APPLIED ANALYSIS

I-Liang Chern

Department of Mathematics
National Taiwan University

Fall, 2018

Contents

Chapter 1

Some Motivation from Calculus of Variations

1.1 Introduction

In mathematical sciences, physical world problems are usually modeled by algebraic equations, differential equations, integral equations, or the extrema (minima or maxima) of some functions or functionals (functions that defined on a function space). For instance, in economics, we minimize cost function under certain constraints. In geometry, we look for geodesics which minimize the arc lengths, surfaces. In mechanics, we find an extremum of so-called action over all possible trajectories. In medical imaging problems, we minimizes some prior functional under the constraint of measurement error. These problems can be viewed as finding extrema (minima or maxima) of some functionals in some function spaces. Likewise, many time-dependent partial differential equations can be viewed as evolution processes in function spaces.

The analytic approach to these problems is to find an appropriate function space and to solve the equations in that space. Most techniques were developed in 18 – 20 centuries. However, students do not need to learn all details of them. More importantly, students need to learn the motivations, the key parts of the techniques and their applications. Thus, this lecture provides a short path to learn these basic analytic techniques. They include

Analytic Techniques

- Some motivation from calculus of variations
- Basic notion of function spaces: Metric spaces, Banach spaces and Hilbert spaces
- Methods of Contraction mapping
- Hilbert spaces: Approximation Theory, Fourier Series
- Compactness

- Bounded Operators in Hilbert spaces, Spectral theory.

Many materials of this lecture note come from the book written by
John K. Hunter and Bruno Nachtergaele, Applied Analysis.
Especially, I will use the homeworks in their book. Nevertheless, I will emphasize more on constructive approach and examples.

1.2 A short story about Calculus of Variations

The development of calculus of variations has a long history. It may goes back to the brachistochrone problem proposed by Johann Bernoulli (1696). This is an ancient Greek problem, which is to find a path (or a curve) connecting two points A and B with B lower than A such that it takes minimal time for a ball to roll from A to B under gravity. Hohann Bernoulli used Fermat principle (light travels path with shortest distance) to prove that the curve for solving the brachistochrone problem is the cycloid.

Euler (1707-1783) and Lagrange (1736-1813) are two important persons in the development of the theory of calculus of variation. I quote two paragraphs below from Wiki for you to know some story of Euler and Lagrange.

"Lagrange was an Italian-French Mathematician and Astronomer. By the age of 18 he was teaching geometry at the Rotal Artillery School of Turin, where he organized a discussion group that became the Turin Academy of Sciences. In 1755, Lagrange sent Euler a letter in which he discussed the Calculus of Variations. Euler was deeply impressed by Lagrange's work, and he held back his own work on the subject to let Lagrange publish first."

"Although Euler and Lagrange never met, when Euler left Berlin for St. Petersburg in 1766, he recommended that Lagrange succeed him as the director of the Berlin Academy. Over the course of a long and celebrated career (he would be lionized by Marie Antoinette, and made a count by Napoleon before his death), Lagrange published a systemization of mechanics using his calculus of variations, and did significant work on the three-body problem and astronomical perturbations."

1.3 Problems from Geometry

Geodesic curves Find the shortest path connecting two points A and B on the plane. Let $y(x)$ be a curve with $(a, y(a)) = A$ and $(b, y(b)) = B$. The geodesic curve problem is to minimize

$$\int_a^b \sqrt{1 + y'(x)^2}\, dx$$

among all paths $y(\cdot)$ connecting A to B.

isoperimetric problem This was an ancient Greek problem. It is to find a closed curve with a given length enclosing the greatest area. Suppose the curve is described by $(x(s), y(s))$, where s is the arc length. We may assume the total length is 2π. Thus, $0 \le s \le 2\pi$. The isoperimetric problem is

$$\max \frac{1}{2} \int_{\mathbb{T}} \left(x(s)\dot{y}(s) - y(s)\dot{x}(s)\right)\, ds$$

subject to

$$\int_{\mathbb{T}} \sqrt{\dot{x}(s)^2 + \dot{y}(s)^2}\, ds = 2\pi.$$

Here, $\mathbb{T} = [0, 2\pi]$ the unit circle. Its solution is the unit circle and the solution to this problem is usually expressed in the form of so-called isoperimetric inequality ($4\pi A \le L^2$). A geometric proof was given by Steiner (1838). An analytic proof was given by Weierstrass and by Edler [1]. The proof by Hurwitz (1902) using Fourier method will be given in later Chapter.

Minimal surface spanned by a given contour Suppose a contour in 3D is given by $h(x, y)$ with $(x, y) \in \partial\Omega$, where Ω is a simple domain in $\mathbb{R}^2$. The problem is to find surface $z(x, y)$ such that

$$z(x, y) = h(x, y) \text{ for } (x, y) \in \partial\Omega,$$

and z minimize the area

$$\int_{\Omega} \sqrt{1 + z_x^2 + z_y^2}\, dx\, dy.$$

The minimal surface problem was studied by Lagrange (1762). Classical examples include plane, catenoids, helicoids. There were many interesting minimal surfaces found in the past two centuries [2].

1.4 Euler-Lagrange Equation

Let us consider the following variational problem: minimize

$$\mathcal{J}[y] := \int_a^b F(x, y(x), y'(x))\, dx$$

subject to the boundary conditions

$$y(a) = y_a, y(b) = y_b.$$

The set

$$\mathcal{A} = \{y : [a, b] \to \mathbb{R}^n \in C^1 | y(a) = y_a, y(b) = y_b\}$$

[1] You can read a review article by Alan Siegel, A historical review of isoperimetric theorem in 2-D, and its place in elementary plan geometry, http://www.cs.nyu.edu/faculty/siegel/SCIAM.pdf. For applications, you may find a book chapter from Fan in http://www.math.ucsd.edu/ fan/research/cb/ch2.pdf

[2] You may found more interesting minimal surfaces from Wiki http://en.wikipedia.org/wiki/Minimal_surface

is called an admissible class. Here, $C^1[a,b]$ denotes the set of functions from $[a,b]$ to $\mathbb{R}^n$ which are continuously differentiable. Given a path $y \in \mathcal{A}$, we consider a variation of this path in the direction of v by

$$y(x,\epsilon) := y(x) + \epsilon v(x).$$

Here, v is a C^1 function with $v(a) = v(b) = 0$ in order to have $y(\cdot,\epsilon) \in \mathcal{A}$ for small ϵ. Such v is called a variation. Sometimes, it is denoted by δy. We can plug $y(\cdot,\epsilon)$ into $\mathcal{J}$. Suppose y is a local minimum, then for any such variation v, $\mathcal{J}[y+\epsilon v]$ takes minimum at $\epsilon = 0$. This leads to a necessary condition:

$$\frac{d}{d\epsilon}\Big|_{\epsilon=0} \mathcal{J}[y+\epsilon v] = 0.$$

Let us compute this derivative

$$\begin{aligned}\frac{d}{d\epsilon}\Big|_{\epsilon=0} \mathcal{J}[y+\epsilon v] &= \int_a^b \frac{\partial}{\partial\epsilon}\Big|_{\epsilon=0} F(x, y(x)+\epsilon v(x), y'(x)+\epsilon v'(x))\, dx \\ &= \int_a^b F_y(x,y(x),y'(x))v(x) + F_{y'}(x,y(x),y'(x))v'(x)\, dx\end{aligned}$$

It is understood that $\partial F/\partial y'$ here means the partial derivative w.r.t. the third variable y'. For instance, suppose $F = {y'}^2/2$, then $\partial F/\partial y' = y'$.

Theorem 1.1 (Necessary Condition)**.** *A necessary condition for $y \in \mathcal{A}$ to be a local minimum of $\mathcal{J}$ is*

$$\boxed{\int_a^b F_y(x,y(x),y'(x))v(x) + F_{y'}(x,y(x),y'(x))v'(x)\, dx = 0} \tag{1.1}$$

for all $v \in C^1[a,b]$ with $v(a) = v(b) = 0$.

If the solution $y \in C^2$, then we can take integration by part on the second term to get

$$\int_a^b F_{y'}(x,y(x),y'(x))v'(x)\, dx = -\int_a^b \frac{d}{dx} F_{y'}(x,y(x),y'(x))v(x)\, dx.$$

Here, we have used $v(a) = v(b) = 0$. Thus, the necessary condition can be rewritten as

$$\int_a^b \left(F_y(x,y(x),y'(x)) - \frac{d}{dx} F_{y'}(x,y(x),y'(x)) \right) v(x)\, dx = 0$$

for all $v \in C^1[a,b]$ with $v(a) = v(b) = 0$. A fundamental theorem of calculus of variations is that

Theorem 1.2. *If $f \in C[a,b]$ satisfies*

$$\int_a^b f(x)v(x)\, dx = 0$$

for all $C^\infty[a,b]$ with $v(a) = v(b) = 0$, then $f \equiv 0$.

Proof. If $f(x_0) \neq 0$ for some $x_0 \in (a,b)$ (say $f(x_0) = C > 0$), then there is small neighborhood $(x_0 - \epsilon, x_0 + \epsilon)$ such that $f(x) > C/2$. We can choose v to be a hump such that $v(x) = 1$ for $|x - x_0| \leq \epsilon/2$ and $v(x) \geq 0$ and $v(x) = 0$ for $|x - x_0| \geq \epsilon$. The test function still satisfies the boundary constraint if ϵ is small enough. Using this v, we get

$$\int_a^b f(x)v(x)\,dx \geq \frac{C\epsilon}{2} > 0.$$

This contradicts to our assumption. We conclude $f(x_0) = 0$ for all $x_0 \in (a,b)$. Since f is continuous on $[a,b]$, we also have $f(a) = f(b) = 0$ by continuity of f. □

Thus, we obtain the following stronger necessary condition.

Theorem 1.3. *A necessary condition for a local minimum y of $\mathcal{J}$ in $\mathcal{A} \cap C^2$ is*

$$\boxed{\frac{\delta \mathcal{J}}{\delta y} := F_y(x, y(x), y'(x)) - \frac{d}{dx} F_{y'}(x, y(x), y'(x)) = 0.} \tag{1.2}$$

Equation 1.2 is called the Euler-Lagrange equation for the minimization problem $\min \mathcal{J}[y]$.

Example For the problem of minimizing arc length, the functional is

$$\mathcal{J}(y) = \int_a^b \sqrt{1 + y'^2}\,dx,$$

where $y(a) = y_0, y(b) = y_1$. The corresponding Euler-Lagrange equation is

$$-\frac{d}{dx} L_{y'} = \frac{d}{dx}\left(\frac{y'}{\sqrt{1+y'^2}}\right) = 0.$$

This yields

$$\frac{y'}{\sqrt{1+y'^2}} = Const.$$

Solving y', we further get

$$y' = C \text{ (a constant)}.$$

Hence $y = Cx + D$. Applying boundary condition, we get

$$C = \frac{y_1 - y_0}{b - a}, D = \frac{by_0 - ay_1}{b - a}.$$

Thus, the minimal arc length curve is a straight line.

1.5 Problems from Mechanics

Least action principle In classical mechanics, the motion of particles in $\mathbb{R}^3$ is described by

$$m\ddot{x} = -\nabla V(x) = F.$$

Here, $V(x)$ is a potential and F is the (conservative) force. This is called Newton's mechanics. Typical examples of potentials are the harmonic potential $V(x) = \frac{k^2}{2}|x|^2$ for a mass-spring system, and Newtonian potential $V(x) = -\frac{G}{|x|}$ for solar-planet system. Here, k is the spring constant, G, the gravitation constant.

The Newton mechanics was reformulated by Lagrange (1788) in variational form and was originally motivated by describing particle motions under constraints. Let us explain this variational formulation without constraint. First, let us introduce the concept of *virtual velocity* or variation of position. Given a path $x(t)$, $t_0 \le t \le t_1$, consider a family of paths

$$x_\epsilon(t) := x(t,\epsilon) := x(t) + \epsilon v(t), t_0 \le t \le t_1, -\epsilon_0 < \epsilon < \epsilon_0.$$

Here, $v(t)$ is called a virtual velocity and $x_\epsilon(\cdot)$ is called a small variation of the path $x(\cdot)$. Sometimes, we denote $v(\cdot)$, the variation of $x_\epsilon(\cdot)$ by δx. That is, $\delta x := \partial_\epsilon|_{\epsilon=0} x_\epsilon$.

Now, the Newton's law of motion can be viewed as

$$\delta W = (F - m\ddot{x}) \cdot v = 0 \text{ for any virtual velocity } v.$$

The term δW is called the total virtual work in the direction v. The term $F \cdot v$ is the virtual work done by the external force F, while $m\ddot{x} \cdot v$ is the work done by the inertia force. The *d'Alembert principle of virtual work* states that *the virtual work is always zero along physical particle path under small perturbation δx.*

If we integrate it in time from t_0 to t_1 with fixed $v(t_0) = v(t_1) = 0$, then we get

$$\begin{aligned}
0 &= \int_{t_0}^{t_1} -m\ddot{x} \cdot v - \nabla V(x) \cdot v \, d\tau \\
&= \int_{t_0}^{t_1} m\dot{x} \cdot \dot{v} - \partial_\epsilon V(x_\epsilon) \, d\tau \\
&= \int_{t_0}^{t_1} \partial_\epsilon \frac{1}{2} m |\dot{x}_\epsilon|^2 - \partial_\epsilon V(x_\epsilon) \, d\tau \\
&= \partial_\epsilon \int_{t_0}^{t_1} L(x_\epsilon, \dot{x}_\epsilon) \, d\tau = \delta S.
\end{aligned}$$

Here,

$$L(\tau, x, \dot{x}) := \frac{1}{2} m |\dot{x}|^2 - V(x),$$

is called the Lagrangian, and the integral

$$S = \int_{t_0}^{t} L(\tau, x(\tau), \dot{x}(\tau)) \, d\tau$$

is called the *action*. Thus, $\delta S = 0$ *along a physical path*. This is called the *Hamilton principle or the least action principle*. You can show that the corresponding Euler-Language equation is exact the Newton's law of motion. Thus the following formulations are equivalent:

- Newton's equation of motion $m\ddot{x} = -V'(x)$;
- d'Alembert principle of virtual work: $\int_{t_0}^{t_1} m\dot{x} \cdot \dot{v} - V'(x)v\, dt = 0$ for all virtual velocity v;
- Hamilton's least action principle: $\delta \int_{t_0}^{t_1} \frac{m}{2}|\dot{x}|^2 - V(x)\, dt = 0$.

One advantage of variational formulation – first integral One advantage of this variational formulation is that it is easy to find some invariants (or so-called integrals) of the system. One exmple is the existence of first integral.

Theorem 1.4. *When the Lagrangian $L(x, \dot{x})$ is* independent of t*, then the quantity (called the first integral)*

$$I(x, \dot{x}) := \dot{x} \cdot \frac{\partial L}{\partial \dot{x}} - L$$

is independent of t along physical trajectories.

Proof. We differentiate I along a physical trajectory:

$$\begin{aligned} \frac{d}{dt}\left[\dot{x}L_{\dot{x}} - L\right] &= \ddot{x}L_{\dot{x}} + \dot{x}\frac{d}{dt}L_{\dot{x}} - L_x\dot{x} - L_{\dot{x}}\ddot{x} \\ &= \dot{x}\left(\frac{d}{dt}L_{\dot{x}} - L_x\right) = 0. \end{aligned}$$

□

For the Newton mechanics where $L(x, \dot{x}) = \frac{1}{2}m|\dot{x}|^2 - V(x)$, this first integral is indeed the total energy. Indeed, we obtain

$$I(x, \dot{x}) = \frac{1}{2}m|\dot{x}|^2 + V(x).$$

1.6 Method of Lagrange Multiplier

In variational problems, there are usually accompanied with some constraints. As we have seen that the iso-perimetric problem. Lagrange introduced auxiliary variable, called the Lagrange multiplier, to solve these kinds of problems. Below, we use the hanging rope problem to explain the method of Lagrange multiplier.

Hanging rope problem A rope given by $y(x)$, $a \le x \le b$ hangs two end points (a, y_a) and (b, y_b). Suppose the rope has length ℓ and density $\rho(x)$. Suppose the rope is in equilibrium, then it minimizes its potential energy, which is

$$\mathcal{J}[y] = \int_0^\ell \rho g y \, ds = \int_a^b \rho g y \sqrt{1 + y'^2} \, dx.$$

The rope is subject to the length constraint

$$\mathcal{W}[y] = \int_a^b \sqrt{1 + y'^2} \, dx = \ell.$$

Method of Lagrange multiplier In dealing with such problems, it is very much like the optimization problems in finite dimensions with constraints. Let us start with two dimensional examples. Suppose we want to minimize $f(x, y)$ with constraint $g(x, y) = 0$. The method of Lagrange multiplier states that a necessary condition for (x_0, y_0) being such a solution is that, if $\nabla g(x_0, y_0) \neq 0$, then $\nabla f(x_0, y_0) \parallel \nabla g(x_0, y_0)$. This means that there exists a constant λ_0 such that $\nabla f(x_0, y_0) + \lambda_0 \nabla g(x_0, y_0) = 0$. In other words, (x_0, y_0, λ_0) is an extremum of the unconstraint function $F(x, y, \lambda) := f(x, y) + \lambda g(x, y)$. That is, (x_0, y_0, λ_0) solves

$$\frac{\partial F}{\partial x} = 0, \ \frac{\partial F}{\partial y} = 0, \ \frac{\partial F}{\partial \lambda} = 0.$$

The first two is equivalent to $\nabla f(x_0, y_0) \parallel \nabla g(x_0, y_0)$. The last one is equivalent to the constraint $g(x_0, y_0) = 0$. The advantage is that the new formulation is an *unconstrained minimization problem.*

For constrained minimization problem in n dimensions, we have same result. Let $\mathbf{y} = (y^1, ..., y^n)$. $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$. Consider

$$\min f(\mathbf{y}) \quad \text{subject to} \quad g(\mathbf{y}) = 0.$$

A necessary condition for $\mathbf{y}_0$ being such a solution is that, if $\nabla g(\mathbf{y}_0) \neq 0$, then there exists λ_0 such that $(\mathbf{y}_0, \lambda_0)$ is an extremum of the unconstraint function $F(\mathbf{y}, \lambda) := f(\mathbf{y}) + \lambda g(\mathbf{y})$. That is, $(\mathbf{y}_0, \lambda_0)$ solves

$$\frac{\partial F}{\partial \mathbf{y}}(\mathbf{y}_0, \lambda_0) = 0, \quad \frac{\partial F}{\partial \lambda}(\mathbf{y}_0, \lambda_0) = 0.$$

For variational problem, we have much the same. Let us consider a variational problem in an abstract form:

$$\min \mathcal{J}[y] \quad \text{subject to} \quad \mathcal{W}[y] = 0$$

in some admissible class $\mathcal{A} = \{y : [a, b] \to \mathbb{R} | y(a) = y_a, y(b) = y_b\}$ in some function space. We approximate this variational problem to a finite dimensional problem. For any large n, we partition $[a, b]$ into n even subintervals:

$$x_i = a + i\frac{b - a}{n}, i = 0, ..., n.$$

We approximate $y(\cdot) \in \mathcal{A}$ by piecewise linear continuous function $\tilde{y}$ with

$$\tilde{y}(x_i) = y(x_i), i = 0, ..., n.$$

The function $\tilde{y} \in \mathcal{A}$ has an one-to-one correspondence to $\mathbf{y} := (y^1, ..., y^{n-1}) \in \mathbb{R}^{n-1}$. We approximate $\mathcal{J}[y]$ by $J(\mathbf{y}) := \mathcal{J}[\tilde{y}]$, and $\mathcal{W}[y]$ by $W(\mathbf{y}) = \mathcal{W}[\tilde{y}]$. Then the original constrained variational problem is approximated by a constrained optimization problem in finite dimension. Suppose $\mathbf{y}_0$ is such a solution. According to the method of Lagrange multiplier, if $\nabla W(\mathbf{y}_0) \neq 0$, then there exists a λ_0 such that $(\mathbf{y}_0, \lambda_0)$ solves the variational problem: $J(\mathbf{y}) + \lambda W(\mathbf{y})$.

Notice that the infinite dimensional gradient $\delta\mathcal{W}/\delta y$ can be approximated by the finite dimensional gradient $\nabla W(\mathbf{y})$. That is

$$\frac{\delta \mathcal{W}}{\delta y}[y] \approx \frac{\delta \mathcal{W}}{\delta y}[\tilde{y}] = \frac{\partial W}{\partial \mathbf{y}} = \nabla W(\mathbf{y}).$$

We summarize the above intuitive argument as the following theorem.

Theorem 1.5. *If y_0 is an extremum of $\mathcal{J}[\cdot]$ subject to the constraint $\mathcal{W}[y] = 0$, and if $\delta\mathcal{W}/\delta y \neq 0$, then there exists a constant λ_0 such that (y_0, λ_0) is an extremum of the functional $\mathcal{J}[y] + \lambda\mathcal{W}[y]$ with respect to (y, λ).*

***Remark.** A more serious proof is the follows.

1. We consider two-parameter variations

$$z(x) = y(x) + \epsilon_1 h_1(x) + \epsilon_2 h_2(x).$$

The variation h_i should satisfy the boundary conditions: $h_i(a) = h_i(b) = 0$ in order to have z satisfy the boundary conditions: $z(a) = y_a$ and $z(b) = y_b$. For arbitrarily chosen such variations h_i, we should also require ϵ_i satisfying

$$W(\epsilon_1, \epsilon_2) = \mathcal{W}[y + \epsilon_1 h_1 + \epsilon_2 h_2] = 0.$$

On the variational subspaces spanned by h_i, $i = 1, 2$, the functional $\mathcal{J}$ becomes

$$J(\epsilon_1, \epsilon_2) := \mathcal{J}[y + \epsilon_1 h_1 + \epsilon_2 h_2].$$

Thus the original problem is reduced to

$$\min J(\epsilon_1, \epsilon_2) \quad \text{subject to} \quad W(\epsilon_1, \epsilon_2) = 0$$

on this variational subspace. By the method of Lagrange multiplier, there exists a λ such that an extremum of the original problem solves the unconstraint optimization problem $\min J + \lambda W$. This leads to three equations

$$\begin{aligned}
0 &= \frac{\partial}{\partial \epsilon_1}(J + \lambda W) = \left(\frac{\delta \mathcal{J}}{\delta y} + \lambda \frac{\delta \mathcal{W}}{\delta y}\right) \cdot h_1 \\
0 &= \frac{\partial}{\partial \epsilon_2}(J + \lambda W) = \left(\frac{\delta \mathcal{J}}{\delta y} + \lambda \frac{\delta \mathcal{W}}{\delta y}\right) \cdot h_2 \\
0 &= \frac{\partial}{\partial \lambda}(J + \lambda W) = \mathcal{W}[y]
\end{aligned}$$

2. Notice that the Lagrange multiplier λ so chosen, depends on h_1 and h_2. We want o show that it is indeed a constant. This is proved below.

3. Since $\delta\mathcal{W}/\delta y(x) \neq 0$, we choose x_1 where $\delta\mathcal{W}/\delta y(x_1) \neq 0$. For any $x_2 \in (a,b)$, we consider $h_i = \delta(x - x_i)$, $i = 1, 2$. Here, δ is the Dirac delta function. It has the property: for any continuous function f,

   $$\int f(x)\delta(x - x_0)\,dx = f(x_0).$$

   By choosing such h_i, we obtain that there exists a λ_{12} such that

   $$\frac{\delta\mathcal{J}}{\delta y}(x_1) + \lambda_{12}\frac{\delta\mathcal{W}}{\delta y}(x_1) = 0$$

   $$\frac{\delta\mathcal{J}}{\delta y}(x_2) + \lambda_{12}\frac{\delta\mathcal{W}}{\delta y}(x_2) = 0$$

   In other words, the constant

   $$\lambda_{12} = -\frac{\frac{\delta\mathcal{J}}{\delta y}(x_1)}{\frac{\delta\mathcal{W}}{\delta y}(x_1)}.$$

   For any arbitrarily chosen x_2, we get the same constant. Thus, λ_{12} is independent of x_2. In fact, the above formula shows

   $$\frac{\frac{\delta\mathcal{J}}{\delta y}(x_1)}{\frac{\delta\mathcal{W}}{\delta y}(x_1)} = \frac{\frac{\delta\mathcal{J}}{\delta y}(x_2)}{\frac{\delta\mathcal{W}}{\delta y}(x_2)},$$

   for any $x_2 \neq x_1$. This means that there exists a constant λ such that

   $$\frac{\delta\mathcal{J}}{\delta y}(x) + \lambda\frac{\delta\mathcal{W}}{\delta y}(x) = 0 \text{ for all } x \in (a,b).$$

Apply the Lagrange method to the hanging rope problem Let us go back to investigate the hanging rope problem. By the method of Lagrangian multiplier, we consider the extremum problem of new Lagrangian

$$L(y, y', \lambda) = \rho g y\sqrt{1 + y'^2} + \lambda\sqrt{1 + y'^2}.$$

The Lagrangian is independent of x, thus it admits the first integral $L - y'L_{y'} = C$, or

$$(\rho g y + \lambda)\left(\sqrt{1 + y'^2} - \frac{y'^2}{\sqrt{1 + y'^2}}\right) = C.$$

Solving for y' gives

$$y' = \pm\frac{1}{C}\sqrt{(\rho g y + \lambda)^2 - C^2}.$$

Using method of separation of variable, we get

$$\frac{dy}{\sqrt{(\rho g y+\lambda)^2-C^2}}=\pm\frac{dx}{C}.$$

Change variable $u=\rho g y+\lambda$, we get

$$\frac{1}{\rho g}\cosh^{-1}\left(\frac{u}{C}\right)=\pm\frac{x}{C}+C_1.$$

Hence

$$y=-\frac{\lambda}{\rho g}+\frac{C}{\rho g}\cosh\left(\frac{\rho g x}{C}+C_2\right).$$

The constraints C, C_2 and the Lagrange multiplier λ are then determined by the two boundary conditions and the constraint. The shape of this hanging rope is called a *catenary*.

1.7 A problem from spring-mass system

[3]

Consider a spring-mass system which consists of n masses placed vertically between two walls. The n masses and the two end walls are connected by $n+1$ springs. If all masses are zeros, the springs are "at rest" states. When the masses are greater than zeros, the springs are elongated due to the gravitation force. The mass m_i moves down u_i distance, called the displacement. The goal is to find the displacements u_i of the masses m_i, $i=1,...,n$.

In this model, the nodes are the masses m_i. We may treat the end walls are the fixed masses, and call them m_0 and m_{n+1}, respectively. The edges (or the bonds) are the springs. Let us call the spring connecting m_i and m_{i+1} by edge (or spring) i, $i=1,...,n+1$. Suppose the spring i has spring constant c_i. Let us call the downward direction the positive direction.

Let me start from the simplest case: $n=1$ and no bottom wall. The mass m_1 elongates the spring 1 by a displacement u_1. The elongated spring has a *restoration force* $-c_1u_1$ acting on m_1.[4] This force must be balanced with the gravitational force on m_1.[5] Thus, we have

$$-c_1u_1+f_1=0,$$

where $f_1=m_1g$, the gravitation force on m_1, and g is the gravitation constant. From this, we get

$$u_1=\frac{f_1}{c_1}.$$

Next, let us consider the case where there is a bottom wall. In this case, both springs 1 and 2 exert forces upward to m_1. The balance law becomes

$$-c_1u_1-c_2u_1+f_1=0.$$

This results $u_1 = f_1/(c_1 + c_2)$.

Let us jump to a slightly more complicated case, say $n = 3$. The displacements

$$u_0 = 0,\ u_4 = 0,$$

due to the walls are fixed. The displacements u_1, u_2, u_3 cause elongations of the springs:

$$e_i = u_i - u_{i-1}, i = 1, 2, 3, 4.$$

The restoration force of spring i is

$$w_i = c_i e_i.$$

The force exerted to m_i by spring i is $-w_i = -c_i e_i$. In fact, when $e_i < 0$, the spring is shortened and it pushes downward to mass m_i (the sign is positive), hence the force is $-c_i e_i > 0$. On the other hand, when $e_i > 0$, the spring is elongated and it pull m_i upward. We still get the force $-w_i = -c_i e_i < 0$. Similarly, the force exerted to m_i by spring $i + 1$ is $w_{i+1} = c_{i+1} e_{i+1}$. When $e_{i+1} > 0$, the spring $i+1$ is elongated and it pulls m_i downward, the force is $w_{i+1} = c_{i+1} e_{i+1} > 0$. When $e_{i+1} < 0$, it pushes m_i upward, and the force $w_{i+1} = c_{i+1} e_{i+1} < 0$. In both cases, the force exterted to m_i by spring $i + 1$ is w_{i+1}.

Thus, the force balance law on m_i is

$$w_{i+1} - w_i + f_i = 0, i = 1, 2, 3.$$

There are three algebraic equations for three unknowns u_1, u_2, u_3. In principle, we can solve it.

Let us express the above equations in matrix form. First, the elongation:

$$e = Au, \text{ or } \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix} = \begin{pmatrix} 1 & & \\ -1 & 1 & \\ & -1 & 1 \\ & & -1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}$$

the restoration force:

$$w = Ce, \text{ or } \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{pmatrix} = \begin{pmatrix} c_1 & & & \\ & c_2 & & \\ & & c_3 & \\ & & & c_4 \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

the force balance laws:

$$A^t w = f, \text{ or } \begin{pmatrix} 1 & -1 & & \\ & 1 & -1 & \\ & & 1 & -1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix}$$

[3] This section is mainly from Gilbert Strang's book, Computational and Applied Mathematics.

[4] The minus sign is due to the direction of force is upward.

[5] The mass m_1 is in equilibrium.

where A^t is the transpose of A.

We can write the above equations in block matrix form as

$$\begin{pmatrix} C^{-1} & A \\ A^t & 0 \end{pmatrix} \begin{pmatrix} -w \\ u \end{pmatrix} = \begin{pmatrix} 0 \\ -f \end{pmatrix}.$$

This kind of block matrix appears commonly in many other physical systems, for instance, network flows, fluid flows. In fact, any optimization system with constraint can be written in this form. Here, the constraint part is the second equation. We shall come back to this point in the next section.

One way to solve the above block matrix system is to eliminate the variable w and get

$$Ku := A^tCAu = f.$$

The matrix $K := A^tCA$ is a symmetric positive definite matrix. It is called the *stiffness matrix*. For $n = 4$, we get

$$K := A^tCA = \begin{pmatrix} c_1 + c_2 & -c_2 & 0 \\ -c_2 & c_2 + c_3 & -c_3 \\ 0 & -c_3 & c_3 + c_4 \end{pmatrix}$$

Mimimum principle Consider the functional

$$P(u) := \frac{1}{2}(Ku, u) - (f, u),$$

where K is a symmetric positive definite matrix in $\mathbb{R}^n$. The directional derivative of P at u in the direction v is defined as

$$P'(u)v = \left.\frac{d}{dt}\right|_{t=0} P(u + tv)$$

$P'(u)$ is called the gradient (or the first variation) of P at u. We can compute this gradient: [6]

$$\begin{aligned} P'(u)v &= \left.\frac{d}{dt}\right|_{t=0} \frac{1}{2}(K(u + tv), u + tv) - (f, u + tv) \\ &= \frac{1}{2}\Big((Kv, u) + (Ku, v)\Big) - (f, v) \\ &= (Ku - f, v). \end{aligned}$$

Here, we have used K being symmetric. Thus,

$$P'(u) = Ku - f.$$

The second derivative is the Hessian. It is

$$P''(u) = K.$$

[6] Here, I use the following properties: $(f, g)' = (f', g) + (f, g')$. This is because $(f, g) = \sum_i f_i g_i$ and $(f, g)' = \sum_i \Big(f_i' g_i + f_i, g_i'\Big) = (f', g) + (f, g')$.

If u^* is a minimum of $P(v)$, then $P'(u^*) = 0$. This is called the Euler-Lagrange equation of P.

Conversely, If u^* satisfies the Euler-Lagrange equation $Ku^* = f$, then u^* is the minimum of $P(v)$. In fact, for any v, we compute $P(v) - P(u^*)$. We claim

$$P(v) - P(u^*) = \frac{1}{2}(K(v - u^*), v - u^*).$$

To see this, since $P(v)$ is a quadratic function of v, we can complete the squares:

$$\begin{aligned} P(v) - P(u^*) &= \frac{1}{2}(Kv, v) - (f, v) - \frac{1}{2}(Ku^*, u^*) + (f, u^*) \\ &= \frac{1}{2}(Kv, v) - \frac{1}{2}(Ku^*, u^*) - (f, v - u^*) \\ &= \frac{1}{2}(Kv, v) - \frac{1}{2}(Ku^*, u^*) - (Ku^*, v - u^*) \\ &= \frac{1}{2}(Kv, v) + \frac{1}{2}(Ku^*, u^*) - (Ku^*, v) \\ &= \frac{1}{2}(K(v - u^*), v - u^*) \geq 0. \end{aligned}$$

Hence we get that u^* is a minimum. In fact, u^* is the only minimum because $P(v) = P(u^*)$ if and only if $(K(v - u^*), v - u^*) = 0$. Since K is positive definite, we get $v - u^* = 0$.

We conclude the above discussion as the follows.

Theorem 1.6. *Let $P(u) := \frac{1}{2}(Ku, u) - (f, u)$ and K is symmetric positive definite. The vector u^* which minimizes $P(v)$ must satisfy the Euler-Lagrange equation $P'(u^*) = Ku^* - f = 0$. The converse is also true.*

The physical meaning of P is the *total potential energy* of the spring-mass system. Indeed,

$$\frac{1}{2}(CAu, Au) = \sum_{i=1}^{n} \frac{1}{2} c_i (u_i - u_{i-1})^2$$

is the sum of the *potential energy stored in the spring,* whereas the term

$$(f, u) = \sum_{i=1}^{n} f_i u_i$$

is the sum of the *works* done by the mass m_i with displacement u_i for $i = 1, ..., n$. The term $-(f, u)$ is the *gravitational potential* due to the masses m_i with displacements u_i.

1.8 A Problem from Elasticity

Consider a continuous elastic bar[7] of length 1, which is hanged vertically. (it is displaced up and down due to gravity). Set up an x-axis along the bar, so that its positive direction pointing downwards and its origin is located at the top of the elastic bar. Consider any point at x along the bar (the

[7] You may pull back and forth an elastic bar and its length is much bigger than its size of cross-section.

position is at x if no external force present), it is displaced down to $x + u(x)$ because of the action of the external force of gravity[8]. Function $u(x)$ is called the displacement. The stretching at any point is measured by the derivative $e = du/dx$, called the *strain*. If u is a constant, the elastic bar is un-stretched. Otherwise the stretching of the bar produces an internal force called *stress* (one can experience this force easily by pulling the two ends of an elastic bar). By experiments, people find this internal force is proportional to the strain of the bar, i.e.

$$\text{(internal force)} \ \ \sigma(x) = c(x)\,\frac{du}{dx}\,,$$

where $c(x)$ is a constant determined by the elastic material, or a function if the material is inhomogeneous.

To set up the model, we take a small piece of the bar $[x, x + \triangle x]$, its equilibrium requires all forces acted on it to be balanced. We have

$$\Big(ac(x)\frac{du}{dx}\Big)_{x+\triangle x} - \Big(ac(x)\frac{du}{dx}\Big)_{x} + (\rho\triangle x a)g = 0, \tag{1.3}$$

where g is the gravitational constant, a the cross-sectional area, and $\rho(x)$ the density at position x.

Dividing both sides of equation (1.3) by $a\triangle x$, then taking $\Delta x \to 0$, we get

$$-\frac{d}{dx}(c(x)\,\frac{du}{dx}) = f(x) \tag{1.4}$$

where $f(x) = g\,\rho(x)$, the external force per unit length.

The equation (1.4) must come with some appropriate physical boundary conditions to ensure it is well-posed.

Boundary conditions

(a) Both ends of the elastic bar are fixed, so no displacements:

$$u(0) = 0, \quad u(1) = 0.$$

This is called Dirichlet boundary conditions.

(b) Top end of the elastic bar is fixed (no displacement), the other end is free (no internal force since it is in the air):

$$u(0) = 0, \quad w\big|_{x=1} = c(x)\frac{du}{dx}|_{x=1} \ = 0\ .$$

The first is called a Dirichlet boundary condition, the second is called a Neumann boundary condition. The boundary conditions

$$u(0) = 0, \quad \text{or} \quad u(1) = 0$$

[8] Some other external force may be considered.

or

$$c(x)\frac{du}{dx}|_{x=1} = 0$$

are all called homogeneous boundary conditions, while the boundary conditions

$$u(0) = -1, \quad \text{or} \quad u(1) = 2,$$

or

$$c(x)\frac{du}{dx}|_{x=1} = 3$$

are all called non-homogeneous boundary conditions. A physical example of the last boundary condition is that there is a object with weight 3 hanging at the end $x = 1$.

Variational formulation We shall discuss how to derive the variational formulation for the differential equation (1.4) with Dirichlet boundary conditions. The same methodology can be applied to any other second order differential equations such as the Sturm-Liouville systems and more general boundary conditions.

The derivation is standard and simple. To do so, we multiply both sides of equation (1.4) by an arbitrary test function v (called *virtual strain*), satisfying $v(0) = v(1) = 0$ then integrating over $(0, 1)$ gives

$$\int_0^1 \left(-\frac{d}{dx}\big(c(x)\frac{du}{dx}\big)v(x) \right) dx = \int_0^1 f(x)v(x)\, dx\ .$$

Now by integration by parts and use the boundary conditions $v(0) = v(1) = 0$, we have

$$\int_0^1 \left(c(x)\frac{du}{dx}\frac{dv}{dx} \right) dx = \int_0^1 f(x)v\ dx.$$

This leads to the **variational formulation** for the equations (1.4) with Dirichlet boundary condition. Namely,

Find the solution u such that $u(0) = 0$, $u(1) = 0$ and

$$a(u, v) = g(v) \quad \textit{for any } v \textit{ satisfying } v(0) = 0 \textit{ and } v(1) = 0 \tag{1.5}$$

where $a(\cdot, \cdot)$ and $g(\cdot)$ are given by

$$a(u, v) = \int_0^1 \left(c(x)\frac{du}{dx}\frac{dv}{dx} \right) dx,\ g(v) = \int_0^1 f(x)v\, dx.$$

Definition 1.1.

- *A C^2-solution of (1.4) is called a classical solution.*
- *A solution of (1.5) is called a weak solution of (1.4).*

The advantage of this variational formulation (or called weak formulation) is that it involves only first order derivatives of u, not second derivatives in the original differential equation formulation. Thus, it has less regularity constraint on the solution u. Physically, we do encounter discontinuous coefficients. the stiffness function $c(x)$ is discontinuous if it is made of two different materials with different stiffness and connected at a point. We have the following proposition

Proposition 1.1. *Suppose $c(x)$ is discontinuous at $\bar{x}$ and smooth elsewhere. Suppose $u$is continuous and is a C^2 solution of (1.4) on both sides of $\bar{x}$. Then u is a solution of (1.5) if and only if it satisfies the following jump condition across $\bar{x}$:*

$$[cu_x] = 0 \text{ across } \bar{x}.$$

Here, $[f] := f(\bar{x}+) - f(\bar{x}-)$ denotes the jump across $\bar{x}$.

Minimal energy formulation Similar to the argument in classical mechanics, for Dirichlet boundary condition, one can define the energy by

$$E[u] := \frac{1}{2}a(u,u) - g(u),$$

and the admissible class

$$\mathcal{A} = \{u \in C^1[0,1] | u(0) = u(1) = 0\}.$$

Then the Euler-Lagrange corresponding to

$$\min_{u \in \mathcal{A}} E[u]$$

is

$$-\frac{d}{dx}\left(c(x)\frac{du}{dx}\right) = f(x)$$

with $u(0) = u(1) = 0$.

In conclusion, The following three formulations are equivalent for u with $u(0) = u(1) = 0$:

- Minimizing energy

$$\min_u \int_0^1 \frac{1}{2}c(x)u_x^2 - f(x)u(x)\, dx$$

- Variational formulation

$$\int_0^1 (c(x)u_x v_x - f(x)v(x))\, dx = 0 \text{ for all } v \text{ with } v(0) = v(1) = 0$$

- The Euler-Lagrange equation

$$-\frac{d}{dx}(c(x)\,\frac{du}{dx}) = f(x)$$

What is the energy functional corresponding to the boundary condition $u(0) = 0$ and the free-end boundary condition $c(1)u_x(1) = 3$?

Elastic bar model is a continuous limit of the spring-mass system. In the continuous model (1.4), we divide the domain $[0,1]$ into $n+1$ subintervals uniformly, each has length $\Delta x = 1/(n+1)$. We label grid points $i\Delta x$ by x_i. We imagine there are masses m_i at x_i with springs connecting them consecutively. Each spring has length Δx while it is at rest. According to the spring-mass model, we have

$$c_i(u_i - u_{i-1}) - c_{i+1}(u_{i+1} - u_i) = m_i g.$$

where c_i is the spring constant of the spring connecting x_i to x_{i+1}. As $\Delta x \approx 0$ with $x_i \approx x$, we have

$$m_i \approx \rho(x_i)\Delta x,\ c_i \approx c(x_{i-1/2})/\Delta x.$$

Here, ρ is the density. Why the spring constant is proportional to $1/\Delta x$? Think about the problem: Let us connect n springs with the same spring constant, what is the resulting spring constant?

Now, we this approximation, we get that for small Δx, the spring-mass system becomes

$$\frac{1}{\Delta x}\Big(c(x_{i-1/2})(u_i - u_{i-1}) - c(x_{i+1/2})(u_{i+1} - u_i)\Big) = \rho(x_i)\Delta x.$$

As we take $\Delta x \to 0$, we get the equation for the elastic bar:

$$-\frac{d}{dx}\Big(c(x)\frac{d}{dx}u(x)\Big) = f(x),$$

where $f = g\rho$.

Notice that the end displacements u_0 and u_{n+1} satisfy the fix-end boundary conditions

$$u_0 = 0,\ u_{n+1} = 0.$$

which correspond to the boundary condition of $u(\cdot)$ in the elastic bar model:

$$u(0) = 0,\ u(1) = 0.$$

1.9 A Problem from Fluid Mechanics

Let us consider two-dimensional incompressible and irrotational flows. The incompressibility reads

$$\nabla \cdot \mathbf{u} = 0.$$

The irrotationality gives

$$\nabla \times \mathbf{u} = 0.$$

From the second, there exists a function ϕ such that $\mathbf{u} = \nabla\phi$. This together with $\nabla \cdot \mathbf{u} = 0$, we get

$$\nabla^2\phi = 0.$$

Suppose the fluid is outside some domain Ω. Then on the boundary $\partial\Omega$, $\mathbf{u} \cdot n = 0$. Here, n is the outer normal of $\partial\Omega$. This is equivalent to the Neumann boundary condition for ϕ:

$$\nabla\phi \cdot n = 0.$$

At far field, we assume that the flow is at constant velocity $(-U, 0)$. This is equivalent to $\phi(x) = -Ux$ as $|x| \to \infty$. We can subtract $\phi_0 = -Ux$ from ϕ. Let $\Phi := \phi - \phi_0$. Then Φ satisfies

$$\nabla^2\Phi = 0,$$

$$\nabla\Phi \cdot n = -Un_x, \ \Phi(x) \to 0 \text{ as } |x| \to \infty.$$

Here, $n = (n_x, n_y)$ is the outer normal of $\partial\Omega$. To derive the formulation, we multiply a test potential ψ on both sides of $\nabla^2\Phi = 0$, then integrate over the outer domain Ω^c. We require $\psi(\infty) = 0$.

$$\begin{aligned}
0 &= \int_{\Omega^c} \nabla^2\Phi\psi\, dx \\
&= -\int_{\Omega^c} \nabla\Phi \cdot \nabla\psi\, dx + \int_{\Omega^c} \nabla \cdot ((\nabla\Phi)\psi)\ dx \\
&= -\int_{\Omega^c} \nabla\Phi \cdot \nabla\psi\, dx - \int_{\partial\Omega} \psi\nabla\Phi \cdot n\, dx \\
&= -\int_{\Omega^c} \nabla\Phi \cdot \nabla\psi\, dx + \int_{\partial\Omega} \psi U n_x\, dx
\end{aligned}$$

Thus, the variation formulation is to find Φ such that

$$\int_{\Omega^c} \nabla\Phi \cdot \nabla\psi\, dx - \int_{\partial\Omega} \psi U n_x\, dx = 0$$

for all test function $\psi \in C^1(\Omega^c)$ with $\psi(\infty) = 0$.

The optimization formulation is to find

$$\min_{\Phi\in\mathcal{A}} \left\{ \frac{1}{2}\int_{\Omega^c} |\nabla\Phi|^2\, dx + \int_{\partial\Omega} \Phi U n_x\, dx = 0 \right\}$$

$$\mathcal{A} = \{\Phi \in C^1(\Omega^c) | |\Phi(x)| \to 0 \text{ as } |x| \to \infty\}.$$

I leave you to prove that the above three formulations are equivalent when $\Phi \in C^2(\Omega^c)$ and $\Phi(\infty) = 0$.

1.10 A problem from image science – Compressed Sensing

In image science, sometimes the data (image) is very sparse under some representation. For instance, the cartoon image is piecewise smooth. Hence it is sparse if it is represented in wavelets. If the image is expressed as a vector x in $\mathbb{R}^n$ space. The dimension $n = 512^2$ for a 512×512 image. As x is represented in wavelets: $x = \Psi d = \sum_i d_i\psi_i$, most coefficients $\{d_i\}$ are zeros, or very closed to zeros. In this case, we say x is sparse as represented in terms of Ψ.

The data is usually detected by so-called sensing matrix A, which is an $m \times n$ matrix. Each individual sensing is

$$b_i = \sum_j a_{ij}x_j + n_i,$$

where b_i is the data collected, n_i is a noise.

The idea of compressed sensing is that to detect a sparse data x (or d) by an $m \times n$ sensing matrix A with $m << n$. If the noise is Gaussian white with mean 0 and variance ϵ, then we have

$$\|Ax - b\|^2 \le \epsilon.$$

There are infinite many $x \in \mathbb{R}^n$ satisfying the above constraint. Among them, we want to find the one which is most sparse as represented in Ψ. That is,

$$\min_d |d|_0 \text{ subject to } \|A\Psi d - b\|^2 \le \epsilon.$$

Here, the L0 ”norm” is defined to be

$$|d|_0 = \#\{d_i \ne 0\}.$$

Indeed, $|\cdot|_0$ is not a norm. This optimization problem is a non-convex optimization problem. It algorithm is an N-P hard problem. In the theory of compressed sensing, if $A\Psi$ satisfies certain in-coherence condition, then the problem is equivalent to the following L1 minimization problem:

$$\min_d |d|_1 \text{ subject to } \|A\Psi d - b\|^2 \le \epsilon,$$

where

$$|d|_1 := \sum_i |d_i|.$$

This is a convex optimization problem which enjoys polynomial computational complexity and many numerical algorithms are available.

Homeworks 1.1. 1. *Prove Proposition 1.1. Also state and prove this proposition for two dimensional case.*

2. *In the one-dimensional elastic bar model, what is the energy functional corresponding to the boundary condition $u(0) = 0$ and the free-end boundary condition $c(1)u_x(1) = 3$?*

3. *Search some pictures of minimal surfaces and make an album of minimal surfaces. Don’t forget to quote where they are from.*

4. *Derivate the Euler-Lagrange equation for the minimal surface problem in 3D.*

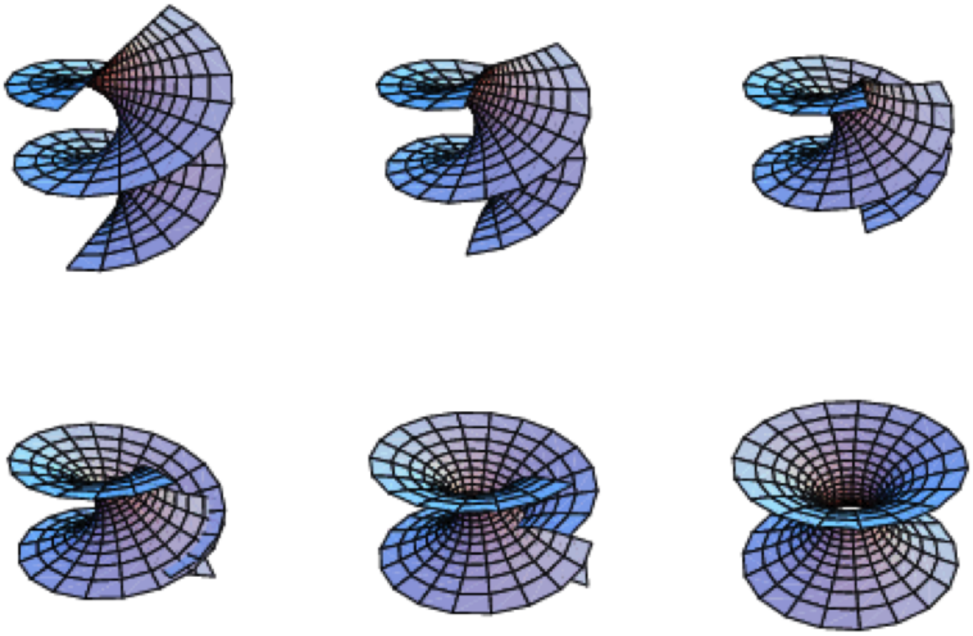

Figure 1.1: Classical minimal surfaces: Helicoid. http://mathworld.wolfram.com/Catenoid.html

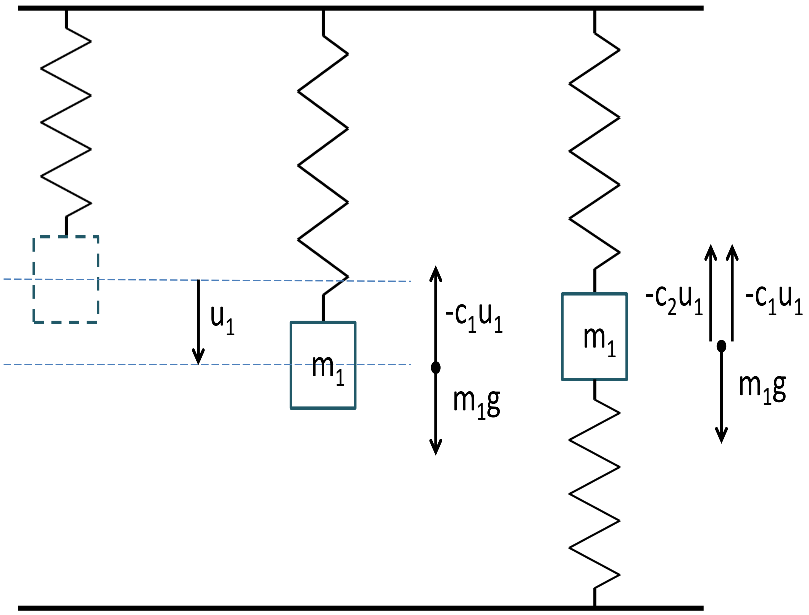

Figure 1.2: The left one is a spring without any mass. The middle one is a spring hanging a mass m_1 freely. The right one is a mass m_1 with two springs fixed on the ceiling and floor.

Chapter 2

Metric Spaces, Banach Spaces

2.1 Metric spaces

2.1.1 History and examples

The French mathematician Maurice Fréchet (1878-1973) introduced metric spaces in 1906 in his dissertation, in which he opened the field of functionals on metric spaces and introduced the notion of compactness [Wiki]. These are important concepts of point set topology.

Definition 2.1. *Given a set X. A metric d is a mapping $d : X \times X \to \mathbb{R}$ satisfying*

(a) $d(x, y) \geq 0$ *for all* $x, y \in X$*, and* $d(x, y) = 0$ *if and only if* $x = y$*;*

(b) $d(x, y) = d(y, x)$ *for all* $x, y \in X$*;*

(c) (triangle inequality) $d(x, y) \leq d(x, z) + d(y, z)$ *for all* $x, y, z \in X$.

A metric space (X, d) *is a set* X *equipped with a metric* d*.*

Examples

1. The sphere S^2 equipped with the Euclidean distance in $\mathbb{R}^3$ is a metric space. The sphere S^2 can also have another metric, the geodesic distance (or the great circle). The geodesic distance $d(x, y)$ is the shortest distance among any path on the sphere connecting x and y .

2. The continuous function space $C[a, b]$ is defined by

$$C[a, b] = \{u : [a, b] \mapsto \mathbb{R} \text{ is continuous}\}$$

with the metric

$$d(u, v) := \sup_{x \in [a,b]} |u(x) - v(x)|.$$

You can check d is a metric.

3. A (undirected) graph $G = (V, E)$ consists of vertex set $V = \{x, y, ...\}$ and edge set $E = \{e = (x, y), ...\}$. Two vertices x, y are called adjacent to each other if there is an edge $e \in E$ connecting them, and their distance is defined to be 1. A path consists of connecting edges. The distance between any two vertices x and y is defined to be the shortest distance along all paths connecting them, if any; otherwise it is defined to be infinity. Let $N = |V|$ be the number of vertices and A be an $N \times N$ matrix whose (i, j) entry is 1 if there is an edge connecting vertices i and j; and is zero otherwise. Then the distance $d(i, j) = \min\{n|(A^n)(i, j) \neq 0\}$. Here, $(A^n)(i, j)$ means the (i, j) entry of the matrix A^n. The graph G with this metric is a typical example of discrete metric space. However, this part is not what we concern in this lecture.

2.1.2 Limits and Continuous Functions

Definition 2.2. *A sequence $\{x_n\}$ in a metric space X is said to converge to $x \in X$ if $d(x_n, x) \to 0$ as $n \to \infty$. That is, all but finite of them cluster at x. In other word, for any $\epsilon > 0$ there exists N such that $d(x_n, x) < \epsilon$ for all $n \geq N$.*

Some basic notions.

- A point x is called a *limit point* of a set A in a metric space X if it is the limit of a sequence $\{x_n\} \subset A$ and $x_n \neq x$.
- The *closure* of a set A in a metric space X is the union of A with all its limit points. We denote it by $\bar{A}$.
- A set A is called *close*d if $A = \bar{A}$.
- The set $B(x, \epsilon) := \{y \in X|d(x, y) < \epsilon\}$ denote the ϵ-ball centered at x.
- A point x is called an *interior point* of a set A if there exists a neighbor $B(x, \epsilon) \subset A$ for some $\epsilon > 0$. The set of all interior points of A is called the *interior* of A and is denoted by A^o.
- A set A is called *open* if $A = A^o$.
- The complement of a set A is $A^c := \{x \in X|x \notin A\}$

One can show the following basic properties

- $(A^o)^o = A^o$
- $\bar{\bar{A}} = \bar{A}$
- $(\bar{A})^c = (A^c)^o$
- Arbitrary union of open sets is open.
- Arbitrary intersection of closed sets is closed.
- Finite union of closed sets is closed.
- Finite intersection of open sets is open.

Examples

1. The sequence $\{(-1)^n + \frac{1}{n}\}$ has no limit.
2. The closure of $\mathbb{Q}$ in $\mathbb{R}$ is $\mathbb{R}$.
3. $\mathbb{R}$ is both open and closed.

Some limit properties in $\mathbb{R}$

1. **Infimum and limit infimum for a set** Let A be a set in $\mathbb{R}$. We have the following definitions.

   (a) b is a low bound of A: if $b \leq x$ for any $x \in A$.

   (b) m is the infimum of A, or the greatest low bound (g.l.b.) of A: if (a) m is a low bound of A, (b) $b \leq m$ for any low bound b of A. We denote it by $\inf A$.

   (c) m is the limit inferior (or limit infimum): if m is the infimum of the set of the limit points of A; we denote it by $\liminf A$. If the limit point set of A is empty, then we define $\liminf A = \infty$.

   (d) We have the following property: if A has a lower bound, then the following statements are equivalent: $m = \liminf A \Leftrightarrow$ for any $\epsilon > 0$, (a) all $x \in A$ but finite many satisfies $m - \epsilon < x$ and (b) there exists at least one $x \in A$ such that $x < m + \epsilon$.

2. **Supremum and limit superior for a set** Let A be a set in $\mathbb{R}$. We have the following definitions.

   (a) u is a low bound of A: if $u \geq x$ for any $x \in A$.

   (b) M is the supremum of A, or the least upper bound (l.u.b.) of A: if (a) M is a upper bound of A, (b) $u \geq M$ for any upper bound u of A. We denote it by $\sup A$.

   (c) M is the limit superior (or limit supremum): if M is the supremum of the set of the limit points of A. We denote it by $\limsup A$. If the limiting set is empty, we define $\limsup A = -\infty$.

   (d) We have the following property: if A has an upper bound, then the following statements are equivalent: $M = \limsup A \Leftrightarrow$ for any $\epsilon > 0$, (a) all $x \in A$ but finite many satisfies $M + \epsilon > x$ and (b) there exists at least one $x \in A$ such that $x > M - \epsilon$.

3. **Infimum and limit infimum for a sequence** Let (x_n) be a sequence. Then the definition of infimum and liminf of (x_n) is just to treat them as a set. The definition of liminf is equivalent to

$$m = \liminf x_n \Leftrightarrow m = lim_{n\to\infty} \inf_{m \geq n} x_m \Leftrightarrow m = \sup_{n \geq 0} \inf_{m \geq n} x_m.$$

Examples

1. Let $x_n = (-1)^n - 1/n, n \geq 1$. Then $\inf\{x_n\} = -2$ and $\liminf_{n\to\infty} x_n = -1$.
2. Let $A = \{\sin x | x \in (-\pi/2, \pi/2)\}$. Then $\inf A = \liminf A = -1$.

Continuous functions

Definition 2.3. *Let f be a function which maps (X, d_X) into (Y, d_Y). We say f is continuous at a point $x_0 \in X$ if for any $\epsilon > 0$ thee exists a $\delta > 0$ such that*

$$d_Y(f(x), f(x_0)) < \epsilon \textit{ whenever } d_X(x, x_0) < \delta.$$

Roughly speaking, f is continuous at x_0 means that whenever x is close to x_0, the corresponding $f(x)$ has to be close to $f(x_0)$. This definition is indeed equivalent to the following two definitions. Their proofs are left to you to get familiar with the ϵ-δ language for limit theory.

Definition 2.4. *We say that f is sequentially continuous at a point $x_0 \in X$ if for any sequence $(x_n)_{n=1}^{\infty}$ with $x_n \to x_0$, we have $f(x_n) \to f(x_0)$ as $n \to \infty$.*

Definition 2.5. *We say f is continuous at $x_0 \in X$ if $f^{-1}(V)$ is open for every open neighborhood V in Y containing $f(x_0)$.*

The ϵ-δ definition for continuity is the most general formulation of continuity in metric space. A more restricted but more quantitative definition is the following order of continuity. The relative closeness of $f(x)$ to $f(x_0)$ with respect to $d_X(x, x_0)$ can be measured by

$$d_Y(f(x), f(x_0)) \le \omega(d_X(x, x_0)),$$

where $\omega(t)$ is a non-negative increasing function, and $\omega(t) \to 0$ as $t \to 0$. For instance, the function $|x|^\alpha \sin(1/x)$ $(\alpha > 0)$ is continuous at $x = 0$. The order of continuity can be measured by $\omega(t) = |t|^\alpha$. Thus, the continuity can be measured by some majorant function $\omega(\cdot)$. But the continuity is independent of its oscillation. The oscillation can be measured from the derivative of the function, or local variation of the function. Among the majorant functions, $\omega(t) = |t|^\alpha \to 0$ for $\alpha > 0$. It converges fast if α is large, and slow if α is close to 0. The majorant function $\omega(t) = 1/\ln|t|$ converges to 0 very slowly as $|t| \to 0$, as compared with $|t|^\alpha$.

Exercise. Use ϵ-δ argument to show that x^2, $1/x$, $\sin(1/x)$ are continuous on $(0, 1)$.

Infimum and limit infimum of a function

1. Let $f : (X, d) \to \mathbb{R}$. Then

   $$m = \inf_{x \in X} f(x) := \inf\{f(x) | x \in X\};$$

   and

   $$\liminf_{x \to \bar{x}} f(x) := \lim_{\delta \to 0+} \inf_{d(x, \bar{x}) < \delta} f(x).$$

2. The above definition is equivalent to: (a) for any $\epsilon > 0$, there exists a $\delta > 0$ such that $m - \epsilon < f(x)$ for all $d(x, \bar{x}) < \delta$; (b) for any ϵ, there exists a $\delta > 0$ and an x with $d(x, \bar{x}) < \delta$ such that $f(x) < m + \epsilon$.

3. Let
$$f(x) = \begin{cases} |x| & x \neq 0 \\ -1 & x = 0. \end{cases}$$
Then $\liminf_{x \to 0} f(x) = 0$.

Definition 2.6. *A function $f : (X, d) \to \mathbb{R}$ is called lower semi-continuous (l.s.c.) if for every $x \in X$,*
$$\liminf_{y \to x} f(y) \geq f(x)$$

There is an equivalent way to check the lower semi-continuity by epigraph. It is defined to be
$$\text{epi} f := \{(x, t) \in X \times \mathbb{R} | f(x) \leq t\}$$
Then a function is l.s.c. if and only if its epigraph is closed in $X \times \mathbb{R}$.

2.1.3 Completions of metric spaces

Definition 2.7. *A sequence $\{x_n\}$ in a metric space X is called a Cauchy sequence if all but finite of them cluster. This means that: for any $\epsilon > 0$, there exists an N such that $d(x_n, x_m) < \epsilon$ for any $n, m \geq N$.*

Definition 2.8. *A metric space is called complete if all Cauchy sequences in X converge.*

Examples

1. $\mathbb{R}^n$, $\mathbb{C}^n$ are complete metric spaces.
2. $\mathbb{Q}^n$ equipped with the metric $d(x, y) := \|x - y\|_2$ is not complete. But the completion of $\mathbb{Q}^n$ in $\mathbb{R}^n$ is $\mathbb{R}^n$.

Given a metric space (X, d), there is a natural way to extend it to a complete and smallest metric space $(\tilde{X}, \tilde{d})$, which means that

1. There is an imbedding $\imath : X \to \tilde{X}$. This means that $\imath$ is one-to-one.
2. The restriction of $\tilde{d}$ on $\imath(X)$ is identical to d. That is, $\tilde{d}(\imath x, \imath y) = d(x, y)$.
3. $(\tilde{X}, \tilde{d})$ is complete.
4. $\imath(X)$ is dense in $\tilde{X}$, that is, $\overline{\imath(X)} = \tilde{X}$.

In applications, we would like to work on complete spaces, which allow us to take limit. If a metric space is not complete, we can take its completion. The completion of an incomplete space is mimic to the completion of $\mathbb{Q}$ in $\mathbb{R}$. You imagine that any real number can be approximated by rational sequences. This approximation sequence can be constructed in many ways. For instance, let $x \in \mathbb{R}$ be represented by

$$x = \sum_{i=-m}^{\infty} a_i p^{-i},$$

where $p > 1$ is a positive integer, m an integer, and $0 \le a_i < p$ are integers. We choose

$$x_n = \sum_{i=-m}^{n} a_i p^{-i}.$$

Then (x_n) is a Cauchy sequence and approaches x. Certainly there are infinite many Cauchy sequences approaching the same x. We say that they are equivalent. In other words, $(x_n) \sim (y_n)$ if $x_n - y_n \to 0$ as $n \to \infty$. The collection of all those Cauchy sequences which approach the same real number x is called an equivalence class. Any particular Cauchy in this equivalence is called a representation of the real number. Thus, we may identify a real number x to the equivalent class of Cauchy sequence associated with it. This correspondence is one-to-one and onto. Thus, $\mathbb{R}$ can be viewed as the set of all these equivalent classes.

The completion of an abstract metric space (X, d) mimic to the above process. Its construction goes as below.

1. Define
$$\tilde{X} := \{(x_n)_{n\in\mathbb{N}} \text{ is a Cauchy sequence in } X\}/\sim,$$
where the equivalence relation is defined by [1]
$$(x_n) \sim (y_n) \text{ if and only if } d(x_n, y_n) \to 0 \text{ as } n \to \infty.$$
Thus, the element $\tilde{x} \in \tilde{X}$ is the set of all Cauchy sequences $\{(x_n)\}$ in which all of them are equivalent.

2. Given $\tilde{x}$ and $\tilde{y}$, choose any two representation (x_n) and (y_n) from $\tilde{x}$ and $\tilde{y}$ respectively, define
$$\tilde{d}(\tilde{x}, \tilde{y}) := \lim_{n\to\infty} d(x_n, y_n).$$

3. Given $x \in X$, define the Cauchy sequence (x_n) with $x_n = x$ for all n. The equivalent class that containing this Cauchy sequence (x_n) is denoted by $\imath(x)$. This is a natural imbedding from X to $\tilde{X}$.

One can show that in the above construction: (i) The relation $\sim$ is an equivalent relation, (ii) $\tilde{d}$ is well-defined, (iii) $\tilde{X}$ is complete, and (iv) $\imath(X)$ is dense in $\tilde{X}$.

[1] A relation $\sim$ is called an equivalent relation in a set X if (i)$x \sim x$, (ii) if $x \sim y$ then $y \sim x$, (iii) if $x \sim y$ and $y \sim z$, then $x \sim z$. An equivalent class $\tilde{x} := \{y \in X | y \sim x\}$.

The function space $C[a, b]$ In applications, especially ODEs, we often encounter that the solution is at least continuous in time. This motives us to study the function space

$$C[a, b] := \{u : [a, b] \to \mathbb{R} \text{ is continuous.}\}$$

Given $u, v \in C[a, b]$, we define

$$d(u, v) := \sup_{x\in[a,b]} |u(x) - v(x)|.$$

Theorem 2.1. *$C[a, b]$ is complete.*

Proof. Suppose $\{u_n\}$ is a Cauchy sequence in $C[a, b]$. For any $\epsilon > 0$, there exists an $N(\epsilon) > 0$ such that

$$\sup_{x\in[a,b]} |u_n(x) - u_m(x)| < \epsilon$$

for every $n, m > N$. For each fixed $x \in [a, b]$, $\{u_n(x)\}$ is a Cauchy sequence in $\mathbb{R}$. Thus, $u_n(x)$ converges to a limit, called $u(x)$. This convergence is indeed uniform in x. In fact, we can take $m \to \infty$ in the above formula to get

$$\sup_{x\in[a,b]} |u_n(x) - u(x)| \le \epsilon.$$

for every $n > N$. Next, we show that u is continuous at every point $x_0 \in [a, b]$. For any $\epsilon > 0$, we have seen that there is N such that $\sup_{x\in[a,b]} |u_N(x) - u(x)| < \epsilon$. On the other hand, u_N is continuous at x_0. Thus, there exists a $\delta > 0$, which depends on u_N, ϵ and x_0, such that

$$|u_N(x) - u_N(x_0)| < \epsilon \text{ for } |x - x_0| < \delta.$$

Thus,

$$|u(x) - u(x_0)| \le |u(x) - u_N(x)| + |u_N(x) - u_N(x_0)| + |u_N(x_0) - u(x_0)| < 3\epsilon.$$

This shows u is continuous at an arbitrary point $x_0 \in [a, b]$. □

Exercise.

- Can you replace $C[a, b]$ by $C(a, b)$ in the above theorem? Here, $C(a, b)$ includes all continuous functions from (a, b) to $\mathbb{R}$ with finite sup norm. In this definition, $1/x$ is not in $C(0, 1)$ but $\sin(1/x)$ does.

- Consider the subset $\mathcal{A}$ in $C(a, b)$ to be the set of those functions which have finite limits at the boundary points a and b. What is the relation between $\mathcal{A}$ and $C[a, b]$?

- Is $C(0, \infty)$ with the above metric a complete metric space?

2.2 Banach spaces

2.2.1 Normed linear space – A space where we can do calculus

The metric space has no algebraic structure. A natural extension of Euclidean space structure is the normed linear space, in which calculus can be introduced. A set X with addition and scalar multiplication is called a linear space (or vector space).

Definition 2.9. *A linear space X over a field $\mathbb{R}$ (or $\mathbb{C}$) has addition and scalar multiplication operations which satisfy*

(a) for all $x, y, z \in X$, $x + y = y + x$; $(x + y) + z = x + (y + z)$; there exists a zero vector 0 such that $x + 0 = x$; for all $x \in X$, there exists a unique $(-x)$ such that $x + (-x) = 0$;

(b) for any $x, y \in X$, any $\lambda, \mu \in \mathbb{R}$, $1x = x$, $(\lambda + \mu)x = \lambda x + \mu x$, $\lambda(x + y) = \lambda x + \lambda y$, $\lambda(\mu x) = (\lambda\mu)x$;

Definition 2.10. *A norm $\|\cdot\|$ on a linear space X is a mapping $X \to \mathbb{R}$ satisfying*

(a) $\|x\| \geq 0$ for all $x \in X$ and $\|x\| = 0$ if and only if $x = 0$;

(b) $\|\lambda x\| = |\lambda|\|x\|$ for all $\lambda \in \mathbb{R}$ and $x \in X$;

(c) (triangle inequality) $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in X$.

A normed linear space $(X, \|\cdot\|)$ is a linear space X equipped with a norm $\|\cdot\|$.

Definition 2.11. *A complete normed linear space is called a Banach space.*

Properties

- A normed linear space is a metric space equipped with the metric $d(x, y) = \|x - y\|$.
- A metric in a linear space defines a norm if it satisfies the translational invariant property ($d(x - z, y - z) = d(x, y)$) and the homogeneity property ($d(\lambda x, 0) = \lambda d(x, 0)$).
- The unit ball in a normed linear space is convex (triangle inequality).
- In a finite dimensional normed space, all norms are equivalent. Here, two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ in a normed linear space X are said to be equivalent if there exists two positive constants C_1, C_2 such that
$$C_1\|x\|_1 \leq \|x\|_2 \leq C_2\|x\|_1$$
for all $x \in X$.

Examples

1. The $\mathbb{R}^n$ space equipped with the Euclidean norm

$$\|x\|_2 = (|x_1|^2 + \cdots + |x_n|^2)^{1/2}$$

is a Banach space.

2. The $\mathbb{R}^n$ space equipped with the p-norm:

$$\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}, 1 \leq p < \infty$$

are Banach spaces. Furthermore, one can show that

$$\|x\|_\infty := \max_i |x_i|$$

is a norm, and

$$\|x\|_p \to \|x\|_\infty, \text{ as } p \to \infty.$$

Notice that $\|x\|_p$ with $0 \leq p < 1$ is not a norm, but it can measure the sparsity of x. Indeed,

$$\|x\|_0 := \#\{x_i \neq 0\},$$

which measure the sparsity of x, and $\|x\|_p \to \|x\|_0$ as $p \to 0$.

3. The set of matrices

$$\mathcal{M}_{m\times n} := \{A : \mathbb{R}^n \mapsto \mathbb{R}^m \text{ is linear}\}$$

equipped with the Frobenious norm defined by

$$\|A\|_F := \left(\sum_{ij} |A_{ij}|^2\right)^{1/2}$$

is a Banach space.

4. The $\ell^p(\mathbb{N})$ $(1 \leq p < \infty)$ space is the set

$$\ell^p(\mathbb{N}) := \{x : \mathbb{N} \to \mathbb{R} | \sum_{i=1}^{\infty} |x_i|^p < \infty\}$$

equipped with the norm

$$\|x\|_p := (|x_1|^p + |x_2|^p + \cdots)^{1/p}.$$

Similar to the finite dimensional case, we define

$$\|x\|_\infty := \sup_i |x_i|$$

is a norm and $\|x\|_p \to \|x\|_\infty$ (as $p \to \infty$) if they exist. Indeed, one can prove that ℓ^p, $1 \leq p \leq \infty$ are Banach spaces.

5. The set of continuous functions
$$C[a,b] := \{u : [a,b] \to \mathbb{R} \text{ is continuous}\}$$
is a linear space. We define the sup norm by
$$\|u\|_\infty := \max_{x\in[a,b]} |u(x)|.$$
Then $(C[a,b], \|\cdot\|_\infty)$ is a Banach space.

6. $(C[a,b], \|\cdot\|_p)$, $1 \le p < \infty$ is not complete. A simple example is that the sequence of continuous functions
$$u_n(x) := \tanh(nx), x \in [-1,1]$$
tends (in all $\|\cdot\|_p, 1 \le p < \infty$) to
$$u(x) = \begin{cases} -1 & \text{for } x < 0 \\ 0 & \text{for } x = 0 \\ 1 & \text{for } x > 0 \end{cases}$$
which is not in $C[-1,1]$.

The Completion of normed linear spaces

1. The completion of $\mathbb{Q}$ in $\mathbb{R}$ under absolute value norm $|\cdot|$ is $\mathbb{R}$.

2. Let
$$C^1[a,b] := \{u : [a,b] \mapsto \mathbb{R}, u, u' \text{ are continuous}\}$$
Then $C^1[a,b]$ is complete under the norm
$$|u|_{1,\infty} := \sup_x |u(x)| + \sup_x |u'(x)|.$$
But $C^1[a,b]$ is not complete under the sup norm $|u|_\infty := \sup |u(x)|$. Its completion under sup norm is $C[a,b]$. .

3. The completion of $C[a,b]$ under the norm
$$\|u\|_1 := \int_a^b |u(x)|\, dx$$
is called the L^1-space, and is denoted by $L^1(a,b)$. It is the set of all *Lebesgue integrable* functions on (a,b).

4. The completion of $(C[a,b], \|\cdot\|_p)$, $1 \le p < \infty$ is the L^p space
$$L^p(a,b) := \{u : [a,b] \to \mathbb{R} | \int_a^b |u(x)|^p\, dx < \infty\}$$
where the above integration is in the Lebesgue sense.

5. The function $1/|x|^\alpha$ is in $L^p(-1,1)$ for $0 < \alpha p < 1$.
6. Is the function $\sin(1/x)$ in $L^p(-1,1)$ for $1 \le p \le \infty$?
7. For which α the corresponding $|x|^{-\alpha}\sin(1/|x|) \in L^p(-1,1)$?

2.2.2 Approximation and Basis

In function spaces, we want to approximate general functions in terms of linear combination of some simple known functions. This linear combination is usually in terms of infinite series, but countable. A set I is called countable if it is either finite many or there is an one-to-one correspondence between I and $\mathbb{N}$. One can check that $\mathbb{Z} \times \mathbb{Z}$ is countable and thus $\mathbb{Q}$ is also countable because a rational number r can be represented by p/q with $(p,q) \in \mathbb{Z} \times \mathbb{Z}$. In $\mathbb{R}$, we want to approximate a real number by an (countable) infinite series. For instance, we may approximate $r \in [0,1]$ by

$$r = \sum_{n=1}^{\infty} a_n 2^{-n}.$$

where $a_n \in \{0,1\}$. Each finite sub series is an element in $\mathbb{Q}$. This motivates the following definition.

Definition 2.12. *A metric space X is said to be separable if there is a countable set $A \subset X$ such that $\bar{A} = X$.*

$C[0,1]$ **is separable**

1. The Bernstein polynomials are

$$b_{\nu,n}(x) := \left(\begin{array}{c} n \\ \nu \end{array} \right) x^\nu (1-x)^{n-\nu}, \nu = 0, ..., n.$$

They are in the space

$$P_n := \{p(x) | p \text{ is a polynomial and } deg(p) \le n\}$$

The space P_n has dimension $n+1$. Since $b_{\nu,n}$, $\nu = 0, ..., n$ are independent, They form a basis of P_n.

2. The set of Bernstein polynomials with rational coefficients

$$\mathcal{A} = \{\sum_{\nu=0}^{n} a_\nu b_{\nu,n} | a_\nu \in \mathbb{Q}, n \ge 0\}$$

is countable and is dense in $C[0,1]$. Indeed, let $f \in C[0,1]$, then

$$B_n(f) := \sum_{\nu=0}^{n} f\left(\frac{\nu}{n}\right) b_{\nu,n}(x)$$

converges to f in $\|\cdot\|_\infty$. The key parts of the proof are

(a) $b_{\nu,n} > 0$ and $\sum_{\nu=0}^{n} b_{\nu,n}(x) = 1$.

(b) The difference $f(x) - B_n(f)$ has the following estimates:

$$\begin{aligned} |f(x) - B_n(f)| &= \left| \sum_{\nu=0}^{n} (f(x) - f(\nu/n))\, b_{\nu,n}(x) \right| \\ &\le \sum_{\nu=0}^{n} |f(x) - f(\nu/n)|\, b_{\nu,n}(x) \\ &= \left(\sum_{|\nu/n - x| \le \delta} + \sum_{|\nu/n - x| > \delta} \right) |f(x) - f(\nu/n)|\, b_{\nu,n}(x) \end{aligned}$$

(c) The Bernstein polynomial $b_{\nu,n}(x)$ concentrates at $\nu/n \sim x$ as $n \to \infty$. More precisely, for any fixed small δ,

$$\sum_{|\nu/n - x| \ge \delta} b_{\nu,n}(x) \to 0$$

as $n \to \infty$.

(d) f is uniformly continuous on $[0,1]$. That is, for any $\epsilon > 0$, there exists a $\delta > 0$ such that $|f(x) - f(y)| < \epsilon$ whenever $|x - y| < \delta$.

I leave you to fill in the gaps. [2]

[2] The Berstein polynomial has the following probability interpretation.

(a) Let X be the random variable of one binormial trial with probability x of success. That is, $P(X = 1) = x$ and $P(X = 0) = (1 - x)$. If we perform two independent Bernoulli trials, denote X_i the outcome of the ith trial. The sample space $\Omega_2 = \{(1,1), (1,0), (0,1), (0,0)\}$. Here, (a_1, a_2) denotes that $X_1 = a_1$ and $X_2 = a_2$.The probability of $S_2 = X_1 + X_2$ is

$$P(S_2 = 2) = x^2, P(S_2 = 1) = 2x(1-x), P(S_2 = 0) = (1-x)^2.$$

For n independent Bernoulli trials, the number of elements that have ν times success is $n!/(\nu!(n-\nu)!)$. Thus, the probability of ν times successes is

$$P(S_n = \nu) = \frac{n!}{\nu!(n-\nu)!} x^{\nu}(1-x)^{n-\nu} = b_{\nu,n}(x).$$

(b) The expectation of a random X is defined to be $E(X) := \sum_{\nu} \nu P(X = \nu)$.

(c) The weak form of law of large number states that: Let X be the random variable of the Bernoulli trial and $E(X) = x$. Let $S_n = X_1 + \cdots + X_n$ where X_i are all independent and with identical distribution as X, then

$$\lim_{n\to\infty} E\left(\frac{S_n}{n}\right) = x.$$

(d) The key is the Chybeshev inequality: $P(|S_n/n - x| > \delta) \le \frac{\sigma^2}{\delta^2 n} \to 0$ as $n \to \infty$. Here, $\sigma^2 = E(|X - x|^2)$ the variance. I shall show a special case: If $E(X) = 0$, then $P(|X| \ge \delta) \le \sigma^2/\delta$. Let $F(x) = P(X \le x)$ be the

Definition 2.13. *Let $(X, \|\cdot\|)$ be a separable Banach space. A set $\{e_n\}_{n=1}^\infty$ is called a Schauder basis of X if for every $x \in X$, there is a unique representation of x in terms of $\{e_n\}$ by*

$$x = \sum_{n=1}^\infty a_n e_n.$$

Examples

1. In ℓ^p, $1 \le p < \infty$, let $e_1 = (1, 0, 0, \cdots)$, $e_2 := (0, 1, 0, \cdots)$, $e_2 := (0, 0, 1, 0, \cdots)$,etc. Then Given a point $x \in \ell^p$, $x = (x_1, x_2, \cdots)$. Let

   $$x^n := (x_1, ..., x_n, 0, 0, \cdots).$$

   Then

   $$\|x^n - x\|_p \to 0.$$

   In other words, $\{e_n\}_{n=1}^\infty$ is a Schauder basis in ℓ^p.

2. In the next section, we will see that the Fourier functions

   $$\{e_n := e^{2\pi i n x}\}_{n=-\infty}^\infty$$

   form a basis in $C(\mathbb{T}) := \{u : [0, 1] \to \mathbb{C} \text{ periodic}\}$.

3. In finite element method, the solution is approximated by piecewise linear function. We shall construct the corresponding basis.

   Let us consider the domain $[0, 1]$. For any $n \in \mathbb{N}$, we partition $[0, 1]$ into 2^n subintervals evenly. The points $x_{nk} = 2^{-n}k$, $k = 0, ..., 2^n$ are called the nodal points of the partition. Given $u \in C[0, 1]$, let u_n be the continuous function with $u_n(x_{nk}) = u(x_{nk})$ and linear on each subinterval $(x_{n,k}, x_{n,k+1})$, $k = 0, ..., 2^n - 1$. The function u_n can approximate u in sup norm. Indeed, since u is uniformly continuous on $[0, 1]$, we have for any $\epsilon > 0$, there exists a

probability distribution function of X.

$$\begin{aligned} P(|X| \ge \delta) &= \int_{|x|/\delta \ge 1} dF(x) \\ &= \int_{|x|/\delta \ge 1} \frac{|x|^2}{\delta^2} dF(x) \\ &\le \int \frac{|x|^2}{\delta^2} dF(x) \\ &= \frac{\sigma^2}{\delta} \end{aligned}$$

We apply this Chybeshev inequality to S_n/n. We may assume $E(X) = x = 0$, using independence of X_i, we get the variance of S_n/n is σ^2/n.

(e) $E(f(S_n/n)) = \sum_{\nu=0}^n f(\nu/n) P(S_n = \nu)$.

$\delta > 0$ such that $|u(x) - u(y)| < \epsilon$ whenever $|x - y| < \delta$. We choose N such that $2^{-N} < \delta$. Now, for any $n \geq N$, there exists an k such that $x_{n,k} \leq x \leq x_{n,k+1}$. Hence we have

$$|u(x) - u_n(x)| \leq |u(x) - u(x_{n,k})| + |u_n(x_{n,k}) - u_n(x_{n,k+1})| \leq 2\epsilon.$$

Thus, $\|u_n - u\|_\infty \to 0$ as $n \to \infty$.

We shall write u_n in terms of a basis. Consider the hat function

$$\phi(x) = \begin{cases} x+1 & \text{for } -1 \leq x < 0 \\ 1-x & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

We can perform scaling and translation and produce

$$\phi_{n,k} := \phi(2^n x - k).$$

This function is centered at x_{nk} with support $(x_{n,k-1}, x_{n,k+1})$. The $\phi_{n,0}$ and $\phi_{n,2^n}$ are the boundary nodal functions, whereas $\phi_{n,k}$, $k = 1, ..., 2^n - 1$ the interior nodal functions. The piecewise linear function u_n defined above can be represented in terms of $\phi_{n,k}$:

$$u_n(x) := \sum_{k=0}^{2^n} u(x_{nk})\phi_{nk}(x), \text{ where } x_{nk} := 2^{-n}k.$$

Furthermore, the hat function ϕ has the following scaling property

$$\phi(x) = \frac{1}{2}\phi(2x+1) + \phi(2x) + \frac{1}{2}\phi(2x-1).$$

For interior nodal functions, we then have

$$\phi_{n-1,k} = \frac{1}{2}\phi_{n,2k-1} + \phi_{n,2k} + \frac{1}{2}\phi_{n,2k+1}$$

for $k = 1, ..., 2^{n-1} - 1$. For boundary nodal functions, we have

$$\phi_{n-1,0} = \phi_{n,0} + \frac{1}{2}\phi_{n,1},$$

and

$$\phi_{n-1,2^{n-1}} = \phi_{n,2^n} + \frac{1}{2}\phi_{n,2^n-1},$$

Let us consider the space

$$V^n := \text{span}\{\phi_{nk}, k = 0, ..., 2^n\}.$$

The dimension of V^n is $2^n + 1$. From the scaling property of ϕ, the space V^n are nested:

$$V^0 \subset V^1 \subset \cdots \subset V^n \subset \cdots$$

Let us define

$$\psi_{n-1,k} = \phi_{n,2k-1}, k = 1, ..., 2^{n-1},$$

and the space

$$W^{n-1} = \text{span}\{\psi_{n-1,k} | k = 1, ..., 2^{n-1}\}.$$

Then $V^n = V^{n-1} + W^{n-1}$. Indeed, the inversion: $\{\phi_{n-1,k}, \psi_{n-1,k}\} \to \{\phi_{n,k}\}$ is given by

$$\phi_{n,2k-1} = \psi_{n-1,k}, k = 1, ..., 2^{n-1}$$

$$\phi_{n,2k} = \phi_{n-1,k} - \frac{1}{2}\left(\psi_{n-1,k} + \psi_{n-1,k+1}\right), k = 1, ..., 2^{n-1} - 1$$

$$\phi_{n,0} = \phi_{n-1,0} - \frac{1}{2}\psi_{n-1,1}, \phi_{n,2^n} = \phi_{n-1,2^{n-1}} - \frac{1}{2}\psi_{n-1,2^{n-1}},$$

The dimensions of V^{n-1} and W^{n-1} are $2^{n-1} + 1$ and 2^{n-1}. Their sum is $2^n + 1$, which is the dimension of V^n. Since $V^n \to C[0,1]$, we then expect

$$C[0,1] = V^0 + W^0 + W^1 + \cdots .$$

We thus expect

$$\{\phi_{0,0}, \phi_{0,1}, \psi_{n,k}, k = 1, ..., 2^n, n \geq 0\}$$

forms a Schauder basis in $C[0,1]$.

2.3 Linear Operators in Banach Spaces, Basic

Definition 2.14. *Let $(X, \|\cdot\}_X)$ and $(Y, \|\cdot\|_Y)$ be two normed linear spaces. A linear map $A : X \to Y$ is called bounded if there exists a constant C such that*

$$\|Ax\|_Y \leq C\|x\|_X$$

for all $x \in X$.

Lemma 2.1. *A linear map A is bounded if and only if it is continuous.*

Proof. 1. If A is bounded, then

$$\|Ax - Ay\|_Y = \|A(x - y)\|_Y \leq C\|x - y\|_X.$$

This shows A is continuous.

2. If A is continuous, then in particular, it is continuous at 0. This means that for any $\epsilon > 0$, there exists a $\delta > 0$ such that $\|Az\|_Y \leq \epsilon$ whenever $\|z\|_X \leq \delta$. Now for any $x \in X$, we rescale it by letting $z = (\delta/\|x\|_x)x$. Then $\|z\| = \delta$. Hence we have $\|Az\|_Y \leq \epsilon$. Or equivalently,

$$\left\|A\left(\frac{\delta}{\|x\|_X}x\right)\right\|_Y \leq \epsilon, \text{ or } \|Ax\|_Y \leq \frac{\epsilon}{\delta}\|x\|_X.$$

□

The operator norm of a bounded linear operator is define to be

$$\|A\| := \sup \frac{\|Ax\|_Y}{\|x\|_X} = \sup_{\|x\|_X=1} \|Ax\|_Y.$$

Examples

1. In $\mathbb{R}^n$, let $l(x) := a \cdot x$, where a an n-vector. If we use the $\|\cdot\|_2$ norm, then the corresponding operator norm of l is exactly $\|a\|_2$.

2. What is the corresponding operator norm of the operator $l(x) = a \cdot x$ in ℓ^p for $1 \le p \le \infty$?

3. Let $a \in [0,1]$. The mapping $u \mapsto u(a)$ is a bounded mapping from $C[0,1] \to \mathbb{R}$. Find its operator norm. Let us denote this operator by δ_a. What is the operator norm corresponding to $\alpha\delta_a + \beta\delta_b$, where $a, b \in [0,1]$ and $\alpha, \beta \in \mathbb{R}$.

4. Let A be an $m \times n$ matrix mapping from$\mathbb{R}^n$ to itself. Find the corresponding operator norm of A when $\mathbb{R}^n$ is equipped with ℓ^1-norm. Do the same thing if $\mathbb{R}^n$ is equipped with ℓ^∞. Find the operator norm of the mapping $Ku(x) = \int_0^1 g(x,y)u(y)\,dy$ in L^1, L^2 and $C_0[0,1]$.

5. The mapping
$$Ku(x) := \int_0^x u(y)\,dy$$
is a bounded mapping from $C[0,1]$ to itself. It is also a bounded mapping from $L^p(0,1)$ to itself.

6. Let $g(x,y) : [0,1] \times [0,1] \to \mathbb{R}$ be continuous. The operator
$$Ku(x) := \int_0^1 g(x,y)u(y)\,dy$$
is a bounded operator from $C[0,1]$ to itself. We may also think K is a mapping from $L^1(0,1)$ to itself. Find the corresponding operator norm. Do the same thing for $L^\infty(0,1)$.

7. A concrete example is
$$g(x,y) = \begin{cases} x(1-y) & \text{for } 0 \le x \le y \le 1 \\ y(1-x) & \text{for } 0 \le y \le x \le 1. \end{cases} \tag{2.1}$$

8. The differentiable operator D maps e^{inx} to ine^{inx}. Then $\{e^{inx}\}$ is a bounded sequence in $C(\mathbb{T})$, but $\{De^{inx}\}$ is not. Thus D is not bounded in $C(\mathbb{T})$. Same proof for all $L^p(\mathbb{T})$ as we treat D in $L^p(\mathbb{T})$.

Kernel and Range Let X and Y be normed linear spaces and let $A : X \to Y$ be a linear map. The kernel $N(A) := \{x \in X | Ax = 0\}$, and the range $R(A) := \{Ax | x \in X\}$.

1. A is 1-1 if and only if $N(A) = \{0\}$.

2. If A is bounded, then $N(A)$ is closed.

3. Let the matrix $A = (a_1, \cdots, a_n)$, where $a_1, .., a_n$ be column vectors in $\mathbb{R}^m$. Let the operator $Ax := \sum_{j=1}^n x_j a_j$ is a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$. Then $R(A) = \text{Span}\{a_1, ..., a_n\}$.

4. Let
$$g(x,y)=\sum_{i=1}^n \psi_i(y)\phi_i(x)$$
The operator
$$Ku(x)=\int_0^1 g(x,y)u(y)\,dy=\sum_{i=1}^n\int_0^1 \psi_i(y)u(y)\,dy\phi_i(x)$$
is a projection of u onto the space spanned by $\{\phi_1,...,\phi_n\}$. It is a bounded operator. Furthermore, both the kernel and range are closed in $C[0,1]$.

5. The shift operator from $\ell^\infty(\mathbb{Z})$ to itself defined by
$$(Tx)_n=x_{n+1}.$$
The shift operator is a bounded operator. Further, $N(T)=\{0\}$ and $R(T)=\ell^\infty(\mathbb{Z})$. However, in $\ell^\infty(\mathbb{N})$, we define
$$(Tx)_n=x_{n+1} \text{ for } n\ge 1.$$
In this case,
$$N(T)=\{(x_1,0,0,...)|x_1\in\mathbb{R}\}.$$
and $R(T)=\ell^\infty(\mathbb{N})$.

6. Consider $Ku=\int_0^x u(y)\,dy$ in $C[0,1]$. Then $N(K)=\{0\}$. For $\int_0^x u(y)\,dy=0$ implies $u\equiv 0$. But
$$R(A)=\{u\in C^1[0,1]|u(0)=0\}$$
which is not closed in $C[0,1]$.

7. In the space $C[0,1]$, consider $Ku=\int_0^x u(y)\,dy$ and $A=I+K$. Then $Au=0$ implies $u(x)+\int_0^x u(y)\,dy=0$. Differentiate it in x, we obtain $u'+u=0$. This leads to $u(x)=Ce^{-x}$. Thus, $N(A)=\{Ce^{-x}|C\in\mathbb{R}\}$. Notice that if we restrict to the space $\{u\in C[0,1]|u(0)=0\}$, then $N(A)=\{0\}$.

   Next, for any $f\in C[0,1]$, we look for a solution $u\in C[0,1]$ such that $Au=f$. Formally, we differentiate $Au=f$ and get
$$u'+u=f'.$$
   By using integration factor, we get
$$(e^y u)'=e^y f'.$$
   Integrate this equation, we get
$$e^x u(x)-u(0)=\int_0^x e^y f'(y)\,dy=e^x f(x)-f(0)-\int_0^x e^y f(y)\,dy.$$

Thus,

$$u(x) = e^{-x}(u(0) - f(0)) + f(x) - \int_0^x e^{-x+y} f(y)\, dy.$$

In this expression, we don't need to require f' exists. Thus, $R(A) = C[0,1]$.

8. Similar to the above. Consider $Ku = \int_0^1 g(x,y)u(y)\, dy$ where g is given by (2.1) and $A = I + K$. Then $N(A) = \{0\}$ and $R(A) = C[0,1]$.

Open mapping theorem The following theorem due to Banach is not so obvious in infinite dimensions. See Lax's Functional Analysis for proof.

Theorem 2.2 (Open mapping theorem)**.** *Let X and Y are two Banach spaces. If $A : X \to Y$ is bounded and onto, then A is an open map, which means that it maps open sets to open sets.*

Example

1. The shift operator (after quotion the kernel) is a 1-1 onto bounded linear map.
2. The Fredholm operator $I + K$ (K is an integral operator) is also such kind of operators.

The open mapping theorem is equivalent to the following bounded inverse theorem.

Theorem 2.3 (Bounded inverse theorem)**.** *Let X and Y be two Banach spaces. If $A : X \to Y$ is a bounded linear and bijective, then A^{-1} is also bounded.*

In fact, we can write it in a little more general form, and the bounded inverse theorem is a corollary of it.

Theorem 2.4. *Let X and Y be Banach spaces. Let $A : X \to Y$ be bounded and linear. Then the following two statements are equivalent:*

(a) A is 1-1 and closed range;

(b) there exists a constant C such that $\|Ax\| \geq C\|x\|$ for all $x \in X$.

Proof.

1. Let $Y_1 = R(A)$. If (a) holds, then by the open mapping theorem, $A^{-1} : Y_1 \to X$ is a bounded linear map. Thus, there exists a constant C such that for and $y \in R(A)$, we have $\|A^{-1}y\| \leq C\|y\|$. Thus, for any $x \in X$, $Ax \in R(A)$, we have $\|x\| \leq C\|Ax\|$.
2. Conversely, Suppose $\{y_n\}$ is a sequence in $R(A)$ and $y_n \to y$. Then there exist $x_n \in X$ such that $y_n = Ax_n$. From $\|x\| \leq C\|Ax\|$, we get that $\{Ax_n\}$ is a Cauchy sequence implies that $\{x_n\}$ is also a Cauchy sequence. Hence $x_n \to x$ for some $x \in X$. By the continuity of A, we get $Ax = \lim Ax_n = \lim y_n = y$. Thus, $R(A)$ is closed.

□

This is maily used for Fredholm opertors.

The open mapping theorem, the bounded inverse theorem and the following closed graph theorem are equivalent.

Definition 2.15. *A mapping $T : X \to Y$ is called closed graph if its graph $\{(x, Tx) | x \in X\}$ is closed in $X \times Y$.*

Theorem 2.5 (Closed graph theorem)**.** *Let X and Y be Banach spaces. A linear map $T : X \to Y$ which is closed graph is also a bounded map.*

Homeworks 2.1.

1. *Suppose a set $A \subset \mathbb{R}$ has a lower bound. Show that $m = \liminf A \Leftrightarrow$ if and only if for any $\epsilon > 0$, (a) all $x \in A$ but finite many satisfies $m - \epsilon < x$ and (b) there exists at least one $x \in A$ such that $x < m + \epsilon$.*

2. *Show that a function $f : (X, d) \to \mathbb{R}$ is lower semi-continuous if and only if its epigraph is closed in $X \times \mathbb{R}$.*

3. *Is $(C(a, b), \| \cdot \|_\infty)$ complete?*

4. *Show that $\|x\|_p \to \|x\|_\infty$ as $p \to \infty$ for $x \in \mathbb{R}^n$ and $\ell^p(\mathbb{N}) \cap \ell^\infty(\mathbb{N})$.*

5. *Find the operator norm of a matrix $A : (\mathbb{R}^n, \| \cdot \|_1) \to (\mathbb{R}^n, \| \cdot \|_1)$. Do the same thing by replacing $\| \cdot \|_1$ by $\| \cdot \|_\infty$.*

6. *Let $g(x, y) : [0, 1] \times [0, 1] \to \mathbb{R}$ be continuous. Show that the operator*

$$Ku(x) := \int_0^1 g(x, y)u(y)\, dy$$

is a bounded operator from $C[0, 1]$ to itself. Find the corresponding operator norm.

Chapter 3

Method of contraction mapping

3.1 Motivation

An important technique to solve systems of equations, ODEs, PDEs is to construct an iterative procedure to generate approximate solutions and find their limit. If the iterative procedure is given by

$$x_{n+1} = Tx_n$$

for some map T, then we look for a fixed point of T. The simplest case is when T is a contraction map, which means $|Tx - Ty| \le \rho|x - y|$ for some $0 < \rho < 1$. In this case, the iterators $\{x_n\}$ converges linearly.

Example Let us solve $x^2 = a$, $a > 0$ by Newton's method, which is an iterative method. Suppose we have found the nth iterator x_n, which is an approximation of the root, we approximate the equation $f(x) = x^2 - a = 0$ by a linear equation

$$g(x) := f(x_n) + f'(x_n)(x - x_n) = 0$$

It is easy to solve the linear equation. Its root is our next iterator x_{n+1}. That is

$$x_{n+1} = x_n - f'(x_n)^{-1} f(x_n).$$

With this, we get

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right).$$

We define

$$Tx = \frac{1}{2}\left(x + \frac{a}{x}\right).$$

Then the solution of $x^2 = a$ is a fixed point of T.

In order to show that T has a fixed point, we check

- $T : [\sqrt{a}, \infty) \to [\sqrt{a}, \infty)$

- $|Tx - Ty| = \frac{1}{2}\left|x + \frac{a}{x} - y - \frac{a}{y}\right| = \frac{1}{2}\left|1 - \frac{a}{xy}\right||x-y|$. We see that if both $x, y \geq \sqrt{a}$, then $|1 - \frac{a}{xy}| < 1$.

Hence, T is a contraction map from $[\sqrt{a}, \infty) \to [\sqrt{a}, \infty)$. Then one can show that the sequence generated by the iteration

$$x_{n+1} = Tx_n$$

converges. Such a method is called method of contraction maps. In this chapter, we will give a general theory and provide many applications.

Remark If the initial point $0 < x_0 < \sqrt{a}$, then one can show that $x_1 > \sqrt{a}$. Then the following iterations fall into the region of convergence $[\sqrt{a}, \infty)$. Thus we conclude that for any $x_0 > 0$, the Newton's method converges for this case.

Exercise What is the value of the following continued fraction?

$$x = \frac{1}{a + \frac{1}{a + \frac{1}{a + \cdots}}}.$$

Applications

- Intermediate value theorem
- Inverse function theorem
- Jacobi method for solving systems
- Local existence of ODEs
- global existence of ODEs
- Existence of stable manifold
- Integral equations

3.2 Method of contraction mapping

The method of contraction mapping in metric space was proposed by Banach in 1921 in his Ph.D thesis. It is an abstract setting of the method of iteration which was developed long ago by Jacobi, Gauss, Picard, etc.

Definition 3.1. *Let (X, d) be a complete metric space. A mapping $T : X \to X$ is called a contraction mapping if there exists a constant $0 \leq \rho < 1$ such that*

$$d(Tx, Ty) \leq \rho\, d(x, y).$$

It is easy to see that T is continuous in X.

Theorem 3.1 (Banach fixed-point theorem)**.** *If $T : (X, d) \to (X, d)$ is a contraction map, then it has a unique fixed point $\bar{x} \in X$*

Proof. We start from any point x_0 and generate the iterates $x_{n+1} = Tx_n$ for $n \geq 0$. We show that $\{x_n\}$ is a Cauchy sequence in X. We have for $n > m$,

$$\begin{aligned} d(x_n, x_m) &= d(T^{m+(n-m)}x_0, T^m x_0) \\ &\leq \rho^m d(T^{n-m}x_0, x_0) \\ &\leq \rho^m \left[d(T^{n-m}x_0, T^{n-m-1}x_0) + d(T^{n-m-1}x_0, T^{n-m-2}x_0) + \cdots + d(Tx_0, x_0) \right] \\ &\leq \rho^m \left[\sum_{k=0}^{n-m-1} \rho^k \right] d(x_1, x_0) \\ &\leq \left(\frac{\rho^m}{1-\rho} \right) d(x_1, x_0). \end{aligned}$$

This shows that $\{x_n\}$ is a Cauchy sequence. From the completeness of X, we get that $\{x_n\}$ has a limit $\bar{x} \in X$. By taking the limit $n \to \infty$ in the equation $x_{n+1} = Tx_n$, we get $T\bar{x} = \bar{x}$.

If $\bar{x}$ and $\bar{y}$ are two fixed points of T in X, then

$$d(\bar{x}, \bar{y}) = d(T\bar{x}, T\bar{y}) \leq \rho d(\bar{x}, \bar{y})$$

Since $0 \leq \rho < 1$, we get $d(\bar{x}, \bar{y}) = 0$ and thus $\bar{x} = \bar{y}$. □

One can estimate the convergence rate of these iterates.

$$\begin{aligned} d(x_n, \bar{x}) &= d(Tx_{n-1}, T\bar{x}) \\ &\leq \rho\, d(x_{n-1}, \bar{x}) \\ &\vdots \\ &\leq \rho^n d(x_0, \bar{x}) \end{aligned}$$

This shows that the geometric convergence like a power sequence because $\rho < 1$.

In many applications, the equations we want to solve depend on a parameter. For example, solving $F(x, \lambda) = 0$, where λ is a parameter. In this case, we want to know how the corresponding solution depends on the parameter. For instance, we want to know how the solution of $x^2 = a$ depends on a. Let us call the parameter λ and denote the parameter space by Λ, which is assumed to be a metric space.

Theorem 3.2. *Consider a parameter-dependent contraction mapping $T : \Lambda \times X \to X$ such that*

(1) T is continuous in both λ and x

(2) For each $\lambda \in \Lambda$, $T(\lambda, \cdot)$ is a contraction with contraction ratio ρ and $0 \leq \rho < 1$ independent of λ.

Then the fixed point $x(\lambda)$ of the contraction mapping $T(\lambda,\cdot)$ is also continuous in $\lambda \in \Lambda$.

Proof. Let $x_n(\lambda)$ be the iterators generated by $T(\lambda,\cdot)$ starting from an arbitrarily chosen initial $x_0(\lambda)$, which is required to be continuous. From the continuity of T, we get each $x_n(\cdot)$ is also continuous. Since the contraction ratio ρ is independent of λ, we have that the convergence of $x_n(\lambda)$ in n is uniform with respect to λ. The limits of $n \to \infty$ and $\lambda \to \lambda_0$ can be interchanged. Thus, the limit function $\lim_{n\to\infty} x_n(\lambda)$ is a continuous function. □

Exercise Consider the discrete logistic map

$$x_{n+1} = \lambda x_n(1-x_n) := F_\lambda(x_n).$$

What is the value of λ in which $x = 0$ is the fixed point? What are the fixed points of $F_\lambda^2 := F_\lambda \circ F_\lambda$?

3.3 Solving large linear systems

It is impractical to solve large linear system

$$Ax = b$$

by so-called direct methods such as Gaussian elimination or LU decomposition because the number of operations required is of $O(N^3)$, where N is the size of x. Instead, iterative methods are usually favored. The main idea is to sacrifice a little bit accuracy, but gain the speed of convergence. This is harmless because many such systems are obtained from PDEs, in which a discretization error has already been introduced. One important class of iterative methods is to decompose A into

$$A = M - N,$$

and to solve it by the following iterative procedure

$$Mx^{n+1} - Nx^n = b.$$

The matrix M is the major part and N the minor part of A, respectively. It also requires that M is easily to invert.

One class of matrix which is easy to find such a decomposition is the diagonally dominant matrices:

$$|a_{ii}| > \sum_{j\neq i} |a_{ij}|, \text{ for all } i.$$

In this class, we choose $M = \text{diag}(a_{11},\cdots,a_{nn})$ and $N = A - M$. We can solve $Ax = b$ by

$$Mx^{n+1} = Nx^n + b$$

or,

$$x^{n+1} = M^{-1}Nx^n + M^{-1}b.$$

The sequence $\{x^n\}$ converges if the operator norm of the mapping $T := M^{-1}N$ (called the iteration operator) is less than 1 by the fixed point theorem. We estimate $\|T\|$ as the follows.

$$T_{ij} = -\begin{pmatrix} 0 & \frac{a_{12}}{a_{11}} & \frac{a_{13}}{a_{11}} & \cdots & \frac{a_{1n}}{a_{11}} \\ \frac{a_{21}}{a_{22}} & 0 & \frac{a_{23}}{a_{22}} & \cdots & \frac{a_{2n}}{a_{22}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{a_{n1}}{a_{nn}} & \frac{a_{n2}}{a_{nn}} & \frac{a_{n3}}{a_{nn}} & \cdots & 0 \end{pmatrix}.$$

The operator $T : (\mathbb{R}^n, |\cdot|_\infty) \to (\mathbb{R}^n, |\cdot|_\infty)$ has the following estimate

$$\begin{aligned} |Tx|_\infty &= \max_i \sum_j |T_{ij}x_j| \\ &\le \max_i \sum_j |T_{ij}|(\max_k |x_k|) \\ &= \max_i \sum_j |T_{ij}||x|_\infty. \end{aligned}$$

By our assumption,

$$\max_i \sum_j |T_{ij}| = \rho < 1.$$

Thus, $\|T\| = \rho < 1$.

3.4 Solving system of algebraic equations

Suppose we want to solve the equation $f(x) = 0$ in $\mathbb{R}^n$, where $f : \mathbb{R}^n \to \mathbb{R}^n$ is a C^1 function. We want to design an iterative procedure to solve this equation. The iteration is to design $c \ne 0$ such that the iterators

$$x_{n+1} = x_n - c(x_n)f(x_n)$$

converge. Or in other words, we want the map

$$Tx = x - c(x)f(x)$$

to be a contraction map. You may think that $c(x_n)$ is a step size of a discretized ODE: $\dot{x} = -f(x)$. We see that a root of $f(x) = 0$ corresponds to a fixed point of T. For example, suppose $f'(\bar{x})$ is non-singular, where $\bar{x}$ is the root. Then we can choose

$$c(x) = f'(\bar{x})^{-1}$$

In this case,

$$T'(\bar{x}) = 0.$$

By the continuity of T', we can have a small domain in the neighborhood of $\bar{x}$ such that T is contraction. In practice, we don't known $\bar{x}$, nor $f'(\bar{x})$. But we may know a rough estimate of $f'(\bar{x})$. If so, we use this one for $c(x)$. We can also choose $c(x) = f'(x)^{-1}$. This leads to the Newton's method.

Theorem 3.3 (Inverse function theorem). *If $f(x_0) = y_0$ and f is C^1 in a neighborhood of x_0 with $f'(x_0)$ being nonsingular. Then there exist a small neighborhood V of y_0 and a small neighborhood U of x_0 such that $f : U \to V$ is invertible.*

Proof. To find the inverse map of f for any $y \sim y_0$, we need to solve

$$f(x) = y$$

in a neighborhood of x_0. We may express $y = y_0 + r$ and $x = x_0 + e$. Using $f(x_0) = y_0$, we get the perturbed equation

$$f(x_0 + e) - f(x_0) = r$$

for $e \sim 0$. Suppose $f'(x_0) = A$. Using Taylor expansion, we get

$$Ae + g(e) = r$$

where $g(e) = f(x_0 + e) - f(x_0) - Ae = o(e)$. We design the iteration procedure

$$Ae^{n+1} + g(e^n) = r,$$

or

$$e^{n+1} = -A^{-1}g(e^n) + A^{-1}r := Te^n$$

to solve the perturbed equation.

In order to apply the method of contraction, we want to find $\eta > 0$ and $\delta > 0$ such that for any $|r| < \eta$, we have T is a contraction from $|e| \le \delta$ into itself. If $|r| < \eta$ and $|e| \le \delta$, then

$$|Te| \le \|A^{-1}\|(|g(e)| + |r|) \le \|A^{-1}\|(o(\delta) + \eta).$$

We require $|Te| \le \delta$. This gives

$$\|A^{-1}\|(o(\delta) + \eta) \le \delta.$$

On the other hand,

$$\begin{aligned} Te_1 - Te_2 &= A^{-1}(g(e_1) - g(e_2)) \\ &= A^{-1}(f(x_0 + e_1) - f(x_0 + e_2) - A(e_1 - e_2)) \\ &= A^{-1}((f'(\tilde{x}) - A)(e_1 - e_2)) \end{aligned}$$

Here, we have used $f \in C^1$ in a neighborhood of x_0 and applied the mean value theorem, and $\tilde{x}$ is a point between $x_0 + e_1$ and $x_0 + e_2$. From the continuity of $f'(x)$ at x_0, we can choose $\delta > 0$ such that whenever $|e_1|, |e_2| \le \delta$, we can get

$$\|A^{-1}\|\|f'(\tilde{x}) - f'(x_0)\| \le 1/2.$$

This gives another constraint on δ. With these two constraints, we choose δ and η such that T is a contraction from $|e| \le \delta$ to itself for every $|r| < \eta$. □

Remarks.

1. The inverse function f^{-1} of the above theorem is also in C^1 and $(f^{-1})' = (f')^{-1}$.

2. The assumption of the existence of the inverse function can be relaxed to $f'(x)$ is continuous at x_0 instead of $f \in C^1$ in a neighborhood of x_0.

3. A corollary of the inverse function theorem is the implicit function theorem. It states: Let $F : \mathbb{R}^{n+m} \to \mathbb{R}^m$ be in C^1 in a neighborhood of (x_0, y_0), and at which $F(x_0, y_0) = 0$. Suppose $F_y(x_0, y_0)$ is invertible. Then there exists a unique function $y = g(x)$ in a neighborhood of x_0 such that $F(x, g(x)) = 0$ in this neighborhood. Moreover, $g \in C^1$ and $g'(x) = F_y^{-1} F_x(x, g(x))$.

3.5 Solving ODEs

Consider the ODE

$$y'(t) = f(t, y)$$

with initial condition

$$y(t_0) = y_0.$$

Let $I = \{|t - t_0| \leq T\}$ with some $T > 0$. We assume $f : I \times \bar{B}_R(y_0) \to \mathbb{R}^n$ is continuous in (t, y) and is Lipschitz continuous in y uniformly with respect to $t \in I$. This means that there exists a constant $L \geq 0$ such that for any $y_1, y_2 \in \bar{B}_R$ and $t \in I$.

$$|f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2|.$$

Theorem 3.4. *Suppose f is as above. Then there exist a $\delta > 0$ and a unique solution of the above ODE for $|t - t_0| \leq \delta$.*

Proof. Let $J := [t_0 - \delta, t_0 + \delta]$, where $\delta > 0$ is to be determined. Consider the space $C(J) = \{y : J \to \mathbb{R}^n$ is continuous.$\}$ equipped with the sup norm

$$\|y\|_\infty := \sup_{t \in J} |y(t)|.$$

Then $C(J)$ is a Banach space. For any $y(\cdot) \in C(J)$, we consider the mapping

$$Ty = y_0 + \int_{t_0}^{t} f(s, y(s))\, ds.$$

Notice that Ty is still a continuous function. Consider the closed ball,

$$X := \{y \in C(J) | \|y - y_0\|_\infty \leq R\}$$

We shall choose a δ properly such that T is a contraction mapping from X to X. Firstly, let

$$M := \sup\{|f(t, y)| | t \in I, |y - y_0| \leq R\}.$$

We have

$$|Ty(t) - y_0| = |\int_{t_0}^{t} f(s, y(s))\, ds| \le M\delta.$$

If we choose $M\delta \le R$, then T maps X into X. Secondly,

$$\begin{aligned} \|Ty_1 - Ty_2\|_\infty &= \sup_{t\in J} |\int_{t_0}^{t} f(s, y_1(s)) - f(s, y_2(s))\, ds| \\ &\le \sup_{t\in J} \int_{t_0}^{t} L|y_1(s) - y_2(s)|\, ds| \\ &\le L\delta\|y_1 - y_2\|_\infty \end{aligned}$$

This gives another condition on δ. Combining the two conditions, we can choose

$$\delta = \min\left\{\frac{R}{M}, \frac{1}{2L}\right\},$$

then T is a contraction from X to X with contraction ratio $1/2$.

From the contraction mapping theorem, we get a unique solution. Such a solution satisfies the integral equation

$$y(s) = y_0 + \int_{t_0}^{t} f(s, y(s))\, ds.$$

We see that $y(\cdot) \in C(J)$ implies $f(\cdot, y(\cdot))$ is continuous. Hence $\int_{t_0}^{t} f(s, y(s))\, ds$ is continuously differentiable. Thus, y is continuously differentiable in J^o. We differentiate this integral equation in t and get $y'(t) = f(t, y(t))$. □

Remarks.

1. The iteration $y^{n+1} = Ty^n$ in the proof of the existence of ODE is called Picard iteration. A rationael behind this iteration is the follows. For "short time", the term y' is the major term, while $f(t, y)$ is a minor term. Indeed, if we perform a rescaling: $t = \epsilon\hat{t}$, then the rescaled equation becomes

$$\frac{1}{\epsilon}\frac{dy}{d\hat{t}} = f(\epsilon\hat{t}, y).$$

Comparing the two terms above, the left-hand side is "more important" than the right-hand side for short period of time. More precisely, we can rewrite the differential equation in integral form

$$y(t) - y_0 = \int_{t_0}^{t} f(s, y(s))\, ds$$

Among these terms, the right-hand side is relatively less important. Thus, the iteration procedure becomes

$$y_{n+1}(t) = y_0 + \int_{t_0}^{t} f(s, y_n(s))\, ds.$$

2. In the above local existence, the existence time period δ depends only on R, M and L. If f is *local Lipschitz* in y[1], then as long as the solution can exist at $(t_1, y(t_1))$ it can always be extended. This leads to that either the trajectory $y(t) \to \infty$ in finite time, or it stays bounded as $t \to \infty$. Thus, a global existence relies on so called *a priori estimate*. If we can show that for any $T > 0$, there exists $R > 0$ (which can depend on T) such that

$$\|y(t)\| \le R, \text{ for } 0 \le t \le T.$$

   This together with the local existence gives global solution for $0 \le t < \infty$. Such an estimate is called *a priori estimate*. We shall study a priori estimate later.

3. The uniqueness of the integral equation

$$y(t) = y_0 + \int_{t_0}^{t} f(s, y(s))\, ds$$

   follows from the uniqueness of the fixed point. One can also use the Gronwall inequality to prove the uniqueness of the ODE directly, as shown below. Suppose y_1 and y_2 are two C^1 solutions with the same initial data at t_0. Then

$$y_1' = f(t, y_1), \;\; y_2' = f(t, y_2).$$

   Subtracting these two, we get

$$|y_1' - y_2'| \le |f(t, y_1) - f(t, y_2)| \le L|y_1 - y_2|.$$

   Notice that for a C^1-function η, we have $|\eta|' \le |\eta|'$, by triangle inequality. Choose $\eta(t) := |y_1(t) - y_2(t)|$. Then

$$\eta' \le L\eta.$$

   Use method of integration factor, we get

$$\left(e^{-Lt}\eta\right)' \le 0.$$

   This leads to

$$\eta(t) \le \eta(t_0) e^{L(t-t_0)}.$$

   Since $\eta(t_0) := y_1(t_0) - y_2(t_0) = 0$, we obtain $\eta(t) \equiv 0$. This proves the uniqueness.

4. Many local existence theorems in PDE theory are based on the method of contraction mapping. For example, local existence of systems of hyperbolic equations in one dimension, see F. John, PDE, systems of first order hyperbolic equations, method of characteristics.

[1] This means that for any y_0 there exists $R > 0$ such that f is Lipschitz in $\overline{B_R(y_0)}$.

Examples

1. The ODE $y' = \sqrt{y}$ with $y(0) = 0$ is not unique.

2. How about the ODE: $y' = 1/\ln y$ with $\lim_{t\to 0+} y(t) = 0$?

3. The solution $y' = y^2$ blows up for any $y(0) = y_0 > 0$.

3.6 Continuous dependence on parameter of solutions of ODE

Let us denote the solution of the ODE with initial data (τ, ξ) by $y(\cdot, \tau, \xi)$. That is, $y(\tau, \tau, \xi) = \xi$.

Theorem 3.5. *Under the same assumption of f in the local existence theorem above, the solution $y(t, \tau, \xi)$ of the ODE $y' = f(t, y)$ is a continuous function in (t, τ, ξ) in a neighborhood of (t_0, t_0, y_0). That is, the solution $y(\cdot, \tau, \xi)$ continuously depends on its initial data (τ, ξ).*

Proof. 1. Following the proof of the local existence theorem, let $y^*(\cdot) := y(\cdot, t_0, y_0)$. There exists ϵ_1 such that the set

$$\{(\tau, y) | |\tau - t_0| \le \delta, |y - y^*(\tau)| \le \epsilon_1\} \subset J \times \bar{B}_R(y_0).$$

We choose $\epsilon = \epsilon_1 e^{-2L\delta}$ and let

$$\bar{U} := \{(\tau, \xi) | |\xi - y^*(\tau)| \le \epsilon, \tau \in J\}.$$

We define the space

$$X = \{y : J \times \bar{U} \to \mathbb{R}^n \text{ in } C \text{ and } d(y, y^*) \le \epsilon\}.$$

where the metric d is defined by

$$d(y_1, y_2) := \sup_{t\in J, (\tau,\xi)\in\bar{U}} e^{-2L|t-\tau|} |y_1(t, \tau, \xi) - y_2(t, \tau, \xi)|.$$

2. For $y \in X$, define

$$(Ty)(t, \tau, \xi) := \xi + \int_\tau^t f(s, y(s, \tau, \xi))\, ds.$$

3. We show that $Ty \in X$ if $y \in X$. Given $y \in X$, we have

$$\begin{aligned}
|Ty(t,\tau,\xi) - y^*(t)| &= \left|\xi - y^*(\tau) + \int_\tau^t f(s, y(s,\tau,\xi)) - f(s, y^*(s))\,ds\right| \\
&\le |\xi - y^*(\tau)| + |\int_\tau^t |f(s, y(s,\tau,\xi)) - f(s, y^*(s))|\ ds| \\
&\le \epsilon + |\int_\tau^t L|y(s,\tau,\xi) - y^*(s)|\,ds| \\
&\le \epsilon + |\int_\tau^t Le^{2L|s-\tau|}|\epsilon\,ds \\
&= \frac{\epsilon}{2} + \frac{\epsilon}{2}e^{2L|t-\tau|} \\
&\le \epsilon e^{2L|t-\tau|}.
\end{aligned}$$

   This proves $d(Ty, y^*) \le \epsilon$.

4. We show
$$d(Ty_1, Ty_2) \le \frac{1}{2}d(y_1, y_2).$$
   where $y_1, y_2 \in X$. Let us abbreviate $y_i(t,\tau,\xi)$ by $y_i(t)$.

$$\begin{aligned}
|Ty_1(t) - Ty_2(t)| &= |\int_\tau^t (f(s, y_1(s,\tau,\xi)) - f(s, y_2(s,\tau,\xi)))\,ds| \\
&\le |\int_\tau^t L|y_1(s) - y_2(s)|\,ds| \\
&\le |\int_\tau^t Le^{2L|s-\tau|}d(y_1, y_2)\,ds \\
&= \frac{1}{2}e^{2L|t-\tau|}d(y_1, y_2)
\end{aligned}$$

   Thus, we have
$$d(Ty_1, Ty_2) \le \frac{1}{2}d(y_1, y_2).$$

5. We apply the parameter-dependent fixed point theorem to get that the fixed point $y(t,\tau,\xi)$ is continuous in the parameter $(\tau,\xi) \in \bar{U}$.

□

Remark. In the above proof, the rationael behind the definition of the weighted distance is the follows. In the estimate of y, we control f by a linear function (Using Lipschitz continuity), we then expect that the growth of y is controlled by a factor $e^{L|t-\tau|}$ at time t. Second, the factor 2 in $e^{2L|t-\tau|}$ gives us a room to control the contraction ratio to be $1/2$.

3.7 *Local structure of ODE near a hyperbolic point

Consider the ODE

$$\dot{x} = f(x)$$

where $x \in \mathbb{R}^2$. Suppose $f(0) = 0$, i.e. 0 is an equilibrium point. An equilibrium point is called hyperbolic if no eigenvalue of $f'(0)$ is pure imaginary. Let us consider the following simple example

$$\begin{cases} \dot{x}_1 = x_1 + g_1(x_1, x_2), \\ \dot{x}_2 = -x_2 + g_2(x_1, x_2). \end{cases}$$

Here $g_i(x) = O(|x|^2)$. The point $(0, 0)$ is a hyperbolic equilibrium. Let us consider the linearized equation of this example:

$$\begin{cases} \dot{x}_1 = x_1 \\ \dot{x}_2 = -x_2. \end{cases}$$

Its solutions are

$$\begin{cases} x_1(t) = x_1^0 e^t, \\ x_2(t) = x_2^0 e^{-t}. \end{cases}$$

The orbit starting from $x_2^0 = 0$ stays on $x_2 = 0$ and converge to 0 as $t \to -\infty$, while those starting from any point with $x_1^0 = 0$ go to 0 as $t \to \infty$. The line $x_2 = 0$ is called a unstable manifold, while the line $x_1 = 0$ is called a stable manifold. All other orbits forms part of a hyperbola: $x_1 x_2 = x_1^0 x_2^0$ and leave the unstable manifold and move toward the stable manifold as $t \to \infty$. This is the orbit structure of the linearized equation near the equilibrium.

In the theory of ODE, it can be shown that the structure of orbits of the nonlinear equations is closed to that of the linearized equation (i.e. $g_i = 0$, $i = 1, 2$). This persistence of the structure is called structure stability of a hyperbolic equilibrium. I shall not give a complete theory here. I will just show the existence of unstable manifold here, which is a key step of the stability theory.

Let $\phi^t(x)$ denote the solution of $\dot{x} = f(x)$ with $x(0) = x$. Given $\epsilon > 0$, a manifold M_ϵ^s is called a stable manifold at 0 if

$$W_\epsilon^s(0) = \{x | \, |\phi^t(x)| \le \epsilon \text{ for all } t \ge 0\}.$$

Similarly, a manifold M_ϵ^u is called an unstable manifold at 0 if

$$W_\epsilon^u(0) = \{x | \, |\phi^t(x)| \le \epsilon \text{ for all } t \le 0\}.$$

Such stable and unstable manifolds exist locally. Let us show the existence of unstable manifold below. This unstable manifold will be represented as $\eta = h(\xi)$. It has the properties: $h(0) = 0$ and $h'(0) = 0$. That is, it is tangent to the unstable manifold of the linearized equation T $(0, 0)$.

We start from an initial data $(x_1(0), x_2(0)) = (\xi, \eta)$ for the ODE:

$$\begin{cases} \dot{x}_1 = x_1 + g_1(x_1, x_2) \\ \dot{x}_2 = -x_2 + g_2(x_1, x_2), \end{cases}$$

Using method of integration factor, the first equation becomes

$$(e^{-t} x_1)' = e^{-t} g_1(x_1, x_2).$$

Integrate it, we get

$$x_1(t) = e^t\xi + \int_0^t e^{t-s} g_1(x_1(s), x_2(s))\, ds.$$

For the second equation, we use the integration factor e^t and integrate it from $-\infty$ to t. This is because we expect $x_2(t) \to 0$ as $t \to -\infty$. Hence we have

$$x_2(t) = \int_{-\infty}^t e^{-t+s} g_2(x_1(s), x_2(s))\, ds.$$

Now we consider the space

$$X_\epsilon = \{x : (-\infty, 0] \to \mathbb{R}^2, \|x\|_\infty \le \epsilon\}$$

equipped with the $\|\cdot\|_\infty$ and the mapping

$$(T_\xi x)(t) = \left(\begin{array}{l} e^t\xi + \int_0^t e^{t-s} g_1(x_1(s), x_2(s))\, ds \\ \int_{-\infty}^t e^{-t+s} g_2(x_1(s), x_2(s))\, ds \end{array} \right).$$

Assume $|\xi| \le \epsilon/2$. Using $g_i(x) = O(|x|^2)$ (say $|g_i| \le C|x|^2$), we get

$$|T_\xi x(t)| \le \epsilon/2 + C\|x\|_\infty^2, \text{ for } t \le 0.$$

Thus, we choose ϵ such that

$$\frac{\epsilon}{2} + C\epsilon^2 \le \epsilon,$$

then T_ξ maps X_ϵ into X_ϵ.

Next, it is easy to check T_ξ is a contraction mapping in X_ϵ with contraction ratio $1/2$ if we choose ϵ with $C\epsilon \le 1/2$. Hence T_ξ has a fixed point (x_1^*, x_2^*) in X_ϵ. This solution satisfies the condition:

$$x_1^*(0) = \xi,\ x_2^*(t) \to 0 \text{ as } t \to -\infty.$$

Let

$$\eta := x_2^*(0) = \int_{-\infty}^0 g_2(x_1^*(s), x_2^*(s))\, ds. \tag{3.1}$$

This defines η as a function of ξ, say $\eta = h(\xi)$. Then $h(\cdot)$ is defined on $(-\epsilon/2, \epsilon/2)$. Its graph is the unstable manifold M_ϵ^u. Namely, for $(\xi, \eta) \in M_\epsilon^u$, the solution of the above ODE with $x_1^*(0) = \xi$ and $x_2^*(0) = \eta$ tends to $(0, 0)$ as $t \to -\infty$.

Remark. In fact, $h(\cdot)$ is an analytic function. Further, $h'(0) = 0$. I leave the proof of $h'(0) = 0$ to the reader because it is not a direct application of the contraction mapping.

Example Consider the following example

$$\begin{array}{rcl}\dot{x} & = & -x+x^2-2xy+y^3\\ \dot{y} & = & y+x^3.\end{array}$$

The stable manifold can be expressed as $\eta = h(\xi)$ near $(0,0)$. On this manifold,

$$\dot{\eta} = h'(\xi)\dot{\xi}.$$

This implies

$$\eta+\xi^3 = h'(\xi)(-\xi+\xi^2-2\xi\eta+\eta^3)$$

Since $h(\xi)$ is analytic and $h'(0)=0$, we can expand

$$h(\xi) = a_2\xi^2+a_3\xi^3+\cdots.$$

By plugging this expression into the above equation, we obtain

$$a_2\xi^2+a_3\xi^3+\xi^3 = (2a_2\xi+3a_3\xi^2)\cdot(-\xi+\xi^2-2\xi(a_2\xi^2)+(a_2\xi^2)^3)+O(\xi^4).$$

Equate the coefficients of the same powers, we get

$$a_2=-2a_2,\ a_3+1=2a_2-3a_3.$$

Thus, $a_2=0$ and $a_3=-1/4$.

Remark. Consider the ODE: $\dot{x}=f(x)$ in $\mathbb{R}^n$ and suppose 0 is an equilibrium point. That is, $f(0)=0$. The spectra $\sigma(f'(0))$ is the set of all eigenvalues of $f'(0)$. They are classified into σ_s, σ_u and σ_c depending the real part of the eigenvalue λ is less tha, greater than or equals 0, respectively. Suppose the invariant spaces corresponding to σ_s, σ_u and σ_c are Π_u, Π_s and Π_c, respectively. Just similar to the stable manifold and unstable manifold, we can also define the centered manifold as the follows

$$M_\epsilon^c := \{x||\phi^t(x)|\le\epsilon, \forall t\in\mathbb{R}\}$$

where ϕ^t is the solution of the above ODE with $\phi^0(x)=x$. The centered manifold theorem states that M_ϵ^c exists locally, has the representation:

$$\eta = h(\xi)$$

for $\xi\in\Pi_c$ and $\eta\in\Pi_s+\Pi_u$. Moreover, the tangent space of M_ϵ^c at 0 is Π_c.

3.8 A priori estimate for solutions of ODE

Global Lipschitz condition

Theorem 3.6. *Consider the ODE in $\mathbb{R}^n$:*

$$\dot{y} = f(t, y).$$

Assume f satisfies

$$|f(t, y)| \le a(t) + b(t)|y|,$$

where $a(\cdot)$ and $b(\cdot)$ are integrable on $[0, T]$ for some $T > 0$. Then the solution exists up to T and has the following estimate

$$|y(t)| \le e^{B(t)}|y(0)| + \int_0^t e^{B(t)-B(s)} a(s)\, ds$$

where $B(t) = \int_0^t b(s)\, ds$.

Proof. We have

$$|y|' \le |y'| \le |f(t, y)| \le a(t) + b(t)|y|.$$

Let $B(t) := \int_0^t b(s)\, ds$. Consider the integration factor $e^{-B(t)}$. We have

$$\left(e^{-B(t)}|y|\right)' \le e^{-B(t)} a(t).$$

Integrate this inequality from 0 to t, we get

$$e^{-B(t)}|y(t)| - |y(0)| \le \int_0^t e^{-B(s)} a(s)\, ds.$$

This leads to

$$|y(t)| \le e^{B(t)}|y(0)| + \int_0^t e^{B(t)-B(s)} a(s)\, ds$$

Thus, as long as $a(\cdot)$ and $b(\cdot)$ are integrable on $[0, T]$, then $|y(t)|$ remains bounded on $[0, T]$. Then from local existence theorem, it can be extended beyond T. Thus, the solution exists on $[0, \infty)$. □

Lyapunov functional

Theorem 3.7. *Consider the ODE in $\mathbb{R}^n$:*

$$y' = f(y),\ y(0) = y_0.$$

Suppose there exists a function Φ such that

$$\nabla\Phi(y) \cdot f(y) \le 0,$$

and $\Phi(y) \to \infty$ as $y \to \infty$. Then the solution exists on $[0, \infty)$.

Proof. Consider $\Phi(y(t))$. It is a non-increasing function because

$$\frac{d}{dt}\Phi(y(t)) = \nabla\Phi(y(t)) \cdot f(y(t)) \le 0$$

Thus,

$$\Phi(y(t)) \le \Phi(y(0))$$

Since $\Phi(y) \to \infty$ as $y \to \infty$, the set

$$\{y|\Phi(y) \le \Phi(y_0)\}$$

is a bounded set. If the maximal existence of interval is $[0, T)$ with $T < \infty$, then $y(\cdot)$ is bounded in $[0, T)$ and can be extended to T. By the local existence of ODE, we can always extend $y(\cdot)$ to $T + \epsilon$. This is a condiction. Hence $T = \infty$. □

As an example, let us consider a damping system

$$\ddot{x} + \gamma\dot{x} = -V'(x)$$

where V is a trap potential, which means that $V(x) \to \infty$ as $|x| \to \infty$. By multiplying $\dot{x}$ both sides, we obtain

$$\frac{dE}{dt} = -\gamma|\dot{x}|^2 \le 0$$

Here,

$$E(t) := \frac{1}{2}|\dot{x}|^2 + V(x)$$

is the energy. The term $\gamma|\dot{x}|^2$ is called the energy dissipation rate. We integrate the above equation from 0 to t, drop the dissipation term to get

$$E(t) \le E(0), \text{ for all } t > 0.$$

This gives a priori estimate of solution

$$\frac{1}{2}|\dot{x}(t)|^2 + V(x(t)) \le E(0).$$

This implies both $\dot{x}$ and x are bounded, because of the property of V.

3.9 Solving a simple boundary-value problem

We consider

$$-u'' + q(x)u = f, \; x \in (0, 1)$$
$$u(0) = 0, \; u(1) = 0.$$

Such a problem occur commonly in quantum mechanics, wave propagation, etc. The function u is called a wave function and q a potential. We can solve this equation for $q \equiv 0$ first. Then solve this equation for small q. Such a method is called a perturbation method.

To solve $-u'' = f$ with $u(0) = u(1) = 0$, we integrate it once to get

$$u'(y) = -\int_1^y f(s)\,ds + C_1.$$

Next, we integrate it from 0 to x and use $u(0) = 0$ to get

$$u(x) = -\int_0^x \int_1^y f(s)\,ds\,dy + C_1 x.$$

By integration-by-part,

$$\begin{aligned} -\int_0^x \int_1^y f(s)\,ds\,dy &= -\int_0^x F(y)\,dy \\ &= -\left[yF(y)\right]_0^x + \int_0^x yF'(y)\,dy \\ &= x\int_x^1 f(y)\,dy + \int_0^x yf(y)\,dy \end{aligned}$$

From $u(1) = 0$, we obtain

$$C_1 = -\int_0^1 yf(y)\,dy.$$

Hence

$$u(x) = \int_0^x y(1-x)f(y)\,dy + \int_x^1 x(1-y)f(y)\,dy$$

Let us define

$$g(x,y) := \begin{cases} x(1-y) & \text{if } 0 \le x \le y \le 1 \\ y(1-x) & \text{if } 0 \le y \le x \le 1. \end{cases}$$

Then the solution above can be represented as

$$u(x) = \int_0^1 g(x,y)f(y)\,dy.$$

Such a g is called the Green's function associated with the operator $-d^2/dx^2$ in $(0,1)$ with Dirichlet boundary condition. It satisfies

$$-\frac{d^2}{dx^2} g(x,y) = \delta(x,y), g(0,y) = g(1,y) = 0.$$

Next, we consider the perturbed problem:

$$\begin{gathered} -u'' + q(x)u = f,\ x \in (0,1) \\ u(0) = 0,\ u(1) = 0. \end{gathered}$$

We can rewrite it as an integral equation

$$u(x)+\int_0^1 g(x,y)q(y)u(y)\,dy=\int_0^1 g(x,y)f(y)\,dy.$$

This equation has the form

$$(I-K)u=h$$

where

$$Ku(x):=\int_0^1 g(x,y)q(y)u(y)\,dy$$

$$h(x):=\int_0^1 g(x,y)f(y)\,dy.$$

Now we can apply method of contraction to show the existence of the solution. We choose the metric space to be $X=\{u\in C[0,1]|u(0)=u(1)=0\}$. This is a complete metric space. We define $Tu=Ku+h$ for $u\in X$. It is easy to see that $Tu\in X$. To see T is a contraction, we have

$$\begin{aligned}
\|Ku\|_\infty &:= \sup_{0\le x\le 1}\left|\int_0^1 g(x,y)q(y)u(y)\,dy\right| \\
&\le \sup_{0\le x\le 1}\int_0^1 |g(x,y)q(y)|\,dy\|u\|_\infty \\
&\le \sup_{0\le x\le 1}\int_0^1 |g(x,y)|\,dy\|q\|_\infty\|u\|_\infty \\
&= \frac{1}{8}\|q\|_\infty\|u\|_\infty.
\end{aligned}$$

Hence, if $\|q\|_\infty<8$, then T is a contraction in X. Consequently, it has a unique fixed point in X.

3.10 Remarks

More general fixed point theorems based only on topological arguments.

Theorem 3.8 (Intermediate value theorem). *If $T:[a,b]\to[a,b]$ is continuous, then T has a fixed point in $[a,b]$.*

Theorem 3.9 (Brouwer fixed point theorem). *If $T:K\to K$ is continuous and K is a convex and compact in $\mathbb{R}^n$, then T has a fixed point in K.*

Theorem 3.10 (Schauder fixed point theorem). *If $T:K\to K$ is continuous and K is a convex and compact in a Banach space, then T has a fixed point in K.*

I leave you to find applications for these general theorems.

Homeworks 3.1.

1. The Gauss-Seidel method decomposes a matrix $A = M - N$, then solve the equation $Ax = b$ by the iteration $Mx^{n+1} = Nx^n + b$. Here, M is the lower triangular part of A and $-N$ the strict upper triangular part of A. Show the convergence result for diagonally dominant matrix A.

2. Consider the equation
$$f(x) := x - \frac{A}{x}$$
   (a) Show how best to choose a polynomial
$$c(x) = a + bx^2$$
   so that the iteration scheme for $\sqrt{A}$
$$x_{n+1} = x_n + c(x_n)f(x_n)$$
   converges most rapidly in the neighborhood of the solution.

   (b) Estimate the rapidity of convergence.

3. Find an iterative procedure to solve
$$2x + 3y + \sin x^2 + \tan(xy) = 0$$
   for (x, y) near $(0, 0)$.

4. Consider the equation
$$-(p(x)u')' = f, x \in (0, 1)$$
   with $u(0) = u(1) = 0$. The function p is a piecewise constant function. That is
$$p(x) = \begin{cases} a_1 & \text{for } x < \bar{x} \\ a_2 & \text{for } x > \bar{x} \end{cases}$$
   where $0 < \bar{x} < 1$ is a discontinuity of p. At this point, we require
$$[u]_{\bar{x}} := u(\bar{x}+) - u(\bar{x}-) = 0,\ [pu']_{\bar{x}} := p(\bar{x}+)u'(\bar{x}+) - p(\bar{x}-)u'(\bar{x}-) = 0.$$
   Find the Green's function of this system.

5. Hunter's book: pp. 78: Ex 3.1.

6. Hunter's book: pp. 79: Ex 3.6.

Chapter 4

Hilbert Spaces

4.1 Hilbert Spaces, Basic

Hilbert spaces were studied in the first decade of the 20th century by David Hilbert, Erhard Schmidt, and Frigyes Riesz, later by von Neumann on operator theory.

4.1.1 Inner product structure

Definition 4.1. *Let X be a complex linear space. An inner product $(\cdot,\cdot)$ is a bilinear form: $X \times X \to \mathbb{C}$ which satisfies*

(a) $(x,x) \ge 0$ *and* $(x,x) = 0$ *if and only if* $x = 0$,

(b) $(x,y) = \overline{(y,x)}$,

(c) $(x, \alpha y + \beta z) = \alpha(x,y) + \beta(x,z)$.

The linear space X equipped with the inner product $(\cdot,\cdot)$ is called an inner product space.

Examples.

1. The space $\mathbb{C}^n$ with
$$(x,y) := \sum_i \overline{x_i} y_i$$
is an inner product space.

2. Let A be a symmetric positive definite matrix in $\mathbb{R}^n$. Define
$$\langle x,y\rangle_A := (x,Ay)$$
Then $\langle\cdot,\cdot\rangle_A$ is an inner product in $\mathbb{R}^n$.

3. The space $C[0,1]$ with the inner product

$$(f,g) := \int_0^1 \overline{f(t)}g(t)\,dt$$

is an inner product space.

4. The space $L^2(0,1)$ is the completion of $C[0,1]$ with the above inner product. It is the space of all square (Lebesgue) integrable functions.

5. Let $\mathbb{T}$ be the unit circle and

$$L^2(\mathbb{T}) := \{f : \mathbb{T} \to \mathbb{C} \mid \int_{\mathbb{T}} |f(t)|^2\,dt < \infty\}$$

It is the space of all square summable and periodic functions.

6. The $\ell^2(\mathbb{N})$ is defined to be

$$\ell^2(\mathbb{N}) := \{x \mid x = (x_1, x_2, \cdots), \sum_{i=1}^{\infty} |x_i|^2 < \infty\}$$

The $\ell^2(\mathbb{N})$ is an inner product space with the inner product

$$(x,y) := \sum_{i=1}^{\infty} \overline{x_i} y_i.$$

Similarly, we define

$$\ell^2(\mathbb{Z}) := \{x \mid x : \mathbb{Z} \to \mathbb{C}, \sum_{i=-\infty}^{\infty} |x_i|^2 < \infty\}.$$

It is also an inner product space.

7. Let $w_n > 0$ be a positive sequence. Define

$$\ell^2_w := \{x \mid x : \mathbb{N} \to \mathbb{C}, \sum_{i=1}^{\infty} w_i |x_i|^2 < \infty\}$$

The inner product is defined to be

$$(x,y) := \sum_{i=1}^{\infty} w_i x_i \overline{y_i}.$$

8. Let $w : (a, b) \to \mathbb{R}^+$ be a positive continuous function. Define the space

$$L^2_w(a, b) := \{f : (a, b) \to \mathbb{C} \mid \int_a^b |f(x)|^2 w(x)\, dx < \infty\}$$

and equip it with the inner product

$$(f, g) := \int_a^b \overline{f(x)} g(x)\, w(x)\, dx.$$

$L^2_w(a, b)$ is an inner product space.

In an inner product space X, We define $\|x\| = \sqrt{(x, x)}$. Then $(X, \|\cdot\|)$ is a normed space. The key is the Cauchy-Schwarz inequality

Theorem 4.1. *Let X be an inner product space. For any $x, y \in X$, we have*

$$|(x, y)| \le \|x\|\|y\|$$

Proof. From non-negativity of $(\cdot, \cdot)$, we get

$$0 \le (x + ty, x + ty) = \|x\|^2 + 2Re(x, y)t + \|y\|^2t^2 \text{ for all } t \in \mathbb{R}.$$

From this, we obtain

$$|Re(x, y)|^2 \le \|x\|^2\|y\|^2.$$

This is one form of Cauchy-Schwarz. We claim that

$$|Re(x, y)| \le \|x\|\,\|y\| \text{ for any } x, y \in X$$

if and only if

$$|(x, y)| \le \|x\|\,\|y\| \text{ for any } x, y \in X.$$

Suppose (x, y) is not real, we choose a phase ϕ such that $e^{i\phi}(x, y)$ is real. Now we replace x by $e^{i\phi}x$. Then

$$|Re(e^{i\phi}x, y)| \le \|x\|\,\|y\|$$

But the left-hand side is $|(x, y)|$. This proves one direction. The other direction is trivial. □

The triangle inequality is equivalent to the Cauchy-Schwarz inequality. In fact, we have

$$\|x + y\|^2 = \|x\|^2 + 2Re(x, y) + \|y\|^2$$

while

$$(\|x\| + \|y\|)^2 = \|x\|^2 + 2\|x\|\,\|y\| + \|y\|^2.$$

By comparing the two equations, we get that the triangle inequality is equivalent to the Cauchy-Schwarz inequality. In fact, the following statements are equivalent:

(a) For any $x, y \in \mathcal{H}$, $Re(x, y) \le \|x\|\,\|y\|$;

(b) For any $x, y \in \mathcal{H}$, $|(x, y)| \le \|x\|\,\|y\|$;

(c) For any $x, y \in \mathcal{H}$, $\|x + y\| \le \|x\| + \|y\|$.

Remark. If we are care about the cosine law, that is

$$\|x-y\|^2 = \|x\|^2 + \|y\|^2 - 2\|x\|\,\|y\|\cos\theta,$$

then we should define the angle between two vectors x and y by

$$\cos\theta := \frac{Re(x,y)}{\|x\|\,\|y\|}.$$

However, this creates a problem, the orthogonality in this sense may not have $(x,y)=0$. This is not what we want. . So, we define the acuate angle between two vectors x and y by

$$\cos\theta := \frac{|(x,y)|}{\|x\|\,\|y\|},$$

and we give up the traditional cosine law.

Definition 4.2. *A complete inner product space is called a Hilbert space.*

In the aforementioned inner product space, the ℓ^2, $L^2(a,b)$, $L^2_w(\mathbb{R})$ are Hilbert spaces.

Proposition 4.1 (Parallelogram law)**.** *A normed linear space is an inner product space if and only if*

$$\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2, \textit{ for all } x, y \in X.$$

Proof. Suppose a norm satisfies the parallelogram law, we define

$$(x,y) := \frac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2 - i\|x+iy\|^2 + i\|x-iy\|^2\right).$$

We leave the reader to check the parallelogram law implies the bilinearity of the inner product. □

4.1.2 Sobolev spaces

The H^1 space. Let (a,b) be a finite interval on $\mathbb{R}$. We recall that

$$L^2(a,b) = \{u : (a,b) \to \mathbb{C} | \int_a^b |u(x)|^2\,dx < \infty\}.$$

Here, the integrability is in the Lebesgue sense. One may think $L^2(a,b)$ to be the completion of $C[a,b]$ under L^2-norm. Thus, $L^2(a,b)$ is a Hilbert space. Similarly, we define

$$H^1(a,b) = \{u : (a,b) \to \mathbb{C} | \int_a^b \left(|u(x)|^2 + |u'(x)|^2\right)\,dx < \infty\}$$

with the inner product

$$(u,v) = \int_a^b \left(\bar{u}v + \overline{u'}v'\right)\,dx.$$

Think why $H^1(a,b)$ is complete?[1] Indeed, it is the completion of $C^1[a,b]$, or $C^\infty[a,b]$ under the above inner product.

[1] If $u_n \to u$ in $L^2(a,b)$ and $u_n' \to v$ in $L^2(a,b)$, then we have $u' = v$.

The H_0^1 spaces. Let

$$H_0^1(a,b) = \{u \in H^1(a,b) | u(a) = u(b) = 0\}.$$

You may wonder why can we define $u(a)$ and $u(b)$ for those function $u \in H^1(a,b)$? In fact, for any two points x_1 and x_2 near a, we can express

$$|u(x_2) - u(x_1)| = |\int_{x_1}^{x_2} u'(x)\,dx| \le \left(\int_{x_1}^{x_2} 1^2 dx\right)^{1/2} \left(\int_{x_1}^{x_2} |u'(x)|^2 dx\right)^{1/2} \le (x_2 - x_1)^{1/2} \|u'\|,$$

which tends to zero as $x_1, x_2 \to a$. Thus, it is meaningful to take $\lim_{x\to a} u(x)$. Alternatively, $H_0^1(a,b)$ is the completion of $C_0^\infty[a,b]$ under the above inner product. Here, $C_0^\infty[a,b]$ are those C^∞ function on $[a,b]$ satisfying zero boundary condition. I shall take this as a fact.

In $H_0^1(a,b)$, we can define another inner product

$$\langle u, v \rangle := \int_a^b \overline{u'(x)} v'(x)\,dx.$$

To see this, we need to check that $\langle u, u \rangle = 0$ implies $u \equiv 0$. From $\int_a^b |u'(x)|^2\,dx = 0$, we get that $u'(x) \equiv 0$ on (a,b). Hence, u is a linear function on (a,b). With $u(a) = u(b) = 0$, we conclude that $u \equiv 0$.

Now, in H_0^1, we have two norms, one is

$$\|u\|_1^2 \equiv \|u\|^2 + \|u'\|^2;$$

the other is

$$\|u\|_2^2 \equiv \|u'\|^2.$$

Here, $\|\cdot\|$ is the L^2-norm. We claim that these two norms are equivalent in H_0^1. [2] For the aforementioned two norms in H_0^1, we have

$$\|u'\|^2 \le \|u\|^2 + \|u'\|^2.$$

Thus, $\|u\|_2 \le \|u\|_1$. On the other hand, the other inequality $\|u\|_1 \le C_2\|u\|_2$ is followed from the theorem below.

Theorem 4.2 (Poincaré inequality). *There exists a constant $C > 0$ such that for any $u \in H_0^1(a,b)$, we have*

$$\|u\|^2 \le C\|u'\|^2. \tag{4.1}$$

[2] Two norms $\|\cdot\|_1$, $\|\cdot\|_2$ are equivalent in a normed space X if there exist two positive constants C_1, C_2 such that for any $u \in X$, we have

$$C_1\|u\|_2 \le \|u\|_1 \le C_2\|u\|_2.$$

Proof. From the fundamental theorem of calculus,

$$u(x) = u(a) + \int_a^x u'(y)\,dy = \int_a^x u'(y)\,dy.$$

Thus,

$$\begin{aligned} |u(x)|^2 &= \left|\int_a^x u'(y)\,dy\right|^2 \\ &\leq \left(\int_a^x 1^2\,dy\right)\left(\int_a^x |u'(y)|^2\,dy\right) \\ &\leq (x-a)\left(\int_a^b |u'(y)|^2\,dy\right) \end{aligned}$$

Here, we have used Cauchy-Schwarz inequality. We integrate x over (a,b) to get

$$\int_a^b |u(x)|^2\,dx \leq \frac{(b-a)^2}{2}\int_a^b |u'(y)|^2\,dy.$$

This proves the Poincaré inequality in one dimension. □

Remark

1. You can see from the proof that we don't need both boundary conditions $u(a) = u(b) = 0$. In fact, $u(x)$ is determined by $u(a)$ and $u'(x), a < x < b$. Thus, the Poincaré inequality is also valid by just assuming $u(a) = 0$. Indeed, it is also valid by assuming

$$\int_a^b u(x)\,dx = 0.$$

2. Dimension analysis for the Poincaré inequality: Let us denote the dimension of length by L, the variable u by U. We denote these by $[x] = L$ and $[u] = U$. The L^2 norm of u has dimension

$$[\|u\|] = (U^2L)^{1/2} = UL^{1/2}.$$

The dimension of L^2 norm of u' is

$$\left[\|u'\|\right] = (UL^{-1})L^{1/2} = UL^{-1/2}$$

By comparing the dimensions on both sides of the Poincaré inequality, we get that the constant has the dimension L.

Best constant in Poincaré inequality To find the best constant C in the Poincaré inequality, we look for the following minimum

$$\min_{u(a)=u(b)=0} \frac{\int_a^b u'(x)^2\,dx}{\int_a^b u(x)^2\,dx}$$

This problem is equivalent to

$$\min_{u(a)=u(b)=0} \int_a^b u'(x)^2\,dx \text{ subject to } \int_a^b u(x)^2\,dx = 1.$$

By the method of Lagrange multiplier, there exists a λ such that

$$\delta\left(\int_a^b u'(x)^2\,dx - \lambda \int_a^b u(x)^2\,dx\right) = 0.$$

The corresponding Euler-Lagrange equation is

$$-u'' - \lambda u = 0$$

with the two boundary condition $u(a) = u(b) = 0$. This is a standard eigenvalue problem. The minimal value of λ is the first eigenvalue of $-D^2$ with the Dirichlet boundary condition. The corresponding eigenvector and eigenvalue are

$$u(x) = \sin\left(\frac{x-a}{b-a}\pi\right), \ \lambda = \left(\frac{\pi}{b-a}\right)^2.$$

Thus, the best constant is

$$C = \frac{1}{\sqrt{\lambda}} = \frac{b-a}{\pi}.$$

Exercise: The weighted Sobolev space. Let $w(x) > 0$ on $[a, b]$. Define the inner product

$$\langle u, v\rangle_w := \int_a^b u'(x)\overline{v'(x)}w(x)\,dx$$

and the corresponding norm $\|u'\|_w^2 := \langle u, u\rangle_w$. Let

$$H^1_{w,0}(a, b) := \{u : (a, b) \to \mathbb{C} | \|u'\|_w < \infty, u(a) = u(b) = 0\}$$

Then the space $H^1_{0,w}(a, b) = H^1_0$ and the norm $\|u'\|_w$ is equivalent to $\|u'\|$.

Homeworks 4.1. *pp. 144-145: 6.1, 6.3, 6.4, 6.5*

4.2 Projection

Projections in Banach spaces

Definition 4.3. *(a) A projection P in a Banach space X is a linear mapping from X to X satisfying $P^2 = P$.*

(b) The direct sum of two subspaces $\mathcal{M}$ and $\mathcal{N}$ in a Banach space X is defined to be

$$\mathcal{M} \oplus \mathcal{N} := \{x + y \mid x \in \mathcal{M},\ y \in \mathcal{N}\}.$$

Theorem 4.3. *If P is a projection on linear space X, then $X = Ran\, P \oplus Ker\, P$, and $RanP \cap KerP = \{0\}$.*
Conversely, if $X = \mathcal{M} \oplus \mathcal{N}$ and $\mathcal{M} \cap \mathcal{N} = \{0\}$, then any $x \in X$ can be uniquely represented as $x = y + z$ with $y \in \mathcal{M}$ and $z \in \mathcal{N}$. Furthermore, the mapping $P : x \mapsto y$ is a projection.

Proof. $(\Rightarrow)$

1. We first show that $x \in Ran\, P \Leftrightarrow x = Px$. $(\Leftarrow)$ If $x = Px$ clearly $x \in RanP$. $(\Rightarrow)$If $x \in RanP$, then $x = Py$ for some $y \in \mathcal{H}$. From $P^2 = P$, we get $Px = P^2 y = Py = x$.
2. Next, if $x \in RanP \cap KerP$, then $x = Px = 0$. Hence, $RanP \cap KerP = \{0\}$.
3. Finally, we can decompose $x \in X$ into
$$x = Px + (x - Px).$$
The part $Px \in RanP$. The other part $x - Px \in Ker\, P$ because $P(x - Px) = Px - P^2 x = 0$.

$(\Leftarrow)$

1. If $x = y_1 + z_1 = y_2 + z_2$ with $y_i \in \mathcal{M}$ and $z_i \in \mathcal{N}$, then $y_1 - y_2 = z_2 - z_1$ and it is in $\mathcal{M} \cap \mathcal{N}$. Thus, $y_1 = y_2$ and $z_1 = z_2$.
2. For $y \in \mathcal{M}$, $Py = y$. For any x, $Px \in \mathcal{M}$, hence $P(Px) = Px$.

□

Remarks.

1. If P is a projection, so is $I - P$.
2. We have $RanP = Ker(I - P)$ and $KerP = Ran(I - P)$.
3. A projection in a Banach space needs not be continuous in general.

Theorem 4.4. *Let X be a Banach space and P is a projection in X.*

(a) If P is continuous, then both $KerP$ and $RanP$ are closed.

(b) *On the other hand, if Y is closed subspace and there exists a closed subspace Z such that $X = Y \oplus Z$. Then the projection $P : x \mapsto y$ is continuous, where $x = y + z$ is the decomposition of x with $y \in Y$ and $z \in Z$.*

Proof. 1. We show the graph of P is closed. That is, if $x_n \to x$ and $y_n := Px_n \to y$, then $y \in Y$ and $x - y \in Z$. From the decomposition, we have $y_n \in Y$ and $z_n := x_n - y_n \in Z$. From the closeness of Y, we get $y \in Y$. From $x_n \to x$ and $y_n \to y$, we get $x_n - y_n$ converges to $x - y$. From the closeness of Z, we get $x - y \in Z$. Thus, $x = y + (x - y)$ with $y \in Y$ and $x - y \in Z$.

2. The theorem follows from the closed graph theorem: A closed graph linear map A from Banach space X to Banach space Y is also continuous.

□

Orthogonal projections in Hilbert spaces

Theorem 4.5 (Orthogonal Projection Theorem)**.** *Let $\mathcal{H}$ be a Hilbert space and let $\mathcal{M} \subset \mathcal{H}$ be a* closed *linear subspace of $\mathcal{H}$. Then*

(a) *for any $x \in \mathcal{H}$, there exists a unique $y \in \mathcal{M}$ such that*

$$\|x - y\| = \min_{z \in \mathcal{M}} \|x - z\|;$$

(b) *$(x - y) \perp \mathcal{M}$;*

(c) *the mapping $P : x \mapsto y$ is a projection.*

Proof. (a) Let us denote

$$\ell = \inf_{z \in \mathcal{M}} \|x - z\|^2.$$

Let $\{y_n\}$ be a minimal sequence of $\|\cdot\|^2$ in $\mathcal{M}$. That is $y_n \in \mathcal{M}$ and $\lim_{n\to\infty} \|y_n - x\|^2 = \ell$. Then from the parallelogram law

$$\|y_m - y_n\|^2 = 2\|y_m - x\|^2 + 2\|y_n - x\|^2 - 4\|\frac{y_m + y_n}{2} - x\|^2.$$

The first two terms tend to 4ℓ as $n, m \to \infty$, while the last term is greater than 4ℓ by the definition of ℓ. This implies $\{y_n\}$ is a Cauchy sequence in $\mathcal{M}$ hence it has a limit y in $\mathcal{M}$.

The uniqueness follows from the parallelogram law as follows. Suppose y_1 and y_2 are two such solutions, that is $\|y_i - x\|^2 = \ell$. Then by the parallelogram law,

$$\|y_1 - y_2\|^2 = 2\|y_1 - x\|^2 + 2\|y_2 - x\|^2 - 4\|\frac{y_1 + y_2}{2} - x\|^2 \le 4\ell - 4\ell = 0.$$

(b) From
$$\|x-y\|^2 \le \|x-y-tz\|^2 = \|x-y\|^2 - 2Re(x-y,tz) + |t|^2\|z\|^2$$
for all $t \in \mathbb{C}$ and $z \in \mathcal{M}$, we choose $t = \epsilon e^{i\phi}$ so that
$$Re(x-y,tz) = |t||(x-y,z)|.$$
Then we get
$$\epsilon|(x-y,z)| \le \epsilon^2|z|^2.$$
Taking $\epsilon \to 0+$, we get $(x-y,z) = 0$.

(c) Let $\mathcal{N} = \{z \in \mathcal{H} | z \perp \mathcal{M}\}$. From (b), we get that $\mathcal{H} = \mathcal{M} \oplus \mathcal{N}$ and $\mathcal{M} \cap \mathcal{N} = \{0\}$. Thus, the mapping $P : x \mapsto y$ is a projection onto $\mathcal{M}$.

□

Let $\mathcal{H}$ be a Hilbert space. $\mathcal{M} \subset \mathcal{H}$ be a subset. Define the orthogonal complement of $\mathcal{M}$ by
$$\mathcal{M}^\perp := \{x \in \mathcal{H} | x \perp y \text{ for all } y \in \mathcal{M}\}.$$
The orthogonal complement of a subset $\mathcal{M}$ in $\mathcal{H}$ is a closed linear subspace. From the orthogonal projection theorem, we have
$$\mathcal{H} = \mathcal{M} \oplus \mathcal{M}^\perp.$$
Any $x \in \mathcal{H}$ can be decomposed into $x = y + z$ with $y \in \mathcal{M}$ and $x - y \in \mathcal{M}^\perp$. The projection $P : x \mapsto y$ is called an orthogonal projection.

Corollary 4.1. *If $\mathcal{M}$ is a closed subspace of a Hilbert space $\mathcal{H}$, then $\left(\mathcal{M}^\perp\right)^\perp = \mathcal{M}$.*

Proof. We only prove $\left(\mathcal{M}^\perp\right)^\perp \subset \mathcal{M}$. Suppose $x \in \left(\mathcal{M}^\perp\right)^\perp$. That is, $(x,w) = 0$ for all $w \in \mathcal{M}^\perp$. By the orthogonal projection theorem, we can decompose $x = y + z$ with $y \in \mathcal{M}$ and $z \in M^\perp$. Then $0 = (x,w) = (y+z,w) = (z,w)$ for all $w \in \mathcal{M}^\perp$. Since $z \in \mathcal{M}^\perp$, we can take $w = z$ and get $(z,z) = 0$. Hence, $x = y \in \mathcal{M}$. This proves $\left(\mathcal{M}^\perp\right)^\perp \subset \mathcal{M}$. □

Theorem 4.6. *Let $P : \mathcal{H} \to \mathcal{H}$ be a projection. The following two statements are equivalent:*

(a) $(Px_1, x_2) = (x_1, Px_2)$ for all $x_1, x_2 \in \mathcal{H}$;

(b) $\mathcal{H} = Ran\,P \oplus Ker\,P$ and $Ran\,P \perp Ker\,P$.

Proof. (a) ⇒ (b): For any $x \in RanP$, then $x = Py$ for some $y \in \mathcal{H}$. Then for any $z \in Ker\,P$,
$$(x,z) = (Py,z) = (y,Pz) = 0.$$
Hence, $RanP \perp Ker\,P$.

(b) ⇒ (a): For any $x_1, x_2 \in \mathcal{H}$, they can be uniquely decomposed into
$$x_1 = y_1 + z_1,\ x_2 = y_2 + z_2,\ \text{with } y_i \in \mathcal{M}, z_i\mathcal{M}^\perp.$$
Thus,
$$(Px_1, x_2) = (y_1, y_2) = (x_1, Px_2).$$

□

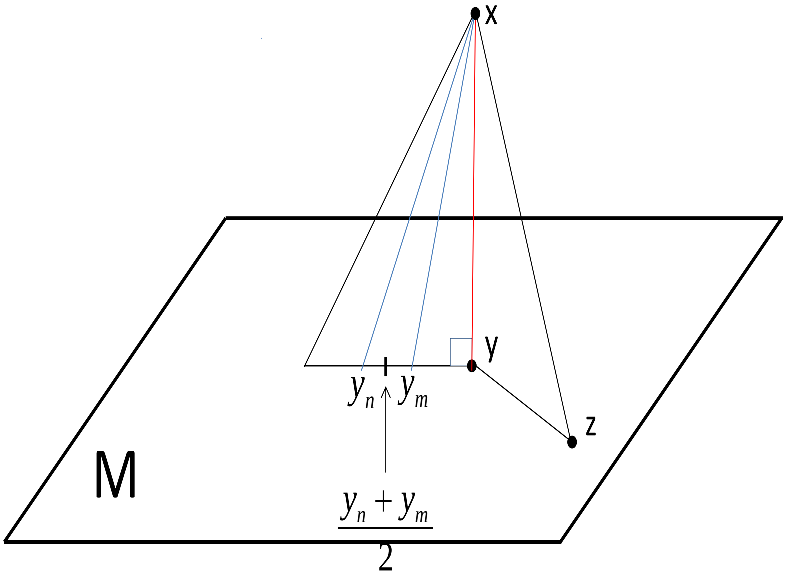

Figure 4.1: Orthogonal projection of x onto a closed subspace $\mathcal{M}$. $\{y_n\}$ are minimal sequence.

Example 1. Given a vector $y \in \mathcal{H}$. Define $P : x \mapsto (y, x)\frac{y}{\|y\|^2}$. Then $RanP = \langle\{y\}\rangle$, the space spanned by y, and $KerP = y^{\perp}$.

Example 2. Given n independent vectors $\{v_1, \cdots, v_n\}$ in $\mathcal{H}$. Let $\mathcal{M} = \langle\{v_1, \cdots, v_n\}\rangle$. Given any $x \in \mathcal{H}$, the orthogonal projection y of x on $\mathcal{M}$ satisfies:

$$y = \text{arg min}\{\frac{1}{2}\|x - z\|^2 \mid z \in \mathcal{M}\}.$$

The corresponding Euler-Lagrange equation is $(x - y) \perp \mathcal{M}$. Since $y \in \mathcal{M}$, we can express y as $y = \sum_{i=1}^{n} \alpha_i v_i$. The condition $(x - y) \perp \mathcal{M}$ is equivalent to $(x - y, v_i) = 0$, $i = 1, \cdots, n$. This leads to the following $n \times n$ system of linear equations

$$\sum_{j=1}^{n} (v_i, v_j)\alpha_j = (x, v_i), i = 1, \cdots, n.$$

From the independence of $\{v_1, \cdots, v_n\}$, we can get a unique solution of this equation.

4.3 Riesz Representation Theorem

Dual space The set

$$\mathcal{H}^* := \{\ell : \mathcal{H} \to \mathbb{C} \text{ bounded linear functional}\}$$

forms a linear space and is called the dual space of $\mathcal{H}$. For an element $\ell \in \mathcal{H}$, we define its norm by

$$\|\ell\| := \sup \frac{\|\ell(x)\|}{\|x\|} = \sup_{\|x\|=1} |\ell(x)|.$$

Then $\mathcal{H}^*$ is a normed linear space. I left you to prove that $\mathcal{H}^*$ is complete.

A typical example of bounded linear functional is the follows. Given a $y \in \mathcal{H}$, the mapping

$$\ell_y(x) := (y, x)$$

is a bounded linear functional, by Cauchy-Schwarz inequality. Its norm $\|\ell_y\| \le \|y\|$. On the other hand, by choosing $x = y/\|y\|$, we obtain

$$\|\ell_y\| \ge |\ell_y(y/\|y\|)| = \|y\|.$$

We thus conclude $\|\ell_y\| = \|y\|$.

An important theorem is the Riesz representation theorem which states that every bounded linear functional on $\mathcal{H}$ must be in this form. In other word, $\mathcal{H}^*$ is isometric to $\mathcal{H}$.

Theorem 4.7 (Riesz representation theorem)**.** *Let ℓ be a bounded linear functional on a Hilbert space $\mathcal{H}$. Then there exists a unique $y \in \mathcal{H}$ such that*

$$\ell(x) = (y, x).$$

Proof. We suppose $\ell \neq 0$. Our goal is to find y such that $\ell(x) = (y, x)$. We first notice that such y must be in $(Ker\ \ell)^\perp$ and $P : x \mapsto (y, x)y/\|y\|^2$ is an orthogonal projection.

Let $\mathcal{N} = Ker\ \ell$. Then $\mathcal{N}$ is closed and $\mathcal{N} \neq \mathcal{H}$. Hence there exists a $z_1 \notin \mathcal{N}$. By the orthogonal projection theorem, there exists a $y_1 \in \mathcal{N}$ and $z := (z_1 - y_1) \perp \mathcal{N}$. From $z_1 \notin \mathcal{N}$, we get $z \neq 0$. Let

$$Px := \frac{\ell(x)}{\ell(z)} z.$$

Then P is an orthogonal projection, i.e. $P^2 = P$ and $RanP \perp Ker\ P$, because $RanP = \{\alpha z | \alpha \in \mathbb{C}\}$ and $Ker\ P = Ker\ \ell = \mathcal{N}$. The latter is due to $Px = 0$ if and only if $\ell(x) = 0$. With these, we get that

$$\mathcal{H} = \{\alpha z | \alpha \in \mathbb{C}\} \oplus Ker\ \ell.$$

Hence, any $x \in \mathcal{H}$ can be represented uniquely by

$$x = \alpha z + m, \text{ with } m \in Ker\ell,\ \alpha = (z, x)/\|z\|^2.$$

We have

$$\ell(x) = \ell(\alpha z) = \frac{(z, x)}{\|z\|^2}\ell(z) = (y, x).$$

where, $y := \frac{\ell(z)}{\|z\|^2} z$. We have shown the existence of y such that $\ell(x) = (y, x)$.

For the uniqueness, suppose there are y_1 and y_2 such that $\ell_{y_1} = \ell_{y_2}$. That is,

$$(y_1, x) = (y_2, x), \text{ for all } x \in \mathcal{H}.$$

Choose $x = y_1 - y_2$, we obtain $\|y_1 - y_2\| = 0$. □

Application of Riesz representation theorem. Now, we consider the Poisson equation on a bounded domain $\Omega \subset \mathbb{R}^n$:

$$(P) : \triangle u = f \text{ in } \Omega, u = 0 \text{ on } \partial\Omega.$$

This problem can be reformulated as the following weak form:

$$(WP): \text{ Find } u \in H_0^1(\Omega) \text{ such that } (\nabla u, \nabla v) = -(f, v), \text{ for all } v \in C_0^1.$$

Lemma 4.1 (Poincaré inequality)**.** *Let $\Omega \subset \mathbb{R}^n$ be a smooth bounded domain. Then there exists a constant C such that for $u \in H_0^1(\Omega)$, we have*

$$\|u\|_2 \le C\|\nabla u\|_2.$$

Existence of Dirichlet problem.

Theorem 4.8. *Let $\Omega \subset \mathbb{R}^n$ be a smooth bounded domain. Let $f \in L^2(\Omega)$. Then (WP) has a unique solution in $H_0^1(\Omega)$.*

Proof. From the Poincaré's inequality, we see that

$$\langle u, v\rangle_1 := (\nabla u, \nabla v) := \int_\Omega \overline{\nabla u(x)} \cdot \nabla v(x)\, dx$$

defines an inner product in $H_0^1(\Omega)$. On the other hand, for $f \in L^(\Omega)$,

$\ell_v := (f, v)$ is a bounded linear map in both L^2 and H_0^1:

$$|(f, v)| \le \|f\|\|v\| \le C\|f\|\,\|\nabla v\|$$

Thus, by the Riesz representation theorem, there exists a unique $u \in H_0^1(\Omega)$ such that

$$\langle u, v\rangle_1 = (\nabla u, \nabla v) = (-f, v)$$

for all $v \in H_0^1(\Omega)$. □

Homeworks 4.2. 1. *pp. 212, Ex. 8.3,*

2. *pp. 212, Ex. 8.5*

3. *Assuming $u \in C^2(\overline{\Omega})$. Then u satisfies (P) if and only if it satisfies (WP).*

4. *Prove the Poincaré's inequality for the case when $u \in C_0^1(\Omega)$, $\Omega \subset \mathbb{R}^d$ a bounded domain.*

4.4 Error estimates for finite element method

Let us consider the Poisson equation in one dimension:

$$-u'' = f \text{ on } (a,b),\ u(a) = u(b) = 0. \tag{4.2}$$

We shall find an approximate solution by finite element method. First, we discretize the space $[a,b]$ and define the finite element functions. We chosse an $n > 0$. Let $h := (b-a)/n$ the mesh size, $x_i = a+ih, i = 0, \cdots, n$ the grid point. Define the finite element function $\phi_i(x)$ to be $\phi_i(x_j) = \delta_{ij}$ and $\phi(x)$ is continuous and piecewise linear. Let

$$V_h = \langle \phi_1, \cdots, \phi_{n-1} \rangle.$$

It is called the finite element space. An element $v \in V_h$ is a continuous and piecewise linear function and is uniquely expressed by

$$v(x) = \sum_{i=1}^{n-1} v(x_i)\phi_i(x).$$

Next, we find an approximate solution $u_h \in V_h$. We express u_h by

$$u_h(x) = \sum_{i=1}^{n-1} U_i\phi_i(x)$$

We project the equation (4.2) onto V_h:

$$(-u'' - f, v) = 0, \text{ for all } v \in V_h$$

This leads to the following equations for $U = (U_1, \cdots, U_{n-1})^T$:

$$\langle u_h, \phi_i \rangle_1 = (f, \phi_i), i = 1, \cdots, n-1.$$

Or

$$\sum_{j=1}^{n-1} (\phi_i', \phi_j')U_j = (f, \phi_i), i = 1, \cdots, n-1.$$

We can compute (ϕ_i, ϕ_j) directly and obtain the matrix $A = (\phi_i', \phi_j')_{(n-1)\times(n-1)}$ as

$$A = \frac{1}{h}\text{diag}(-1, 2, -1)$$

This matrix is invertible. So, we can invert it and find U.

Finally, we study the error of the approximate solution u_h. Let u be the exact solution and $e_h := u - u_h$ be the true error. Since both u and u_h satisfy

$$(u', v') = (f, v),\ (u_h', v') = (f, v) \text{ for all } v \in V_h,$$

we obtain

$$(e_h', v') = 0 \text{ for all } v \in V_h.$$

That is, $(u - u_h) \perp_1 V_h$. This is equivalent to say that u_h is the $\langle \cdot, \cdot \rangle_1$-orthogonal projection of u on V_h. Thus,

$$\|u' - u_h'\|_2 \le \|u' - v'\|_2 \text{ for all } v \in V_h.$$

In particular, we can choose $v \in V_h$ that equals u at $x_1, \cdots, x_{n-1}$. That is,

$$v = \pi_h u := \sum_{i=1}^{n-1} u(x_i)\phi_i,$$

then

$$\|u' - u_h'\|_2 \le \|u' - (\pi_h u)'\|_2. \tag{4.3}$$

Thus, the true error is controlled by the approximation error.

Approximation error in terms of $\|u''\|_\infty$ It is easy to see that π_h is a projection. If $u \in C^2$, then in each cell (x_i, x_{i+1}), the projection error $w(x) = u(x) - \pi_h u(x)$ satisfies $w(x_i) = w(x_{i+1}) = 0$. By applying Rolle's theorem twice, we get that for any $x \in (x_i, x_{i+1})$, there exists an $\xi_i \in (x_i, x_{i+1})$ such that

$$w(x) = \frac{w''(\xi_i)}{2}(x - x_i)(x - x_{i+1}).$$

This leads to

$$|w(x)| \le \frac{h^2}{8} \max_{\xi \in (x_i, x_{i+1})} |w''(\xi)|.$$

Hence

$$\begin{aligned}
\int_a^b |w(x)|^2\,dx &= \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} |w(x)|^2\,dx \\
&\le \sum_{i=0}^{n-1} h \left(\frac{h^2}{8}\right)^2 \max_{x \in (x_i, x_{i+1})} |w''(x)|^2 \\
&\le (b-a) \left(\frac{h^2}{8}\right)^2 \max_{x \in [a,b]} |u''(x)|^2.
\end{aligned}$$

Here, we have used that $w''(x) = u''(x)$ on each subinterval (x_i, x_{i+1}). Hence,

$$\|u - \pi_h u\|_2 \le \sqrt{b-a}\frac{h^2}{8}\|u''\|_\infty$$

We can also estimate $u' - (\pi_h u)'$ by mean value theorem. First, there exists a $\zeta_1 \in (x_i, x_{i+1})$ such that $u'(\zeta_1) = (u(x_{i+1} - u(x_i))/h$. For any $x \in (x_i, x_{i+1})$, there exists $\zeta_2 \in (x_i, x_{i+1})$ such that $u'(x) - u'(\zeta_1) = u''(\zeta_2)(x - \zeta_1)$. Therefore, we get

$$u'(x) - (\pi_h u)'(x) = u'(x) - \frac{u(x_{i+1}) - u(x_i)}{h} = u''(\zeta_2)(x - \zeta_1).$$

Notice that $(\pi_h u)'(x) = \frac{u(x_{i+1}) - u(x_i)}{h}$ for $x \in (x_i, x_{i+1})$. Hence, we obtain

$$\begin{aligned} \int_a^b |u' - (\pi_h u)'|^2 \, dx &= \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} |u' - (\pi_h u)'|^2 \, dx \\ &\le \sum_{i=0}^{n-1} h \, h^2 \max_{x \in [a,b]} |u''(x)|^2 \\ &= (b-a) h^2 \|u''\|_\infty^2 \end{aligned}$$

Approximation error in terms of $\|u''\|_2$ The estimate above is in terms of $\|u''\|_\infty$. It is desirable to estimate in terms of $\|u''\|_2$. That is, we want to estimate $\|u - \pi_h u\|_2$ in terms of $\|u''\|_2$. To do so, we should use the integral representation of error of the Lagrange interpolation. We recall that for $w(x_i) = w(x_{i+1}) = 0$, w has the representation:

$$w(x) = h^2 \int_{x_i}^{x_{i+1}} g\left(\frac{x - x_i}{h}, \frac{y - x_i}{h}\right) w''(y) \, dy$$

$$w'(x) = h \int_{x_i}^{x_{i+1}} g_x\left(\frac{x - x_i}{h}, \frac{y - x_i}{h}\right) w''(y) \, dy$$

where g is the Green's function of d^2/dx^2 on (x_i, x_{i+1}). Thus, we can estimate $\|w\|_2$ in terms of $\|w''\|_2$ on (x_i, x_{i+1}). Namely,

$$|w(x)|^2 \le h^4 \left(\int_{x_i}^{x_{i+1}} |g\left(\frac{x - x_i}{h}, \frac{y - x_i}{h}\right)|^2 \, dy \right) \left(\int_{x_i}^{x_{i+1}} |w''(y)|^2 \, dy \right).$$

$$\begin{aligned} \int_{x_i}^{x_{i+1}} |w(x)|^2 \, dx &\le \int_{x_i}^{x_{i+1}} \int_{x_i}^{x_{i+1}} |g\left(\frac{x - x_i}{h}, \frac{y - x_i}{h}\right)|^2 \, dy \, dx \int_{x_i}^{x_{i+1}} |w''(y)|^2 \, dy \\ &\le \frac{1}{90} h^4 \int_{x_i}^{x_{i+1}} |w''(y)|^2 \, dy. \end{aligned}$$

As we sum over $i = 1, \cdots, n-1$, we get

$$\|w\|_2 \le \frac{1}{\sqrt{90}} h^2 \|w''\|_2.$$

Similarly, we get

$$\|w'\|_2 \le \frac{1}{\sqrt{6}} h \|w''\|_2.$$

Theorem 4.9. *For $u \in H^2(a,b) \cap H_0^1[a,b]$, the interpolation error has the following estimates*

$$\|u - \pi_h u\|_2 \le \frac{1}{\sqrt{90}} h^2 \|u''\|_2,$$

$$\|u' - (\pi_h u)'\|_2 \le \frac{1}{\sqrt{6}} h \|u''\|_2.$$

True error of the finite element method

Theorem 4.10. *For the finite element method for problem (4.2), the true error $u - u_h$ has the following estimate*

$$\|u' - u_h'\|_2 \le \|u' - (\pi_h u)'\|_2 \le \frac{1}{\sqrt{6}} h \|u''\|_2,$$

$$\|u - u_h\|_2 \le \frac{1}{6} h^2 \|u''\|_2.$$

Proof. The first estimate follows from the previous theorem. For the second, the trick is called duality argument. Let $e_h = u - u_h$. We find the function ϕ_h such that $\phi_h'' = -e_h$ and $\phi(a) = \phi(b) = 0$. Then

$$(e_h, e_h) = -(e_h, \phi_h'') = (e_h', \phi_h') = (e_h', \phi_h' - (\pi_h \phi_h)').$$

Here, I have used

$$(e_h', v') = 0 \text{ for all } v \in Ran(\pi_h).$$

Applying interpolation estimate to ϕ_h, we get

$$\|e_h\|^2 \le \|e_h'\| \|(\phi_h - \pi_h \phi_h)'\| \le \frac{1}{\sqrt{6}} \|e_h'\| \, h \, \|\phi_h''\| = \frac{1}{\sqrt{6}} h \|e_h'\| \, \|e_h\|$$

Hence, we get

$$\|e_h\|_2 \le \frac{1}{\sqrt{6}} h \|e_h'\|_2 \le \frac{1}{6} h^2 \|u''\|_2.$$

□

Homeworks 4.3. *1. The error function w on each interval (x_i, x_{i+1}) satisfies $w(x_i) = w(x_{i+1}) = 0$. w can be estimated in terms of w'' in (x_i, x_{i+1}). This is indeed a generalized Poincaré inequality. You can get best estimate via Fourier sin expansion. Find the best constant and the get the best error estimate.*

Given $x_0 < x_1 < x_2$. Let w be a smooth function satisfying $w(x_i) = 0$ for $i = 0, 1, 2$. Find an integral representation of w in terms of w''' on (x_0, x_2).

Chapter 5

Bases in Hilbert Spaces

5.1 Orthogonal bases, general theory

In this section, we shall discuss how to approximate a point $x \in \mathcal{H}$ in terms of an expansion in an orthogonal set $U = \{u_\alpha | \alpha \in I\}$.

Definition 5.1. *1. A set $U = \{u_\alpha | \alpha \in I\}$ is called an orthogonal set in a Hilbert space $\mathcal{H}$ if any two of them are orthogonal to each other.*

2. It is called an orthonormal set if it is orthogonal and each of them is a unit vector.

For separable Hilbert space (i.e. there exists a countable set $\mathcal{A}$ such that $\overline{\mathcal{A}} = \mathcal{H}$), we can choose U to be countable. But in general, U can be uncountable. The index set I may not be ordered, or may even not be countable. Nevertheless, we can still discuss the meaning of the limit of (uncountable) summation. Consider a set $\{x_\alpha | \alpha \in I\}$ in a Hilbert space $\mathcal{H}$. For a finite set $J \subset I$, let us denote $\sum_{\alpha \in J} x_\alpha$ by S_J.

Definition 5.2. *1. We say that $\sum_{\alpha \in I} x_\alpha$ converges to x unconditionally, if for any $\epsilon > 0$, there exists a finite set $K \subset I$ such that for any finite set J with $K \subset J \subset I$, we have*

$$\|x - \sum_{\alpha \in J} x_\alpha\| < \epsilon.$$

2. The summation $\sum\{x_\alpha | \alpha \in I\}$ is called Cauchy if for any $\epsilon > 0$, there exists a finite set $K \subset I$ such that for any finite set J with $K \subset J \subset I$, we have

$$\|\sum_{\alpha \in J \setminus K} x_\alpha\| < \epsilon.$$

3. It is called absolute Cauchy if for any $\epsilon > 0$, there exists a finite set $K \subset I$ such that for any finite set J with $K \subset J \subset I$, we have

$$\sum_{\alpha \in J \setminus K} \|x_\alpha\| < \epsilon.$$

Remark

1. It is clear that if $\sum_{\alpha\in I}$ converges, then it is Cauchy.

2. If $\sum_{\alpha\in I} x_\alpha$ is Cauchy, then for any $n \in \mathbb{N}$, there exists a finite K_n such that for any $\alpha \notin K_n$, $\|x_\alpha\| < 1/n$. Then for any $\alpha \notin \bigcup_{n\in\mathbb{N}} K_n$, $\|x\| < 1/n$ for all n. Thus, $x_\alpha = 0$. Since $\bigcup_{n\in\mathbb{N}} K_n$ is countable, we conclude that there are at most countable nonzero x_α. In this case, we can select $J_n = \bigcup_{j\le n} K_j$, then S_{J_n} converges to $\sum_{\alpha\in I} x_\alpha$.

3. It is clearly that absolute Cauchy implies Cauchy.

Example For non-separable Hilbert space, we consider quasi-periodic functions on $\mathbb{R}$. A function is called quasi-periodic if

$$f(t) = \sum_{k=1}^{n} a_k e^{i\omega_k t}$$

where $n \in \mathbb{N}$, $a_k \in \mathbb{C}$ and $\omega_k \in \mathbb{R}$ are arbitrary. For quasi-periodic functions f and g, we define

$$(f, g) := \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \overline{f(t)} g(t)\, dt.$$

One can show that this inner product is equivalent to

$$(f, g) = \lim_{T\to\infty} \frac{1}{T} \int_{t_0}^{t_0+T} \overline{f(t)} g(t)\, dt,$$

which is independent of t_0 for quasi-periodic functions. Let $\mathcal{H}$ be the completion of all quasi-periodic functions with the above inner product. They are called the L^2-almost periodic functions. The set

$$U = \{e^{i\omega t} | \omega \in \mathbb{R}\}$$

is a uncountable orthonormal set in $\mathcal{H}$.

Lemma 5.1. *Let $\{x_\alpha | \alpha \in I\}$ be an orthogonal set. Then $\sum_{\alpha\in I} x_\alpha$ converges if and only if $\sum_{\alpha\in I} \|x_\alpha\|^2$ converges. In this case,*

$$\left\| \sum_{\alpha\in I} x_\alpha \right\|^2 = \sum_{\alpha\in I} \|x_\alpha\|^2 \tag{5.1}$$

Proof. We check that (a) $\sum_{\alpha\in I} x_\alpha$ is Cauchy if and only if (b) $\sum_{\alpha\in I} \|x_\alpha\|^2$ is Cauchy. This is because if (a) is true, which means that for any $\epsilon > 0$, there exists a finite set $K \subset I$ such that for any $J \subset I \setminus K$, we have $\| \sum_{i\in J} x_\alpha \|^2 < \epsilon$. But

$$\left\| \sum_{\alpha\in J} x_\alpha \right\|^2 = \sum_{\alpha\in J} \|x_\alpha\|^2.$$

This is equivalent to say that $\sum_{\alpha\in I} \|x_\alpha\|^2$ is Cauchy. In this case, (5.1) follows from the continuity of the norm. □

Theorem 5.1 (Bessel's inequality). *Let $U := \{u_\alpha | \alpha \in I\}$ be an orthonormal set in a Hilbert space $\mathcal{H}$. Then for any $x \in \mathcal{H}$, we have*

(a) $\sum_{\alpha \in I} |(u_\alpha, x)|^2 \le \|x\|^2$*;*

(b) $\bar{x} = \sum_{\alpha \in I} (u_\alpha, x) u_\alpha$ *converges unconditionally;*

(c) $(x - \bar{x}) \perp U$*.*

(d) Let $\langle U \rangle$ denote the finite linear span of the set U and $\overline{\langle U \rangle}$ the closure of $\langle U \rangle$. Then $x \in \overline{\langle U \rangle}$ if and only if $x = \sum_{\alpha \in I} (u_\alpha, x) u_\alpha$.

(e) The subspace $\overline{\langle U \rangle}$ has the characterization

$$\overline{\langle U \rangle} = \{\sum_{\alpha \in I} a_\alpha u_\alpha \mid \sum_{\alpha \in I} |a_\alpha|^2 < \infty\}$$

Proof. (a) Let us denote $(u_\alpha, x) u_\alpha$ by x_α. Then $\{x_\alpha | \alpha \in I\}$ is an orthogonal set. From

$$0 \le \|x - \sum_{\alpha \in J} (u_\alpha, x) u_\alpha\|^2 = \|x\|^2 - \sum_{\alpha \in J} |(u_\alpha, x)|^2,$$

where J is any finite subset of I, we have $\sum_{\alpha \in J} \|x_\alpha\|^2 \le \|x\|^2$ for all finite set $J \subset I$. Let

$$M = \sup\{\sum_{\alpha \in J} \|x_\alpha\|^2 \; | J \subset I \text{ is a finite set.}\}.$$

Then for any $\epsilon > 0$, there exists a finite set $K \subset I$ such that $M - \epsilon < \sum_{\alpha \in K} \|x_\alpha\|^2$. Now for any finite set $J \subset I \setminus K$, we have

$$M - \epsilon < \sum_{\alpha \in K} \|x_\alpha\|^2 \le \sum_{\alpha \in K \cup J} \|x_\alpha\|^2 = \left(\sum_{\alpha \in K} + \sum_{\alpha \in J} \|x_\alpha\|^2 \le M \right)$$

This gives

$$\sum_{\alpha \in J} \|x_\alpha\|^2 < \epsilon.$$

Thus, $\sum_{\alpha \in I} \|x_\alpha\|^2$ is Cauchy. It converges and has an upper bound $\|x\|^2$.

(b) From (a) and Lemma 5.1, we get the convergence of the un-order sum $\bar{x} := \sum_{\alpha \in I} (u_\alpha, x) u_\alpha$.

(c) We use continuity of inner product: for any $u_\beta \in U$,

$$\left\langle x - \sum_{\alpha \in I} (u_\alpha, x) u_\alpha, u_\beta \right\rangle = (x, u_\beta) - \sum_{\alpha \in I} \overline{(u_\alpha, x)} (u_\alpha, u_\beta) = (x, u_\beta) - \overline{(u_\beta, x)} = 0.$$

(d) From(b), $\bar{x} \in \overline{\langle U \rangle}$. Thus, we have $x \in \overline{\langle U \rangle} \Leftrightarrow (x - \bar{x}) \in \overline{\langle U \rangle}$. From (c), $x - \bar{x} \perp \langle U \rangle$. Thus, $x \in \overline{\langle U \rangle} \Leftrightarrow x - \bar{x} = 0$. Hence any $x \in \overline{\langle U \rangle}$ can be expressed as $x = \sum (u_\alpha, x) u_\alpha$.

(e) If $\sum_{\alpha \in I} a_\alpha u_\alpha$ with $\sum_{\alpha \in I} |a_\alpha|^2 < \infty$, then using the same argument of (b), we get the unorder sum $\sum_{\alpha \in I} a_\alpha u_\alpha$ converges. And hence it is in $\overline{\langle U \rangle}$. On the other hand, we have seen from (d) that any element x in $\overline{\langle U \rangle}$ can be expressed as

$$x = \sum_{\alpha \in I} (u_\alpha, x) u_\alpha,$$

with $\sum_{\alpha \in I} |(u_\alpha, x)|^2 < \infty$.

□

Theorem 5.2. *Let $U := \{u_\alpha | \alpha \in I\}$ be an orthonormal set in a Hilbert space $\mathcal{H}$. Then the following conditions are equivalent:*

(a) $(x, u_\alpha) = 0$ for all $\alpha \in I$ implies $x = 0$;

(b) Any $x \in \mathcal{H}$ can be represented as $x = \sum_{\alpha \in I} (u_\alpha, x) u_\alpha$;

(c) $\overline{\langle U \rangle} = \mathcal{H}$;

(d) The norm of any $x \in \mathcal{H}$ can be characterized by $\|x\|^2 = \sum_{\alpha \in I} |(u_\alpha, x)|^2$;

(e) U is a maximal orthonormal set.

Proof. We see that (a) $\Leftrightarrow$ $\langle U \rangle^\perp = \{0\} \Leftrightarrow \overline{\langle U \rangle}^\perp = 0 \Leftrightarrow \mathcal{H} = \overline{\langle U \rangle}$. The latter follows from the decomposition theorem $\mathcal{H} = \mathcal{M} \oplus \mathcal{M}^\perp$ for any closed subspace $\mathcal{M}$. This together with the previous theorem show the equivalent from (a) to (c). The equivalence between (b) and (d) follows from Lemma 5.1. To prove (e)$\Leftrightarrow$ (a), let us suppose there is a $v \notin U$ and $v \perp U$. That is, U is not maximal since $U \subset V := \{v\} \cup U$. Then from (a), $v = 0$. Conversely, U is maximal means that any orthonormal set V with $U \subset V$, then $U = V$. In other word, if there is a $v \in V \backslash U$, then $v = 0$. But this is the statement (a). □

An orthonormal set in $\mathcal{H}$ satisfying one of the statements of this theorem is called an orthonormal basis of $\mathcal{H}$. By (b), any element $x \in \mathcal{H}$ can be represented as $x = \sum_{\alpha \in I} c_\alpha u_\alpha$. By (d), this representation must be unique. And by (b), it is also represented as

$$x = \sum_{\alpha \in I} (u_\alpha, x) u_\alpha.$$

Theorem 5.3. *Any Hilbert space $\mathcal{H}$ has an orthonormal basis. If $\mathcal{H}$ is separable, then the orthonormal basis is countable.*

Proof. We consider the set $\mathcal{S}$ of all orthonormal sets in $\mathcal{H}$ and order them by the inclusion relation. We find that $\mathcal{S}$ has the property: any totally partial order family has an upper bounded. A total partial order family $\{U|U \in \mathcal{A}\}$ means any two of its elements, say U and V, is either $U \subset V$ or $V \subset U$. We see that its union $V = \cup_{U\in\mathcal{A}} U$ is an upper bound. With this property, by Zorn's lemma in the set theory, the set $\mathcal{S}$ has a maximal element. By the previous theorem, it is an orthonormal basis.

Next, we assume $\mathcal{H}$ is separable. That means there is a countable set $\mathcal{A} = \{v_i|i \in \mathbb{N}\}$ such that $\overline{\mathcal{A}} = \mathcal{H}$. Now, we construct a sequence of nested subspaces V_n and its basis $\{w_1, w_2, \cdots, w_n\}$ by the following procedure. Let $w_1 = v_1$, $V_1 = \langle\{v_1\}\rangle$. If $v_2 \notin V_1$, then define $w_2 = v_2$ and $V_2 = \langle\{v_1, v_2\}\rangle$. Otherwise, we skip v_2 and continue this process. Either we can find the next one, or we have exhausted all elements in $\mathcal{A}$. For the latter case, $\mathcal{H}$ is finite dimension. For the former case, we continue this process and select an infinite countable subset $\mathcal{B} := \{w_1, w_2, \cdots\}$ from $\mathcal{A}$ such that $\{v_1, \cdots, v_n\} \subset \langle\{w_1, \cdots, w_n\}\rangle \subset \langle\mathcal{B}\rangle$ for all n. Thus, $\mathcal{A} \subset \langle\mathcal{B}\rangle$ From $\overline{\mathcal{A}} = \mathcal{H}$, we get $\overline{\langle\mathcal{B}\rangle} = \mathcal{H}$.

We can construct an orthonormal set $\{u_1, u_2, \cdots\}$ from the independent set $\{w_1, w_2, \cdots\}$ by the Gram-Schmidt orthonormalization procedure. It is an induction procedure. We choose $u_1 = w_1/\|w_1\|$. Let

$$z_2 = w_2 - (w_2, u_1)u_1$$

and $u_2 = z_2/\|z_2\|$. Suppose $\{u_1, \cdots, u_n\}$ are found. We define

$$z_{n+1} = w_{n+1} - \sum_{i=1}^{n}(w_{n+1}, u_i)u_i$$

and $u_{n+1} = z_{n+1}/\|z_{n+1}\|$. Then, by construction, we have

$$\langle\{u_1, \cdots, u_n\}\rangle = \langle\{w_1, \cdots, w_n\}\rangle$$

for all n. Consequently

$$\langle\{w_1, \cdots, w_n\}\rangle \subset \overline{\langle\{u_1, u_2, \cdots\}\rangle}$$

By taking $n \to \infty$ and taking closure on the left-hand side, we get

$$\mathcal{H} = \overline{\langle u_1, u_2, \cdots\rangle}$$

□

For concrete Hilbert space such as $L^2(\Omega)$ with Ω having certain symmetry, one can find some natural orthonormal basis. Here are some important examples:

- In $\ell^2(\mathbb{N})$, the Cartesian unit vectors are defined as

$$e_1 := (1, 0, \cdots), e_2 := (0, 1, 0, \cdots), \cdots.$$

These Cartesian unit vectors form an orthonormal basis in $\ell^2(\mathbb{N})$.

- In $L^2(\mathbb{T})$, the trigonometric functions $\{e^{in\theta}/\sqrt{2\pi}\ |n \in \mathbb{Z}\}$ is an orthonormal basis.
- In the space $L^2_w(-1,1)$ with $w(x) = (1-x^2)^{-1/2}$, the Tchebyshev polynomials
$$T_n(x) := \cos\left(n \cos^{-1}(x)\right)$$
is an orthogonal basis. Indeed, the projection : $\theta \to x$ by $x = \cos\theta$ from the upper unit circle $\mathbb{T}^+$ to $[-1,1]$ is 1-1 and onto. The measure $d\theta$ on $\mathbb{T}$ induces the measure $(1-x^2)^{-1/2}dx$ on $(-1,1)$. The trigonometric functions $\cos n\theta$ correspond to the Tchebyshev polynomials on $(-1,1)$. You can check by induction that T_n are polynomials.
- In $L^2(-1,1)$ the Legendre polynomials can be constructed from the polynomials $\{1, x, x^2, \cdots\}$ by the Gram-Schmidt orthogonalization procedure.
- Hermite polynomials form orthogonal basis in $L^2_w(-\infty,\infty)$ with $w(x) = e^{x^2/2}$.
- Haar basis. the Haar function on $\mathbb{R}$ is defined to be
$$\psi(x) = \begin{cases} 1 & \text{for } 0 \le x < 1/2 \\ -1 & \text{for } 1/2 \le x < 1 \\ 0 & \text{otherwise.} \end{cases}$$
We can translate and rescale ψ to get
$$\psi_{j,k}(x) := 2^{j/2}\psi(2^j x - k),\ j,k \in \mathbb{Z}.$$
Then you can check $(\psi_{j,k}, \psi_{\ell,m}) = \delta_{j,\ell}\delta_{k,m}$. Indeed, $\mathcal{A} := \{\psi_{j,k}|j,k \in \mathbb{Z}\}$ is an orthonormal basis in $L^2(\mathbb{R})$. We shall leave the proof of $\mathcal{A}$ being an orthonormal set as an exercise. We shall discuss the completeness of $\mathcal{A}$ in later chapter.

Homeworks 5.1. *1. pp. 145-147: 6.12,*

2. Ex. 6.13,

3. Ex. 6.14

4. Let ψ be the Haar function on $\mathbb{R}$. Show that the set $\{\psi_{j,k}|j,k \in \mathbb{Z}\}$ is an orthonormal set in $L^2(\mathbb{R})$.

5.2 The Fourier basis in $L^2(\mathbb{T})$

5.2.1 Definition and examples

Definition We study Fourier expansion for 2π-periodic functions. Suppose f is a 2π-periodic function. Let us expand f as
$$f(x) \sim \sum_{k=-\infty}^{\infty} a_k e^{ikx}.$$

By taking the following inner product, defined by

$$(f,g) := \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\overline{g(x)}\,dx,$$

with e^{imx}, we find that

$$a_m = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-imx}\,dx.$$

a_m's are called the Fourier coefficients, or Fourier multiples. m is the wave number. We denote a_m by $\hat{f}_m$.

Examples

1. $$f(x) = \begin{cases} 1 & \text{for } 0 < x < \pi \\ -1 & \text{for } -\pi < x < 0 \end{cases}$$

2. $f(x) = \frac{1}{\pi}|x|$

5.2.2 Basic properties

A 2π-periodic function can be identified as a function on circle, which is $\mathbb{T} = \mathbb{R}/(2\pi\mathbb{Z})$. Some important properties of Fourier transform are

- The differentiation becomes a multiplication under Fourier transform. It is also equivalent to say that the differential operator is diagonalized in Fourier basis.
- The convolution becomes a multiplication under Fourier transform.

Differentiation

Lemma 5.2. *If $f \in C^1[\mathbb{T}]$, then*

$$\widehat{f'}_k = ik\hat{f}_k.$$

Proof.

$$\begin{aligned}
\widehat{f'}_k &= \frac{1}{2\pi}\int_0^{2\pi} f'(x)e^{-ikx}\,dx \\
&= \frac{1}{2\pi}\,e^{-ikx}f(x)\Big|_{x=0}^{x=2\pi} - \frac{1}{2\pi}\int_0^{2\pi}(-ik)e^{-ikx}f(x)\,dx \\
&= ik\hat{f}_k.
\end{aligned}$$

Here, we have used the periodicity of f in the last step. □

Convolution If f and g are in $L^2(\mathbb{T})$, we define the convolution of f and g by

$$(f*g)(x) = \int_{\mathbb{T}}\int_{\mathbb{T}} f(x-y)g(y)\,dy.$$

Lemma 5.3. *If $f,g \in C(\mathbb{T})$, then*

$$\left(\widehat{f*g}\right)_k = 2\pi \hat{f}_k \hat{g}_k.$$

Proof.

$$\begin{aligned}
\left(\widehat{f*g}\right)_k &= \frac{1}{2\pi}\int_{\mathbb{T}} f*g(x)e^{-ikx}\,dx \\
&= \frac{1}{2\pi}\int_{\mathbb{T}}\int_{\mathbb{T}} f(x-y)g(y)\,dye^{-ikx}\,dx \\
&= \frac{1}{2\pi}\int_{\mathbb{T}}\int_{\mathbb{T}} f(x-y)e^{-ik(x-y)}g(y)\,dye^{-iky}\,dx \\
&= \frac{1}{2\pi}\int_{\mathbb{T}}\left(\int_{\mathbb{T}} f(x-y)e^{-ik(x-y)}\,dx\right)g(y)e^{-iky}\,dy \\
&= \frac{1}{2\pi}\int_{\mathbb{T}}\left(\int_{\mathbb{T}} f(x)e^{-ikx}\,dx\right)g(y)e^{-iky}\,dy \\
&= 2\pi \hat{f}_k \hat{g}_k.
\end{aligned}$$

Here, we have used Fubini theorem. □

Remarks

1. The above two lemmae are also valid for f,g are in L^2. Their proofs are based on the L^2 convergence of the Fourier series for nice functions and the fact that nice functions are dense in L^2.

2. Many solutions of differential equations are expressed in convolution forms. For instance $-u'' = f$ in $\mathbb{T}$, its solution can be expressed as $u = g*f$, where g is the Green's function of $-d^2/dx^2$ on $\mathbb{T}$.

3. In image processing, a blurred image is modelled by

$$z(x) = \int k(x-y)f(y)\,dy$$

where $f(y)$ is the original image, z the blurred image, and

$$k(x) = \frac{1}{2\pi\sigma^2}e^{-|x|^2/2\sigma^2}$$

the blur operator.

Regularity and decay If f is smooth, then its Fourier coefficients decays very fast. Indeed, by taking integration by part n times, we have

$$\begin{aligned}\hat{f}_k &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-ikx}\,dx \\ &= \frac{1}{(-ik)^n}\frac{1}{2\pi}\int_{-\pi}^{\pi} f^{(n)}(x)e^{-ikx}\,dx\end{aligned}$$

Thus, if $f \in C^n$, we see $\hat{f}_k = O(|k|^{-n})$.[1] This can also be observed by the following arguments. We notice that

$$\hat{f}_k = -\frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-ik(x+\pi/k)}\,dx$$

Hence,

$$\begin{aligned}\hat{f}_k &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-ikx}\,dx \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{f(x)-f(x-\pi/k)}{2}e^{-ikx}\,dx \\ &:= \frac{1}{2\pi}\int_{-\pi}^{\pi} D_{\pi/k}f(x)e^{-ikx}\,dx \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} D_{\pi/k}^{n}f(x)e^{-ikx}\,dx\end{aligned}$$

Here, $D_{\pi/k}$ is a backward finite difference operator. Thus, $\hat{f}_k$ measures the oscillation of f at scale π/k. If f is smooth, then $D^n_{\pi/k}f = O(|k|^{-n})g(x)$ with g being uniformly bounded in k. Thus, $\hat{f}_k = O(|k|^{-n})$. Indeed we have better result:

Lemma 5.4. *If $f \in C^n(\mathbb{T})$, then $\hat{f}_k = o(|k|^{-n})$.*

We shall only need to show that $\hat{f}_k \to 0$ as $|k| \to \infty$ for continuous function f. The rest for high derivative cases can be obtained by taking integration by part. We have seen that

$$\begin{aligned}\hat{f}_k &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-ikx}\,dx \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{f(x)-f(x-\pi/k)}{2}e^{-ikx}\,dx\end{aligned}$$

When f is continuous on $\mathbb{T}$, it is uniformly continuous on $\mathbb{T}$. Thus, for any $\epsilon > 0$, we can find $K > 0$ such that for all $|k| > K$ we have

$$\left|\frac{f(x)-f(x-\pi/k)}{2}\right| < \epsilon.$$

[1] If fact, we shall see later from the Riemann-Lebesgue lemma that $\hat{f}_k = o(|k|^{-n})$.

From this, we obtain

$$|\hat{f}_k| \le \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{f(x)-f(x-\pi/k)}{2}e^{-ikx}\right|\,dx < \epsilon.$$

When f is not smooth, say in L^1, we still have $\hat{f}_k \to 0$ as $|k| \to \infty$. This is the following Riemann-Lebesgue lemma.

Lemma 5.5 (Riemann-Lebesgue)**.** *If f is in $L^1(a,b)$, then*

$$\hat{f}_k := \int_a^b f(x)e^{-ikx}\,dx \to 0, \text{ as } |k| \to \infty.$$

Proof. 1. For $f \in L^1(a,b)$, we have

$$|\hat{f}_k| \le \|f\|_1 \text{ for all } k.$$

2. Any function $f \in L^1(a,b)$ can be approximated by a continuous function $g \in C[a,b]$ in the L^1 sense. That is, for any $\epsilon > 0$, there exists $g \in C[a,b]$ such that $\|f-g\|_1 < \epsilon$.

3. For $g \in C[a,b]$, we have: for any $\epsilon > 0$, there exists a $K > 0$ such that for $|k| > K$, we have

$$|\hat{g}_k| < \epsilon.$$

Combining these two, we get

$$|\hat{f}_k| \le |\hat{g}_k| + |\hat{f}_k - \hat{g}_k| \le |\hat{g}_k| + \|f-g\|_1 < 2\epsilon.$$

Thus, $|\hat{f}_k| \to 0$ as $|k| \to \infty$.

□

Remarks.

1. If f is a Dirac delta function, we can also define its Fourier transform

$$\hat{f}_k = \frac{1}{2\pi}\int_{-\pi}^{\pi}\delta(x)e^{-ikx}\,dx = \frac{1}{2\pi}.$$

In this case, $\delta \notin L^1$ and $\hat{\delta}_k = 1/2\pi$ does not converge to 0 as $|k| \to \infty$.

2. If f is a piecewise smooth function with finite many jumps, then it holds that $\hat{f}_k = O(1/k)$. One may consider f has only one jump first. Then f is a superposition of a step function g and a smooth function h. We have seen that $\hat{h}_k$ decays fast. For the step function g, we have $\hat{g}_k = O(1/k)$.

5.2.3 Convergence Theory

Let denote the partial sum of the Fourier expansion by f_N:

$$f_N(x) := \sum_{k=-N}^{N} \hat{f}_k e^{ikx}.$$

We shall show that under proper condition, f_N will converge to f. The convergence is in the sense of uniform convergence for smooth functions, in L^2 sense for L^2 functions, and in pointwise sense for BV functions.

Convergence theory for Smooth functions

Theorem 5.4. *If f is a 2π-periodic, C^∞-function, then for any $n > 0$, there exists a constant C_n such that*

$$|f_N(x) - f(x)| \le C_n N^{-n}. \tag{5.2}$$

Proof. We can express f_N in convolution form: $f_N = D_N * f$:

$$\begin{aligned} f_N(x) &:= \sum_{|k|\le N} \hat{f}_k e^{ikx} \\ &= \sum_{|k|\le N} \frac{1}{2\pi}\int_{-\pi}^{\pi} f(y)e^{ik(x-y)}\,dy \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\sin(N+\frac{1}{2})(x-y)}{\sin(\frac{1}{2}(x-y))} f(y)\,dy \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\sin(N+\frac{1}{2})t}{\sin\frac{t}{2}} f(x+t)\,dt \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} D_N(t) f(x+t)\,dt \end{aligned}$$

Here,

$$D_N(x) := \sum_{|k|\le N} e^{ikx} = \frac{\sin(N+1/2)x}{\sin(x/2)}$$

is called the Dirichlet kernel. Using $D_N(x)dx = \pi$, we have

$$\begin{aligned} f_N(x) - f(x) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\sin(N+\frac{1}{2})t}{\sin\frac{t}{2}}(f(x+t)-f(x))\,dt \\ &:= \frac{1}{2\pi}\int_{-\pi}^{\pi} \sin((N+\frac{1}{2})t)g(x,t)\,dt \end{aligned}$$

The function $g(x,t) := (f(x+t)-f(x))/\sin(t/2) = \int_0^1 f'(x+st)\,ds \cdot t/\sin(t/2)$ is 2π periodic and in C^∞. We can apply integration-by-part n times to arrive

$$f_N(x) - f(x) = (N+\frac{1}{2})^{-n}\frac{(-1)^{n/2}}{2\pi}\int_{-\pi}^{\pi} \partial_t^n g(t)\sin((N+\frac{1}{2})t)\,dt$$

for even n. Similar formula for odd n. Thus, we get

$$\sup_x |f_N(x)-f(x)| \le CN^{-n}\left|\int \partial_t^n g(x,t)\sin((tN+\frac{1}{2})t)\,dt\right| = O(N^{-n}).$$

This completes the proof. □

Remark. The constant C_n, which depends on $\int |g^{(n)}|\,dt$, is in general not big, as compared with the term N^{-n}. Hence, the approximation (5.2) is highly efficient for smooth functions. For example, $N = 20$ is sufficient in many applications. The accuracy property (5.2) is called spectral accuracy.

5.2.4 L^2 Convergence Theory

The L^2 convergence theory states that:

Theorem 5.5. *If $f \in L^2(\mathbb{T})$, then the Fourier expansion $f_N(f) \to f$ in $L^2(\mathbb{T})$. In other word, $\{e^{ikx} | k \in \mathbb{Z}\}$ constitutes an orthonormal basis in $L^2(\mathbb{T})$.*

Proof. In the proof below, I shall use the fact that $C^\infty(\mathbb{T})$ is dense in $L^2(\mathbb{T})$. I shall not prove this theorem. We can prove the L^2 convergence by the following two equivalent arguments.

1. We have seen that C^∞ can be approximated by trigonometric polynomials. That is, $C^\infty(\mathbb{T}) \subset \overline{\langle U\rangle}$, where $U = \{e^{ikx} | k \in \mathbb{Z}\}$. From $\overline{C^\infty(\mathbb{T})} = L^2(\mathbb{T})$, we get $L^2(\mathbb{T}) = \overline{\langle U\rangle}$.

2. Alternatively, we show that if $f \in L^2(\mathbb{T})$ and $f \perp e^{ikx}$ for all $k \in \mathbb{Z}$, then $f = 0$. For any $f \in C^\infty(\mathbb{T})$, if $(f, e^{ikx}) = 0$ for all $k \in Z$, from its finite Fourier expansion f_N, which is zero, converges to f, we get that $f \equiv 0$. Thus, $U^\perp \cap C^\infty(\mathbb{T}) = \{0\}$. For arbitrary $f \in L^2(\mathbb{T})$, suppose $f \perp e^{ikx}$ for all k. We regularize f by $f_\epsilon := \rho_\epsilon * f \to f$ in $L^2(\mathbb{T})$. But the Fourier coefficients $\widehat{f_\epsilon}$ are
$$\left(\widehat{\rho_\epsilon * f}\right)_k = (\widehat{\rho}_\epsilon)_k\,(\hat{f})_k = 0.$$
Thus, $f_\epsilon \perp e^{ikx}$ for all $k \in \mathbb{Z}$. This together with $f_\epsilon \in C^\infty(\mathbb{T})$ give $f_\epsilon \equiv 0$. Since $f_\epsilon \to f$ in $L^2(\mathbb{T})$, we get $f \equiv 0$ also. We conclude that $f \perp U$ implies $f = 0$.

□

Remark By the general theorem of orthogonal basis in Hilbert space, we have seen that (a) $U^\perp = \{0\} \Leftrightarrow$ (b) $\overline{\langle U\rangle} = L^2(\mathbb{T}) \Leftrightarrow$ (c) Parvesal equality: $\|f\|^2 = \sum_k |\hat{f}_k|^2$. Yet, we shall state and prove them below.

The Fourier transform maps a 2π-periodic function f into its Fourier coefficients $(\hat{f}_k)_{k=-\infty}^\infty$. We may view the Fourier transform maps $L^2(\mathbb{T})$ space into ℓ^2 space. The function spaces L^2 and ℓ^2 are defined below.

$$L^2(\mathbb{T}) := \{f \mid f \text{ is } 2\pi \text{ periodic and } \int_{-\pi}^{\pi} |f(x)|^2\,dx < \infty\}$$

with the inner product

$$(f,g) := \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\overline{g(x)}\,dx$$

and L^2-norm: $\|f\| = \sqrt{(f,f)}$. The space $\ell^2(\mathbb{Z})$ is defined as

$$\ell^2(\mathbb{Z}) := \{(a_k)_{k=-\infty}^\infty \mid \sum_{k=-\infty}^{\infty} |a_k|^2 < \infty\}.$$

with inner product $(a,b) := \sum_k a_k\overline{b_k}$.

From this, we have for any N,

$$0 \le (f - f_N, f - f_N) = \|f\|^2 - \sum_{|k|\le N} |\hat{f}_k|^2.$$

This gives

$$\sum_{k=-\infty}^{\infty} |\hat{f}_k|^2 \le \|f\|^2. \tag{5.3}$$

This is the Bessel inequality. It says that the Fourier transform maps continuously from $L^2(\mathbb{T})$ to $\ell^2(\mathbb{Z})$.

Theorem 5.6 (Isometry property)**.** *The Fourier transform is an isometry from $L^2(\mathbb{T})$ to $\ell^2(\mathbb{Z})$:*

$$(f,g) = \sum_k \hat{f}_k\overline{\hat{g}_k}.$$

Proof. To show this, we first assume that f is a smooth function. We can apply the convergence theorem for f. This yields

$$\begin{aligned}(f,g) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\overline{g(x)}\,dx\\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_k \hat{f}_k e^{ikx}\overline{g(x)}\,dx\\ &= \sum_k \hat{f}_k\overline{\hat{g}_k}.\end{aligned}$$

In the last equality, the summation in k converges fast and is independent of x (from smoothness of f). This implies that we can interchange the integration in x and the summation in k.

To show this formula is also valid for all $f, g \in L^2$, we approximate f by $f_\epsilon := \rho_\epsilon * f$, which are in C^∞ and converge to f in L^2. The isometry property is valid for f_ϵ and g: $(f_\epsilon, g) = (\widehat{f_\epsilon}, \hat{g})$. As $\epsilon \to 0$,

$$|(f_\epsilon - f, g)| \le \|f_\epsilon - f\|\|g\| \to 0,$$

and

$$|(\widehat{f_\epsilon} - \hat{f}, \hat{g})| \le \|\widehat{f_\epsilon} - \hat{f}\|\|\hat{g}\| \le \|f_\epsilon - f\|\|g\| \to 0.$$

The last inequality is from the Bessel inequality. Thus, we obtain $(f, g) = (\hat{f}, \hat{g})$. □

The isometry property says that the Fourier transformation preserves the inner product. When $g = f$ in the above isometry property, we obtain the following Parseval identity.

Corollary 5.2 (Parseval identity)**.** *For $f \in L^2$, we have*

$$\|f\|^2 = \sum_k |\hat{f}_k|^2.$$

Theorem 5.7 (L^2-convergence theorem)**.** *If $f \in L^2$, then*

$$f_N = \sum_{k=-N}^{N} \hat{f}_k e^{ikx} \to f \textit{ in } L^2.$$

Proof. First, the sequence $\{f_N\}$ is a Cauchy sequence in L^2. This follows from $\|f_N - f_M\| = \sum_{N \le |k| < M} |\hat{f}_k|^2$ and the Bessel inequality. Suppose f_N converges to g. Then it is easy to check that the Fourier coefficients of $f - g$ are all zeros. From the Parvesal identity, we have $f = g$. □

5.2.5 BV Convergence Theory

A function is called a BV function (or a function of finite total variation) on an interval (a, b), if for any partition $\pi = \{a = x_0 < x_1 < \cdots < x_n = b\}$,

$$\|f\|_{BV} := \sup_\pi \sum_i |f(x_i) - f(x_{i-1})| < \infty.$$

An important property of BV function is that its singularity can only be jump discontinuities, i.e., at a discontinuity, say, x_0, f has both left limit $f(x_0-)$ and right limit $f(x_0+)$.

Further, any BV function f can be decomposed into $f = f_0 + f_1$, where f_0 is a piecewise constant function andf_1 is absolutely continuous (i.e. f_1 is differentiable and f_1' is integrable). The jump points of f_0 are countable. The BV-norm of f is exactly equal to

$$\|f\|_{BV} = \sum_i |[f(x_i)]| + \int |f_1'(x)|\, dx.$$

where x_i are the jump points of f (also f_0) and $[f(x_i)] := f(x_i+) - f(x_i-)$ is the jump of f at x_i.

Theorem 5.8 (Fourier inversion theorem for BV functions). *If f is in BV (function of bounded variation), then*

$$f_N(x) := \sum_{k=-N}^{N} \hat{f}_k e^{ikx} \to \frac{1}{2}(f(x+) + f(x-)).$$

Proof. Recall that

$$\begin{aligned} f_N(x) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} D_N(x-y)f(y)\,dy \\ &= \left(\int_{-\pi}^{0} + \int_{0}^{\pi}\right) D_N(t)f(x+t)\,dt \\ &= f_N^+(x) + f_N^-(x). \end{aligned}$$

Here, $D_N(x) = \sum_{|k|\le N} e^{ikx} = \frac{\sin(N+1/2)x}{\sin(x/2)}$. Using $\int_0^\pi \frac{\sin(N+1/2)x}{\sin(x/2)}\,dx = \pi$, we have

$$\begin{aligned} f_N^+(x) - \frac{1}{2}f(x+) &= \frac{1}{2\pi}\int_0^\pi \frac{\sin(N+\frac{1}{2})t}{\sin\frac{t}{2}}(f(x+t) - f(x))\,dt \\ &:= \frac{1}{2\pi}\int_0^\pi \sin((N+\frac{1}{2})t)g(t)\,dt \end{aligned}$$

From f being in BV, the function $g(t)$ is in $L^1(0,\pi)$. By the Riemann-Lebesgue lemma, $f_N^+(x) - \frac{1}{2}f(x+) \to 0$ as $N \to \infty$. Similarly, we have $f_N^-(x) - f(x-) \to 0$ as $N \to \infty$. □

Gibbs phenomena In applications, we encounter piecewise smooth functions frequently. In this case, the approximation is not uniform. An overshoot and undershoot always appear across discontinuities. Such a phenomenon is called Gibbs phenomenon. Since a BV function can be decomposed into a piecewise constant function and a smooth function, we concentrate to the case when there is only one discontinuity. The typical example is the function

$$f(x) = \begin{cases} 1 & \text{for } 0 < x < \pi \\ -1 & \text{for } -\pi < x < 0 \end{cases}$$

The corresponding f_N is

$$f_N(x) = \frac{1}{2\pi}\int_{x-\pi}^{x} \frac{\sin((N+\frac{1}{2})t)}{\sin(t/2)}\,dt - \frac{1}{2\pi}\int_{x}^{x+\pi} \frac{\sin((N+\frac{1}{2})t)}{\sin(t/2)}\,dt$$

First, we show that we may replace $\frac{1}{2\sin(t/2)}$ by $\frac{1}{t}$ with possible error $o(1/N)$. This is because the function $\frac{1}{t} - \frac{1}{2\sin(t/2)}$ is in C^1 on $[-\pi,\pi]$ and the Riemann-Lebesgue lemma. Thus, we have

$$\begin{aligned} f_N(x) &= \frac{1}{\pi}\int_{x-\pi}^{x} \frac{\sin((N+\frac{1}{2})t)}{t}\,dt - \frac{1}{\pi}\int_{x}^{x+\pi} \frac{\sin((N+\frac{1}{2})t)}{t}\,dt + o(1/N) \\ &= \frac{1}{\pi}\int_{(x-\pi)(N+1/2)}^{x(N+1/2)} \operatorname{sinc}(t)\,dt - \frac{1}{\pi}\int_{x(N+1/2)}^{(x+\pi)(N+1/2)} \operatorname{sinc}(t)\,dt + o(1/N). \end{aligned}$$

Here, the function sinc$(t) := \sin(t)/t$. It has the following properties:

$$\int_0^\infty \text{sinc}(t)\,dt = \pi/2.$$

For any $z > 0$,

$$\int_z^\infty \text{sinc}(t)\,dt = O\left(\frac{1}{z}\right)$$

To see the latter inequality, we rewrite

$$\int_z^\infty \text{sinc}(t)\,dt = \left(\int_z^{n\pi} + \sum_{k\ge n}\int_{n\pi}^{(n+1)\pi}\right)\text{sinc}(t)\,dt$$

where $n = [z/\pi] + 1$. Notice that the series is an alternating series. Thus, the series is bounded by its leading term, which is of $O(1/z)$. Let us denote the integral $\int_0^z \text{sinc}(t)\,dt$ by Si(z).

To show that the sequence f_N does not converge uniformly, we pick up $x = z/(N+1/2)$ with $z > 0$. After changing variable, we arrive

$$\begin{aligned} f_N(\frac{z}{(N+1/2)}) &= \frac{1}{\pi}\int_{z-(N+1/2)\pi}^{z} \text{sinc}(t)\,dt - \frac{1}{\pi}\int_z^{z+(N+1/2)\pi} \text{sinc}(t)\,dt + o(1/N) \\ &= \frac{1}{\pi}\int_{-\infty}^{z} \text{sinc}(t)\,dt - \frac{1}{\pi}\int_z^{\infty} \text{sinc}(t)\,dt + O(1/(z+N)) + O(1/(z-N)) \\ &= \frac{2}{\pi}\int_0^z \text{sinc}(t)\,dt + (1/(z+N)) + O(1/(z-N)) \\ &= 1 - \frac{2}{\pi}\int_z^\infty \text{sinc}(t)\,dt + (1/(z+N)) + O(1/(z-N)) \end{aligned}$$

In general, for function f with arbitrary jump at 0, we have

$$\begin{aligned} f_N(\frac{z}{(N+1/2)}) &= f(0+) - \frac{[f]}{\pi}\int_z^\infty \text{sinc}(t)\,dt + (1/(z+N)) + O(1/(z-N)) \\ &= f(0+) + O(1/z) + O(1/(z-N)). \end{aligned}$$

where, the jump $[f] := f(0+) - f(0-)$.

We see that the rate of convergence is slow if $z = N^\alpha$ with $0 < \alpha < 1$. This means that if the distance of x and the nearest discontinuity is $N^{-1+\alpha}$, then the convergent rate at x is only $O(N^{-\alpha})$. If the distance is $O(1)$, then the convergent rate is $O(N^{-1})$. This shows that the convergence is not uniform.

The maximum of Si(z) indeed occurs at $z = \pi$ where

$$\frac{1}{\pi}\text{Si}(\pi) \approx 0.58949$$

This yields

$$f_N(\frac{\pi}{N+1/2}) = f(0+) + 0.08949\,(f(0+) - f(0-)).$$

Hence, there is about 9% overshoot. This is called Gibbs phenomenon.

5.2.6 Fourier Expansion of Real Valued Functions

We have

$$\hat{f}_n = \frac{1}{2\pi}\int_{\mathbb{T}} f(x)e^{-inx}\,dx, \hat{f}_{-n} = \frac{1}{2\pi}\int f(x)e^{inx}\,dx.$$

Thus, when f is real valued,

$$\hat{f}_n = \overline{\hat{f}_{-n}}.$$

If we express $\hat{f}_n = \frac{1}{2}(a_n - ib_n)$, where $a_n, b_n \in \mathbb{R}$, then $\hat{f}_{-n} = \frac{1}{2}(a_n + ib_n)$ and

$$\begin{aligned} f(x) &= \sum_{n\in\mathbb{Z}} \hat{f}_n e^{inx} \\ &= \frac{1}{2}a_0 + \frac{1}{2}\sum_{n=1}^{\infty}(a_n - ib_n)e^{inx} + \frac{1}{2}\sum_{n=1}^{\infty}(a_n + ib_n)e^{-inx} \\ &= \frac{1}{2}a_0 + \sum_{n=1}^{\infty}(a_n\cos nx + b_n \sin nx) \end{aligned}$$

Here,

$$\begin{aligned} \frac{1}{2}(a_n - ib_n) &= \frac{1}{2\pi}\int_{\mathbb{T}} f(x)e^{-inx}\,dx \\ &= \frac{1}{2\pi}\int_{\mathbb{T}} f(x)\left(\cos nx - i\sin nx\right)\,dx. \end{aligned}$$

Thus,

$$a_n = \frac{1}{2\pi}\int_0^{2\pi} f(x)\cos nx\,dx,\ b_n = \frac{1}{2\pi}\int_0^{2\pi} f(x)\sin nx\,dx.$$

The functions $\{\cos nx, \sin nx\}$ are orthogonal to each other. But

$$\frac{1}{2\pi}\int_0^{2\pi}\cos^2 nx\,dx = \frac{1}{2\pi}\int_0^{2\pi}\sin^2 nx\,dx = \frac{1}{2} \text{ for all } n.$$

The Parseval equality reads

$$\frac{1}{2\pi}\int_{\mathbb{T}} f(x)^2\,dx = 2\sum_n \left(a_n^2 + b_n^2\right).$$

Homeworks 5.2.

1. Derive the Fourier expansion formula for periodic functions with period L.

2. What is the limit of the above Fourier expansion formula as $L \to \infty$.

3. Derive the Fourier expansion for the following functions: $f(x) = |x| - 1/2$ for $|x| \leq 1$ and f is a periodic function with period 2.

4. What is the convergence rate of the above function in L^2 and pointwise convergence rate at $x = 0$?

5.3 Applications of Fourier expansion

5.3.1 Characterization of Sobolev spaces

Let $H^m(\mathbb{T})$ be the completion of $C^\infty(\mathbb{T})$ under the norm

$$\|u\|_{H^m}^2 := \|u\|^2 + \|u'\|^2 + \cdots + \|u^{(m)}\|^2.$$

From $\widehat{u'}_k = ik\hat{u}_k$, we get

$$\widehat{u^{(m)}}_k = (ik)^m \hat{u}_k.$$

From Parseval equality, we obtain

$$\|u\|^2 = \sum_{k\in\mathbb{Z}} |\hat{u}_k|^2, \cdots, \|u^{(m)}\|^2 = \sum_{k\in\mathbb{Z}} |k|^{2m}|\hat{u}_k|^2.$$

Thus, we have

$$\|u\|_{H^m}^2 = \sum_{k\in\mathbb{Z}} (1 + |k|^2 + \cdots + |k|^{2m})|\hat{u}_k|^2.$$

The regularity of u is characterized by $u, ..., u^{(m)} \in L^2$. On the other hand, it is also described by

$$\sum_{k\in\mathbb{Z}} (1 + |k|^2 + \cdots + |k|^{2m})|\hat{u}_k|^2 < \infty,$$

which is an equivalent way to characterize the decay of $\hat{u}_k$.

Remark. Notice that for a fixed $m \geq 0$, the following quantities are equivalent

$$(1 + |k|)^{2m} \sim (1 + |k|^2)^m \sim (1 + |k|^{2m}) \text{ for all } k \in \mathbb{Z}.$$

This means that there are positive constants C_i such that

$$(1 + |k|)^{2m} \leq C_1(1 + |k|^2)^m \leq C_2(1 + |k|^{2m}) \leq C_3(1 + |k|)^{2m} \text{ for all } k \in \mathbb{Z}.$$

Thus, the Sobolev norm $\|u\|_{H^m}^2$ with $m \geq 0$ is equivalent to

$$\sum_{k\in\mathbb{Z}} (1 + |k|^{2m})|\hat{u}_k|^2.$$

We can also define Sobolev space with negative exponent m by

$$\|u\|_{H^m}^2 = \sum_{k\in\mathbb{Z}} (1 + |k|)^{2m}|\hat{u}_k|^2.$$

When m is large enough, the Sobolev space $H^m(\mathbb{T})$ can be embedded into $C(\mathbb{T})$. This is the following theorem.

Theorem 5.9. *For $m > 1/2$, we have $H^m(\mathbb{T}) \subset C(\mathbb{T})$.*

Proof. 1. For any smooth function f, we have

$$\begin{aligned} |f(x)| &= \left|\sum_k \hat{f}_k e^{ikx}\right| \\ &\le \sum_k \left|(1+|k|)^{-m}(1+|k|)^m \hat{f}_k\right| \\ &\le \left(\sum_k (1+|k|)^{-2m}\right)^{1/2} \left(\sum_k (1+|k|)^{2m}|\hat{f}_k|^2\right)^{1/2} \end{aligned}$$

When $m > 1/2$, then

$$\sum_{k\in\mathbb{Z}} (1+|k|)^{-2m} < \infty$$

Thus, we obtain

$$\|f\|_\infty \le C\|f\|_{H^m}$$

2. For any $f \in H^m(\mathbb{T})$, we can approximate f by f_N in H^m. (Check by yourself) From

$$\|f_N - f_M\|_\infty \le C\|f_N - f_M\|_{H^m}$$

we get that f_N is Cauchy in uniform norm. Thus it converges to f in $\|\cdot\|_\infty$.

□

5.3.2 Heat equation on a circle

We can solve the heat flow on circle exactly. This problem indeeds motivated Fourier invent the Fourier expansion. Let us consider

$$u_t = u_{xx}, x \in \mathbb{T}$$

with initial data

$$u(x,0) = f(x).$$

If we expand $u(x,t) = \sum_{n\in\mathbb{Z}} u_n(t)e^{inx}$, then, formally,

$$\sum_{n\in\mathbb{Z}} \dot{u}_n e^{inx} = \sum_{n\in\mathbb{Z}} -n^2 u_n e^{inx}.$$

Since $\{e^{inx}|n \in \mathbb{Z}\}$ are independent, we get

$$\dot{u}_n = -n^2 u_n.$$

Thus,

$$u_n(t) = u_n(0)e^{-n^2 t}.$$

At $t \to 0$, we expect $u_n(0) = \hat{f}_n$. Thus, we define the function

$$u(x,t) = \sum_{n\in\mathbb{Z}} \hat{f}_n e^{-n^2 t} e^{inx}.$$

In the following, we need to check:

1. u, u_t and u_{xx} exist and $u_t = u_{xx}$ for $t > 0$ and $x \in \mathbb{T}$;
2. $u(\cdot,t) \to f$ in $L^2(\mathbb{T})$ as $t \to 0+$.

- Proof of (1). We show that u_x exists here. It is clearly that $\sum_{n\in\mathbb{Z}} \hat{f}_n e^{-n^2t} e^{inx}$ converges absolute and uniformly w.r.t. x for $t > 0$, as long as $\hat{f}_n$ grows at most algebraically in n. Since $\sum_{n\in\mathbb{Z}} ine^{-n^2t}\hat{f}_n e^{inx}$ converges absolute and uniformly w.r.t. x for $t > 0$. This implies u is differentiable in x and the differentiation can be interchange with the infinite summation:

  $$\partial_x u = \partial_x \sum_{n\in\mathbb{Z}} \hat{f}_n e^{-n^2t} = \sum_{n\in\mathbb{Z}} \hat{f}_n e^{-n^2t} ine^{inx}.$$

  Similar proof for the existence of u_{xx} and u_t for $t > 0$. Since the Fourier coefficients of u_t and u_{xx} are identical on $t > 0$, we thus get $u_t = u_{xx}$.

- Proof of (2). Let us denote $u(\cdot,t)$ by $T(t)f$ and itself Fourier transform $\hat{u}_k(t)$ by $\hat{T}f$, or $\widehat{T}(t)\widehat{f}$. T is a linear operator from $L^2(\mathbb{T})$ to itself, while $\widehat{T}(t)$ a linear operator in $\ell^2(\mathbb{Z})$. We have

  $$\widehat{T}(t)\widehat{f}_n = e^{-n^2t}\hat{f}_n.$$

  Our goal is to prove $T(t)f \to f$ in $L^2(\mathbb{T})$. By the isometry property of the Fourier transform, this is equivalent to $\widehat{T}(t)\widehat{f} \to \widehat{f}$ in $\ell^2(\mathbb{Z})$. We have

  $$\begin{aligned}\lim_{t\to 0+} \|\widehat{T}(t)\widehat{f} - \widehat{f}\|_2^2 &= \lim_{t\to 0+} \sum_{n\in\mathbb{Z}} |(e^{-n^2t} - 1)^2|\widehat{f}_n|^2 \\ &= \sum_{n\in\mathbb{Z}} \lim_{t\to 0+} |(e^{-n^2t} - 1)^2|\widehat{f}_n|^2 = 0.\end{aligned}$$

  The interchange of $\sum$ and $\lim$ here is due to the dominant convergence theorem and the convergence of

  $$\sum_{n\in\mathbb{Z}} |(e^{-n^2t} - 1)^2|\widehat{f}_n|^2 \le 2^2 \sum_{n\in\mathbb{Z}} |\widehat{f}_n|^2 < \infty.$$

  is uniform w.r.t. t.

5.3.3 Solving Laplace equation on a disk

We consider the Laplace equation

$$u_{xx} + u_{yy} = 0$$

on the domain

$$\Omega : x^2 + y^2 < 1,$$

with the Dirichlet boundary condition:

$$u = f \text{ on } \partial\Omega.$$

In the polar coordinate, the equation has the form:

$$u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\theta\theta} = 0.$$

The boundary condition is

$$u(1,\theta) = f(\theta), \theta \in \mathbb{T}.$$

The solution is expanded as

$$u(r,\theta) = \sum_{n\in\mathbb{Z}} u_n(r)e^{in\theta}.$$

Plug this into the Laplace equation, we get

$$\sum_{n\in\mathbb{Z}} \left(u_n'' + \frac{1}{r}u_n' - \frac{n^2}{r^2} \right) e^{in\theta} = 0.$$

This leads to

$$u_n'' + \frac{1}{r}u_n' - \frac{n^2}{r^2} = 0 \text{ for all } n \in \mathbb{Z}.$$

The two independent solutions are $u_n = r^n$ or $u_n = r^{-n}$. However, the one with negative power will not satisfy the finiteness of u at $r = 0$. Thus, we obtain

$$u(r,\theta) = \sum_{n\in\mathbb{Z}} a_n r^{|n|} e^{in\theta}.$$

At $r = 1$, we get

$$f(\theta) = \sum_{n\in\mathbb{Z}} a_n e^{in\theta}.$$

Thus,

$$a_n = \frac{1}{2\pi}\int_{\mathbb{T}} f(\theta)e^{-in\theta}\,d\theta.$$

For $r < 1$, the L^2 norm of the infinite series $\sum_{n\in\mathbb{Z}} a_n r^{|n|} e^{in\theta}$ is bounded by $\sum_{n\in\mathbb{Z}} |a_n|^2$, uniformly in $r < 1$. Thus, from dominant convergence theorem, we have

$$\lim_{r\to 1-} u(r,\cdot) = f(\cdot) \text{ in } L^2(\mathbb{T}).$$

If we differentiate the infinite series in r term-by-term, we get

$$\sum_{n\in\mathbb{Z}} a_n |n| r^{|n|-1} e^{in\theta}.$$

This infinite series converges absolutely and uniformly for $r \leq r_0$ for any fixed $r_0 < 1$. This implies that u is differentiable in r and the differentiation can be performed term-by-term in the infinite series:

$$\partial_r u = \sum_{n\in\mathbb{Z}} a_n |n| r^{|n|-1} e^{in\theta}.$$

By the same argument, we get $\partial_{rr} u$ and $\partial_{\theta\theta} u$ exist and u satisfies the Laplace equation in polar coordinate form.

Alternatively, we can write the above summation in convolution form:

$$u(r,\theta) = \int_{\mathbb{T}} g(r,\theta-\phi) f(\phi)\, d\phi,$$

where

$$\begin{aligned} g(r,\theta) &= \frac{1}{2\pi} \sum_{n\in\mathbb{Z}} r^{|n|} e^{in\theta} \\ &= \frac{1}{2\pi}\left(\frac{1}{1-re^{i\theta}} + \frac{e^{-i\theta}}{1-e^{-i\theta}}\right) \\ &= \frac{1}{2\pi}\frac{1-r^2}{1-2r\cos\theta + r^2} \end{aligned}$$

The function g is called the Poisson kernel. It is infinitely differentiable for $r < 1$. This implies $g * f \in C^\infty(\Omega)$.

5.3.4 Hurwitz's proof for isoperimetric inequality (see Hunter's book)

The isoperimetric inequality involves to find the maximal area enclosed by a simple closed curve with given perimeter. If the perimeter is L, the area is A, then the isoperimeter inequality is

$$4\pi A \leq L^2.$$

The equality holds when the closed curve is a circle. There are many proofs of this inequality. In 1902, Hurwitz provided a proof using Fourier expansion. Let us show his proof here as an application of Fourier expansion. Let the closed curve is given by $(x, y) = (f(s), y(s))$, where s is the arc length. We may assume the length of the curve is 2π, otherwise we rescale it by (x, y) by $(2\pi x/L, 2\pi y/L)$. Since s is the arc length, we have

$$\dot{f}(s)^2 + \dot{g}(s)^2 = 1.$$

The area of the enclosed region is given by

$$A = \frac{1}{2}\int_{\mathbb{T}} f(s)\dot{g}(s) - g(s)\dot{f}(s)\, ds.$$

Our goal is to maximize A subject to the perimeter constraint

$$\int_{\mathbb{T}} \dot{f}(s)^2 + \dot{g}(s)^2\, ds = 2\pi.$$

We expand f and g in Fourier series:

$$f(s) = \sum_{n=-\infty}^{\infty} \hat{f}_n e^{ins},\ g(s) = \sum_{n=-\infty}^{\infty} \hat{g}_n e^{ins}$$

From the Parvesal equality:

$$\frac{1}{2\pi}\int_{\mathbb{T}} |\dot{f}(s)|^2\, ds = \sum_{n\in\mathbb{Z}} n^2 |\hat{f}_n|^2.$$

Thus, we have

$$1 = \sum_{n\in\mathbb{Z}} n^2 \left(|\hat{f}_n|^2 + |\hat{g}_n|^2\right).$$

For the area functional, we get

$$\frac{A}{\pi} = \frac{1}{2\pi}\int_{\mathbb{T}} f\dot{g} - g\dot{f}\, ds = \sum_{n\in\mathbb{Z}} \hat{f}_n \overline{in\hat{g}_n} - \hat{g}_n \overline{in\hat{f}_n} = -2\sum_{n\in\mathbb{Z}} nIm(\overline{\hat{g}_n}\hat{f}_n).$$

Subtracting these two series, we get

$$1 - \frac{A}{\pi} = \sum_{n\neq 0} \left(|n\hat{f}_n - i\hat{g}_n|^2 + |n\hat{g}_n + i\hat{f}_n|^2 + (n^2-1)(|\hat{f}_n|^2 + |\hat{g}_n|^2)\right).$$

We then get

$$1 - \frac{A}{\pi} \geq 0.$$

The equality holds only when $\hat{f}_n = \hat{g}_n = 0$ for all $n \geq 2$ and $\hat{f}_1 = i\hat{g}_1$. Plug this into the arc length constraint, we get

$$\hat{f}_1 = \frac{1}{\sqrt{2}}e^{i\delta},\ \hat{g}_1 = \frac{i}{\sqrt{2}}e^{i\delta}.$$

Thus,

$$f(s) = x_0 + \cos(s+\delta),\ g(s) = y_0 + \sin(s+\delta).$$

Homework Derive the Euler-Lagrange for the constrained maximization problem:

$$\max \frac{1}{2}\int_{\mathbb{T}} f(s)\dot{g}(s) - g(s)\dot{f}(s)\, ds$$

subject to

$$\int_{\mathbb{T}} \sqrt{\dot{f}(s)^2 + \dot{g}(s)^2}\, ds = 2\pi.$$

Show the isoperimetric inequality.

5.3.5 Von Neumann stability analysis for finite difference methods

In numerical PDEs, the stability analysis is a crucial step to the convergence theory of a numerical scheme. Below, I shall demonstrate the von Neumann stability analysis for heat equation in one dimension. It is a L^2 stability analysis suitable for for (the interior part of) numerical PDEs with constant coefficients.

Let us consider the heat equation:

$$u_t = u_{xx}$$

in one dimension with initial data $u(x, 0) = f(x)$. Let $h = \Delta x$, $k = \Delta t$ be the spatial and temporal mesh sizes. Define $x_j = jh$, $j \in \mathbb{Z}$ and $t^n = nk$, $n \geq 0$. Let us abbreviate $u(x_j, t^n)$ by u_j^n. We shall approximate u_j^n by U_j^n, where U_j^n satisfies some finite difference equations.

- Spatial discretization: The simplest one is to use the centered finite difference approximation for u_{xx}:

  $$u_{xx} = \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + O(h^2) := D_{x,+}D_{x,-}u + O(h^2).$$

  Here, the notation $(D_{x,+}u)_j := (u(x_{j+1}) - u(x_j))/h$ is the forward finite difference, $(D_{x,-}u)_j = (u(x_j) - u(x_{j-1}))/h$ the backward finite difference. You can check that

  $$D_{x,+}D_{x,-}u_j = a\left((u_{j+1} - u_j) - (u_j - u_{j-1})\right)/h^2.$$

  The spatial discretization results in the following systems of ODEs

  $$\dot{U}_j(t) = \frac{U_{j+1}(t) - 2U_j(t) + U_{j-1}(t)}{h^2}$$

  or in vector form

  $$\dot{U} = \frac{1}{h^2}AU$$

  where $U = (U_0, U_1, ...)^t$, $A = \text{diag}\,(1, -2, 1)$.

- Temporal discretization: We can apply numerical ODE solvers

  - Forward Euler method:

    $$U^{n+1} = U^n + \frac{k}{h^2}AU^n \tag{5.4}$$

  - Backward Euler method:

    $$U^{n+1} = U^n + \frac{k}{h^2}AU^{n+1} \tag{5.5}$$

  - 2nd order Runge-Kutta (RK2):

    $$U^{n+1} - U^n = \frac{k}{h^2}AU^{n+1/2},\ U^{n+1/2} = U^n + \frac{k}{2h^2}AU^n \tag{5.6}$$

  - Crank-Nicolson:

    $$U^{n+1} - U^n = \frac{k}{2h^2}(AU^{n+1} + AU^n). \tag{5.7}$$

These linear finite difference equations can be solved formally as

$$U^{n+1} = GU^n$$

where

- Forward Euler: $G = 1 + \frac{k}{h^2}A$,
- Backward Euler: $G = (1 - \frac{k}{h^2}A)^{-1}$,
- RK2: $G = 1 + \frac{k}{h^2}A + \frac{1}{2}\left(\frac{k}{h^2}\right)^2 A^2$
- Crank-Nicolson: $G = \frac{1+\frac{k}{2h^2}A}{1-\frac{k}{2h^2}A}$

For the Forward Euler, We may abbreviate it as

$$U_j^{n+1} = G(U_{j-1}^n, U_j^n, U_{j+1}^n), \tag{5.8}$$

where

$$G(U_{j-1}, U_j, U_{j+1}) = U_j + \frac{k}{h^2}(U_{j-1} - 2U_j + U_{j+1})$$

Stability and Convergence for the Forward Euler method Our goal is to show under what condition can U_j^n converges to $u(x_j, t^n)$ as the mesh sizes $h, k \to 0$. To see this, we first see the error produced by a true solution by the finite difference equation. Plug a true solution $u(x,t)$ into (5.4). We get

$$u_j^{n+1} - u_j^n = \frac{k}{h^2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right) + k\tau_j^n \tag{5.9}$$

where

$$\tau_j^n = D_{t,+}u_j^n - (u_t)_j^n - (D_+D_-u_j^n - (u_{xx})_j^n) = O(k) + O(h^2).$$

Let e_j^n denote for $u_j^n - U_j^n$. Then subtract (5.4) from (5.9), we get

$$e_j^{n+1} - e_j^n = \frac{k}{h^2}\left(e_{j+1}^n - 2e_j^n + e_{j-1}^n\right) + k\tau_j^n. \tag{5.10}$$

This can be expressed in operator form:

$$e^{n+1} = \mathbf{G}e^n + k\tau^n. \tag{5.11}$$

$$\begin{aligned}
\|e^n\| &\le \|\mathbf{G}e^{n-1}\| + k\|\tau^{n-1}\| \\
&\le \|\mathbf{G}^2e^{n-2}\| + k(\|\mathbf{G}\tau^{n-2}\| + \|\tau^{n-1}\|) \\
&\le \|\mathbf{G}^ne^0\| + k(\|\mathbf{G}^{n-1}\tau^0\| + \cdots + \|\mathbf{G}\tau^{n-2}\| + \|\tau^{n-1}\|)
\end{aligned}$$

Suppose $\mathbf{G}$ satisfies the *stability* condition

$$\|\mathbf{G}^nU\| \le C\|U\|$$

for some C independent of n. Then

$$\|e^n\| \le C\|e^0\| + C\max_m |\tau^m|.$$

If the local truncation error has the estimate

$$\max_m \|\tau^m\| = O(h^2) + O(k)$$

and the initial error e^0 satisfies

$$\|e^0\| = O(h^2),$$

then so does the global true error satisfies

$$\|e^n\| = O(h^2) + O(k) \text{ for all } n.$$

The above analysis leads to the following definitions.

Definition 5.3. *A finite difference method is called consistent if its local truncation error τ satisfies*

$$\|\tau_{h,k}\| \to 0 \textit{ as } h, k \to 0.$$

Definition 5.4. *A finite difference scheme $U^{n+1} = \mathbf{G}_{h,k}(U^n)$ is called stable under the norm $\|\cdot\|$ in a region $(h,k) \in R$ if*

$$\|\mathbf{G}^n_{h,k}U\| \le C\|U\|$$

for all n with nk fixed. Here, C is a constant independent of n.

Definition 5.5. *A finite difference method is called convergence if the true error*

$$\|e_{h,k}\| \to 0 \textit{ as } h, k \to 0.$$

In the above analysis, we have seen that for forward Euler method for the heat equation,

$$\text{stability } + \text{ consistency } \Rightarrow \text{ convergence.}$$

L^2 **Stability – von Neumann Analysis** Since we only deal with smooth solutions in this section, the L^2-norm or the Sobolev norm is a proper norm to our stability analysis. For constant coefficient and scalar case, the von Neumann analysis (via Fourier method) provides a necessary and sufficient condition for stability. For system with constant coefficients, the von Neumann analysis gives a necessary condition for statbility. For systems with variable coefficients, the Kreiss' matrix theorem provides characterizations of stability condition.

Below, we give L^2 stability analysis. We use two methods, one is the energy method, the other is the Fourier method, that is the von Neumann analysis. We describe the von Neumann analysis below.

Given $\{U_j\}_{j\in\mathbb{Z}}$, we define

$$\|U\|^2 = \sum_j |U_j|^2$$

and its Fourier transform

$$\hat{U}(\xi) = \frac{1}{2\pi} \sum U_j e^{-ij\xi}.$$

The advantages of Fourier method for analyzing finite difference scheme are

- the shift operator is transformed to a multiplier:

  $$\widehat{TU}(\xi) = e^{i\xi}\hat{U}(\xi),$$

  where $(TU)_j := U_{j+1}$;
- the Parseval equility

  $$\begin{aligned} \|U\|^2 &= \|\hat{U}\|^2 \\ &\equiv \int_{-\pi}^{\pi} |\hat{U}(\xi)|^2 \, d\xi. \end{aligned}$$

If a finite difference scheme is expressed as

$$U_j^{n+1} = (GU^n)_j = \sum_{i=-l}^{m} a_i (T^i U^n)_j,$$

then

$$\widehat{U^{n+1}}(\xi) = \widehat{G}(\xi)\widehat{U^n}(\xi).$$

From the Parseval equality,

$$\begin{aligned} \|U^{n+1}\|^2 &= \|\widehat{U^{n+1}}\|^2 \\ &= \int_{-\pi}^{\pi} |\widehat{G}(\xi)|^2 \, |\widehat{U^n}(\xi)|^2 \, d\xi \\ &\le \max_{\xi} |\widehat{G}(\xi)|^2 \int_{-\pi}^{\pi} |\widehat{U^n}(\xi)|^2 \, d\xi \\ &= |\widehat{G}|_\infty^2 \|U\|^2 \end{aligned}$$

Thus a sufficient condition for stability is

$$|\widehat{G}|_\infty \le 1. \tag{5.12}$$

Conversely, suppose $|\widehat{G}(\xi_0)| > 1$, from$\widehat{G}$ being a smooth function in ξ, we can find ϵ and δ such that

$$|\widehat{G}(\xi)| \ge 1 + \epsilon \text{ for all } |\xi - \xi_0| < \delta.$$

Let us choose an initial data U^0 in ℓ^2 such that $\widehat{U^0}(\xi) = 1$ for $|\xi - \xi_0| \le \delta$. Then

$$\begin{aligned} \|\widehat{U^n}\|^2 &= \int |\widehat{G}|^{2n}(\xi)|\widehat{U^0}|^2 \\ &\ge \int_{|\xi-\xi_0|\le\delta} |\widehat{G}|^{2n}(\xi)|\widehat{U^0}|^2 \\ &\ge (1+\epsilon)^{2n}\delta \to \infty \text{ as } n \to \infty \end{aligned}$$

Thus, the scheme can not be stable. We conclude the above discussion by the following theorem.

Theorem 5.10. *A finite difference scheme*

$$U_j^{n+1} = \sum_{k=-l}^{m} a_k U_{j+k}^n$$

with constant coefficients is stable if and only if

$$\widehat{G}(\xi) := \sum_{k=-l}^{m} a_k e^{-ik\xi}$$

satisfies

$$\max_{-\pi\le\xi\le\pi} |\widehat{G}(\xi)| \le 1. \tag{5.13}$$

Homeworks.

1. Compute the $\widehat{G}$ for the schemes: Forward Euler, Backward Euler, RK2 and Crank-Nicolson.

5.3.6 Weyl's ergodic theorem

In discrete dynamical systems, the dynamics of x^n is characterized by an iterative map $x^{n+1} = F(x^n)$. The simplest example is $x^n \in \mathbb{T}$ and $F(x^n) = x^n + 2\pi\gamma$. This is arisen from integrable system. It is the Poincare section map of the continuous integrable system:

$$(x, y) \mapsto (x(t), y(t)) := (x + 2\pi\omega_x t, y + 2\pi\omega_y t),\ (x, y) \in \mathbb{T}^2,\ \omega_x, \omega_y \in \mathbb{R}.$$

The cross section is taken to be $y \equiv 0(\mod 2\pi)$. Then the trajectory with $y^0 = 0$ will revisit $y = 0$ at time with $2\pi\omega_y t = 2\pi n$, that is, $t^n = n/\omega_y$. The corresponding $x(t^n) = x + 2\pi n\omega_x/\omega_y$. If we call $\gamma := \omega_x/\omega_y$, then the map: $x(t^n) \mapsto x(t^{n+1})$ is the above linear discrete map.

For a continuous function $f : \mathbb{T} \to \mathbb{C}$, we are interested in two kinds of averages:

- phase space average: $\langle f\rangle_s := \frac{1}{2\pi}\int_{\mathbb{T}f(x)\,dx}$,
- time average: $\langle f\rangle_t := \lim_{N\to\infty} \frac{1}{N+1}\sum_{n=0}^{N} f(x^n)$.

Theorem 5.11 (Weyl ergodic 1916). *If γ is irrational, then*

$$\langle f\rangle_t(x^0) = \langle f\rangle_s \tag{5.14}$$

for all $f \in C(\mathbb{T})$ and for all $x^0 \in \mathbb{T}$.

Proof. 1. It is easy to check that (5.14) holds for all $f = e^{imx}$:

$$\begin{aligned}\frac{1}{N+1} f(x^n) &= \frac{1}{N+1}\sum_{n=0}^{N} e^{im(x^0+2\pi n\gamma)}\\ &= \frac{e^{imx^0}}{N+1}\sum_{n=0}^{N} e^{2\pi i m\gamma n}\\ &= \frac{e^{imx^0}}{N+1}\left(\frac{1-e^{2\pi i m\gamma(N+1)}}{1-e^{2\pi i m\gamma}}\right).\end{aligned}$$

Thus,

$$\langle e^{imx}\rangle_t = \lim_{N\to\infty} \frac{1}{N+1}\left(\frac{1-e^{2\pi i m\gamma(N+1)}}{1-e^{2\pi i m\gamma}}\right) = 0.$$

Here, we have used $e^{2\pi i m\gamma} \neq 1$ for irrational γ. On the other hand,

$$\langle e^{imx}\rangle_s = \frac{1}{2\pi}\int_{\mathbb{T}} e^{imx}\,dx = 0.$$

Thus, $\langle e^{imx}\rangle_t = \langle e^{imx}\rangle_s$. With this, (5.14) also holds for all trigonometric polynomials.

2. The trigonometric polynomials are dense in $C(\mathbb{T})$.

3. Given any $f \in \mathbb{C}(\mathbb{T})$, for any $\epsilon > 0$, there exists a trigonometric polynomial p such that $\|f-p\|_\infty < \epsilon$. We have

$$\left|\frac{1}{N+1}\sum_{n=0}^{N} f(x^n) - \langle f\rangle_s\right| \le 2\epsilon + \left|\frac{1}{N+1}\sum_{n=0}^{N} p(x^n) - \langle p\rangle_s\right|$$

Taking limit sup, we get

$$\limsup_{N\to\infty}\left|\frac{1}{N+1}\sum_{n=0}^{N} f(x^n) - \langle f\rangle_s\right| \le 2\epsilon.$$

Thus, (5.14) also holds for any $f \in C(\mathbb{T})$.

□

Chapter 6

Compactness

6.1 Compactness in metric space

6.1.1 Motivation and brief history

[1] The concept of compactness plays an important role in analysis. There are many notions for compactness. Here, I will mention the most fundamental two. The first one states that: a set is called compact if any its open cover has finite subcover. It is motivated from a fundamental question in analysis: on which domain a local property can also be a global property. For instance, on which domain a continuous function is indeed uniform continuous. This concept (open cover) was introduced by Dirichlet in his 1862 lectures, which were published in 1904. This concept was repeatedly introduced by Heine (1872), Borel (1895) and continuously developed by Cousin (1895), Lebesgue (1898), Alexandroff and Uryson (1929). The main theorem is the Heine-Borel theorem which states that a set in $\mathbb{R}^n$ is compact if and only if it is closed and bounded.

The concept of the second notion is called sequential compactness, which was started from the Bolzano(1817)-Weistrass(1857) theorem. It states that a bounded sequence in $\mathbb{R}^n$ always has a convergence subsequence. It can be proven by bi-section method. The original Bolzano theorem was indeed a lemma to prove the extremal value theorem: a continuous function on a closed bounded interval always attains its extremal value. It was generalized to the function space $(C[a,b], |\cdot|_\infty)$ by Arzelà(1882-1883)-Ascoli(1883-1884) which states that a bounded and equi-continuous family of functions has convergence subsequence. Its generalization to $C(K)$ was done by Fréchet (1906) who also introduced the name of compactness.

Most theorems will not be proven here because it is hard to learn these abstract theorems without knowing their motivations. Thus, I will emphasize on the motivations and applications.

[1] Reference: You may find the history in [Compact Space, Wiki] and in the paper: [Manya Raman Sundström, A pedagogical history of compactness, ArXiv: 1006.4131v1 (2010).]

6.1.2 From local to global – Finite cover property

Many properties of functions are local such as the continuity, boundedness, local integrability, etc, while some are global such as convexity, coerciveness ($f(x) \to \infty$ as $|x| \to \infty$). It is clearly that the continuity is a local property. The following definition introduces the concept of uniform continuity, which is a global concept.

Definition 6.1. *We say $f : K \subset X \to Y$ is uniformly continuous on K if for any $\epsilon > 0$ there exists a $\delta > 0$ such that for any $x, y \in K$ with $d_X(x, y) < \delta$, we have $d_Y(f(x), f(y)) < \epsilon$.*

In this definition, the δ depends on ϵ and K, but is independent to any particular point in K. Clearly, the function $1/|x|$ is not uniformly continuous in $(0, 1]$ but uniformly continuous on $[a, 1]$ for any fixed $a > 0$. The key that a local property (such as continuity) can be a global (uniform) property on a set K depends on a particular property of K. It is called the compactness. To give its definition, first we give definition of open cover. Given a set K in (X, d), a collection of open sets $\{U_\alpha\}_{\alpha\in\mathcal{A}}$ is called an open cover of K if all U_α are open and

$$\bigcup_{\alpha\in\mathcal{A}} U_\alpha \supset K.$$

Definition 6.2. *A set K in a metric space (X, d) is said to be compact if any open cover $\{U_\alpha\}_{\alpha\in\mathcal{A}}$ of K has a finite sub-cover.*

Theorem 6.1 (Heine-Borel). *In $\mathbb{R}^n$, a set K is compact if and only if it is closed and bounded.*

I shall not prove this theorem here. You can find the proof in many textbooks. For instance, [Rudin's Principle of Mathematical Analysis, pp. 40]. Instead, I shall prove the following theorem which motivates the definition of Heine-Borel property..

Theorem 6.2. *Any continuous function on a compact set K is uniformly continuous on K.*

Proof. For any $\epsilon > 0$, for any $x \in K$, there exists $\delta_x(\epsilon) > 0$ such that $d(f(y), f(x)) < \epsilon$ whenever $d(x, y) < \delta_x$. Now, we choose $U_x = B(x, \delta_x/2)$. Then $\{U_x\}_{x\in K}$ is an open cover of K. By the Heine-Borel property, there exists a finite sub-cover. This means that there exists $x_1, ..., x_n \in K$ such that $\bigcup_{i=1}^n U_{x_i} \supset K$. We choose $\delta = \min_{i=1,...,n} \delta_{x_i}/2$. Now, for any $x, y \in K$ with $d(x, y) < \delta$, there exists a $k \in \{1, ..., n\}$ such that $x \in U_{x_k}$. This means that $d(x, x_k) < \delta_{x_k}/2$. Since $d(x, y) < \delta \le \delta_{x_k}/2$, we have $d(y, x_k) \le d(y, x) + d(x, x_k) \le \delta_{x_k}$. By the continuity of f at x_k, we then have

$$d(f(x), f(y)) \le d(f(x), f(x_k)) + d(f(x_k), f(y)) < 2\epsilon.$$

□

Remarks

1. The set $[0, \infty)$ is not compact. Thus, a continuous function $f : [0, \infty) \to \mathbb{R}$ is in general not uniformly continuous.

2. Is $\sin(x^2)$ uniformly continuous on $[0, \infty)$?
3. What is a natural condition you should impose on f so that it is uniformly continuous on $[0, \infty)$?
4. How about the same question for function on $(0, 1]$?

6.1.3 Sequential compactness

The other notion is the sequential compactness, which is arisen from finding extremal point of a continuous function. One often faces to check whether a sequence of approximate solutions converges or not. This motivates of the following definition.

Definition 6.3. *A set K in a metric space (X, d) is sequentially compact if every sequence in K has a convergent subsequence with limit in K. It is called pre-compact if its closure is sequentially compact.*

Theorem 6.3 (Bolzano-Weistrass). *A set in $\mathbb{R}^n$ is sequentially compact if and only if it is closed and bounded.*

Proof. I will only show that if an infinite sequence $\{x_n\} \subset [0, 1]$ then it has a convergent subsequence. The method below is called the bisection method. It is constructive and applicable for many other problems too. First, (x_n) is infinite many in $[0, 1]$. Let us partition $[0, 1]$ into $[0, 1/2]$, $[1/2, 1]$. One of them contains subsequence with infinite many elements. Let us call this subinterval $[a_1, b_1]$ and this subsequence (x_{1n}). We continuously perform partition and selection. Eventually, we get a nested intervals $[a_1, b_1] \supset [a_2, b_2] \supset \cdots$ with $b_k - a_k = 2^{-k}$ and subsequence $(x_{kn}) \subset [a_k, b_k]$ with infinite many elements. Since the nested intervals squeeze to zero length, we can just choose a point $y_k \in \{x_{kn}\}$ and $y_k \notin \{y_1, \cdots, y_{k-1}\}$, then (y_k) is a Cauchy sequence and converges to some point in $[0, 1]$ by the completeness of $\mathbb{R}$ and the closed-ness of $[0, 1]$. □

Theorem 6.4. *In metric space, a subset K is compact if and only if it is sequentially compact.*

For proof, see Hunter's book, pp. 24-27.

Remarks

1. In the theory of point set topology, the compactness implies the sequential compactness, but not vice versa. The sequential compactness is equivalent to so-called countable compactness. Its definition is: any countable open cover has finite cover. The compactness property (any open cover has finite cover) is a stronger property. It guarantees a local property becomes a global one. In the case of metric spaces, the compactness, the countable compactness and the sequential compactness are equivalent.
2. In general metric spaces, the boundedness is replaced by so-called total boundedness. A metric space (X, d) is totally bounded if $\forall \epsilon > 0$, $\exists x_1, ..., x_n$ such that $X = \cup_i B(x_i, \epsilon)$.
3. A metric space (X, d) is compact if and only if it is totally bounded and complete.
4. A totally bounded metric space is separable.

6.1.4 Applications in metric spaces

Compactness is preserved by continuous map

Theorem 6.5. *Let $f : (X, d_X) \to (Y, d_Y)$ be continuous. Let $K \subset X$ be a compact subset. Then $f(K)$ is also compact.*

Proof. If $(f(x_n))$ is a sequence in $f(K)$, then (x_n) is a sequence in K, which has a convergent subsequence (x_{n_k}) and converges to $x \in K$. By the continuity of f, we have $f(x_{n_k}) \to f(x) \in f(K)$. □

Theorem 6.6 (Extreme Value Theorem)**.** *Let $f : (K, d) \to \mathbb{R}$ be continuous, where (K, d) is a compact metric space. Then f attains both maximum and minimum in K.*

Proof. Since $f(K)$ is compact in $\mathbb{R}$, it is bounded and closed. Thus, it has both maximum and minimum. □

Homeworks 6.1.

1. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called coercive if

$$\lim_{|x|\to\infty} f(x) = \infty.$$

   Show that a function $f\mathbb{R}^n \to \mathbb{R}$ which is lower semi-continuous and coercive attains its infimum.

6.2 Compact sets in Banach spaces and Hilbert spaces

- The characterization of compact set in $C(K)$ is the Arzelà-Ascoli theorem. Here, K is a compact set in a metric space. It basically says that any bounded and equi-continuous sequence in $C(K)$ has convergent subsequence. One of its application is to prove existence theorem for ODEs $\dot{x} = f(t, x)$, where f is only continuous in x and L^1 in t.

- Characterization of compact set in $L^2(\Omega)$ is the Rellich theorem. It states that any bounded set in $H^1(\Omega)$ is pre-compact in $L^2(\Omega)$. The condition on Ω is *bounded* and $\partial\Omega$ is Lipschitz continuous.

Here, we recall that a set is pre-compact if its closure is compact.

6.2.1 Compact set in $C(K)$

We will show in this section that a family of functions $\mathcal{F} \subset C[a, b]$ is pre-compact if it is bounded in $C^1[a, b]$. The key part is the boundedness of f' for all $f \in \mathcal{F}$. This implies that there exists an L such that $|f(x) - f(y)| \le L|x - y|$ for all $f \in \mathcal{F}$. This condition can be weaken to the following equi-continuity.

Definition 6.4. *Let X, Y be metric spaces. A family of function $\mathcal{F} \subset C(X,Y)$ is said to be equicontinuous at x if for any $\epsilon > 0$ there exists a $\delta > 0$ such that for any $f \in \mathcal{F}$, we have $d(f(y), f(x)) < \epsilon$ whenever $d(x,y) < \delta$. The family $\mathcal{F}$ is said to be equi-continuous in X if it is equicontinuous at every $x \in X$. It is called uniformly equi-continuous if the above δ can be chosen independent of x.*

Theorem 6.7. *Let K be a compact metric space. If $\mathcal{F} \subset C(K)$ is equi-continuous, then it is uniformly equi-continuous.*

The proof is a simple modification of Theorem 6.2.

Theorem 6.8 (Arzelà-Ascoli)**.** *Let K be a compact metric space. A subset $\mathcal{F} \subset C(K)$ is pre-compact if and only if it is bounded and equi-continuous.*

Proof. $(\Rightarrow)$

1. We show that if $\mathcal{F}$ is unbounded, then it cannot be pre-compact. Since $\mathcal{F}$ is unbounded, we can find $f_n \in \mathcal{F}$ such that $\|f_{n+1}\|_\infty \geq \|f_n\|_\infty + 1$. Then $\|f_n\|_\infty \to \infty$. If $\{f_n\}$ has a convergent subsequence $\{f_{n_k}\}$, then $n_k \to \infty$ and $\|f_{n_k}\|_\infty$ is bounded. This is a contradiction.

2. We show that if $\mathcal{F}$ is precompact, then $\mathcal{F}$ is equicontinuous. For any $\epsilon > 0$,

$$\overline{\mathcal{F}} \subset \bigcup_{f \in \mathcal{F}} B_{\epsilon/3}(f).$$

From compactness of $\overline{\mathcal{F}}$, there exist $\{f_1, ..., f_n\}$ such that

$$\overline{\mathcal{F}} \subset \bigcup_{i=1}^{n} B_{\epsilon/3}(f_i).$$

Each f_i is uniformly continuous on K, thus there exist $\delta_i > 0$ such that

$$|f_i(x) - f_i(y)| < \epsilon/3 \text{ whenever } d(x,y) < \delta_i.$$

We choose

$$\delta = \min_{1 \leq i \leq n} \delta_i.$$

Then for every $f \in \mathcal{F}$, there exists f_i such that $\|f - f_i\|_\infty < \epsilon/3$. For any $d(x,y) < \delta$,

$$|f(x) - f(y)| \leq |f(x) - f_i(x)| + |f_i(x) - f_i(y)| + |f_i(y) - f(y)| < \epsilon.$$

Thus, $\mathcal{F}$ is equicontinuous.

$(\Leftarrow)$ We show that if $\mathcal{F}$ is bounded and equicontinuous, then it is pre-compact, i.e. any sequence $\{f_n\}$ in $\mathcal{F}$ has convergent subsequence.

1. First, from K being a compact metric space, then it is separable. Thus, there exist countable $\{x_i\}_{i\in\mathbb{N}}$ which is dense in K.

2. From boundedness of $\{f_n(x_1)\}$ in $\mathbb{R}$, there exists a subsequence from $\{f_n\}$, called $\{f_{1,n}\}$, such that $\{f_{1,n}(x_1)\}$ converges. Repeating this process, we can choose subsequence $\{f_{2,n}\}$ from $\{f_{1,n}\}$ such that $\{f_{2,n}(x_2)\}$ converges, and so on. Eventually, we obtain nested subsequence $\{f_{k,n}\}$ such that $\{f_{k,n}\} \subset \{f_{k-1,n}\}$ and $f_{k,n}(x_k)$ converges. In fact, $f_{k,n}(x_i)$ converges for all x_i with $i \le k$.

3. Let $g_k = f_{k,k}$. $\{g_k\}$ is a subsequence of $\{f_n\}$. Then for any $i \le k$, $\{g_k\}$ is a subsequence of $\{f_{i,n}\}$. Thus, $\{g_k(x_i)\}$ converges.

4. Since $\mathcal{F}$ is equicontinuous, there exists a $\delta > 0$ such that
$$|g_k(x) - g_k(y)| < \frac{\epsilon}{3} \text{ whenever } d(x,y) < \delta.$$

5. Since $\{x_i\}$ is dense in K, we have
$$K \subset \bigcup_{i=1}^{\infty} B_\delta(x_i).$$
From compactness of K, there exists a finite subset, call it $\{x_1, ..., x_p\}$ such that
$$K \subset \bigcup_{i=1}^{p} B_\delta(x_i).$$

6. Since $\{g_k(x_i)\}$ are Cauchy, there exists an N such that for any $i = 1, ..., p$ and for any $n, m \ge N$, we have
$$|g_n(x_i) - g_m(x_i)| < \epsilon/3.$$

7. Now, for any $x \in K$, there exists x_i such that $x \in B_\delta(x_i)$. We have for any $n, m \ge N$
$$|g_n(x) - g_m(x)| \le |g_n(x) - g_n(x_i)| + |g_n(x_i) - g_m(x_i)| + |g_m(x_i) - g_m(x)| < \epsilon.$$
Thus, $\{g_k\}$ converges uniformly in K.

□

6.2.2 Compact sets in Hilbert spaces

Roughly speaking, the compact set in a Hilbert space is almost like a bounded set in a finite dimensional subspace. We have the following abstract theorem, which says that the "tail" of a compact set has to be equally small.

Theorem 6.9. *Let $\mathcal{H}$ be a separable Hilbert space. Suppose $D \subset \mathcal{H}$ is a bounded set. If in addition, there is an orthonormal basis $\{e_n\}$ of $\mathcal{H}$ such that for any $\epsilon > 0$, there exists an N such that*

$$\sum_{n=N+1}^{\infty} |(x, e_n)|^2 < \epsilon \text{ for all } x \in D, \tag{6.1}$$

then D is precompact in $\mathcal{H}$.

Proof. Let (x_n) be a sequence in D. By the boundedness of D, the sequence (x_n) is bounded, say by 1. Let $V_n = \langle e_1, \cdots, e_n \rangle$ and P_n be the orthogonal projection onto V_n. We shall use diagonal process to construct a convergent subsequence of (x_n).

By assumption, given $\epsilon = 1/k$, there exists an N_k, which satisfies $N_k \geq N_{k-1}$ and

$$\sum_{N_k+1}^{\infty} |(x, e_n)|^2 \leq \frac{1}{k} \text{ for all } x \in D.$$

The sequence $(P_{N_1}(x_n))$ is bounded in the finite dimensional space V_{N_1}, by Hein-Borel theorem, it has a convergent subsequence $(P_{N_1} x_{1,n})$. In fact, we select $(x_{1,n})$ such that

$$\|P_{N_1}(x_{1,n} - x_{1,m})\|^2 \leq \frac{1}{n} \text{ for all } n < m.$$

Next, we consider $P_{N_2}(x_{1,n})$ in V_{N_2}. We can choose convergent subsequence $(P_{N_2} x_{2,n})$, $n \geq 2$ such that

$$\|P_{N_2}(x_{2,n} - x_{2,m})\|^2 \leq \frac{1}{n} \text{ for all } n < m.$$

Continuing this process, we construct $(x_{k,n})_{n \geq k}$, which is a subsequence of $(x_{k-1,n})_{n \geq k-1}$, such that

$$\|P_{N_k}(x_{k,n} - x_{k,m})\|^2 \leq \frac{1}{n} \text{ for all } n < m.$$

Now, we claim $x_{k,k}$ is a Cauchy sequence. For any large k, l with $l > k$, we have $x_{l,l} \in \{x_{k,n} | n \geq k\}$ and

$$\begin{aligned} \|x_{k,k} - x_{l,l}\|^2 &= \|P_{N_k}(x_{k,k} - x_{l,l})\|^2 + \|(I - P_{N_k})(x_{k,k} - x_{l,l})\|^2 \\ &\leq \|P_{N_k}(x_{k,k} - x_{l,l})\|^2 + \frac{2}{k} \\ &\leq \frac{3}{k}. \end{aligned}$$

This shows $(x_{k,k})$ is a Cauchy sequence in $\mathcal{H}$. Hence D is precompact. □

The Sobolev space $H^s(\mathbb{T})$, $s > 0$, is defined to be

$$\{u \in L^2(\mathbb{T}) | \sum_n (1 + |n|^2)^s |(u, e_n)|^2 < \infty\}$$

Here, $e_n = \frac{1}{\sqrt{2\pi}} e^{inx}$. The Sobolev norm of u is defined by

$$\|u\|_s^2 = \sum_n (1 + |n|^2)^s |(u, e_n)|^2$$

Theorem 6.10. *A bounded set D in $H^s(\mathbb{T})$ with $s > 0$ is a precompact set in $L^2(\mathbb{T})$.*

Proof. For, if $u \in D$, then

$$\sum_{N+1}^{\infty} |(u, e_n)|^2 \le \frac{1}{1+N^{2s}} \sum_{N+1}^{\infty} (1+|n|^2)^s |(u, e_n)|^2 \le \frac{1}{1+N^{2s}} \|u\|_s^2$$

Thus, D has uniformly small tails. Hence it is precompact in $L^2(\mathbb{T})$. □

Remarks

1. Consider the generalization of Aszelà-Ascoli theorem to continuous functions in the whole space $\mathbb{R}^d$. We can compactize $\mathbb{R}^n$ by $\mathbb{R}^n_* := \mathbb{R}^n \cup \{\infty\}$. The topology is generated by the union of the neighborhoods of ∞ and the topology of $\mathbb{R}^n$. The topological space $\mathbb{R}^n_*$ is compact. Indeed, for any open cover $\{U_\alpha | \alpha \in \mathcal{A}\} = \mathbb{R}^n_*$, there exists U_β which is open and covers ∞. This means that there exists an $M > 0$ such that the outer open ball $\overline{B_M(0)}^c := \{x | \, |x| > M\} \subset U_\beta$. On the other hand, $\{U_\alpha | \alpha \in \mathcal{A}\}$ is also an open cover of the compact set $\overline{B_M(0)}$. Thus, there exists a finite subcover $\alpha_1, ..., \alpha_n$. These together with U_β covers $\mathbb{R}^n_*$.

   Alternatively, one can identify $\mathbb{R}^n_*$ with S^n in $\mathbb{R}^{n+1}$ by the polar stereographic transform: $x \in \mathbb{R}^n_* \mapsto (x', z') \in S^n$ with

   $$\frac{|x'|}{|x|} = \frac{1-z'}{1}, \; |x'|^2 + z'^2 = 1.$$

   These give

   $$z' = \frac{|x|^2 - 1}{|x|^2 + 1}, \; x' = (1 - |z'|)x.$$

   The polar stereographic projection induces a natural metric on $\mathbb{R}^n_*$.

   With this, then we can define $C(\mathbb{R}^n_*)$ and the equi-continuity for $\mathcal{F} \subset C(\mathbb{R}^n_*)$ as usual. In fact, the equi-continuity at ∞ is: for any $\epsilon > 0$, there exists an $M > 0$ such that for any $f \in \mathcal{F}$ and for any $x \in \mathbb{R}^n$ with $|x| > M$, we have

   $$|f(x) - f(\infty)| < \epsilon.$$

2. Consider $L^2(\mathbb{R})$. Let $\alpha, \beta > 0$, $M > 0$. Define

   $$D = \{u \in L^2(\mathbb{R}) \,|\, \int (1+|x|^2)^\alpha |u(x)|^2 \, dx + \int (1+|\xi|^2)^\beta |\hat{u}(\xi)|^2 \, d\xi \le M\}$$

   Show that D is a precompact set in $L^2(\mathbb{R})$. Hint: use Haar basis.

3. In quantum mechanics, one considers the space $\mathcal{H} := H^1(\mathbb{R}^3) \cap L^2_V(\mathbb{R}^3)$, where $V : \mathbb{R}^3 \to \mathbb{R}^+$ satisfying

   $$V(x) \to \infty \text{ as } |x| \to \infty.$$

One can define the norm

$$\|u\|_{\mathcal{H}}^2 := \|u\|_{H^1(\mathbb{R}^3)}^2 + \|u\|_{L_V^2}^2 := \int_{\mathbb{R}^3} \left(|\nabla u|^2 + |u|^2 + V(x)|u|^2\right)\, dx.$$

Show that a bounded set in $\mathcal{H}$ is pre compact in $L^2(\mathbb{R}^3)$.

6.3 Weak Convergence

A sequence (x_n) in a Hilbert space $\mathcal{H}$ is said to converge weakly to $x \in \mathcal{H}$ if

$$(x_n, y) \to (x, y) \text{ for all } y \in \mathcal{H}.$$

We denote it by $x_n \rightharpoonup x$. It is clear that if $x_n \to x$ in $\mathcal{H}$, then, by Cauchy-Schwarz, $x_n \rightharpoonup x$. We call $x_n \to x$ in norm the strong convergence. Below, we give some typical examples of weak convergence.

1. $\sin nx \rightharpoonup 0$ in $L^2[a,b]$. This is due to the Riemann-Lebesgue lemma: If $f \in L^1(a,b)$, then $\int_a^b f(x) \sin nx\, dx \to 0$ as $n \to \infty$. The sequence $\{\sin nx\}$ has no cluster point:

$$\int_a^b |\sin(nx)\sin(mx)|^2\, dx \geq C(b-a).$$

2. In $\ell^2(\mathbb{N})$, let $e_n = (0, \cdots, 1, 0, \cdots)$ be the standard basis in $\ell^2(\mathbb{N})$. Then $e_n \rightharpoonup 0$. Further, $\|e_n - e_m\|^2 = 2$. Therefore, $\{e_n\}$ has no cluster point.

3. Let $\mathcal{H}$ be a Hilbert space and $\{e_n\}$ be an orthonormal set. Then $e_n \rightharpoonup 0$ as $n \to \infty$. This is due to the Bessel's inequality: for any x,

$$\sum_{n=1}^{\infty} |(x, e_n)|^2 \leq \|x\|^2,$$

Hence, $(x, e_n) \to 0$.

4. In $\ell^2(\mathbb{N})$, the unbounded sequence $\{ne_n\}$ does not converge to 0 weakly. For, choosing $x = \left(n^{-3/4}\right)_{n=1}^{\infty} \in \ell^2(\mathbb{N})$, but $(x, ne_n) = n^{1/4}$ does not not converge.

5. In $L^2[0,1]$, consider the Haar basis $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$, $0 \leq k < 2^j$, $j = 0, 1, \cdots$. Then $\psi_{jk} \rightharpoonup 0$, as $j \to \infty$.

6. In $L^2(\mathbb{R})$, the Haar basis $\psi_{jk} \rightharpoonup 0$ as $k \to \infty$. In particular, consider the box function $B_0(x) = 1$ for $0 \leq x < 1$ and 0, otherwise. Then $B_0(x-n) \rightharpoonup 0$ as $n \to \infty$.

7. In $L^2[0,1]$, consider

$$f_n(x) = \begin{cases} \sqrt{n} & \text{if } 0 \le x < 1/n \\ 0 & \text{if } 1/n \le x \le 1 \end{cases}$$

Then $f_n \rightharpoonup 0$. This is because for any $g \in L^2(0,1))$,

$$\begin{aligned} \int_0^1 g(x) f_n(x)\, dx &= \sqrt{n} \int_0^{1/n} 1 \cdot g(x)\, dx \\ &\le \sqrt{n} \left(\int_0^{1/n} 1^2\, dx \right)^{1/2} \left(\int_0^{1/n} |g(x)|^2 \Delta x \right)^{1/2} \\ &= \int_0^{1/n} |g(x)|^2\, dx \to 0. \end{aligned}$$

8. Let $f_n(x) := n^{1/4} e^{-nx^2}$. Then $f_n \rightharpoonup 0$ in $L^2(\mathbb{R})$.

Remarks.

(i) For functions in physical space such as $L^2[0,1]$ or $L^2(\mathbb{R})$, typical weak convergences are either oscillation, concentration, or escape to infinite.

(ii) When $\|x\|^2 < \liminf \|x_n\|^2$, we say that energy is lost in the weak convergence process.

Theorem 6.11. *Let $\mathcal{H}$ be a Hilbert space and $\mathcal{D}$ is a dense subset in $\mathcal{H}$. Then a sequence (x_n) in $\mathcal{H}$ converges to x weakly if and only if*

(a) $\|x_n\| \le M$ *for some constant* M*;*

(b) $(x_n, y) \to (x, y)$ *as* $n \to \infty$ *for all* $y \in \mathcal{D}$.

Proof. ($\Leftarrow$): Let $z \in \mathcal{H}$, we want to show that $(x_n, z) \to (x, z)$ as $n \to \infty$. Since $\mathcal{D}$ is dense in $\mathcal{H}$, for any $\epsilon > 0$, there exists $y \in \mathcal{D}$ such that $\|z - y\| < \epsilon$. On the other hand, $(x_n, y) \to (x, y)$. Thus, there exists an N such that for any $n \ge N$, $(x_n - x, y)| < \epsilon$. Now,

$$\begin{aligned} |(x_n - x, z)| &\le |(x_n - x, y)| + |(x_n - x, z - y)| \\ &\le |(x_n - x, y)| + (\|x_n\| + \|x\|)\|y - z\| \\ &\le \epsilon + (M + \|x\|)\epsilon. \end{aligned}$$

Thus, $(x_n - x, z) \to 0$ as $n \to \infty$.

($\Rightarrow$): (a) is followed from the uniform boundedness theorem below which basically says that pointwise boundedness of a family of bounded linear functional $\phi_n \in \mathcal{H}^*$ implies uniform boundedness. In the present case, define $\phi_n(x) := (x_n, x)$. Then for any $x \in \mathcal{H}$, $\phi_n(x)$ converges, thus is bounded. By the uniform boundedness principle, $\|\phi_n\|$, or equivalently $\|x_n\|$, are bounded. (b) is trivial. □

The following theorem is the Uniform Boundedness Principle, or the Banach-Steinhaus theorem.

Theorem 6.12 (Banach-Steinhaus)**.** *Let X be a Banach space and let $\mathcal{F} \subset X^*$. Then the point wise boundedness of $\mathcal{F}$ implies uniform boundedness of $\mathcal{F}$. This means that: if for any $x \in X$, there exists an M_x such that $\|\phi(x)\| \le M_x$ for all $\phi \in \mathcal{F}$, then there exists an M such that $\|\phi\| \le M$ for all $\phi \in \mathcal{F}$.*

Theorem 6.13. *If $x_n \rightharpoonup x$, then $\|x\| \le \liminf_{n\to\infty}\|x_n\|$. If, in addition, $\|x_n\| \to \|x\|$, then $x_n \to x$.*

Proof. The first result is obtained by

$$\|x\|^2 = \lim(x, x_n) \le \|x\| \liminf_{n\to\infty}\|x_n\|$$

The second result comes from

$$\|x_n - x\|^2 = \|x_n\|^2 - (x_n, x) - (x, x_n) + \|x\|^2.$$

□

Remarks

1. When $\|x\|^2 < \liminf \|x_n\|^2$, we say that there is an energy lost. For instance, those $e_n \rightharpoonup 0$ in $\ell^2(\mathbb{N})$ but $\|e_n\| = 1$. You can also see that the concentration examples. For instance, $f_n = n^{1/4}e^{-nx^2}$. Each of them has finite energy, but their weak limit is 0, which has zero energy.

Theorem 6.14 (Mazur)**.** *If $x_n \rightharpoonup x$ in a Hilbert $\mathcal{H}$, then there is a sequence $\{y_n\}$ which are finite convex combination of $\{x_n\}$ such that $y_n \to x$ (strongly).*

Proof. We may assume $x = 0$. We will choose a subsequence $(x_{n_k})_{k\ge1}$ of (x_n), then take average of them to produce the sequence (y_k). We start from $n_1 = 1$. From $(x_n, x_{n_1}) \to 0$, we can choose n_2 such that $|(x_{n_2}, x_{n_1})| \le 1/2$. Next, from $(x_n, x_{n_i}) \to 0$, $i = 1, 2$, we pick n_3 such that

$$|(x_{n_3}, x_{n_i})| \le \frac{1}{3} \text{ for } i = 1, 2.$$

Given $n_1, \cdots, n_k$, we pick n_{k+1} such that

$$|(x_{n_{k+1}}, x_{n_i})| \le \frac{1}{k+1}, \text{ for all } i = 1, \cdots, k.$$

Now, we define

$$y_k = \frac{1}{k}(x_{n_1} + \cdots + x_{n_k}).$$

Then

$$\begin{aligned} \|y_k\|^2 &= \frac{1}{k^2}\sum_{i=1}^{j-1}\|x_{n_i}\|^2 + \frac{2}{k^2} Re \sum_{j=1}^{k}\sum_{i=1}^{j-1}(x_{n_i}, x_{n_j}) \\ &\le \frac{M^2}{k} + \frac{2}{k^2}\sum_{j=1}^{k}\sum_{i=1}^{j-1}\frac{1}{j} \le \frac{M^2+2}{k}. \end{aligned}$$

Thus, $y_k \to 0$. □

Examples

1. We have seen that $e^{-inx} \rightharpoonup 0$ from Riemann-Lebesgue lemma. Define

$$F_N(x) = \frac{1}{2N+1} \sum_{n=-N}^{N} e^{inx}.$$

Then $F_N \to 0$ strongly in $L^2(\mathbb{T})$.

2. The sequence e_n converges to 0 weakly in $\ell^2(\mathbb{N})$. Their averages $y_n = \frac{1}{n}(1, 1, \cdots, 1, 0, 0, \cdots)$ have norm $\|y_n\| = 1/\sqrt{n}$ which converges to 0 strongly.

3. Let $f_n \in L^2(0,1)$ given by

$$f_n(x) = \begin{cases} \sqrt{n} & \text{if } 0 \le x < 1/n \\ 0 & \text{if } 1/n \le x \le 1 \end{cases}$$

Let $g_n(x) = \frac{1}{n}\left(f_1 + \cdots + f_n\right)$. Is $g_n \to 0$ strongly in $L^2(0,1)$?

4. Let $f_n(x) = n^{1/4} e^{-nx^2}$. Then $f_n \rightharpoonup 0$ in $L^2(\mathbb{R})$. What is $g_n = (f_1 + \cdots + f_n)/n$? Does it converge to 0 strongly?

Remark If H is a convex function in $\mathbb{R}^n$ and if $\nabla u_n \rightharpoonup \nabla u$ in $L^2(\Omega)$, then

$$\liminf_{n\to\infty} \int_\Omega H(\nabla u_n)\, dx \ge \int_\Omega H(\nabla u)\, dx.$$

Theorem 6.15 (Banach-Alaoglu). *A closed bounded set in a Hilbert space $\mathcal{H}$ is weakly compact.*

Proof. We shall prove the theorem for separable Hilbert spaces. We show that if (x_n) is bounded, then it has a weakly convergent subsequence. We use diagonal process. Suppose $\mathcal{D} = \{y_n | n \in \mathbb{N}\}$ is dense in $\mathcal{H}$. Then (x_n, y_1) is bounded in $\mathbb{C}$. By Hein-Borel theorem, there exists a subsequence $(x_{1,k})$ such that $(x_{1,k}, y_1)$ converges. We repeat this process, there exists a subsequence $(x_{2,k})$ of $(x_{1,k})$ such that $(x_{2,k}, y_2)$ converges. Continuing in this way, we obtain $(x_{m,k})_{k\ge 0}$ such that $(x_{m,k}, y_i), k \ge 0$ converges for all $1 \le i \le m$. Taking the diagonal subsequence $(x_{k,k})$, we claim $(x_{k,k}, y)$ converges for all $y \in \mathcal{H}$. If $y \in \mathcal{D}$, then $y = y_n$ for some n. Taking $k > n$, we have $(x_{k,k}, y_n)$ converges as $k \to \infty$. Thus, $(x_{k,k}, y)$ converges for any $y \in \mathcal{D}$. This limit defines a linear functional ℓ on $\mathcal{D}$ by

$$\ell(y) = \lim_{k\to\infty} (x_{k,k}, y).$$

From (x_n) being bounded , say by M, by our assumption, we have $\|x_{k,k}\| \le M$. Hence, ℓ is bounded, thus continuous, on $\mathcal{D}$ and has unique extension to $\mathcal{H}$ with norm bounded by M. By the

Riesz representation theorem, there exists an $x \in \mathcal{H}$ such that $\ell(y) = (x, y)$. Now, for any $y \in \mathcal{H}$, we choose $y' \in \mathcal{D}$ such that $y' \to y$. Then

$$\begin{aligned}
\lim_{k\to\infty} (x_{k,k}, y) &= \lim_{k\to\infty} \lim_{y'\to y} (x_{k,k}, y') \\
&= \lim_{y'\to y} \lim_{k\to\infty} (x_{k,k}, y') \\
&= \lim_{y'\to y} (x, y') \\
&= (x, y).
\end{aligned}$$

That the interchange of two limits is allowed is due to the uniform boundedness of $(x_{k,k})$. □

Remark Let $\Omega \subset \mathbb{R}^d$ be a smooth bounded domain. If u_n is bounded in $H^1(\Omega)$, then u_n has a convergent subsequence u_{n_k} which converges to u strongly in L^2. Moreover, $u \in H^1$ and $\nabla u_{n_k} \rightharpoonup \nabla u$.

6.4 Direct method in Calculus of Variations

We introduce the direct method in calculus of variations as an application. For simplicity, we consider quadratic minimization problem.

6.4.1 Dirichlet problem

Let us consider the following variational formulation of the Dirichlet problem: Find $u \in H_0^1(\Omega)$ such that it minimizes the Dirichlet integral

$$J[u] := \int_\Omega \left(\frac{1}{2} |\nabla u(x)|^2 - f(x)u(x) \right) dx,$$

where $f \in L^2(\Omega)$.

1. In $H_0^1(\Omega)$, we use $\|u\|_{H^1} := \|\nabla u\|_{L^2}$. From the Poincarè inequality, there exists a constant C such that for any $u \in H_0^1$, $\|u\|_{L^2} \le C\|\nabla u\|_{L^2}$. Hence, the usual H^1 norm defined by $\|u\|_{H^1} := \|u\|_{L^2} + \|\nabla u\|_{L^2}$ is equivalent to $\|\nabla u\|_{L^2}$.

2. We show that $J[u]$ is coercive, i.e. there exists an $\alpha, \beta > 0$ such that $J[u] \ge \alpha\|\nabla u\|^2 - \beta$. This is mainly due to Poincarè inequality again. We have

$$\begin{aligned}
|(f, u)| &\le \|f\| \cdot \|u\| \\
&\le \|f\| C \|\nabla u\| \\
&\le \epsilon \|\nabla u\|^2 + \frac{C^2}{\epsilon} \|f\|^2
\end{aligned}$$

and

$$\begin{aligned}
J[u] &\ge \frac{1}{2} \|\nabla u\|^2 - \|f\| \cdot \|u\| \\
&\ge \left(\frac{1}{2} - \epsilon \right) \|\nabla u\|^2 - \frac{C^2}{\epsilon} \|f\|^2
\end{aligned}$$

By choosing $\epsilon > 0$ small enough, we have $J[u] \geq \alpha\|\nabla u\|^2 - \beta$.

3. From the coerciveness of $J[u]$ in H_0^1, we see that $J[u]$ is bounded from below. Hence
$$\inf\{J[u]|u \in H_0^1\} = m > -\infty$$
exists. Now, we choose a sequence $u_n \in H_0^1$ such that
$$\lim_{n\to\infty} J[u_n] = m.$$
Such a sequence is called a minimal sequence. From the coerciveness of $J[u]$, we obtain
$$\alpha\|\nabla u_n\|^2 \leq J[u_n] + \beta$$
for all n. Thus, $\|\nabla u_n\|$ is bounded. Hence, the sequence is bounded in $H_0^1(\Omega)$.

4. By the compact embedding of $H_0^1(\Omega)$ in $L^2(\Omega)$ and the Banach-Alaoglu theorem, we can choose a subsequence of (u_n), still denoted by (u_n), such that $u_n \to u$ in $L^2(\Omega)$ and $u_n \rightharpoonup u$ weakly in $H_0^1(\Omega)$.

5. From $u_n \rightharpoonup u$ in H_0^1, we get
$$\|\nabla u\|^2 \leq \liminf_{n\to\infty} \|\nabla u_n\|^2.$$
From $u_n \to u$ in L^2, we get $(f, u_n) \to (f, u)$. Combining these two together, we get
$$J[u] \leq \liminf_{n\to\infty} J[u_n] = m$$
Thus, the minimum is attained.

□

6.4.2 Phase field model for binary systems

Consider a binary system, composed of two species A and B. Let u_A, u_B be their concentrations. A phenomenological theory describing the equilibrium state is through a coarse grained Gibb's free energy $\psi(u, T)$, which is defined to be
$$\psi(u_A, u_B, T) = \alpha u_A u_B + T(S_A + S_B),$$
where T is the temperature and S_A and S_B are the entropy. Usually, $S_i = u_i \log u_i$ for $i = A, B$. The term $\alpha u_A u_B$ represents the interaction energy between species A and B. They are repulsive when $\alpha > 0$ and attractive when $\alpha < 0$. Notice that
$$u_A + u_B = 1$$
from conservation of volume. Thus, we can use $u = u_A$ only and the Gibb's free energy is
$$\psi(u, T) = \alpha u(1-u) + T\left(u \log u + (1-u)\log(1-u)\right), 0 \leq u \leq 1.$$

For the repulsive case, the interaction energy is concave downward, while the entropy term is concave upward. Therefore, there exists a critical temperature T_c, such that for $T > T_c$, the function $\psi(u, T)$ is strictly convex in u and has a unique minimum. While for $T < T_c$, $\psi(u, T)$ has two local minima and one maximum (i.e. it is a double well). The equilibrium is obtained by

$$\min \int_\Omega \psi(u(x), T)\, dx \text{ subject to } \int_\Omega u(x)\, dx = u_m|\Omega|.$$

Suppose the concentration of A is uniform, then there exist a unique minimum for $T > T_c$. However, there are two minima u_a and u_b for $T < T_c$ with $\psi'(u_a) = \psi'(u_b) = 0$. Let $\Omega_a \subset \Omega$ and $\Omega_b = \Omega \setminus \Omega_a$. Ω_a is chosen to satisfy

$$u_a|\Omega_a| + u_b|\Omega_b| = u_m.$$

We define

$$u(x) = \begin{cases} u_a, & x \in \Omega_a \\ u_b, & x \in \Omega_b \end{cases}$$

Then u is a minimum. Thus, this problem is ill-posed. The solution is not unique. Indeed, it is quite flexible to choose the subdomain Ω_a. Cahn-Hilliard modify the free energy by adding a gradient term $\gamma|\nabla u|^2$ where $\gamma > 0$ so that the free energy becomes

$$\Psi(u, T) = \psi(u, T) + \frac{\gamma}{2}|\nabla u|^2.$$

The equilibrium state is

$$\min_{u\in\mathcal{A}} \mathcal{E}(u) \text{ subject to } \int_\Omega u(x)\, dx = u_m|\Omega|.$$

where

$$\mathcal{E}(u) = \int_\Omega \left(\psi(u(x)) + \frac{\gamma}{2}|\nabla u|^2\right)\, dx$$

The function class $\mathcal{A}$ is called the admissible class. It should be chosen so that

(a) Boundary condition is satisfied. In our case, it is $u_n(x) = 0$, for $x \in \partial\Omega$;

(b) The energy $\mathcal{E}$ is finite. This requires $u \in H^1(\Omega)$ and $0 \le u \le 1$.

In this case, one can apply the direct method in Calculus of Variations to show the existence and the phase transition phenomena.

Homeworks 6.2. *1. For general convex function, which is coercive, strongly lower semicontinuous (l.s.c), the minimum is also attained. Hint: From Mazur theorem, strongly l.s.c plus convexity of J implies J is also weakly l.s.c. The coerciveness of J implies bounded set is weakly precompact. Combining these two together gives the existence of minimum.*

2. pp. 214: Ex. 8.22

Chapter 7

Bounded Linear Operators in a Hilbert Space

7.1 Examples of bounded operators

We have seen many examples of bounded and unbounded operators in Chapter 2. We will review some and provide more examples.

1. The differential operator: $D : u \to u'$ is a unbounded operator, while its inverse:
$$Ku(x) = \int_0^x u(y)\,dy = \int_0^1 g(x,y)u(y)\,dy$$
is a bounded operator from $C[0,1]$ to $C[0,1]$. Here, $g(x,y) = 1$ if $0 \le y < x$ and 0 otherwise. You can see that it is also a bounded map from $L^2(0,1)$ to itself.

2. Consider $D^2 u = f$ on $(0,1)$ with the boundary condition: $u(0) = u(1) = 0$. Its inversion can be represented as
$$u(x) = \int_0^1 g(x,y)f(y)\,dy, \quad g(x,y) = \left\{ \begin{array}{ll} x(1-y) & \text{for } 0 < x < y < 1 \\ y(1-x) & \text{for } 0 < y < x < 1 \end{array} \right.$$
The mapping $f \mapsto u$ is a bounded linear operator from $L^2(0,1)$ to itself (Check by yourself).

3. Consider $- \triangle u = f$ in $\mathbb{R}^3$ with $u(x) \to 0$ as $|x| \to \infty$. Then u is given by
$$u(x) = \int \frac{1}{4\pi} \frac{1}{|x-y|} f(y)\,dy$$
The mapping $f \mapsto u$ is a bounded linear map from $L^2(\mathbb{R}^3)$ to $L^2(\mathbb{R}^3)$.

4. Blur operator
$$Kf(x) = \int K(x,y)f(y)\,dy.$$

Here $K(x, y)$ is called the blur kernel. If $K(\cdot, \cdot)$ is bounded, then K is a bounded operator from L^2 to itself. In many cases, K is translational invariant, i.e. $K(x, y) = K(x - y)$. In this case, we can represent it as a multiplier in the Fourier space:

$$\widehat{Kf}(\xi) = \widehat{K}(\xi)\hat{f}(\xi).$$

5. Potential Theory: Consider the potential equation

$$\triangle u = 0 \text{ in } \Omega \subset \mathbb{R}^2,$$

Let $K(x - y) = -\frac{1}{2\pi} \ln |x - y|$ be the fundamental solution of $-\triangle$ in $\mathbb{R}^2$. That is,

$$- \triangle_y K(x - y) = \delta(x - y).$$

Now, we multiply $- \triangle_y u = 0$ by K, integrate it over a domain $\Omega_{x,\epsilon}$ in y, and perform integration by part, then take $\epsilon \to 0$. Here, $\Omega_{x,\epsilon}$ is $\Omega \setminus B_\epsilon(x)$ and $\epsilon << 1$ so that $B_\epsilon(x) \subset \Omega$ if $x \in \Omega$. Then we will get

$$\int_{\partial\Omega} \left(K(x - y)u_n(y) - \frac{\partial K(x - y)}{\partial n_y} u(y) \right) dy = u(x). \tag{7.1}$$

When $x \in \partial\Omega$, there is only half of the ball $B_\epsilon(x)$ lies inside Ω and we get

$$\int_{\partial\Omega} \left(K(x - y)u_n(y) - \frac{\partial K(x - y)}{\partial n_y} u(y) \right) dy = \frac{u(x)}{2}. \tag{7.2}$$

Here, the data $u(x)$ and $u_n(x)$ for $x \in \partial\Omega$ are the limits from *inside* of Ω. The formula (7.1) provides us a representation of $u(x)$ for $x \in \Omega^o$ in terms of the boundary data $u(y)$ and $u_n(y)$, $y \in \partial\Omega$. On the other hand, (7.2) gives a relation between the Dirichlet data $u(x)$ and Neumann data $u_n(x)$ on the boundary. For instance, suppose we are given a Dirichlet data $u(x) = f(x)$ for $x \in \partial\Omega$. Then we can solve u_n from

$$\int_{\partial\Omega} K(x - y)u_n(y)\, dy = \int_{\partial\Omega} \frac{\partial K(x - y)}{\partial n_y} f(y)\, dy + \frac{f(x)}{2}.$$

This is called the Fredholm integral equation of first kind. On the other hand, if we are given Neumann data $u_n = g$ on the boundary, then we can find u on the boundary from

$$\frac{u(x)}{2} + \int_{\partial\Omega} \frac{\partial K(x - y)}{\partial n_y} u(y)\, dy = \int_{\partial\Omega} K(x - y)g(y)\, dy.$$

This is called the Fredholm integral equation of the second kind. The mapping: $u \mapsto u_n$ on the boundary is a linear map, called the Dirichlet-to-Neumann map. Similarly, its inverse map is a linear map and will be shown that it is a bounded linear map. We will see that this linear map is a bounded operator from $L^2(\partial\Omega)$ to itself.

6. Radon transform. Given a smooth, compact supported function f in $\mathbb{R}^2$ with support in $B_1(0)$, given any $\theta \in S^2$ and $s \in \mathbb{R}$, define

$$(Rf)(\theta, s) = \int_{\theta^\perp} f(s\theta + y)\, dy$$

where $\theta^\perp = \{y|(\theta, y) = 0\}$. The transformation R is called the Radon transform. We will see that it is a bounded operator from $L^2(B_1)$ to $L^2(B_1)$, where B_1 is the unit disk on $\mathbb{R}^2$.

7. Hilbert transform:

$$(Hf)(x) := P.V. \frac{1}{\pi} \int_{\mathbb{R}} \frac{f(y)}{x - y}\, dy := \lim_{\epsilon \to 0+} \int_{\mathbb{R} \setminus (x-\epsilon, x+\epsilon)} \frac{f(y)}{x - y}\, dy.$$

The Fourier representation of Hilbert transform is

$$\widehat{Hf}(\xi) = -\text{sign}(\xi) \hat{f}(\xi).$$

Let u be the the harmonic function on the upper half plane with Dirichlet data f on $y = 0$. Let v be the complex conjugate of u. Then $v(x, 0+)$ is the Hilbert transform of f.

Issues we are interested The issues we are concerned are:

- Soving $Au = f$ for existence, uniqueness and stability.
- Solving the least square problem: Find $\min \|Au - f\|^2$.
- Eigen-expansion of A.

7.2 Preliminaries

Operator norm . Let $\mathcal{H}$ and $\mathcal{K}$ be Banach spaces. Let $A : \mathcal{H} \to \mathcal{K}$ be a linear operator. We recall that a linear operator A is called bounded if there exists an M such that $\|Ax\| \le M\|x\|$ for all $x \in \mathcal{H}$. It is easy to see that a linear operator A is bounded if and only if it is continuous. For a bounded operator A, we define its operator norm $\|A\|$ by

$$\|A\| := \sup_{\|x\| \neq 0} \frac{\|Ax\|}{\|x\|}.$$

Kernel and Range We denote the kernel of A by $N(A)$ and range by $R(A)$. From the boundedness of A, we get $N(A)$ is closed. The range of a bounded linear map may not be closed. For example, the mapping $Kf = \int_0^x f(y)\, dy$ maps $C^0[0,1]$ to $C^0[0,1]$. But the range is $C^1[0,1]$ with $u(0) = 0$, which is not closed in $C^0[0,1]$ (prove it).

Proposition 1. *Let A be a bounded linear map from Banach spaces $\mathcal{H}$ to $\mathcal{K}$. The following two statements are equivalent:*

(i) there exists a constant $c > 0$ such that $\|Ax\| \ge c\|x\|$ for all $x \in \mathcal{H}$;

(ii) $R(A)$ is closed, and $N(A) = \{0\}$.

Proof. (i) ⇒ (ii). If $\{Ax_n\}$ is a Cauchy sequence in $R(A)$, then from the assumption $\|Ax\| \ge c\|x\|$, we get that $\{x_n\}$ is also a Cauchy sequence and it converges to $x \in \mathcal{H}$. The continuity of A implies $Ax_n \to Ax$. Thus, $R(A)$ is closed. Also, from $\|Ax\| \ge c\|x\|$, we get that $Ax = 0$ implies $x = 0$. Thus, $N(A) = \{0\}$.
We prove (ii) ⇒ (i). From (ii), $R(A)$ is a Banach space and $A : \mathcal{H} \to R(A)$ is 1-1 and onto. Hence A^{-1} exists from a Banach space $R(A)$ to $\mathcal{H}$. The open mapping theorem states that a bounded linear map from a Banach space onto another Banach space maps open sets to open sets. Thus, we get A^{-1} is continuous, hence it is bounded. That is, there exists a constant c_1 such that any $y \in R(A)$, $\|A^{-1}y\| \le c_1\|y\|$. Or, equivalently, $c_1\|Ax\| \ge \|x\|$ for any $x \in \mathcal{H}$. □

Example Fredholm operators are typical example of bounded operators that have closed range. Let us consider the following concrete example. In the space $C[0,1]$, we consider $Ku = \int_0^x u(y)\,dy$. Then $Au = 0$ implies $u(x) + \int_0^x u(y)\,dy = 0$. Differentiate it in x, we obtain $u' + u = 0$. This leads to $u(x) = Ce^{-x}$. Thus, $N(A) = \{Ce^{-x} | C \in \mathbb{R}\}$. Notice that if we restrict to the space $\mathcal{C}_0 := \{u \in C[0,1] | u(0) = 0\}$, then $N(A) = \{0\}$. Thus, $A : \mathcal{C} \to C[0,1]$ has zero kernel. Next, for any $f \in C[0,1]$, we look for a solution $u \in C[0,1]$ such that $Au = f$. Formally, we differentiate $Au = f$ and get

$$u' + u = f'.$$

By using integration factor, we get

$$(e^y u)' = e^y f'.$$

Integrate this equation, we get

$$e^x u(x) - u(0) = \int_0^x e^y f'(y)\,dy = e^x f(x) - f(0) - \int_0^x e^y f(y)\,dy.$$

Thus,

$$u(x) = e^{-x}(u(0) - f(0)) + f(x) - \int_0^x e^{-x+y} f(y)\,dy.$$

In this expression, we don't need to require f' exists. Thus, $R(A) = C[0,1]$. We conclude that $A : \mathcal{C} \to C[0,1]$ is closed range with zero kernel.

The set of bounded operators . The set

$$B(\mathcal{H},\mathcal{K}) = \{A : \mathcal{H} \to \mathcal{K} \text{ is a bounded linear operator}\}$$

forms a normed linear space with the above operator norm. It is indeed a Banach space from the completeness of the space $\mathcal{K}$. For, if $\{A_n\}$ is Cauchy in operator norm, then for any $x \in \mathcal{H}$, it holds that $A_n x$ is also a Cauchy in $\mathcal{K}$ and hence converges to a unique point y in $\mathcal{K}$. Thus, we can define $Ax = y$. We leave the rest of the proof to the readers.

Dual space. Let $\mathcal{H}$ be a Hilbert space over $\mathbb{C}$. We recall that the set $B(\mathcal{H}, \mathbb{C})$ is called the dual space of $\mathcal{H}$ and is denoted by $\mathcal{H}^*$. The *Riesz representation theorem* says that $\mathcal{H}^*$ is isometric to $\mathcal{H}$. That is, for any $\ell \in B(\mathcal{H}, \mathbb{C})$, there exists a unique $y \in \mathcal{H}$ such that $\ell(x) = (y, x)$ for all $x \in \mathcal{H}$.

Adjoint operator . Let $A : \mathcal{H} \to \mathcal{K}$ be a bounded operator. We can define the adjoint operator $A^* : \mathcal{K}^* \to \mathcal{H}^*$ by

$$(A^*y^*)(x) = y^*(Ax) \text{ for all } y^* \in \mathcal{K}^*, x \in \mathcal{H}.$$

By the Riesz representation theorem, we can identify each $y^* \in \mathcal{K}^*$ as a point $y \in \mathcal{K}$ by $y^*(z) = (y, z)$ for all $z \in \mathcal{K}$. Thus, in terms of the inner products in $\mathcal{H}$ and $\mathcal{K}$, the adjoint operator $A^* : \mathcal{K} \to \mathcal{H}$ is defined as

$$(A^*y, x) = (y, Ax) \text{ for all } x \in \mathcal{H}, y \in \mathcal{K}.$$

The adjoint operator A^* is well-defined for the following reason. The linear map $\ell(x) = (y, Ax)$ is bounded. By Riesz representation theorem, there exists a unique $z \in \mathcal{H}$ such that $(y, Ax) = (z, x)$ for all $x \in \mathcal{H}$. We then define $A^*y = z$.

One can show that

1. $(A^*)^* = A$
2. $(AB)^* = B^*A^*$.
3. If $A : \mathcal{H} \to \mathcal{K}$ is a bounded operator, then $A^* : \mathcal{K} \to \mathcal{H}$ is also a bounded operator and $\|A^*\| = \|A\|$:

$$\sup_{\|y\|=1} \|A^*y\| = \sup_{\|y\|=1} \sup_{\|x\|=1} (A^*y, x) = \sup_{\|x\|=1} \sup_{\|y\|=1} (y, Ax) = \sup_{\|x\|=1} \|Ax\| = \|A\|.$$

Examples.

1. Let $A : \mathbb{C}^n \to \mathbb{C}^m$ be a linear map with matrix representation $A = (a_{ij})_{m\times n}$. Then its dual operator A^* has the matrix representation $A^* = (\bar{a}_{ji})_{n\times m}$.
2. Let $K : L^2[0, 1] \to L^2[0, 1]$ be
$$Kf(x) = \int_0^1 k(x, y) f(y)\, dy$$
Then
$$K^*f(x) = \int_0^1 \overline{k(y, x)} f(y)\, dy$$
3. Let S be the shift operator in $\ell^2(\mathbb{N})$ defined by
$$S(x_1, x_2, \cdots) = (0, x_1, x_2, \cdots).$$
Then
$$S^*(x_1, x_2, \cdots) = (x_2, x_3, \cdots).$$

If $A : \mathcal{H} \to \mathcal{H}$ satisfies $(Ax, y) = (x, Ay)$, i,e, $A = A^*$, we call it self-adjoint.

7.3 Solving $Ax = b$ and its least-squares solution

The solvability of $Ax = b$ relies heavily on to property of its dual A^*. We have the following theorem.

Theorem 7.1. *Let $A : \mathcal{H} \to \mathcal{K}$ be a bounded linear map. It holds*

(i) $N(A^*) = R(A)^\perp$, $\overline{R(A)} = N(A^*)^\perp$;

(ii) $N(A) = R(A^*)^\perp$, $\overline{R(A^*)} = N(A)^\perp$;

(iii) $A : \mathcal{H} = N(A) \oplus \overline{R(A^*)} \longrightarrow N(A^*) \oplus \overline{R(A)}$, *with* $A : \overline{R(A^*)} \to R(A)$ *being one-to-one and onto.*

Proof. (i) First, we show $N(A^*) \subset R(A)^\perp$. Suppose $y \in N(A^*)$. For any $x \in \mathcal{H}$, we have $(Ax, y) = (x, A^*y) = 0$. Thus, $y \perp R(A)$, or equivalently, $y \in R(A)^\perp$.
Next, we show $R(A)^\perp \subset N(A^*)$. If $z \in R(A)^\perp$, it means that $(z, Ax) = 0$ for all $x \in \mathcal{H}$. This implies $(A^*z, x) = 0$ for all $x \in \mathcal{H}$. Thus, $A^*z = 0$. Hence, $R(A)^\perp \subset N(A^*)$. By taking orthogonal complement, we get

$$N(A^*)^\perp \subset R(A)^{\perp\perp} = \overline{R(A)}.$$

This completes the proof of (i).

(ii) We apply (i) to A^* to get $N(A^{**})^\perp = \overline{R(A^*)}$. Taking orthogonal complement and using $A^{**} = A$, we get $N(A) = \overline{R(A^*)}^\perp = R(A^*)^\perp$.

(iii) Since A is bounded, we get that $N(A)$ is a closed subspace of $\mathcal{H}$. By the orthogonal projection theorem, $\mathcal{H} = N(A) \oplus N(A)^\perp$. Similarly, A^* is also bounded. Thus, $\mathcal{K} = N(A^*) \oplus N(A^*)^\perp$. From (i) and (ii), we have $N(A)^\perp = \overline{R(A^*)}$ and $N(A^*)^\perp = \overline{R(A)}$. If $x \in \overline{R(A^*)}$ with $Ax = 0$, then $x \in N(A) \cap N(A)^\perp = \{0\}$. Hence A is 1-1 on $\overline{R(A^*)}$. The onto part for $A : \overline{R(A^*)} \to R(A)$ is trivial.

□

Remarks.

1. The necessary condition for the solvability of $Ax = b$ is $b \perp N(A^*)$. Usually, it is easier to solve the homogeneous equation such as $A^*y = 0$. If $R(A)$ is also closed, then this is also a sufficient condition. In finite dimensions, $R(A)$ is always closed. So in finite dimension case, solvability of $Ax = b$ if and only if $b \perp N(A^*)$.

2. Another application example is $A = I - K$, where K is a compact operator. In this case, $R(A)$ is also closed. We will prove this later. So, $b \perp N(A^*)$ is a necessary and sufficient condition for the solvability of $Ax = b$.

3. In the case of $R(A)$ being closed, in particular, $N(A^*) = \{0\}$ if and only if $Ax = b$ is solvable. Thus, the existence of $Ax = b$ is equivalent to the uniqueness of $A^*y = 0$.

4. In general, $R(A)$ may not be closed in the infinite dimensional case. In this case, $b \perp N(A^*)$ does not imply the solvability of $Ax = b$. As an example, we consider the multiplication operator $M : L^2(0,1) \to L^2(0,1)$ defined by
$$Mf(x) = xf(x).$$
Then $M^* = M$. For any $f \in L^2(0,1)$, if $Mf = 0$, then $xf(x) = 0$. This implies $f(x) = 0$ almost everywhere. Hence $f = 0$ in $L^2(0,1)$. This shows $N(M) = N(M^*) = \{0\}$. Let $g \equiv 1$ on $(0,1)$. $g \in L^2(0,1)$. But the only function f satisfying $xf \equiv 1$ is $1/x$ which is not in $L^2(0,1)$. We can conclude that $R(A)$ is not closed.

5. Another example is that A is an integral operator, say $Af(x) = \int_0^x f(y)\,dy$. Then $A : L^2(0,1) \to L^2(0,1)$. The range of A are those differentiable function with $f(0) = 0$. It is clear that $R(A) \neq L^2(0,1)$. The step function $H(x - 1/2)$ is not in the range of A.

We shall come back to the applications of this theorem after we learn compact operator and singular value decomposition.

Least-squares solutions We have seen that a necessary condition for solvability of $Ax = b$ is $b \perp N(A^*)$. In the case $b \notin N(A^*)^\perp$, we can find the least-squares solution:
$$x^\dagger := \text{arg min}\|Ax - b\|^2$$
The least squares solution may not exist. As in the above example:let us replace
$$Mf(x) := xf(x) = 1$$
in $L^2(0,1)$ by finding
$$\inf \int_0^1 |xf(x) - 1|^2\,dx.$$
This solution, if exists, must be $f(x) = 1/x$, which is not in $L^2(0,1)$.

To find condition for the existence of the least squares solutions, we decompose b into
$$b = \hat{b} + b^\perp \in \overline{R(A)} + N(A^*).$$
Notice that this decomposition is unique. Now, we have
$$\|Ax - b\|^2 = \|Ax - \hat{b}\|^2 + \|b^\perp\|^2.$$
Thus,
$$\inf \|Ax - b\|^2 = \inf \|Ax - \hat{b}\|^2 + \|b^\perp\|^2 = \|b^\perp\|^2.$$
If $\hat{b} \in R(A)$, then there exists a $\hat{x} \in \mathcal{H}$ such that $A\hat{x} = \hat{b}$. Thus, $\hat{x}$ is a least squares solution. If $N(A) = \{0\}$, then the solution is unique. Otherwise, we can find a unique $x^\dagger \in N(A)^\perp$ such that A is 1-1 from $N(A)^\perp \to R(A)$. Any least squares solution satisfies $A\hat{x} = \hat{b}$. We can decompose

$\hat{x} = x_1 + x_2$ with $x_2 \in N(A)$ and $x_1 \in N(A)^\perp$. Then we obtain $Ax_1 = \hat{b}$ and $x_1 \in N(A)^\perp$. Thus, $x_1 = x^\dagger$.

We conclude that given $b \in \mathcal{K}$, the existence of least squares solution of $Ax = b$ if and only if $b \in R(A) + N(A^*)$. These is a unique $x^\dagger \in N(A)^\perp$ which is the unique least squares solution in $N(A)^\perp$. The general least squares solutions have the form $\hat{x} = x^\dagger + N(A)$.

We will see the solvability in more detail later for the cases: (i) A is a Fredholm operator and (ii) A is a compact operator.

7.4 Unitary operators

Definition 7.1. *A linear map $U : \mathcal{H} \to \mathcal{K}$ is called unitary (or orthogonal) if it is invertible and $(Ux, Uy)_\mathcal{K} = (x, y)_\mathcal{H}$ for all $x, y \in \mathcal{H}$.*

In other words, a unitary map is 1-1, onto and preserves inner product. Thus, we have $\|Ux\| = \|x\|$. This implies $\|U\| = 1$. Two spaces $\mathcal{H}$ and $\mathcal{K}$ are called isometric if there is a unitary map between them. We then identify $\mathcal{K}$ as $\mathcal{H}$.

Proposition 2. *(a) A linear map $U : \mathcal{H} \to \mathcal{H}$ is unitary if and only if (b) $U^*U = UU^* = I$.*

Proof. (b) ⇒ (a). If $Ux = 0$, then $x = U^*Ux = U^*0 = 0$. Thus, $N(U) = \{0\}$. Similarly, $N(U^*) = \{0\}$. This shows U and U^* are 1-1. For any $x \in \mathcal{H}$, $U(U^*x) = x$. This shows U is onto. Finally, $(Ux, Uy) = (x, U^*Uy) = (x, y)$. This shows U is inner product preserving. This shows (b) ⇒ (a).
(a) ⇒ (b). From $(Ux, Uy) = (x, y)$, we get $(U^*Ux, y) = (x, y)$ for all $x, y \in \mathcal{H}$. This implies $U^*Ux = x$ for all x. Thus, we have $U^*U = I$. This together with the 1-1 and onto property, we get $U^{-1} = U^*$. This leads to $UU^* = I$. □

Examples of unitary operators

1. A linear map Q in $\mathbb{R}^n$ is orthogonal if and only if $Q^TQ = I$.
2. Given a self-adjoint matrix A in $\mathbb{C}^n$ (i.e. $A - A^* = 0$), define
$$U = e^{iA} = \sum_{n=0}^\infty \frac{(iA)^n}{n!}.$$
Then U is a unitary map in $\mathbb{C}^n$. First, we recall that e^{iA} is well-defined. [1] Using $(A^n)^* = (A^*)^n$, we get
$$U^* = \sum_{n=0}^\infty \frac{(-iA^*)^n}{n!} = e^{-iA^*}.$$
Because $A - A^* = 0$, we get $AA^* = A^2 = A^*A$. Thus, A commutes with A^*. From this, we get $UU^* = e^{iA}e^{-iA^*} = e^{iA-iA^*} = I$.

[1] When A is a matrix, then $\|A\|$ is finite in $B(\mathbb{C}^n, \mathbb{C}^n)$. Thus, the series $\sum_{n=0}^\infty \frac{A^n}{n!}$ converges absolutely and uniformly in $B(\mathbb{C}^n, \mathbb{C}^n)$. With this property, one can show that If $AB = BA$, then $e^{A+B} = e^Ae^B = e^Be^A$.

3. Let $H = -\frac{1}{2}\nabla^2 + V(x)$ be the Schrödinger operator. It is a self-adjoint *unbounded* operator. We shall see later that e^{-itH} is a unitary operator.

4. Suppose $\mathcal{H}$ is a Hilbert space. Let $\{u_n\}_{n\in\mathbb{N}}$ and $\{v_n\}_{n\in\mathbb{N}}$ be two orthonormal bases of $\mathcal{H}$. We define a linear map U on the basis by

$$Uu_n = \lambda_n v_n, \text{ for all } n \in \mathbb{N}$$

where $\lambda_n \in \mathbb{C}$ and $|\lambda_n| = 1$. Then U is a unitary map. For, if $u = \sum_n \alpha_n u_n$, then

$$\|Uu\|^2 = \|\sum_n \alpha\lambda_n v_n\|^2 = \sum_n |\alpha_n|^2 = \|u\|^2.$$

5. Consider the harmonic oscillator is quantum system

$$i\partial_t\psi = H\psi, \psi(0) = \psi_0.$$

The Hamiltonian $H = -d^2/dx^2 + x^2$. Its eigenvalues are $\lambda_n = n^2$, eigenstates are the Hermite polynomials: $u_n = H_n(x)e^{-x^2/2}$. They constitute an orthonormal basis in $L^2(\mathbb{R})$. Its solution $\psi(t) := U(t)\psi_0$ is

$$U(t)\psi_0 := \sum_{n=0}^{\infty} e^{-i\lambda_n t}\alpha_n u_n(x), \text{ where } \psi_0 = \sum_{n=0}^{\infty}\alpha_n u_n(x).$$

The operator $U(t)$ is a unitary operator.

6. Periodic Hilbert transforml: $H : L^2(\mathbb{T}) \to L^2(\mathbb{T})$ is defined by

$$\widehat{(Hf)}_n = i \operatorname{sign} n \hat{f}_n.$$

Or equivalently, $He^{inx} = i(\operatorname{sign} n)e^{inx}$. See see that the weight

$$|i \operatorname{sign} n| = \begin{cases} 1 & \text{when } n \neq 0 \\ 0 & \text{when } n = 0. \end{cases}$$

Thus, $H(1) = 0$ and its kernel is the space spanned by 1, i.e. $N(H) = \langle 1 \rangle$. Its orthogonal complement is

$$\mathcal{H} = \{f \in L^2(\mathbb{T}) | \int_0^{2\pi} f(x)\, dx = 0\}.$$

Then, H is a unitary map in $\mathcal{H}$.

7. The translation operator $T_a : L^2(\mathbb{T}) \to L^2(\mathbb{T})$ defined by

$$(T_a f)(x) = f(x - a)$$

is unitary. The set $\{T_a | a \in \mathbb{T}\}$ is a unitary representation of the additive group $\mathbb{T}$.

8. The Fourier transform $F : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ is a unitary map.
9. The Hilbert transform $H : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ is defined by

$$\widehat{(Hf)}(\xi) = i\,\frac{\xi}{|\xi|}\hat{f}(\xi)$$

is a unitary map.

The mean ergodic theorem. The statistical behavior of a deterministic dynamical system can be characterized by a probabilistic average.

Theorem 7.2 (von Neumann)**.** *Let U be a unitary operator on a Hilbert space $\mathcal{H}$. Let $\mathcal{M}$ be its invariant space, i.e. $\mathcal{M} = \{x \in \mathcal{H} | Ux = x\}$. Let P be the orthogonal projection onto $\mathcal{M}$. Then, for all $x \in \mathcal{H}$, we have*

$$\lim_{N\to\infty} \frac{1}{N+1}\sum_{n=0}^{N} U^n x = Px. \tag{7.3}$$

Proof. 1. From the definition, we have $N(I-U) = R(P) = \mathcal{M}$. We can decompose $\mathcal{H} = N(P) \oplus R(P)$. If $x \in R(P)$, then

$$\frac{1}{N+1}\sum_{n=0}^{N} U^n x = \frac{1}{N+1}\sum_{n=0}^{N} x = x = Px.$$

Thus, (7.3) holds for $x \in R(P)$. So, we only need to prove (7.3) for $x \in N(P)$.

2. Since U is unitary, $Ux = x$ if and only if $x = U^*Ux = U^*x$. Thus, $N(I-U) = N(I-U^*) = \mathcal{M}$.

3. From Theorem 7.1,

$$N(P) = N(I-U)^\perp = N(I-U^*)^\perp = \overline{R(I-U)}.$$

4. For $x \in R(I-U)$, $x = (I-U)y$ for some $y \in \mathcal{H}$.

$$\begin{aligned} \frac{1}{N+1}\sum_{n=0}^{N} U^n x &= \frac{1}{N+1}\sum_{n=0}^{N}(U^n - U^{n+1})y \\ &= \frac{1}{N+1}(y - U^{N+1}y) \\ &\to 0, \text{ as } N \to \infty. \end{aligned}$$

5. For every $x \in N(P)$, we can find $x_k \in R(I-U)$ to approximate x. Hence,

$$\begin{aligned} \left\|\frac{1}{N+1}\sum_{n=0}^{N} U^n x\right\| &\le \left\|\frac{1}{N+1}\sum_{n=0}^{N} U^n (x - x_k)\right\| + \left\|\frac{1}{N+1}\sum_{n=0}^{N} U^n x_k\right\| \\ &\le \|x - x_k\| + \left\|\frac{1}{N+1}\sum_{n=0}^{N} U^n x_k\right\|. \end{aligned}$$

By taking $N \to \infty$, then $k \to \infty$, it follows that (7.3) holds for $x \in N(P)$.

□

Now, we study a deterministic discrete dynamical system on a set Ω. We associate a probability measure $\mathcal{P}$ on Ω. That is, $(\Omega, \mathcal{P})$ is a probability space. A mapping $T : \Omega \to \Omega$ is measure preserving if $\mathcal{P}(T^{-1}A) = \mathcal{P}(A)$ for all measurable set $A \subset \Omega$. We shall study statistical behavior of the iterative map $x^{n+1} = Tx^n$. An important quantity is the average, which is measured by

$$\frac{1}{N+1}\sum_{n=0}^{N} f(x^n)$$

for any continuous function f. However, we may not have topology on Ω. Therefore, alternative way is

$$\frac{1}{N+1}\sum_{n=0}^{N} f \circ T^n.$$

The mapping T then induces an operator U on $L^2(\Omega, \mathcal{P})$ by $f \to f \circ T$. Since T is measure preserving, we have

$$\int_\Omega \overline{f \circ T(x)} g \circ T(x)\, d\mathcal{P}(x) = \int_\Omega \overline{f(x)} g(x)\, d\mathcal{P}(x).$$

That is, $\langle Uf, Ug\rangle = \langle f, g\rangle$. Thus, U is unitary.

Definition 7.2. *A 1-1, onto, measure preserving map T on $(\Omega, \mathcal{P})$ is ergodic if the only function $f \in L^2(\Omega, \mathcal{P})$ such that $f = f \circ T$ are the constant functions.*

This definition states that T is ergodic if the invariant of U is the constant functions, which is spanned by the constant function $f \equiv 1$. Then the von Neumann ergodic theorem implies that:

$$\frac{1}{N+1}\sum_{n=0}^{N} f \circ T^n \to \langle f, 1\rangle 1, \text{ as } N \to \infty,$$

where $\langle f, 1\rangle = \int_\Omega f(x)\, d\mathcal{P}(x)$.

Theorem 7.3. *Let $T : \Omega \to \Omega$ be a 1-1, onto, measure preserving on a probability space $(\Omega, \mathcal{P})$. Assume T is ergodic on Ω. Then for any $f \in L^2(\Omega, \mathcal{P})$,*

$$\frac{1}{N+1}\sum_{n=0}^{N} f \circ T^n \to \int_\Omega f(x)\, d\mathcal{P}(x)$$

in $L^2(\Omega, \mathcal{P})$.

Since L^2 convergence implies convergence almost everywhere, we obtain that

$$\frac{1}{N+1}\left(f(x^0) + f(x^1) + \cdots + f(x^N)\right) \to \int_\Omega f(x)\, d\mathcal{P}(x)$$

for almost all $x^0 \in \Omega$ with $x^{n+1} := Tx^n$.

Remarks.

1. The general ergodic theorem holds for $f \in L^1(\Omega, \mathcal{P})$, due to Birkhoff.

2. If T is only measure preserving, then the limit

$$\lim_{N\to\infty} \frac{1}{N+1} \sum_{n=0}^{N} f(T^n x^0)$$

still exists for almost all $x^0 \in \Omega$.

7.5 Compact operators

Definition 7.3. *A linear map $A : \mathcal{H} \to \mathcal{K}$ is compact if it maps bounded set into a precompact set in $\mathcal{K}$.*

Examples of compact operators.

1. Consider the Poisson equation $- \triangle u = f$ in a bounded domain $\Omega \subset \mathbb{R}^n$ with Dirichlet boundary condition. By the Riesz representation theorem, for any $f \in L^2(\Omega)$, there exists a unique $u \in H_0^1(\Omega)$ such that $(\nabla u, \nabla v) = (f, v)$ for all $v \in H_0^1(\Omega)$. The mapping $K = (-\triangle)^{-1}$ maps f to u is a bounded operator from $L^2(\Omega)$ to $H_0^1(\Omega)$. Since $H_0^1(\Omega)$ is compact embedding into $L^2(\Omega)$, we get K is a compact operator from $L^2(\Omega)$ to $L^2(\Omega)$. The operator Kf has an integral representation:

$$u(x) = \int_\Omega G(x, y) f(y)\, dy$$

G is called the Green function. G satisfies

(a) $G(x, y) = G(y, x)$

(b) $- \triangle G(\cdot, y) = \delta(\cdot - y)$,

(c) $G(x, y) = 0$ for $x \in \partial\Omega$.

The Green function has the following unique representation

$$G(x, y) = \frac{1}{4\pi} \frac{1}{|x-y|} + h(x, y)$$

where $h(x, y) = h(y, x)$, $h(\cdot, y)$ is harmonic in Ω and $h(x, y) = -\frac{1}{4\pi}\frac{1}{|x-y|}$ for $x \in \partial\Omega$.

2. An operator with finite dimensional range is compact.

3. A diagonal operator $A : \ell^2(\mathbb{N}) \to \ell^2(\mathbb{N})$ defined by

$$A(x_1, x_2, x_3, \cdots) = (\lambda_1 x_1, \lambda_2 x_2, \lambda_3 x_3, \cdots)$$

is compact if and only if $\lambda_n \to 0$ as $n \to \infty$.

Proposition 3. *(a) Let $A : \mathcal{H}_1 \to \mathcal{H}_2$ and $B : \mathcal{H}_2 \to \mathcal{H}_3$ be bounded linear operators. If one of them is compact, then AB is also compact.*

(b) Let $A_n : \mathcal{H} \to \mathcal{H}$ be compact operators and $A_n \to A$ uniformly. Then A is also compact.

Proof. (a) The key is that a bounded operator (which is continuous) maps a compact set into a compact.

(b) We use diagonal process to prove (b). Let (x_n) be a bounded sequence in $\mathcal{H}$. We claim that we can find a convergent subsequence of (Ax_n). Since A_m are all compact. We start from $m = 1$, the sequence $(A_1 x_i)$ is bounded and hence has a convergent subsequence $(A_1 x_{1,n})$. The sequence $A_2 x_{1,n}$ is again bounded, thus we can find subsequence $(x_{2,n})$ of $(x_{1,n})$ such that $A_2 x_{2,n}$ converges. We continuous this process to get a subsequence $(x_{m,n})$ of $(x_{m-1,n})$ with $(A_m x_{m,n})_{n\in\mathbb{N}}$ converges. Then we take the subsequence $(x_{m,m})$ which has the property: for every $k \in N$, the sequence $(A_k x_{m,m})_{m\in\mathbb{N}}$ is a Cauchy sequence. We claim that the sequence $(Ax_{m,m})$ is also a Cauchy sequence. For any $\epsilon > 0$, we can find k large enough such that $\|A - A_k\| < \epsilon$.With this k, we can find N such that for $n, m \ge N$, $\|A_k x_{m,m} - A_k x_{n,n}\| < \epsilon$. Combining these two, we get

$$\|Ax_{m,m} - Ax_{n,n}\| < 3\epsilon.$$

This shows $Ax_{m,m}$ is a Cauchy sequence, hence Ax_n has convergent subsequence.

□

Theorem 7.4 (Schauder)**.** *If $K : \mathcal{H} \to \mathcal{K}$ is compact, then so is its dual K^*.*

Proof. Suppose (y_n) is a sequence in $\mathcal{K}$ with $\|y_n\| \le 1$. We want to show $(K^* y_n)$ has convergent subsequence. Let B be the unit ball in $\mathcal{H}$ and $C = \overline{KB}$. From compactness of K, we get that C is compact. Now, (y_n) is not only continuous on C, in fact, they are equi-continuous on C because their norms are bounded by 1. By Arzela-Ascoli theorem, (y_n) has a subsequence, still denoted by (y_n), which converges on C. That is, for any small $\epsilon > 0$, we have

$$\sup_{z\in C} |(y_n - y_m, z)| < \epsilon,$$

if m, n are large enough. But this is the same as

$$\sup_{\|x\|\le 1} |(y_n - y_m, Kx)| = \sup_{\|x\|\le 1} |(K^*(y_n - y_m), x)| = \|K^* y_n - K^* y_m\|$$

Hence, $(K^* y_n)$ is a Cauchy sequence in $\mathcal{H}$. This shows $(K^* y_n)$ has Cauchy subsequence. □

Hilbert-Schmidt operator Let Ω be a measurable set in $\mathbb{R}^d$. Let us consider the following integral operator in $L^2(\Omega)$:

$$Ku(x) = \int_\Omega k(x, y)u(y)\, dy$$

We assume

$$\int_\Omega \int_\Omega |k(x,y)|^2 \, dy \, dx < \infty.$$

Then

$$|Ku(x)|^2 \le \left(\int_\Omega k(x,y)|^2 \, dy \right) \left(\int_\Omega |u(y)|^2 \, dy \right)$$

Hence

$$\int |Ku(x)|^2 \, dx \le \left(\int_\Omega \int_\Omega k(x,y)|^2 \, dy \, dx \right) \left(\int_\Omega |u(y)|^2 \, dy \right)$$

This shows that

$$\|K\|^2 \le \int_\Omega \int_\Omega |k(x,y)|^2 \, dy \, dx$$

Such operator is called a Hilbert-Schmidt operator. We define the Hilbert-Schmidt norm of K by

$$\|K\|_{HS}^2 = \int_\Omega \int_\Omega |k(x,y)|^2 \, dy \, dx.$$

We claim that Hilbert-Schmidt operator is compact. To see this, let $\{e_n(y)\}$ be an orthonormal basis in $L^2(\Omega)$. Since $k \in L^2(\Omega \times \Omega)$, we have for almost $x \in \Omega$, $k(x, \cdot) \in L^2(\Omega)$ and we can expand $k(x,y)$ in terms of $e_n(y)$ as

$$k(x,y) = \sum_n a_n(x) e_n(y)$$

By Parseval equality

$$\|k(x,\cdot)\|^2 = \sum_n |a_n(x)|^2$$

We integrate in x to get

$$\int_\Omega \int_\Omega |k(x,y)|^2 \, dy \, dx = \sum_n \int_\Omega |a_n(x)|^2 \, dx$$

Here, dominant convergence theorem is used. Now, we approximate the operator K by $K_n u(x) = \int_\Omega k_N(x,y) u(y) \, dy$ with

$$k_N(x,y) = \sum_{n \le N} a_n(x) e_n(y).$$

Clearly, K_N has finite dimensional range and hence a compact. We claim that K_N converges to K uniformly. For

$$\|K - K_N\|^2 \le \int_\Omega \int_\Omega |k(x,y) - k_N(x,y)|^2 \, dy \, dx = \sum_{n>N} \int_\Omega |a_n(x)|^2 \, dx$$

Since

$$\sum_{n=1}^\infty \int_\Omega |a_n(x)|^2 = \int_\Omega \int_\Omega |k(x,y)|^2 \, dy \, dx < \infty,$$

we get $\|K - K_N\|^2 \to 0$ as $N \to \infty$. Thus, K_N converges to K uniformly. Since all K_N are compact operators, we get that K is also a compact operator.

7.6 Fredholm Operators

In applications, we encounter the operator $T := I - K$, where K is a compact operator, frequently. Indeed, K is treated as a perturbation relative to the identity operator. Below, we analyze its kernel and range.

Theorem 7.5. *Let K be a compact operator in a Hilbert space $\mathcal{H}$. Let $T = I - K$. The following statements hold.*

(i) $N(T)$ is finite dimensional.

(ii) There is an integer m such that $N(T^k) = N(T^m)$ for all $k \geq m$.

(iii) $R(T)$ is closed.

Remarks From Schauder's theorem, K is compact if and only if K^* is compact. From duality principle and (iii), we get $N(T)^\perp = R(T^*)$ and $N(T^*)^\perp = R(T)$.

Theorem 7.6 (Fredholm Alternative)**.** *Let $T = I - K$. Then one of the following is true:*

(a) either $Tu = f$ has a solution for every $f \in \mathcal{H}$,

(b) or $T^*v = 0$ has a nontrivial solution.

The statement (a) is $R(T) = \mathcal{H}$, which is equivalent to $N(T^*) = \{0\}$. The statement (b) is $R(T^*) \neq \{0\}$. In applications, if $N(T^*) \neq \{0\}$, then the solvability for $Tu = f$ is $f \perp N(T^*)$.

Proof. (i) If $N(T)$ is not finite dimensional, then we can construct an orthonormal set $\{e_n\}_{n\in\mathbb{N}}$ in $N(T)$. Since $T = I - K$ and $Te_n = 0$, we have $e_n = Ke_n$. From K being compact, $(Ke_n)_{n\in N}$, hence $(e_n)_{n\in\mathbb{N}}$, is precompact. But this is impossible because $(e_n, e_m) = 0$ for all $m \neq n$.

(ii) If there exists m such that $N(T^{m+1}) = N(T^m)$, then for all $k > m$, $N(T^k) = N(T^m)$. We prove by induction. We will only show that if $N(T^{m+1}) = N(T^m)$, then $N(T^{m+2}) = N(T^{m+1})$. It is obvious that $N(T^{m+1}) \subset N(T^{m+2})$. Conversely, if $x \in N(T^{m+2})$, then $Tx \in N(T^{m+1})$. By our assumption $N(T^{m+1}) = N(T^m)$. Hence $T^m(Tx) = 0$. Thus, we get $x \in N(T^{m+1})$.

Next, if for every $m \in \mathbb{N}$, $N(T^m)$ is a proper subspace of $N(T^{m+1})$, then for all $m \in \mathbb{N}$, we can find unit vector $e_{m+1} \perp N(T^m)$ and $e_{m+1} \in N(T^{m+1})$. Then $(e_m)_{m\in\mathbb{N}}$ is an orthonormal set. Take $m < n$, from $T = I - K$, we have

$$Ke_n - Ke_m = e_n - Te_n - e_m + Te_m.$$

The last three terms: $e_m \in N(T^m) \subset N(T^{n-1})$; $e_n \in N(T^n)$ implies $Te_n \in N(T^{n-1})$; and $Te_m \in N(T^{m-1}) \subset N(T^{n-1})$. Thus, the last three terms are in $N(T^{n-1})$. Hence,

$$\|Ke_n - Ke_m\| = \|e_n - (Te_n + e_m - Te_m)\| \geq 1$$

for all $m < n$. This contradicts to the compactness of K.

(iii) Suppose $(y_n = (I - K)x_n)$ is a convergent sequence in $R(I - K)$. We want to show that there is an x and a subsequence of (x_n) which converges to x. Let first assume $N(T) = \{0\}$. Suppose $\|x_n\|$ is unbounded. Pick up the subsequence, still call it (x_n), whose norm tends to ∞. Consider $z_n = x_n/\|x_n\|$. We have

$$\frac{y_n}{\|x_n\|} = z_n - Kz_n.$$

Since y_n is a convergence, $y_n/\|x_n\| \to 0$. (z_n) is a bounded sequence now, hence Kz_n has a convergent subsequence, still denote it by (z_n), which converges to z. Then we have

$$0 = z - Kz.$$

Hence $z \in N(T)$. By our assumption, $z = 0$. But z is the limit of z_n and $\|z_n\| = 1$ for all n. This is impossible. Hence the assumption $\|x_n\|$ is unbounded is impossible.

Now (x_n) is a bounded sequence. Hence (Kx_n) has convergent subsequence, still denote by Kx_n. From $y_n = x_n - Kx_n$, we see both (y_n) and (Kx_n) converge. Thus, (x_n) also converges.

Finally, we do not make the assumption $N(T) = \{0\}$. Let z_k be the projection of x_k on $N(T)$. Consider $w_k = (x_k - z_k) \perp N(T)$. Then

$$y_k = Tx_k = T(x_k - z_k) = Tw_k.$$

We now replace $\mathcal{H}$ by $N(T)^\perp$. By the previous argument, we get (w_k) has a subsequence converges to $w \in N(T)^\perp$. Hence $y = \lim y_k = \lim Tw_k = Tw$. This shows that $R(T)$ is closed.

□

Least-Squares Solutions Let K be a compact operator in $\mathcal{H}$ and let $T = I - K$. Consider the equation

$$Tu = f$$

and suppose $f \notin N(T^*)^\perp$. In this case, we look for a solution which minimizes $\|Tu - f\|^2$. Such a solution is called the least-squares solution. We can decompose $\mathcal{H}$ in the domain and range as the follows:

$$T : N(T) \oplus R(T^*) \to N(T^*) \oplus R(T).$$

This is because both $R(T^*)$ and $R(T)$ are closed. Further, T is 1-1 and onto from $R(T^*) \to R(T)$ and has a bounded inverse.

Now, given $f \in \mathcal{H}$, we decompose $f = f^* + f^\perp$, with $f^* \in R(T)$ and $f^\perp \in N(T^*)$. With f^*, there exists a unique $u^* \in R(T^*)$ such that $Tu^* = f^*$. Then the solution u^* is the least-squares solution. For any $u \in \mathcal{H}$, we can decompose $u = v + u^\perp$, with $v \in R(T^*)$ and $u^\perp \in N(T)$. Then $Tu = Tv$. We have

$$\begin{aligned}\|Tu - f\|^2 &= \|Tv - f^*\|^2 + \|f^\perp\|^2 \geq \|f^\perp\|^2 \\ &= \|f^\perp\|^2 + \|Tu^* - f^*\|^2 = \|Tu^* - f\|^2.\end{aligned}$$

This shows that u^* is a least-squares solution.

Homeworks 7.1.

Consider the Dirichlet problem in a bounded smooth domain $\Omega \subset \mathbb{R}^3$:

$$\triangle u(x) = 0, \; x \in \Omega$$

$$u(x) = f(x), \; x \in \partial\Omega$$

Let

$$k(x, y) = \frac{1}{2\pi} \frac{(x - y, n_y)}{|x - y|^3}, x \in \Omega, y \in \partial\Omega,$$

where n_y is the outer normal of $\partial\Omega$ at y. The goal of boundary integral method is to find a function ϕ defined on $\partial\Omega$ such that $u(x) = \int_{\partial\Omega} k(x, y)\, dy$. The problems below are the steps toward this goal.

1. If $x_0 \in \partial\Omega$, show that

$$u(x) \to -\phi(x_0) + \int_{\partial\Omega} k(x_0, y)\phi(y)\, dy, \text{ as } x \to x_0 \text{ from inside.}$$

2. The operator $K\phi(x) := \int_{\partial\Omega} k(x, y)\phi(y)\, dy$ is a compact operator in $L^2(\partial\Omega)$.
   Hint: Consider the regularization

$$k_\delta(x, y) = \frac{1}{2\pi} \frac{(x - y, n_y)}{|x - y|^3 + \delta}, \delta > 0.$$

   Try to show $K_\delta \to K$ and K_δ is compact.

3. Show that

$$(-I + K)\phi = f$$

   has a unique solution for $f \in L^2(\partial\Omega)$.

Chapter 8

Spectrum of Bounded Operators

8.1 Spectrum and resolvent

In this chapter, we want to represent a bounded operator as a direct sum of simple operators on smaller spaces. Such a small space is called an invariant space. If the operator restricted to this invariant space is a scalar multiple of a identity, such invariant subspace is called an eigenspace.

Definition 8.1. *(a) Let A be a bounded operator on a Hilbert space $\mathcal{H}$. A subspace $\mathcal{K}$ is called invariant under A if $A\mathcal{K} \subset \mathcal{K}$.*

(b) If $Ax = \lambda x$ for some $\lambda \in \mathbb{C}$ and $x \in \mathcal{H}$, then λ is called an eigenvalue of A and x the corresponding eigenvector. The space $\langle x \rangle$ is an invariant subspace.

(c) A complex number λ is said to be in the resolvent set $\rho(A)$ is that the operator $(\lambda - A)$ is a bijection with a bounded inverse $R_\lambda(A) := (\lambda - A)^{-1}$.

(d) We call $\sigma(A) := \mathbb{C} \backslash \rho(A)$ the spectrum of A.

Remarks.

1. From the bounded inverse theorem (or the open mapping theorem) that a *bijective* bounded operator from a Banach space *onto* another Banach space has bounded inverse. In finite dimensional space $\mathcal{H}$, the complement of resolvent set is the collection of all eigenvalues. In infinite dimensional space, it contains others.

2. If $|\lambda| > \|A\|$, then $\lambda \in \rho(A)$. This is because

$$(\lambda - A)^{-1} = \lambda^{-1}\left(I - \frac{A}{\lambda}\right)^{-1} = \lambda^{-1}\sum_{n=0}^{\infty}\left(\frac{A}{\lambda}\right)^n$$

This series converges uniformly and absolutely because $\|A/\lambda\| < 1$.

Definition 8.2. *Let A be a bounded linear operator on a Hilbert space $\mathcal{H}$.*

(a) The point spectrum of A are those $\lambda \in \sigma(A)$ such that $\lambda - A$ is not 1-1, that is, $N(\lambda - A) \neq \{0\}$. The point spectrum consists of those eigenvalues. We denote this set by $\sigma_p(A)$.

(b) The continuous spectrum of A consists of those λ such that $(\lambda - A)$ is 1-1 but not onto, and $R(\lambda - A)$ is dense in $\mathcal{H}$.We denote this set by $\sigma_c(A)$.

(c) The residual spectrum of A consists of those λ such that $(\lambda - A)$ is 1-1, not onto, and $R(\lambda - A)$ is not dense in $\mathcal{H}$.We denote this set by $\sigma_r(A)$.

Example 1 Let $\mathcal{H} = L^2(0,1)$. Define

$$Mu(x) = xu(x).$$

Then $M : \mathcal{H} \to \mathcal{H}$ is a bounded operator with $\|M\| = 1$. We claim that $\sigma(M) = \sigma_c(M) = [0,1]$. If $\lambda \in \mathbb{C}\backslash[0,1]$, then $(\lambda - M)$ has bounded inverse. Namely, $(\lambda - x)u(x) = f(x)$ gives $u(x) = (\lambda - x)^{-1}f(x)$ which is in $L^2(0,1)$ provided $f \in L^2(0,1)$. This shows $\sigma(M) \subset [0,1]$. On the other hand, if $\lambda \in [0,1]$, we first find that $(\lambda - M)$ is 1-1. This is because $Mu = \lambda u$ for some $\lambda \in [0,1]$, then $u(x) = 0$ a.e.. Thus M is 1-1 and M has no eigenvalue. For $\lambda \in [0,1]$, $(\lambda - M)$ is not onto, because the nonzero constant function $g(x) = c$ is in $L^2(0,1)$ while its pre image of $\lambda - M$ is $c/(\lambda - x)$ with is not in $L^2(0,1)$. This shows that $\sigma(M) = [0,1]$. Finally, we claim that for any $\lambda \in [0,1]$, $R(\lambda - M)$ is dense in $L^2(0,1)$. To see this, for any $f \in L^2(0,1)$, we define

$$f_n(x) = \begin{cases} f(x) & \text{if } |x - \lambda| > 1/n, \\ 0 & \text{if } |x - \lambda| \le 1/n. \end{cases}$$

Then $f_n \to f$ in $L^2(0,1)$. Notice that every $f_n \in R(\lambda - M)$ because $(\lambda - x)^{-1}f_n(x) \in L^2(0,1)$. We conclude that $\sigma(M) = \sigma_c(M)$.

Example 2 We can also define the resolvent and spectrum of an unbounded operator by the same way. For instance, $A = -\partial^2$ is an unbounded operator in $L^2(\mathbb{R})$. For any $\lambda \in \mathbb{C}$, we can solve

$$(\lambda + \partial_x^2)u = f$$

by using Fourier transform:

$$(\lambda - |\xi|^2)\hat{u}(\xi) = \hat{f}(\xi).$$

This yields

$$\hat{u}(\xi) = \frac{1}{\lambda - |\xi|^2}\hat{f}(\xi).$$

If $\lambda \in \mathbb{C}$ and $\lambda \notin [0,\infty)$, then $\hat{u} \in L^2(\mathbb{R})$ provided $\hat{f} \in L^2(\mathbb{R})$. From Parseval equality: $\|f\|^2 = \|\hat{f}\|^2$, we conclude that $\lambda + \partial_x^2$ has a bounded inverse in $L^2(\mathbb{R})$ for any $\lambda \in \mathbb{C} \setminus [0,\infty)$. On the other hand, for any $\lambda \in [0,\infty)$, we claim that $(\lambda + \partial^2)$ is 1-1, but not onto and $R(\lambda + \partial^2)$ is dense in $L^2(\mathbb{R})$. These can easily be checked by taking the Fourier transform. The operator $\lambda + \partial^2$ becomes $\lambda - |\xi|^2$. Using the same argument as example 1, we get that $\lambda - |\xi|^2$ in $L^2(\mathbb{R}_\xi)$ is 1-1, not onto, but the range is dense in $L^2(\mathbb{R}_\xi)$. Thus, $\sigma(-\partial^2) = \sigma_c(-\partial^2) = [0,\infty)$.

Example 3 Consider the shift operators $\ell^2(\mathbb{N}) \to \ell^2(\mathbb{N})$ defined by

$$R(x_1, x_2, ...) = (0, x_1, x_2, ...),\ L(x_1, x_2, x_3, ...) = (x_2, x_3, ...).$$

1. It is easy to see that $L^* = R$ and $R^* = L$. And $LR = I$ but $RL \neq I$.
2. Both L and R satisfy $\|R\| = 1$ and $\|L\| = 1$. Hence we have both $\sigma(R), \sigma(L) \subset \{\lambda \mid |\lambda| \leq 1\}$.
3. We claim that $0 \in \sigma_r(R) = \{0\}$. This is because $N(R) = \{0\}$ but $e_1 \notin R(R)$. Thus, the range of R is not dense in $\ell^2(\mathbb{N})$. Hence, $0 \in \sigma_r(R)$.
4. For any $|\lambda| < 1$, we claim λ is an eigenvalue of L. We solve $(\lambda - L)x = 0$, which is

   $$\lambda x_n - x_{n+1} = 0, n = 1, 2, ...$$

   This yields $x = (1, \lambda, \lambda^2, \lambda^3, ...)x_1$, which is in ℓ^2 if and only if $|\lambda| < 1$. Thus,

   $$\{\lambda |\, |\lambda| < 1\} = \sigma_p(L).$$

5. We claim that $N(\lambda - L) = \{0\}$ for $|\lambda| = 1$. When $(\lambda - L)x = 0$, we have $\lambda x_n - x_{n+1} = 0$ for $n \geq 1$. This gives $x = x_1(1, \lambda, \lambda^2, ...)$. Such $x \in \ell^2$ with $|\lambda| = 1$ yields $x_1 = 0$. Thus, $N(\lambda - L) = \{0\}$ for $|\lambda| = 1$.
6. We claim that $N(\lambda - R) = \{0\}$ for $\lambda \neq 0$. When $\lambda \neq 0$, $(\lambda - R)x = 0$ means

   $$\lambda(x_1, x_2, x_3, ...) - (0, x_1, x_2, ...) = (0, 0, 0, ...).$$

   This yields that $\lambda x_1 = 0$, $x_n = \lambda^{-1} x_{n-1}$ for $n \geq 2$. Thus, $x = 0$.
7. To study $R(\lambda - R)$, we solve

   $$\lambda(x_1, x_2, x_3, ...) - (0, x_1, x_2, ...) = (y_1, y_2, y_3, ...).$$

   This yields

   $$\lambda x_1 = y_1, \lambda x_n - x_{n-1} = y_n, n \geq 2.$$

   If we choose $y = e_1$. This gives

   $$x_1 = 1/\lambda, x_n = \lambda^{-n+1}, n \geq 2.$$

   If $|\lambda| \leq 1$ and $y = e_1$, then $x \notin \ell^2(\mathbb{N})$. Thus, $R(\lambda - R) \neq \ell^2(\mathbb{N})$.
8. Similarly, we solve $(\lambda - L)x = y$ with $|\lambda| = 1$ and $y = e_1$, we find

   $$x_n = \lambda^{n-1} x_1 + \lambda^{n-2}, n \geq 2.$$

   Thus, $e_1 \notin R(\lambda - L)$ for $|\lambda| = 1$.

9. From duality principle,
$$\overline{R(\lambda^* - R)} = N(\lambda - L)^\perp = \ell^2(\mathbb{R}) \text{ for } |\lambda| = 1$$
and
$$\overline{R(\lambda^* - L)} = N(\lambda - R)^\perp = \ell^2(\mathbb{R}) \text{ for } \lambda \neq 0.$$
Thus, we have
$$\sigma_c(L) = \{\lambda \mid |\lambda| = 1\}, \sigma_c(R) = \{\lambda \mid |\lambda| \le 1, \lambda \neq 0\}$$
We summary them as
$$\sigma_p(L) = \{\lambda \mid |\lambda| < 1\}, \sigma_c(L) = \{\lambda \mid |\lambda| = 1\}, \sigma_r = \phi;$$
$$\sigma_p(R) = \phi, \sigma_c(R) = \{\lambda \mid |\lambda| \le 1, \lambda \neq 0\},\ \sigma_r(R) = \{0\}.$$

Operator-valued analytic function We consider the space of all bounded operators in $\mathcal{H}$, denoted by $B(\mathcal{H})$. It is a Banach space under operator norm. Given $A \in B(\mathcal{H})$, we consider $R_\lambda(A) = (\lambda - A)^{-1}$, which is an operator-valued function defined on $\rho(A)$.

Proposition 4. *Let A be a bounded operator in a Hilbert space $\mathcal{H}$.*

(a) The resolvent set $\rho(A)$ is an open set in $\mathbb{C}$.

(b) $R_\lambda(A) : \rho(A) \to B(\mathcal{H})$ is analytic.

(c) For any $\lambda, \mu \in \rho(A)$, $R_\lambda(A)$ and $R_\mu(A)$ commute and
$$R_\lambda(A) - R_\mu(A) = (\mu - \lambda) R_\mu(A) R_\lambda(A).$$

(d) $\sigma(A) \subset \{z \mid |z| \le \|A\|\}$.

(e) $\sigma(A)$ is not empty.

Proof. (a) We begin with formal computation. Suppose $\lambda \in \rho(A)$. For $\mu \sim \lambda$,
$$\begin{aligned}\frac{1}{\mu - A} &= \frac{1}{\mu - \lambda + \lambda - A} = (\lambda - A)^{-1} \frac{1}{1 - \left(\frac{\lambda - \mu}{\lambda - A}\right)} \\ &= \left(\frac{1}{\lambda - A}\right)\left[\sum_{n=0}^\infty \left(\frac{\lambda - \mu}{\lambda - A}\right)^n\right] \\ &= R_\lambda(A)\left[\sum_{n=0}^\infty (\lambda - \mu)^n R_\lambda(A)^n\right]\end{aligned}$$
Since $\|R_\lambda(A)^n\| \le \|R_\lambda(A)\|^n$, the series
$$\sum_{n=0}^\infty (\lambda - \mu)^n R_\lambda(A)^n$$

converges uniformly if
$$|\lambda - \mu| < \|R_\lambda(A)\|^{-1}.$$

One can check that

$$(\mu - A)\left[\sum_{n=0}^{\infty}(\lambda-\mu)^n R_\lambda(A)^{n+1}\right] = I = \left[\sum_{n=0}^{\infty}(\lambda-\mu)^n R_\lambda(A)^{n+1}\right](\mu - A).$$

This shows that $\mu \in \rho(A)$ if $|\mu - \lambda| < \|R_\lambda(A)\|$. Thus, $\rho(A)$ is open.

(b) Since $R_\mu(A)$ has a power series expansion, it is analytic.

(c) We have

$$\begin{aligned} R_\lambda(A) - R_\mu(A) &= R_\lambda(A)(\mu - A)R_\mu(A) - R_\lambda(A)(\lambda - A)R_\mu(A) \\ &= (\mu - \lambda)R_\lambda(A)R_\mu(A). \end{aligned}$$

Interchanging λ and μ in the above formula shows that $R_\lambda(A)$ and $R_\mu(A)$ commute.

(d) Notice that
$$R_\lambda(A) = \frac{1}{\lambda}\sum_{n=0}^{\infty}\left(\frac{A}{\lambda}\right)^n \tag{8.1}$$

If $|\lambda| > \|A\|$, then the series converges. Hence $\lambda \in \rho(A)$. Or equivalently, $\sigma(A) \subset \{\lambda \mid |\lambda| \le \|A\|\}$.

(e) If $\sigma(A)$ is empty, then $R_\lambda(A)$ is analytic on entire $\mathbb{C}$. From (8.1), we see that $R_\lambda(A) \to 0$ as $\lambda \to \infty$. Thus, $R_\lambda(A)$ is a bounded entire function. By Liouville theorem, $R_\lambda(A)$ would be a constant and hence zero. This is a contradiction.

□

Definition 8.3. *The spectral radius* $r(A) := \sup\{|\lambda|\, |\lambda \in \sigma(A)\}$.

Proposition 5. *If A is a bounded operator, then* $r(A) = \lim_{n\to\infty} \|A^n\|^{1/n}$.

Proof. Let $a_n = \log \|A^n\|$. We want to show that a_n/n conveges. Since $\|A^{m+n}\| \le \|A^m\|\,\|A^n\|$, we have
$$a_{m+n} \le a_m + a_n, \text{ and } a_{pm} \le pa_m.$$
We write $n = pm + q$ with $0 \le q < m$. It follows that
$$a_n \le pa_m + a_q.$$
Dividing this formula by n. Taking $n \to \infty$ with m fixed. We get $p/n \to 1/m$. Hence
$$\limsup_{n\to\infty} \frac{a_n}{n} \le \frac{a_m}{m}.$$

Then we take $m \to \infty$ to obtain

$$\limsup_{n\to\infty} \frac{a_n}{n} \le \liminf_{m\to\infty} \frac{a_m}{m}.$$

This shows $\lim_{n\to\infty} a_n/n$ exists.

Let us denote $\lim_{n\to\infty} \|A^n\|^{1/n}$ by ρ. We shall show $r(A) = \rho$.

First, we expand $(\lambda - A)^{-1}$ formally as a Neumann series

$$\frac{1}{\lambda} \sum_{n=0}^{\infty} \left(\frac{A}{\lambda}\right)^n.$$

From the definition of ρ, we see that this Neumann series converges if $\rho/|\lambda| < 1$ and diverges if $\rho/|\lambda| > 1$. For if $\rho/|\lambda| < 1$, it means that we can find an N such that for any $n \ge N$,

$$\frac{\|A^n\|^{1/n}}{|\lambda|} < \eta < 1.$$

Or

$$\frac{\|A^n\|}{|\lambda|^n} < \eta^n, \text{ with } \eta < 1.$$

Thus, the Neumann series converges. Similar augument for the divergence proof. Thus, we get $\{\lambda \mid |\lambda| > \rho\} \subset \rho(A)$ from the convergence argument, and $\{\lambda \mid |\lambda| < \rho\} \subset \sigma(A)$ from the divergence argument. These two imply $\rho = r(A)$. □

Remark. Notice that $r(A) = 0$ does not imply $A = 0$. For instant, the matrix

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

has zero spectral radius but it is not zero. A bounded operator is called nilponent if $r(A) = 0$.

Homeworks 8.1.

1. *Ex. 9.2*
2. *Ex. 9.6*
3. *Ex. 9.7*
4. *Ex. 9.8*
5. *Ex. 9.10*

8.2 The spectrum of bounded self-adjoint operators

Definition 8.4. *A linear operator A in a Hilbert space $\mathcal{H}$ is self-adjoint if $A^* = A$.*

The operator $Mu(x) := xu(x)$ defined in the last subsection is a self-adjoint operator.

Proposition 6. *If A is a bounded self-adjoint operator on a Hilbert space $\mathcal{H}$ and $\mathcal{M}$ is an invariant subspace of A, then $\mathcal{M}^\perp$ is also an invariant subspace of A.*

Proof. If $x \in \mathcal{M}^\perp$, we need to check $Ax \perp \mathcal{M}$. For any $y \in \mathcal{M}$,

$$(Ax, y) = (x, Ay) = 0.$$

Thus, $Ax \in \mathcal{M}^\perp$. □

This proposition says that A can be *decoupled* on $\mathcal{M}$ and $\mathcal{M}^\perp$.

Proposition 7. *The eigenvalues of a bounded self-adjoint operator on a Hilbert space $\mathcal{H}$ are real, and the eigenvectors associated with different eigenvalues are orthogonal.*

Proof. Suppose $\lambda \in \mathbb{C}$ is an eigenvalue of A associated with an eigenvector $x \neq 0$.

$$\lambda(x, x) = (x, Ax) = (Ax, x) = \overline{\lambda}(x, x).$$

Since $x \neq 0$, we conclude $\lambda = \overline{\lambda}$.

If $Ax = \lambda x$ and $Ay = \mu y$, with $\lambda \neq \mu$ and $\lambda, \mu \in \mathbb{R}$, then

$$\lambda(x, y) = (Ax, y) = (x, Ay) = \mu(x, y)$$

Since $\lambda \neq \mu$, we obtain $(x, y) = 0$. □

Proposition 8. *If A is a bounded self-adjoint operator on a Hilbert space $\mathcal{H}$, then*

(a) $\|A\| = \sup_{\|x\|=1} |(Ax, x)|$;

(b) The spectral radius $r(A) = \|A\|$.

Proof. (a) 1. Let us denote $\sup_{\|x\|=1} |(Ax, x)|$ by α. From

$$|(Ax, x)| \leq \|A\| \|x\|^2,$$

we get $\alpha \leq \|A\|$.

2. To prove the reverse inequality, we first show that

$$\|A\| = \sup\{|(Ax, y)| \mid \|x\| = 1, \|y\| = 1\}.$$

This follows from

$$\|A\| = \sup_{\|x\|=1} \|Ax\| = \sup_{\|x\|=1} \sup_{\|y\|=1} |(Ax, y)|. \tag{8.2}$$

3. We use polarization formula

$$\begin{aligned}(Ax,y) &= \frac{1}{4}[(A(x+y),x+y)-(A(x-y),x-y)\\ &\quad -i(A(x+iy),x+iy)+i(A(x-iy),x-iy)]\end{aligned}$$

Since A is self-adjoint, the first two terms are real and the last two are pure imaginary. We replace y by $e^{i\phi}$ so that $(Ax, e^{i\phi}y)$ is real. Then the last two imaginary terms vanish. Hence we get

$$\begin{aligned}|(Ax,y)|^2 &= \frac{1}{16}|(A(x+y),x+y)-(A(x-y),x-y)|^2\\ &\le \frac{\alpha^2}{16}(\|x+y\|^2+\|x-y\|^2)^2\\ &= \frac{\alpha^2}{4}(\|x\|^2+\|y\|^2)^2.\end{aligned}$$

Here, we have used the parallelogram laws. Using this in (8.2), we obtain $\|A\| \le \alpha$.

(b) 1. First, we show $\|A^2\| = \|A\|^2$. We have

$$\|A\|^2 = \sup_{\|x\|=1}(Ax,Ax) = \sup_{\|x\|=1}|(A^2x,x)| = \|A^2\|$$

Here, we have used self-adjointness of A. From this, we can get $\|A\|^{2^m} = \|A^{2^m}\|$
2. We have shown that $r(A) = \lim_{n\to\infty}\|A^n\|^{1/n}$. In particular, we choose a subsequence with $n = 2^m$, then we get $r(A) = \|A\|$.

□

Theorem 8.1. *If A is a bounded self-adjoint operator on a Hilbert space $\mathcal{H}$, then*

(a) $\sigma(A) \subset [-\|A\|, \|A\|]$;

(b) The residual spectrum of A is empty.

Proof. 1. We want to show that for $\lambda = a+ib$ with $b \neq 0$, it holds that $\lambda - A$ is invertible. For any $x \in \mathcal{H}$,

$$\begin{aligned}\|(A-\lambda)x\|^2 &= ((A-\lambda)x,(A-\lambda)x)\\ &= ((A-a)x,(A-a)x)+((-ib)x,(-ib)x)\\ &\quad +(Ax,(-ib)x)+((-ib)x,Ax)\\ &= \|(A-a)x\|^2+b^2\|x\|^2\\ &\ge b\|x\|^2.\end{aligned}$$

This shows that $\lambda - A$ is 1-1 and has closed range.

2. If $R(\lambda - A) = \mathcal{H}$, then $\lambda - A$ has bounded inverse, hence $\lambda \in \rho(A)$.

3. If $R(\lambda - A) \neq \mathcal{H}$, then, because it is closed, there exists $z \neq 0$ such that $z \perp R(\lambda - A)$. That is,
$$0 = (z, (\lambda - A)x) = ((\overline{\lambda} - A)z, x), \text{ for all } x \in \mathcal{H}.$$
Here, we have used $A^* = A$. This implies $\overline{\lambda}$ is an eigenvalue of A. Since the eigenvalues of bounded self-adjoint operator must be real, we get $\lambda \in \mathbb{R}$. This is a contradiction.

4. We have shown that $r(A) = \|A\|$ and $\sigma(A) \subset \mathbb{R}$. Combine these two, we get $\sigma(A) \subset [-\|A\|, \|A\|]$.

5. If $\lambda \in \sigma_r(A)$, then $\overline{R(\lambda - A)} \neq \mathcal{H}$. Hence there exists a $z \neq 0$ such that $z \perp R(\lambda - A)$. By the same argument above, we get that $\overline{\lambda}$ is an eigenvalue. Since $\sigma(A) \subset \mathbb{R}$, we see that $\lambda = \overline{\lambda}$ and $\lambda \subset \sigma_p(A) \cap \sigma_r(A)$, which is an empty set. Thus, $\sigma_r(A)$ is empty.

□

8.3 The spectrum of compact operators

Given a compact operator K in a Hilbert space $\mathcal{H}$, we shall study the operator $\lambda - K$. We may normalize it by $I - \frac{K}{\lambda}$.

Corollary 8.3. *Let K be a compact operator in a Hilbert space $\mathcal{H}$. Then the only possible continuous spectrum is zero.*

8.4 Spectral theorem for compact, self-adjoint operators

Theorem 8.2. *If A is a compact self-adjoint operatorin a Hilbet space $\mathcal{H}$, then the only accumulation points of eigenvalues of A is zero.*

Proof. Suppose $\lambda_n \neq 0$ are eigenvalues of A, $\lambda_n \to \lambda$. Suppose $\lambda \neq 0$, we want to get a contradiction. Let e_n be the associated unit eigenvectors. They form an orthonormal set. Let $f_n = e_n/\lambda_n$. Then f_n are bounded because λ_n are bounded away from zero. The sequence $(Af_n) = (e_n)$ cannot have a Cauchy subsequence. This contradicts to the compactness of A. □

Theorem 8.3 (Rayleigh Principle)**.** *Suppose A is a compact, self-adjoint operator in a Hilbert space $\mathcal{H}$. Then the maximum*
$$\sup_{\|x\|=1} |(Ax, x)| = \|A\|$$
is attained. The maximal point is an eigenvalue of A.

Proof. We have seen that $\sup_{\|x\|=1} |(Ax, x)| = \|A\|$ for bounded self-adjoint operator. Suppose (x_n) be a sequence of unit vectors such that $(Ax_n, x_n) \to \lambda$, where $\lambda = \|A\|$ or $-\|A\|$. From the boundedness of (x_n) and compactness of A, (Ax_n) has a convergent subsequence. We still denote it by (Ax_n) and $Ax_n \to y$. We claim y is an eigenvector.

First, $y \neq 0$, otherwise $(Ax_n, x_n) \to 0$ and leads to $\|A\| = 0$, and the statement is trvially true. Second, we check

$$\begin{aligned}
\|Ay - \lambda y\|^2 &= \lim_{n\to\infty} \|(A-\lambda)Ax_n\|^2 \\
&\le \|A\|^2 \lim_{n\to\infty} \|(A-\lambda)x_n\|^2 \\
&= \|A\|^2 \lim_{n\to\infty} [\|(Ax_n\|^2 - 2\lambda(Ax_n, x_n) + \lambda^2\|x_n\|^2] \\
&\le \|A\|^2 \lim_{n\to\infty} [\|A\|^2\|x_n\|^2 - 2\lambda(Ax_n, x_n) + \lambda^2\|x_n\|^2] \\
&= \|A\|^2[\lambda^2 - 2\lambda^2 + \lambda^2] = 0.
\end{aligned}$$

□

Theorem 8.4 (Spectral theorem for compact, self-adjoint operators)**.** *Let A be a compact, self-adjoint operator on a Hilbert space $\mathcal{H}$. Then A can be diagonalized in the following sense:*

1. *there are countable (finite or infinite) eigenvalues λ_n which can be ordered so that $(|\lambda_n|)$ is a nonincreasing sequence and $\lambda_n \to 0$;*

2. *The eigenspaces $\mathcal{E}_n$ associated with those nonzero eigenvalues λ_n are finite dimensional.*

3. *Let P_n be the orthogonal projection onto $\mathcal{E}_n$. Then*

$$A = \sum_n \lambda_n P_n.$$

Proof. 1. We shall find the eigenspaces by successively applying the Rayleigh principle. We begin with $A_1 = A$ and $\mathcal{N}_1 = \mathcal{H}$. By the Rayleigh principle, there exists a unit vector e_1 such that $A_1 e_1 = \lambda_1 e_1$ and $|\lambda_1| = \|A_1\|$. Let $\mathcal{E}_1 = \langle e_1 \rangle$ and $\mathcal{N}_2 = \mathcal{E}_1^\perp$. Hence, $\mathcal{N}_1 = \mathcal{E}_1 \oplus \mathcal{N}_2$.

2. Since A_1 is self-adjoint, $\mathcal{N}_2$ is also an invariant subspace of A. We define P_1 be the orthogonal projection onto $\mathcal{E}_1$ and A_2 be the restriction of A_1 on $\mathcal{N}_2$. Then

$$A_1 = A_1 \circ (P_1 + (I - P_1)) = \lambda_1 P_1 + A_2,$$

and $A_2 : \mathcal{N}_2 \to \mathcal{N}_2$ is compact and self-adjoint. Since A_2 is a restriction of A_1, we get $\|A_2\| \le \|A_1\|$. We apply Rayleigh principle to A_2 on $\mathcal{N}_2$ to get λ_2, e_2 and $\mathcal{E}_2 := \langle e_2 \rangle$. We have

$$|\lambda_2| = \|A_2\| \le \|A_1\| = |\lambda_1|,$$

and $e_2 \perp e_1$.

3. We continue this process inductively: Given the decomposition $\mathcal{N}_{n-1} = \mathcal{E}_{n-1} \oplus \mathcal{N}_n$ and the operator $A_{n-1} = \lambda_{n-1} P_{n-1} + A_n$, we apply Rayleigh principle to A_n on $\mathcal{N}_n$ to get λ_n and e_n; define $\mathcal{E}_n = \langle e_n \rangle$; $\mathcal{N}_{n+1}$ to be the orthogonal complement of $\mathcal{E}_n$ in $\mathcal{N}_n$; and define A_{n+1} to be the restriction of A_n on $\mathcal{N}_{n+1}$. Then we get $|\lambda_n| = \|A_n\|$, $e_n \perp e_i$ for all $i < n$, and $A_n = \lambda_n P_n + A_{n+1}$.

4. If $A_{n+1} = 0$ for some n, then $A = \sum_{i=1}^{n} \lambda_i P_i$ and $N(A) = \mathcal{N}_{n+1}$.

5. If $A_n \neq 0$ for all n, then there are infinite many eigenvalues λ_n, $n \in \mathbb{N}$. Each of them has only finite multiplicity. By Theorem 8.2 and monotonicity of $|\lambda_n|$, we get $\lambda_n \to 0$.

6. For every n,
$$A = \sum_{i=1}^{n} \lambda_i P_i + A_{n+1}.$$
Since $\|A_{n+1}\| = |\lambda_{n+1}| \to 0$ as $n \to \infty$, we get
$$\left\| A - \sum_{i=1}^{n} \lambda_i P_i \right\| \to 0.$$

7. The range of A is
$$R(A) = \{\sum_{n=1}^{\infty} \lambda_n a_n e_n \mid \sum_{n=1}^{\infty} |a_n|^2 < \infty.\}$$
Hence, its closure $\mathcal{M} := \overline{R(A)}$ is
$$\mathcal{M} = \{\sum_{n=1}^{\infty} b_n e_n \mid \sum_{n=1}^{\infty} |b_n|^2 < \infty\}.$$

8. We claim $N(A) = \mathcal{M}^{\perp}$. If $x \perp \mathcal{M}$, then $x \perp e_n$ for all $n \in \mathbb{N}$. Hence
$$Ax = \sum_{n=1}^{\infty} \lambda_n (x, e_n) e_n = 0.$$
Conversely, $Ax = 0$ implies $(x, e_n) = 0$ for all $n \in \mathbb{N}$. Hence, $x \perp \mathcal{M}$.

□

Singular Value Decomposition for Compact Operators We have seen that for bounded operator $A : \mathcal{H} \to \mathcal{K}$, the space $\mathcal{H}$ and $\mathcal{K}$ can be decomposed such that
$$A : N(A) \oplus \overline{R(A^*)} \to N(A^*) \oplus R(A)$$
and $A : \overline{R(A^*)} \to R(A)$ is 1-1 and onto. Below, we show that for compact operator, we can find orthonormal bases in $\overline{R(A^*)}$ and $\overline{R(A)}$ such that A can be represented as a diagonal matrix. Such a decomposition is important in image processing, inverse problems,etc.

Theorem 8.5 (Singular Value Decomposition of Compact Operators)**.** *Let $A : \mathcal{H} \to \mathcal{K}$ be a compact operator, where $\mathcal{H}$ and $\mathcal{K}$ are Hilbert spaces. Then there exist an orthonormal bases $\{u_i\}$ in $N(A)^{\perp}$ and $\{v_i\}$ in $N(A^*)^{\perp}$ and $\mu_i > 0$ such that*
$$Au_i = \mu_i v_i$$

Proof. 1. The operator $B = A^*A$ is a compact, self-adjoint operator in $\mathcal{H}$. Hence, B has eigenvalues λ_i and eigenvectors u_i such that (u_i) constitutes an orthonormal basis of $N(B)^\perp$.

2. We notice that
$$\lambda_i(u_i, u_i) = (A^*Au_i, u_i) = (Au_i, Au_i) > 0.$$
so we can define $\mu_i = \sqrt{\lambda_i}$.

3. We also notice that $A^*Ax = 0 \Rightarrow (A^*Ax, x) = 0 \Rightarrow (Ax, Ax) = 0$, hence $N(A^*A) \subset N(A)$. On the other hand, if $Ax = 0$, then $A^*Ax = 0$. Thus, $N(A) \subset N(A^*A)$. We get $N(B) = N(A^*A) = N(A)$ and
$$\overline{R(A^*)} = N(A)^\perp = N(B)^\perp.$$
Thus, $\{u_i\}$ constitutes an orthonormal basis in $N(A)^\perp$.

4. Define
$$v_i = \frac{1}{\mu_i} Au_i.$$
We have $Au_i = \mu_i v_i$,
$$A^*v_i = A^*Au_i/\mu_i = \lambda_i/\mu_i u_i = \mu_i u_i.$$
From
$$(v_i, v_j) = \frac{1}{\mu_i \mu_j}(Au_i, Au_j) = \frac{1}{\mu_i \mu_j}(A^*Au_i, u_j) = \frac{\mu_i}{\mu_j}(u_i, u_j)$$
we see that (v_i) is an orthonormal set in $R(A)$. In fact,
$$R(A) = \left\{ \sum_{n=1}^\infty \mu_n a_n v_n \,\Big|\, \sum_{n=1}^\infty |a_n|^2 < \infty \right\} = \left\{ b \in N(A^*)^\perp \Big| \sum_{n=1}^\infty |\mu_n|^{-2} |(b, v_n)|^2 < \infty \right\}$$
and $\mu_n \to 0$, we get that
$$\overline{R(A)} = \left\{ \sum_{n=1}^\infty b_n v_n \,\Big|\, \sum_{n=1}^\infty |b_n|^2 < \infty \right\}.$$

□

8.5 Ill-posed problems

Let A be a compact operator from $\mathcal{H}$ to $\mathcal{K}$. We are interested to solve the problem
$$Ax = b.$$

In application, for example, the operator A is a blur operator in image processing, the Radon transform in computed tomography, etc. In general, we call A the sensing operator, b the measured data and x the data to be restored. The measured data should be $b \in R(A)$, otherwise there is no solution. Unfortunately, the measured data usually contains noise, namely, we collect $b^\delta = b + n$, where n is a noise. We assume $\|n\| \leq \delta$ in our Hilbert space. Our goals are

- To solve $Ax = b$ with $b \in R(A)$ (noise free problem).
- To solve $Ax^\delta = b^\delta$ with $\|b^\delta - b\| \le \delta$
- To find least square solution for general $b^\delta \in \mathcal{K}$.

In many situations, the sensing operator is compact because the sensing process is sort of averaging process. In this case, the singular value μ_n of A tends to 0. This is the source of ill-posedness. Namely, a small perturbation of b causes a large variation of x.

Noise free problem Let A have singular value decomposition $(\mu_i, u_i, v_i)_{i=1}^\infty$. We have seen that $b \in R(A)$ if and only if

$$\sum_{i=1}^\infty \mu_i^{-2} |(b, v_i)|^2 < \infty.$$

The solution $x^\dagger$ is

$$x^\dagger = \sum_{i=1}^\infty \mu_i^{-1}(b, v_i) u_i \in N(A)^\perp \subset \mathcal{H}.$$

We denote $x^\dagger$ by $A^\dagger b$, called the pseudo-inverse of A.

Noisy problem The above solution is not stable under small perturbation of b. Suppose the perturbation is $b^\delta = b + \delta v_n$. Then $A^\dagger b^\delta = A^\dagger b + \delta/\mu_n u_n$. We see that $A^\dagger x^\delta \to \infty$ as $n \to \infty$ because $\mu_n \to 0$. Thus, $A^\dagger$ is not bounded on $R(A)$.

We look for least squares solution

$$\inf \|Ax - b\|^2.$$

As we have seen before that this least squares solution exists if and only if

$$b = \hat{b} + b^\perp \in R(A) + N(A^*),$$

or if and only if $\hat{b} \in R(A)$. The corresponding least squares solution is $A^\dagger \hat{b}$. For simplicity, we assume $N(A^*) = \{0\}$. Otherwise, we just replace b by $\hat{b}$.

We are given $b \in \overline{R(A)}$. We know that $\inf \|Ax - b\|^2$ has no solution unless $b \in R(A)$. Now, instead, we consider the regularized problem

$$\inf F_\alpha(x) := \alpha\|x\|^2 + \|Ax - b\|^2, \alpha > 0.$$

This is a convex optimization problem. The functional F_α is strictly convex, ontinuous, and coercive. Thus, it has a unique minimum in $\mathcal{H}$. Alternatively, the corresponding Euler-Lagrange equation is

$$\alpha x + A^* A x = A^* b.$$

This equation is called the normal equation. The operator $T_\alpha := \alpha I + A^* A$ is self-adjoint, positively definite, with eigenvalues bounded above 0. Thus, T_α has bounded inverse in $\mathcal{H}$. Let us denote it by $(\alpha I + A^* A)^{-1}$. The corresponding solution

$$x_\alpha := (\alpha I + A^* A)^{-1} A^* b.$$

In terms of singular value decomposition of A, we project the normal equation into u_i:

$$\left(\alpha + \mu_i^2\right)(x_\alpha, u_i) u_i = \mu_i (b, v_i) u_i.$$

We obtain

$$(x_\alpha, u_i) = \frac{\mu_i}{\alpha + \mu_i^2}(b, v_i).$$

Thus, the regularized solution x_α is given by

$$x_\alpha = \sum_{i=1}^{\infty} \frac{\mu_i}{\alpha + \mu_i^2}(b, v_i) u_i.$$