varphi}\,dx.$$

Deduce that for any $u\in C^1(\overline{\Omega})$,

$$u\in H^1_0(\Omega) \implies u|_{\partial\Omega}=0.$$

6. *Assuming that $d\geq 3$, find an element of $H^1_0(\Omega)$ which does not admit a continuous representative.*

Exercise 1.3. (Weak solutions are strong solutions).
Let $u_{\mathrm{w}}\in H^1_0(\Omega)$ satisfy

$$a_k(u_{\mathrm{w}},v)=L(v)\quad\text{for all } v\in H^1_0(\Omega).$$

Suppose that u_{w} has a smooth representative u_{s} ("s" for "strong"). Then show that u_{s} is a solution of (1.3).

Exercise 1.4. (Properties of a_k).
We admit that the functions $\widetilde{n}:\Omega\to\mathbb{C}$ and $\tilde{\rho}:\Omega\to\mathbb{C}^{d\times d}$ obtained by the PML truncation are smooth and that $\widetilde{\rho}$ satisfies the following property: there exists $c>0$ such that the inequality

$$\mathrm{Re}\big(\widetilde{\rho}(x)\xi\cdot\xi\big)\geq c\|\xi\|^2$$

holds for all $\xi\in\mathbb{C}^d$ and $x\in\Omega$. Show that for all $k_0>0$, there exist constants $C_0,c_1,C_2>0$ such that the estimates

(i) $|a_k(u,v)|\leq C_0\|u\|_{H^1_k}\|v\|_{H^1_k}$

(ii) $\mathrm{Re}(a_k(u,u))\geq c_1\|u\|^2_{H^1_k}-C_2\|u\|^2_{L^2}$.

hold for all $k\geq k_0$ and $u,v\in H^1_0(\Omega)$. Here, $\|\cdot\|_{H^1_k}$ is the semiclassical Sobolev norm, *defined by*

$$\|u\|^2_{H^1_k}:=\|u\|^2_{L^2(\Omega)}+k^{-2}\|\nabla u\|^2_{L^2(\Omega)}\,, \tag{1.6}$$

where, for $u\in H^1_0(\Omega)$, ∇u is the weak gradient of u defined in Exercise 1.2.

Remark 1.1 (k-dependent norm). The rescaling by k^{-2} on the gradient term in (1.6) is the "natural" scaling allowing to formulate conveniently the results that will follow. The basic reason behind this is that solutions of the Helmholtz equations "oscillate at frequency k", and with this scaling, the contributions from $\|u\|_{L^2}$ and $\|\nabla u\|_{L^2}$ are balanced for such functions (indeed, an exact balance is achieved by plane waves $e^{i\boldsymbol{k}\cdot x}$ with $|\boldsymbol{k}|=1$).

1.2 Abstract Helmholtz problems

Let us recall two fundamental results. The first one is the celebrated theorem of Lax and Milgram from their 1954 paper [32].

Theorem 1.1 (The Lax-Milgram theorem)

Let V be a Hilbert space and $B : V \times V \to \mathbb{C}$ be a *bounded sesquilinear form*, i.e. $u \mapsto B(u, v)$ is linear, $v \mapsto B(u, v)$ is anti-linear, and

$$|B(u, v)| \leq C\|u\|_V \|v\|_V \quad \forall u, v \in V.$$

Moreover, suppose that B is *coercive*, in the sense that there exists $\alpha > 0$ such that

$$\mathrm{Re}(B(u, u)) \geq \alpha \|u\|_V^2 \quad \forall u \in V.$$

Then, for any continuous anti-linear form $f : V \to \mathbb{C}$, there exist a unique $u \in V$ such that

$$B(u, v) = f(v) \quad \forall v \in V. \tag{1.7}$$

Exercise 1.5. (Proof of the Lax-Milgram theorem).

(i) Show that if a solution $u \in V$ of the variational problem (1.7), then it is unique.

(ii) Show that there exists $F \in V$ and an injective, bounded linear operator $A : V \to V$ satisfying

$$(F, v)_V = f(v), \quad (Au, v)_V = B(u, v) \quad \forall u, v \in V,$$

where $(\cdot, \cdot)_V$ denotes the inner product on V.

(iii) Show that $\mathrm{Ran}(A)$ is closed (where $\mathrm{Ran}(A)$ denotes the range of A). (Hint: use the the coercivity assumption and the fact that V is complete)

(iv) Show that $(\mathrm{Ran}(A))^\perp = \{0\}$.

(v) Conclude.

Next, we state a result concerning Fredholm operators of index 0.[10] Recall that given two Hilbert spaces H_1 and H_2, a bounded linear operator $K : H_1 \to H_2$ is *compact* if, given any bounded sequence $(u_n)_{n\in\mathbb{N}}$ of elements of H_1, there exists a subsequence of $(Ku_n)_{n\in\mathbb{N}}$ which converges in H_2.[11]

Theorem 1.2 (Fredholm operators of index 0)

Let H_1, H_2 be two Hilbert spaces. Let $A_0 : H_1 \to H_2$ be a bounded linear operator and let $K : H_1 \to H_2$ be a compact operator. Suppose that

(i) A_0 is an isomorphism

(ii) $A_0 + K$ is injective

[10] The celebrated book [30] by Tosio Kato contains a general treatment of such operators in Chapter IV, §5. A more elementary presentation can be found in [44, Chapter 5].

[11] The definition of compact operator, and the result of Theorem 1.2 are usually stated in the more general setting of "topological vector spaces" instead of Hilbert spaces, but we do need this here.

Then $A_0 + K$ is an isomorphism.

Idea/Reference for the proof. Since A_0 is an isomorphism, we have $\mathrm{ind}(A_0) = 0$, where, for a linear operator $L : E \to F$,

$$\mathrm{ind}(L) := \dim(\mathrm{Ker}(L)) - \dim(F/\mathrm{Ran}(L))$$

provided both dimensions are finite (i.e., if L is a Fredholm operator). A key property is that the index is invariant by compact perturbation [44, Theorem 5.10]: here, this tells us that $\mathrm{ind}(A_0+K) = 0$. Since $A_0 + K$ is injective, it follows that $\dim(F/\mathrm{Ran}(A_0 + K)) = 0$, that is, $A_0 + K$ is also surjective. Thus, $A_0 + K$ is a bijective bounded linear operator, and the conclusion follows from the bounded inverse theorem (see, e.g., [41, Corollary 2.12]). □

We now follow [24] and introduce an abstract setting which contains the Helmholtz problem of §1.1 as a particular case, as well as other variants of it (e.g., other boundary conditions/truncation methods, etc.). Let $\Omega \subset \mathbb{R}^d$ be an open set and let $\mathcal{H} := L^2(\Omega)$, with its usual norm and inner product denoted by $\|\cdot\|_{\mathcal{H}}$ and $\langle\cdot,\cdot\rangle_{\mathcal{H}}$, respectively.

Definition 1.1 (The spaces $\mathcal{H}^n_k$)

For every $k > 0$, let $(\mathcal{H}^n_k)_{n\in\mathbb{N}}$ be a scale of Hilbert spaces. For each $n \in \mathbb{N}$, denote by $\|\cdot\|_{\mathcal{H}^n_k}$ the norm on $\mathcal{H}^n_k$. We assume that

$$\mathcal{H}^0_k = \mathcal{H} = L^2(\Omega)\,, \qquad \forall k > 0$$

with equal norms, and for all $k, k' > 0$, $\mathcal{H}^n_k = \mathcal{H}^n_{k'} =: \mathcal{H}^n$ (as vector spaces), although the norms may differ for $k \neq k'$. Moreover, for all $n \in \mathbb{N}$, we assume that there is a continuous and dense inclusion $\mathcal{H}^{n+1}_k \subset \mathcal{H}^n_k$ with $\|u\|_{\mathcal{H}^n_k} \leq \|u\|_{\mathcal{H}^{n+1}_k}$.

One can think of the superscript n as indicating the "level of regularity", or "number of derivatives" that elements of the space $\mathcal{H}^n_k$ admit, and of the subscript k as the fact that, in the norm $\|\cdot\|_{\mathcal{H}^n_k}$, the derivatives of order j are scaled by a factor k^{-j}.

We identify $\mathcal{H}$ with its anti-dual in the canonical way. In fact, in what follows, every element $u \in \mathcal{H}$ will also be regarded as an element of $(\mathcal{H}^n_k)^*$ for any $n \in \mathbb{N}$, with its action on $\mathcal{H}^n_k$ given by

$$\langle u, v\rangle := \langle u, v\rangle_{\mathcal{H}} \qquad \forall v \in \mathcal{H}^n_k.$$

This identification is legitimate thanks to the injectivity in Exercise 1.6 below (which comes from the density of the embeddings $\mathcal{H}^n_k \subset \mathcal{H}$). Under these identification, one has the embedding $\mathcal{H} \subset (\mathcal{H}^n_k)^*$ for all $n \in \mathbb{N}$ with

$$\|u\|_{(\mathcal{H}^n_k)^*} \leq \|u\|_{\mathcal{H}} \qquad \forall n \in \mathbb{N}.$$

We will also denote $(\mathcal{H}^n_k)^*$ by $\mathcal{H}^{-n}_k$ and think of it as a space where we "miss" n derivatives to be in L^2. Thus, $(\mathcal{H}^n_k)_{n\in\mathbb{Z}}$ is a Hilbert scale suited to measure the "regularity", or "number of derivatives", of a function.

$$\ldots \subset \mathcal{H}^{-2}_k \subset \mathcal{H}^{-1}_k \subset \mathcal{H} \subset \mathcal{H}^1_k \subset \mathcal{H}^2_k \subset \ldots$$

$$\text{less regular} \quad \longleftrightarrow \quad \text{more regular}$$

Furthermore, we will use the notation

$$\langle L, u\rangle := L(u) \quad \text{and} \quad \langle u, L\rangle := \overline{\langle L, u\rangle} \qquad \forall (L,u) \in (\mathcal{H}_k^n)^* \times \mathcal{H}_k^n$$

for any $n \in \mathbb{N}$.

Exercise 1.6. (Identification map).
Show that the map $I : \mathcal{H} \to (\mathcal{H}_k^n)^*$ defined by

$$\langle Iu, v\rangle := \langle u, v\rangle_{\mathcal{H}} \qquad \forall (u,v) \in \mathcal{H} \times \mathcal{H}_k^n,$$

is injective and satisfies $\|I\| \leq 1$.

For every $k > 0$, we consider a sesquilinear form

$$a_k : \mathcal{H}_k^1 \times \mathcal{H}_k^1 \to \mathbb{C}$$

and introduce the following assumptions.

Assumption 1.3 (k-uniform Continuity)

The sesquilinear forms a_k are *k-uniformly bounded*, that is, for all $k_0 > 0$, there exists $C_0(k_0) > 0$ such that, for all $k \geq k_0$,

$$\sup_{u,v \in \mathcal{H}_k^1 \setminus \{0\}} \frac{|a_k(u,v)|}{\|u\|_{\mathcal{H}_k^1} \|v\|_{\mathcal{H}_k^1}} \leq C_0(k_0). \tag{1.8}$$

We denote by $\|a_k\|$, the *norm of* a_k, the quantity that appears in the left-hand side.

Assumption 1.4 (Gårding inequality)

For all $k_0 > 0$, there exist $c_{\mathrm{Ga}}(k_0), C_{\mathrm{Ga}}(k_0) > 0$ such that the *Gårding inequality*

$$\mathrm{Re}(a_k(u,u)) \geq C_{\mathrm{Ga}}(k_0)\|u\|^2_{\mathcal{H}_k^1} - c_{\mathrm{Ga}}(k_0)\|u\|^2_{\mathcal{H}}$$

holds for all $k \geq k_0$ and all $u \in \mathcal{H}_k^1$.

Assumption 1.5 (Compact injection)

The embedding $\mathcal{H}_k^1 \subset \mathcal{H}$ is *compact*, that is, every bounded sequence in $\mathcal{H}_k^1$ admits a subsequence that converges in $\mathcal{H}$.

Assumption 1.6 (Injectivity)

For all $k > 0$, if $u \in \mathcal{H}_k^1$ satisfies $a_k(u,v) = 0$ for all $v \in \mathcal{H}_k^1$, then $u = 0$.

Assumption 1.4 owes its name to Lars Gårding, see [26, Theorem 2.1]. In what follows, we sometimes omit the dependence in k_0 of C_0, c_{Ga} and C_{Ga} from the notation.

Exercise 1.7. (Compactness of the identification map).
Show that under Assumption 1.5, the identification map $I : \mathcal{H} \to (\mathcal{H}_k^n)^*$ from Exercise 1.6 is compact when $n \geq 1$.

Remark 1.2 (Validity of Assumptions (1.3)-(1.6) in practice). In the setting of §1.1, we would choose $\mathcal{H}_k^1$ to be the space $H_0^1(\Omega)$ endowed with the norm $\|\cdot\|_{H_k^1}$

$$u \mapsto \sqrt{\|u\|^2_{L^2(\Omega)} + k^{-2}\|\nabla u\|^2_{L^2(\Omega)}}$$

as in Exercise 1.4. Then,

1. Assumptions 1.3 and 1.4 are satisfied as shown in Exercise 1.4.
2. Assumption 1.5, i.e., the compactness of the embedding $H_0^1(\Omega) \subset L^2(\Omega)$, holds by the Rellich theorem ([38] in German, see also [1, Theorem 6.3]).
3. Assumption 1.6 is shown via a unique continuation principle, as discussed in [23, Section 4.2].

Remark 1.3 (Concrete version of the scale $(\mathcal{H}_k^n)_{n\in\mathbb{N}}$). In the setting of §1.1, the $\mathcal{H}_k^n$ norm would be defined by

$$\|u\|^2_{\mathcal{H}_k^n} := \sum_{|\alpha|\leq n} k^{-|\alpha|}\|\partial^\alpha u\|^2_{L^2(\Omega)},$$

and the space $\mathcal{H}_k^n$ as the adherence of $C^\infty(\overline{\Omega})$ in $L^2(\Omega)$ for this norm, intersected with $\mathcal{H}_k^1$.

Given a Hilbert space H, we denote by H^* the *anti-dual* of H, i.e., the vector space of continuous *anti-linear* forms $T : H \to \mathbb{C}$, equipped with the norm

$$\|T\|_{H^*} := \sup_{u\in H\setminus\{0\}} \frac{|T(u)|}{\|u\|_H}$$

Definition 1.2 (Helmholtz operator $P(k)$)

For all $k > 0$, the *Helmholtz operator* $P(k) : \mathcal{H}_k^1 \to (\mathcal{H}_k^1)^*$ is defined by

$$\langle P(k)u, v\rangle := a_k(u, v) \qquad \forall u, v \in \mathcal{H}_k^1.$$

Given $f \in (\mathcal{H}_k^1)^*$, the *Helmholtz problem* is to find $u \in \mathcal{H}_k^1$ such that

$$P(k)u = f. \tag{1.9}$$

Observe that for all $k > 0$, $\|P(k)\|_{\mathcal{H}_k^1\to(\mathcal{H}_k^*)} = \|a_k\|$ (with the latter defined in (1.8)). In particular, $P(k)$ is a (k-uniformly) bounded linear map.

Theorem 1.7 (Well-posedness of the Helmholtz problem)

Suppose that Assumptions (1.3)-(1.6) hold. Then, for all $k > 0$, the Helmholtz operator $P(k)$ is an isomorphism.

Proof. Let $k_0 > 0$ and let C_0, c_{Ga} and C_{Ga} be as in Assumptions 1.3 and 1.4. Let $m > c_{\mathrm{GA}}$ and let $a_k^+ : \mathcal{H}_k^1 \times \mathcal{H}_k^1 \to \mathbb{C}$ be the sesquilinear form defined by

$$a_k^+(u,v) := \langle (P(k) + mI)u, v\rangle = a_k(u,v) + m\langle u, v\rangle_{\mathcal{H}}\,, \qquad u, v \in \mathcal{H}_k^1$$

where I is the identification map from Exercise 1.6. By Assumption 1.3,

$$|a_k^+(u,v)| \le (C_0 + m)\|u\|_{\mathcal{H}_k^1}\|v\|_{\mathcal{H}_k^1}$$

and by Assumption 1.4,

$$\mathrm{Re}(a_k^+(u,u)) \ge C_{\mathrm{Ga}}\|u\|^2_{\mathcal{H}_k^1}; \tag{1.10}$$

that is, a_k^+ is bounded and coercive. Thus, by the Lax-Milgram theorem, the variational problem

$$\text{Find } u \in \mathcal{H}_k^1 \text{ such that for all } v \in \mathcal{H}_k^1, \quad a_k^+(u,v) = \langle f, v\rangle. \tag{1.11}$$

admits a unique solution for every $f \in (\mathcal{H}_k^1)^*$. By definition of a_k^+, this means that for any $f \in (\mathcal{H}_k^1)^*$, we can find u such that $(P(k) + mI)u = f$; hence, $P(k) + mI$ is surjective. The coercivity (1.10) also immediately implies that $P(k)+mI$ is injective. Therefore, the bounded linear map $P(k)+mI$ is bijective, hence an isomorphism by the bounded inverse theorem [41, Corollary 2.12].

Since I is compact (by Exercise 1.7) and $P(k)$ is injective by Assumption 1.6, we deduce that $P(k)$ is an isomorphism by the Fredholm theorem (Theorem 1.2). □

Definition 1.3 (Helmholtz resolvent)

For all $k > 0$, we define $R(k) : (\mathcal{H}_k^1)^* \to \mathcal{H}_k^1$ the *Helmholtz resolvent*:

$$R(k) := P(k)^{-1} : (\mathcal{H}_k^1)^* \to \mathcal{H}_k^1.$$

We will denote

$$\rho(k) := \sup_{\|f\|_{\mathcal{H}} \le 1} \|R(k)f\|_{\mathcal{H}}.$$

Let $R(k)^*$ be the adjoint of $R(k)$. Identifying $\mathcal{H}_k^1$ with its bidual (this is possible since $\mathcal{H}_k^1$ is a Hilbert space, and in particular, reflexive), $R(k)^*$ maps $(\mathcal{H}_k^1)^*$ to $\mathcal{H}_k^1$ and satisfies

$$\big\langle R(k)u, v\big\rangle = \big\langle u, R(k)^* v\big\rangle \qquad \forall u, v \in (\mathcal{H}_k^1)^*. \tag{1.12}$$

Exercise 1.8. (Lower bound on $\rho(k)$).

Let $\Omega \subset \mathbb{R}^d$ be a non-empty open set. Suppose that there is a subset $U \subset \Omega$ such that for all $k > 0$, $C_c^\infty(U) \subset \mathcal{H}_k^1$ and that for $u, v \in C_c^\infty(U)$, $a_k(\cdot,\cdot)$ is given by

$$a_k(u,v) = \int_U (k^{-2}\nabla u \cdot \nabla v - uv)\, dx\,.$$

1. *Show that for any $u \in C_c^\infty(U)$, $P(k)u = -k^{-2}\Delta u - u$.*

2. *Let*

$$R_U := \sup\{R > 0 \mid U \text{ contains a ball of radius } R\}.$$

Show that there exists $C_d > 0$ depending only on the dimension d such that for all $k > 0$, there exists $u(k) \in C_c^\infty(U)$ satisfying

$$\|P(k)u(k)\|_{L^2(\Omega)} \leq \frac{C_d}{\langle kR_U \rangle} \|u(k)\|_{L^2(\Omega)}$$

where $\langle \cdot \rangle$ is the "Japanese bracket", defined by $\langle X \rangle := (1 + \|X\|^2)^{1/2}$.

3. *Deduce that $\rho(k) \geq \frac{\langle kR_U \rangle}{C_d}$.*

Exercise 1.9. (Norm of $R(k)$ from $(\mathcal{H}_k^1)^*$ to $\mathcal{H}_k^1$).
Under assumptions (1.3)-(1.6). Show that for all $k_0 > 0$, there exists $C > 0$ such that for all $k \geq k_0$,

$$\|R(k)\|_{(\mathcal{H}_k^1)^* \to \mathcal{H}_k^1} \leq C(1 + \rho(k)).$$

Hint: use the Gårding inequality. As a first step, show that $\|R(k)\|_{\mathcal{H} \to \mathcal{H}_k^1} \leq C(1 + \rho(k))$.

Remark 1.4. Exercise 1.8 shows a first example of the deep relationship that exists between $\rho(k)$ and the geometry of the propagation domain. We will present this aspect in more details in Chapter 3.

1.3 Galerkin approximation

Let $f \in (\mathcal{H}_k^1)^*$ and let $u \in \mathcal{H}_k^1$ be the unique solution of the Helmholtz problem $P(k)u = f$; observe that by definition of $P(k)$, u can be equivalently defined as the unique solution of the variational problem

$$\text{Find } u \in \mathcal{H}_k^1 \text{ such that} \quad a_k(u, v) = \langle f, v \rangle \quad \forall v \in \mathcal{H}_k^1.$$

The basic principle of the Galerkin approximation[12] is to solve the same variational problem, but in a finite-dimensional subspace.

Definition 1.4 (Galerkin approximation)

Let $k > 0$, let $u \in \mathcal{H}_k^1$. Given a (usually, finite-dimensional) subspace $V_h \subset \mathcal{H}_k^1$, we say that u_h is a *Galerkin approximation of u in V_h* if it is a solution of the variational problem

$$\text{Find } u_h \in V_h \text{ such that} \quad a_k(u_h, v_h) = \langle f, v_h \rangle \quad \forall v_h \in V_h \tag{1.13}$$

where $f := P(k)u$.

Let us make a few comments about this definition:

[12] for Boris G. Galerkin, who proposed this method in 1915 in a paper related to the equations of elasticity [21]

1. If u_h is a Galerkin approximation of u, then the *Galerkin error* $u - u_h$ satisfies the fundamental property
$$a_k(u - u_h, v_h) = \big\langle P(k)(u - u_h), v_h \big\rangle = 0, \qquad \forall v_h \in V_h\,; \tag{1.14}$$
this is known as **Galerkin orthogonality**. In fact, (1.14) can be used as an equivalent definition of a Galerkin approximation of u.

2. The Galerkin problem (1.13) can be solved in practice as follows: introduce a basis $\{\phi_1, \dots, \phi_N\}$ of V_h, and let $A \in \mathbb{C}^{N\times N}$ and $F \in \mathbb{C}^N$ be defined by
$$A_{ij} := a_k(\phi_j, \phi_i)\,, \quad F_i := \langle f, \phi_i \rangle, \quad 1 \le i, j \le N.$$
Then, if a solution of (1.13) u_h exists, by linearity, the vector U_h of its coefficients in the basis $\{\phi_i\}_{1\le i\le N}$ satisfies the linear system of equations
$$AU_h = F.$$
Typically, the linear system coefficients A_{ij} and F_i are computed efficiently to high precision via numerical quadrature methods. If the linear system is non-singular, its unique solution $A^{-1}F$ can also be obtained by standard algorithms (in practice, this step is the main computational bottleneck when k becomes large, especially because the linear system becomes very large and sign-indefinite. We won't develop this here).

Exercise 1.10. (Uniqueness implies existence).
Show that the following assertions are equivalent:

(i) The only Galerkin approximation of $u = 0$ is $u_h = 0$.

(ii) Any $u \in \mathcal{H}^1_k$ admits a unique Galerkin approximation u_h in V_h.

Hint: use that V_h is finite-dimensional.

1.4 Why should the Galerkin error be small?

A Galerkin approximation may or may not exist, and is not necessarily unique. It is also not clear at first glance what makes it a good candidate to approximate u. A simple result in this direction is Céa's lemma [15]. Its proof shows in its simplest form one of the fundamental mechanisms allowing to exploit Galerkin orthogonality to obtain a bound on the Galerkin error.

Lemma 1.8 (Céa's lemma)

Suppose that the sesquilinear form a_k satisfies
$$\gamma(k) := \inf_{u\in\mathcal{H}^1_k} \frac{|a_k(u,u))|}{\|u\|^2_{\mathcal{H}^1_k}} > 0. \tag{1.15}$$

Then, any $u \in \mathcal{H}^1_k$ admits a unique Galerkin approximation u_h, and it satisfies the estimate

$$\|u - u_h\|_{\mathcal{H}^1_k} \leq \frac{\|a_k\|}{\gamma(k)} \inf_{v_h \in V_h} \|u - v_h\|_{\mathcal{H}^1_k}, \tag{1.16}$$

where $\|a_k\|$ is as in (1.8).

Proof. It suffices to show that if a Galerkin approximation of u exists, then it satisfies the estimates (1.16); indeed, applied to $u = 0$, this estimate implies $u_h = 0$, and existence and uniqueness for all $u \in \mathcal{H}^1_k$ then follow from Exercise 1.10.

If a Galerkin approximation exists, then by Galerkin orthogonality (1.14),

$$a_k(u - u_h, u - u_h) = a_k(u - u_h, u - v_h) + \underbrace{a_k(u - u_h, \overbrace{v_h - u_h}^{\in V_h})}_{=0}.$$

Thus, by definition of γ,

$$|a_k(u - u_h, u - v_h)| = |a_k(u - u_h, u - u_h)| \geq \gamma \|u - u_h\|^2_{\mathcal{H}^1_k}. \tag{1.17}$$

On the other hand, by definition of $\|a_k\|$,

$$|a_k(u - u_h, u - v_h)| \leq \|a_k\| \|u - u_h\|_{\mathcal{H}^1_k} \|u - v_h\|_{\mathcal{H}^1_k}. \tag{1.18}$$

The combination of eqs. (1.17) and (1.18) implies

$$\|u - u_h\|_{\mathcal{H}^1_k} \leq \frac{\gamma}{\|a_k\|} \|u - v_h\|_{\mathcal{H}^1_k},$$

and the estimate (1.16) follows since $v_h \in V_h$ was arbitrary. □

Remark 1.5.

1. The estimate (1.16) implies that the ratio $\|u - u_h\|_{\mathcal{H}^1_k} / \inf_{v_h \in V_h} \|u - v_h\|_{\mathcal{H}^1_k}$ is bounded by a constant *independent of the space* V_h (but which may depend on k). This very strong property is known as "quasi-optimality": it guarantees that **if V_h contains a good approximation of u, the Galerkin method will find it** (up to a constant). In practice, one can systematically *ensure* that V_h contains such a good approximation by defining V_h as a space of piecewise polynomial functions on a sufficiently fine grid of Ω (see §2.1 below).

2. By assumption 1.3, $\|a_k\|$ is k-uniformly bounded. Hence, if (1.15) holds where $\gamma(k)$ is k-uniformly bounded as well, then (1.16) in fact gives a k-uniform bound on the above ratio. This is known as **k-uniform quasi-optimality**. It is a *key* property, as it guarantees that the departure from optimality of the Galerkin approximation will not explode as $k \to \infty$ (i.e., at high-frequency).

The main drawback of Céa's lemma is the assumption (1.15): it is *not* satisfied for large k by typical Helmholtz problems (see Exercise 1.11). Nevertheless, it holds "up to a lower order term" since the Gårding inequality (Assumption 1.4) gives

$$|a_k(u, u)| \geq c_{\mathrm{GA}} \|u\|^2_{\mathcal{H}^1_k} - C_{\mathrm{GA}} \|u\|^2_{\mathcal{H}}$$

and the $\mathcal{H}$ norm is "weaker" than the $\mathcal{H}^1_k$ norm. The main result of this section, due to Schatz [43], exploits this fact to obtain a sufficient condition for k-uniform quasi-optimality to hold. For this, we follow [42] by introducing the following key quantity. Recall the definition of $R(k)^*$ from (1.12).

Definition 1.5 (Adjoint approximability constant)

For any $k > 0$, the *adjoint approximability constant* $\eta(k, V_h)$ is defined by

$$\eta(k, V_h) := \|(\mathrm{Id} - \Pi_h)R(k)^*\|_{\mathcal{H}\to\mathcal{H}^1_k}$$

where $\Pi_h : \mathcal{H}^1_k \to \mathcal{H}^1_k$ is the $\mathcal{H}^1_k$-orthogonal projection onto V_h.

In words, the adjoint approximability measures the ability of the space V_h to approximate in the $\mathcal{H}^1_k$ norm the solutions u' of the adjoint Helmholtz problem

$$P(k)^* u' = f'$$

with data f' of unit $\mathcal{H}$-norm. The size of $\eta(k, V_h)$ is thus determined by the balance between (i) the approximation power of the space V_h (how "fine" the space V_h is) and (ii) the growth of the adjoint resolvent $R(k)^*$.

Theorem 1.9 (The Schatz argument)

Suppose that Assumptions (1.3)-(1.6) hold and let $k_0 > 0$. There exists $\eta_0 > 0$ such that for any $k \geq k_0$, if the approximation space V_h satisfies

$$\eta(k, V_h) \leq \eta_0, \tag{1.19}$$

then every function $u \in \mathcal{H}^1_k$ admits a unique Galerkin approximation u_h in V_h. Moreover, the error $u - u_h$ satisfies the estimates

$$\|u - u_h\|_{\mathcal{H}} \leq \eta(k, V_h)\|a_k\| \cdot \|u - u_h\|_{\mathcal{H}^1_k}, \tag{1.20}$$

$$\|u - u_h\|_{\mathcal{H}^1_k} \leq 2\frac{\|a_k\|}{c_{\mathrm{Ga}}} \inf_{v_h \in V_h} \|u - v_h\|_{\mathcal{H}^1_k}. \tag{1.21}$$

Remark 1.6. The estimate (1.21) implies a k-uniform quasi-optimality, but with *only if* $\eta(k, V_h)$ *is sufficiently small*. We will see that this can be a **severe price to pay** for large k. Namely, when V_h is chosen as a space of piecewise polynomials on a mesh of Ω, the condition $\eta(k, V_h) \leq \eta_0$ can require the mesh to be drastically finer than the minimum requirement to make

$$\inf_{v_h \in V_h} \|u - v_h\|_{\mathcal{H}^1_k} \leq \varepsilon\,,$$

for some desired error level ε. The reason for this is that there can exist a $\mathcal{H}$-normalized function f' for which the solution of the adjoint problem $P(k)^* u' = f'$ is not well approximated by V_h, i.e., $\inf_{v_h \in V_h} \|u' - v_h\| \gg 1$, even though $\inf_{v_h \in V_h} \|u - v_h\| \leq \varepsilon$.

Proof. As in the proof of Céa's lemma, it suffices to show that if a Galerkin approximation of u exists, then it satisfies the estimates (1.20) and (1.21).

Hence, suppose that u_h is a Galerkin approximation of u. Then

$$\begin{aligned}\|u-u_h\|_{\mathcal{H}}^2 &= \big\langle R(k)P(k)(u-u_h), u-u_h\big\rangle \\ &= \big\langle P(k)(u-u_h), R(k)^*(u-u_h)\big\rangle \\ &= \big\langle P(k)(u-u_h), R(k)^*(u-u_h)-v_h\big\rangle\end{aligned}$$

where we used duality and the Galerkin orthogonality to subtract an arbitrary element $v_h \in V_h$ in the right argument of the duality pairing. Since $v_h \in V_h$ is arbitrary, this gives

$$\|u-u_h\|_{\mathcal{H}}^2 \le \|a_k\|\|u-u_h\|_{\mathcal{H}_k^1} \underbrace{\inf_{v_h\in V_h} \|R_k^*(u-u_h)-v_h\|_{\mathcal{H}_k^1}}_{=\|(\mathrm{Id}-\Pi_h)R(k)^*(u-u_h)\|_{\mathcal{H}_k^1}},$$

(recalling that $\|P(k)\|_{\mathcal{H}_k^1\to(\mathcal{H}_k^1)^*} = \|a_k\|$) and the estimate (1.20) follows by definition of $\eta(k,V_h)$.

In turn, to show (1.21), we combine the previous estimate with the Gårding inequality (Assumption 1.4), the k-uniform continuity (Assumption 1.3), and (1.19):

$$\begin{aligned}\|u-u_h\|_{\mathcal{H}_k^1}^2 &\le c_{\mathrm{GA}}^{-1}\left(a_k(u-u_h,u-u_h)+C_{\mathrm{Ga}}\|u-u_h\|_{\mathcal{H}}^2\right) \\ &\le c_{\mathrm{GA}}^{-1}\left(a_k(u-u_h,u-u_h)+C_{\mathrm{Ga}}\eta(k,V_h)^2\|a_k\|^2\|u-u_h\|_{\mathcal{H}_k^1}^2\right) \\ &\le c_{\mathrm{GA}}^{-1}\left(a_k(u-u_h,u-u_h)+C_{\mathrm{Ga}}\eta_0^2C_0^2\|u-u_h\|_{\mathcal{H}_k^1}^2\right)\end{aligned}$$

Substracting the last term and using Galerkin orthogonality,

$$\begin{aligned}\|u-u_h\|_{\mathcal{H}_1^k}^2\left(1-\eta_0^2C_0^2\frac{C_{\mathrm{Ga}}}{c_{\mathrm{Ga}}}\right) &\le c_{\mathrm{Ga}}^{-1}a_k(u-u_h,u-u_h) \\ &= c_{\mathrm{Ga}}^{-1}a_k(u-u_h,u-v_h) \\ &\le \frac{\|a_k\|}{c_{\mathrm{Ga}}}\|u-u_h\|_{\mathcal{H}_k^1}\|u-v_h\|_{\mathcal{H}_k^1}.\end{aligned}$$

We obtain (1.21) by choosing $\eta_0 := \frac{1}{C_0}\sqrt{\frac{c_{\mathrm{Ga}}}{2C_{\mathrm{Ga}}}}$ and taking the infimum for $v_h \in V_h$. □

The previous result shows that, for analysing the Galerkin method, it is necessary to understand the adjoint approximability constant $\eta(k,V_h)$, as well as the best approximation error $\inf_{v_h\in V_h}\|u-v_h\|$. These quantities depend on the choice of V_h, and we now have to be more specific about it in order to continue the theory. This is the object of the next chapter.

Exercise 1.11. (Indefiniteness of a_k).
Let $\Omega\subset\mathbb{R}^d$ be a non-empty open set. Suppose that $C_c^\infty(\Omega)\subset\mathcal{H}_k^1$ and that for $u,v\in C_c^\infty(\Omega)$, a_k is given by

$$a_k(u,v)=\int_\Omega (k^{-2}\nabla u\cdot\nabla v - uv)\,dx\,,$$

Show that for k large enough, there exists $u_1,u_2\in C_c^\infty(\Omega)$ such that

$$a_k(u_1,u_1)<0<a_k(u_2,u_2).$$

Deduce that the quantity $\eta(k)$ defined in (1.15) vanishes for k large enough.

Chapter 2

The finite-element method and the pollution effect

We will restrict our attention to a specific Galerkin method, namely, the *finite-element method*.[1] This is when the approximation spaces V_h are built as piecewise polynomial functions on a "mesh" of the computational domain.

The goal of this chapter is to give sharp bounds on the error $u - u_h$ when the finite-element method is applied to a Helmholtz problem in terms of the wavenumber k, the mesh-size h and the polynomial degree p. The main result is Theorem 2.7 below. A key step will be to estimate the behavior of the adjoint approximability constant $\eta(k, V_h)$ of Definition 1.5 for piecewise polynomial spaces. It is already apparent on the definition

$$\eta(k, V_h) = \|(\mathrm{Id} - \Pi_h) R(k)^* \|_{\mathcal{H} \to \mathcal{H}^1_k}$$

that this involves two phenomena:

(i) the best polynomial approximation error is smaller for more regular functions, and

(ii) the solution of a Helmholtz problem (more generally, of an elliptic PDE) is more regular than the data.

We will see that the finite-element method is plagued by the so-called "pollution effect": roughly, this means that, perhaps surprisingly, a mesh-size of order $h \lesssim k^{-1}$ is generally insufficient to obtain an accurate approximation of a Helmholtz problem with wavenumber k.

2.1 Finite-element spaces

The finite-element method is a special case of a Galerkin method, where the approximation space V_h is chosen as a special space involving polynomials. More precisely, one starts by constructing a

[1] developed around the 1960s, building on early ideas from Courant [14] (1943) among others. Its development was especially stimulated by the possibilities offered by the appearance of computers. In France, its mathematical analysis for the approximation of elliptic PDEs was pioneered by P. Ciarlet, see [12]. A standard modern reference is [6].

mesh Ω_h of the computational domain Ω, i.e., a partition of Ω into a finite set of (possibly curved) simplices (or sometimes other shapes). As an illustration, a simple mesh of a planar domain Ω is represented in Figure 2.1. The simplices are called *elements* of Ω_h. It is customary to denote

$$h := \max_{K \in \Omega_h} h_K \,, \quad h_K := \text{diameter}(K)$$

the *mesh-width* of Ω_h. Since h determines in a large part the accuracy of the approximation, the subscript h is traditionally used to indicate an object related to the mesh (u_h, V_h, Ω_h, etc.).

A *finite-element scheme* V is an assignment, to any mesh Ω_h, of a finite-dimensional subspace $V_h = V(\Omega_h)$ of $\mathcal{H}^1$. For example, the standard Lagrange finite-element of order p is the scheme V^p defined by

$$V^p(\Omega_h) := \big\{ u_h \in \mathcal{H}^1 \;\mid\; (u_h)|_K \text{ is a polynomial of degree } p,\ \forall K \in \Omega_h \big\}.$$

There are many other popular finite-element schemes, and we will not review them here.

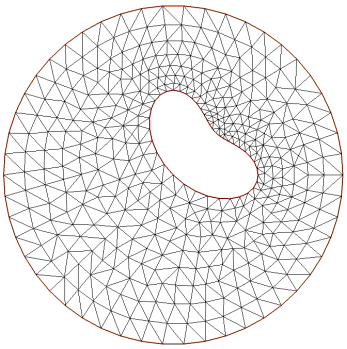


Figure 2.1: A triangular mesh Ω_h of a planar domain Ω (delimited by the red borders).

In practice, the ability of a finite element space to approximate functions does not just depend on h and p, but also on the "quality" of the mesh. A mesh of good quality is one where the elements are not too distorted/elongated. This is described by the so-called *shape-regularity constant*

$$\gamma(\Omega_h) := \max_{K \in \Omega_h} \frac{h_K}{\rho_K}$$

where, for every K, ρ_K is the in-radius of K (i.e., the radius of the inscribed ball). There are efficient methods to obtain arbitrarily refined meshes of a given computational domain, with uniformly bounded shape-regularity constants.

We now state the general approximation property satisfied by standard finite-element schemes. Recall the scale of spaces $(\mathcal{H}^n_k)_{n \in \mathbb{N}}$ from Definition 1.1. It is not our aim here to present the proof of these results: we take them as requirements for admissible finite-element schemes.

Definition 2.1 (Finite-element scheme of order p)

Given $p \in \mathbb{N}$, we say that the finite-element scheme V is *of order* p if it has the following **approximation property** (see [6, Theorem 4.4.20]): for every $k_0, \gamma_0 > 0$, there exists $C > 0$ such that the estimate

$$\inf_{v_h \in V(\Omega_h)} \|u - v_h\|_{\mathcal{H}_k^m} \leq C(hk)^{\ell - m} \|u\|_{\mathcal{H}_k^\ell} \tag{2.1}$$

holds for any $\ell \in \{1, \dots, p+1\}$, $m \in \{0, \dots, \ell\}$, $u \in \mathcal{H}_k^\ell$, and any mesh Ω_h satisfying $\gamma(\Omega_h) \leq \gamma_0$ of meshwidth h.

2.2 Elliptic regularity

We now formulate our fifth key assumption on the Helmholtz problems under consideration (again following [24]). Let

$$\mathscr{P}(k) := \frac{P(k) + P(k)^*}{2} \quad : \quad \mathcal{H}_k^1 \to (\mathcal{H}_k^1)^*$$

where $P(k)^* : \mathcal{H}_k^1 \to (\mathcal{H}_k^1)^*$ is the adjoint of $P(k)$.

Assumption 2.1 (Elliptic regularity for $\mathscr{P}(k)$ and $P(k)^*$)

For every $k_0 > 0$ and $n \in \mathbb{N} \setminus \{0\}$, there is a constant $C_{\text{ell}} > 0$ such that the following property holds for all $k \geq k_0$:

$$\forall u \in \mathcal{H}_k^1\,, \quad Qu \in \mathcal{H}_k^n \implies \Big(u \in \mathcal{H}_k^{n+2} \quad \text{and} \quad \|u\|_{\mathcal{H}_k^{n+2}} \leq C_{\text{ell}} \Big(\|u\|_{\mathcal{H}_k^1} + \|Qu\|_{\mathcal{H}_k^n} \Big) \Big),$$

where Q is either one of the operators $\mathscr{P}(k)$ or $P(k)^*$.

Remark 2.1 (Elliptic regularity in concrete settings). Elliptic regularity is a general property of elliptic second-order partial differential equations. In the setting of §1.1, with the spaces $(\mathcal{H}_k^n)_{n \in \mathbb{N}}$ defined as in Remark 1.3, the elliptic regularity holds if the coefficients $\tilde{\rho}, \tilde{n}$ and the boundary $\partial\Omega$ are smooth, see [33, Theorem 4.18 (i)] (note that the requirement of the boundary values required by this theorem – i.e., that $\gamma u \in H^{r+\frac{3}{2}}(\partial\Omega)$ – is fulfilled since $\gamma u = 0$ due to the Dirichlet boundary condition in $\mathcal{H}_k^1$).

Exercise 2.1. (Mapping properties of $R(k)^*$).
Suppose that Assumptions (1.3)-(1.6) and (2.1) hold. Show that for every $n \in \mathbb{N}$ and $k > 0$, $R(k)^*$ maps $\mathcal{H}_k^n$ to $\mathcal{H}_k^{n+2}$ continuously with

$$\forall k_0 > 0\,, \exists C > 0 : \quad \|R(k)^*\|_{\mathcal{H}_k^n \to \mathcal{H}_k^{n+2}} \leq C(1 + \rho(k)). \tag{2.2}$$

(Hint: Proceed by induction, using Exercise 1.9 for the initialization).

Lemma 2.2 (First bound on the adjoint-approximability constant)

Suppose that Assumptions (1.3)-(1.6) and (2.1) hold and let V^p be a finite-element scheme of order $p \geq 1$ and let $k_0, \gamma_0 > 0$. Then, there exists $C > 0$ such that the estimate

$$\eta(k, V_h) \leq C(1 + \rho(k))hk \tag{2.3}$$

holds for any $k > k_0$, $V_h = V^p(\Omega_h)$ with Ω_h any mesh satisfying $\gamma(\Omega_h) \leq \gamma_0$.

Proof. Let $f \in \mathcal{H}$. Applying the approximation property (2.1) to $R(k)^* f$, we find

$$\|(\mathrm{Id} - \Pi_h)R(k)^* f\|_{\mathcal{H}^1_k} \leq C(hk)^{2-1}\|R(k)^* f\|_{\mathcal{H}^2_k}$$

and the conclusion follows by using the estimate (2.2) with $n = 0$. □

It follows from Lemma 2.2 and the Schatz argument (Theorem 1.9) that a Galerkin approximation exists and is k-uniformly quasi-optimal provided that $(1 + \rho(k))hk$ is sufficiently small. In what follows, we show the following stronger bound:

Lemma 2.3 (Sharp bound on the adjoint-approximability constant)

Under the same assumptions as Lemma 2.2, there exists $C > 0$ such that the estimate

$$\eta(k, V_h) \leq C\left((hk) + \rho(k)(hk)^p\right) \tag{2.4}$$

holds for all $k \geq k_0$ and $V_h = V^p(\Omega_h)$ with Ω_h any mesh satisfying $\gamma(\Omega_h) \leq \gamma_0$.

This immediately implies the following result

Corollary 2.4 (Error bound in the "asymptotic regime")

Suppose that Assumptions (1.3)-(1.6) and (2.1) hold, let V^p be finite-element scheme of order $p \geq 1$ and let $k_0, \gamma_0 > 0$. There exists $\varepsilon > 0$ and $C > 0$ such that for all $k \geq k_0$, if Ω_h is a mesh of Ω satisfying $\gamma(\Omega_h) \leq \gamma_0$ and

$$(1 + \rho(k))(hk)^p \leq \varepsilon, \tag{2.5}$$

then every $u \in \mathcal{H}^1_k$ admits a unique Galerkin approximation u_h in $V^p(\Omega_h)$ which satisfies

$$\|u - u_h\|_{\mathcal{H}^1_k} \leq C \inf_{v_h \in V^p(\Omega_h)} \|u - v_h\|_{\mathcal{H}^1_k}.$$

Remark 2.2 (Sharpness of Corollary 2.4). It is known in practice that Corollary 2.4 is sharp, and we illustrate this with Numerical experiments in Figure 2.2. This means that to ensure k-uniform quasi-optimality, it is *not* sufficient to take $h \lesssim k^{-1}$: the mesh needs to be more refined than expected, especially when $\rho(k)$ is large. Only when the stronger condition (2.5) is satisfied, do we enter the *asymptotic regime*, namely, the regime in which **the finite-element approximation is essentially the best approximation**.

The fact that the condition $h \lesssim k^{-1}$ is *not* enough to guarantee k-uniform quasi-optimality is known as the **pollution effect**. A larger polynomial order mitigates this effect, since Corollary 2.4 ensures that it is sufficient to choose $h \sim k^{-1}\rho(k)^{-\frac{1}{p}}$ to obtain a k-uniform bound on the ratio (Galerkin error)/(Best approximation error). In fact, it can be shown that one may completely get rid of the pollution effect by choosing not just h, but also p as a function of k, with $p \gtrsim \log k$ and $hk/p \lesssim 1$ [34].[2]

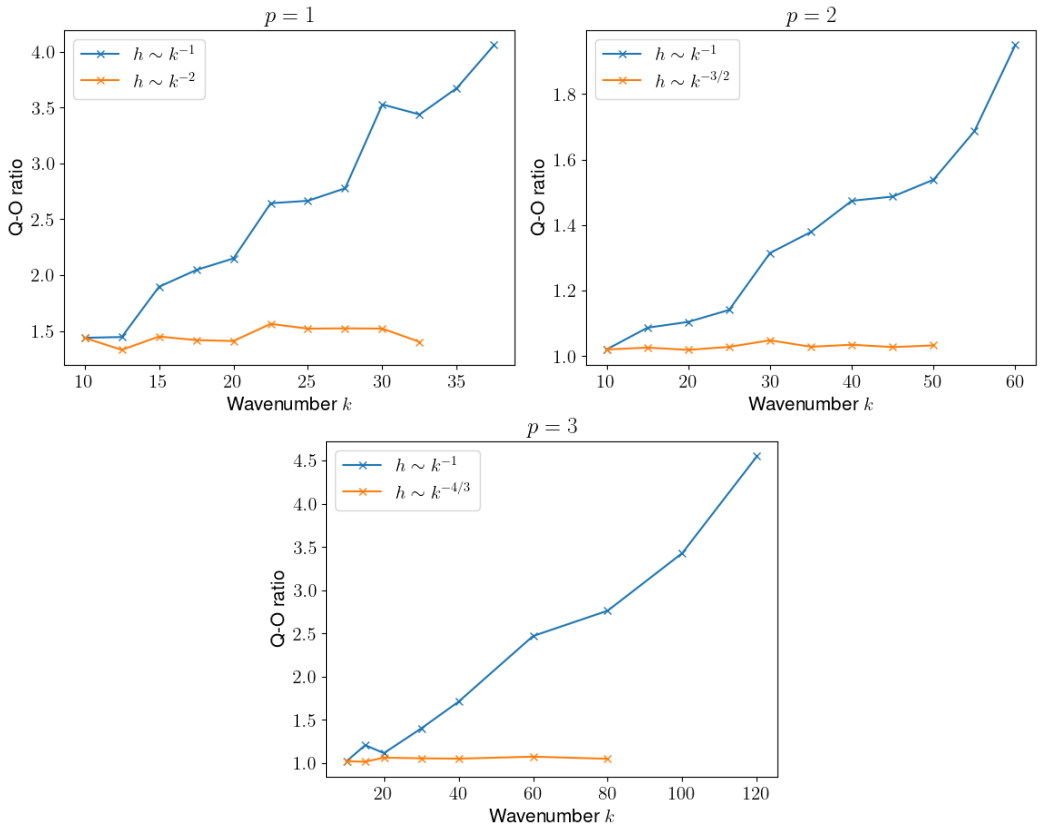


Figure 2.2: Numerical experiments for the resolution of a Helmholtz problem $(-k^{-2}\Delta - 1)u = f$ in $\mathbb{R}^d$ with the Sommerfeld condition (1.1) (truncated to a bounded domain via PML truncation), where the exact solution is $u = \chi(x)e^{ikx_1}$, $\chi = e^{-\sigma\|x^2\|}$ with $\sigma = 10$. For this problem, one can show that $ck \le \rho(k) \le Ck$. The problem is solved numerically for a sequence of wavenumbers k, with a finite-element scheme of order $p = 1, 2$ or 3, either with a mesh-size $hk = \text{Cst}$ (i.e., $h \sim k^{-1}$) or $(hk)^p\rho(k) = \text{Cst}$ (i.e., $h \sim k^{-(p+1)/p}$). In each case, we plot the quasi-optimality ("Q-O") ratio, defined as (Galerkin error)/(Best approximation error). The fact that the blue curves are increasing is a manifestation of the pollution effect, and the fact that the orange curves stay bounded is predicted by Corollary 2.4. Missing data on the orange curves is because of memory limitation.

[2]Pollution is defined more precisely as the fact that more than $O(k^d)$ degrees of freedom are necessary to obtain k-uniform quasi-optimality, and one can show that for $p \gtrsim \log k$ and $hk/p \lesssim 1$, the number of degrees of freedom is indeed of that order.

2.3 Frequency splitting of the resolvent

The main idea for the proof of Lemma 2.3 is the following phenomenon: the behaviour of the Helmholtz resolvent (and similarly for its adjoint) is only "bad" at **low frequencies**. But low-frequency functions (i.e., functions that are slowly varying) are very well-approximated by piecewise polynomials.

The reason why the Helmholtz resolvent is well-behaved on high-frequency functions can be understood informally on the following example. Consider the Helmholtz equation in free space

$$(-k^{-2}\Delta - 1)u = f \qquad \text{in } \mathbb{R}^d,$$

with the Sommerfeld radiation condition (1.1). Via the Fourier transform, the resolvent operator $(-k^{-2}\Delta - 1)^{-1}$ can essentially be seen as the multiplication by $1/(k^{-2}|\xi|^2 - 1)$ in Fourier space. In particular, the "troubles", i.e., the only region of Fourier space where the resolvent can be large, are restricted to the set $\{k^{-1}|\xi| \lesssim 1\}$ i.e., the *low-frequencies*.

In our setting, we will formalize this idea by constructing a "low-frequency cutoff" $S(k)$ (i.e., $S(k)$ removes frequencies $\lesssim k$) such that the perturbation

$$P^\sharp(k) := P(k) + S(k)$$

has a well-behaved inverse $R^\sharp(k) := [P(k) + S(k)]^{-1}$. We can then split the resolvent as

$$\begin{aligned} R(k) &= R^\sharp(k) + R(k)[P^\sharp(k) - P(k)]R^\sharp(k) \\ &= R^\sharp(k) + R(k)S(k)R^\sharp(k). \end{aligned} \tag{2.6}$$

The way to think about this splitting is that $R^\sharp(k)$ is the "high-frequency part of $R(k)$" which is well-behaved, and the second term $R(k)S(k)R^\sharp(k)$ has the "bad behavior", but restricted to low-frequencies thanks to $S(k)$.

We now give a precise statement of the properties of the operators $S(k)$ and $P^\sharp(k)$, but postpone their construction to a later stage.

Proposition 2.5 (Operators $S(k)$ and $P^\sharp(k)$)

Suppose that Assumptions (1.3)-(1.6) and (2.1) hold and let $k_0 > 0$. Then, for every $k \geq k_0$, there exists a bounded self-adjoint operator

$$S(k) : \mathcal{H} \to \mathcal{H}$$

with the following properties

1. $S(k)$ is *smoothing*, in the sense that for every $k \geq k_0$ and $n \in \mathbb{N}$, $S(k)$ maps $(\mathcal{H}^n_k)^*$ to $\mathcal{H}^n_k$ with
$$\sup_{k \geq k_0} \|S(k)\|_{(\mathcal{H}^n_k)^* \to \mathcal{H}^n_k} < \infty.$$
2. The operator $P^\sharp(k) := P(k) + S(k)$ is *k-uniformly coercive*: there exists $c^\sharp > 0$ such that for all $k \geq k_0$,
$$\mathrm{Re}\langle P^\sharp(k)u, u\rangle \geq c^\sharp \|u\|^2_{\mathcal{H}^1_k}.$$

3. The resolvent operator $R^\sharp(k) := P^\sharp(k)^{-1}$ *gains two derivatives k-uniformly*, in the sense that for all $n \in \mathbb{N}$, $R^\sharp(k)$ maps $\mathcal{H}_k^n$ to $\mathcal{H}_k^{n+2}$ and

$$\sup_{k \geq k_0} \|R^\sharp(k)\|_{\mathcal{H}_k^n \to \mathcal{H}_k^{n+2}} < \infty.$$

We obtain a better bound for $\eta(k, V_h)$ as an immediate consequence.

Corollary 2.6 (Bound for $\eta(k, V_h)$ via frequency splitting)

Under the same assumptions as Lemma 2.2, the exists $C > 0$ such that the estimate

$$\eta(k, V_h) \leq C\left((hk) + \rho(k)(hk)^p\right) \tag{2.7}$$

holds for all $k \geq k_0$ and $V_h = V^p(\Omega_h)$ with Ω_h any mesh satisfying $\gamma(\Omega_h) \leq \gamma_0$.

Proof. Using the splitting (2.6) (taking the adjoint) and the definition of $\eta(k, V_h)$, and applying the triangle inequality

$$\eta(k, V_h) \leq \|(\mathrm{Id} - \Pi_h)R^\sharp(k)^*\|_{\mathcal{H} \to \mathcal{H}_k^1} + \|(\mathrm{Id} - \Pi_h)R^\sharp(k)^* S(k) R(k)^*\|_{\mathcal{H} \to \mathcal{H}_k^1}.$$

Thus by the approximation property (2.1),

$$\eta(k, V_h) \leq C(hk)\|R^\sharp(k)\|_{\mathcal{H} \to \mathcal{H}_k^2} + C(hk)^p \|R^\sharp(k) S(k) R(k)\|_{\mathcal{H} \to \mathcal{H}_k^{p+1}}.$$

By Proposition 2.5, we have $\|R^\sharp(k)\|_{\mathcal{H} \to \mathcal{H}_k^2} \leq C$ and

$$\begin{aligned}\|R^\sharp(k) S(k) R(k)\|_{\mathcal{H} \to \mathcal{H}_k^{p+1}} &\leq \|R^\sharp(k)\|_{\mathcal{H}_k^{p-1} \to \mathcal{H}_k^{p+1}} \|S(k)\|_{(\mathcal{H}_k^{p-1})^* \to \mathcal{H}_k^{p-1}} \|\mathrm{Id}\|_{\mathcal{H} \to (\mathcal{H}_k^{p-1})^*} \|R(k)\|_{\mathcal{H} \to \mathcal{H}} \\ &\leq C\rho(k)\end{aligned}$$

by the definition of $\rho(k)$ and the continuous embedding $\mathcal{H} \subset (\mathcal{H}_k^{p-1})^*$ (Exercise 1.6). □

2.4 Pre-asymptotic regime: the elliptic projection argument

In this section, we show the main result of this chapter, which improves the result of Corollary 2.4:

Theorem 2.7 (Error bound in the "pre-asymptotic regime")

Suppose that assumptions (1.3)-(1.6) and (2.1) hold, let V^p be a finite-element scheme of order $p \geq 1$, and let $k_0, \gamma_0 > 0$. There exits $\varepsilon > 0$ and $C > 0$ such that for all $k \geq k_0$, if Ω_h is a mesh of Ω satisfying $\gamma(\Omega_h) \leq \gamma_0$ and

$$(1 + \rho(k))(hk)^{2p} \leq \varepsilon\,,$$

then every $u \in \mathcal{H}^1_k$ admits a unique Galerkin approximation u_h in $V^p(\Omega_h)$, and

$$\|u - u_h\|_{\mathcal{H}^1_k} \leq C(1 + \rho(k)(hk)^p) \inf_{v_h \in V^p(\Omega_h)} \|u - v_h\|_{\mathcal{H}^1_k}. \tag{2.8}$$

Remark 2.3 (Comments on 2.4).

(i) This result is part of Theorem 1.1 from [24], and we refer to §1.2 of that reference for a discussion of its history, which started in dimension $d = 1$ and for $p = 1$, for "non-trapping problems" (this corresponds to cases where $\rho(k) \lesssim k$) in the seminal works by Ihlenburg and Babuška [28, 29].

(ii) In the asymptotic regime (that is, when the condition (2.5) holds), the bound (2.8) does not give any new information compared to Corollary 2.4, and the criterion for obtaining k-uniform quasi-optimality remains the same. However, Theorem 2.7 gives new information in the so-called "pre-asymptotic regime", that is, when $\rho(k)(hk)^p$ is not small, but $\rho(k)(hk)^{2p}$ is. In particular, using the approximation property (2.1), we see that for $u \in \mathcal{H}^{p+1}_k$,

$$\|u - u_h\|_{\mathcal{H}^1_k} \leq C(1 + \rho(k)(hk)^p)(hk)^p \|u\|_{\mathcal{H}^{p+1}_k}$$

and thus, under the assumptions of Theorem 2.7, we obtain a k-uniform bound on the *relative error*

$$\frac{\|u - u_h\|_{\mathcal{H}^1_k}}{\|u\|_{\mathcal{H}^{p+1}_k}}.$$

(iii) Even if one only cares about the asymptotic regime k-uniform quasi-optimality, the proof of Theorem 2.7 introduces a key argument (the "elliptic projection argument") that will be our starting point in the next chapter.

Definition 2.2 (Elliptic projection $\Pi^\sharp_h$)

Let Assumptions (1.3)-(1.6) and (2.1) hold and let $k_0 > 0$. For any $k \geq k_0$, let $V_h \subset \mathcal{H}^1_k$. The *elliptic projection* onto V_h is the unique bounded linear operator $\Pi^\sharp_h : \mathcal{H}^1_k \to V_h$ satisfying

$$\langle P^\sharp(k)u_h, (\mathrm{Id} - \Pi^\sharp_h)v\rangle = 0 \qquad \forall (u_h, v) \in V_h \times \mathcal{H}^1_k,$$

where $P^\sharp(k)$ is as in Proposition 2.5.

Due to the k-uniform coercivity of $P^\sharp(k)$, the operator $\Pi^\sharp_h$ essentially computes the best approximation in V_h with respect to the $\mathcal{H}^1_k$ norm. This is the object of Exercise 2.2 below.

Exercise 2.2. (Aubin-Nitsche trick for $\Pi^\sharp_h$).

1. *Show that the operator $\Pi^\sharp_h$ is well-defined and that for every $k_0 > 0$, there exists $C > 0$ such that*

$$\|(\mathrm{Id} - \Pi^\sharp_h)u\|_{\mathcal{H}^1_k} \leq C \inf_{v_h \in V_h} \|u - v_h\|_{\mathcal{H}^1_k}.$$

(Hint: use Céa's lemma).

2. *Using the definition of $\Pi_h^\sharp$, show that for any $\xi \in \mathcal{H}_k^1$,*

$$|\langle \xi, (\mathrm{Id} - \Pi_h^\sharp)u\rangle| \leq \|(\mathrm{Id} - \Pi_h^\sharp)u\|_{\mathcal{H}_k^1} \inf_{w_h \in V_h} \|R^\sharp \xi - w_h\|_{\mathcal{H}_k^1}.$$

(Hint: start as in the Schatz argument)

3. *Using the two previous questions, deduce that, under the assumptions of Theorem 2.7,*

$$\|(\mathrm{Id} - \Pi_h^\sharp)u\|_{(\mathcal{H}_k^{p-1})^*} \leq C(hk)^p \inf_{v_h \in V_h} \|u - v_h\|_{\mathcal{H}_k^1}.$$

where C does only depends on k_0, γ_0 and V^p. This method is sometimes called the "Aubin-Nitsche trick", see [2, 37].

Proof of Theorem 2.7. We first show that if $(1+\rho(k))(hk)^{2p}$ is sufficiently small, then

$$\|u - u_h\|_{(\mathcal{H}_k^{p-1})^*} \leq C(1+\rho(k))(hk)^p \inf_{v_h \in V^p(\Omega_h)} \|u - v_h\|_{\mathcal{H}_k^1}. \tag{2.9}$$

For this, we fix $\xi \in \mathcal{H}_k^{p-1}$ and compute (starting as in the Schatz argument)

$$\begin{aligned}
\langle u - u_h, \xi\rangle &= \langle R(k)P(k)(u-u_h), \xi\rangle && \text{(def. of } R(k)\text{)}\\
&= \langle P(k)(u-u_h), R(k)^*\xi\rangle && \text{(duality)}\\
&= \langle P(k)(u-u_h), (\mathrm{Id} - \Pi_h^\sharp)R(k)^*\xi\rangle && \text{(Galerkin orthogonality (1.14))}\\
&= \langle P^\sharp(k)(u-u_h), (\mathrm{Id} - \Pi_h^\sharp)R(k)^*\xi\rangle\\
&\qquad - \langle S(k)(u-u_h), (\mathrm{Id} - \Pi_h^\sharp)R(k)^*\xi\rangle && \text{(def. of } S \text{ and } P^\sharp\text{)}\\
&= \langle P^\sharp(k)(u-v_h), (\mathrm{Id} - \Pi_h^\sharp)R(k)^*\xi\rangle\\
&\qquad - \langle S(k)(u-u_h), (\mathrm{Id} - \Pi_h^\sharp)R(k)^*\xi\rangle && \text{(def. of } \Pi_h^\sharp\text{)}
\end{aligned}$$

where v_h is an arbitrary element of $V^p(\Omega_h)$ (the fact that we have gotten a v_h in the last step is the main gain from using the elliptic projection) Using the boundedness of $P^\sharp(k) : \mathcal{H}_k^1 \to (\mathcal{H}_k^1)^*$ and $S(k) : (\mathcal{H}_k^{p-1})^* \to \mathcal{H}_k^{p-1}$ (from Proposition 2.9), we deduce

$$\begin{aligned}
|\langle u - u_h, \xi\rangle| \leq{}& C\|u - v_h\|_{\mathcal{H}_k^1} \|(\mathrm{Id} - \Pi_h^\sharp)R(k)^*\xi\|_{\mathcal{H}_k^1}\\
&+ C\|u - u_h\|_{(\mathcal{H}_k^{p-1})^*} \|(\mathrm{Id} - \Pi_k^\sharp)R(k)^*\xi\|_{(\mathcal{H}_k^{p-1})^*}.
\end{aligned}$$

Applying the properties of $\Pi_h^\sharp$ from Exercise 2.2, it follows that

$$\begin{aligned}
|\langle u - u_h, \xi\rangle| \leq{}& C\|u - v_h\|_{\mathcal{H}_k^1} \|(\mathrm{Id} - \Pi_h)R(k)^*\xi\|_{\mathcal{H}_k^1}\\
&+ C(hk)^p \|u - u_h\|_{(\mathcal{H}_k^{p-1})^*} \|(\mathrm{Id} - \Pi_h)R(k)^*\xi\|_{\mathcal{H}_k^1}.
\end{aligned}$$

where $\Pi_h : \mathcal{H}_k^1 \to \mathcal{H}_k^1$ is the $\mathcal{H}_k^1$-orthogonal projection onto $V^p(\Omega_h)$. By the approximation property (2.1) and the norm estimates for $R(k)^*$ from Exercise 1.9 we obtain

$$|\langle u - u_h, \xi\rangle| \leq C(hk)^p \|u - v_h\|_{\mathcal{H}_k^1} \|R(k)^*\xi\|_{\mathcal{H}_k^{p+1}}$$

$$+ C(hk)^{2p}\|u-u_h\|_{(\mathcal{H}_k^{p-1})^*}\|R(k)^*\xi\|_{\mathcal{H}_k^{p+1}}.$$
$$\leq C(1+\rho(k))(hk)^p\left(\|u-v_h\|_{\mathcal{H}_k^1}+(hk)^p\|u-u_h\|_{(\mathcal{H}_k^{p-1})^*}\right)\|\xi\|_{\mathcal{H}_k^{p-1}}.$$

By taking the supremum over $\xi\in\mathcal{H}_k^{p-1}\setminus\{0\}$, we deduce that

$$\|u-u_h\|_{(\mathcal{H}_k^{p-1})^*}\leq C(1+\rho(k))(hk)^p\|u-v_h\|_{\mathcal{H}_k^1}+C(1+\rho(k))(hk)^{2p}\|u-u_h\|_{(\mathcal{H}_k^{p-1})^*}, \tag{2.10}$$

from which (2.9) follows since v_h was arbitrary.

To obtain the bound (2.8), we now write

$$\begin{aligned}
\|u-u_h\|_{\mathcal{H}_k^1}^2 &\leq C|\langle P^\sharp(k)(u-u_h),(u-u_h)\rangle| && \text{(coercivity of } P^\sharp)\\
&\leq C|\langle P(k)(u-u_h),(u-u_h)\rangle|+C|\langle S(k)(u-u_h),(u-u_h)\rangle| && \text{(def. of } P^\sharp)\\
&\leq C|\langle P(k)(u-u_h),(u-v_h)\rangle|+\|u-u_h\|_{(\mathcal{H}_k^{p-1})^*}^2 &&\\
&\qquad\qquad \text{(Galerkin orth. and mapping properties of } S(k))\\
&\leq \|u-u_h\|_{\mathcal{H}_k^1}\|u-v_h\|_{\mathcal{H}_k^1}+\|u-u_h\|_{(\mathcal{H}_k^{p-1})^*}^2 &&
\end{aligned}$$

and the conclusion follows by taking the supremum over $v_h\in V^p(\Omega_h)$ and using the bound (2.9) from the previous step. □

2.5 Construction of $S(k)$

We now give the construction of the operator $S(k)$ of Proposition 2.5. We have seen that $S(k)$ can be viewed as a low-frequency cutoff; hence to define it, we must be more precise about what this means. The trouble is that we cannot directly use the Fourier transform (since we are not on $\mathbb{R}^d$). Instead, we will diagonalize the Helmholtz operator – or more precisely, its real part $\mathscr{P}(k)$ – and use the eigenfunctions to play the role of the Fourier basis $e^{ix\cdot\xi}$, and the associated eigenvalues to play the role of the Fourier variable. This is the object of the next proposition

Proposition 2.8 (Spectral decomposition of $\mathscr{P}(k)$)

Suppose that Assumptions (1.3)-(1.6) and (2.1) hold and let $k_0>0$. Then there exists $c,C>0$ such that, for every $k\geq k_0$, there exists a sequence of real numbers

$$c_{\mathrm{Ga}}(k_0)-C_{\mathrm{Ga}}(k_0)\leq\lambda_0\leq\lambda_1\leq\dots$$

with $\lambda_j\to\infty$ as $j\to\infty$, and there exists a family $(u_j)_{n\in\mathbb{N}}\in(\mathcal{H}_k^1)^{\mathbb{N}}$ of associated eigenfunctions

$$\mathscr{P}(k)u_j=\lambda_j u_j,$$

such that

(i) $(u_j)_{j\in\mathbb{N}}$ is a Hilbert basis of $\mathcal{H}$,

(ii) $\forall u\in\mathcal{H}\,,\quad u\in\mathcal{H}_k^1\iff\displaystyle\sum_{j=0}^{\infty}(C_{\mathrm{Ga}}+\lambda_j)|\langle u,u_j\rangle|^2<\infty,$

(iii) $\forall u \in \mathcal{H}^1_k\,, \quad c\|u\|^2_{\mathcal{H}^1_k} \leq \sum_{j=0}^{\infty}(C_{\mathrm{Ga}}+\lambda_j)|\langle u, u_j\rangle|^2 \leq C\|u\|^2_{\mathcal{H}^1_k}.$

Moreover, for any $u \in \mathcal{H}^1_k$, one has

$$\mathscr{P}(k)u = \sum_{j=0}^{\infty} \lambda_j \langle u, u_j\rangle u_j \tag{2.11}$$

with the series converging in $(\mathcal{H}^1_k)^*$.

Exercise 2.3. **(Proof of Proposition 2.8).**

1. *Show that*
$$\operatorname{Re} a^+_k(u,v) := \big\langle (\mathscr{P}(k)+C_{\mathrm{Ga}})u, v\big\rangle$$
defines an equivalent inner product on $\mathcal{H}^1_k$.

2. *Show that there exists a bounded linear operator* $A(k) : \mathcal{H} \to \mathcal{H}^1_k$ *such that, for all* $f \in \mathcal{H}$,
$$(\operatorname{Re} a^+_k)(A(k)f, v) = \langle f, v\rangle$$
Show that $A(k)$ *can be viewed as a compact operator from* $\mathcal{H}$ *to* $\mathcal{H}$, *which is positive definite (i.e.,* $\langle A(k)u, u\rangle > 0$ *for every* $u \in \mathcal{H}$*) and satisfies* $\|A(k)\|_{\mathcal{H}\to\mathcal{H}} \leq \frac{1}{c_{\mathrm{Ga}}(k_0)}$.

3. *By the spectral theorem for compact self-adjoint operators applied to* $A(k)$, *there exists* $(\mu_j)_{j\in\mathbb{N}}$ *a sequence of positive numbers converging to* 0, *and* $(u_j)_{j\in\mathbb{N}}$ *a Hilbert basis of* $\mathcal{H}$ *such that* $A(k)u_j = \mu_j u_j$. *Show that* $u_j \in \mathcal{H}^1_k$ *and* $\mathscr{P}(k)u_j = \lambda_j u_j$ *with* $\lambda_j := \frac{1}{\mu_j} - C_{\mathrm{Ga}}$.

4. *and that* $(\sqrt{\mu_j}u_j)_{j\in\mathbb{N}}$ *is an orthonormal family of* $\mathcal{H}^1_k$ *for this inner product.*

5. *Show that* $(\sqrt{\mu_j}u_j)_{j\in\mathbb{N}}$ *is a Hilbert basis of* $\mathcal{H}^1_k$ *and deduce properties (i) and (ii).*

6. *Show that for any* $u, v \in \mathcal{H}^1_k$ *and* $N \in \mathbb{N}$,
$$\left\langle \sum_{j=0}^{N} \lambda_j \langle u, u_j\rangle, v \right\rangle \leq C\|u\|_{\mathcal{H}^1_k}\|v\|_{\mathcal{H}^1_k}$$
where C *is independent of* N *and* k, *and deduce that the series* (2.11) *converges in* $(\mathcal{H}^1_k)^*$.

7. *Conclude.*

We may interpret the decomposition (2.11) by viewing $\mathscr{P}(k)$ as a "pointwise multiplier" in the basis $(u_j)_{j\in\mathbb{N}}$, with its "symbol" given by λ_j. This opens the possibility to define other multipliers by

$$Xu := \sum_{j=0}^{\infty} \hat{x}_j \langle u, u_j\rangle u_j.$$

for some sequence of complex numbers $(\hat{x}_j)_{j=0}^{\infty}$. "Low-frequency cutoffs" will be operators of this form with only a finite number of non-zero coefficients $\hat{x}_j$. It will be convenient to think of them as functions of $\mathscr{P}(k)$, with $\hat{x}_j$ equal to $f(\lambda_j)$.

Definition 2.3 (Functions of $\mathscr{P}$)

Let $f : \mathbb{R} \to \mathbb{C}$ be bounded. For every $k > 0$ we define $f\big(\mathscr{P}(k)\big) : \mathcal{H} \to \mathcal{H}$ by

$$f\big(\mathscr{P}(k)\big)u := \sum_{j=0}^{\infty} f(\lambda_j)\langle u, u_j\rangle u_j$$

where λ_j and u_j are as in Proposition 2.8.

We will often omit the wavenumber k and simply write $\psi(\mathscr{P})$.

Exercise 2.4. (Elementary properties of $f(\mathscr{P})$).

1. *Show that the set of finite linear combinations of the functions u_j is dense in $\mathcal{H}$.*
2. *Deduce that for all $k > 0$,*
$$\|f(\mathscr{P})\|_{\mathcal{H}\to\mathcal{H}} \leq \sup_{x\in\mathbb{R}} |f(x)|$$
3. *Show that for f bounded, $f(\mathscr{P})^* = \overline{f}(\mathscr{P})$.*
4. *Show that for all $k \geq k_0$,*
$$\|f(\mathscr{P})\|_{\mathcal{H}_k^{-1}\to\mathcal{H}} + \|f(\mathscr{P})\|_{\mathcal{H}\to\mathcal{H}_k^1} \leq \sup_{x\in[c_{\mathrm{Ga}}-C_{\mathrm{Ga}},\infty)} |(x + C_{\mathrm{Ga}})^{1/2} f(x)|\,.$$
5. *Show that for $r_z(x) := (x - z)^{-1}$, we have $r_z(\mathscr{P}) = (\mathscr{P}(k) - z)^{-1}$ on $\mathcal{H}$, and deduce from the previous questions that*
$$\|(\mathscr{P}(k) - z)^{-1}\|_{\mathcal{H}\to\mathcal{H}} \leq \frac{1}{|\Im(z)|}.$$
$$\|(\mathscr{P}(k) - z)^{-1}\|_{\mathcal{H}_k^{-1}\to\mathcal{H}_k^1} \leq C\frac{\langle z\rangle}{|\Im(z)|}$$
where C only depends on k_0.

Bounds on the norms of $f(\mathscr{P})$ in higher norms can be obtained using the elliptic regularity assumption. This is the object of the next proposition.

Proposition 2.9 (Mapping properties of $\psi(\mathscr{P})$)

For all real-valued $\psi \in C_c^\infty(\mathbb{R})$, the operator $\psi(\mathscr{P})$ is a bounded self-adjoint operator on $\mathcal{H}$. It maps $\mathcal{H}_k^{-n}$ to $\mathcal{H}_k^n$ for any $n \in \mathbb{N}$, and

$$\forall n \in \mathbb{N}\,,\ \forall k_0 > 0\,,\quad \sup_{k\geq k_0} \|\psi(\mathscr{P})\|_{\mathcal{H}_k^{-n}\to\mathcal{H}_k^n} < \infty.$$

Proof. 1. By Exercise 2.4, for any compactly supported real-valued function ψ, the operator $\psi(\mathscr{P})$ is bounded and self-adjoint on $\mathcal{H}$, and maps $\mathcal{H}$ to $\mathcal{H}^1_k$ boundedly. In particular, $\mathscr{P}\psi(\mathscr{P}) : \mathcal{H} \to (\mathcal{H}^1_k)^*$ is well-defined. For all $j \in \mathbb{N}$,

$$\mathscr{P}\psi(\mathscr{P})u_j = \lambda_j \psi(\lambda_j) u_j$$

and thus by density, $\mathscr{P}\psi(\mathscr{P}) = (x\psi)(\mathscr{P})$. Therefore, we have in fact $\mathscr{P}\psi(\mathscr{P}) : \mathcal{H} \to \mathcal{H}^1_k$ with k-uniform norm (since $x \mapsto x\psi(x)$ is again compactly supported).

2. We claim for all $n \in \mathbb{N}$ and for any $\psi \in C^\infty_c(\mathbb{R})$, $\psi(\mathscr{P})$ maps $\mathcal{H}$ to $\mathcal{H}^n_k$ with

$$\sup_{k \geq k_0} \|\psi(\mathscr{P})\|_{\mathcal{H} \to \mathcal{H}^n_k} < \infty. \tag{2.12}$$

This is true for $n = 0$ and $n = 1$ by what precedes. Now arguing by induction, suppose it is true for some given $n \geq 0$, and let $u \in \mathcal{H}$. Then $\psi(\mathscr{P})u \in \mathcal{H}^1_k$ and $\mathscr{P}\psi(\mathscr{P})u = (x\psi)(\mathscr{P})u \in \mathcal{H}^n_k$, and thus $\psi(\mathscr{P})u \in \mathcal{H}^{n+2}_k$ by elliptic regularity (Assumption 2.1), with

$$\|\psi(\mathscr{P})u\|_{\mathcal{H}^{n+2}_k} \leq C_{\mathrm{ell}}(\|\psi(\mathscr{P})u\|_{\mathcal{H}^1_k} + \|(x\psi)(\mathscr{P})u\|_{\mathcal{H}^n_k}) \leq C\|u\|_{\mathcal{H}}$$

where C does not depend on k.

3. From step 2 and the fact that $\psi(\mathscr{P})$ is self-adjoint, it follows by duality that $\psi(\mathscr{P})$ maps $(\mathcal{H}^n_k)^* \to \mathcal{H}$ with a k-uniformly bounded norm.

4. Finally, let $\widetilde{\psi} \in C^\infty_c(\mathbb{R})$ be such that $\widetilde{\psi} \equiv 1$ on $\operatorname{supp}\psi$, so that $\psi = \psi\widetilde{\psi}$. Then,

$$\|\psi(\mathscr{P})\|_{(\mathcal{H}^n_k)^* \to (\mathcal{H}^n_k)} = \|\psi(\mathscr{P})\widetilde{\psi}(\mathscr{P})\|_{(\mathcal{H}^n_k)^* \to (\mathcal{H}^n_k)} \leq \|\psi^\sharp(\mathscr{P})\|_{\mathcal{H} \to \mathcal{H}^n_k} \|\widetilde{\psi}(\mathscr{P})\|_{(\mathcal{H}^n_k)^* \to \mathcal{H}} \leq C$$

where C is independent of k by steps 2 and 3. The result follows by taking $\psi = \psi^\sharp$. □

Using the previous tools, we can now define the operator $S(k)$ of Proposition 2.5 and establish its properties.

Definition 2.4 (The operators $S(k)$ and $P^\sharp(k)$)

For any $k_0 > 0$, there exists $\psi^\sharp \in C^\infty_c(\mathbb{R}^d)$ such that

$$x + \psi^\sharp(x) \geq \frac{x + C_{\mathrm{Ga}}}{2} \qquad \forall x \in [c_{\mathrm{Ga}} - C_{\mathrm{Ga}}, +\infty).$$

For any $k \geq k_0$, we define $S(k) : \mathcal{H} \to \mathcal{H}$ and $P^\sharp(k) : \mathcal{H}^1_k \to (\mathcal{H}^1_k)^*$ by

$$S(k) := \psi^\sharp(\mathscr{P}), \quad P^\sharp(k) := P(k) + S(k).$$

The mapping properties of $S(k)$ stated in Proposition 2.5 follow immediately from Proposition 2.9.

To obtain those of $P^\sharp(k)$, the idea is that (at least informally for now), the "symbol" of the operator $\mathrm{Re}(P^\sharp(k))$ is $\lambda + \psi^\sharp(\lambda)$. Hence, in view of the norm equivalence of Proposition 2.8 (iii), the definition of $\psi^\sharp$ ensures that it is coercive on $\mathcal{H}^1_k$. Elliptic regularity can then be used (in a similar way as in Exercise 2.1) to lift these mapping properties in higher norms.

Proposition 2.10 (Properties of $P^\sharp(k)$)

For every $k_0 > 0$, there exists $C^\sharp(k_0) > 0$ such that

$$\mathrm{Re}\langle P^\sharp(k)u, u\rangle \geq C^\sharp(k_0)\|u\|^2_{\mathcal{H}^1_k}.$$

The operator $P^\sharp(k) : \mathcal{H}^1_k \to (\mathcal{H}^1_k)^*$ is an isomorphism, and its inverse $R^\sharp(k)$ maps $\mathcal{H}^{n-1}_k$ to $\mathcal{H}^{n+1}_k$ for all $n \in \mathbb{Z}$ with

$$\forall k_0, \quad \sup_{k\geq k_0} \|R^\sharp(k)\|_{\mathcal{H}^{n-1}_k \to \mathcal{H}^{n+1}_k} < \infty.$$

Proof. 1. Let $(u_j)_{j\in\mathbb{N}}$ and $(\lambda_j)_{j\in\mathbb{N}}$ be as in Proposition 2.8. Let

$$u = \sum_{j=0}^{+\infty} \alpha_j u_j.$$

where $(\alpha_j)_{j\in\mathbb{N}}$ has finitely many coefficients. By definition of $\psi^\sharp$,

$$\begin{aligned}
\mathrm{Re}\langle P^\sharp(k)u, u\rangle &= \langle (\mathscr{P} + \psi^\sharp(\mathscr{P}))u, u\rangle \\
&= \sum_{j=0}^{+\infty} (\lambda_j + \psi^\sharp(\lambda_j))|\alpha_j|^2 \\
&\geq \sum_{j=0}^{+\infty} \frac{1}{2}(\lambda_j + C_{\mathrm{Ga}})|\alpha_j|^2 \\
&\geq c\|u\|^2_{\mathcal{H}^1_k}
\end{aligned}$$

for some $c > 0$ independent of k by Proposition 2.8 (iii). The same follows for any $u \in \mathcal{H}^1_k$ by density, showing the first claim.

2. By the Lax-Milgram theorem, $P^\sharp$ is an isomorphism from $(\mathcal{H}^1_k)^*$ to $\mathcal{H}^1_k$. Moreover, for $u \in \mathcal{H}^n_k$, $n \geq 0$, we have $\mathscr{P}(k)R^\sharp(k)u = (P^\sharp(k) - S(k))u = u - S(k)u \in \mathcal{H}^n_k$ so that, by elliptic regularity, $R^\sharp(k)u \in \mathcal{H}^{n+2}_k$ and

$$\|R^\sharp(k)u\|_{\mathcal{H}^{n+2}_k} \leq C_{\mathrm{ell}}(\|u\|_{\mathcal{H}^1_k} + \|u - S(k)\|_{\mathcal{H}^n_k}) \leq C\|u\|_{\mathcal{H}^n_k}$$

by the mapping properties of $S(k) = \psi^\sharp(\mathscr{P})$ (Proposition 2.9).

3. The previous items show the claimed mapping properties for $n \geq 0$, and the ones for $n \leq 0$ follow by duality.

□

At a later stage, we will also need estimates for the resolvent $(\mathscr{P}(k) - z)^{-1}$. We record this result here since its proof involves the ingredients used in this paragraph. The technique of the proof is similar to the one used in Proposition 2.10.

Proposition 2.11 (Mapping properties of $(\mathscr{P}(k)-z)^{-1}$)

For all $z \in \mathbb{C} \setminus \mathbb{R}$ and for all $n \in \mathbb{Z}$, the operator $(\mathscr{P}(k)-z)^{-1}$ maps $\mathcal{H}^n_k$ to $\mathcal{H}^{n+2}_k$ and for every $k_0 > 0$ and $n \in \mathbb{Z}$, there exists $C > 0$ such that the estimate

$$\|(\mathscr{P}(k)-z)^{-1}\|_{\mathcal{H}^{n-1}_k \to \mathcal{H}^{n+1}_k} \leq C \frac{\langle z \rangle^{1+\lfloor n/2 \rfloor}}{|\Im(z)|}$$

holds for all $k \geq k_0$ and $z \in \mathbb{C} \setminus \mathbb{R}$.

Proof. The case $n = 0$ is shown in Exercise 2.4. For $n = 1$, we have by elliptic regularity

$$\begin{aligned}
\|(\mathscr{P}(k)-z)^{-1}u\|_{\mathcal{H}^2_k} &\leq C(\|(\mathscr{P}(k)-z)^{-1}u\|_{\mathcal{H}} + \|\mathscr{P}(k)(\mathscr{P}(k)-z)^{-1}\|_{\mathcal{H}}) \\
&\leq C\Big(\|u\|_{\mathcal{H}} + \langle z \rangle \|(\mathscr{P}(k)-z)^{-1}u\|_{\mathcal{H}}\Big) \\
&\leq C\frac{\langle z \rangle}{|\Im(z)|}\|u\|_{\mathcal{H}}
\end{aligned}$$

by Exercise 2.4. Next, assume that the claimed mapping property holds from $\mathcal{H}^{n-1}_k$ to $\mathcal{H}^{n+1}_k$ for some $n \geq 0$. Then for $u \in \mathcal{H}^{n+1}_k$, by elliptic regularity,

$$\begin{aligned}
\|(\mathscr{P}(k)-z)^{-1}u\|_{\mathcal{H}^{n+3}_k} &\leq C\left(\|(\mathscr{P}(k)-z)^{-1}u\|_{\mathcal{H}^1_k} + \|\mathscr{P}(k)(\mathscr{P}(k)-z)^{-1}u\|_{\mathcal{H}^{n+1}_k}\right) \\
&\leq C\left(\frac{\langle z \rangle}{|\Im(z)|}\|u\|_{\mathcal{H}} + \|u\|_{\mathcal{H}^{n+1}_k} + |z|\|(\mathscr{P}(k)-z)^{-1}u\|_{\mathcal{H}^{n+1}_k}\right) \\
&\leq C\frac{\langle z \rangle^{2+\lfloor \frac{n}{2} \rfloor}}{|\Im(z)|} \\
&= C\frac{\langle z \rangle^{1+\lfloor \frac{n+2}{2} \rfloor}}{|\Im(z)|}.
\end{aligned}$$

By induction, this gives the result when $n \geq 0$, and the case $n \leq 0$ follows by duality. □

Chapter 3

Non-uniform meshes defined by billiard trajectories

The analysis of the previous chapter tacitly assumed that the mesh Ω_h is *quasi-uniform*, in the sense that all elements are of comparable diameter. Indeed, all estimates are formulated in terms of the *global, maximal mesh size* h. However, one may consider meshes Ω_h with varying elements sizes. For instance, if we introduce a cover $\Omega = \Omega_1 \cup \ldots \cup \Omega_N$, we may consider meshes Ω_h with distinct *local mesh sizes* h_j in each region Ω_j (with some transition in the overlaps). The motivation for considering non-uniform meshes can be seen from numerical experiments such as the one displayed in Figure 3.1. A non-uniform mesh which could seem more suited for this computation is shown for example in Figure 3.2. The following question then arises: how to choose the parameters $h_1, \ldots, h_N$ with respect to k ? Can the results of the previous chapter be improved for such meshes?

An affirmative answer to this question was given in [4] and the goal of this chapter is to present it. Roughly speaking, given a covering $\Omega = \Omega_1 \cup \ldots \cup \Omega_N$ into N open sets, the main result of [4] is a bound of the *local Galerkin errors* in terms of the *local best approximation errors* in each region. A corollary of this result is that, for *trapping problems* (problems for which the resolvent $\rho(k)$ grows faster than k) there are meshes which strongly violate the criterion "$(hk)^p\rho(k)$ sufficiently small", but nevertheless achieve k-uniform quasi-optimality. Such meshes can be constructed by taking into account the properties of billiard trajectories in Ω.

3.1 The Helmholtz resolvent and billiard trajectories

In the previous chapter, we have presented the pollution effect: the fact that the (sharp) sufficient condition for k-uniform quasi-optimality is $(hk)^p\rho(k) \lesssim 1$, and not just $hk \lesssim 1$. Clearly, the severity of the pollution is dictated by the growth of $\rho(k)$, and for concrete Helmholtz problems, this intimately depends on the geometry of Ω – or more precisely, on the behavior of billiard trajectories, or *rays*, in Ω.

Let us discuss this relationship more precisely for a Helmholtz problem of the form

$$\left\{ \begin{array}{rcll} -k^{-2}\operatorname{div}(A(x)\nabla u) - n(x)u & = & f & \text{on } \mathbb{R}^d. \\ \dfrac{\partial u}{\partial r} - iku & = & o(r^{-(d-1)/2}) & \text{when } r \to \infty. \end{array} \right. \tag{3.1}$$

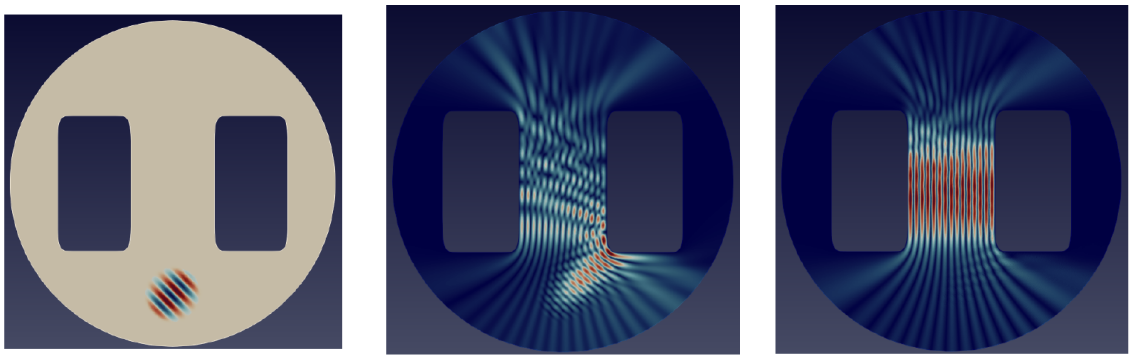


Figure 3.1: Helmholtz problem with an obstacle consisting of two "mirrors". Left: data f (real part). Middle: finite-element solution u_h (real part). Right: error $|u - u_h|$ (computed using a reference solution on a finer mesh with larger p). The scales in each plot are different, but what matters is the relative scale. The error is concentrated between the mirrors, even though the solution is not particularly large there. Would it be useful to refine more the mesh between the mirrors?

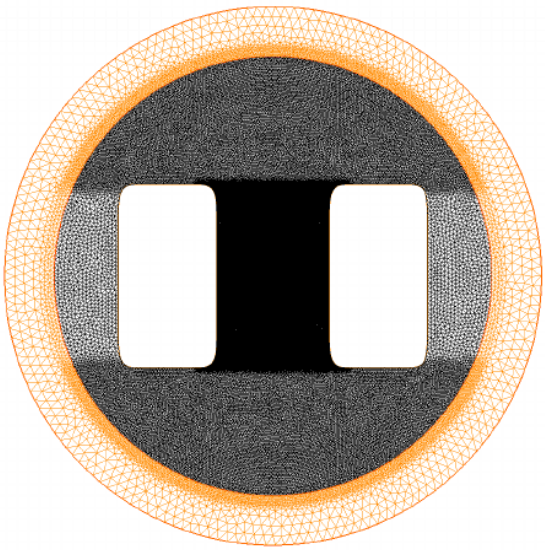


Figure 3.2: A mesh that seemes to be more adapted for computing the solution of the Helmholtz problem shown in Figure 3.1. Can we prove error estimates taking into account the non-uniform mesh size?

where A and n are smooth and bounded functions such that $A \equiv I$ and $n \equiv 1$ outside of a compact set (here, $\Omega = \mathbb{R}^d$ is unbounded, but we will consider the "truncated resolvent" $\chi R(k)\chi$, see below). We assume that $A(x) \geq c\mathrm{Id}$ and that $n(x) \geq 0$. To the Helmholtz operator, we associate the *Hamiltonian*

$$H(q,p) := \langle A(q)p, p\rangle - n(q)$$

(with q the position and p the impulsion) and consider the *billiard trajectories*, $p, q : \mathbb{R} \to \mathbb{R}^d$ defined by the ODE system

$$\dot{q} = \frac{\partial H}{\partial p}(q,p)\,, \quad \dot{p} = -\frac{\partial H}{\partial q}(q,p)\,, \quad (p(0), q(0)) = (p_0, q_0)\,, \tag{3.2}$$

where the dot denotes the time derivative. These equations are uniquely and globally solvable on $\mathbb{R}$ owing to the smoothness and boundedness of A and n. When $A \equiv I$ and $n \equiv 1$, they simply become $\dot{q} = 2p$, $\dot{p} = 0$, and the solution is thus

$$q(t) = q_0 + 2p_0 t\,, \quad p(t) = p_0\,, \quad t \in \mathbb{R},$$

that is, a straight-line "ray" issued from p_0 with constant velocity $2p_0$. For Helmholtz problems involving boundaries, these trajectories are defined as above away from boundaries, and continued via the Snell-Descartes laws when reaching a boundary.[1] A reason why rays appear naturally when $k \to \infty$ is presented in Exercise 3.1 below.

Depending on the geometry, it can occur that a billiard trajectory remains "trapped" in a compact set for all positive and negative times (as illustrated in Figure 3.3 in a case involving a boundary). We call the **cavity**, denoted by $\mathcal{K}$, the set of points in $\mathbb{R}^d$ such that *there is at least one trapped ray passing through* x with $H(p_0, q_0) = 0$. If the cavity is not empty, we say that

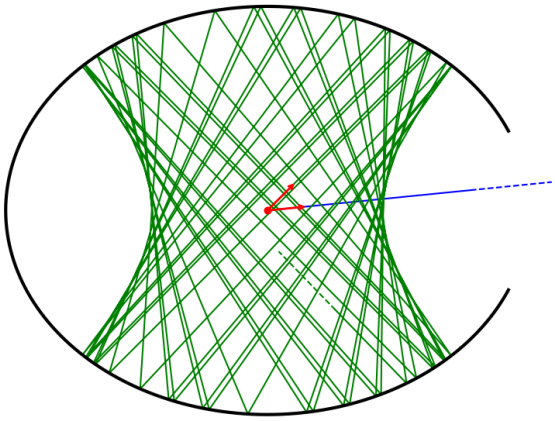


Figure 3.3: Rays in an elliptic cavity (with reflection by Snell-Descartes' law at the boundary). The ray displayed in blue is not trapped forward, whereas the green ray is trapped (in both directions). The red dot is in the cavity since it has at least a trapped trajectory passing over it.

the Helmholtz problem is **trapping**, and **non-trapping** otherwise. Roughly speaking, the growth

[1] To be precise, one must use the more complicated notion of *generalized bicharacteristic flow* to account for "glancing rays" along boundaries, see [27, Section 24.3].

of the resolvent $\rho(k)$ is dictated by "how many" trajectories are trapped, and how "stable" this trapping is under perturbations.

Let us state more precisely some known results in this direction. For a fixed $\chi \in C_c^\infty(\mathbb{R}^d)$, let $\rho_\chi(k) := \|\chi R(k)\chi\|_{L^2\to L^2}$ where $R(k)$ is the solution operator $f \mapsto u$ of the Helmholtz problem (3.1).

1. If the problem is *non-trapping*, then there exists $c, C > 0$ such that
$$k \lesssim \rho_\chi(k) \lesssim k,$$
see, e.g., [40, 9], [19, Theorem 4.43]. The hidden constants are related to the longest time for which a ray remains in the support of χ, see [25].

2. If the problem is *trapping*, there exists $\chi \in C_c^\infty(\mathbb{R}^d)$ and $C > 0$ such that
$$\rho_\chi(k) \gtrsim k\log(k),$$
see [5], and this estimate is sharp (the lower bound is achieved in the presence of boundaries, for instance when the propagation domain is the complement of two convex obstacles).

3. In any case, there exists $C > 0$ such that
$$\rho_\chi(k) \lesssim e^{Ck}$$
[7, 47] and this estimate is sharp (in the presence of boundaries, the upper bound is achieved by elliptic cavities like the one in Figure 3.3). However, the polynomial bound
$$\rho_\chi(k) \lesssim k^{5n/2+2+\varepsilon}$$
holds with $\varepsilon > 0$ arbitrarily small, for all wavenumbers $k \in \mathbb{R}_+ \setminus \mathcal{J}$ for a set $\mathcal{J}$ of arbitrarily small Lebesgue measure [31].

A further link between billiard trajectories and the resolvent operator can be seen by considering $\|\chi_1 R(k)\chi_2\|_{L^2\to L^2}$ where χ_1 and χ_2 are smooth compactly supported functions. This quantity characterizes how large the solution u can be on $\operatorname{supp}\chi_1$ if the data f is supported on $\operatorname{supp}\chi_2$. In general, i.e., if we don't specify any conditions on χ_1 and χ_2, we just have the estimate
$$\|\chi_1 R(k)\chi_2\|_{L^2\to L^2} \lesssim \rho_\chi(k)\,, \tag{3.3}$$
where χ is compactly supported and equal to 1 in a large ball containing the supports of χ_1 and χ_2. Here, we think of $\rho_\chi(k)$ as describing the "strength of the trapping", with $\rho(k) \lesssim k$ for non-trapping problems and $\rho(k) \lesssim e^{Ck}$ for strongly trapping problems. But once again, billiard trajectories provide more informations. In the next statement, we fix a subset $\mathcal{J} \subset \mathbb{R}_+$ such that ρ_χ is polynomially bounded on $\mathbb{R}_+ \setminus \mathcal{J}$.

1. If $\operatorname{supp}\chi_1$ and $\operatorname{supp}\chi_2$ are dynamically isolated (no billiard trajectory starting in $\operatorname{supp}\chi_1$ attains a neighborhood of $\operatorname{supp}\chi_2$) and if in addition, $\operatorname{supp}\chi_1$ is dynamically isolated from the cavity, then
$$\|\chi_1 R(k)\chi_2\|_{L^2\to L^2} = O(k^{-\infty})\,,$$
in the sense that for all N, there is C_N such that the left-hand side is bounded by $C_N k^{-N}$ for all k in $\mathbb{R}_+ \setminus \mathcal{J}$ (this follows from propagation of singularities, see [19, Theorem E.47], see also [23, Theorem 5.10] for a statement in a simpler setting, and the fact that for all $f \in \mathcal{H}$, $R(k)\chi_2 f$ is "outgoing", i.e., super-algebraically small away from outgoing directions in phase-space, see [23, Lemma 5.18]. See also [4, Theorems C.3 and C.4]).

2. If the supports of χ_1 **and** χ_2 do not intersect the cavity, then in fact

$$\|\chi_1 R(k)\chi_2\|_{L^2\to L^2} \lesssim k,$$

for all $k \in \mathbb{R}_+ \setminus \mathcal{J}$, see [8, 11]. In other words, if both supports are away from the cavity, one obtains the non-trapping bound, *as if there were no cavity.*

3. If the support of χ_1 **or** χ_2 do not intersect the cavity, then

$$\|\chi_1 R(k)\chi_2\|_{L^2\to L^2} \lesssim \sqrt{k\rho_\chi(k)}, \tag{3.4}$$

for all $k \in \mathbb{R}_+ \setminus \mathcal{J}$, see [16]. Thus, we get an estimate "in-between" the non-trapping estimate $O(k)$, and the "worst possible" growth $\rho_\chi(k)$ in (3.3).

For Helmholtz problems with boundaries and/or truncated by PML boundaries, some of these results can be extended. For general statements, including the case of domains with boundaries, we refer to [19]. Corresponding statements can also be shown for Helmholtz problems truncated by a PML, see [4, §4, Appendix C]. In the PML region, essentially nothing propagates (rays attaining this region can be thought as being "absorbed", or escaping to infinity). Let us state these results in the case of the model Helmholtz problem of §1.1. We define the following *dynamical regions*, where trajectories are defined with respect to the problem (1.2) (i.e., before PML truncation)

Definition 3.1 (Dynamical regions)

1. The *cavity*, $\mathcal{K} \subset \overline{\Omega}_+$, is the set of points $x \in \overline{\Omega}_+$ for which there is at least one billiard trajectory passing over x, and which remains in a compact set for all positive **and** negative times.

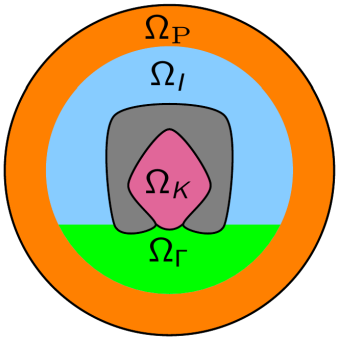
2. The *visible set* (from the cavity), $\mathcal{V} \subset \overline{\Omega}_+$, is the set of points $x \in \overline{\Omega}_+$ for which there is at least one billiard trajectory passing over x and remaining in a compact set for all positive **or** negative times. Observe that

$$\mathcal{K} \subset \mathcal{V}.$$

3. The *invisible set* (from the cavity) $\mathcal{I} \subset \overline{\Omega}_+$ is the set of points $x \in \overline{\Omega}_+$ for which all billiard trajectories passing over x escape any given compact set in a finite time, both in the future and the past. Thus,

$$\mathcal{I} = \overline{\Omega}_+ \setminus \mathcal{V}.$$

We apply PML truncation on a bounded computational domain Ω and let $\Omega_{\mathcal{P}}$ be an open neighbourhood of the PML truncation boundary which is strictly contained in the PML region. We then use the notation from Chapters 1 and 2. Let $\Omega_{\mathcal{K}}, \Omega_{\mathcal{V}}, \Omega_{\mathcal{V}}$ be open neighbourhoods of, respectively, $\mathcal{K}, \mathcal{V} \setminus (\mathcal{K} \cup \Omega_{\mathcal{P}})$ and $\mathcal{I} \setminus \Omega_{\mathcal{P}}$ in Ω, see Figure 3.4. Estimates on $\chi_1 R(k)\chi_1$, for k in a set where $\rho(k)$ is polynomially bounded, depending on the locations are then gathered in the Table in Figure 3.4.



$\operatorname{supp}\chi_1 \backslash \operatorname{supp}\chi_2$	$\Omega_\mathcal{K}$	$\Omega_\mathcal{V}$	$\Omega_\mathcal{I}$	$\Omega_\mathcal{P}$
$\Omega_\mathcal{K}$	$\rho(k)$	$\sqrt{k\rho(k)}$	$O(k^{-\infty})$	$O(k^{-\infty})$
$\Omega_\mathcal{V}$	$\sqrt{k\rho(k)}$	k	k	1
$\Omega_\mathcal{I}$	$O(k^{-\infty})$	k	k	1
$\Omega_\mathcal{P}$	$O(k^{-\infty})$	1	1	1

Figure 3.4: **Left:** a computational domain with a PML truncation boundary (outer circle) and an obstacle (in gray). The neighbourhoods $\Omega_\mathcal{K}$, $\Omega_\mathcal{V}$, $\Omega_\mathcal{I}$ associated to the dynamical regions of Definition 3.1 are represented in red, green and blue, respectively. The region represented in orange must lie strictly inside the PML region. Any ray attaining the PML can thought as being "absorbed". This figure was created by Jeffrey Galkowski. **Right:** bounds on $\|\chi_1 R(k)\chi_2\|_{\mathcal{H}\to\mathcal{H}}$ up to k-uniformly bounded constants, for all $k \notin \mathcal{J}$, with ρ polynomially bounded on $\mathbb{R}_+ \setminus \mathcal{J}$, when $\operatorname{supp}\chi_1, \operatorname{supp}\chi_2$ are subsets of $\Omega_\star$, $\star \in \{\mathcal{K}, \mathcal{V}, \mathcal{I}, \mathcal{P}\}$.

Exercise 3.1. (Rays and eikonal equation).
Look for approximate solutions of (3.1) *with* $f = 0$ *in the form*

$$u(x) = e^{ikS(x)}$$

where $S(x) = S_0(x) + k^{-1}S_1(x) + \ldots$, *and check that, at leading order,* S_0 *must obey the* eikonal equation *(also known as* Hamilton-Jacobi equation*):*

$$H(x, \nabla S_0(x)) = 0 \qquad \forall x \in \mathbb{R}^d.$$

Suppose that $q(t)$ *solves the ODE*

$$\dot{q} = \frac{\partial H}{\partial p}(q, \nabla S_0(q)), \quad q(0) = q_0$$

and let $p(t) = \nabla S_0(q(t))$. *Show that* $t \mapsto (q(t), p(t))$ *solves the Hamilton equations* (3.2).

Remark 3.1 (The Schrödinger equation). Many of the results stated above are also valid for (and in fact, often stated in terms of) Schrödinger problems, where the operator is $P(h) = -h^2\Delta + V(x)$ and $V(x) \geq 0$ is a potential with sufficient decay at infinity. The results then concern the growth of the resolvent $(P(h) - E)^{-1}$ where E is an "energy level". In this case, the Hamilton equations correspond to the motion of a classical particle (i.e., following Newtonian mechanics) under a potential V.

3.2 Propagation of errors in the finite-element method

Let us pretend for a moment that Galerkin are *locally quasi-optimal*, that is, for any $k_0 > 0$ and any subsets $U \Subset \widetilde{U} \subset \Omega$, there is a constant $C > 0$ independent of k such that

$$\|u - u_h\|_{\mathcal{H}^1_k(U)} \leq C \inf_{v_h \in V^p(\Omega_h)} \|u - w_h\|_{\mathcal{H}^1_k(\widetilde{U})} \tag{3.5}$$

for all $k \geq k_0$, $u \in \mathcal{H}^1_k$, Ω_h a sufficiently fine mesh and $u_h \in V^p(\Omega_h)$ a Galerkin approximation of u. Here, the local $\mathcal{H}^1_k$ norm in a subset $U \subset \Omega$, $\|\cdot\|_{\mathcal{H}^1_k(U)}$, is defined by

$$\|u\|_{\mathcal{H}^1_k(U)} := \inf \left\{ \|\tilde{u}\|_{\mathcal{H}^1_k} \;\middle|\; \tilde{u} \in \mathcal{H}^1_k \text{ coincides with } u \text{ on } U \right\}.$$

Consider a cover

$$\Omega = \Omega_1 \cup \ldots \cup \Omega_N \,,$$

and for every $j = 1, ..N$, let $\Omega'_j \Subset \Omega_j$.[2] Then $\inf_{v_h \in V_h} \|u - w_h\|_{\mathcal{H}^1_k(\Omega_j)} \lesssim C(h_j k)^p \|u\|_{\mathcal{H}^{p+1}_k}$, and thus, since $u = R(k)f$, one would have by (3.5)

$$\begin{pmatrix} \|u - u_h\|_{\mathcal{H}^1_k(\Omega'_1)} \\ \vdots \\ \|u - u_h\|_{\mathcal{H}^1_k(\Omega'_N)} \end{pmatrix} \lesssim \begin{pmatrix} (h_1 k)^p \|\chi_1 R(k)\|_{\mathcal{H}^{p-1}_k \to \mathcal{H}^{p+1}_k} \\ \vdots \\ (h_N k)^p \|\chi_N R(k)\|_{\mathcal{H}^{p-1}_k \to \mathcal{H}^{p+1}_k} \end{pmatrix} \|f\|_{\mathcal{H}^{p-1}_k}$$

where $\chi_j \in C^\infty_c(\Omega)$ are such that $\chi_j \equiv 1$ on Ω_j. In such a situation, the best choice of local mesh-sizes h_j would simply be dictated by the rate of growth of $\|\chi_j R(k)\|$.

However, local quasi-optimality does not hold in general: in many cases, the Galerkin error in a given region is influenced by the mesh-size in some other region. To illustrate this, we use the following toy numerical experiment from [3]. We consider the Helmholtz "impedance" problem

$$-k^{-2}\Delta u - u = 0 \quad \text{in } \Omega \quad \text{and} \quad k^{-1}\partial_n u - iu = g \quad \text{on } \partial\Omega, \tag{3.6}$$

in a rectangular domain Ω, with data g chosen so that the exact solution is the plane-wave e^{ikx_1}. We solve this problem using three meshes Ω^1_h, Ω^2_h and Ω^3_h. The meshes $\Omega^{1/2}_h$ are quasi-uniform, with mesh sizes $h_2 \ll h_1$ (i.e., Ω^2_h is more refined than Ω^1_h). On the other hand, Ω^3_h is a mesh with sizes h_1 in the left-hand half and h_2 in the right-hand half (with some transition region in between) as in Figure 3.5 below.

The Galerkin errors in each case are plotted in Figure 3.6. As expected, the numerical approximation by the finite-element method with the mesh Ω^2_h is much more accurate than with the mesh Ω^1_h (top right and top left panels). But with the mesh Ω^3_h, the error in the right-hand part of the mesh is not the local best approximation error (the error is about 10^3 times larger than on the mesh Ω^2_h with the same local mesh size); rather, it appears to be dominated by an errors propagating from the left region.

The results for the same experiment, but choosing this time the data g in (3.6) such that the solution is e^{ikx_2}, are shown in Figure 3.7. They show that, even if the solution does not involve propagation from the left to the right, the error can still have this property.

3.3 Sketch of a localized argument

We now consider a Helmholtz problem satisfying the assumptions of Chapters 1 and 2, and introduce a cover

$$\Omega = \Omega_1 \cup \ldots \cup \Omega_N$$

[2] that is, Ω_j is contained in a compact subset of Ω_j.

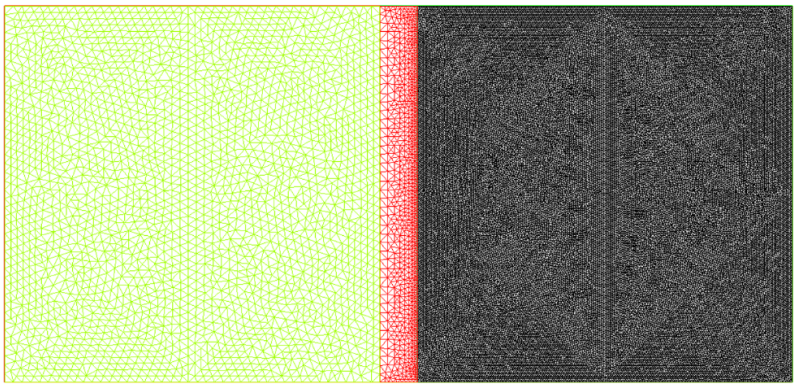


Figure 3.5: The mesh Ω_h^3.

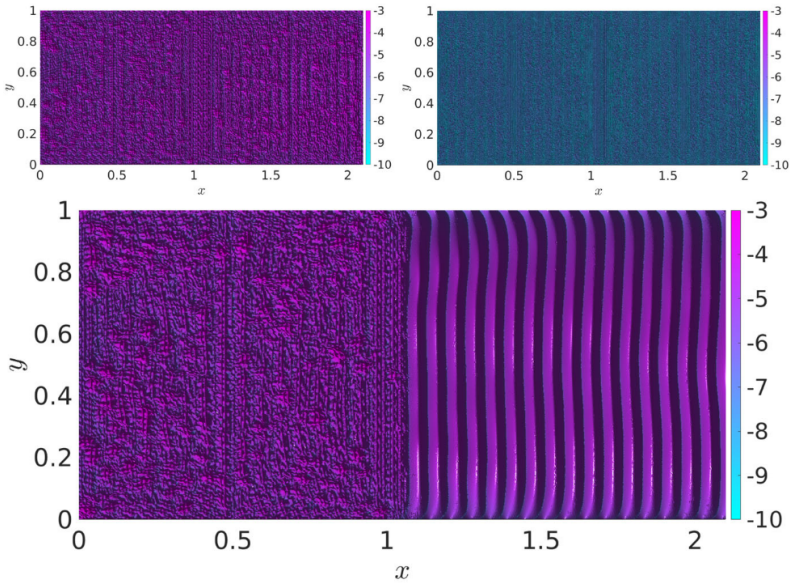


Figure 3.6: Plot of the quantity $\log_{10}(10^{-12} + |\mathrm{Re}(u - u_h)|)$ for the finite-element approximation of (3.6), on the meshes Ω_h^1 (quasi-uniform mesh with size h_1, top left panel), Ω_h^2 (quasi-uniform mesh with size $h_2 \ll h_1$, top right panel), and Ω_h^3 (non-uniform mesh with size h_1 on the left half and h_2 on the right half, bottom panel).

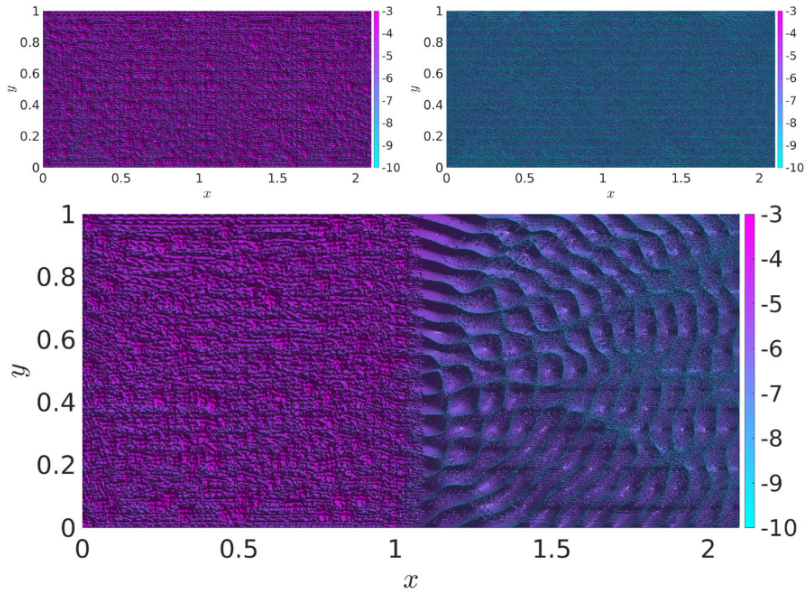


Figure 3.7: Same as Figure 3.6, but with g in (3.6) chosen such that the solution is e^{ikx_2}.

by open subsets $\Omega_j \subset \Omega$. Let $\phi_1, \dots, \phi_N$ be a smooth partition of unity (p. o. u.) subordinate to this covering, i.e.,

$$\sum_{j=1}^{N} \phi_j = 1, \quad \text{and} \quad \phi_j \in C^\infty(\overline{\Omega}), \ \operatorname{supp} \phi_j \subset \Omega_j \cup \partial\Omega, \quad 1 \le j \le N.$$

Let $\chi_j \in C^\infty(\overline{\Omega})$, $j = 1, \dots, N$ such that

$$\operatorname{supp} \chi_j \subset \Omega_j \cup \partial\Omega \quad \text{and} \quad \phi_j \prec \chi_j$$

where the notation "$\prec$" is defined by

$$u \prec \tilde{u} \iff \operatorname{dist}(\operatorname{supp} u, \operatorname{supp}(1 - \tilde{u})) > 0\,,$$

that is, $\tilde{u}$ is equal to 1 on a neighbourhood of $\operatorname{supp} u$ (with some positive margin).

Let $k > 0$, $u \in \mathcal{H}^1_k$, $V_h \subset \mathcal{H}^1_k$ and suppose that u_h is a Galerkin approximation of u in V_h. Our aim is to estimate $\chi_j(u - u_h)$. We start by doing this in a negative norm, with the idea to use the coercivity of $P(k)$ up to $S(k)$ to convert this estimate into an estimate in the $\mathcal{H}^1_k$ norm (as in the proof of Theorem 2.7). Thus, given $i \in \{1, \dots, N\}$ and a test function $\xi \in \mathcal{H}^{p-1}_k$, let us compute

$$\begin{aligned}
\langle \chi_i(u - u_h), \xi \rangle &= \langle u - u_h, \chi_i \xi \rangle \\
&= \langle P(k)(u - u_h), R(k)^* \chi_i \xi \rangle && \text{(duality)} \\
&= \langle P(k)(u - u_h), (I - \Pi_h^\sharp) R(k)^* \chi_i \xi \rangle && \text{(Galerkin orth.)}
\end{aligned}$$

$$
\begin{aligned}
&= \big\langle P^\sharp(k)(u-u_h), (I-\Pi_h^\sharp)R(k)^*\chi_i\xi\big\rangle \\
&\quad - \big\langle S(k)(u-u_h), (I-\Pi_h^\sharp)R(k)^*\chi_i\xi\big\rangle && \text{(Defs. of } P^\sharp(k),\ S(k)\text{)} \\
&= \sum_{j=1}^N \big\langle P^\sharp(k)(u-u_h), (I-\Pi_h^\sharp)\phi_j R(k)^*\chi_i\xi\big\rangle \\
&\quad - \sum_{j=1}^N \big\langle S(k)(u-u_h), (I-\Pi_h^\sharp)\phi_j R(k)^*\chi_i\xi\big\rangle && (\{\phi_i\} \text{ p. o. u}) \\
&= \sum_{j=1}^N \big\langle P^\sharp(k)(u-v_{h,j}), (I-\Pi_h^\sharp)\phi_j R(k)^*\chi_i\xi\big\rangle \\
&\quad - \sum_{j=1}^N \big\langle S(k)(u-u_h), (I-\Pi_h^\sharp)\phi_j R(k)^*\chi_i\xi\big\rangle && \text{(def. of } \Pi_h^\sharp\text{)}
\end{aligned}
$$

where $v_{h,1},\dots,v_{h,N}$ are arbitrary elements of V_h. To continue this sketch, we now assume that

$$\text{the operators } S(k),\ P^\sharp(k),\ \Pi_h^\sharp \text{ and } R^\sharp(k) \text{ are } \textit{local} \tag{3.7}$$

Here, we say that an operator $A:\mathcal{H}\to\mathcal{H}$ is local if

$$\chi_1 \perp \chi_2 \implies \chi_1 A \chi_2 = 0\,, \tag{3.8}$$

where $\chi_1\perp\chi_2$ means that the supports of χ_1 and χ_2 are at a positive distance from each other. The assumption (3.7) is usually not satisfied, but under proper assumptions, these operators are in fact "pseudolocal", that is, roughly speaking, they satisfy the property (3.8) up to $O(k^{-\infty})$ remainders that can be successfully controlled in the later stages of the proof. We will give an idea of the proof of this pseudo-locality for the operators $S(k)$ (that of $P^\sharp(k)=P(k)+S(k)$ follows because $P(k)$ will be local by assumption) and $R^\sharp(k)$ in Chapter 4 (see Theorem 4.2); for the pseudo-locality of $\Pi_h^\sharp$, we refer to [4, Section 7].

Observe that

$$\phi \prec \chi \iff \phi \perp (1-\chi).$$

Thus, since adjoints/products of local operators are again local, we deduce that

$$P^\sharp(k)^*(I-\Pi_h^\sharp)\phi_j = \chi_j P^\sharp(k)^*\chi_j(I-\Pi_h^\sharp)\phi_j \quad\text{and}\quad S(k)^*(I-\Pi_h^\sharp)\phi_j = \chi_j S(k)^*\chi_j(I-\Pi_h^\sharp)\phi_j$$

that is, roughly speaking, we can "move ϕ_j across" these operators. Therefore,

$$
\begin{aligned}
\langle \chi_i(u-u_h),\xi\rangle = &\sum_{j=1}^N \big\langle P^\sharp(k)\chi_j(u-v_{h,j}), \chi_j(I-\Pi_h^\sharp)\phi_j R(k)^*\chi_i\xi\big\rangle \\
&- \sum_{j=1}^N \big\langle S(k)\chi_j(u-u_h), \chi_j(I-\Pi_h^\sharp)\phi_j R(k)^*\chi_i\xi\big\rangle.
\end{aligned}
$$

We now estimate the right-hand side by using the mapping properties of $P^\sharp(k)$ and $S(k)$ (Propositions 2.5 and 2.10):

$$\|\chi_i(u-u_h)\|_{(\mathcal{H}_k^{p-1})^*} \le C\sum_{j=1}^N \Big(\|\chi_j(u-v_{h,j})\|_{\mathcal{H}_k^1} \cdot \|\chi_j(I-\Pi_h^\sharp)\chi_j R_{ij}(k)^*\xi\|_{\mathcal{H}_k^1}$$

$$+ \|\chi_j(u-u_h)\|_{(\mathcal{H}_k^{p-1})^*} \cdot \|\chi_j(I-\Pi_h^\sharp)\chi_j R_{ij}(k)^*\xi\|_{(\mathcal{H}_k^{p-1})^*}\Big)$$

where

$$R_{ij}(k) := \phi_i R(k)\phi_j \tag{3.9}$$

is the *localized resolvent from j to i*. One can see that the *local adjoint approximability operators*,

$$\eta_{ij} := \chi_j(I-\Pi_h^\sharp)\chi_j \cdot R_{ij}(k)^* \qquad 1 \le i,j \le N, \tag{3.10}$$

will play a role analogous to the adjoint approximability constant η of Definition 1.5. In the definition of η_{ij}, we see the interaction between

(i) the operators $R_{ij}(k)$, which reflect the **strength of propagation between subdomains**, and

(ii) the operators $\chi_j(I-\Pi_h^\sharp)\chi_j$, which reflect the **local approximation power** of V_h on Ω_j

(Indeed, concerning the second point, recall from the proof of Theorem 2.7 and Exercise 2.2 that, due to the k-uniform coercivity of $P_k^\sharp$, the operator $\Pi_h^\sharp$ essentially computes the best approximation in V_h).

Let us define two matrices $\mathbf{B}, \mathbf{W} \in \mathbb{R}^{N\times N}$ by

$$\mathbf{B}_{ij} := C\|\eta_{ij}\|_{\mathcal{H}_k^{p-1}\to\mathcal{H}_k^1}\,, \quad \mathbf{W}_{ij} := C\|\eta_{ij}\|_{\mathcal{H}_k^{p-1}\to(\mathcal{H}_k^{p-1})^*}\,. \tag{3.11}$$

With adaptations of the arguments of Chapter 2, and under standard local assumptions for the finite-element scheme, one will obtain that, up to $O(k^{-\infty}(hk)^p)$ terms,

$$\mathbf{B}_{ij} \lesssim (1+\rho_{ij}(k))(h_j k)^p\,, \quad \mathbf{W}_{ij} \lesssim (1+\rho_{ij}(k))(h_j k)^{2p},$$

where $\rho_{ij}(k) := \|R_{ij}(k)\|_{L^2\to L^2}$ is the norm of the resolvent from j to i.

Let $\|\underline{u-u_h}\|_{(\mathcal{H}_k^{p-1})^*}$ be the column vector of local Galerkin errors $\|\chi_i(u-u_h)\|_{(\mathcal{H}_k^{p-1})^*}$, and similarly $\|\underline{u-v_h}\|_{\mathcal{H}_k^1}$ the vector of local best approximation errors, whose components are $\|\chi_i(u-v_{h,j})\|_{\mathcal{H}_k^1}$. We then arrive arrive at the matrix system of inequalities

$$\|\underline{u-u_h}\|_{(\mathcal{H}_k^{p-1})^*} \le \mathbf{B}\|\underline{u-v_h}\|_{\mathcal{H}_k^1} + \mathbf{W}\|\underline{u-u_h}\|_{(\mathcal{H}_k^{p-1})^*}\,, \tag{3.12}$$

which is a localized version of the estimate that we have shown in the case of a uniform mesh (2.10). In (3.12), "$\le$" must be understood in the component-wise sense. If the condition

$$\sum_{n=0}^{\infty} \mathbf{W}^n < \infty \tag{3.13}$$

holds, then $(I-\mathbf{W})^{-1}$ exists and has positive coefficients, so we obtain

$$\|\underline{u-u_h}\|_{(\mathcal{H}_k^{p-1})^*} \le (I-\mathbf{W})^{-1}\mathbf{B}\|\underline{u-v_h}\|_{\mathcal{H}_k^1}\,, \tag{3.14}$$

i.e., a bound on the local Galerkin errors (in a negative norm) by the local best approximation errors.

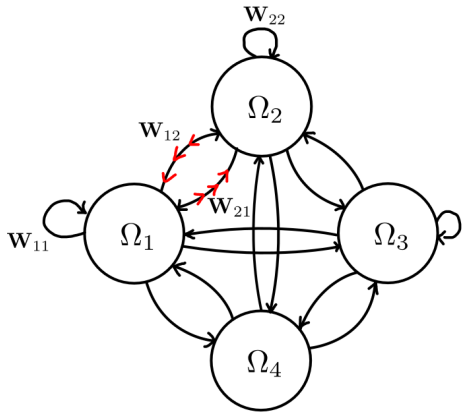


Figure 3.8: Graph representing the propagation of errors between subdomains. The black arrows indicate the weight carried by the edge from i to j, and the red arrows emphasize that this weight corresponds to errors flowing in the opposite directions (from j to i).

To go further, we now give a sufficient condition for (3.13) to hold. For this, we view $\mathbf{W}$ as the adjacency matrix of a directed graph, with N nodes representing $\Omega_1, \dots, \Omega_N$, and with the edge from node i to j carrying the weight $\mathbf{W}_{ij}$. This graph can be thought as a representation of the error propagation between subdomains, with the weight $\mathbf{W}_{ij}$ indicating the amount of error travelling from Ω_j to Ω_i, in view of (3.12). Take note that the error on the edge $i \to j$ thus flows from j to i, see Figure 3.8.

Recall that $(\mathbf{W}^n)_{ij}$ is equal to the sum of the weights of all paths[3] of length n from node i to j. Therefore, the sum converges if for any i and j, the sum of weights of *all* paths from i to j is finite. Clearly, this is not possible if there is a *loop* in the graph with weight > 1 (by taking this loop a large number of times, one can construct a path from i to j with an arbitrarily large weight). Conversely, if the sum c_∞ of the weights of all *simple loops*[4] is < 1, then the sum is finite and in this case, one has the bound

$$(I - \mathbf{W})^{-1} \leq \frac{1}{1 - c_\infty} \mathscr{T}\,, \quad \mathscr{T}_{ij} := \text{sum of weights of all } \textit{direct path} \text{ from } i \text{ to } j,^{5}$$

see Exercise 3.2, [4, Appendix B]. The loop condition can be interpreted intuitively as follows: if there is a loop with weight > 1, the numerical error can be amplified infinitely by propagating along this loop an arbitrary amount of times.

Thus, under the condition that $c_\infty < 1$, and with the assumption (3.7), we obtain the estimate

$$\|\underline{u - u_h}\|_{(\mathcal{H}_k^{p-1})^*} \leq \frac{1}{1 - c_\infty} \mathscr{T}\mathbf{B} \|\underline{u - v_h}\|_{\mathcal{H}_k^1}\,.$$

[3] A path from i to j is a finite sequence of edges of the form $(i \to i_2)(i_2 \to i_3)\dots(i_n \to j)$. Its length is the number of its edges, and its weight is the product $\mathbf{W}_{ii_2} \dots \mathbf{W}_{i_n j}$ of the weights of its edges. The empty path is considered a path from i to i for each i.

[4] A simple loop is a path from a node to itself, which visits every node other than the origin/end exactly once.

[5] A direct path is one that does not visit twice the same node, with the convention that the empty path is a direct path (in other words, $\mathscr{T}$ has ones on its diagonal).

To translate this result in terms of the mesh sizes h_j, the main remaining work is then to estimate the coefficients of $\mathbf{W}$ and $\mathbf{B}$, i.e., the norms of the operators η_{ij} from (3.10).

Exercise 3.2. (Loops in weighted graphs).
Let $N \in \mathbb{N}$ and let $\mathcal{G}$ be a directed graph with nodes $1, \dots, N$, with a weight $\mathbf{W}_{ij} \geq 0$ on the edge $(i \to j)$. For $i, j \in \{1, \dots, N\}$, denote by $\mathbb{P}_{ij}$ the set of all paths from i to j $\mathbb{D}_{ij} \subset \mathbb{P}_{ij}$ the set of direct paths from i to j (with $\mathbb{D}_{ii} = \{\mathbf{0}\}$ by convention, with $\mathbf{0}$ standing for the empty path), and let $\mathbb{SL}$ be the set of simple loops. Given a path $p = (i_1 \to i_2)(i_2 \to i_3) \dots (i_n \to i_{n+1})$, let

$$|p| := n\,, \quad w(p) := \mathbf{W}_{i_1 i_2} \dots \mathbf{W}_{i_n i_{n+1}}\,,$$

the length and the weight of the path, respectively.

1. *Show that*
$$(\mathbf{W}^n)_{ij} = \sum_{p \in \mathbb{P}_{ij}\,,\ |p|=n} w(p).$$

2. *Show that there exists an* injective *map*
$$\mathcal{D}ec : \mathbb{P}_{ij} \to \mathbb{D}_{ij} \times (\mathbb{SL}^{(\mathbb{N})})$$
(where $A^{(\mathbb{N})}$ stands for the set of finite sequences of elements of A) which is weight-preserving, *that is, for all $p \in \mathbb{P}_{ij}$*
$$w(p) = w[\mathcal{D}ec(p)]$$
with, for any $(q, (\ell_1, \dots, \ell_n)) \in \mathbb{D} \times (\mathbb{SL}^{(\mathbb{N})})$,
$$w\big[(q, (\ell_1, \ell_2, \dots, \ell_n))\big] = w(q) w(\ell_1) \dots w(\ell_n).$$
(Hint: construct $\mathcal{D}ec$ by "removing loops one by one" until only a direct path is left).

3. *Deduce that if*
$$c_\infty := \sum_{\ell \in \mathbb{SL}} w(\ell) < 1\,,$$
then $I - \mathbf{W}$ is invertible, that its inverse has non-negative coefficients, and
$$\sum_{p \in \mathbb{D}_{ij}} w(p) \leq \big((I - \mathbf{W})^{-1}\big)_{ij} \leq \frac{1}{1 - c_\infty} \sum_{p \in \mathbb{D}_{ij}} w(p).$$

3.4 Local error estimate

We now state the main result of [4] in the setting of the model Helmholtz problem of §1.1, with then $P(k)$, $R(k)$ and $\rho(k)$ defined as in Definitions 1.2 and 1.3.

We let the space $\mathcal{H}^1_k$ be defined as in Remark 1.2. Recall the dynamical regions $\mathcal{K}, \mathcal{V}, \mathcal{I}$ from Definition 3.1. Recall the neighbourhoods Ω_i, $i = 1, ..., 4$, with $i = 1, 2, 3, 4$ corresponding to $\mathcal{K}, \mathcal{V}, \mathcal{I}, \mathcal{P}$, respectively (we also write $\Omega_\mathcal{K}$ instead of Ω_1, and so on).

We denote by V^p be the Lagrange finite-element scheme of order p. For a mesh Ω_h, we introduce the following measure of "local uniformity at scale ε"

$$\nu(\Omega_h,\varepsilon) := \sup_{x\in\Omega} \sup_{\substack{K,K'\in\Omega_h \\ K\cap B(x,\varepsilon)\neq\emptyset \\ K'\cap B(x,\varepsilon)\neq\emptyset}} \frac{h_K}{h_{K'}}\,.$$

Let

$$h_\star := \Big(\sup_{K\in\Omega_h\,,\ K\cap\Omega_\star\neq\emptyset} h_K \Big)\,, \qquad \star\in\{\mathcal{K},\mathcal{V},\mathcal{I},\mathcal{P}\}$$

(we also write $h_1 = h_{\mathcal{K}}$ and so on) and $h := \max_{K\in\Omega_h} h_K$. Let $\mathbf{W}$ and $\mathbf{B}$ and $\mathcal{T}$ be the 4×4 matrices defined by

$$\mathbf{B}_{ij} := (1+\rho_{ij}(k))(h_j k)^p\,, \quad \mathbf{W}_{ij} := (1+\rho_{ij}(k))(h_j k)^{2p} \quad \text{with} \quad \rho_{ij}(k) := \|\mathbf{1}_{\Omega_i} R(k) \mathbf{1}_{\Omega_j}\|_{L^2\to L^2},$$

$$\mathscr{T}_{ij} := \sum_{p\in\mathbb{D}_{ij}} \mathbf{W}_{ij}\,.$$

Observe that the quantities ρ_{ij} are bounded via the table in the right panel of Figure 3.4, allowing to effectively compute bounds on $\mathbf{B}$, $\mathbf{W}$ and (thus) $\mathscr{T}$ in terms of the local mesh sizes and k.

Theorem 3.1 (Local error estimate for the model Helmholtz problem)

Given $k_0>0$, $N>0$, $\beta>0$, $\mathcal{J}\subset\mathbb{R}_+$ such that ρ is polynomially bounded on $\mathbb{R}_+\setminus\mathcal{J}$, $\Omega'_\star \Subset \Omega_\star$ for each $\star\in\{\mathcal{K},\mathcal{V},\mathcal{I},\mathcal{P}\}$, there exists $\varepsilon>0$ and $C>0$ such that the following holds.
For all $k\in(k_0,+\infty)\setminus\mathcal{J}$ and for any mesh Ω_h of Ω satisfying $\gamma(\Omega_h)+\nu(\Omega_h,k^{-1})\le\beta$ and

$$(h_{\mathcal{K}}k)^{2p}\rho(k) + (h_{\mathcal{V}}k)^{2p}k + (h_{\mathcal{I}}k)^{2p}k + (h_{\mathcal{P}}k)^{2p} \le \varepsilon\,, \tag{3.15}$$

every $u\in\mathcal{H}^1_k$ admits a unique Galerkin approximation $u_h\in V^p(\Omega_h)$ and the estimate

$$\begin{pmatrix} \|u-u_h\|_{\mathcal{H}^1_k(\Omega'_{\mathcal{K}})} \\ \|u-u_h\|_{\mathcal{H}^1_k(\Omega'_{\mathcal{V}})} \\ \|u-u_h\|_{\mathcal{H}^1_k(\Omega'_{\mathcal{I}})} \\ \|u-u_h\|_{\mathcal{H}^1_k(\Omega'_{\mathcal{P}})} \end{pmatrix} \le C\,(I+\mathscr{T}\mathbf{B}) \begin{pmatrix} \|u-v_{h,\mathcal{K}}\|_{\mathcal{H}^1_k(\Omega_{\mathcal{K}})} \\ \|u-v_{h,\mathcal{V}}\|_{\mathcal{H}^1_k(\Omega_{\mathcal{V}})} \\ \|u-v_{h,\mathcal{I}}\|_{\mathcal{H}^1_k(\Omega_{\mathcal{I}})} \\ \|u-v_{h,\mathcal{P}}\|_{\mathcal{H}^1_k(\Omega_{\mathcal{P}})} \end{pmatrix} + Ck^{-N}\|u-v_h\|_{\mathcal{H}^1_k} \tag{3.16}$$

holds for any $v_h\in V^p(\Omega_h)$ and $v_{h,\star}\in V^p(\Omega_h)$ with $\star\in\{\mathcal{K},\mathcal{V},\mathcal{I},\mathcal{P}\}$.

Remark 3.2 (Comments on Theorem 3.1).

1. The propagation graph corresponding to the matrix $\mathbf{W}$ is displayed in Figure 3.9. The condition that (3.15) holds with $\varepsilon>0$ small enough is necessary and sufficient to ensure $c_\infty<1$ (with c_∞ the sum of weights of all simple loops) for all $k\in(k_0,\infty)\setminus\mathcal{J}$. Indeed,

without this condition, one of the loops $(i \to i)$ will be > 1 for k large enough. On the other hand, if the condition holds, the only edge carrying a weight $\gtrsim 1$ is $(\mathcal{K} \to \mathcal{V})$ (with the weight $(h_{\mathcal{V}}k)^{2p}\sqrt{k\rho}$). If a simple loop ℓ contains this edge, it must also contain either the edge $(\mathcal{V} \to \mathcal{K})$, or an edge with a $O(k^{-\infty})$, hence smaller, weight. Thus, the weight of ℓ is bounded by a k-independent multiple of

$$(h_{\mathcal{V}}k)^{2p}\sqrt{k\rho(k)} \cdot (h_{\mathcal{K}}k)^{2p}\sqrt{k\rho(k)} = (h_{\mathcal{V}}k)^{2p}k \cdot (h_{\mathcal{K}}k)^{2p}\rho(k) \leq \varepsilon^2.$$

2. The numerical experiments in [4] suggest that the bound in Theorem 3.1 is sharp.

3. The main result in [4] is stronger than the one stated here for several reasons, including the following: (i) it holds for an arbitrary number of subdomains, (ii) it weakens the restrictions over the mesh-size in the PML region (one only needs a constant number of dofs per wavelength in the PML), (iii) it gives bounds on the high- and low-frequencies of the Galerkin error and (iv) it gives bounds not only on the $\mathcal{H}^1_k$-norm of the error, but also the $\mathcal{H}$-norm and negative norms. We will ignore these improvements in what follows for the sake of conciseness.

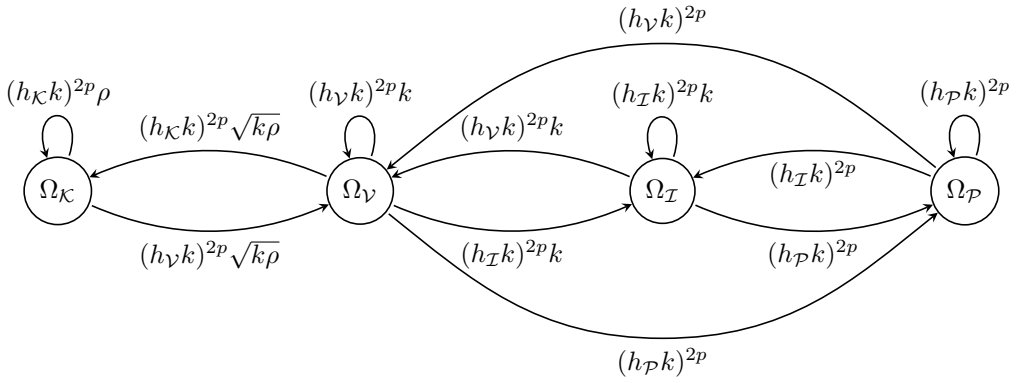


Figure 3.9: The graph showing propagation of errors for the decomposition into $\Omega_{\mathcal{K}}$, $\Omega_{\mathcal{V}}$, $\Omega_{\mathcal{I}}$, and $\Omega_{\mathcal{P}}$.

Corollary 3.2 (Sufficient condition for k-uniform quasi-optimality)

If, in addition to the assumptions of Theorem 3.1,

$$(h_{\mathcal{K}}k)^p\rho(k) + (h_{\mathcal{V}}k)^p\sqrt{k\rho} + (h_{\mathcal{I}}k)^pk + (h_{\mathcal{P}}k)^pk \leq \varepsilon \tag{3.17}$$

then

$$\|u - u_h\|_{\mathcal{H}^1_k} \leq C \inf_{v_h \in V^p(\Omega_h)} \|u - v_h\|_{\mathcal{H}^1_k}.$$

Proof. The condition (3.17) ensures that all coefficients of $\mathbf{B}$ and $\mathbf{T}$ are bounded. The conclusion follows by using the triangle inequality and choosing $v_{h,\mathcal{K}} = v_{h,\mathcal{V}} = v_{h,\mathcal{I}} = v_{h,\mathcal{P}} = v_h$, with v_h the best approximation of u in $V^p(\Omega_h)$. □

Chapter 4

Pseudo-locality results

In §3.3, we introduced the assumption in (3.8) requiring that the operators $P^\sharp(k)$, $S(k)$, $R^\sharp(k)$ and $\Pi_h^\sharp$ be local, but this is not quite true. In this chapter, we prove that, under appropriate assumptions on repeated commutators between $P(k)$ and cutoff functions (see Definition 4.2), the operators $S(k)$ and $R^\sharp(k)$ are then *pseudo-local*; that is, roughly, for any cutoffs $\chi \perp \psi$,

$$\chi S(k)\psi = O(k^{-\infty}), \quad \chi R^\sharp(k)\psi = O(k^{-\infty}).$$

We will only show the interior pseudo-locality, i.e., we will show this in the case where the cutoffs χ and ψ are supported away from the boundaries, see Remark 4.2.

4.1 Order notation and interior cutoffs

In this section, for all $k > 0$, we denote $\mathcal{H}_k^{-\infty} := \cup_{n\in\mathbb{Z}}\mathcal{H}_k^n$. Let $\mathscr{L}$ the vector spaces of families $\{L(k)\}_{k>0}$ such that for each $k > 0$, $L(k) : \mathcal{H}_k^{-\infty} \to \mathcal{H}_k^{-\infty}$ is a linear operator. Sometimes, we write $L(k)$ to denote the whole family (we will do this for instance with $P(k)$, $P^*(k)$ and $\mathscr{P}(k)$). For $A, B \in \mathscr{L}$, we denote by $AB \in \mathscr{L}$ the element of $\mathscr{L}$ defined by $\{A(k)B(k)\}_{k>0}$. For $A \in \mathscr{L}$, let $\mathrm{ad}_A : \mathscr{L} \to \mathscr{L}$ the linear operator defined by

$$\mathrm{ad}_A B := AB - BA.$$

Definition 4.1 (Order notation)

Given $L \in \mathscr{L}$ and $f : \mathbb{R}_+ \to \mathbb{R}_+$ we write

$$L = O_m(f)$$

if, $L(k)$ maps $\mathcal{H}_k^n$ to $\mathcal{H}_k^{n-m}$ for all $k > 0$ and $n \in \mathbb{Z}$, and, for all $k_0 > 0$ and all $n \in \mathbb{Z}$, there exists $C(k_0, n)$ such that the estimate

$$\|L(k)u\|_{\mathcal{H}_k^{n-m}} \leq Cf(k)\|u\|_{\mathcal{H}_k^n}$$

holds for all $k \geq k_0$ and $u \in \mathcal{H}^n_k$. If $A = O_m(k^M)$ for all $m < 0$ or for all $M < 0$, we then write

$$A = O_{-\infty}(k^M)\,, \quad A = O_m(k^{-\infty})\,,$$

respectively. If both properties hold, we write

$$A = O_{-\infty}(k^{-\infty}).$$

Observe that if $A = O_{m_1}(f_1)$ and $B = O_{m_2}(f_2)$, then $AB = O_{m_1+m_2}(f_1 f_2)$

Remark 4.1 (Order of $S(k)$ and $R^\sharp(k)$). Observe that by the definition of $S(k)$ in Definition 2.4, and by Propositions 2.9 and 2.10, one has

$$S(k) = \psi^\sharp(\mathscr{P}) = O_{-\infty}(1)\,, \quad R^\sharp(k) = O_{-2}(1).$$

Moreover, by Exercises 1.9, 2.1 and duality,

$$R(k) = O_{-2}((1 + \rho(k)).$$

Assumption 4.1 (Mapping properties of $P(k)$)

The operator $P(k)$ satisfies $P(k) = O_2(1)$.

It follows by duality that $P(k)^* = O_2(1)$ and similarly for $\mathscr{P}(k)$.

Definition 4.2 (Interior cutoffs)

We say that $\chi \in \mathscr{L}$ is an *interior cutoff* if it satisfies

(i) $\chi = O_0(1)$

(ii) For all $N \in \mathbb{N}$,

$$\mathrm{ad}^N_\chi Q = O_{-N+2}(k^{-N}) \quad \text{and} \quad \mathrm{ad}^N_{\chi^*} Q = O_{-N+2}(k^{-N})$$

where Q is any one of the operators $P(k)$, $P^*(k)$ and $\mathscr{P}(k)$.

Remark 4.2 (Boundaries). The assumption 4.1 does not hold in concrete settings if we choose the spaces $\mathcal{H}^n_k$ as in Remark 1.3. This is due to the fact that the action of $P(k)$ "removes boundary conditions". For a similar reason, for general smooth cutoffs χ, one will not have

$$\mathrm{ad}^N_\chi P(k) = O_{-N+2}(k^{-N})$$

due to boundary contributions.

Nevertheless, an appropriate assumption is

$$\psi P(k) \psi = O_2(1)$$

if $\psi \in C^\infty_c(\Omega)$. Moreover, for any cutoff χ supported away from boundaries, one can find $\psi \in C^\infty_c(\Omega)$ with $\chi \prec \psi$, and in this case,

$$\mathrm{ad}^N_\chi P(k) = \mathrm{ad}^N_\chi (\psi P(k) \psi)$$

and then, the estimate $O_{-N+2}(k^{-N})$ is indeed satisfied in concrete settings (see Exercise 4.1).

To properly show pseudo-locality up to the boundary, one needs to introduce another scale of spaces, and construct cutoff functions with a special behaviour at the boundary. We will not attempt to do this here.

Exercise 4.1. (Commutators with differential operators).
Let $\Omega \subset \mathbb{R}^d$ be a non-empty open set and for every $k > 0$, let $L(k) : C^\infty(\overline{\Omega}) \to C^\infty(\overline{\Omega})$ be the differential operator defined by

$$L(k)u(x) = \sum_{|\alpha| \leq N} a_\alpha(x) k^{-|\alpha|} \partial^\alpha u(x)$$

for some functions $a_\alpha \in C^\infty(\overline{\Omega})$.

Let $\chi \in C_c^\infty(\Omega)$ and $M \in \mathbb{N}$ and $n \geq \max(0, M - N)$. Show that there exists $C > 0$ such that for any $u \in C^\infty(\overline{\Omega})$ and for all $k > 0$,

$$\|(\mathrm{ad}_\chi^M L)u\|_{H_k^n(\Omega)} \leq C k^{-M} \|u\|_{H_k^{n+N-M}(\Omega)}$$

where the norm $H_k^n(\Omega)$ is defined by $\|u\|^2_{H_k^n(\Omega)} := \sum_{|\alpha| \leq n} k^{-2|\alpha|} \|\partial^\alpha u\|^2_{L^2(\Omega)}$.

Definition 4.3 (Separated operators)

We say $A, B \in \mathscr{L}$ are *separated* if there exists an interior cutoff χ such that

$$A(I - \chi) = 0 \quad \text{and} \quad \chi B = 0.$$

We think of A and B as two cutoff functions with disjoint, compact supports in Ω.

4.2 Pseudolocality of $S(k)$ and $R^\sharp(k)$

We can now state the main result of this chapter.

Theorem 4.2 (Pseudo-locality of S and $R^\sharp(k)$)

Suppose that Assumption (1.3)-(1.6), (2.1) and (4.1) hold. Let S and $R^\sharp$ be defined as in Definition 2.4. Let $A, B \in \mathscr{L}$ be separated and such that

$$A = O_0(1)\,, \quad B = O_0(1)\,.$$

Then,

$$AS(k)B = O_{-\infty}(k^{-\infty})\,, \quad AR^\sharp(k)B = O_2(k^{-\infty}).$$

Theorem 4.2 is obtained as a consequence of the following lemma:

Lemma 4.3 (Commutator estimates for $S(k)$ and $R^\sharp(k)$)

Let the assumptions of Theorem 4.2 be satisfied. Let χ be an interior cutoff. Then, for any $N \in \mathbb{N}$,

$$\mathrm{ad}_\chi^N S(k) = O_{-\infty}(k^{-N}) \quad \text{and} \quad \mathrm{ad}_\chi^N R^\sharp(k) = O_{-2}(k^{-N}).$$

Proof of Theorem 4.2 using Lemma 4.3. By assumption, there exists an interior cutoff χ such that $A = A\chi$ and $\chi B = 0$. Thus,

$$AS(k)B = (A\chi)S(k)B = AS(k)(\chi B) + A(\mathrm{ad}_\chi S(k))B = A(\mathrm{ad}_\chi S(k))B.$$

By repeating this argument N times, and using Lemma 4.3, we obtain

$$AS(k)B = A(\mathrm{ad}_\chi^N S(k))B = O_{-\infty}(k^{-N}).$$

Since this is true for any N this shows the pseudo-locality of $S(k)$. The reasoning for $R^\sharp(k)$ is identical. □

We now set out to prove Lemma 4.3. For this, the central tool will be the Helffer-Sjöstrand formula, that we recall now. The name of the formula is from Bernard Helffer and Johannes Sjöstrand. We refer to [18, Chapter 8] for historical notes about this formula, and to [48] for the proof of the particular form of this formula stated here. The function w in the formula is obtained as (a multiple of) a almost-analytic extension of f, as in [48, Theorem 3.6].

Proposition 4.4 (The Helffer-Sjöstrand formula)

For every smooth compactly supported function $f : \mathbb{R} \to \mathbb{R}$, there exists a continuous function $w : \mathbb{C} \to \mathbb{C}$ such that if $\mathscr{A}$ is a self-adjoint operator on a Hilbert space, then

$$f(\mathscr{A}) = \int_{\mathbb{C}} w(z)(\mathscr{A} - z)^{-1} dm_{\mathbb{C}}(z)$$

where $dm_{\mathbb{C}}(x + iy) = dx\, dy$ and for every $M \in \mathbb{N}$, there exists κ_M such that

$$|w(z)| \leq \kappa_M \langle z \rangle^{-2M} |\Im(z)|^M \qquad \text{for all} z \in \mathbb{C}. \tag{4.1}$$

Let us now summarize the strategy for proving Lemma 4.3. By the Helffer-Sjöstrand formula,

$$\mathrm{ad}_\chi^N f(\mathscr{P}) = \int_{\mathbb{C}} w(z) \mathrm{ad}_\chi^N (\mathscr{P}(k) - z)^{-1} dm_{\mathbb{C}}(z).$$

For $N = 1$, one has

$$\mathrm{ad}_\chi (\mathscr{P}(k) - z)^{-1} = (\mathscr{P}(k) - z)^{-1} (\mathrm{ad}_\chi \mathscr{P}(k)) (\mathscr{P}(k) - z)^{-1},$$

and more generally, for $N \geq 1$, it will be possible to express $\mathrm{ad}_\chi^N (\mathscr{P}(k) - z)^{-1}$ in terms of $\mathrm{ad}_\chi^N \mathscr{P}(k)$. Since we can bound the latter when χ is an interior cutoff (by Definition 4.2), we will thus obtain bounds on $\mathrm{ad}_\chi^N (\mathscr{P}(k) - z)^{-1}$ by using the mapping properties of $(\mathscr{P}(k) - z)^{-1}$ (see Proposition

2.11). If we can do this with some control over the variable z, we can then integrate these estimates in the Helffer-Sjöstrand formula to obtain a bound for $\mathrm{ad}_\chi f(\mathscr{P})$. Applying this with $f = \psi^\sharp$ (and recalling that $S(k) = \psi^\sharp(\mathscr{P})$), this allows to bound the commutator involving $S(k)$ in Lemma 4.3.

The control with respect to z is the object of the next lemma. To state it, we use the notation that $A = O_m(f(k,n,z))$ where $z \in U \subset \mathbb{C}$. This means that for all $k_0 > 0$, there is C such that for all $k \geq k_0$, $n \in \mathbb{Z}$ and all $z \in U$,

$$\|Au\|_{\mathcal{H}_k^{n-m}} \leq C f(k,n,z) \|u\|_{\mathcal{H}_k^n}.$$

Lemma 4.5 (Estimates commutators with inverses)

Let $U \subset \mathbb{C}$. Suppose that $X = O_m(1)$ and for every $z \in U$, let $Y_z, Y_z^* : \mathcal{H}_k^{n+2} \to \mathcal{H}_k^n$ be invertible. Furthermore, suppose that there are $L_n \geq 0$ such that

(a) for all $z \in U$,

$$Y_z^{-1} = O_{-2}\big(C_1(z)\langle z\rangle^{L_n}\big), \quad (Y_z^*)^{-1} = O_{-2}\big(C_1(z)\langle z\rangle^{L_n}\big)$$

(b) for all $z \in U$,

$$\mathrm{ad}_X^N Y_z = O_{2+N(m-1)}\big(k^{-N} C_2(z)\big), \quad \mathrm{ad}_{X^*}^N Y_z^* = O_{2+N(m-1)}\big(k^{-N} C_2(z)\big)$$

for some functions $C_1, C_2 : U \to \mathbb{R}_+$. Then for all $N \in \mathbb{N}$, $n \in \mathbb{Z}$, there is M_n such that, for all $z \in U$,

$$\mathrm{ad}_X^N Y_z^{-1} = O_{-2+N(m-1)}\big(k^{-N}(1+C_1(z))^{N+1}(1+C_2(z))^N \langle z\rangle^{M_n}\big).$$

Proof. The basic idea is that $\mathrm{ad}_X^N Y_z^{-1}$ is equal to a linear combination of terms of the form $Y_z^{-1}(\mathrm{ad}_X^{i_1} Y_z)Y_z^{-1}(\mathrm{ad}_X^{i_2} Y_z)Y_z^{-1} \dots Y_z^{-1}(\mathrm{ad}_X^{i_M} Y_z)Y_z^{-1}$, and the next definitions formalize this more precisely.[1]

We will prove the lemma by showing the estimate for $\mathrm{ad}_X^N Y_z^{-1}$ acting on elements of $\mathcal{H}$ and then (using the second parts of assumptions (a) and (b)) argue by duality to act on $\mathcal{H}_k^{-n}$.

An operator $a_N : \mathcal{H} \to \mathcal{H}$ is called an (N,z)-*atom* if either

(i) $N = 0$ and $a_N = 1$, or

(ii) $a_N = (\mathrm{ad}_X^N Y_z)Y_z^{-1}$, or

(iii) $a_N = a_i a_j$ where a_i is an (i,z)-atom and a_j is an (j,z)-atom with $i + j = N$ and $1 \leq i, j \leq N-1$.

[1] It is in fact possible to give a full closed-form expression for $\mathrm{ad}_X^N Y^{-1}$ involving sums of compositions of quantities of the form $(\mathrm{ad}_X^i Y)$ and Y^{-1}. However, the formula and its proof, involving sums over all possible ordered partitions of $\{1,\dots,N\}$, are slightly cumbersome and for the present purposes, this would be more information than actually needed.

An operator t_N is called an (N,z)-*term* if it is of the form

$$t_N = \sum_{j=1}^{J} \sigma_j Y_z^{-1} a_{N,j}$$

where $J \in \mathbb{N}$, σ_j are real coefficients and $a_{N,j}$ are (N,z)-atoms. For example,

$$t_5 = Y_z^{-1}(\mathrm{ad}_X^5 Y_z)Y_z^{-1} - Y_z^{-1}(\mathrm{ad}_X^2 Y_z)Y_z^{-1}(\mathrm{ad}_X^3 Y_z)Y_z^{-1}$$

is a $(5,z)$-term. Notice that if t_i and t_j are (i,z)- and (j,z)-terms, then $t_i Y_z t_j$ is an $(i+j,z)$-term.

It follows immediately from assumptions (a) and (b), by induction on N, that if $t_N(z)$ is a (N,z)-term for all $z \in U$, then

$$t_N(z) = (1+C_1(z))^{N+1}(1+C_2(z))^N O_{-2+N(m-1)}(k^{-N}\langle z\rangle^{M_n}; \mathcal{H}_k \to \mathcal{H}_k).$$

Thus it remains to show that for all $z \in U$, $\mathrm{ad}_X^N Y_z^{-1} : \mathcal{H} \to \mathcal{H}$ is an (N,z)-term. For this, it suffices to prove that, for all $N \in \mathbb{N}$,

$$t_N \text{ is an } (N,z)\text{-term} \quad \Longrightarrow \quad \mathrm{ad}_X t_N \text{ is an } (N+1,z)\text{-term.} \tag{4.2}$$

By linearity, it is enough to prove (4.2) in the case where $t_N = Y_z^{-1} a_N$ for some (N,z)-atom a_N. We consider separately the three cases (i), (ii), (iii) above in the definition of an (N,z)-atom.

Case (i): If $a_N = 1$, then

$$\mathrm{ad}_X t_N = \mathrm{ad}_X Y_z^{-1} = XY_z^{-1} - Y_z^{-1}X = Y_z^{-1}Y_z XY_z^{-1} - Y_z^{-1}XY_zY_z^{-1} = -Y_z^{-1}(\mathrm{ad}_X Y_z)Y_z^{-1}$$

which is a $(1,z)$-term acting on u. This shows the implication (4.2) for $N=0$, and in the following cases, we fix $N \geq 1$ and proceed by induction assuming that it holds for all $i \leq N-1$.

Case (ii): If $a_N = (\mathrm{ad}_X^N Y_z)Y_z^{-1}$, then

$$\mathrm{ad}_X t_N = \quad (\mathrm{ad}_X Y_z^{-1})(\mathrm{ad}_X^N Y_z)Y_z^{-1} + Y_z^{-1}(\mathrm{ad}_X^{N+1} Y_z)Y_z^{-1} + Y_z^{-1}(\mathrm{ad}_X^N Y_z)(\mathrm{ad}_X Y_z^{-1}).$$

The second term on the right-hand side is an $(N+1,z)$-term. The first term on the right-hand side can be rewritten as

$$-Y_z^{-1}\underbrace{(\mathrm{ad}_X Y_z)Y_z^{-1}}_{(1,z)\text{-atom}}\underbrace{(\mathrm{ad}_X^N Y_z)Y_z^{-1}}_{(N,z)\text{-atom}}.$$

This is thus an $(N+1,z)$-term. Similarly, the third term is an $(N+1,z)$-term, and thus $\mathrm{ad}_X t_N$ is an $(N+1,z)$-term.

Case (iii): If $a_N = a_i a_j$ then, since $a_j : \mathcal{H} \to \mathcal{H}$,

$$t_N = Y_z^{-1} a_i a_j = Y_z^{-1} a_i Y_z Y_z^{-1} a_j = t_i Y_z t_j$$

where $t_i := Y_z^{-1} a_i$ and $t_j := Y_z^{-1} a_j$ are (i,z)- and (j,z)-terms, respectively, with $i+j=n$. Thus

$$\mathrm{ad}_X t_N = (\mathrm{ad}_X t_i)Y_z t_j + t_i(\mathrm{ad}_X Y_z)t_j + t_i Y_z(\mathrm{ad}_X t_j).$$

The first term is an $(N+1,z)$-term by the induction hypothesis. Similarly, the last term is an $(N+1,z)$-term. The middle term can be rewritten as

$$t_i(\mathrm{ad}_X Y_z)t_j = \overbrace{t_i Y_z \underbrace{Y_z^{-1}(\mathrm{ad}_X Y_z)Y_z^{-1}}_{(1,z)\text{-term}} Y_z t_j}^{(i+(j+1),z)\text{-term}}{}^{(j+1,z)\text{-term}}$$

which is an $(N+1,z)$-term. This concludes the proof. □

We can now complete the proof of Lemma 4.3, and thus, of Theorem 4.2.

Proof of Lemma 4.3. 1. Let $f \in C_c^\infty(\mathbb{R})$. By the Helffer–Sjöstrand formula,

$$\mathrm{ad}_\chi^N f(\mathscr{P}) = \int_{\mathbb{C}} w(z)\mathrm{ad}_\chi^N(\mathscr{P}(k)-z)^{-1}\, dm_{\mathbb{C}}(z).$$

By Lemma 4.5 with $X=\chi$, $U=\mathbb{C}\setminus\mathbb{R}$, $Y_z=(\mathscr{P}(k)-z)$, the definition of interior cutoffs, and the resolvent estimate of Proposition 2.11,

$$\mathrm{ad}_\chi^N(\mathscr{P}(k)-z)^{-1} = \left(1+\frac{\langle z\rangle}{|\Im(z)|}\right)^N O_{-2+N(m-1)}(k^{-N}\langle z\rangle^{M_n}).$$

Therefore,

$$\mathrm{ad}_\chi^N f(\mathscr{P}) = O_{-2+N(m-1)}\left(k^{-N}\int_{\mathbb{C}^N} w(z)\langle z\rangle^{M_n}\left(1+\frac{\langle z\rangle}{|\Im(z)|}\right)^N dm_{\mathbb{C}}(z)\right).$$

The bound (4.1) on w implies that the integral is finite, and thus, for all $f\in C_c^\infty(\mathbb{R})$,

$$\mathrm{ad}_\chi^N f(\mathscr{P}) = O_{-2+N(m-1)}(k^{-N}). \tag{4.3}$$

2. We now upgrade the regularity index from $-2+N(m-1)$ to $-\infty$ by induction on N. For $N=0$,

$$\mathrm{ad}_\chi^0 f(\mathscr{P}) = f(\mathscr{P}) = O_{-\infty}(1) \tag{4.4}$$

by Proposition 2.9. Next fix an integer $N\geq 1$ and suppose that for all $i\leq N-1$ and all $g\in C_c^\infty(\mathbb{R})$,

$$\mathrm{ad}_\chi^i g(\mathscr{P}) = O_{-\infty}(k^{-i}).$$

Let $f\in C_c^\infty(\mathbb{R})$ and let f_1 and f_2 such that $f=f_1f_2$. Thus $f(\mathscr{P})=f_1(\mathscr{P})f_2(\mathscr{P})$ and thus, using the Leibniz identity

$$\mathrm{ad}_X^N(YZ) = \sum_{i=0}^N \binom{N}{i}(\mathrm{ad}_X^i Y)(\mathrm{ad}_X^{N-i}Z),$$

we obtain

$$\mathrm{ad}_\chi^N(f_1(\mathscr{P})f_2(\mathscr{P}))$$

$$= f_1(\mathscr{P})(\mathrm{ad}_\chi^N f_2(\mathscr{P})) + (\mathrm{ad}_\chi^N f_1(\mathscr{P}))f_2(\mathscr{P}) + \sum_{i=1}^{N-1} \binom{N}{i} (\mathrm{ad}_\chi^i f_1(\mathscr{P}))(\mathrm{ad}_\chi^{N-i} f_2(\mathscr{P})).$$

Bounding the first two terms on the right-hand side by (4.4) and (4.3), and bounding the third term by the induction hypothesis, we obtain that

$$\begin{aligned}\mathrm{ad}_\chi^N (f_1(\mathscr{P})f_2(\mathscr{P})) &= O_{-\infty}(1)O_{-2+N(m-1)}(k^{-N}) + \sum_{i=1}^{N-1} O_{-\infty}(k^{-i})O_{-\infty}(k^{-N+i}) \\ &= O_{-\infty}(k^{-N}).\end{aligned}$$

This completes the induction, showing that for any $f \in C_c^\infty(\mathbb{R})$,

$$\mathrm{ad}_\chi^N f(\mathscr{P}) = O_{-\infty}(k^{-N}).$$

Applying this with $f = \psi^\sharp$, this shows the claimed commutator estimate for $S(k)$.

3. In turn, we obtain the commutator estimate involving $R^\sharp(k)$ by applying Lemma 4.5 with $X = \chi$, $U = \{1\}$ and $Y_1 = P^\sharp(k)$. Indeed, for assumption (a), the required estimate is given by Proposition 2.10, while for assumption (b),

$$\mathrm{ad}_\chi^N P^\sharp(k) = \mathrm{ad}_\chi^N P(k) + \mathrm{ad}_\chi^N S(k) = O_{2+N(m-1)}(k^{-N}) + O_{-\infty}(k^{-N})$$

by the definition of spatial cutoffs (Definition 4.2) and by the previous step. □

.

Bibliography

[1] R. A. Adams and J. J. F. Fournier. *Sobolev spaces*, volume 140. Elsevier, 2003.

[2] 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-Scienze Fisiche e Matematiche*, 21(4):599–637, 1967.

[3] M. Averseng, J. Galkowski, and E. A. Spence. Helmholtz FEM solutions are locally quasi-optimal modulo low frequencies. *Adv. Comput. Math.*, 50(6):Paper No. 112, 48, 2024.

[4] M. Averseng, J. Galkowski, and E. A. Spence. Non-uniform finite-element meshes defined by ray dynamics for Helmholtz problems. *arXiv preprint arXiv:2506.15630*, 2025.

[5] J.-F. Bony, N. Burq, and T. Ramond. Minoration de la résolvante dans le cas captif. *Comptes Rendus. Mathématique*, 348(23-24):1279–1282, 2010.

[6] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008.

[7] 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 Math.*, 180(1):1–29, 1998.

[8] N. Burq. Lower bounds for shape resonances widths of long range Schrödinger operators. *American journal of mathematics*, 124(4):677–735, 2002.

[9] N. Burq. Semi-classical estimates for the resolvent in nontrapping geometries. *International Mathematics Research Notices*, 2002(5):221–241, 2002.

[10] J.-P. Bérenger. A perfectly matched layer for the absorption of electromagnetic waves. *Journal of computational physics*, 114(2):185–200, 1994.

[11] F. Cardoso and G. Vodev. Uniform estimates of the resolvent of the Laplace-Beltrami operator on infinite volume Riemannian manifolds. II. *Ann. Henri Poincaré*, 3(4):673–691, 2002.

[12] P. G. Ciarlet. *The finite element method for elliptic problems*. Studies in Mathematics and its Applications, Vol. 4. North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978.

[13] D. Colton and R. Kress. *Integral equation methods in scattering theory*. SIAM, 2013.

[14] R. Courant. Variational methods for the solution of problems of equilibrium and vibrations. *Bull. Amer. Math. Soc.*, 49:1–23, 1943.

[15] J. Céa. Approximation variationnelle des problèmes aux limites. *Ann. Inst. Fourier*, 14(2):345–444, 1964.

[16] K. Datchev and A. Vasy. Propagation through trapped sets and semiclassical resolvent estimates. *Ann. Inst. Fourier (Grenoble)*, 62(6):2347–2377 (2013), 2012.

[17] E. De Giorgi. *Sull'analiticità delle estremali degli integrali multipli*. Consiglio Nazionale delle Ricerche, 1956.

[18] M. Dimassi and J. Sjöstrand. *Spectral asymptotics in the semi-classical limit*, volume 268 of *Lond. Math. Soc. Lect. Note Ser.* Cambridge: Cambridge University Press, 1999.

[19] S. Dyatlov and M. Zworski. *Mathematical theory of scattering resonances*, volume 200 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2019.

[20] D. M. Eidus. The principle of limit amplitude. *Russian Mathematical Surveys*, 24(3):97, 1969.

[21] B. G. Galerkin. Beams and plates. *Vestnik Ingenerov*, 19:897–908, 1915.

[22] J. Galkowski, D. Lafontaine, and E. A. Spence. Perfectly-matched-layer truncation is exponentially accurate at high frequency. *SIAM Journal on Mathematical Analysis*, 55(4):3344–3394, 2023.

[23] J. Galkowski and E. A. Spence. Numerical analysis of the high-frequency Helmholtz equation using semiclassical analysis. *arXiv preprint arXiv:2511.15287*, 2025.

[24] J. Galkowski and E. A. Spence. Sharp preasymptotic error bounds for the Helmholtz-fem. *SIAM Journal on Numerical Analysis*, 63(1):1–22, 2025.

[25] J. Galkowski, E. A. Spence, and J. Wunsch. Optimal constants in nontrapping resolvent estimates and applications in numerical analysis. *Pure and Applied Analysis*, 2(1):157–202, 2019.

[26] L. Gårding. Dirichlet's problem for linear elliptic partial differential equations. *Mathematica Scandinavica*, pages 55–72, 1953.

[27] L. Hörmander. *The analysis of linear partial differential operators. III: Pseudo-differential operators.* Classics in Mathematics. Springer, Berlin, 1994 edition, 2007.

[28] 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.

[29] 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 Journal on Numerical Analysis*, 34(1):315–358, 1997.

[30] T. Kato. *Perturbation theory for linear operators*, volume 132. Springer Science & Business Media, 2013.

[31] 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.

[32] P. D. Lax and A. N. Milgram. Parabolic equations. In *Contributions to the theory of partial differential equations*, Ann. of Math. Stud., no. 33, pages 167–190. Princeton Univ. Press, Princeton, NJ, 1954.

[33] W. C. H. McLean. *Strongly elliptic systems and boundary integral equations*. Cambridge university press, 2000.

[34] J. M. Melenk and S. Sauter. Convergence analysis for finite element discretizations of the Helmholtz equation with Dirichlet-to-Neumann boundary conditions. *Mathematics of Computation*, 79(272):1871–1914, 2010.

[35] C. S. Morawetz. The limiting amplitude principle. *Communications on Pure and Applied Mathematics*, 15(3):349–361, 1962.

[36] J. Nash. Continuity of solutions of parabolic and elliptic equations. *American Journal of Mathematics*, 80(4):931–954, 1958.

[37] J. Nitsche. Ein kriterium für die quasi-optimalität des ritzschen verfahrens. *Numerische Mathematik*, 11(4):346–348, 1968.

[38] F. Rellich. Ein satz über mittlere konvergenz. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, 1930:30–35, 1930.

[39] F. Rellich. Über das asymptotische Verhalten der Lösungen von $\Delta u + \lambda u = 0$ in unendlichen Gebieten. *Jber. Deutsch. Math.-Verein.*, 53:57–65, 1943.

[40] D. Robert and H. Tamura. Semi-classical estimates for resolvents and asymptotics for total scattering cross-sections. *Ann. Inst. Henri Poincaré, Phys. Théor.*, 46:415–442, 1987.

[41] W. Rudin. *Functional analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, Inc., New York, second edition, 1991.

[42] S. A. Sauter. A refined finite element convergence theory for highly indefinite Helmholtz problems. *Computing*, 78(2):101–115, 2006.

[43] A. H. Schatz. An observation concerning Ritz-Galerkin methods with indefinite bilinear forms. *Mathematics of Computation*, 28(128):959–962, 1974.

[44] M. Schechter. *Principles of functional analysis.*, volume 36 of *Grad. Stud. Math.* Providence, RI: American Mathematical Society (AMS), 2nd ed. edition, 2001.

[45] S. L. Sobolev. *Some applications of functional analysis in mathematical physics*, volume 90 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1991. Translated from the third Russian edition by Harold H. McFaden.

[46] A. Sommerfeld. Die greensche funktion der schwingungslgleichung. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 21:309–352, 1912.

[47] G. Vodev. On the exponential bound of the cutoff resolvent. *Serdica Math. J.*, 26(1):49–58, 2000.

[48] M. Zworski. *Semiclassical analysis*, volume 138 of *Grad. Stud. Math.* Providence, RI: American Mathematical Society (AMS), 2012.