Lecture Notes
Spectral Theory

Thierry Ramond
Université Paris Sud & Ritsumeikan University
e-mail: thierry.ramond@math.u-psud.fr

March 1, 2015

Contents

Foreword

This graduate course is intended as an introduction to the spectral theory of unbounded operators. We shall give a detailed exposition of the general theory, and illustrate the notions with numerous examples of operators, mainly from quantum mechanics. The last part of the course will be devoted to recent developpements around the notion of pseudospectrum for non-selfadjoint operators. We aim in particular to review some numerical experiments from the book of M. Embree and L.Trefetthen.

Main References :

[Br] Brezis, H., Analyse Fonctionnelle, Dunod, 1999.

[Da] Davies, E.B. , Spectral Theory of Differential Operators, Cambrdge University Press, 1995.

[He] Helffer, B. , Spectral Theory And Its Applications (Cambridge Studies In Advanced Mathematics), Cambridge University Press 2013.

[Te] Teschl, G. , Mathematical Methods in Quantum Mechanics, with application to Schrödinger operators, Graduate Studies in Mathematics, Volume 99, 2009.

[TrEm] Trefethen, Lloyd N., Embree M., Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators, Princeton University Press, 2005.

Plan of the lecture :

Chapter 1 Introduction to quantum mechanics (2 lectures)

Chapter 2 Hilbert spaces (3 lectures)

Chapter 3 Spectral Theory for Bounded Operators (4 lectures)

Chapter 4 Unbounded operators (3 lectures)

Chapter 5 Examples of Non-Selfadjoint Operators and Pseudospectrum (3 lectures)

Chapter 1

A brief introduction to quantum mechanics

Spectral theory has been initiated at the early beginnings of the 20th century by D. Hilbert, but it has been set to its present form (almost) some years later by J. Von Neumann in close relation with the development of quantum mechanics. Our aim in this course is to present general ideas from spectral theory and illustrate them by examples coming from physics. In particular, the emphasis will be put on self-adjoint operators (a generalization of the notion of hermitian matrix, say). However the last chapter of the course will deal with recent advances in the spectral study of non-selfadjoint operators, and the notion of pseudo-spectrum. In this brief introduction, we try to explain how one has been led to the Schrödinger equation, and raise some natural questions that we shall answer in later chapters.

1.1 Hamiltonian Classical Mechanics

According to Newton's law, the trajectory $\mathbb{R} \ni t \mapsto x(t) \in \mathbb{R}^n$ of a particle of mass m under a force field $F(x)$, that we suppose for clarity to be time-independent, satisfies

$$mx''(t) = F(x(t)). \tag{1.1.1}$$

The derivative $x'(t)$ is the speed of the particle at time t, and $x''(t)$ is its acceleration. Equivalently, (1.1.1) can be written as

$$\begin{cases} x'(t) = \dfrac{1}{m}\xi(t), \\ \xi'(t) = F(x(t)), \end{cases} \tag{1.1.2}$$

where $\xi(t) := mx'(t)$ is the momentum of the particle. In that setting, the curve $(x(t), \xi(t))$, which is called the phase space trajectory of the classical particle, appears to be an integral curve

of the vector field $H : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^n$ given by

$$H(x,\xi) = \begin{pmatrix} \xi/m \\ F(x) \end{pmatrix}. \tag{1.1.3}$$

Taking into account how these equations transform if we change coordinates, it seems to be important to distinguish between the variables x and ξ. This leads to consider $\mathcal{H}$ as a function on $T^*\mathbb{R}^n$ instead of $\mathbb{R}^n \times \mathbb{R}^n$, with values in the space $T(T^*\mathbb{R}^n)$ of tangent vectors to $T^*\mathbb{R}^n$. Let us say that this distinction is meaningful mainly when there are physical reasons to consider that the space in which the particle moves is a manifold M (for example a circle or a sphere…) instead of $\mathbb{R}^n$.

When the force field comes from a potential, ie $F(x) = -\nabla V(x)$ for some smooth fonction $V : \mathbb{R}^n \to \mathbb{R}$, the energy $p(t) = p(x(t), \xi(t))$ of the particle at time t, defined as the sum of its kinetic energy and of its potential energy,

$$p(t) = \frac{1}{2m}\xi(t)^2 + V(x(t)), \tag{1.1.4}$$

is constant along the trajectory $t \mapsto \exp(tH)(x,\xi)$. Indeed

$$\partial_t p(t) = \frac{1}{m}\xi(t) \cdot \partial_t \xi(t) + \nabla V(x(t)) \cdot \partial_t x(t) = 0,$$

thanks to (1.1.2).

Another interesting point is that the vector field $\mathcal{H}$ in (1.1.3) can be defined in terms of the function $p(x,\xi) = \xi^2 + V(x)$ in (1.1.4), through the formula:

$$H(x,\xi) = H_p(x,\xi) = \partial_\xi p(x,\xi)\partial_x - \partial_x p(x,\xi)\partial_\xi, \tag{1.1.5}$$

and H_p is called the Hamiltonian field associated to p.

1.2 Quantum Mechanics - Schrödinger picture

At the beginning of the 20th century, some physical experiments had led to results that classical mechanics could not explain, and that even seem to be in conflict with one another. For example, the study of the radiations emitted by a so-called black-body, or the discovery of the photoelectric effect, seem to suggest that light is constituted of particles, with given energy, that were named quanta. On the other hand Young's experiment (also called double-slit experiment) seems to show that lights do behave like a wave, that may generate diffraction patterns.

E. Schrödinger, reasoning by analogy with optics, proposed a model that permits to predict with an extraordinary accuracy the results of these experiments, and was, therefore, universally adopted by physicists.

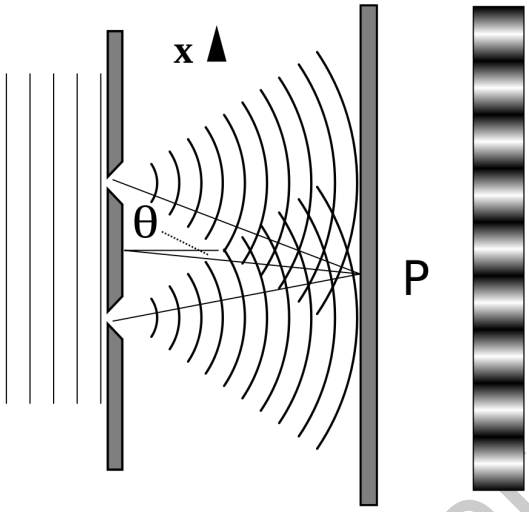

Figure 1.1: Young's Experiment

1.2.1 Plane waves and the Schrödinger equation

We call plane wave with wave vector k and angular velocity ω (in french: pulsation, 脈動) the function

$$\phi(t,x) = e^{i(k\cdot x - \omega t)}. \tag{1.2.6}$$

Notice that for fixed t, $x \mapsto \phi(t,x)$ is constant on each hyper-plane perpendicular to k: one says that the plane wave propagates in the direction of k.

If we suppose that this plane wave describes a quantum particle with momentum ξ, the only reasonable choice is to set

$$\xi = \hbar k \tag{1.2.7}$$

for some real constant $\hbar$, that we call Planck's constant (as a master of fact, Planck's constant is $h = 2\pi\hbar$). Then we call

$$E = h\nu = \hbar\omega \tag{1.2.8}$$

the energy of the particle, where $\nu = 2\pi\omega$ is the frequency of the plane wave, and (1.2.6) becomes

$$\phi(t,x) = e^{i(\xi\cdot x - Et)/\hbar} \tag{1.2.9}$$

Now the kinetic energy E_c of the particle has to be $E_c = \frac{\xi^2}{2m}$, so that

$$E_c\phi(t,x) = -\frac{\hbar^2}{2m}\Delta\phi(t,x), \quad \text{where } \Delta = \sum_{j=1}^{n}\partial_j^2. \tag{1.2.10}$$

If further the particle is placed in a potential $V(x)$, its total energy E is the sum of V and of its kinetic energy. Since, thanks to (1.2.9), $E\phi(t,x) = \frac{\hbar}{i}\partial_t\phi(t,x)$, we obtain Schrödinger's equation

$$i\hbar\partial_t\phi(t,x) = -\frac{\hbar^2}{2m}\Delta\phi(t,x) + V(x)\phi(t,x). \tag{1.2.11}$$

In this equation, we see Schrödinger's operator :

$$P(x,hD_x) = -\frac{\hbar^2}{2m}\Delta + V(x), \tag{1.2.12}$$

a second order, linear, partial differential operator. Notice that it may be formally obtained by replacing ξ by $hD_x = \frac{h}{i}\partial_x$ in the expression of the classical energy function $p(x,\xi)$, a procedure that is called quantization of the classical observable p.

Then Schrödinger's postulate is that quantum particles are associated with normalized solutions $\phi(t,x)$ of this equation, that is solutions such that for each fixed t,

$$\|\phi(t,.)\|_{L^2} = (\int |\phi(t,x)|^2 dx)^{1/2} = 1. \tag{1.2.13}$$

The function $\phi(t,x)$ is then called wave function of the particle, and the quantity $|\phi(t,x)|^2$ is interpreted as the density of probability of presence of the particule. In other words,

$$\|\mathbf{1}_\Omega(x)\phi(t,.)\|_{L^2(\mathbb{R}^n)} = (\int \mathbf{1}_\Omega(x)|\phi(t,x)|^2 dx)^{1/2}$$

is the probability of presence of the particle in the region $\Omega \subset \mathbb{R}^n$ at time t.

Notice that the plane wave (1.2.9) is not in $L^2(\mathbb{R}^n)$. However, any function of $L^2(\mathbb{R}^n)$ can be written as a superposition of plane waves (or wave packet)

$$\phi(t,x) = \frac{1}{(2\pi\hbar)^n}\int \psi(t,\xi)e^{i(\xi\cdot x - Et)/\hbar}d\xi,$$

by the Fourier-Plancherel theorem. One may in particular consider a plane wave as the "limit" of a gaussian wave packet:

$$g(t,x) = \int e^{ix\cdot\xi/\hbar}e^{-itE/\hbar}e^{-(\xi-\eta)^2/2\hbar}d\xi.$$

A more serious difficulty, is that Schrödinger's equation (1.2.11) does not make sense for a function $\phi(t,.)$ in $L^2(\mathbb{R}^n)$. Even if we may consider $P(x,hD_x)\phi(t,.)$ as distribution, the equality (1.2.11) asks if $P(x,hD_x)\phi(t,.)$ is an $L^2(\mathbb{R}^n)$ function. We are therefore immediately led to the notion of unbounded operators, that is operators on a Hilbert space $\mathcal{H}$ whose domain of operation is not the whole $\mathcal{H}$. The mathematical theory of such operators is due to Von Neumann at the

beginning of the years 1930. One may also notice that distribution's theory by Laurent Schwartz came more than 20 years later.

Let us finish this very brief discussion with the presentation of the stationary Schrödinger operator. When the potential V does not depend on t, it is meaningful to search for solutions of (1.2.11) with separate variables t and x, that is of the form $\phi(t,x) = a(t)u(x)$. We get that, for a certain constant $E \in \mathbb{C}$, which is the energy of the particle,

$$a(t) = e^{-iEt/h}, \quad \text{and } P(x, hD_x)u(x) = Eu(x), \tag{1.2.14}$$

The equation for u appears as an equation for eigenvalues of the (unbounded) operator P, and the possible energies E of a quantum particle are therefore eigenvalues of the stationary Schrödinger operator. Of course the fact that E may be a complex number is somewhat puzzling if we think at physical interpretation, and we shall come back to this point in these notes. For now, let us notice that when V is real valued,

$$\forall u \in \mathcal{C}_0^\infty(\mathbb{R}^n),\ \langle Pu, u\rangle_{L^2} = \langle u, Pu\rangle_{L^2}, \tag{1.2.15}$$

so that $Pu = Eu$ implies that $E \in \mathbb{R}$. When (1.2.15) holds, we say that the operator P is symmetric.

When the set of eigenvalues of P is a discrete set, we obtain the so-called quantization of the energy levels, compatible with the experiments leading to the description of quantum particles as individualized corpuscles.

1.2.2 Quantum Observables and the Uncertainty Principle

To the classical energy function $p(x,\xi)$, we have associated the Schrödinger operator $P(x, hD_x)$, which eigenvalues should be the different possible energies of quantum stationary states. In a general way, we would like to be able to associate such an operator to any reasonable function q of (x,ξ), called "classical observable". There are many ways to do so, mainly because the operators associated with $q(x,\xi) = x$ (the position operator) and $p(x,\xi) = \xi$ (the impulsion operator) do not commute. From this elementary fact follows the well-known Uncertainty Principle that we review now.

• The quantum observable (the operator) associated to the position function $x_j(x,\xi) = x_j$ is simply the operator X_j of multiplication by x_j. Here again, we have to deal with an unbounded operator: for $u \in L^2$, it is not always true that $X_j u$ is in L^2. The average j-th coordinate of the position of the particle described by the wave function $\phi(t,x)$ is defined as

$$\langle X_j\rangle_\phi = \langle x_j\phi(t,x), \phi(t,x)\rangle_{L^2} = \int x_j|\phi(t,x)|^2\, dx. \tag{1.2.16}$$

One may notice that this quantity is the average of the function x_j for the measure $|\phi(t,x)|^2 dx$.

• As for a plane wave (1.2.6), the quantum observable associated with the momentum function $\xi_j(x,\xi) = \xi_j$ is the operator $\Xi_j = \frac{\hbar}{i}\partial_j = \hbar D_j$, here again an unbounded operator on L^2. The average j-th coordinate of the momentum of the particle described by the wave function $\phi(t,x)$ is

$$\begin{aligned} \langle \Xi_j \rangle_\phi &= \langle \frac{\hbar}{i}\partial_j \phi(t,x), \phi(t,x)\rangle_{L^2} = \int \frac{\hbar}{i}\partial_j\phi(t,x)\overline{\phi(t,x)}\, dx \\ &= \frac{1}{(2\pi\hbar)^n}\int \mathcal{F}_\hbar(\frac{\hbar}{i}\partial_j\phi(t,.))(\xi)\overline{\mathcal{F}_\hbar(\phi(t,.))(\xi)}\, d\xi \\ &= \frac{1}{(2\pi\hbar)^n}\langle \xi_j \mathcal{F}_\hbar\phi(t,.), \mathcal{F}_\hbar\phi(t,.)\rangle, \end{aligned} \tag{1.2.17}$$

where $\mathcal{F}_\hbar$ is the semiclassical Fourier transform defined, for example for ϕ in $\mathcal{S}(\mathbb{R}^n)$, by

$$\mathcal{F}_\hbar\phi(\xi) = \int e^{-ix\cdot\xi/\hbar}\phi(x)dx, \quad \mathcal{F}_\hbar^{-1}\phi(x) = \frac{1}{(2\pi h)^n}\int e^{ix\cdot\xi/\hbar}\phi(\xi)d\xi. \tag{1.2.18}$$

Above, we have used Parseval's formula

$$\int \phi(x)\overline{\psi(x)}dx = \frac{1}{(2\pi h)^n}\int \mathcal{F}_h(\phi(\xi))\overline{\mathcal{F}_h(\psi)(\xi)}d\xi, \tag{1.2.19}$$

and the relation

$$\mathcal{F}_\hbar(\hbar D_j u) = \xi_j \mathcal{F}_\hbar(u)(\xi). \tag{1.2.20}$$

It is a simple computation to show that the commutator $[X_j, \Xi_j] = X_j\Xi_j - \Xi_j X_j$ is the operator $i\hbar I$, where we have denoted I the identity operator on $L^2(\mathbb{R}^n)$: for any $\phi \in L^2$ such that this computation has a meaning,

$$[X_j, \Xi_j]\phi = i\hbar\phi. \tag{1.2.21}$$

The following result is a general formulation of the uncertainty principle:

Lemma 1.2.1 Let A_1 and A_2 be two quantum observables (i.e. two self-adjoint operators). Let $u \in L^2$ such that $\|u\|_{L^2} = 1$, and, for $j = 1, 2$, $\delta_j = (\langle A_j^2\rangle_u - (\langle A_j\rangle_u)^2)^{1/2}$ the standard deviation of the observable A_j in the state u. If $[A_1, A_2] = i\hbar I$, then

$$\delta_1\delta_2 \geq \frac{\hbar}{2} = \frac{h}{4\pi}\cdot$$

Proof: One may consider only the case where $\langle A_j\rangle_u = 0$, moving to the observables $\tilde{A}_j = A_j - \langle A_j\rangle_u I$. Then

$$\delta_1^2\delta_2^2 = \|A_1u\|^2\, \|A_2u\|^2 \geq |\langle A_1u, A_2u\rangle|^2.$$

Since A_1 is self-adjoint, we have

$$\begin{aligned}\langle A_1u, A_2u\rangle &= \langle u, A_1A_2u\rangle \\ &= \frac{1}{2}\langle u, (A_1A_2 + A_2A_1)u\rangle + \frac{1}{2}\langle u, (A_1A_2 - A_2A_1)u\rangle \\ &= \frac{1}{2}\langle u, (A_1A_2 + A_2A_1)u\rangle + i\frac{\hbar}{2}.\end{aligned}$$

Moreover $(A_1A_2 + A_2A_1)$ is symmetric, so that the first term of the RHS is real, and finally

$$\delta_1^2\delta_2^2 \geq |\langle A_1u, A_2u\rangle|^2 = \frac{1}{4}\langle u, (A_1A_2 + A_2A_1)u\rangle^2 + \frac{\hbar^2}{4} \geq \frac{\hbar^2}{4}.$$

□

The above lemma applies to the operators X_j and Ξ_j: we shall see later on that they are self-adjoint operators. But for these two, one may prefer the following more explicit proof of the uncertainty principle: for $u \in L^2(\mathbb{R}^n)$

$$\frac{\hbar}{2}\|u\|_{L^2}^2 \leq \|X_ju\|_{L^2}\,\|\hbar D_ju\|_{L^2}. \tag{1.2.22}$$

Indeed, we can get (1.2.22) computing $\|\lambda X_ju + i\hbar D_ju\|^2$ for $\lambda \in \mathbb{R}$, when we notice that since it is of constant sign, it has a negative discriminant as a function of λ. The main role is of course still played by the relation $[X_j, \Xi_j] = i\hbar I$:

$$\begin{aligned}\int [X_j, \Xi_j]u(x)\overline{u(x)}dx &= \int \hbar D_ju(x)\overline{x_ju(x)}dx - \int \hbar D_j(x_ju)\overline{u(x)}dx \\ &= -2i\,\mathrm{Im}\int x_ju(x)\overline{\hbar D_ju(x)}dx,\end{aligned}$$

thus

$$\begin{aligned}\|\lambda X_ju + i\hbar D_ju\|^2 &= \lambda^2\|X_ju\|^2 + 2\lambda\,\mathrm{Re}\int x_ju(x)\overline{i\hbar D_ju(x)}dx + \|hD_ju\|^2 \\ &= \lambda^2\|X_ju\|^2 - \lambda\hbar\|u\|^2 + \|hD_ju\|^2,\end{aligned}$$

and

$$\hbar^2\|u\|^4 - 4\|X_ju\|^2\|hD_ju\|^2 \leq 0.$$

At last, notice that (1.2.22) can be written another way, closer to that given in Lemma 1.2.1: for $(x_0, \xi_0) \in \mathbb{R}^n \times \mathbb{R}^n$, writing (1.2.20) for the function $v(x) = e^{i\xi_0\cdot x}u(x + x_0)$ and using the fact that the semiclassical Fourier transform is an isometry (up to a constant factor) in $L^2(\mathbb{R}^n)$, we have

$$\frac{\hbar}{2}\|u\|_{L^2}^2 \leq \|(x - x_0)_ju\|\;\|(\xi - \xi_0)_j\mathcal{F}_hu\|. \tag{1.2.23}$$

1.3 An example of energy quantization

In this section we shall describe the possible energy levels of a particle placed in a potential well. Our aim is to illustrate the importance of spectral theory on a particularly simple model, where (almost) no sophisticated tools are necessary. We consider the harmonic oscillator

$$P_{osc}(x, \hbar D) = (\hbar D)^2 + V(x), \tag{1.3.24}$$

where

$$V(x) = \sum_{j=1}^{d} \mu_j x_j^2 \text{ with } \mu_j > 0 \text{ for all } j = 1 \dots d. \tag{1.3.25}$$

We postpone the precise definition of the operator P_{osc} on $L^2(\mathbb{R}^n)$, and we only look for its eigenvalues, i.e. the E's in $\mathbb{C}$ such that Ker$(P_{osc} - E) \neq \{0\}$ in L^2.

We proceed by separation of variables: thanks to the particular form of the potential, we look for solutions of the equation $P_{osc}u = Eu$ that can be written as

$$u(x_1, x_2, \dots x_d) = u_1(x_1)u_2(x_2) \dots u_d(x_d). \tag{1.3.26}$$

We obtain a diagonal system of ordinary differential equations for the unknown functions u_j:

$$-\hbar^2 \partial_j^2 u_j + \mu_j x_j^2 u_j = E_j u_j, \tag{1.3.27}$$

where the E_j should sum to E, and we study each of these equations separately. They are of the form

$$P_{\hbar,\mu} u = Eu, \quad P_{\hbar,\mu} = (\hbar D)^2 + \mu x^2. \tag{1.3.28}$$

Performing the change of variable $x \mapsto y(x) = \mu^{1/4} \dfrac{x}{\sqrt{\hbar}}$, we obtain the equation

$$-v''(y) + y^2 v(y) = \frac{E}{\hbar\sqrt{\mu}} v(y), \tag{1.3.29}$$

where $v(y) = u(x)$. Therefore, we are led to the study of the differential operator Q on $L^2(\mathbb{R})$ given by

$$Q(x, \hbar D) = D^2 + x^2, \tag{1.3.30}$$

and to the associated eigenvalue equation

$$Qu = Eu. \tag{1.3.31}$$

There are many ways to solve this equation, and we have chosen to base our discussion on the following two remarks:

• For $u \in \mathcal{C}_0^\infty(\mathbb{R})$, we have $\langle Qu, u\rangle \in \mathbb{R}$, or even $\langle Qu, u\rangle \geq \|u\|_{L^2}^2$. Indeed, integrating by parts, we see that

$$\langle Qu, u\rangle = \int (D^2 + x^2)u(x)\bar{u}(x)dx = \|Du\|^2 + \|xu\|^2 \geq 2\|Du\|\,\|xu\| \geq \|u\|^2, \tag{1.3.32}$$

where the last inequality is nothing else than the uncertainty principle ($\hbar = 1$ in the present settings). Notice that $\langle Qu, u\rangle = \langle u, Qu\rangle$ for $u \in \mathcal{C}_0^\infty(\mathbb{R})$, a dense subset of $L^2(\mathbb{R})$: the unbounded operator Q is thus said to be symmetric.

Using again the density of $\mathcal{C}_0^\infty(\mathbb{R})$ in $L^2(\mathbb{R})$, we can deduce that if $u \in L^2(\mathbb{R})$ is a solution of (1.3.31) with $\|u\|_{L^2} = 1$, then $E \in [1, +\infty[$. As a matter of fact, this property is only true for u in L^2 that can be written as limit of a sequence (u_n) in $\mathcal{C}_0^\infty(\mathbb{R})$ such that $Qu_n \to Qu$ in L^2. This gives some insight to what a good choice of domain for the unbounded operator P_{osc} should be.

• We can write $Q = L^+L^- + 1 = L^-L^+ - 1$, where L^+ and L^- are the so-called creation and annihilation operators respectively:

$$L^+ = -\partial_x + x, \quad L^- = \partial_x + x. \tag{1.3.33}$$

Indeed, we have, for example:

$$L^+L^-u = (-\partial_x + x)(\partial_x u + xu) = -u'' + [x, \partial_x]u + x^2u = Qu - u.$$

Notice also that $\langle L^-u, v\rangle = \langle u, L^+v\rangle$, so that $\langle \phi, L^+L^-\phi\rangle = \|L^-\phi\|^2 \geq 0$. This is another proof of the inequality $\langle Qu, u\rangle \geq \|u\|^2$.

Suppose now that $u_E \in L^2$ is a normalized eigenfunction ($\|u_E\|_{L^2} = 1$) of the operator L^+L^- for the eigenvalue $E \geq 0$, i.e. $L^+L^-u_E = Eu_E$. We have

$$EL^-u_E = L^-(L^+L^-)u_E = (L^-L^+)L^-u_E = (L^+L^- + 2)L^-u_E, \tag{1.3.34}$$

so that L^-u_E is an eigenvector of L^+L^- for the eigenvalue $E - 2$, unless $L^-u_E = 0$. One may also notice that $\|L^-u_E\|^2 = \langle u_E, L^+L^-u_E\rangle = E$. Therefore, L^+L^- cannot have an eigenvalue which is not an even integer, otherwise $(L^-)^n u_E$ would be for n large enough an eigenfunction for a negative eigenvalue.

Let us put it the other way round: the equation $L^-u_0 = 0$ has a unique solution in L^2, given by

$$u_0(x) = \pi^{-1/4}e^{-x^2/2}. \tag{1.3.35}$$

Thus u_0 is an eigenfunction for L^+L^- for the eigenvalue 0. Then $\tilde{u}_2 = L^+u_0$ is an eigenvector for the eigenvalue 2, with norm

$$\|\tilde{u}_2\|^2 = \langle L^+u_0, L^+u_0\rangle = \langle u_0, L^-L^+u_0\rangle = \langle u_0, (L^+L^- + 2)u_0\rangle = 2.$$

Thus $u_2(x) = \frac{1}{\sqrt{2}}L^+u_0$ is a normalized eigenvector. By induction, we see that the set of eigenvalues of L^+L^- is $2\mathbb{N}$, and that the normalized eigenvector associated to $E = 2n$ is

$$u_n = \frac{1}{\sqrt{2n(2n-2)(2n-4)\dots 2}}(L^+)^n u_0 = \frac{1}{2^{n/2}\sqrt{n!}\pi^{1/4}}(L^+)^n(e^{-x^2/2}).$$

Finaly, since $Q = L^+L^- + 1$, the set of the eigenvalues of the operator Q is $2\mathbb{N}+1$, and u_n is the eigenfunction associated to $E = 2n+1$.

Now we come back to the operator $P_{h,\mu}$. The set of eigenvalules of $P_{h,\mu}$ is $\sigma_p = \{(2n+1)h\sqrt{\mu}, n \in \mathbb{N}\}$, and the eigenfunction associated to $E = (2n+1)h\sqrt{\mu}$ is

$$\psi_n(x) = C_n(\hbar)u_n(\mu^{1/4}\frac{x}{\sqrt{\hbar}}),$$

where $C_n(\hbar) = h^{-1/4}\mu^{1/8}$ is chosen so that $\|\psi_n\| = 1$. It is easy to see that ψ_n can also be written

$$\psi_n(x) = \frac{1}{2^{n/2}\sqrt{n!}\pi^{1/4}}h^{-1/4}\mu^{1/8}H_n(\mu^{1/4}\frac{x}{\sqrt{\hbar}})e^{-\sqrt{\mu}x^2/2\hbar}, \tag{1.3.36}$$

where H_n is a polynomial of degree n. The H_n are the so-called Hermite polynomials, given by the relation

$$H_n(x) = e^{x^2/2}(x-\partial_x)^n(e^{-x^2/2}) = (-1)^n e^{x^2}\partial_x^n(e^{-x^2}).$$

Note that

Lemma 1.3.1 *The vector space $\mathcal{H}_0$ generated by the functions ψ_n is dense in $L^2(\mathbb{R})$.*

Proof: Of course, $\mathcal{H}_0$ is also generated by $\mathcal{D} = \{x \mapsto x^k e^{-x^2/2}, k \in \mathbb{N}\}$. But if $\psi \in \mathcal{D}^\perp$, we have for all $n \in \mathbb{N}$ and all $\xi \in \mathbb{R}$,

$$\int \sum_{k=0}^{n} \frac{(-i\xi x)^k}{k!} e^{-x^2/2}\bar{\psi}(x)dx = 0.$$

Thus, using Lebesgue's dominated convergence theorem, $\mathcal{F}_{\hbar,x\to\xi}(e^{-x^2/2}\bar{\psi}(x)) = 0$ and $\psi = 0$. Therefore $\mathcal{D}^\perp = \{0\}$, so that $\overline{\mathcal{D}} = L^2$ (see Corollary 2.2.8). □

At last, we consider the full harmonic oscillator P_{osc} on $L^2(\mathbb{R}^n)$. We have shown that the set of eigenvalues of P_{osc} contains

$$\sigma(P_{osc}) = \{\lambda(\alpha) = \sum_{j=1}^{d}\sqrt{\mu_j}(2\alpha_j+1)\hbar, \alpha \in \mathbb{N}^d\},$$

and that the eigenvector associated to $\lambda(\alpha)$ is

$$u_\alpha(x,\hbar) = \hbar^{-d/4}(\mu_1\dots\mu_d)^{1/8}H_n(\mu^{1/4}\frac{x_1}{\sqrt{\hbar}})\dots H_n(\mu^{1/4}\frac{x_d}{\sqrt{\hbar}})e^{-\sum_{j=1}^{d}\sqrt{\mu}x_j^2/2\hbar}$$

It will follow from the material of the next chapters that P_{osc} has no other eigenvalue. We will also see that the whole spectrum of P_{osc} is the set $\sigma(P_{osc})$ given above.

Chapter 2

Hilbert spaces

2.1 Scalar Products

Let $\mathcal{H}$ be a vector space on $\mathbb{C}$.

Definition 2.1.1 A linear form [resp. anti-linear form] ℓ on $\mathcal{H}$ is a mapping $\ell : \mathcal{H} \to \mathbb{C}$ such that

$$\forall x, y \in \mathcal{H}, \forall \lambda \in \mathbb{C},\ \ell(x+y) = \ell(x) + \ell(y),\ \text{and}\ \ell(\lambda x) = \lambda \ell(x)\ [\text{resp. } \ell(\lambda x) = \bar{\lambda}\ell(x)].$$

Definition 2.1.2 A sesquilinear form on $\mathcal{H}$ is a mapping $s : \mathcal{H} \times \mathcal{H} \to \mathbb{C}$ such that for all $y \in \mathcal{H}$, $x \mapsto s(x, y)$ is linear and $x \mapsto s(y, x)$ is anti-linear. If moreover $s(x, y) = \overline{s(y, x)}$, the sesquilinear form s is said to be Hermitian .

Notice that when the sesquilinear form s is Hermitian, $s(x, x) \in \mathbb{R}$ for any $x \in \mathcal{H}$. Using the following identity, we can easily see that it is a necessary and sufficient condition:

Proposition 2.1.3 (Polarization identity) Let s be a Hermitian sesquilinear form on $\mathcal{H}$. For all $(x, y) \in \mathcal{H} \times \mathcal{H}$,

$$4s(x, y) = s(x+y, x+y) - s(x-y, x-y) + is(x+iy, x+iy) - is(x-iy, x-iy).$$

Notice in particular that a Hermitian sesquilinear form is completely determined by its values on the diagonal of $\mathcal{H} \times \mathcal{H}$.

Remark 2.1.4 For a real symmetric bilinear form b, the polarization identity reads

$$4b(x,y) = b(x+y, x+y) - b(x-y, x-y).$$

Definition 2.1.5 A (Hermitian) scalar product is a Hermitian sesquilinear form s such that $s(x,x) \geq 0$ for all $x \in \mathcal{H}$, and $s(x,x) = 0 \Leftrightarrow x = 0$.

Proposition 2.1.6 When s is a Hermitian scalar product, the Cauchy-Schwarz inequality holds:

$$\forall x, y \in \mathcal{H},\ |s(x,y)| \leq \sqrt{s(x,x)}\sqrt{s(y,y)},$$

as well as the triangular (or Minkowski's) inequality:

$$\forall x, y \in \mathcal{H},\ \sqrt{s(x+y, x+y)} \leq \sqrt{s(x,x)} + \sqrt{s(y,y)}.$$

Proof: Let $x, y \in \mathcal{H}$. Denote θ the argument of the complex number $s(x,y)$, so that $|s(x,y)| = e^{-i\theta} s(x,y)$. For any $\lambda \in \mathbb{R}$, we have

$$s(x + \lambda e^{i\theta} y\,,\, x + \lambda e^{i\theta} y) \geq 0.$$

Therefore, for any $\lambda \in \mathbb{R}$,

$$\begin{aligned} 0 &\leq s(x,x) + s(x, \lambda e^{i\theta} y) + s(\lambda e^{i\theta} y, x) + s(\lambda e^{i\theta} y, \lambda e^{i\theta} y) \\ &\leq s(x,x) + 2\lambda\, \mathsf{Re}(e^{-i\theta} s(x,y)) + \lambda^2 s(y,y) \\ &\leq s(x,x) + 2\lambda |s(x,y)| + \lambda^2 s(y,y). \end{aligned}$$

Since this 2nd order polynomial has constant sign, its discriminant is negative, that is

$$|s(x,y)|^2 - s(x,x)\, s(y,y) \leq 0,$$

which is the Cauchy-Schwarz inequality.

Minkowski's inequality is then a simple consequence of the Cauchy-Schwarz inequality

$$\begin{aligned} s(x+y, x+y) &= s(x,x) + 2\,\mathsf{Re}\, s(x,y) + s(y,y) \\ &\leq s(x,x) + 2|\,\mathsf{Re}\, s(x,y)| + s(y,y) \\ &\leq s(x,x) + 2\sqrt{s(x,x)}\sqrt{s(y,y)} + s(y,y) \\ &\leq (\sqrt{s(x,x)} + \sqrt{s(y,y)})^2. \end{aligned}$$

□

In particular the map $\|\cdot\| : x \mapsto \sqrt{s(x,x)}$ is a norm on $\mathcal{H}$, and for all $x, y \in \mathcal{H}$, we have

$$|s(x,y)| \leq \|x\| \, \|y\|.$$

Thus the scalar product is a continuous map from $\mathcal{H} \times \mathcal{H}$ to $\mathbb{C}$ for the topology defined by its associated norm.

Definition 2.1.7 A Hilbert space is a pair $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ where $\mathcal{H}$ is a vector space on $\mathbb{C}$, and $\langle\cdot,\cdot\rangle$ is a Hermitian scalar product on $\mathcal{H}$, such that $\mathcal{H}$ is complete for the associated norm $\|\cdot\|$.

Example 2.1.8 – The space $\mathbb{C}^n$, equipped with the scalar product

$$\langle x, y\rangle = \sum_{j=1}^{n} x_j \overline{y_j}$$

is a Hilbert space.

– The space $\ell^2(\mathbb{C})$ of sequences (x_n) such that $\sum |x_n|^2 < +\infty$, equipped with the scalar product $\langle (x_n), (y_n)\rangle = \sum_n x_n \overline{y_n}$ is a Hilbert space.

– The space $L^2(\Omega)$ of square integrable functions on the open set $\Omega \subset \mathbb{R}^n$, equipped with the scalar product

$$\langle f, g\rangle_{L^2} = \int f(x)\overline{g(x)}dx,$$

is a Hilbert space. This is one of the main achievement of Lebesgue's integration theory.

Exercise 2.1.9 Prove that $\ell^2(\mathbb{C})$ is a Hilbert space: Let (x^n) be a Cauchy sequence in $\ell^2(\mathbb{C})$. Denote $x^n = (x_i^n)_{i\in\mathbb{N}}$.

1. Show that the sequence $(x_i^n)_{n\in\mathbb{N}}$ is a Cauchy sequence of $\mathbb{C}$. Denote x_i its limit.

2. Show that the sequence $x = (x_i)$ belongs to $\ell^2(\mathbb{C})$, and that (x^n) converges to x.

2.2 Orthogonality

Definition 2.2.1 Let $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ be a Hilbert space, and A a subset of $\mathcal{H}$. The orthogonal complement to A is the set $A^\perp$ given by

$$A^\perp = \{x \in \mathcal{H},\ \forall a \in A,\ \langle x, a\rangle = 0\}.$$

In the case where $A = \{x\}$, $A^\perp$ is the set of vectors that are orthogonal to x.

Proposition 2.2.2 For any subset A of $\mathcal{H}$, $A^\perp$ is a closed subspace of $\mathcal{H}$. Moreover $A^\perp = (\bar{A})^\perp$.

Proof: For each $a \in A$, the set $\{a\}^\perp$ is closed, since the map $x \mapsto \langle x, a\rangle$ is continuous. Thus $A^\perp$ is the intersection of a family of closed set, therefore a closed set. Now $0 \in A^\perp$, and if $x_1, x_2 \in A^\perp$, we have $\langle \lambda_1 x_1 + \lambda_2 x_2, a\rangle = \lambda_1\langle x_1, a\rangle + \lambda_2\langle x_2, a\rangle = 0$ for any $a \in A$, so that $A^\perp$ is indeed a subspace of $\mathcal{H}$.

Since $A \subset \bar{A}$, we have $(\bar{A})^\perp \subset A^\perp$. On the other hand let $b \in A^\perp$. For $a \in \bar{A}$, there exists a sequence (a_n) of vectors in A such that $(a_n) \to a$. Now

$$\langle a, b\rangle = \lim_{n\to+\infty} \langle a_n, b\rangle = 0,$$

so that $b \in (\bar{A})^\perp$. □

Lemma 2.2.3 (Pythagore's theorem) Let $\{x_1, x_2, \ldots, x_n\}$ be a family of pairwise orthogonal vectors. Then

$$\|x_1 + x_2 + \cdots + x_n\|^2 = \|x_1\|^2 + \|x_2\|^2 + \cdots + \|x_n\|^2.$$

Proof: Indeed

$$\|x_1 + x_2 + \cdots + x_n\|^2 = \langle \sum_{j=1}^n x_j, \sum_{k=1}^n x_k\rangle = \sum_{j=1}^n \sum_{k=1}^n \langle x_j, x_k\rangle = \sum_{j=1}^n \|x_j\|^2.$$

□

Lemma 2.2.4 (Parallelogram's law) Let x_1 and x_2 be two vectors of the Hilbert space $\mathcal{H}$. Then

$$2\|x_1\|^2 + 2\|x_2\|^2 = \|x_1 + x_2\|^2 + \|x_1 - x_2\|^2.$$

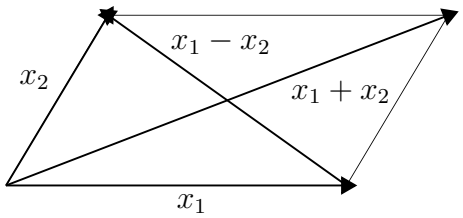

Figure 2.1: Parallelogram's law

The proof of this lemma is straightforward, but it is interesting to notice that the parallelogram identity holds if and only if the norm $\|\cdot\|$ comes from a scalar product, that is the map

$$(x_1, x_2) \mapsto \frac{1}{4}(\|x_1 + x_2\|^2 - \|x_1 - x_2\|^2 + i\|x_1 + ix_2\|^2 - i\|x_1 - ix_2\|^2)$$

is a scalar product whose associated norm is $\|\cdot\|$.

Exercise 2.2.5 Prove it.

Proposition 2.2.6 (Orthogonal Projection) Let $\mathcal{H}$ be a Hilbert space, and F a closed subspace of $\mathcal{H}$. For any x in $\mathcal{H}$, there exists a unique vector Πx in F such that

$$\forall f \in F,\ \|x - \Pi x\| \leq \|x - f\|.$$

This element Πx is called the orthogonal projection of x onto F, and it is characterized by the property

$$\Pi x \in F \text{ and } \forall f \in F,\ \langle x - \Pi x, f\rangle = 0.$$

Moreover the map $\Pi : x \mapsto \Pi x$ is linear, $\Pi^2 = \Pi$, and $\|\Pi x\| \leq \|x\|$.

Notice that the proposition states in particular that Πx is the only element of F such that $x - \Pi x$ belongs to $F^\perp$.

Proof: – First of all we suppose only that F is a convex, closed subset of $\mathcal{H}$. Let $x \in \mathcal{H}$ be fixed, and denote $d = \inf_{f\in F} \|x - f\|$ the distance between x and F.

If f_1 and f_2 are two vectors in F, then, since F is convex, $(f_1 + f_2)/2$ also belongs to F. Therefore $\|(f_1 + f_2)/2\| \geq d$. On the other hand the parallelogram law says that

$$\|\frac{(f_1 + f_2)}{2}\|^2 + \|\frac{(f_1 - f_2)}{2}\|^2 = \frac{1}{2}(\|f_1\|^2 + \|f_2\|^2),$$

so that

$$0 \leq \|\frac{(f_1 - f_2)}{2}\|^2 \leq \frac{1}{2}(\|f_1\|^2 + \|f_2\|^2) - d^2.$$

Now for $n \in \mathbb{N}$, we define

$$F_n = \{f \in F,\ \|x - f\|^2 \leq d^2 + \frac{1}{n}\}.$$

The sets F_n are closed, and non-empty by the definition of d, and they form a decreasing sequence of sets. Moreover, if f_1, f_2 belong to F_n, then

$$\|\frac{(f_1 - f_2)}{2}\|^2 \leq \frac{1}{2}(\|f_1 - x\|^2 + \|f_2 - x\|^2) - d^2 \leq \frac{1}{n}.$$

Thus the diameter of the F_n tends to 0, and their intersection, which is the set of points in F at distance d of x contains at most one point.

At last, for all $n \in \mathbb{N}$ we pick $x_n \in F_n$. For all $p < q$ in $\mathbb{N}$, we have $F_q \subset F_p$ therefore

$$\|x_p - x_q\| \leq \frac{1}{p},$$

which proves that (x_n) is a Cauchy sequence, therefore converges to some $\Pi x \in F$, such that $\|x - \Pi x\| = d$.

Concerning the characterization of Πx, we notice that for all $t \in [0, 1]$ and all $f \in F$, we have $(1 - t)\Pi x + tf \in F$ by the convexity assumption on F. Thus

$$\|\Pi x - x\|^2 \leq \|((1 - t)\Pi x + tf) - x\|^2 \leq \|(\Pi x - x) + t(f - \Pi x))\|^2.$$

Thus, for all $f \in F$ and all $t \in [0, 1]$,

$$0 \leq t^2\|f - \Pi x\|^2 + 2t\,\mathrm{Re}\langle \Pi x - x, f - \Pi x\rangle.$$

Dividing by t and choosing $t = 0$ gives

$$\mathrm{Re}\langle x - \Pi x, f - \Pi x\rangle \leq 0.$$

Reciprocally if $\mathrm{Re}(x - y, f - y) \leq 0$ for all $f \in F$, then, for all $f \in F$

$$\|x - f\|^2 = \|x - y + y - f\|^2 \geq \|x - y\|^2,$$

so that $y = \Pi x$.

– Now we make the assumption that F is a closed vector space. Since a vector space is convex, the previous proof holds. Moreover, in the last part, we have now

$$0 \leq |t|^2\|f - \Pi x\|^2 + 2\,\mathrm{Re}\,\bar{t}\langle \Pi x - x, f - \Pi x\rangle.$$

for all $t \in \mathbb{C}$ since the line $\{(1 - t)\Pi x + tf,\ t \in \mathbb{C}\}$ belongs to F. Therefore, in that case, Πx is characterized by the property (noticing that $f - \Pi x$ describe F as f describes F),

$$\forall f \in F,\ \langle x - \Pi x, f\rangle = 0.$$

The linearity of the map Π then follows: for $x_1, x_2 \in \mathcal{H}$ and $\lambda_1, \lambda_2 \in \mathbb{C}$ we have, for all $f \in F$,

$$\lambda_1\langle x_1 - \Pi x_1, f\rangle = 0 \quad \text{and} \quad \lambda_2\langle x_2 - \Pi x_2, f\rangle = 0$$

so that, for all $f \in F$,

$$\langle \lambda_1 x_1 + \lambda_2 x_2 - (\lambda_1 \Pi x_1 + \lambda_2 \Pi x_2), f\rangle = 0,$$

and $\Pi(\lambda_1 x_1 + \lambda_2 x_2) = \lambda_1 \Pi x_1 + \lambda_2 \Pi x_2$.

Since $\Pi x = x$ when $x \in F$, it is clear that $\Pi^2 = \Pi$, therefore we are left with the proof that $\|\Pi x\| \leq \|x\|$. But this is obvious since for all $f \in F$, and in particular for $f = 0$, we have, by Pythagore's theorem

$$\|x - f\|^2 = \|x - \Pi x\|^2 + \|\Pi x - f\|^2$$

and thus $\|x - f\| \geq \|\Pi x - f\|$. □

Corollary 2.2.7 If F is a closed subspace of $\mathcal{H}$, then

$$F \oplus F^\perp = \mathcal{H}.$$

Proof: For $x \in \mathcal{H}$, we can write $x = \Pi x + (I - \Pi)x = x_1 + x_2$. Since $x_1 \in F$ and $x_2 = x - \Pi x$ is orthogonal to F, we have $\mathcal{H} = F + F^\perp$, and it remains to show that the sum is direct. If $0 = x_1 + x_2$ with $x_1 \in F$ and $x_2 \in F^\perp$, then Pythagore's theorem give

$$0 = \|x_1\|^2 + \|x_2\|^2,$$

so that $x_1 = x_2 = 0$. □

Corollary 2.2.8 A subspace F of $\mathcal{H}$ is dense in $\mathcal{H}$ if and only if $F^\perp = \{0\}$. Moreover $(F^\perp)^\perp = \bar{F}$.

Proof: Since $\bar{F}$ is a closed subspace of $\mathcal{H}$, we have

$$H = \bar{F} \oplus (\bar{F})^\perp = \bar{F} \oplus F^\perp,$$

and the first statement follows.

It is clear that $F \subset (F^\perp)^\perp$. Since $(F^\perp)^\perp$ is a closed set, this implies that $\bar{F} \subset (F^\perp)^\perp$. On the other hand, let $x \in (F^\perp)^\perp$, and denote Πx its projection onto $\bar{F}$. We have

$$\|x - \Pi x\|^2 = \langle x - \Pi x, x - \Pi x\rangle = \langle x - \Pi x, x\rangle - \langle x - \Pi x, \Pi x\rangle = 0.$$

Indeed, $\langle x - \Pi x, \Pi x\rangle = 0$ since $\Pi x \in \bar{F}$, and $\langle x - \Pi x, x\rangle = 0$ since $x - \Pi x \in (\bar{F})^\perp = F^\perp$. Thus $x = \Pi x \in \bar{F}$. □

2.3 Riesz's theorem

A linear form $\ell : \mathcal{H} \to \mathbb{C}$ is continuous if there exists $C > 0$ such that

$$\forall x \in \mathcal{H},\ |\ell(x)| \leq C\|x\|.$$

Proposition 2.3.1 (Riesz's representation Theorem) Let ℓ be a continuous linear form on $\mathcal{H}$. There exists a unique $y = y(\ell) \in \mathcal{H}$ such that

$$\forall x \in \mathcal{H},\ \ell(x) = \langle x, y\rangle.$$

Moreover

$$|||\ell||| := \sup_{x\in\mathcal{H}, x\neq 0} \frac{|\ell(x)|}{\|x\|} = \|y(\ell)\|.$$

Proof: The uniqueness part of the statement is easy, and we concentrate on the existence part. We denote $\mathrm{Ker}\,\ell = \{x \in \mathcal{H},\ \ell(x) = 0\}$ the kernel of ℓ. Since ℓ is continuous, it is a closed subspace of $\mathcal{H}$, and we denote by Π the orthogonal projection onto $\mathrm{Ker}\,\ell$. If $\ell = 0$, we can take $y(\ell) = 0$. Otherwise, there exists $z \in \mathcal{H}$ such that $\ell(z) \neq 0$, which means that $w = z - \Pi z \neq 0$. Therefore we can set

$$y = y(\ell) = \frac{\overline{\ell(w)}}{\|w\|^2} w.$$

Notice in particular that $\ell(y) = \|y\|^2$. As a matter of fact, y spans $(\mathrm{Ker}\,\ell)^\perp$. Indeed if $x \in (\mathrm{Ker}\,\ell)^\perp$, we have

$$\ell(x - \frac{\ell(x)}{\ell(y)} y) = 0$$

therefore $x - \dfrac{\ell(x)}{\ell(y)} y \in \mathrm{Ker}\,\ell \cap (\mathrm{Ker}\,\ell)^\perp = \{0\}$, so that $x = \dfrac{\ell(x)}{\ell(y)} y$.

Thus, again since $H = \mathrm{Ker}\,\ell \oplus (\mathrm{Ker}\,\ell)^\perp$, any $x \in \mathcal{H}$ can be written

$$x = \Pi x + \lambda y,$$

for some $\lambda \in \mathbb{C}$. Then $\ell(x) = \lambda \ell(y)$ and

$$\langle x, y\rangle = \langle \Pi x + \lambda y, y\rangle = \langle \Pi x, y\rangle + \lambda \frac{\ell(w)^2}{\|w\|^2} = \lambda \|y\|^2 = \ell(x).$$

□

2.4 Lax-Milgram's theorem

Proposition 2.4.1 (Lax-Milgram) Let $\mathcal{H}$ be a Hilbert space on $\mathbb{C}$, and $a(x,y)$ a sesquilinear form on $\mathcal{H}$. We assume that

i) The sesquilinear form a is continuous , i.e. there exists $M > 0$ such that $|a(x,y)| \leq M\|x\|\,\|y\|$ for all $x, y \in \mathcal{H}$.

ii) The sesquilinear form a is coercive , i.e. there exists $c > 0$ such that $|a(x,x)| \geq c\|x\|^2$ for all $x \in \mathcal{H}$.

Then, for any continuous linear form ℓ on $\mathcal{H}$, there exists a unique $y \in \mathcal{H}$ such that

$$\forall x \in \mathcal{H},\ \ell(x) = a(x,y).$$

Moreover $\|y\| \leq \|\ell\|/c$.

Proof: For any $y \in \mathcal{H}$, the linear form $x \mapsto a(x,y)$ is continuous. Thanks to Riesz theorem, there exists a unique $A(y) \in \mathcal{H}$ such that

$$\forall x \in \mathcal{H},\ a(x,y) = \langle x, A(y)\rangle.$$

The map $A : y \mapsto A(y)$ is linear, since, for all $x \in \mathcal{H}$,

$$\langle x, A(\alpha_1 y_1 + \alpha_2 y_2\rangle = a(x, \alpha_1 y_1 + \alpha_2 y_2) = \overline{\alpha_1} a(x,y_1) + \overline{\alpha_2} a(x,y_2) = \langle x, \alpha_1 A(y_1) + \alpha_2 A(y_2)\rangle.$$

The map A is also continuous since we have $\langle A(y), A(y)\rangle = a(A(y), y) \leq M\|A(y)\|\|y\|$, so that

$$\|A(y)\| \leq M\|y\|.$$

Now let ℓ be a continuous linear form on $\mathcal{H}$. There exists $z \in \mathcal{H}$ such that

$$\forall x \in \mathcal{H},\ \ell(x) = \langle x, z\rangle.$$

Therefore we are left with the equation $A(y) = z$ for a given $z \in \mathcal{H}$, and we are going to show that it has a unique solution, namely that A is a bijection on $\mathcal{H}$.

Since a is coercive, one has

$$c\|y\|^2 \leq |a(y,y)| \leq |\langle y, A(y)\rangle| \leq \|A(y)\|\,\|y\|,$$

so that

$$\|A(y)\| \geq c\|y\|, \tag{2.4.1}$$

and A is 1 to 1.

Moreover Ran A is a closed subspace of $\mathcal{H}$. Indeed if $(v_j) \in$ Ran A converges to v in $\mathcal{H}$, setting $v_j = Au_j$, we obtain thanks to (2.4.1),

$$c\|u_p - u_q\| \leq \|v_p - v_q\|.$$

So (u_j) is a Cauchy sequence, and converges to some $u \in \mathcal{H}$. Since A is continuous, one has

$$v = \lim_{j\to+\infty} v_j = \lim_{j\to+\infty} A(u_j) = A(\lim_{j\to+\infty} u_j) = Au,$$

and $v \in$ Ran A.

Eventually if $x \in (\text{Ran}\, A)^\perp$, we have $0 = |\langle A(x), x\rangle| \geq c\|x\|^2$, so that $(\text{Ran}\, A)^\perp = \{0\}$, and Ran $A = \overline{\text{Ran}\, A} = ((\text{Ran}\, A)^\perp)^\perp = \mathcal{H}$. □

2.5 The Dirichlet problem

As an illustration, we apply now the Lax-Milgram theorem to prove the existence as well as the uniqueness of the solution of the so-called Dirichlet problem.

Let us start with a 1d problem. We want to solve the following problem on $I =]0, 1[$,

$$\begin{cases} -u'' + Vu = f, \\ u(0) = u(1) = 0, \end{cases} \tag{2.5.2}$$

where the potential V is in $L^\infty(I)$ and the source term f is in $L^2(I)$. When V and f are continuous, a classical solution is a function $u \in \mathcal{C}^2(\bar{I})$ such that for all $x \in I$, $-u''(x) + V(x)u(x) = f(x)$. Obviously, when $f \in L^2$ is not continuous, this can not hold for any $u \in \mathcal{C}^2(I)$. We are thus lead to change to another notion of solution: a function $u \in \mathcal{C}^1(I)$ is a weak solution of (2.5.2) when

$$\forall \varphi \in \mathcal{C}_0^1(I),\ \int_I u'(x)\varphi'(x)dx + \int_I V(x)u(x)\varphi(x)dx = \int_I f(x)\varphi(x)dx. \tag{2.5.3}$$

Notice that this integral formulation can be obtained for $\mathcal{C}^2$ functions by multiplying the differential equation in (2.5.2) by $\varphi(x)$ and integrating by parts: a classical solution is of course a weak solution. As a matter of fact, since $\mathcal{C}_0^\infty(I)$ is dense in $\mathcal{C}_0^1(I)$, we may, and we will, replace $\mathcal{C}_0^1$ by $\mathcal{C}_0^\infty$ in the definition of weak solutions.

We can further extend the notion of solution. Indeed let us introduce the set $H^1(I)$ as the set of functions u in $L^2(I)$ for which there exists $v \in L^2(I)$ such that, for all $\varphi \in \mathcal{C}_0^\infty(I)$,

$$\int_I u(x)\varphi'(x)dx = -\int_I v(x)\varphi(x)dx.$$

For such functions, v is called the weak derivative of u. When u is $\mathcal{C}^1(\bar{I})$, the weak derivative of u is nothing else than its derivative in the classical sense.

For $u \in H^1(I)$ we can read (2.5.3) as

$$\forall \varphi \in \mathcal{C}_0^\infty(I),\ \int_I v(x)\varphi'(x)dx + \int_I u(x)\varphi(x)dx = \int_I f(x)\varphi(x)dx.$$

There is still one problem to solve: there is nothing like the value of an L^2 function at the boundary of I, that is at 0 and 1 here, since those functions are only defined almost everywhere. We need to restrict again our set of possible solutions to a subset of $H^1(I)$.

First of all, one can prove that the space $H^1(I)$, endowed with the scalar product

$$\langle u_1, u_2\rangle_{H^1} = \langle u_1, u_2\rangle_{L^2} + \langle u_1', u_2'\rangle_{L^2},$$

where we have denoted u_j' the weak derivative of u_j, is a Hilbert space. The associated norm is of course

$$\|u\|_{H^1} = \sqrt{\|u\|_{L^2}^2 + \|u'\|_{L^2}^2}.$$

We denote $H_0^1(I)$ the closure of $\mathcal{C}_0^\infty(I)$ in $H^1(I)$, which is then also a Hilbert space. In this particular 1d case, we can easily characterize $H_0^1(I)$.

Proposition 2.5.1 Let $I =]0,1[\subset \mathbb{R}$. If $f \in H^1(I)$, then f is a continuous function on $[0,1]$. The set $H_0^1(I)$ is the subset of f's in $H^1(I)$ such that $f(0) = f(1) = 0$.

Proof: To make some computations clearer, we work with $I =]-a, a[$. For $f \in \mathcal{H}^1(I)$, we have $f' \in L^2(I) \subset L^1(I)$. Thus the function $g : I \to \mathbb{C}$ given by

$$g(x) = \int_{-a}^{x} f'(t)dt$$

is continuous. Morever $g' - f' = 0$ so that $g - f$ is a constant function. Since g can be extended as a continuous function on $[-a, a]$, f too.

The function $x \mapsto |f(x)|$ is continuous on $[-a, a]$, therefore it has a minimum at a point $b \in [-a, a]$. Since

$$2a|f(b)|^2 = \int_{-a}^{a} |f(b)|^2 dt \leq \int_{-a}^{a} |f(t)|^2 dt,$$

we have $\sqrt{2a}|f(b)| \leq \|f\|_{L^2}$. At last, since

$$f(x) = f(b) + \int_b^x f'(t)dt,$$

we get

$$|f(x)| \leq \frac{1}{2\sqrt{a}}\|f\|_{L^2} + \sqrt{2a}\|f'\|_{L^2} \leq C\|f\|_{H^1}.$$

In particular, the linear form δ_x is continuous on $H^1(I)$ for any $x \in [-a, a]$.

Now we have seen that the linear forms $\delta_{\pm a}$ are continuous on $H^1(I)$, and vanishes on $\mathcal{C}_0^\infty(I)$. Thus if $f \in H_0^1(I)$, we have $f(-a) = f(a) = 0$. Conversely, let $f \in \mathcal{H}^1(I)$ such that $f(a) = f(-a) = 0$. Let also g be the function which is equal to f on $[-a, a]$ and is 0 everywhere else on $\mathbb{R}$. We have $g' = f'1_[-a, a]$, so that $g' \in L^2(\mathbb{R})$, and $g \in H^1(\mathbb{R})$. For $\lambda < 1$, the sequence $g_\lambda = g(x/\lambda)$ tends to f in $H^1(I)$ when $\lambda \to 1$, and the support if g_λ is contained in $[-a\lambda, a\lambda] \subset I$. If (χ_ε) is a standard mollifier, $g_\lambda * \chi_\varepsilon$ belongs to $\mathcal{C}_0^\infty(I)$ for any $\varepsilon > 0$ small enough, and converges to g_λ in $H^1(\mathbb{R})$. Thus $g_\lambda \in \mathcal{H}_0^1(I)$ and $f \in \mathcal{H}_0^1(I)$. □

Remark 2.5.2 The orthogonal F of $H_0^1(I)$ in $H^1(I)$ is the subspace of functions u such that

$$-u'' + u = 0$$

in the weak sense. Indeed, the function u belongs to F if and only if for all $\varphi \in \mathcal{H}_0^1(I)$, therefore, by density, if and only if for all $\varphi \in \mathcal{C}_0^\infty(I)$,

$$0 = (u, \bar{\varphi})_{H^1} = \int_I u\varphi dx + \int_I u'\varphi' dx.$$

We are going to prove that, if $V \geq 0$ for example, the problem (2.5.2) above has a unique weak solution in $H_0^1(I)$. To do so, we apply Lax-Milgram's theorem to the sesquilinear form a defined on $H_0^1(I) \times H_0^1(I)$ by

$$a(u, v) = \int_I u'(x)\bar{v}'(x)dx + \int_I V(x)u(x)\bar{v}(x)dx,$$

and the linear form

$$\ell(u) = \int f(x)u(x)dx.$$

Let us prove that they are continuous. For $u \in \mathcal{H}_0^1(I)$, we have clearly

$$|\ell(u)| \leq \|f\|_{L^2}\|u\|_{L^2} \leq \|f\|_{L^2}\|u\|_{H^1}.$$

If moreover $v \in \mathcal{H}_0^1(I)$,

$$|a(u, v)| \leq \|u'\|_{L^2}\|v'\|_{L^2} + \|V\|_{L^\infty}\|u\|_{L^2}\|v\|_{L^2} \leq (1 + \|V\|_{L^\infty})\|u\|_{H^1}\|v\|_{H^1}.$$

Now we verify the coercivity of a. For $u \in \mathcal{H}_0^1(I)$, we have, since $V \geq 0$,

$$|a(u, u)| = \int_I |u'|^2 dx + \int_I V(x)|u(x)|^2 dx \geq \int_I |u'(x)|^2 dx.$$

The coercivity of a thus rely on the following well-known Poincaré inequality that we shall prove in the n-dimensionnal case below (see Proposition 2.5.3). It says that, in the present settings, there exists a constant $C > 0$ such that

$$\int_I |u(x)|^2 dx \leq C \int_I |u'(x)|^2 dx.$$

We go back to our 1d problem, and prove that a is indeed coercive. Using Poincaré's inequality, we obtain

$$|a(u,u)| \geq \int_I |u'(x)|^2 dx \geq \frac{1}{2}\int_I |u'(x)|^2 dx + \frac{1}{2C}\int_I |u(x)|^2 dx \geq c\|u\|^2_{H^1},$$

for $c = \min(1/2, 1/2C)$, which is the estimate we need.

Let us go back to the general Dirichlet problem in $\mathbb{R}^n$, $n \geq 2$. Let Ω be an open, regular bounded subset of $\mathbb{R}^n$. Let also $(a_{ij}(x))_{1\leq i,j\leq n}$ a family of functions in $L^\infty(\Omega, \mathbb{C})$. We suppose that there exists a constant $c > 0$ such that

$$\forall x \in \Omega, \forall \xi \in \mathbb{C}^n,\ c|\xi|^2 \leq \mathrm{Re}(\sum_{i,j} a_{ij}(x)\xi_i\overline{\xi_j}) \leq \frac{1}{c}|\xi|^2$$

Then we denote Δ_a the differential operator defined, for $\varphi \in \mathcal{C}^\infty(\Omega)$, by

$$\Delta_a(\varphi) = \sum_{i,j=1}^n \partial_i(a_{i,j}(x)\partial_j\varphi)$$

Notice that when $A = Id$, Δ_a is nothing else than the usual Laplacian.

The Dirichlet problem on Ω can be stated as follows: for $f \in L^2(\Omega)$, find $u \in L^2(\Omega)$ such that

$$\begin{cases} -\Delta_a u = f, \\ u_{|\partial\Omega} = 0. \end{cases}$$

The case of the equation $-\Delta_a u + Vu = f$ for a non-negative, bounded potential V can be handled the same way, but we choose $V = 0$ for the sake of clarity.

As in dimension 1, we shall work in the Sobolev space $H^1(\Omega)$, which is the space of functions $u \in L^2(\Omega)$ such that, for any $j \in \{1, \ldots, n\}$, there exists an element v_j in $L^2(\Omega)$ with, for any $\psi \in \mathcal{C}^\infty_0(\Omega)$,

$$\int u\partial_j\psi dx = -\int v_j\psi dx.$$

We denote $\nabla u = (v_1, \ldots, v_n)$ the weak gradient of $u \in \mathcal{H}^1(\Omega)$. The space H^1 endowed with the scalar product

$$(u,v)_{H^1} = (u,v)_{L^2} + (\nabla u, \nabla v)_{L^2}$$

is a Hilbert space, and the associated norm is

$$\|\varphi\|_{H^1} = \sqrt{\|\varphi\|^2_{L^2} + \|\nabla\varphi\|^2_{L^2}}.$$

Eventually, the space $H_0^1(\Omega)$ is defined as the closure for the H^1 norm of the vector space $\mathcal{C}_0^\infty(\Omega)$ of smooth, compactly supported functions on Ω:

$$H_0^1 = \{u \in L^2(\Omega),\ \exists (u_n) \subset \mathcal{C}_0^\infty(\Omega), \|u_n - u\|_{H^1} \to 0,\ \text{as } n \to +\infty\}.$$

First, we prove the

Proposition 2.5.3 (Poincaré's inequality) Let $\Omega \subset \mathbb{R}^n$ an open subset, bounded in one direction. There exists a constant $C > 0$ such that

$$\forall u \in \mathcal{H}_0^1(\Omega), \int_\Omega |u|^2 dx \leq C \int_\Omega |\nabla u|^2 dx.$$

Proof: The assumption means that there is an $R > 0$ such that, for example $\Omega \subset \{|x_n| < R\}$. For $\varphi \in \mathcal{C}_0^\infty(\Omega)$, we get

$$\varphi(x', x_n) = \int 1_{[-R, x_n]}(t) \partial_n \varphi(x', t) dt.$$

Using Cauchy-Schwartz inequality we then have,

$$|\varphi(x', x_n)|^2 \leq 2R \int_{-R}^{R} |\partial_n \varphi(x', t)|^2 dt.$$

We integrate this inequality on Ω, and we get

$$\begin{aligned}\int_\Omega |\varphi(x', x_n)|^2 dx &\leq 2R \int_{-R}^{R} \int_{\mathbb{R}^{n-1}} \int_{-R}^{R} |\partial_n \varphi(x', t)|^2 dt dx_n dx' \\ &\leq 4R^2 \int |\partial_n \varphi(x)|^2 dx \leq 4R^2 \int |\nabla \varphi(x)|^2 dx.\end{aligned}$$

The results in $H_0^1(\Omega)$ follows by density. □

Remark 2.5.4 Poincaré's inequality is not true for constant u's. Notice that these functions do not belong to $H_0^1(\Omega)$ for Ω bounded (at least in one direction).

Proposition 2.5.5 Let $\Omega \subset \mathbb{R}^n$ be a bounded open subset. For any $f \in L^2(\Omega)$, the equation $-\Delta_a u = f$ has a unique weak solution in $H_0^1(\Omega)$.

Proof: In the weak sense in $H_0^1(\Omega)$, the equation $-\Delta_a u = f$ means that

$$\forall \varphi \in \mathcal{C}_0^\infty(\Omega),\ \sum_{i,j} \int_\Omega a_{ij}(x) \partial_i u(x) \partial_j \varphi(x) dx = \int_\Omega f \varphi dx.$$

Let us denote $a(v,u)$ the sesquilinear form on $H^1_0(\Omega)\times H^1_0(\Omega)$ given by

$$a(v,u)=\sum_{i,j}\int_\Omega a_{ij}(x)\partial_i v(x)\overline{\partial_j u(x)}dx,$$

and ℓ the linear form on $H^1_0(\Omega)$ given by $\ell(v)=\int fvdx$. The above equation can be written

$$\forall\varphi\in\mathcal{C}^\infty_0(\Omega),\ a(\varphi,\overline{u})=\ell(\varphi),$$

and we want to prove that it has a unique solution $\overline{u}\in\mathcal{H}^1_0(\Omega)$. Thanks to Lax-Milgram's theorem, we only need to prove that a is continuous and coercive.

The continuity comes from the boundedness of the functions a_{ij}, and Cauchy-Schwarz inequality

$$|a(v,u)|\leq\sum_{i,j}\int_\Omega|a_{ij}(x)|\ |\partial_i v(x)||\partial_j u(x)|dx\leq C\sum_{i,j}\|\partial_i v\|_{L^2}\|\partial_j v\|_{L^2}\leq C\|v\|_{H^1}\|u\|_{H^1}.$$

Concerning the coercivity, we have, first for $u\in\mathcal{C}^\infty_0(\Omega)$, then by density for $u\in\mathcal{H}^1_0(\Omega)$,

$$|a(u,u)|\geq|\,\mathsf{Re}\,a(u,u)|=\mathsf{Re}\int_\Omega\left(\sum_{i,j}a_{i,j}\partial_i u\overline{\partial_j u}\right)dx\geq c\int_\Omega\sum_j|\partial_j u|^2dx.$$

It remains to prove that

$$\|\nabla u\|^2_{L^2}:=\int_\Omega\sum_j|\partial_j u|^2dx\geq\|u\|^2_{H^1},$$

which is a consequence of the Poincaré inequality. □

2.A An introduction to the finite elements method

In the previous section, we have seen that Lax-Milgram's theorem permits us to obtain, under suitable assumptions, existence and uniqueness for the solution to the partial differential equation $-\Delta_a u+Vu=f$. As a matter of fact, Lax-Milgram theorem can also be used to obtain approximations of the solution for this equation.

In order to introduce the main ideas, we only consider the 1d case, and the Dirichlet problem for the equation

$$-u''+V(x)u=f(x).$$

on a bounded interval $I=[0,1]\subset\mathbb{R}$. The main idea consists in applying Lax-Milgram's theorem on a finite dimensional subspace G of $H^1_0(]0,1[)$, and construct the corresponding solution. Of course one can expects that the quality of this approximate solution should improve as the dimension of G grows.

We want to find an accurate approximation of the solution $v \in \mathcal{H}_0^1(]0,1[)$ of the problem

$$\forall u \in \mathcal{H}_0^1(]0,1[),\ a(u,v) = \ell(v), \tag{2.A.4}$$

where the sesquilinear form

$$a(u,v) = \int_0^1 u'\bar{v}' + Vu\bar{v}dx \tag{2.A.5}$$

is continuous:

$$|a(u,v)| \leq M\|u\|\ \|v\| \tag{2.A.6}$$

and coercive :

$$c\|u\|^2 \leq |a(u,u)|, \tag{2.A.7}$$

on $H_0^1(]0,1[) \times H_0^1(]0,1[)$.

Let $n \in \mathbb{N}$, and denote $x_0 = 0, x_1 = 1/n, \ldots, x_{n-1} = (n-1)/n, x_n = 1$ the regular subdivision of $[0,1]$ with step $1/n$. We define $n+1$ functions in $\mathcal{C}^0([0,1])$, piecewise linear, by

$$\begin{cases} g_0(0) = 1,\ g_0(x) = 0 \text{ for } x \geq 1/n, \\ g_1(0) = 0,\ g_1(1/n) = 1,\ g_1(x) = 0 \text{ for } x \geq 2/n, \\ \vdots \\ g_j(x) = 0 \text{ for } \ x \leq (j-1)/n,\ g_j(j/n) = 1,\ g_j(x) = 0 \text{ for } x \geq (j+1)/n,\ j = 2, \ldots, n-1 \\ \vdots \\ g_n(x) = 0 \text{ for } \ x \leq (n-1)/n,\ g_n(1) = 1. \end{cases}$$

It is easy to see that the finite elements g_j are linearly independent. Thus they form a basis of the space G_n that they generate, which is the space of continuous functions on $[0,1]$ that are linear on each interval of the form $[j/n, (j+1)/n]$, $j = 0 \ldots n$.

The functions in G_n belong to $H^1(I)$. Indeed they are in $L^2(I)$ since they are continuous, and they are differentiable on $]0,1[$ but perhaps on the set $\{j/n,\ j = 1, \ldots, n-1\}$, which is of measure 0. Moreover their derivative is piecewise constant, therefore belongs to $L^2(]0,1[)$. Using the characterization of $H_0^1(]0,1[)$ in Proposition 2.5.1, we see that the space G_n^0 generated by the finite elements $g_1, g_2 \ldots, g_{n-1}$ is included in $H_0^1(]0,1[)$. In particular, the sesquilinear form a is still continuous and coercive on $G_n^0 \times G_n^0$. Therefore, the problem of finding v such that

$$\forall u \in G_n^0,\ a(u,v) = \int_0^1 fudx \tag{2.A.8}$$

has one and only one solution v_n in G_n^0. What makes this discussion non-void is twofold. First, v_n is a good approximation of the solution to the original problem.

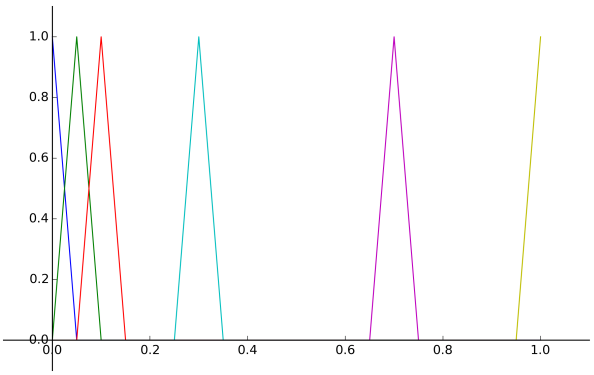

Figure 2.2: Some finite elements g_j for $n = 20$

Proposition 2.A.1 (Céa's Lemma) Let v be the solution of (2.A.8) in $H_0^1(]0,1[)$, and v_n the solution of (2.A.8) in G_n^0. With the constants $M > 0$ and $c > 0$ given in (2.A.6) and (2.A.7) we have

$$\|v - v_n\| \leq \frac{M}{c} \inf_{y \in G_n^0} \|v - y\|.$$

Proof: For any $z \in G_n^0$, we have

$$a(z, v - v_n) = a(z, v) - a(z, v_n) = \ell(z) - \ell(z) = 0.$$

Thus for any $y \in G_n^0$,

$$M\|v - y\| \, \|v - v_n\| \geq |a(v - y, v - v_n)| \geq |a(v - y + y - v_n, v - v_n)| \geq c\|v - v_n\|^2,$$

which proves the lemma. □

This lemma states that, up to the loss $M/c \geq 1$, v_n is the best approximation of u in G_n^0. As a matter of fact since the R.H.S. is not known in general, this result does not seem to give any interesting information. But the idea is, that we may have some a priori estimate on $\|v - y_0\|$ for some well chosen $y_0 \in G_n^0$. A good choice is the function y_0 defined by $y_0(x_j) = v(x_j)$: using this function we can obtain by elementary computations the

Proposition 2.A.2 Let $f \in L^2(I)$, and v the unique solution in H_0^1 of the problem (2.A.4). Let also $v_n \in G_n^0$ the solution of the problem (2.A.8). Then $\|v - v_n\|_{H^1} = \mathcal{O}(\frac{1}{n})$ as $n \to +\infty$.

Second, it is fairly easy to compute v_n (at least with a computer)! Since G_n^0 is spanned by the $(g_j)_{j=1,\dots n-1}$, it is clear that the problem (2.A.8) is equivalent to that of finding $v \in G_n^0$ such that

$$\forall j \in \{1, \dots, n-1\},\ a(g_j, v) = \int_I f g_j dx$$

Now since v_n belongs to G_n^0, we can write

$$v_n = \sum_{k=1}^{n-1} v^k g_k,$$

so that

$$a(g_j, v_n) = \sum_{k=1}^{n-1} \overline{v^k} a(g_j, g_k).$$

Therefore, to compute the coordinates (v_k) of v_n, we only have to solve the $(n-1) \times (n-1)$ linear system

$$AX = B, \text{ with } A = (a(g_j, g_k))_{j,k}, \text{ and } B = (\int_I f g_j dx)_j.$$

Notice that, since supp $g_j \cap$ supp $g_k = \emptyset$ when $|j-k| > 1$, the matrix A is sparse, and in particular tridiagonal.

We have inserted below a small chunk of code in Python that solves the the 1d, second order equation $-u'' + V(x)u = f(x)$ with Dirichlet boundary conditions on $[0,1]$ using the finite elements method.

```
#-------------------------------------------------------------------------
#-------------------------------------------------------------------------
# We solve the equation -u''+Vu=f on [0,1]
# with Dirichlet boundary conditions u(0)=u(1)=0
# using P1 finite elements
# T. Ramond, 2014/06/15
#-------------------------------------------------------------------------

from pylab import *
import numpy as np
from scipy.integrate import quad
from scipy import linalg as la

#-------------------------------------------------------------------------
# Finite elements on [0,1]. Only those that are 0 at 0 and 1.
# numbered from 0 to numpoints-2.

def fe(j,x):
    #print 'j= ', j, ', x= ', x
    N=float(numpoints)
    if (x<j/N):
```

```
        z = 0
    if ((x>=j/N) and (x<=(j+1)/N)):
        z = x*N-j
    if ((x>(j+1)/N) and (x<=(j+2)/N)):
        z =2+j-x*N
    if  (x>(j+2)/N):
        z = 0
    return z

def dfe(j,x):
    N=float(numpoints)
    if (x<j/N):
        z = 0
    if ((x>=j/N) and (x<=(j+1)/N)):
        z = N
    if ((x>(j+1)/N) and (x<=(j+2)/N)):
        z =-N
    if  (x>(j+2)/N):
        z = 0
    return z

#-------------------------------------------------------------------------
# The coefficients of the equation

def f(x):
    return x**2

def V(x):
    return 1

#-------------------------------------------------------------------------
# Some true solutions

# for V(x)=1, f(x)=x**2
def truesolution1(x):
    e=np.exp(1)
    a=(2/e-3)/(e-1/e)
    b=-2-a
    return a*np.exp(x)+(b/np.exp(x)) + x**2+2

# for V(x)=0, f(x)=x**2
def truesolution0(x):
    return x*(1-x**3)/12

#-------------------------------------------------------------------------
# Figure

figure(figsize=(10,6), dpi=80)

#axis
ax = gca()
```

```
ax.spines['right'].set_color('none')
ax.spines['top'].set_color('none')
ax.xaxis.set_ticks_position('bottom')
ax.spines['bottom'].set_position(('data',0))
ax.yaxis.set_ticks_position('left')
ax.spines['left'].set_position(('data',0))
xlim(-.1, 1.1)


#uncomment each line below to draw the finite elements

# Create a new subplot from a grid of 1x2
# subplot(1,2,1)

#plot (X, [fe(0,x) for x in X])
#plot (X, [fe(1,x) for x in X])
#plot (X, [fe(2,x) for x in X])
#plot (X, [dfe(2,x) for x in X])
#plot (X, [fe(6,x) for x in X])
#plot (X, [fe(numpoints-1,x) for x in X])
#plot (X, [fe(numpoints,x) for x in X])

#numpoints=11
#X=linspace(0,1,(numpoints-1)*10)

#for k in range(numpoints-1):
#   plot (X, [fe(k,x) for x in X])

#subplot(1,2,2)

#-------------------------------------------------------------------------
# Solution

# We try different numbers of finite elements.
# For numpoints>16 there seem to be numerical instabilities (?)

for numpoints in range(6,20,5):

    # the matrix A

    A=np.zeros((numpoints-1,numpoints-1))

    def integrand(x,*args):
        return dfe(args[0],x)*dfe(args[1],x)+V(x)*fe(args[0],x)*fe(args[1],x)

    for i in range(numpoints-1):
        for j in range(numpoints-1):
            A[i,j],errA = quad(integrand,0,1, args=(i,j),limit=100)

    # the right-hand side B

```

```
    B=np.zeros(numpoints-1)

    def secondmembre(x,i):
        return fe(i,x)*f(x)

    for j in range(numpoints-1):
        B[j],errB = quad(secondmembre,0,1,args=j)

    # Compute the coordinates of the approximate solution
    # in the finite elements basis

    u=la.solve(A,B)

    # Build the appsolution
    # appsolution(x) = sum u_j*fe_j(x)

    def appsolution(x):
        s=0
        for j in range(numpoints-1):
            s=s+u[j]*fe(j,x)
        return s

    # Plot the approximate solution

    X = np.linspace(0, 1, numpoints, endpoint=True)
    appsolutiongraph=[appsolution(x) for x in X]
    plot (X, appsolutiongraph)

#-----------------------------------------------------------------------
# For comparison: plot the true solution if it is known
# comment lines below if not

Y=linspace(0,1,400)
truesolutiongraph=[truesolution1(x) for x in Y]
plot (Y, truesolutiongraph)

#-----------------------------------------------------------------------
savefig("finite_elements_1d.png",dpi=80)
show()
#-----------------------------------------------------------------------
#-----------------------------------------------------------------------
```

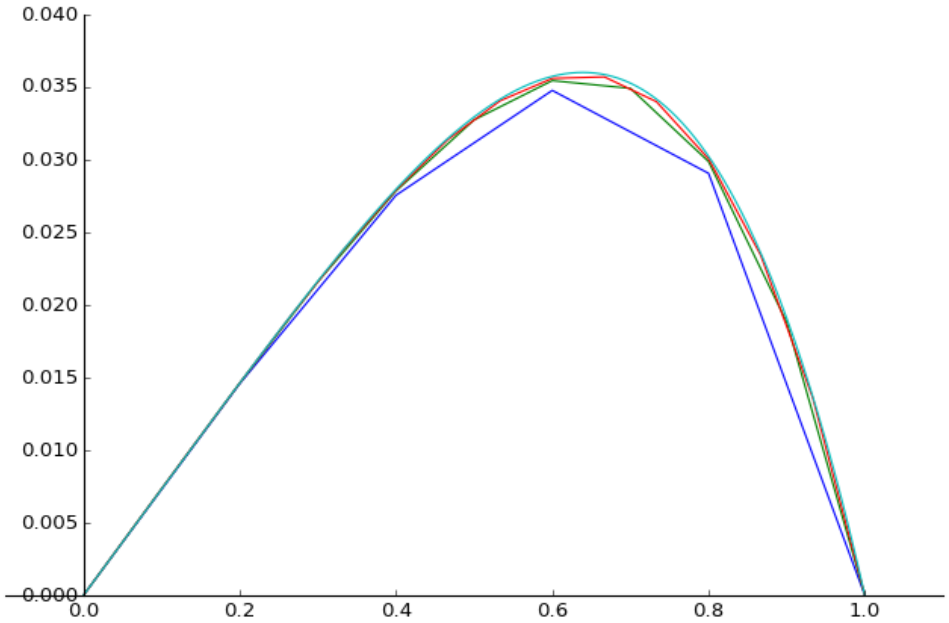

Figure 2.3: Exact and approximate solutions for $-u'' + u = x^2$ on [0,1] with Dirichlet boundary conditions, using P1 finite elements.

Chapter 3

Bounded Operators on Hilbert spaces

Let $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ be a separable Hilbert space on $\mathbb{C}$.

3.1 Definitions

Definition 3.1.1 A bounded operator $T : \mathcal{H} \to \mathcal{H}$ is a linear map such that there exists a constant $C > 0$ satisfying

$$\forall x \in \mathcal{H},\ \|Tx\| \leq C\|x\|.$$

The set of bounded operators on $\mathcal{H}$ is denoted $\mathcal{L}(\mathcal{H})$.

Proposition 3.1.2 A linear operator on $\mathcal{H}$ is bounded if and only if it is continuous.

Proof: If T is bounded, it is 1-Lipschitz, therefore continuous. Conversely, suppose that T is not bounded. There exists a sequence (x_n) such that $\|x_n\| = 1$ and $\|Tx_n\| > n$. Then the sequence (x_n/n) tends to 0, but $\|Tx_n\| \geq 1$, so that T is not continuous. □

For $T \in \mathcal{L}(\mathcal{H})$, the quantity

$$\sup_{x \in \mathcal{H}, \|x\|=1} \|Tx\| = \sup_{x \in \mathcal{H}, x \neq 0} \frac{\|Tx\|}{\|x\|},$$

is finite, and we denote it by $|||T|||$. Notice that $|||T|||$ is also the smallest constant $C \geq 0$ such that the inequality in Definition 3.1.1 holds. It is straightforward to prove the

Proposition 3.1.3 The map $|||\cdot||| : T \mapsto |||T|||$ is a norm on $\mathcal{L}(\mathcal{H})$, and $(\mathcal{L}(\mathcal{H}), |||\cdot|||)$ is a Banach space.

Exercise 3.1.4 Show that for a bounded operator T, one has $\|Tx\| \leq |||T|||\,\|x\|$, and that $|||T_1T_2||| \leq |||T_1|||\,|||T_2|||$ when $T_1, T_2 \in \mathcal{L}(H)$.

Exercise 3.1.5 Let $k \in L^2([0,1]\times[0,1])$, and for $f \in L^2([0,1])$, denote Kf the function given by

$$Kf(x) = \int_0^1 k(x,y)f(y)dy.$$

Show that K is a bounded operator on $\mathcal{H} = L^2([0,1])$. What is its norm?

3.2 Adjoints

Let T be a bounded operator on $\mathcal{H}$, and $y \in \mathcal{H}$ a fixed vector. The map $x \mapsto \langle Tx, y\rangle$ is a continuous, linear form on $\mathcal{H}$, since

$$|\langle Tx, y\rangle| \leq \|Tx\|\,\|y\| \leq |||T|||\,\|x\|\,\|y\|.$$

By Riesz's theorem, there exists $z = z(y) \in \mathcal{H}$ such that, for all $x \in \mathcal{H}$,

$$\langle Tx, y\rangle = \langle x, z\rangle.$$

We shall denote $z = T^*y$, so that we have $\langle Tx, y\rangle = \langle x, T^*y\rangle$.

Proposition 3.2.1 The map $T^* : y \to T^*y$ is a bounded operator on $\mathcal{H}$, $(T^*)^* = T$ and $|||T^*||| = |||T|||$.

Proof: For $y_1, y_2 \in \mathcal{H}$ and $\lambda_1, \lambda_2 \in \mathbb{C}$, we have, for all $x \in \mathcal{H}$,

$$\begin{aligned}\langle x, T^*(\lambda_1 y_1 + \lambda_2 x_2)\rangle &= \langle Tx, \lambda_1 y_1 + \lambda_2 y_2\rangle = \bar\lambda_1\langle Tx, y_1\rangle + \bar\lambda_2\langle Tx, y_2\rangle \\ &= \bar\lambda_1\langle x, T^*y_1\rangle + \bar\lambda_2\langle x, T^*y_2\rangle = \langle x, \lambda_1 T^*y_1 + \lambda_2 T^*y_2\rangle,\end{aligned}$$

which proves that T^* is linear. For $x \in \mathcal{H}$, we have

$$\|T^*x\|^2 = \langle T^*x, T^*x\rangle = \langle TT^*x, x\rangle \leq \|TT^*x\|\,\|x\| \leq |||T|||\,\|T^*x\|\,\|x\|,$$

so that

$$\|T^*x\| \leq |||T|||\,\|x\|.$$

Therefore T^* is bounded, and $|||T^*||| \leq |||T|||$. Now we have $(T^*)^* = T$ since, for all $x, y \in \mathcal{H}$,

$$\langle Tx, y\rangle = \langle x, T^*y\rangle = \overline{\langle T^*y, x\rangle} = \overline{\langle y, (T^*)^*x\rangle} = \langle (T^*)^*x, y\rangle.$$

Thus $|||T||| = |||(T^*)^*||| \leq |||T^*|||$, which finishes the proof that $|||T^*||| = |||T|||$. □

Exercise 3.2.2 Show that $(T_1 + T_2)^* = T_1^* + T_2^*$, $(\lambda T)^* = \bar{\lambda} T^*$ and $(T_1T_2)^* = T_2^*T_1^*$. Show moreover that if T is invertible, so is T^*, and $(T^*)^{-1} = (T^{-1})^*$.

Definition 3.2.3 For $T \in \mathcal{L}(\mathcal{H})$, the bounded operator T^* is called the adjoint of T. When $T = T^*$, we say that T is selfadjoint.

Exercise 3.2.4 Show that the operator K above is selfadjoint if and only if $k(x, y) = \overline{k(y, x)}$. Compare with the case of a self-adjoint (one also says Hermitian) matrix $M \in \mathcal{M}_n(\mathbb{C})$.

Proposition 3.2.5 Let $T \in \mathcal{L}(\mathcal{H})$. We have $\mathrm{Ker}(T^*) = (\mathrm{Ran}\, T)^\perp$.

Proof: A vector x belongs to $\mathrm{Ker}(T^*)$ if and only if $(T^*x, y) = 0$ for any $y \in \mathcal{H}$, that is $(x, Ty) = 0$ for any $y \in \mathcal{H}$, that is again $x \in (\mathrm{Ran}\, T)^\perp$. □

Exercise 3.2.6 Show that if $\lambda \in \mathbb{C}$ is an eigenvalue of a selfadjoint operator, then $\lambda \in \mathbb{R}$.

3.3 Riesz Theorem in Banach spaces

Compact sets in infinite dimensional space can be much more difficult to handle than in the finite dimensional case, where, thanks to Bolzano-Weierstrass theorem, we know that they are the bounded and closed subsets. The following result makes this difference very clear.

Proposition 3.3.1 (Riesz's Theorem) A Banach space $\mathcal{E}$ is of finite dimension if and only if its closed unit ball

$$\overline{B(0,1)} = \{x \in \mathcal{E},\ \|x\| \leq 1\}$$

is compact.

Proof: If $\mathcal{E}$ is of finite dimension, then its closed unit ball, as any bounded closed subset, is compact. Let $\mathcal{E}$ be of infinite dimension, and suppose, by contradiction, that $\overline{B(0,1)}$ is compact. We are going to build a sequence in $\overline{B(0,1)}$ from which we can not extract a convergence subsequence.

Notice first that if $F \subsetneq \mathcal{E}$ is a proper, closed subspace of $\mathcal{E}$, then for any $\varepsilon > 0$, one can find $w \in \mathcal{E}$ such that $\|w\| = 1$ and $d(w, F) \geq 1 - \varepsilon$. Indeed, let $x \in \mathcal{E} \setminus F$, and

$$d = \inf_{y \in F} \|x - y\|$$

be the distance of x to F. Since F is closed, we have $d > 0$. For $\varepsilon > 0$, we can find $x_\varepsilon \in F$ be such that

$$d \leq d(x, x_\varepsilon) \leq \frac{d}{1-\varepsilon}.$$

Now set

$$w = \frac{x - x_\varepsilon}{\|x - x_\varepsilon\|}.$$

We have of course $\|w\| = 1$ and, for any $y \in F$

$$\|w - y\| = \|y - \frac{x - x_\varepsilon}{\|x - x_\varepsilon\|}\| = \frac{1}{d(x, x_\varepsilon)}\|x - x_\varepsilon - \|x - x_\varepsilon\| y\| \geq d \times \frac{1-\varepsilon}{d} \geq 1 - \varepsilon,$$

since $x_\varepsilon - \|x - x_\varepsilon\| y \in F$.

Now since $\mathcal{E}$ is not of finite dimension, there exists a strictly increasing sequence $\mathcal{E}_1 \subsetneq \mathcal{E}_2 \subsetneq \cdots \subsetneq \mathcal{E}_n \subsetneq \mathcal{E}_{n+1} \subsetneq \ldots$, of finite dimensional subspaces of $\mathcal{E}$. Thus we can find a sequence (x_n) with $x_n \in \mathcal{E}_n$ such that $d(x_n, \mathcal{E}_{n-1}) \geq 1/2$. For this sequence we have , for any $p, q \in \mathbb{N}$,

$$d(x_p, x_q) > 1/2,$$

so that no subsequence of (x_n) can converge. This is the required contradiction. □

Exercise 3.3.2 Prove this in two lines for a Hilbert space $\mathcal{H}$, using the orthogonal projection of x on F.

3.4 Weak convergence

Definition 3.4.1 A sequence (x_n) is said to be weakly convergent to x in $\mathcal{H}$ when for any $y \in \mathcal{H}$, the sequence of complex numbers $(\langle x_n, y\rangle)_n$ converges to $\langle x, y\rangle$. In that case we write

$$x_n \rightharpoonup x.$$

If a sequence (x_n) converges to x, it obviously converges weakly to x. When $\mathcal{H}$ is of finite dimension, the converse is also true, therefore this notion is meaningful only for infinite dimensional Hilbert spaces.

It is easy to prove that a convergent sequence is bounded. This is also true for weakly convergent sequences, but it is a equivalent formulation to the principle of uniform boundedness, that we recall now for completeness.

Proposition 3.4.2 (Uniform boundedness principle) Let $\mathcal{E}$ be a Banach space. Let (T_n) be a family of continuous linear operators from $\mathcal{E}$ to some normed space. If for all $x \in \mathcal{E}$, the sequence $(T_n x)$ is bounded, then the sequence $(|||T_n|||)$ is bounded.

Proof: This statement follows from Baire's lemma: a complete metric space X can not be the countable union of closed subsets with empty interior. Let (T_n) be as above, and let us denote, for $p \in \mathbb{N}$,

$$\mathcal{E}_p = \{x \in \mathcal{E},\ \forall n \in \mathbb{N},\ \|T_n x\| \leq p\}.$$

Since for each $x \in \mathcal{E}$, the sequence $(\|T_n x\|)$ is bounded, we have $H = \bigcup_{p\in\mathbb{N}} \mathcal{E}_p$. Thus there exists p_0 such that $\mathring{\mathcal{E}}_{p_0} \neq \emptyset$, that is $x_0 \in \mathcal{E}$ and $r > 0$ such that $\overline{B(x_0, r_0)} \subset \mathcal{E}_{p_0}$. Thus for $u \in \mathcal{E}$ such that $\|u\| \leq r_0$, we have for all $n \in \mathbb{N}$, $\|T_n(x_0 + u)\| \leq p_0$, and

$$\|T_n u\| \leq p_0 + \|T_n x_0\| \leq C_0,$$

since the sequence $(T_n x_0)$ is bounded. This proves that, for any $n \in \mathbb{N}$, we have $|||T_n||| \leq \frac{C_0}{r_0}$. □

Proposition 3.4.3 A weakly convergent sequence in a Hilbert space $\mathcal{H}$ is bounded.

Proof: Let (x_n) be a weakly convergent sequence, and ℓ_n the continuous linear forms defined by $\ell_n(y) = \langle x_n, y\rangle$. The sequence (ℓ_n) satisfies the assumptions of the uniform boundedness principle, therefore there exists $M > 0$ such that

$$\forall n \in \mathbb{N},\ |||\ell_n||| \leq M.$$

i.e.

$$\forall n \in \mathbb{N}, \forall y \in \mathcal{H},\ |\langle x_n, y\rangle| = |\ell_n(y)| \leq M\|y\|$$

In particular for $y = x_n$ we obtain $\|x_n\| \leq M$. □

When $\mathcal{H}$ is not of finite dimension, we have seen that closed and bounded subsets of $\mathcal{H}$ may not be compact. Therefore there exists bounded sequences from which one can not extract a

convergent subsequence. For example, in the Banach space $\mathcal{C}^0([0,1])$ of continuous functions on $[0,1]$, equipped with the norm

$$\|f\|_\infty := \sup_{x\in[0,1]} |f(x)|,$$

the sequence of monomials $(x \mapsto x^n)$ is bounded by 1, but cannot have any other accumulation point than its pointwise limit, that is the function f given by $f(x) = 0$ if $0 \leq x < 1$ and $f(1) = 1$. Since this function f is not continuous, the sequence $(x \mapsto x^n)$ has no convergent subsequence.

The notion of weak convergence can be seen as a remedy for this, as we have the

Proposition 3.4.4 From any bounded sequence in the Hilbert space $\mathcal{H}$, one can extract a weakly convergent subsequence.

This result also holds when $\mathcal{H}$ is only a reflexive Banach space: it is then a consequence of the famous Banach-Alaoglou theorem. For a discussion in that direction, we send the reader for example to the book "Functional Analysis", by W. Rudin. We propose here a proof which is specific to the case of Hilbert spaces.

Proof: Let (x_n) be a bounded sequence in $\mathcal{H}$. For any fixed k, the sequence $((x_k, x_n))_n$ is bounded in $\mathbb{C}$, therefore has a limit point. We denote $(x_{\varphi_0(n)})$ a subsequence of (x_n) such that $((x_0, x_{\varphi_0(n)}))$ converges. Then we denote $(x_{\varphi_1(n)})$ a subsequence of $(x_{\varphi_0(n)})$ such that $(x_1, (x_{\varphi_1(n)}))$ converges, and so on. We set $z_n = x_{\varphi_n(n)}$ (a diagonal procedure), and of course, for all $k \in \mathbb{N}$, $((x_k, z_n))$ converges. We denote F the vector space generated by the x_n. For any $y \in F$, the sequence $((y, z_n))$ converges to some complex number that we denote $\ell(y)$.

If $y \in \overline{F}$, there exists (y_k) a sequence in F such that $(y_k) \to y$. Fix $\varepsilon > 0$. There exists $K \in \mathbb{N}$ such that $\|y - y_k\| \leq \varepsilon/2M$, where $M > 0$ is a upper bound for the (bounded) sequence (z_n). Then there exists $N_\varepsilon \in \mathbb{N}$ such that, for any $n \geq N_\varepsilon$, $|(y_K, z_n)| \leq \varepsilon/2$, and thus

$$|(y, z_n)| \leq |(y_K, z_n)| + |(y - y_K, z_n)| \leq \varepsilon.$$

Therefore, for any $y \in \overline{F}$, the sequence $((y, z_n))$ converges also to some $\ell(y)$. The map

$$\ell : y \in \overline{F} \mapsto \ell(y)$$

is obviously linear, and continuous still since (z_n) is bounded. Therefore, Riesz representation theorem 2.3.1 in the Hilbert space $\overline{F}$ ensures that there exists a unique $x \in \overline{F}$ such that, for all $y \in \overline{F}$,

$$\lim_{n\to+\infty} (y, z_n) = \ell(y) = (y, x).$$

Eventually, for $y \in \mathcal{H}$, we write $y = \Pi_{\overline{F}} y + (I - \Pi_{\overline{F}})y$, where $\Pi_{\overline{F}}$ is the orthogonal projection onto $\overline{F}$, and we have

$$(y, z_n) = (\Pi_{\overline{F}} y, z_n) + ((I - \Pi_{\overline{F}})y, z_n) = (\Pi_{\overline{F}} y, z_n) \to (\Pi_{\overline{F}} y, x) = (y, x),$$

which proves that (z_n) converges weakly. □

3.5 Compact Operators

Definition 3.5.1 A linear operator $T \in \mathcal{L}(\mathcal{H})$ is said to be compact if the image by T of the closed unit ball of H is relatively compact, that is

$$\overline{T(\overline{B(0,1)})} \text{ is compact.}$$

Another way of stating this definition is to say that $T \in \mathcal{L}(\mathcal{H})$ is compact when from any bounded sequence (x_n), one can extract a subsequence (x_{n_k}) such that (Tx_{n_k}) converges.

Notice also that a compact operator is bounded. Indeed, since a compact set is bounded, there exists $M > 0$ such that

$$\forall x \in \overline{B(0,1)}),\ \|Tx\| \leq M,$$

so that

$$\forall x \in \mathcal{H},\ \|Tx\| \leq M\|x\|.$$

Example 3.5.2 If $\mathcal{H}$ is of finite dimension, any linear operator on H is compact. Any operator of finite rank is compact. Indeed, in both cases, $\overline{T(\overline{B(0,1)})}$ is a bounded, closed subset of a finite dimensional vector space, thus a compact set.

Proposition 3.5.3 The set $\mathcal{K}(H)$ of compact operators on $\mathcal{H}$ is a closed subspace of $\mathcal{L}(\mathcal{H})$, and it is a two-sided ideal of $\mathcal{L}(\mathcal{H})$.

Proof: The fact that $\mathcal{K}(H)$ is a subspace of $\mathcal{L}(\mathcal{H})$ follows easily by the characterization of compact operators with bounded sequences. One can also easily see that way that ST and TS are compact operators when T is, and $S \in \mathcal{L}(\mathcal{H})$.

Now let (T_n) be a sequence of compact operators, which converges to $T \in \mathcal{L}(\mathcal{H})$. We want to prove that $T \in \mathcal{K}(H)$. Let (x_n) be a sequence in $\overline{B(0,1)}$. Since T_0 is compact, one can find a subsequence $(x_n^{(0)})$ of (x_n) such that $(T_0 x_n^{(0)})$ converges. Since T_1 is compact, one can find a subsequence $(x_n^{(1)})$ of $(x_n^{(0)})$ such that $(T_1 x_n^{(1)})$ converges. By induction, we can find, for any $k \geq 1$, a subsequence $(x_n^{(k)})$ of $(x_n^{(k-1)})$ such that $(T_k x_n^{(k)})$ converges. Let us denote $(x_{\varphi(n)})$ the sequence given by $x_{\varphi(n)} = x_n^{(n)}$. Of course, $(T_k x_{\varphi(n)})$ converges for all k. For any $p, q \in \mathbb{N}$, we have, for any $k \in \mathbb{N}$,

$$\|Tx_{\varphi(p)} - Tx_{\varphi(q)}\| \leq \|Tx_{\varphi(p)} - T_k x_{\varphi(p)}\| + \|T_k x_{\varphi(p)} - T_k x_{\varphi(q)}\| + \|T_k x_{\varphi(q)} - Tx_{\varphi(q)}\|.$$

For $\varepsilon > 0$, we can find $K \in \mathbb{N}$ such that $\|T - T_K\| \leq \varepsilon/3$. Since the sequence $(T_K x_{\varphi(n)})$ converges, there exists $N_\varepsilon \in \mathbb{N}$ such that, for all $p, q \geq N_\varepsilon$,

$$\|T_K x_{\varphi(p)} - T_K x_{\varphi(q)}\| \leq \varepsilon/3.$$

Thus for any $p, q \geq N_\varepsilon$, we have $\|Tx_{\varphi(p)} - Tx_{\varphi(q)}\| \leq \varepsilon$, and this proves that $(Tx_{\varphi(n)})$ converges, so that T is indeed a compact operator. □

Proposition 3.5.4 Let T be a linear operator on $\mathcal{H}$. The following five properties are equivalent.

i) There exists a sequence (T_n) of finite rank operators on $\mathcal{H}$ such that $\|T_n - T\| \to 0$ as $n \to +\infty$.

ii) T is a compact operator.

iii) $T(\overline{B(0,1)})$ is compact.

iv) For any sequence (x_n) in $\mathcal{H}$ such that $(x_n) \rightharpoonup 0$, we have $(Tx_n) \to 0$.

v) For any orthonormal system (e_n) of $\mathcal{H}$, we have $\|Te_n\| \to 0$.

Proof: – (i) implies (ii) since $\mathcal{K}(H)$ is closed.

– (ii) implies (iii): Suppose (ii). We want to prove that from any sequence (y_n) in $T(\overline{B(0,1)})$, we can extract a convergent subsequence. Let $(x_n) \in \overline{B(0,1)}$ such that $y_n = Tx_n$. Since T is compact, we can extract a subsequence (x_{n_k}) such that (Tx_{n_k}) converges to some $y \in \mathcal{H}$. On the other hand, from Proposition 3.4.4, we can find a subsequence $(x_{n_{k_\ell}})$ which converges weakly to some $x \in \overline{B(0,1)}$. Thus, for any $z \in \mathcal{H}$, we have

$$(Tx_{n_{k_\ell}}, z) = (x_{n_{k_\ell}}, T^*z) \to (x, T^*z) = (Tx, z).$$

Since $Tx_{n_{k_\ell}} \to y$, we have $y = Tx$, so that $y \in T(\overline{B(0,1)})$.

– (iii) implies (iv): Suppose that (x_n) is a sequence converging weakly to 0. We know that there exists $M > 0$ such that

$$\forall n \in \mathbb{N},\ \|x_n\| \leq M.$$

Therefore the sequence (y_n) given by $y_n = x_n/M$ belongs to $\overline{B(0,1)}$, and (Ty_n) is a sequence from the compact subset $T(\overline{B(0,1)})$, therefore possesses limit points. On the other hand, (Ty_n) converges weakly to 0, since, for any $w \in \mathcal{H}$,

$$(w, Ty_n) = (T^*w, y_n) = \frac{1}{M}(T^*w, x_n) \to 0.$$

Thus the only possible limit point ℓ of (Ty_n) is 0, since $(Ty_n, \ell) \to \|\ell\|^2 = 0$, and this proves that (Ty_n), and then (Tx_n) converges to 0.

– (iv) implies (v): Recall that an orthonormal set (e_n) is a set of normed, pairwise orthogonal vectors. For any $y \in \mathcal{H}$, we have

$$\sum_n |(y, e_n)|^2 \leq \|y\|^2$$

Indeed if F_n denotes the space generated by $(e_1, \dots, e_n)$, the vector $y_n = \sum_{k=0}^n (y, e_k) e_k$ is the orthogonal projection of y onto F_n. Therefore $\|y_n\|^2 \leq \|y\|^2$ for all n.

Thus the sequence (e_n) is weakly convergent to 0, so that $\|Te_n\| \to 0$ as $n \to +\infty$.

– (v) implies (i): Suppose that (i) does not hold. Then there exists $\epsilon > 0$ such that $\|T - R\| \geq \varepsilon$ for any finite rank operator R. For $R = 0$ we deduce that $\|T\| \geq \varepsilon$, that is there exists $e_0 \in \mathcal{H}$ such that $\|e_0\| = 1$ and $\|Te_0\| \geq \varepsilon$. Suppose that we have constructed a set $\{e_0, e_1, \dots, e_n\}$ of normed, pairwise orthogonal vectors such that $\|Te_j\| \geq \varepsilon$ for any $j = 1 \dots n$. Denote R_n the projector on the space F_n generated by those vectors. Since TR_n is of finite rank, we have $\|T - TR_n\| \geq \varepsilon$, thus there exists $y_{n+1} \in \mathcal{H}$ such that

$$\varepsilon \|(I - R_n) y_n\| \leq \varepsilon \|y_n\| \leq \|(T - TR_n) y_n\|$$

Therefore, if we set $e_{n+1} = \dfrac{(I - R_n) y_n}{\|(I - R_n) y_n\|}$, we have $\|Te_n\| \geq \varepsilon$ and $e_{n+1} \in F_n^\perp$.

Therefore, by induction, we have build an orthonormal set $\{e_j, j \in \mathbb{N}\}$ for which $\|Te_n\| \geq \varepsilon$ for any n, and this contradicts (v). □

Proposition 3.5.5 If $T \in \mathcal{L}(\mathcal{H})$ is a compact operator, then its adjoint T^* is also compact.

Proof: Let (x_n) be a weakly convergent sequence to 0. Since T^* is continuous, we have

$$T^* x_n \rightharpoonup 0.$$

Thus, since T is compact, $T(T^* x_n) \to 0$. Eventually, since (x_n) is bounded, we have

$$\|T^* x_n\|^2 = (x_n, T(T^* x_n)) \leq \|x_n\| \, \|T(T^* x_n)\| \to 0.$$

□

3.6 Spectrum of self-adjoint compact operators

3.6.1 Definitions

In finite dimensional spaces, a linear map is injective if and only if it is bijective. However, in general, the notion of spectrum of an operator does not coincide with that of the set of its

eigenvalues.

Definition 3.6.1 Let $T \in \mathcal{L}(\mathcal{H})$.

- A complex number λ is an eigenvalue of T when $T - \lambda I$ is not injective.
- A complex number λ is in the spectrum of T when $T - \lambda I$ is not a bijection.

The spectrum of T is often denoted $\sigma(T)$, $\mathrm{sp}(T)$ or $\mathrm{spec}(T)$. The complement $\rho(T) = \mathbb{C} \setminus \sigma(T)$ is called the resolvent set of T.

Notice that, by the open mapping theorem, if $\lambda \in \rho(T)$, then $(T - \lambda I)^{-1}$ is a bounded operator on $\mathcal{H}$.

Example 3.6.2 The right shift operator on $\ell^2(\mathbb{C})$ is injective but not surjective, and the left shift is surjective but not injective. In particular 0 is not an eigenvalue of the right shift, but belongs to its spectrum.

Exercise 3.6.3 Let $T \in \mathcal{L}(L^2(\mathbb{S}^1))$ be the bounded operator defined by

$$T(f)(\theta) = \cos(\theta) f(\theta).$$

It is clear that T has no eigenvalue. Indeed if $T(f) = \lambda f$ then $f = 0$ a.e.. On the other hand, show that $\sigma(T) = [-1, 1]$.

When $\lambda \in \mathbb{C}$ is an eigenvalue of T, the vector space $\mathrm{Ker}(T - \lambda I)$ is called the eigenspace associated to λ. Any non-vanishing element of this eigenspace is called an eigenvector for the eigenvalue λ. The dimension of $\mathrm{Ker}(T - \lambda I)$ is called the geometric multiplicity of the eigenvalue λ. Notice that for matrices (that is, in finite dimension), there is also a notion of for an eigenvalue λ, namely its order as a zero of the characteristic polynomial $P(x) = \det(T - xI)$.

Proposition 3.6.4 The spectrum of a bounded operator is a compact set, included in $\overline{D(0, \|T\|)} \subset \mathbb{C}$.

Proof: First, we are going to prove that the resolvent set $\rho(T)$ is open. So let $z_0 \in \rho(T)$. For $z \in \mathbb{C}$, we have

$$zI - T = (z - z_0)I + z_0 I - T = (z_0 I - T)^{-1}(I + (z - z_0)(z_0 I - T)).$$

Let us denote $M = \|z_0 I - T\|$. For $|z - z_0| < 1/M$, the operator $I + (z - z_0)(z_0 I - T)$ is invertible, with inverse

$$R(z_0, z) = \sum_{k \geq 0} (-1)^k (z - z_0)^k (z_0 I - T)^k.$$

Thus for $|z - z_0| < 1/M$, the operator $zI - T$ is invertible, with inverse $R(z_0, z)(z_0 I - T)$. Therefore $\sigma(T)$ is closed in $\mathbb{C}$.

To end the proof of the proposition, we have to show that $\sigma(T)$ is bounded. But this is obvious since $zI - T = z(I - \frac{1}{z}T)$ is invertible for any z with $|z| > \|T\|$. □

3.6.2 The spectral theorem for self-adjoint compact operators

We study here the spectrum of self-adjoint compact operators, which is the closest case to that of finite self-adjoint matrices.

The following properties of self-adjoint operators, and their proof, does not depend on the dimension, finite or not.

Proposition 3.6.5 Let $T \in \mathcal{L}(\mathcal{H})$, $T \neq 0$, be a self-adjoint operator. Then

i) If $\lambda \in \mathbb{C}$ is an eigenvalue of T, then $\lambda \in \mathbb{R}$.

ii) Two eigenvectors for two different eigenvalues are orthogonal.

iii) If a subspace $F \subset H$ is invariant by T, so is $F^\perp$. This is true in particular for eigenspaces.

Proof: (i) If $Tx = \lambda x$, then

$$\lambda \|x\|^2 = (Tx, x) = (x, Tx) = \overline{\lambda} \|x\|^2,$$

so that $\lambda = \overline{\lambda}$. (ii) If $Tx_1 = \lambda_1 x_1$ and $Tx_2 = \lambda_2 x_2$, we have

$$\lambda_1 (x_1, x_2) = (Tx_1, x_2) = (x_1, Tx_2) = \lambda_2 (x_1, x_2),$$

so that $(x_1, x_2) = 0$ if $\lambda_1 \neq \lambda_2$. (iii) For $y \in F^\perp$, and for any $x \in F$, we have

$$(Ty, x) = (y, Tx) = 0,$$

since Tx belongs to F. Thus $Ty \in F^\perp$. □

Now we turn to operators that are both self-adjoint and compact. We are going to give a precise description of the structure of their spectrum.

Proposition 3.6.6 Let $T \in \mathcal{L}(\mathcal{H})$, $T \neq 0$, be a self-adjoint compact operator. Then any non-zero eigenvalue of T is of finite geometric multiplicity.

Proof: By contradiction, suppose that $E = \mathsf{Ker}(T - \lambda I)$, with $\lambda \neq 0$, has infinite dimension. Then one can find an orthonormal sequence (e_n) in E (by Gram-Schmidt procedure, say), and we have, for $p, q \in \mathbb{N}$,

$$\|Te_p - Te_q\|^2 = |\lambda|^2 \|e_p - e_q\|^2 = 2|\lambda|^2.$$

Therefore no subsequence of (Te_n) can converge, which is absurd since T is compact. □

Of course, if the Hilbert space $\mathcal{H}$ is $\{0\}$, no operator on $\mathcal{H}$ can have an eigenvalue, since there are no non-vanishing vector in $\mathcal{H}$. This is the only case where a self-adjoint compact operator has no eigenvalue, according to the

Proposition 3.6.7 Let $\mathcal{H}$ be a Hilbert space with $H \neq \{0\}$. Suppose $T \in \mathcal{L}(\mathcal{H})$ is a self-adjoint compact operator. Then either $\|T\|$ or $-\|T\|$ is an eigenvalue of T.

Proof: We have seen that $T(\overline{B(0,1)})$ is a compact set. Therefore the continuous function $x \mapsto \|x\|$ has a maximum on this set: there exists $u \in \overline{B(0,1)}$ such that

$$\|Tu\| = \sup_{x \in \overline{B(0,1)}} \|Tx\| = \|T\|.$$

Now take $w \in \mathcal{H}$ such that $(u, w) = 0$ and $\|w\| = 1$. For any $z \in \mathbb{C}$ we have

$$\begin{aligned} \|T\|^2(1 + |z|^2) &\geq (T^2(u + zw), u + zw) \geq \|Tu\|^2 + 2\,\mathsf{Re}(\bar{z}(T^2u, w)) + |z|^2\|Tw\|^2 \\ &\geq \|T\|^2 + 2\,\mathsf{Re}(\bar{z}(T^2u, w)) + |z|^2\|Tw\|^2, \end{aligned}$$

so that for $z \neq 0$,

$$|z|\ \|T\|^2 \geq 2\,\mathsf{Re}(\frac{\bar{z}}{|z|}(T^2u, w)) + |z|\ \|Tw\|^2.$$

Taking successively $z = s$ and then $z = is$ for $s \in \mathbb{R}$, and passing to the limit $s \to 0$ gives $\mathsf{Re}(T^2u, w) = 0$ and $\mathsf{Im}(T^2u, w) = 0$, so that $(T^2u, w) = 0$. Therefore $T^2u \in (\langle u \rangle^\perp)^\perp$, and $T^2u = cu$ for some $c \in \mathbb{C}$. But then

$$\|T\|^2 = \|Tu\|^2 = (T^2u, u) = c\|u\|^2 = c,$$

so that, finally, $T^2u = \|T\|^2 u$. Eventually, we set $v = \|T\|u - Tu$. We have

$$Tv = \|T\|Tu - T^2u = \|T\|Tu - \|T\|^2 u = -\|T\|v,$$

so that either $v = 0$, and $Tu = \|T\|u$, or $v \neq 0$ and this vector is an eigenvector of T for the eigenvalue $-\|T\|$. □

We are now in position to prove the following

Proposition 3.6.8 (The spectral theorem (1)) Let $T \neq 0$ be a self-adjoint compact operator on $\mathcal{H}$. There exists a Hilbertian basis $\{e_n,\ n \in \mathbb{N}\}$ whose elements are eigenvectors of T. The corresponding eigenvalues λ_n are real, and those which are not 0 have finite multiplicity. The spectrum of T contains 0, either as an eigenvalue or not. In any case

$$\sigma(T) = \{0\} \bigcup \{\lambda_n,\ n \in \mathbb{N}\}$$

Finally, 0 is the only accumulation point of $\sigma(T)$.

Proof: Let E_0 be the sum of the eigenspaces associated to $\|T\|$ and $-\|T\|$. It is not reduced to $\{0\}$, it is of finite dimension, and it has an orthonormal basis $\{e_1^{(0)}, \dots, e_{n_0}^{(0)}\}$. Since E_0 is stable by T, so is $H_1 = E_0^\perp$, and we can consider the operator $T_1 \in \mathcal{L}(H_1)$, the restriction of T to H_1. We can then apply the same procedure to T_1, then T_2,... and the procedure stops if and only if at some step N we find a space H_N such that $T|_{H_N} = 0$.

– If it happens, 0 is an eigenvalue of T with infinite multiplicity, and the spectrum of T is the union of $\{0\}$ and a finite set of real, non-vanishing eigenvalues of finite multiplicity. The union of a Hilbertian basis of Ker T and of the $e_j^{(k)}$ for $k \in \{0, \dots, N\}$ and $j \in \{0, \dots, n_k\}$ gives a Hilbertian basis of $\mathcal{H}$. It is also clear that 0 is then the only accumulation point of $\sigma(T)$.

– If it does not happen, let us denote F the closure of the sum of all the eigenspaces E_j. The space F is stable under T, therefore $F^\perp$ is also stable by T. Moreover $T|_{F^\perp}$ has no eigenvalue, which implies that $F^\perp = \{0\}$, so that F is dense in $\mathcal{H}$. In particular the $e_j^{(k)}$ form a Hilbertian basis of $\mathcal{H}$.

Suppose now that $\lambda \neq 0$ is an accumulation point of $\sigma(T)$, and let $(\lambda_n)_n$ a sequence of distincts eigenvalues which converges to λ, and e_n corresponding normed eigenvectors. For any $p, q \in \mathbb{N}$ large enough, we have

$$\|Te_p - Te_q\|^2 = \|\lambda_p e_p - \lambda_q e_q\|^2 = \lambda_p^2 + \lambda_q^2 \geq |\lambda|^2.$$

Thus the sequence (Te_n) has no convergent subsequence, which contradicts the fact that $e_n \in \overline{B(0,1)}$ and T is compact. Now the spectrum of T is bounded and closed, therefore should have a limit point: it can't be any other value than 0. □

The reader may have noticed that we have used the notion of Hilbertian basis of a Hilbert space ($\mathcal{H}$ or Ker T).

Definition 3.6.9 A Hilbertian basis is a set of normed, pairwise orthogonal vectors that is dense in $\mathcal{H}$.

The existence of a Hilbertian basis in a separable Hilbert space can be easily obtained through the well-known Gram-Schmidt orthonormalization procedure.

Exercise 3.6.10 Let T be a self-adjoint, compact and positive operator, that is

$$\forall x \in \mathcal{H},\ (Tx, x) \geq 0.$$

1. Show that the non vanishing eigenvalues of T are positive. We write them as the decreasing sequence

$$\lambda_1 \geq \lambda_2 \geq \dots$$

where the λ_j's are repeated according to their multiplicity.

2. Let E_k be the vector space generated by the eigenvectors $e_1, \dots, e_k$ associated to the eigenvalues $\lambda_1, \dots, \lambda_k$. Show that

$$\lambda_{k+1} = \max_{x \in E_k^{\perp}, \|x\|=1} (Tx, x).$$

3. Show that if F is a vector space of codimension k, $F \cap E_{k+1}$ contains a unit vector.

4. Deduce the Min-Max (or Courant-Fischer) formula: for any $k \geq 0$,

$$\lambda_{k+1} = \min_{\text{codim } F = k} \max_{x \in F, \|x\|=1} (Tx, x)$$

5. Show the same way that for any $k \geq 0$,

$$\lambda_{k+1} = \max_{\dim E = k} \min_{x \in E, \|x\|=1} (Tx, x)$$

3.6.3 The Fredholm alternative

Now we pass to the study of the spectrum of compact operators that are not self-adjoint. We will see that a lot of the spectral theorem for the selfadjoints ones remains true, but the existence of a Hilbertian basis constituted with eigenvectors. The results below come from the study in the half of the 19th century by I. Fredholm of integral equations.

Let us first recall that the codimension of a subspace F of $\mathcal{H}$ is the dimension of any subspace G such that $F \oplus G = H$. In particular, it is the dimension of the quotient space H/F. If $\mathcal{H}$ has finite dimension, then of course $\text{codim } F = \dim H - \dim F$.

Definition 3.6.11 Let $A \in \mathcal{L}(\mathcal{H})$. We say that A is a Fredholm operator (or simply that A is Fredholm) when

i) Its kernel $\mathrm{Ker}(A)$ has finite dimension.

ii) Its range $\mathrm{Ran}(A)$ is closed, and of finite codimension.

The index of A is then defined by $\mathrm{Ind}(A) = \dim \mathrm{Ker}\, A - \mathrm{codim}\, \mathrm{Ran}\, A$.

When $\mathcal{H}$ has finite dimension, any linear operator on $\mathcal{H}$ is Fredholm with index 0.

Proposition 3.6.12 If T is a compact operator on $\mathcal{H}$, then for all $\lambda \neq 0$, $\lambda I - T$ is Fredholm. Moreover

i) $\mathrm{Ran}(\lambda I - T) = \mathrm{Ker}(\overline{\lambda} I - T^*)^\perp$.

ii) $\mathrm{Ker}(\lambda I - T) = \{0\}$ if and only if $\mathrm{Ran}(\lambda I - T) = H$.

iii) $\dim \mathrm{Ker}(\lambda I - T) = \dim \mathrm{Ker}(\overline{\lambda} I - T^*)$

Proof: Suppose $\dim \mathrm{Ker}(\lambda I - T) = +\infty$. Then there exists an infinite Hilbertian basis (e_n) of $\mathrm{Ker}(\lambda I - T)$. But $Te_n = \lambda e_n$, so that

$$\|Te_p - Te_q\| = |\lambda| \|e_p - e_q\| = 2|\lambda|,$$

and no subsequence of (Te_n) can converge, which is a contradiction since $\|e_n\| = 1$ for all n.

To prove that $\mathrm{Ran}(\lambda I - T)$ is closed, we shall prove below that there is a constant $C > 0$ such that

$$\forall x \in \mathrm{Ker}(\lambda I - T)^\perp,\ \|(\lambda I - T)x\| \geq \frac{1}{C}\|x\|. \tag{3.6.1}$$

Suppose for a second that this is true. Let (y_n) be a sequence in $\mathrm{Ran}(\lambda I - T)$ which converges to some $y \in \mathcal{H}$. There exists a sequence (x_n) in $\mathcal{H}$ such that $y_n = (\lambda I - T)x_n$. We can even suppose that $x_n \in \mathrm{Ker}(\lambda I - T)^\perp$, so that

$$\|y_p - y_q\| = \|(\lambda I - T)(x_p - x_q)\| \geq \frac{1}{C}\|x_p - x_q\|.$$

Therefore the sequence (x_n) converges to some $x \in \mathcal{H}$, and $y = (I - T)x$.

Now let us prove (3.6.1). Suppose it is not true. Then there exists a sequence (x_n) in $\mathrm{Ker}(\lambda I - T)^\perp$ such that $\|x_n\| = 1$ and $\|(\lambda I - T)x_n\| \leq \frac{1}{n}$, that is $\lambda x_n - Tx_n \to 0$. From the bounded sequence (x_n) we can extract a weakly convergent subsequence (x_{n_k}), and denoting x its weak limit, we have

$$Tx_{n_k} \to Tx.$$

Thus $\lambda x = Tx$, and $x \in \mathrm{Ker}(\lambda I - T)$, so that

$$0 = (x_{n_k}, x) \to (x, x) = 1,$$

a contradiction. This ends the proof that $\lambda I - T$ is a Fredholm operator.

The point (i) is then easy: we know that $\mathrm{Ker}(A^*)^\perp = \overline{\mathrm{Ran}(A)}$ for any bounded operator A (see Proposition 3.2.5). For $A = \lambda I - T$, this is (i) since $\mathrm{Ran}\, A$ is closed.

Now we prove (ii). Suppose that $\mathrm{Ker}(\lambda I - T) = \{0\}$ but that $\mathrm{Ran}(\lambda I - T) = H_1 \subsetneq H = H_0$. Then for any $n \geq 1$, $H_n = (\lambda I - T)(H_{n-1})$ is a closed subspace, and $H_n \subsetneq H_{n-1}$ since $(\lambda I - T)$ is injective. Now take $x_k \in \mathcal{H}_k \cap H_{k+1}^\perp$ with norm 1. We have, for $p > q$,

$$Tx_p - Tx_q = -(\lambda x_p - Tx_p) + (\lambda x_q - Tx_q) + \lambda(x_p - x_q) = z + x_q$$

with $z \in \mathcal{H}_{q+1}$. Since $x_q \in \mathcal{H}_{k+1}^\perp$, this implies that $\|Tx_p - Tx_q\| \geq 1$. But this is impossible since T is compact.

Conversely, suppose that $\mathrm{Ran}(\lambda I - T) = H$. Then we know by (i) that $\mathrm{Ker}(\overline{\lambda} I - T^*) = \{0\}$. Thus $\mathrm{Ran}(\overline{\lambda} I - T^*)) = H$ by what preceeds. But then $\mathrm{Ker}(\lambda I - T) = \mathrm{Ran}(\overline{\lambda} I - T^*)^\perp = \{0\}$.

We finish with the proof of (iii). First of all, we have $\dim \mathrm{Ker}(\lambda I - T) \geq \dim \mathrm{Ran}(\lambda I - T)^\perp$. Indeed, suppose that this is not true. Then one can find a linear mapping $A : \mathrm{Ker}(\lambda I - T) \to \mathrm{Ran}(\lambda I - T)^\perp$ that is injective but not surjective. We can also extend A to an operator $\tilde{A} : H \to \mathrm{Ran}(\lambda I - T)^\perp$ by setting $\tilde{A}x = 0$ when $x \in \mathrm{Ker}(\lambda I - T)^\perp$. Since $\mathrm{Ker}(\lambda I - T)$ has finite dimension, $\tilde{A}$ has finite range, so $T + \tilde{A}$ is a compact operator. Moreover

$$(\lambda I - (T + \tilde{A}))x = 0 \Rightarrow (\lambda I - T)x = \tilde{A}x \in \mathrm{Ran}(\lambda I - T)^\perp,$$

so that $\mathrm{Ker}(\lambda I - (T + \tilde{A})) = \{0\}$. By (ii), we deduce that $\mathrm{Ran}(\lambda I - (T + \tilde{A}))) = H$. Now for $y \in \mathrm{Ran}(\lambda I - T)^\perp \setminus \mathrm{Ran}\, A$. There exists $x \in \mathcal{H}$ such that

$$(\lambda I - T - \tilde{A})x = y.$$

For any $z = z_1 + z_2 \in \mathcal{H}$, with $z_1 \in \mathrm{Ran}(\lambda I - T)^\perp$ and $z_2 \in \mathrm{Ran}(\lambda I - T)$,

$$(y, z) = (y, z_1) = ((\lambda I - T)x - \tilde{A}x, z_1) = (-\tilde{A}x, z_1) = (-\tilde{A}x, z),$$

thus $y = -\tilde{A}x$, which is impossible.

Now by (i), we have $\dim \mathrm{Ker}(\lambda I - T) \geq \dim \mathrm{Ran}(\lambda I - T)^\perp = \dim \mathrm{Ker}(\overline{\lambda} I - T^*)$, and we get the converse inequality exchanging the roles of T and T^*. □

The following famous result can be easily deduced from this proposition:

Proposition 3.6.13 (Fredholm Alternative) Let $T \in \mathcal{L}(\mathcal{H})$ be a compact operator. Then

- Either $\lambda I - T$ is a bijection: for any $y \in \mathcal{H}$, the equation $\lambda x - Tx = y$ has a unique solution $x \in \mathcal{H}$,
- Or the equation $\lambda x - Tx = 0$ has non-trivial solutions.

In the second case, the equation $\lambda x - Tx = y$ has solutions if and only if $y \in \mathrm{Ker}(\overline{\lambda} I - T^*)^\perp$, and then, the set of solutions is an affine subspace of dimension $\dim \mathrm{Ker}(\overline{\lambda} I - T^*)$.

We can now state the Spectral Theorem for compact operators.

Proposition 3.6.14 (The Spectral Theorem (2)) Let $\mathcal{H}$ be an infinite dimensional Hilbert space. If T is a compact operator on $\mathcal{H}$ then

i) The spectrum of T contains 0.

ii) The spectrum of T but perhaps 0, consists only in eigenvalues of finite multiplicity.

iii) Either $\sigma(T) \subset \{0\}$ is finite, or it is a sequence tending to 0.

Proof: (i) Suppose $0 \notin \sigma(T)$. Then T is invertible, and T^{-1} is bounded. Therefore $I = T \circ T^{-1}$ is compact, since it is the composition of a compact operator with a bounded one. But this can not be true since $\dim H = +\infty$.

(ii) Let $\lambda \in \sigma(T) \setminus \{0\}$. If $\mathrm{Ker}(\lambda I - T) = \{0\}$ then by Fredholm's alternative, $(\lambda I - T)$ is invertible, a contradiction. Therfore λ is an eigenvalue, and we know that $\mathrm{Ker}(\lambda I - T)$ is of finite dimension.

(iii) The proof of the same result for self-adjoint compact operators applies also here. □

Chapter 4

Unbounded operators

Here again, $\mathcal{H}$ is a separable Hilbert space on $\mathbb{C}$.

4.1 Definitions

Definition 4.1.1 An unbounded operator on $\mathcal{H}$ is a pair $(\mathcal{D}, T)$ consisting of a subspace $\mathcal{D}$ of $\mathcal{H}$ and of a linear map $T : \mathcal{D} \to \mathcal{H}$. The space $\mathcal{D}$ is the domain of the unbounded operator.

For example, on the Hilbert space $\mathcal{H} = L^2(\mathbb{R}^n)$ we have the unbounded operators $(\mathcal{C}_0^\infty(\mathbb{R}^n), -\Delta)$, $(\mathcal{C}^2(\mathbb{R}^n), -\Delta)$ or even $(H^2(\mathbb{R}^n), -\Delta)$, where $-\Delta = -\sum_{j=1}^n \partial_j^2$ is the usual Laplacian. Of course these three operators are related. The following definition is meant to clarify this situation:

Definition 4.1.2 If $\mathcal{D}' \subset \mathcal{D}$ and $Tu = T'u$ for all $u \in \mathcal{D}'$, we say that $(\mathcal{D}, T)$ is an extension of $(\mathcal{D}', T')$, and we denote $(\mathcal{D}', T') \subset (\mathcal{D}, T)$.

For example, we have $(\mathcal{C}_0^\infty(\mathbb{R}^n), -\Delta) \subset (\mathcal{C}^2(\mathbb{R}^n), -\Delta) \subset (H^2(\mathbb{R}^n), -\Delta)$.

Definition 4.1.3 An unbounded operator $(\mathcal{D}, T)$ is bounded when the quantity

$$\|T\| = \sup\{\|Tu\|, u \in \mathcal{D}, \|u\| = 1\}$$

is finite.

In that case T is a continuous linear map on $\mathcal{D}$, and if $\mathcal{D}$ is dense in $\mathcal{H}$, T extends uniquely as a bounded operator on $\mathcal{H}$. Unless explicitly stated, we shall always consider unbounded operators $(\mathcal{D}, T)$ with dense domain, and this chapter will mostly deal with unbounded operators with dense domains that are not continuous. What will replace the continuity property is that of closedness.

4.2 Closed operators

Definition 4.2.1 Let $(\mathcal{D}, T)$ be an unbounded operator, and $G = \{(u, Tu),\ u \in \mathcal{D}\}$ its graph. We say that $(\mathcal{D}, T)$ is closed when G is a closed subspace of $\mathcal{H} \times \mathcal{H}$.

Notice that there is no general relation between the closedness of the graph G and of that of its projection $\Pi_1 G$ and $\Pi_2 G$ on each factor of $\mathcal{H} \times \mathcal{H}$ (see Figure 4.1).

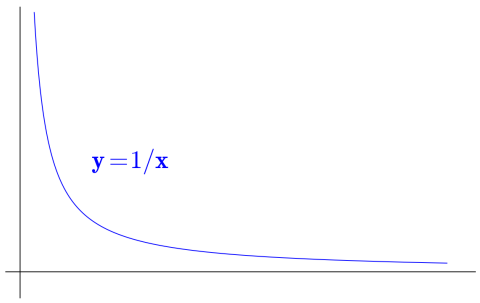

Figure 4.1: A closed graph with two open projections

Examples 4.2.2 – The operator $(\mathcal{C}_0^\infty(\mathbb{R}^n), -\Delta)$ is not closed. Indeed, let us denote G its graph. Let $u \in \mathcal{H}^2(\mathbb{R}^n) \setminus \mathcal{C}_0^\infty(\mathbb{R}^n)$, and let (u_n) be a sequence in $\mathcal{C}_0^\infty(\mathbb{R}^n)$ which tends to u

in $H^2(\mathbb{R}^n)$. Then $-\Delta u_n$ tends to $-\Delta u$ in L^2, so that $((u_n, -\Delta u_n))$ is a sequence in G that converges to $(u, -\Delta u) \in L^2 \times L^2$. However $u \notin \mathcal{C}_0^\infty(\mathbb{R}^n)$, so that $(u, -\Delta u) \notin G$.

– On the other hand, the operator $(H^2(\mathbb{R}^n), -\Delta)$ is closed. Indeed, let us denote G its graph, and let (u_n, v_n) be a sequence in $G \subset H^2(\mathbb{R}^n) \times L^2(\mathbb{R}^n)$ which converges to $(u, v) \in L^2(\mathbb{R}^n) \times L^2(\mathbb{R}^n)$. We have $v_n = -\Delta u_n$, therefore $\hat{v}_n = |\xi|^2 \hat{u}_n$, where $\hat{f}$ denotes the Fourier transform of $f \in L^2(\mathbb{R}^n)$. Since the Fourier transform is an isometry on $L^2(\mathbb{R}^n)$ we have $\hat{v}_n \to \hat{v}$ and $\hat{u}_n \to \hat{u}$ in $L^2(\mathbb{R}^n)$. Therefore $|\xi|^2 \hat{u} = \hat{v}$, which implies that $u \in \mathcal{H}^2(\mathbb{R}^n)$, and $v = -\Delta u$, so that $(u, v) \in G$.

Proposition 4.2.3 An unbounded operator $(\mathcal{D}, T)$ on $\mathcal{H}$ is bounded if and only if $\mathcal{D} = \mathcal{H}$ and $(\mathcal{D}, T)$ is closed.

Proof: If $\mathcal{D} = \mathcal{H}$ and $(\mathcal{D}, T)$ is closed, the closed graph theorem precisely states that $T \in \mathcal{L}(\mathcal{H})$. Conversely suppose $T \in \mathcal{L}(\mathcal{H})$, and let (x_n, y_n) be a sequence in the graph G of T that converges to $(x, y) \in \mathcal{H} \times \mathcal{H}$. We have $y_n = Tx_n$ for all n, therefore $y = Tx$ since T is continuous. Thus G is closed. □

Definition 4.2.4 An unbounded operator is said to be closable when it has a closed extension.

We have just seen that $(\mathcal{C}_0^\infty(\mathbb{R}^n), -\Delta)$ is closable.

Proposition 4.2.5 An unbounded operator is closable if and only if, denoting G its graph, $\overline{G}$ is a graph, i.e. $(u, v_1) \in \overline{G}, (u, v_2) \in \overline{G} \Longrightarrow v_1 = v_2$. Then the operator $(\widetilde{D}, \widetilde{T})$ whose graph is $\overline{G}$ is called the closure of $(\mathcal{D}, T)$.

Proof: Of course if $(\mathcal{D}, T)$ is closable, $(\widetilde{\mathcal{D}}, \widetilde{T})$ is a closed extension of $(\mathcal{D}, T)$. Suppose that there exists a closed operator $(\mathcal{D}', T')$ such that $(\mathcal{D}, T) \subset (\mathcal{D}', T')$. Then the graph $G(T)$ of $(\mathcal{D}, T)$ is included in the graph $G(T')$ of $(\mathcal{D}', T')$, so that we also have $\overline{G(T)} \subset G(T')$, thus $\overline{G(T)}$, as the subgraph of a graph, is a graph. □

Notice that by linearity, an operator is closable if and only if $(0, v) \in \overline{G}$ implies $v = 0$. The closure of a closable operator $(\mathcal{D}, T)$ is of course an extension of $(\mathcal{D}, T)$. Moreover it is the "smallest" closed extension of $(\mathcal{D}, T)$.

Example 4.2.6 The operator $(\mathcal{C}_0^\infty(\mathbb{R}^n), -\Delta)$ is closable, and its closure is $(H^2(\mathbb{R}^n), -\Delta)$. Suppose $(0, v) \in L^2(\mathbb{R}^n) \times L^2(\mathbb{R}^n)$ belongs to $\overline{G}$. Then there is a sequence $((u_n, v_n))$ in $\mathcal{C}_0^\infty(\mathbb{R}^n) \times L^2(\mathbb{R}^n)$ such that $u_n \to 0$, $v_n \to v$ in $L^2(\mathbb{R}^n)$ and $v_n = -\Delta u_n$. Using the Fourier transform as above, we get $v = 0$. This proves that $(\mathcal{C}_0^\infty(\mathbb{R}^n), \Delta)$ is closable. Now we can see the same way that if $(u, v) \in \mathcal{C}_0^\infty(\mathbb{R}^n) \times L^2(\mathbb{R}^n)$ belongs to $\overline{G}$, we have first $u \in L^2$, then $v = -\Delta u \in L^2$, so that $u \in \mathcal{H}^2$. Therefore the closure of $(\mathcal{C}_0^\infty(\mathbb{R}^n), -\Delta)$ is an extension of $(H^2(\mathbb{R}^n), -\Delta)$, which is closed, so is the closure of $(\mathcal{C}_0^\infty(\mathbb{R}^n), -\Delta)$.

Exercise 4.2.7 1. Show that the unbounded operator $(\mathcal{D}, T)$ is closable if and only if, for each $x \in \overline{\mathcal{D}}$ there exists $y \in \mathcal{H}$ such that, for any sequence $(x_n) \subset \mathcal{D}$ such that $x_n \to x$, either the sequence (Tx_n) diverges or it converges to y.

2. Show that, then, the closure of $(\mathcal{D}, T)$ is the unbounded operator $(\widetilde{\mathcal{D}}, \widetilde{T})$ defined by

i) $\widetilde{D}$ is the set of x in $\mathcal{D}$ such that (Tx_n) converges for some (x_n) that converges to x.

ii) For $x \in \widetilde{D}$, $Tx = \lim_{n\to+\infty} Tx_n$ where (x_n) converges to x.

If $(\mathcal{D}, T)$ is an injective unbounded operator, we denote $(\mathcal{D}^{-1}, T^{-1})$ the unbounded operator whose graph is $G^{-1} = \{(x, y) \in \mathcal{H} \times \mathcal{H},\ (y, x) \in G\}$. The operator $(\mathcal{D}^{-1}, T^{-1})$ is called the inverse of $(\mathcal{D}, T)$. Of course $\mathcal{D}^{-1} = \text{Ran}\, T$, and $T^{-1} \circ T$ is the identity operator on $\mathcal{D}$, and $T \circ T^{-1}$ is the identity operator on $\mathcal{D}^{-1}$. Since G^{-1} is closed if and only if G is closed, the operator $(\mathcal{D}^{-1}, T^{-1})$ is closed when $(\mathcal{D}, T)$ is.

4.3 Adjoints

Let $(\mathcal{D}, T)$ be an unbounded operator on $\mathcal{H}$, with dense domain $\mathcal{D}$, and $x \in \mathcal{D}$. There could exist only one $y \in \mathcal{H}$ such that $\forall z \in \mathcal{D}, (x, Tz) = (y, z)$. Indeed if $(y, z) = 0$ for all $z \in \mathcal{D}$, then $y \in \mathcal{D}^\perp = \mathcal{H}^\perp = \{0\}$. The set of x's for which there is one y is denoted $\mathcal{D}^*$, and for $x \in \mathcal{D}^*$, we denote $y = T^*x$. We have defined the adjoint of $(\mathcal{D}, T)$:

Definition 4.3.1 Let $(\mathcal{D}, T)$ be an unbounded operator on $\mathcal{H}$, with dense domain $\mathcal{D}$. The adjoint of $(\mathcal{D}, T)$ is the unbounded operator $(\mathcal{D}^*, T^*)$ given by

i) $\mathcal{D}^* = \{x \in \mathcal{H},\ \exists y \in \mathcal{H}$ such that $\forall z \in \mathcal{D}, (x, Tz) = (y, z)\}$.

ii) For $x \in \mathcal{D}^*$, $T^*x = y$, where y is the element of $\mathcal{H}$ given in the definition of T^*.

Remark 4.3.2 Since $\mathcal{D}$ is dense in $\mathcal{H}$, $\mathcal{D}^*$ is also the set of $x \in \mathcal{H}$ for which there is a constant $C(x) > 0$ such that, for all $z \in \mathcal{D}$, $|\langle x, Tz\rangle| \leq C(x)\|z\|$.

Indeed, for $z \in \mathcal{H}$ and $(z_n) \subset \mathcal{D}$ such that $(z_n) \to z$, the sequence $(\langle x, Tz_n\rangle)$ is a Cauchy sequence in $\mathbb{C}$, and if we denote $\ell_x(z)$ its limit, we can see that ℓ_x is a continuous linear form on $\mathcal{H}$ which extends $z \in \mathcal{D} \mapsto \langle x, Tz\rangle$. Applying Riesz theorem we obtain $\ell_x(z) = (y, z)$ for some unique $y \in \mathcal{H}$.

Proposition 4.3.3 Let $(\mathcal{D}, T)$ be an unbounded operator with dense domain, and (D^*, T^*) its adjoint. We have

$$\mathsf{Ker}\, T^* = (\mathsf{Ran}\, T)^\perp.$$

Proof: Let $x \in \mathsf{Ker}\, T^*$. For $y \in \mathsf{Ran}\, T$ we have, for some $z \in \mathcal{D}$, $(x, y) = (x, Tz) = (T^*x, z) = 0$, so that $x \in (\mathsf{Ran}\, T)^\perp$. Conversely, if $x \in (\mathsf{Ran}\, T)^\perp$, we have $(x, Tz) = 0$ for any $z \in \mathcal{D}$, so that $(x, Tz) = (0, z)$ for any $z \in \mathcal{D}$. Therefore $x \in \mathcal{D}^*$, and $T^*x = 0$. □

Be careful. It is not true in general (that is for operators that are not closed) that any other equality obtained from this one by permutations of $*$ and $\perp$ hold.

Example 4.3.4 The adjoint of $(\mathcal{D}, T) = (\mathcal{C}_0^\infty(\mathbb{R}^d), -\Delta)$ is $(H^2(\mathbb{R}^d), -\Delta)$. For $f \in L^2(\mathbb{R}^d)$, and $\varphi \in \mathcal{C}_0^\infty(\mathbb{R}^d)$, we have $(f, -\Delta\varphi)_{L^2} = \langle -\Delta f, \bar\varphi\rangle$, where we mean the action of the distribution $-\Delta f$ on the test function $\bar\varphi$. But $\varphi \mapsto \langle -\Delta, \bar\varphi\rangle$ extends as a continuous linear form on L^2 if and only if $\Delta f \in L^2$. Thus the domain of the adjoint of $(\mathcal{D}, T)$ is $\mathcal{D}^* = \{u \in L^2, \Delta u \in L^2\} = H^2$. Last, for $f \in \mathcal{H}^2$ and $g \in L^2$ we have $\langle -\Delta f, \bar g\rangle = (-\Delta f, g)_{L^2}$, so that $T^*g = -\Delta g$.

We can also prove this using Fourier transform on L^2 instead of distribution theory. For $f \in L^2(\mathbb{R}^d)$, and $\varphi \in \mathcal{C}_0^\infty(\mathbb{R}^d)$, we have

$$(f, -\Delta\varphi)_{L^2} = (\hat f, \widehat{-\Delta\varphi})_{L^2} = (\hat f, |\xi|^2\hat\varphi)_{L^2} = \int |\xi|^2 \hat f(\xi)\overline{\hat\varphi(\xi)}d\xi.$$

Thus there exists $g \in L^2$ such that $(\hat f, \widehat{-\Delta\varphi})_{L^2} = (g, \varphi)_{L^2} = (\hat g, \hat\varphi)_{L^2}$ for any $\varphi \in \mathcal{C}_0^\infty$ if and only if $\xi \mapsto |\xi|^2\hat f(\xi)$ belongs to L^2, which, since $f \in L^2$, is equivalent to $f \in H^2(\mathbb{R}^n)$. Then $T^*f = g = -\Delta f$.

Exercise 4.3.5 Show that if $(\mathcal{D}_1, T_1) \subset (\mathcal{D}_2, T_2)$, then $(\mathcal{D}_2^*, T_2^*) \subset (\mathcal{D}_1^*, T_1^*)$. Deduce that if $(\mathcal{D}, T) = (\mathcal{D}^*, T^*)$, then $(\mathcal{D}, T)$ has no proper extension.

We are now interested in taking the adjoint of the adjoint of an unbounded operator. It is not clear whether the domain $\mathcal{D}^*$ is dense or not in $\mathcal{H}$. However we have the

Proposition 4.3.6 *Let $(\mathcal{D}, T)$ be an unbounded operator, and G its graph. Let us denote G^* the graph of its adjoint. Then*

$$G^* = J(\overline{G})^\perp, \quad \text{where } J : (u, v) \mapsto (v, -u).$$

In particular $(\mathcal{D}^*, T^*)$ is closed.

Proof: Let $(x, y) \in G^*$, i.e. $x \in \mathcal{D}^*$ and $y = T^*u$. For $(x_0, y_0) \in J(\overline{G})$, there is a sequence $((x_n, y_n))_n \subset G$ such that $(y_n, -x_n) = J(x_n, y_n) \to (x_0, y_0)$. Thus

$$\langle\langle x, y\rangle, (x_0, y_0)\rangle_{\mathcal{H}\times\mathcal{H}} = \lim_{n\to\infty} \langle x, y_n\rangle - \langle y, x_n\rangle = \lim_{n\to\infty} \langle x, Tx_n\rangle - \langle T^*x, x_n\rangle = 0,$$

and $G^* \subset [J(\bar{G})]^\perp$. Conversely, for $(x, y) \in J(\overline{G})^\perp$ and $(x_0, y_0) \in G$, we have $\langle\langle x, y\rangle, (y_0, -x_0)\rangle = 0$, and $\langle x, Tx_0\rangle = \langle y, Tx_0\rangle$, which shows that $x \in \mathcal{D}^*$ and $T^*x = y$. □

Proposition 4.3.7 The space $\mathcal{D}^*$ is dense in $\mathcal{H}$ if and only if $(\mathcal{D}, T)$ is closable. In that case, the adjoint of $(\mathcal{D}^*, T^*)$ is the closure $(\widetilde{\mathcal{D}}, \widetilde{T})$ of $(\mathcal{D}, T)$, that is $(\mathcal{D}^{**}, T^{**}) = (\widetilde{D}, \widetilde{T})$. Moreover $(\widetilde{\mathcal{D}}, \widetilde{T})^* = (\mathcal{D}^*, T^*)$.

Proof: Notice that

$$(0, y_0) \in \bar{G} \Leftrightarrow J(0, y_0) \in J(\bar{G}) \Leftrightarrow \forall \langle x, y\rangle \in G^*, (J(0, y_0), (x, y)) = 0,$$

so that

$$(0, y_0) \in \bar{G} \Leftrightarrow \forall x \in \mathcal{D}^*, \langle y_0, x\rangle = 0.$$

Thus $(0, y_0) \in \bar{G}$ if and only if $y_0 \in (\mathcal{D}^*)^\perp$, and $\mathcal{D}^*$ is dense if and only if T is closable.

At last, for a closable $(\mathcal{D}, T)$, since $J^2 = -I$, Proposition 4.3.6 gives $(G^*)^* = \bar{G}$. □

4.4 Symmetric and selfadjoints unbounded operators

As many operators P from quantum mechanics, the position and momentum operators X_j and Ξ_j are symmetric unbounded operators on $L^2(\mathbb{R}^n)$, that is they satisfy, for $\phi, \psi \in \mathcal{C}_0^\infty(\mathbb{R}^n)$ say,

$$\int P\varphi(x)\overline{\psi(x)}dx = \int \varphi(x)\overline{P\psi(x)}dx.$$

In order to properly define their corresponding time evolution $u(t,x) = e^{-itP/h}u_0(x)$, that is the solution to the Cauchy problem

$$\begin{cases} -ih\partial_t u = Pu(t,x), \\ u(0,x) = u_0, \end{cases}$$

we have to require that P is self-adjoint (Stone's Theorem). It is therefore a natural question to ask if those symmetric operators have a self-adjoint extension, and if they do, if they have a unique one or not.

Definition 4.4.1 *An unbounded operator $(\mathcal{D}, T)$ is symmetric when $(\mathcal{D}, T) \subset (\mathcal{D}^*, T^*)$, that is*

$$\forall x, y \in \mathcal{D},\ (Tx, y) = (x, Ty).$$

Example 4.4.2 The operator $(\mathcal{C}_0^\infty(\mathbb{R}^n), -\Delta)$ on $L^2(\mathbb{R}^n)$ is symmetric, as well as $(\mathcal{C}_0^\infty(\mathbb{R}^n), -\Delta + V(x))$ when $V \in L^\infty(\mathbb{R}^n)$ is a real-valued function. $(\mathcal{C}_0^\infty(\mathbb{R}^n), \partial_j)$ on $L^2(\mathbb{R}^n)$ is not symmetric, but $(\mathcal{C}_0^\infty(\mathbb{R}^n), D_j)$ is, where $D_j = \frac{1}{i}\partial_j$.

Since $(\mathcal{D}^*, T^*)$ is closed, a symmetric operator $(\mathcal{D}, T)$ is closable, and $(\mathcal{D}^*, T^*)$ is an extension of its closure $(\widetilde{\mathcal{D}}, \widetilde{T})$. Notice that $(\widetilde{\mathcal{D}}, \widetilde{T})$ is also symmetric, since $(\widetilde{\mathcal{D}}, \widetilde{T}) \subset (\mathcal{D}^*, T^*) = (\widetilde{\mathcal{D}}, \widetilde{T})^*$.

Definition 4.4.3 An unbounded operator $(\mathcal{D}, T)$ is self-adjoint when $(\mathcal{D}^*, T^*) = (\mathcal{D}, T)$.

Notice that a self-adjoint operator is necessarily symmetric and closed.

Example 4.4.4 The unbounded operator $(H^2(\mathbb{R}^n), -\Delta)$ is self-adjoint. Indeed, we have seen that $(\mathcal{C}_0^\infty(\mathbb{R}^n), -\Delta)$ is closable, with closure $(H^2(\mathbb{R}^n), -\Delta)$. Since, moreover $(\mathcal{C}_0^\infty(\mathbb{R}^n), -\Delta)^* = (H^2(\mathbb{R}^n), -\Delta)$, we have $(H^2(\mathbb{R}^n), -\Delta)^* = (\mathcal{C}_0^\infty(\mathbb{R}^n), -\Delta)^{**} = (H^2(\mathbb{R}^n), -\Delta)$.

Proposition 4.4.5 Let $(\mathcal{D}, T)$ be a symmetric operator. The following assertions are equivalent:

i) $(\mathcal{D}, T)$ is self-adjoint.

ii) $(\mathcal{D}, T)$ is closed and $\mathrm{Ker}(T^* + i) = \mathrm{Ker}(T^* - i) = \{0\}$.

iii) $\mathrm{Ran}(T + i) = \mathrm{Ran}(T - i) = \mathcal{H}$.

Proof: (i) implies (ii): If $(\mathcal{D}, T)$ is self-adjoint then $(\mathcal{D}, T) = (\mathcal{D}^*, T^*)$ so that it is a closed operator. Morever if $\lambda \in \mathbb{C}$ is an eigenvalue of $(\mathcal{D}^*, T^*)$, then, for some $u \neq 0$,

$$\lambda(u,u) = (\lambda u, u) = (T^* u, u) = (u, T^* u) = (u, \lambda u) = \overline{\lambda}(u,u),$$

so that $\lambda \in \mathbb{R}$. Thus $\mathrm{Ker}(T^* + z) = \{0\}$ whenever $z \notin \mathbb{R}$.

(ii) implies (iii): We know that $\mathrm{Ker}((T \pm i)^*) = \mathrm{Ran}(T \pm i)^\perp$. Thus $\overline{\mathrm{Ran}(T \pm i)} = \mathcal{H}$, and the result follows if we show that $\mathrm{Ran}(T \pm i)$ is closed. First, since T is symmetric, for any $x \in \mathcal{D}$ we have

$$\|(T \pm i)x\|^2 = \|Tx\|^2 + (Tx, \pm ix) + (\pm ix, Tx) + \|x\|^2 = \|Tx\|^2 + \|x\|^2.$$

Now let $(y_n) \subset \mathrm{Ran}(T \pm i)$ be a sequence that converges to $y \in \mathcal{H}$. There exists $(x_n) \subset \mathcal{D}$ such that $y_n = (T \pm i)x_n$. From the above inequality we have

$$\|x_p - x_q\| \leq \|(T \pm i)x_p - (T \pm i)x_p\| \leq \|y_p - y_q\|,$$

so that (x_n) is a Cauchy sequence. Denoting x its limit we have $(x_n, y_n) \to \langle x, y \rangle$, and since $(\mathcal{D}, T)$ is closed, $y = (T \pm i)x$ and $y \in \mathrm{Ran}(T \pm i)$.

(iii) implies (i): Let $x \in \mathcal{D}^*$. We want to prove that $x \in \mathcal{D}$. Since $\mathrm{Ran}(T - i) = H$, there exists $y \in \mathcal{D}$ such that $(T - i)y = (T^* - i)x$. Since $(\mathcal{D}, T) \subset (\mathcal{D}^*, T^*)$, we have $y \in \mathcal{D}^*$ and $Ty = T^* y$. Thus $(T^* - i)y = (T^* - i)x$ and $x - y \in \mathrm{Ker}(T^* - i) = \mathrm{Ran}(T + i)^\perp = \{0\}$, so that $x = y \in \mathcal{D}$. □

4.5 Essential self-adjointness

For symmetric operators that are not closed (remember that they are closable), we have also the following important

Definition 4.5.1 A symmetric unbounded operator $(\mathcal{D}, T)$ is essentially self-adjoint when its closure $(\widetilde{\mathcal{D}}, \widetilde{T})$ is self-adjoint.

We have seen that $(\mathcal{C}_0^\infty(\mathbb{R}^d), -\Delta)$ is essentially self-adjoint.

Proposition 4.5.2 If a symmetric unbounded operator $(\mathcal{D}, T)$ is essentially self-adjoint, then it has a unique self-adjoint extension.

Proof: If $(\mathcal{D}', T')$ is a self-adjoint extension of $(\mathcal{D}, T)$, we have $(\widetilde{\mathcal{D}}, \widetilde{T}) \subset (\mathcal{D}', T')$ since $(\mathcal{D}', T')$ is closed. Taking adjoints we get

$$(\mathcal{D}', T') = (\mathcal{D}', T')^* \subset (\widetilde{\mathcal{D}}, \widetilde{T})^* = (\widetilde{\mathcal{D}}, \widetilde{T}).$$

□

Notice that to prove that a symmetric operator $(\mathcal{D}, T)$ is essentially self-adjoint, one may apply Proposition 4.4.5 to its closure $(\widetilde{\mathcal{D}}, \widetilde{T})$, and get the

Proposition 4.5.3 Let $(\mathcal{D}, T)$ be a symmetric unbounded operator. The following properties are equivalent

i) $(\mathcal{D}, T)$ is essentially self-adjoint.

ii) $\mathrm{Ker}(T^* - i) = \mathrm{Ker}(T^* + i) = \{0\}$.

iii) $\mathrm{Ran}(T + i)$ and $\mathrm{Ran}(T - i)$ are dense in $\mathcal{H}$.

Exercise 4.5.4 Prove it.

4.6 Spectrum and resolvent

4.6.1 Spectrum

Definition 4.6.1 Let $(\mathcal{D}, T)$ be an unbounded operator with dense domain. The resolvent set $\rho(T)$ is the set of $z \in \mathbb{C}$ such that $(T - zI) : \mathcal{D} \to \mathcal{H}$ is a bjection, and $(T - zI)^{-1} : \mathcal{H} \to \mathcal{H}$ is a bounded operator.

For a bounded operator $T \in \mathcal{L}(\mathcal{H})$, the open mapping theorem implies that, if $(T - zI) : \mathcal{H} \to \mathcal{H}$ is a bijection, its inverse is automatically bounded. Therefore, as we have already seen, the resolvent set of a bounded operator is the set of $z \in \mathbb{C}$ such that $(T - zI) : \mathcal{D} \to \mathcal{H}$ is a bijection. We recall also that $\sigma(T) \subset \overline{D(0, \|T\|)}$.

Proposition 4.6.2 If the unbounded operator $(\mathcal{D}, T)$ is not closed, then its spectrum is the whole complex plane $\mathbb{C}$.

Proof: If there exists $z \in \rho(T)$, the operator $(T - zI)^{-1}$ belong to $\mathcal{L}(\mathcal{H})$, therefore is closed by Proposition 4.2.3. We have seen at the end of Section 4.2 that this implies that $T - zI$, the inverse of $(T - zI)^{-1}$, is closed, and so is T. □

We introduce now some important subsets of the spectrum of an unbounded operator $(\mathcal{D}, T)$. For $z \in \mathbb{C}$, we have the following possibilities:

- Either $T - zI$ is not injective. Then $z \in \sigma(T)$, and we say that z is an eigenvalue for $(\mathcal{D}, T)$, with associated eigenspace Ker$(T - zI)$. The point spectrum $\sigma_p(T)$ of $(\mathcal{D}, T)$ is the set of eigenvalues of T.
- Or $T - zI$ is injective. Then
  - Either $T - zI$ is not surjective. Then $z \in \sigma(T)$ and
    * Either Ran$(T - zI)$ is dense in $\mathcal{H}$, and we say that z belongs to the continuous spectrum $\sigma_c(T)$ of $(\mathcal{D}, T)$
    * Or Ran$(T - zI)$ is not dense in $\mathcal{H}$, and we say that z belongs to the residual spectrum $\sigma_r(T)$ of $(\mathcal{D}, T)$
  - Or $T - zI$ is surjective. Then $T - zI$ is a bijection and
    * Either $(T - zI)^{-1}$ is not bounded, and $z \in \sigma(T)$. We denote $\sigma'(T)$ the set of such points.
    * Or $(T - zI)^{-1}$ is bounded, and $z \in \rho(T)$.

Notice that for a closed operator $(\mathcal{D}, T)$, we have $\sigma'(T) = \emptyset$. Indeed since $(T - zI)$ is a closed operator, so is $(T - zI)^{-1}$. Since its domain is $\mathcal{H}$, we have seen that $(T - zI)^{-1}$ is automatically bounded.

From now on, we will only consider the spectrum of closed, densely defined unbounded operators $(\mathcal{D}, T)$. Therefore, in particular, we will always have

$$\sigma(T) = \sigma_p(T) \cup \sigma_c(T) \cup \sigma_r(T).$$

4.6.2 The Resolvent

Let $(\mathcal{D}, T)$ be a closed densely defined unbounded operator.

Definition 4.6.3 The map

$$\mathcal{R}_T : z \in \rho(T) \mapsto (T - zI)^{-1} \in \mathcal{L}(\mathcal{H})$$

is called the resolvent of T.

Proposition 4.6.4 The resolvent $\mathcal{R}_T(z)$ of $(\mathcal{D}, T)$ has the following properties

i) $\rho(T^*) = \overline{\rho(T)}$, and $\overline{\mathcal{R}_T(z)} = \mathcal{R}_{T^*}(\bar{z})$,

ii) For $z, z' \in \rho(T)$, $\mathcal{R}_T(z) - \mathcal{R}_T(z') = (z - z')\mathcal{R}_T(z)\mathcal{R}_T(z')$.

iii) For $z, z' \in \rho(T)$, $\mathcal{R}_T(z)\mathcal{R}_T(z') = \mathcal{R}_T(z')\mathcal{R}_T(z)$.

Proof: (i) comes from the relation $\mathrm{Ker}(T^*) = \mathrm{Ran}(T)^\perp$, and from the fact that $(T^*)^{-1} = (T^{-1})^*$.

(ii) is often called "the first resolvent formula ". It stems from a direct computation

$$\begin{aligned}(T-z)^{-1} - (z-z')(T-z)^{-1}(T-z')^{-1} &= (T-z)^{-1}[I - (z-z')(T-z')^{-1}] \\ &= (T-z)^{-1}[I - (z - T + T - z')(T-z')^{-1}] = (T-z')^{-1}.\end{aligned}$$

(iii) follows directly from (ii). □

Proposition 4.6.5 *The resolvent set $\rho(T)$ is open in $\mathbb{C}$, and $\mathcal{R}_T$ is holomorphic on $\rho(T)$. Moreover for $z \in \rho(T)$ we have*

$$\frac{1}{\mathrm{dist}(z, \sigma(T))} \leq \|\mathcal{R}_T(z)\|. \tag{4.6.1}$$

Proof: Let $z_0 \in \rho(T)$. For $z \in \rho(T)$ the first resolvent identity gives, with $\mathcal{R}(z) = \mathcal{R}_T(z)$,

$$\mathcal{R}(z) = \mathcal{R}(z_0) + (z - z_0)\mathcal{R}(z).$$

Iterating, we get by induction that for any $n \in \mathbb{N}$,

$$\mathcal{R}(z) = \sum_{j=0}^{n} (z - z_0)^j \mathcal{R}(z_0)^{j+1} + (z - z_0)^{n+1}\mathcal{R}(z_0)^{n+1}\mathcal{R}(z).$$

Thus, for $z \in \mathbb{C}$ such that $|z - z_0| \leq \|\mathcal{R}(z_0)\|^{-1}$, we set $\mathcal{S}(z) = \sum_{j=0}^{\infty} (z - z_0)^j \mathcal{R}(z_0)^{j+1}$. As the sum of a norm convergent entire series in $\mathcal{L}(\mathcal{H})$, $\mathcal{S}$ is an analytic function in the disk $D(z_0, \|\mathcal{R}(z_0)\|^{-1})$. Moreover

$$\begin{aligned}\mathcal{S}(z)(T-z) = \sum_{j=0}^{\infty} (z-z_0)^j \mathcal{R}(z_0)^{j+1}(T-z) &= \sum_{j=0}^{\infty} (z-z_0)^j \mathcal{R}(z_0)^{j+1}(T - z_0 + z_0 - z) \\ \sum_{j=0}^{\infty} (z-z_0)^j \mathcal{R}(z_0)^j - \sum_{j=0}^{\infty} (z-z_0)^{j+1}\mathcal{R}(z_0)^{j+1} &= I,\end{aligned}$$

and we also have $(T-z)S(z)=I$. At last, for $z_0\in\rho(T)$, if $z\in D(z_0,\|\mathcal{R}(z_0)\|^{-1})$, then $z\in\rho(T)$. Therefore, for any $z\in\sigma(T)$ we have $z\notin D(z_0,\|\mathcal{R}(z_0)\|^{-1})$, or dist$(z_0,z)\geq\|\mathcal{R}(z_0)\|^{-1}$, from which (4.6.1) follows. □

If there is a sequence (ψ_n) in $\mathcal{D}$ such that $\|\psi_n\|=1$ and $\|(T-z)\psi_n\|\to 0$, then z belongs to the spectrum of T. Indeed, if such a sequence exists, z can not belong to $\rho(T)$, otherwise we would have

$$1=\|\psi_n\|=\|R_T(z)(T-z)\psi_n\|\leq C\|(T-z)\psi_n\|\to 0,$$

a contradiction. It follows From the estimate (4.6.1) that this assertion is an equivalence for complex number z that are on the boundary of the spectrum:

Proposition 4.6.6 When $z\in\partial\sigma(T)$, there is a sequence (ψ_n) in $\mathcal{D}$ such that $\|\psi_n\|=1$ and $\|(T-z)\psi_n\|\to 0$.

Proof: Suppose that $z\in\partial\sigma(T)$. Let (z_n) be a sequence in $\rho(T)$ such that dist$(z_n,z)=\frac{1}{n}$. From (4.6.1), there is a sequence $\tilde{\phi}_n$ such that $\|\tilde{\phi}_n\|=1$ and $\|R(z_n)\phi_n\|\geq n$. We set $\phi_n=\tilde{\phi}_n/\|R(z_n)\tilde{\phi}_n\|$. Then $\|\phi_n\|\to 0$ and if $\psi_n=R(z_n)\phi_n$ we have $\|\psi_n\|=1$ and

$$(T-z)\psi_n=(T-z_n)\psi_n+(z_n-z)\psi_n=\phi_n+(z_n-z)\psi_n.$$

Therefore $\|(T-z)\psi_n\|\to 0$. □

4.6.3 The case of selfadjoints unbounded operators

Proposition 4.6.7 Let $(\mathcal{D},T)$ be a closed symmetric operator. Then T is selfadjoint if and only if $\sigma(T)\subset\mathbb{R}$.

Proof: Suppose that $\sigma(T)\subset\mathbb{R}$. Then Ran$(T\pm i)=\mathcal{H}$, so that T is selfadjoint from Proposition 4.4.5. Conversely if $(\mathcal{D},T)$ is selfadjoint, then for $z\in\mathbb{C}\setminus\mathbb{R}$, $(T-z)$ is 1 to 1 since if $z=x+iy\in\mathbb{C}$, we have $\|(T-z)u\|^2=\|(T-x)u\|^2+y^2\|u\|^2\geq y^2\|u\|^2$. Since Ran$(T-z)^\perp=$ Ker$(T-\bar{z})$, we obtain that for $y\neq 0$, Ran$(T-z)$ is dense in $\mathcal{H}$. Still for $y\neq 0$ we get also that $\|(T-z)^{-1}\|\leq 1/|y|$. Thus $\mathbb{C}\setminus\mathbb{R}\subset\rho(T)$. □

Since $\sigma(T)\subset\mathbb{R}$, any element in the spectrum of a selfadjoint operator belongs to its boundary. Thus, we have the following criterion:

Proposition 4.6.8 Let $(\mathcal{D}, T)$ be a selfadjoint operator. $z \in \mathbb{C}$ belongs to the spectrum of T if and only if there is a sequence (ψ_n) in $\mathcal{D}$ such that $\|\psi_n\| = 1$ and $\|(T - z)\psi_n\| \to 0$.

4.7 The spectral theorem for selfadjoint unbounded operators

4.7.1 More on compact selfadjoint operators

Let $K \in \mathcal{L}(\mathcal{H})$ be a compact and selfadjoint operator. We know that there exist a sequence of subspaces $\mathcal{H}_k$ of finite dimension and pairwise orthogonal, and a bounded sequence $(\lambda_k)_k$ of real numbers such that

$$\mathcal{H} = \oplus_{k\in\mathbb{N}} \mathcal{H}_k,$$

and for $u \in \mathcal{H}_k$, $Ku = \lambda_k u$. We denote Π_k the orthogonal projector onto $\mathcal{H}_k$, and E_λ the orthogonal projector onto

$$\mathcal{G}_\lambda = \oplus_{\lambda_k \leq \lambda} \mathcal{H}_k.$$

The family (E_λ) is a spectral family in the following sense:

Definition 4.7.1 A family $(E_\lambda)_{\lambda\in\mathbb{R}}$ of orthogonal projectors on $\mathcal{H}$ is called a spectral family when

i) For all $u \in \mathcal{H}$, $E_\lambda u \to 0$ as $\lambda \to -\infty$, and $E_\lambda u \to u$ as $\lambda \to +\infty$.

ii) For all $\lambda, \mu \in \mathbb{R}$, $E_\lambda E_\mu = E_{\min(\lambda,\mu)}$.

iii) For all $u \in \mathcal{H}$, $E_\lambda u \to E_{\lambda_0} u$ as $\lambda \to \lambda_0^+$.

Indeed, the property (i) comes from the fact that the sequence (λ_k) is bounded, so that $E_\lambda = 0$ for $\lambda < \min \lambda_k$, and $E_\lambda = I$ for $\lambda > \max \lambda_k$. The inclusion $\mathcal{G}_\lambda \subset \mathcal{G}_\mu$ for $\lambda < \mu$ implies (ii). At last if $\lambda_0 \neq 0$, then it is an isolated point in $\sigma(K)$. Thus here exists $\varepsilon > 0$ such that $\mathcal{G}_\lambda = \mathcal{G}_{\lambda_0}$ for $\lambda \in [\lambda_0, \lambda_0 + \varepsilon[$, which impies (iii) for $\lambda_0 \neq 0$. For $\lambda_0 = 0$, it may happen that the eigenvalues of K in a right neighborhood of 0 form a sequence (λ_n), that we can suppose to be decreasing, of eigenvalues such that $\lambda_n > 0$ and $\lambda_n \to 0$ as $n \to +\infty$. If it is not the case, then (iii) follows as in the case $\lambda_0 \neq 0$. Otherwise, for $u \in \mathcal{H}$, if $\lambda \in [\lambda_0, \lambda_n[$,

$$E_\lambda u - E_{\lambda_0} u = \sum_{k>n} \Pi_{\lambda_k} u = \sum_{k>n} \sum_{j=1}^{\dim \mathcal{H}_k} (u, e_j^k) e_j^k,$$

where $(e_j^k)_j$ is a orthonormal basis of $\mathcal{H}_k$. As the rest of a convergent series, the R.H.S. goes to 0 as $n \to +\infty$, and this implies (iii) in that case.

Notice that for any $u, v \in \mathcal{H}$, the function $\lambda \mapsto (E_\lambda u, v)$ is continuous from $\mathbb{R}$ to $\mathbb{R}$, but at the points λ_k, where it is only right continuous and we have

$$\lim_{\lambda \to \lambda_k^-} (E_\lambda u, v) = (E_{\lambda_k} - \Pi_k u, v).$$

In particular, the distributional derivative of the function $\lambda \mapsto (E_\lambda u, v)$ is the compactly supported measure

$$d(E_\lambda u, v) = \sum_k (\Pi_k u, v) \delta_{\lambda_k},$$

and, thus,

$$(u, v) = \sum_k (\Pi_k u, v) = \langle d(E_\lambda u, v), 1 \rangle = \int 1 \, d(E_\lambda u, v) = \int d(E_\lambda u, v),$$

where the last notation -the usual one- comes from the "before Laurent Schwarz" period. The same way, we get

$$(Ku, v) = \sum \lambda_k (\Pi_k u, v) = \langle d(E_\lambda u, v), \lambda \rangle = \int \lambda d(E_\lambda u, v).$$

One may notice that v plays no role in these formula.Therefore they are most often written as

$$u = \int dE_\lambda u \qquad \text{and} \qquad Ku = \int \lambda dE_\lambda u,$$

or even

$$I = \int dE_\lambda \qquad \text{and} \qquad K = \int \lambda dE_\lambda.$$

Then it is very natural to define functions $f(K)$ of compact operators, through the formula

$$f(K) = \int f(\lambda) dE_\lambda.$$

If we stay at the level of basic distributions theory, this formula make sense for $\mathcal{C}^\infty$ functions only. We may further notice that dE_λ (more precisely $d(E_\lambda u, v)$) is a distribution of order 0, so that this functional calculus is well defined for functions that are only continuous. As a matter of fact, one can also define $f(K)$ for measurable functions through the theory of Stieljes integral.

4.7.2 The general case

One of the most important fact of the theory, is that a lot of what we have said for compact self-adjoint operators also holds for unbounded self-adjoint operators.

Proposition 4.7.2 Let $(\mathcal{D}, T)$ be an unbounded selfadjoint operator. There exists a spectral family $(E_\lambda)_{\lambda\in\mathbb{R}}$ such that

i) $\mathcal{D} = \{u \in \mathcal{H}, \int \lambda^2 d(E_\lambda u, u) < \infty\}$,

ii) for all $u, v \in \mathcal{D}$, $(Tu, v) = \int \lambda d(E_\lambda u, v)$.

We have put the relatively long proof of this result in an appendix to this chapter, and in the rest of this chapter, we concentrate on some of its consequences.

Many spectral properties of $(\mathcal{D}, T)$ can be recovered from the knowledge of its spectral family. For example, using Proposition 4.6.8, one can get the

Proposition 4.7.3 Let $(\mathcal{D}, T)$ be a selfadjoint unbounded operator, and (E_λ) its spectral family. Then $\lambda_0 \in \mathbb{R}$ belongs to $\sigma(T)$ if and only if, for any $\varepsilon > 0$,

$$E(]\lambda_0 - \varepsilon, \lambda_0 + \varepsilon[) := \int_{\lambda_0-\varepsilon}^{\lambda_0+\varepsilon} dE_\lambda = E_{\lambda_0+\varepsilon} - E_{\lambda_0-\varepsilon} \neq 0.$$

Exercise 4.7.4 Prove it!

From the spectral theorem, we can deduce also an important inequality, namely

Proposition 4.7.5 Let $(\mathcal{D}, T)$ be a selfadjoint operator. For $u \in \mathcal{D}$ and $z \in \mathbb{R}$, we have

$$\text{dist}(z, \sigma(T))\|u\| \leq \|(T - z)u\|. \tag{4.7.2}$$

Proof: We have

$$\|(T - z)u\|^2 = ((T - z)u, (T - z)u) = \int (\lambda - z) d(E_\lambda u, (T - z)u)$$

and

$$\begin{aligned}\overline{(E_\lambda u,(T-z)u)} = ((T-z)u,E_\lambda u) &= \int_{\mu\in\mathbb{R}} (\mu-z)d(E_\mu u,E_\lambda u)\\ &= \int_{\mu\in\mathbb{R}} (\mu-z)d(E_\lambda E_\mu u,u) = \int_{\mu\in\mathbb{R}} (\mu-z)d(E_{\min(\lambda,\mu)}u,u)\\ &= \int_{-\infty}^{\lambda} (\mu-z)d(E_\mu u,u) + \int_{\lambda}^{+\infty} (\mu-z)d(E_\lambda u,u)\\ &= \int_{-\infty}^{\lambda} (\mu-z)d(E_\mu u,u).\end{aligned}$$

Thus

$$d(E_\lambda u,(T-z)u) = \overline{(\lambda-z)}d(E_\lambda u,u),$$

and, finaly

$$\|(T-z)u\|^2 = \int |\lambda-z|^2 d(E_\lambda u,u),$$

so that

$$\|(T-z)u\|^2 \geq \int \inf|\lambda-z|^2 d(E_\lambda u,u) \geq \inf_{\lambda\in\sigma(T)} |\lambda-z|^2\|u\|^2.$$

□

In particular for $z\in\rho(T)$ and $u=\mathcal{R}_T(z)v$ where $\|v\|=1$, we get $\|\mathcal{R}_T(z)\| \leq \dfrac{1}{\operatorname{dist}(z,\sigma(T))}$. Then, using Proposition 4.6.5, we obtain

$$\|\mathcal{R}_T(z)\| = \frac{1}{\operatorname{dist}(z,\sigma(T))}. \tag{4.7.3}$$

In particular, it is worthwhile to notice that for a selfadjoint operator, we have

$$\forall z\in\mathbb{C}\setminus\mathbb{R}, \|\mathcal{R}_T(z)\| \leq \frac{1}{|\operatorname{Im} z|}.$$

4.7.3 Discrete spectrum and essential spectrum

Definition 4.7.6 The discrete spectrum of an unbounded operator $(\mathcal{D},T)$ is the set of eigenvalues λ of T that are isolated in $\sigma(T)$ and with finite multiplicity ($\dim\operatorname{Ker}(T-\lambda)<+\infty$). We denote it $\sigma_{disc}(T)$, and we call essentiel essential spectrum of T its complement $\sigma_{ess}(T)=\sigma(T)\setminus\sigma_{disc}(T)$.

Le spectre discret est inclus dans le spectre ponctuel défini plus haut, mais l'inclusion inverse est fausse en général. De même, le spectre continu est inclus dans le spectre essentiel sans que la réciproque ne soit toujours vraie.

On voit que $\lambda_0 \in \sigma_{disc}(A)$ si et seulement si il existe $\epsilon > 0$ tel que le projecteur $E(]\lambda_0-\epsilon, \lambda_0+\epsilon[)$ est de rang fini. De même, $\lambda_0 \in \sigma_{ess}(A)$ si et seulement si pour tout $\epsilon > 0$, le projecteur $E(]\lambda_0 - \epsilon, \lambda_0 + \epsilon[)$ n'est pas de rang fini.

On a vu par exemple que $(H^2(\mathbb{R}^d), -\Delta)$ est autoadjoint. Sa famille spectrale est $E_\lambda = \mathcal{F}^{-1}\mathbf{1}_{\xi^2 \leq \lambda}\mathcal{F}$, et son spectre est inclus dans $[0, +\infty[$. On peut montrer aussi que

$$\sigma(-\Delta) = \sigma_{ess}(-\Delta) = [0, +\infty[,$$

par exemple en utilisant la notion de suite de Weyl:

Definition 4.7.7 *Soit $(\mathcal{D}, A)$ un opérateur autoadjoint, et $\lambda \in \mathbb{R}$. On dit qu'une suite (u_n) de $\mathcal{D}$ est une suite de Weyl pour A et λ lorsque $\|u_n\| = 1$, u_n tend vers 0 faiblement et $\|(A - \lambda)u_n\| \to 0$.*

L'intérêt de cette définition réside dans la

Proposition 4.7.8 *$\lambda \in \mathbb{R}$ appartient au spectre essentiel de A si et seulement si il existe une suite de Weyl pour A et λ.*

On peut vérifier que la suite (u_n) définie ci-dessous est une suite de Weyl pour $(H^2(\mathbb{R}^d), -\Delta)$ et λ lorsque $\lambda > 0$ (cf. e.g. [**?**, Section 7.3]):

$$u_n(x) = \mathcal{F}^{-1}_{\xi \to x}(e^{-n^2|\xi-\xi_0|^2}), \quad \lambda = |\xi_0|^2.$$

4.8 Perturbations of self-adjoints operators

D'un point de vue très général, on dit que l'opérateur $A + B$ est une perturbation de l'opérateur A lorsque $A + B$ a les mêmes propriétés que A. On donne ici deux critères concernant les perturbations d'un opérateur autoadjoint A: le premier permet de dire que $A + B$ est encore autoadjoint, et le second que le spectre essentiel de $A + B$ est le même que celui de A. Il faut remarquer que le spectre discret ne peut pas rester stable par perturbation, aussi petite soit-elle.

4.8.1 Kato-Rellich Theorem

Definition 4.8.1 *Soit $(\mathcal{D}_A, A)$ et $(\mathcal{D}_B, B)$ deux opérateurs, avec $\mathcal{D}_A \subset \mathcal{D}_B$. On dit que B est A-borné lorsque pour un $a > 0$, il existe $b > 0$ tel que, pour tout $u \in \mathcal{D}_A$,*

$$\|Bu\| \leq a\|Au\| + b\|u\|$$

La borne inférieure de l'ensemble des $a > 0$ pour lesquels cette propriété est vraie est appelée borne relative de B pour A.

Lorsque A est autoadjoint, en appliquant le théorème du graphe fermé, on peut voir que tout opérateur fermé B tel que $\mathcal{D}_A \subset \mathcal{D}_B$ est A-borné. Ce qui suit repose sur le

Lemma 4.8.2 *Soit $(\mathcal{D}_A, A)$ un opérateur autoadjoint, et $(\mathcal{D}_B, B)$ un opérateur tel que $\mathcal{D}_A \subset \mathcal{D}_B$. B est A-borné si et seulement si il existe $z \in \rho(A)$ tel que $B\mathcal{R}_A(z)$ est un opérateur borné (c'est alors le cas pour tout $z \in \rho(A)$ grâce à la première formule de la résolvante). La borne relative a de B pour A est donnée par*

$$a = \lim_{\lambda \to +\infty} \|B\mathcal{R}_A(\pm i\lambda)\|.$$

Proof: Supposons que $B\mathcal{R}_A(\pm i\lambda)$ soit borné pour un $\lambda > 0$. Par la première formule de la résolvante c'est vrai pour tout $\lambda > 0$, et on note $\|B\mathcal{R}_A(\pm i\lambda)\| = a_\lambda$. On a immédiatement

$$\|Bu\| \leq a_\lambda \|Au\| + \lambda a_\lambda \|u\|,$$

ce qui montre que B est A-borné et que sa borne relative a vérifie pour tout $\lambda > 0$ l'inégalité $a \leq \|B\mathcal{R}_A(\pm i\lambda)\|$, donc

$$a \leq \liminf_{\lambda \to +\infty} \|B\mathcal{R}_A(\pm i\lambda)\|.$$

Réciproquement, supposons que B soit A-borné, de borne relative a. Pour $\epsilon > 0$, il existe $b > 0$ tel que

$$\|B\mathcal{R}_A(\pm i\lambda)u\| \leq (a+\epsilon)\|A\mathcal{R}_A(\pm i\lambda)u\| + b\|\mathcal{R}_A(\pm i\lambda)u\|.$$

Or par le théorème spectral, on a $\|\mathcal{R}_A(\pm i\lambda)u\| \leq \frac{1}{\lambda}\|u\|$ et

$$\|A\mathcal{R}_A(\pm i\lambda)u\|^2 = \int \frac{\mu^2}{\mu^2+\lambda^2} d\langle E_\mu u, u\rangle \leq \|u\|.$$

Donc $\|B\mathcal{R}_A(\pm i\lambda)\|$ est un opérateur borné de norme inférieure à $(a+\epsilon) + b/\lambda$. Ceci étant vrai pour tout $\epsilon > 0$, on voit que $\limsup_{\lambda \to +\infty} \|B\mathcal{R}_A(\pm i\lambda)\| \leq a$. □

Voilà enfin le Théorème de Kato-Rellich.

Proposition 4.8.3 *Soit $(\mathcal{D}_A, A)$ un opérateur autoadjoint (resp. essentiellement autoadjoint), et $(\mathcal{D}_B, B)$ un opérateur symétrique A-borné de borne relative inférieure à 1. Alors $(\mathcal{D}_A, A+B)$ est autoadjoint (resp. essentiellement autoadjoint).*

Proof: Supposons $(\mathcal{D}_A, A)$ autoadjoint. D'après le lemme précédent, il existe $\lambda > 0$ tel que $\|B\mathcal{R}_A(\pm i\lambda)\| < 1$, et donc $I+B\mathcal{R}_A(\pm i\lambda)$ est inversible. Or $(A+B\pm i\lambda) = (I+B\mathcal{R}_A(\pm i\lambda))(A\pm i\lambda)$, donc $(A + B \pm i\lambda)$ est d'image dense. □

4.8.2 Weyl's theorem

Definition 4.8.4 *Soit $(\mathcal{D}_A, A)$ un opérateur fermé et $(\mathcal{D}_B, B)$ un opérateur tel que $\mathcal{D}_A \subset \mathcal{D}_B$. On dit que B est A-compact lorsqu'il existe $z \in \rho(A)$ tel que $B\mathcal{R}_A(z)$ est compact (c'est alors le cas pour tout $z \in \rho(A)$ grâce à la première formule de la résolvante).*

Si B est A-compact, B est A-borné de borne relative 0. Cela découle du Lemme 4.8.2 et de l'identité $B\mathcal{R}_A(i\lambda) = (B\mathcal{R}_A(i))((A+i)\mathcal{R}_A(i\lambda))$: le premier opérateur est compact, et le second tend vers 0 fortement quand $\lambda \to +\infty$ (par exemple avec le théorème spectral). Le théorème de Kato-Rellich peut donc s'appliquer dans ce cas.

Proposition 4.8.5 *Théorème de Weyl.*
Si $(\mathcal{D}_A, A)$ est un opérateur autoadjoint, et $(\mathcal{D}_B, B)$ un opérateur symétrique A-compact, alors $(\mathcal{D}_A, A + B)$ est autoadjoint et

$$\sigma_{ess}(A + B) = \sigma_{ess}(A).$$

Remark 4.8.6 *Le théorème de Weyl sert aussi sous la forme suivante: s'il existe $z \in \rho(A + B) \cap \rho(A)$ tel que $\mathcal{R}_{A+B}(z) - \mathcal{R}_A(z)$ est compact, alors $\sigma_{ess}(A + B) = \sigma_{ess}(A)$. Cet énoncé entraine le précédent compte tenu de la seconde identité de la résolvante:*

$$\mathcal{R}_{A+B}(z) = -\mathcal{R}_A(z)B\mathcal{R}_{A+B}(z) = -\mathcal{R}_{A+B}(z)B\mathcal{R}_A(z).$$

Proof: On prouve le théorème sous la forme énoncée dans la remarque. Soit $\lambda \in \sigma_{ess}(A + B)$, $\lambda \neq z$, et (u_n) une suite de Weyl pour $A + B$ et λ. On va montrer que $(v_n = \mathcal{R}_A(z)u_n)$ (après

normalisation) est une suite de Weyl pour A et λ. D'abord (v_n) tend faiblement vers 0, mais pas fortement (donc est normalisable) puisque

$$\lim_{n\to\infty} \|v_n\| = \lim_{n\to\infty} \|\mathcal{R}_{A+B}(z)u_n\| = |\lambda - z|^{-1} \neq 0.$$

De plus on a

$$\begin{aligned}(A-\lambda)v_n &= (A-\lambda)\mathcal{R}_A(z)u_n = u_n + (z-\lambda)\mathcal{R}_A(z)u_n \\ &= u_n + (z-\lambda)\mathcal{R}_{A+B}(z)u_n - Ku_n = \mathcal{R}_{A+B}(z)(A+B-\lambda)u_n - Ku_n,\end{aligned}$$

où $K = \mathcal{R}_{A+B}(z) - \mathcal{R}_A(z)$ est compact par hypothèse, ce qui montre que $\|(A-\lambda)v_n\| \to 0$. La réciproque s'obtient en échangeant les rôles de $A+B$ et A.

□

Ce théorème permet en particulier de montrer que si $V \in L^\infty(\mathbb{R}^d)$ tend vers 0 à l'infini, alors le spectre essentiel de l'opérateur de Schrödinger $P = -h^2\Delta + V$ est le même que celui de $-h^2\Delta$, i.e. $\sigma_{ess}(P) = [0, +\infty[$. On a en effet le

Lemma 4.8.7 *Si $V \in L^\infty(\mathbb{R}^d)$ tend vers 0 à l'infini, alors V est $-\Delta$ compact.*

Proof: Il s'agit de montrer que $V(-\Delta+1)^{-1}$ est compact. Puisque $(-\Delta+1)^{-1}$ est continu de $H^2(\mathbb{R}^d)$ dans L^2, il suffit de montrer que $V : H^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ est un opérateur compact. Soit (V_k) la suite d'opérateurs définis par

$$V_k : u \mapsto \phi(\frac{x}{k})V(x)u(x), \quad \phi \in \mathcal{C}_0^\infty(B(0,1)).$$

Rappelons que pour $\phi \in \mathcal{S}(\mathbb{R}^d)$, l'application $H^s(\mathbb{R}^d) \ni u \mapsto \phi u \in \mathcal{H}^t(\mathbb{R}^d)$ est compacte pour $t > s$. Chaque V_k est donc un opérateur compact de H^2 dans L^2 puisque composé de la multiplication par $\phi_k = \phi(\frac{\cdot}{k}) \in \mathcal{S}$, qui est compacte de H^2 dans L^2, et de la multiplication par $V \in L^\infty$ qui est continue de L^2 dans L^2. Enfin on voit que

$$\|V_k - V\|_{\mathcal{L}(H^2,L^2)} \leq \sup_{|x|>k} |V(x)|,$$

ce qui montre que (V_k) tend fortement vers V quand V tend vers 0 à l'infini. □

Résumons: lorsque $V \in L^\infty(\mathbb{R}^d, \mathbb{R})$ et $V(x) \to 0$ quand $|x| \to \infty$, $(\mathcal{C}_0^\infty, P = -h^2\Delta + V)$ est un opérateur essentiellement autoadjoint. Son spectre essentiel est $[0, +\infty[$, et P peut avoir des valeurs propres négatives, isolées et de multiplicité finie. Le seul point d'accumulation possible de l'ensemble des valeurs propres négative est 0.

4.A Proof of the spectral theorem

4.A.1 The Cayley Transform

4.B Exercises

Exercise 4.B.1 Show that the un bounded operator $(\mathcal{C}_0^\infty(\mathbb{R}), T)$ on $L^2(\mathbb{R})$ defined by

$$T(\varphi) = [x \mapsto \varphi(0)e^{-x^2}]$$

is not closable.

Chapter 5

Pseudospectrum

Index