EM11 Functional Analysis

Version 0.9.3

Hong-Yan ZHANG
hongyan.siae@gmail.com

Sino-european Institute of Aviation Engineering (SIAE)
Civil Aviation University of China (CAUC)
Tianjing 300300, P. R. China

November 4, 2014

Contents

Part I

Preliminaries

Chapter 1

Measure and Integral

1.1 Monotone Functions

1.1.1 Fundamental Results

Let f be an increasing function defined on the real line $(-\infty, +\infty)$. Thus for any two real numbers x_1 and x_2,

$$x_1 < x_2 \Longrightarrow f(x_1) \leq f(x_2). \tag{1.1}$$

The notation "$t \uparrow x$" means "$t < x, t \to x$" and "$t \downarrow x$" means "$t > x, t \to x$".

There are some properties for the increasing function as follows:

① For each x, both **unilateral limits**

$$\lim_{t \uparrow x} f(t) = f(x-), \quad \lim_{t \downarrow x} f(t) = f(x+) \tag{1.2}$$

exist and are finite. Furthermore the **limits at infinity**

$$\lim_{t \downarrow -\infty} f(t) = f(-\infty), \quad \lim_{t \uparrow +\infty} f(t) = f(+\infty)$$

exist; the former may be $-\infty$, the latter may be $+\infty$.[1] This follows from monotonicity; indeed

$$f(x-) = \sup_{-\infty < t < x} f(t),$$
$$f(x+) = \sup_{x < t < +\infty} f(t).$$

② For each x, f is continuous iff

$$f(x-) = f(x) = f(x+).$$

③ The only possible kind of discontinuity of an increasing function is a jump. If there is a jump at x, we call x a **point of jump** of f and the number $f(x+) - f(x-)$ the **size of the jump** or simply "the jump" at x.

④ The set of discontinuities of f is countable.

⑤ Let f_1 and f_2 be two increasing functions and D a set that is (everywhere) dense in $(-\infty, +\infty)$. Suppose that

$$\forall x \in D : f_1(x) = f_2(x),$$

then f_1 and f_2 have the same points of jump of the same size, and they coincide except possibly at some of these points of jump.

[1] Usually, the $+\infty$ may be denoted as ∞ for simplicity.

⑥ If we put

$$\forall x : \hat{f} = f(x+),$$

then $\hat{f}$ is increasing and right continuous everywhere.

⑦ Let f be increasing on D, and define $\hat{f}$ on $(-\infty, +\infty)$ as follows:

$$\forall x : \hat{f}(x) = \inf_{x<t\in D} f(t).$$

Then $\hat{f}$ is increasing and right continuous everywhere.

Let D be dense in $(-\infty, +\infty)$, and suppose that f is a function with the domain D. We may speak of the monotonicity, continuity, uniform continuity, and so on of f on its domain of definition if in the usual definitions we restrict ourselves to the points of D. Even if f is defined in a larger domain, we may still speak of these properties "on D" by considering the "restriction of f to D".

1.1.2 Double Limit Lemma

Lemma 1. *Let $\{C_{jk} : j \in \mathbb{N}, k \in \mathbb{N}\}$ be a doubly indexed array of real numbers with the following properties:*

- *for each fixed j, the sequence $\{C_{jk} : k \in \mathbb{N}\}$ is increasing in k;*
- *for each fixed k, the sequence $\{C_{jk} : j \in \mathbb{N}\}$ is increasing in j.*

Then we have

$$\lim_j \uparrow \lim_k \uparrow C_{jk} = \lim_k \uparrow \lim_j \uparrow C_{jk} \leq +\infty.$$

1.2 Classes of Sets

1.2.1 Notations

Let Ω be an "abstract space", namely a nonempty set of elements to be called "points" and denoted generically by ω. Some of the usual operations and relations between sets, together with the usual notation, are given below

Table 1.1: Concepts for Set Theory

Jargon	**Notation**	**Remark**
universal set (abstract space)	Ω	全集(抽象空间)
empty set	$\emptyset$	空集
union	$E \cup F, \quad \bigcup_n E_n$	并
intersection	$E \cap F = EF, \quad \bigcap_n E_n$	交
complement	$E^c = \Omega \setminus E$	补
difference	$E \setminus F = E \cap F^c = E - F$	差
symmetric difference	$E\Delta F = (E \setminus F) \cup (F \setminus E)$	对称差
singleton (point)	$\{\omega\}$	单点集合
containing	$E \subset F, \quad F \supset E$	not excluding $E = F$
	$\mathcal{A} \subset \mathcal{B}, \quad \mathcal{B} \supset \mathcal{A}$	not excluding $\mathcal{A} = \mathcal{B}$
belonging	$\omega \in E$	for element in a set
	$E \in \mathcal{A}$	for set in a collection

A nonempty collection $\mathcal{A}$ of subsets of Ω may have certain "closure properties". Let us list some of those used below; note that *j is always an index for a countable set and that commas as well as semicolons are used to denote "conjunctions" of premises.*

(i) $E \in \mathcal{A} \Longrightarrow E^c \in \mathcal{A}$.

(ii) $E_1 \in \mathcal{A}, E_2 \in \mathcal{A} \Longrightarrow E_1 \cup E_2 \in \mathcal{A}$.

(iii) $E_1 \in \mathcal{A}, E_2 \in \mathcal{A} \Longrightarrow E_1 \cap E_2 \in \mathcal{A}$.

(iv) $\forall n \geq 2 : E_j \in \mathcal{A}, 1 \leq j \leq n \Longrightarrow \bigcup_{j=1}^{n} E_j \in \mathcal{A}$.

(v) $\forall n \geq 2 : E_j \in \mathcal{A}, 1 \leq j \leq n \Longrightarrow \bigcap_{j=1}^{n} E_j \in \mathcal{A}$.

(vi) $E_j \in \mathcal{A}; E_j \subset E_{j+1}, 1 \leq j < \infty \Longrightarrow \bigcup_{j=1}^{\infty} E_j \in \mathcal{A}$.

(vii) $E_j \in \mathcal{A}; E_j \supset E_{j+1}, 1 \leq j < \infty \Longrightarrow \bigcap_{j=1}^{\infty} E_j \in \mathcal{A}$.

(viii) $E_j \in \mathcal{A}, 1 \leq j < \infty \Longrightarrow \bigcup_{j=1}^{\infty} E_j \in \mathcal{A}$.

(ix) $E_j \in \mathcal{A}, 1 \leq j < \infty \Longrightarrow \bigcap_{j=1}^{\infty} E_j \in \mathcal{A}$.

(x) $E_1 \in \mathcal{A}, E_2 \in \mathcal{A}, E_1 \subset E_2 \Longrightarrow E_2 \setminus E_1 \in \mathcal{A}$.

It follows from simple set algebra that under (i):

- (ii) $\Longleftrightarrow$ (iii)
- (vi) $\Longleftrightarrow$ (vii)
- (viii) $\Longleftrightarrow$ (ix)
- (ii) implies (iv) by induction
- (iii) implies (v) by induction
- (viii) implies (ii) and (vi)
- (ix) implies (iii) and (vii)

Definition 2. *A nonempty collection $\mathcal{F}$ of subsets of Ω is called a **field** or **algebra** iff (i) and (ii) hold. It is called a **monotone class** (M.C.) iff (vi) and (vii) hold. It is called a **Borel field** (B.F.) or σ-field iff (i) and (viii) hold.*

Theorem 3. *A field is a B.F. if and only if it is also an M.C.*

Proof.

- The "only if" part is trivial.

Table 1.2: Comparison of Field, B.F. and M.C.

Concept	**Requirements**	**Remark**
field/algebra	$E \in \mathcal{A} \Longrightarrow E^c \in \mathcal{A}$ $E_1 \in \mathcal{A}, E_2 \in \mathcal{A} \Longrightarrow E_1 \cup E_2 \in \mathcal{A}$	域/代数
monotone class	$E_j \in \mathcal{A}; E_j \subset E_{j+1}, 1 \le j < \infty \Longrightarrow \bigcup_{j=1}^{\infty} E_j \in \mathcal{A}$ $E_j \in \mathcal{A}; E_j \supset E_{j+1}, 1 \le j < \infty \Longrightarrow \bigcap_{j=1}^{\infty} E_j \in \mathcal{A}$	单调类
Borel field (σ-field/algebra)	$E \in \mathcal{A} \Longrightarrow E^c \in \mathcal{A}$ $E_j \in \mathcal{A}, 1 \le j < \infty \Longrightarrow \bigcup_{j=1}^{\infty} E_j \in \mathcal{A}$	Borel 域 (σ-域/代数)

- To prove the "if" part, we show that (iv) and (vi) imply (viii). Let $E_j \in \mathcal{A}$ for $1 \le j < \infty$, then
$$F_n = \bigcup_{j=1}^{n} E_j \in \mathcal{A}$$
by (iv), which holds in a field, $F_n \subset F_{n+1}$ and
$$\bigcup_{j=1}^{\infty} E_j = \bigcup_{j=1}^{\infty} F_j,$$
hence $\bigcup_{j=1}^{\infty} \in \mathcal{A}$ by (vi).

1.2.2 Inverse Image

Let $f : \Omega \to \bar{\Omega}$ be a mapping, consider the "inverse mapping" $f^{-1} : \bar{\Omega} \to \Omega$, defined as follows:

$$\forall A \subset \bar{\Omega} : f^{-1}(A) = \{\omega \in \Omega : f(\omega) \in B\}$$

Theorem 4. *For any function $f : \Omega \to \bar{\Omega}$, the inverse mapping has the following properties:*

- $f^{-1}(A^c) = \left(f^{-1}(A)\right)^c$
- $f^{-1}\left(\bigcup_\alpha A_\alpha\right) = \bigcup_\alpha f^{-1}(A_\alpha)$
- $f^{-1}\left(\bigcap_\alpha A_\alpha\right) = \bigcap_\alpha f^{-1}(A_\alpha)$

1.2.3 Borel's Covering Theorem

Theorem 5 (Heine-Borel)**.** *Let $[a, b]$ be a compact interval, and $(a_j, b_j), j \in \mathbb{N}$, be bounded open intervals, which may intersect arbitrarily, such that*

$$[a, b] \subset \bigcup_{j=1}^{\infty} (a_j, b_j). \tag{1.3}$$

Then there exists a finite integer ℓ such that when ℓ is substituted for ∞ in the above, the inclusion remains valid, i.e.,

$$\exists \ell, 1 \le \ell < \infty, \quad [a, b] \subset \bigcup_{j=1}^{\ell} (a_j, b_j).$$

In other words, a finite subset of the original infinite set of open intervals suffices to do the covering. The Heine-Borel theorem is also called Borel's covering theorem.

1.2.4 Indicator

Definition 6. *For each $\Delta \subset \Omega$, the function $\mathbb{1}_\Delta(\cdot)$ defined as follows:*

$$\forall \omega \in \Omega : \quad \mathbb{1}_\Delta(\omega) = \begin{cases} 1, & \text{if } \omega \in \Delta; \\ 0, & \text{if } \omega \in \Omega \setminus \Delta. \end{cases}$$

*is called the **indicator** or **indicator function** of Δ.*

A countable partition of Ω is a countable family of disjoint sets $\{\Lambda_j\}$ with $\Lambda_j \in \mathcal{F}$ for each j and such that $\Omega = \bigcup_j \Lambda_j$. We have then

$$1 = \mathbb{1}_\Omega = \sum_j \mathbb{1}_{\Lambda_j}.$$

There are some interesting properties for the indicator function as follows:

-
-

1.2.5 Monotone Class Theorems

There are some related concepts and results:

① The collection $\mathcal{B}$ of all subsets of Ω is a B.F. called the **total Borel Field**.

② The collection of the two sets $\{\emptyset, \Omega\}$ is a B.F. called the **trivial Borel Field**.

③ If A is any index set and if for every $\alpha \in A$, $\mathcal{F}_\alpha$ is a B.F. (or M.C.) then the intersection

$$\bigcap_{\alpha \in A} \mathcal{F}_\alpha$$

of all these B.F.'s (or M.C.'s), namely the collection of sets each of which belongs to all $\mathcal{F}_\alpha$, is also a B.F. (or M.C.).

④ Given any nonempty collection $\mathcal{C}$ of sets, there is a **minimal** B.F. (or field, or M.C.) containing it; this is just the intersection of all B.F.'s (or fileds, or M.C.'s) containing $\mathcal{C}$, of which there is at least one, namely the total B.F. $\mathcal{B}$. The minimal B.F. (or field, or M.C.) is also said to be **generated by** $\mathcal{C}$.

- In particular if $\mathcal{F}_0$ is a field there is a minimal B.F. (or M.C.) containing $\mathcal{F}_0$.

Theorem 7. *Let $\mathcal{F}_0$ be a field, $\mathcal{G}$ the minimal M.C. containing $\mathcal{F}_0$, $\mathcal{F}$ the minimal B.F. containing $\mathcal{F}_0$, then $\mathcal{F} = \mathcal{G}$.*

Proof.

- Since a B.F. is an M.C., we have $\mathcal{F} \supset \mathcal{G}$. To prove $\mathcal{F} \subset \mathcal{G}$ it is sufficient to show that $\mathcal{G}$ is a B.F. Hence by Theorem 3 it is sufficient to show that $\mathcal{G}$ is a field. We shall show that it is closed under intersection and complementation. Define two classes of subsets of $\mathcal{G}$ as follows:

$$\mathcal{C}_1 = \{E \in \mathcal{G} : E \cap F \in \mathcal{G}, \forall F \in \mathcal{F}_0\}$$
$$\mathcal{C}_2 = \{E \in \mathcal{G} : E \cap F \in \mathcal{G}, \forall F \in \mathcal{F}\}$$

The identities

$$F \cap \left(\bigcup_{j=1}^{\infty} E_j \right) = \bigcup_{j=1}^{\infty} (F \cap E_j)$$
$$F \cap \left(\bigcap_{j=1}^{\infty} E_j \right) = \bigcap_{j=1}^{\infty} (F \cap E_j)$$

show that both $\mathcal{C}_1$ and $\mathcal{C}_2$ are M.C.'s. Since $\mathcal{F}_0$ is closed under intersection and contained in $\mathcal{G}$, it is clear that $\mathcal{F}_0 \subset \mathcal{C}_1$. Hence $\mathcal{G} \subset \mathcal{C}_1$ by the minimality of $\mathcal{G}$ and so $\mathcal{G} = \mathcal{C}_1$. This means for any $F \subset \mathcal{F}_0$ and $E \in \mathcal{G}$ we have $F \cap E \in \mathcal{G}$, which in turn means $\mathcal{F}_0 \subset \mathcal{C}_2$. Hence $\mathcal{G} = \mathcal{C}_2$ and this means $\mathcal{G}$ is closed under intersection.

- Next, define another class of subsets of $\mathcal{G}$ as follows:

$$\mathcal{C}_3 = \{E \in \mathcal{G} : E^c \in \mathcal{G}\}$$

The (DeMorgan) identities

$$\left(\bigcup_{j=1}^{\infty} E_j \right)^c = \bigcap_{j=1}^{\infty} E_j^c$$
$$\left(\bigcap_{j=1}^{\infty} E_j \right)^c = \bigcup_{j=1}^{\infty} E_j^c$$

show that $\mathcal{C}_3$ is a M.C. Since $\mathcal{F}_0 \subset \mathcal{C}_3$, it follows as before that $\mathcal{G} = \mathcal{C}_3$, which means $\mathcal{G}$ is closed under complementation. ■

Corollary 8. *Let $\mathcal{F}_0$ be a field, $\mathcal{F}$ the minimal B.F. containing $\mathcal{F}_0$; $\mathcal{C}$ a class of sets containing $\mathcal{F}_0$ and having the closure properties (vi) and (vii), then $\mathcal{C}$ contains $\mathcal{F}$.*

The theorem above is one of a type called monotone class theorems. They are among the most useful tools of measure theory, and serve to extend certain relations which are easily verified for a special class of sets or functions to a larger class. Many versions of such theorems are known, as given below.

Theorem 9. *Let $\mathcal{D}$ be a class of subsets of Ω having the closure property (iii); let $\mathcal{A}$ be a class of sets containing Ω as well as $\mathcal{D}$, and having the closure properties (vi) and (x). Then $\mathcal{A}$ contains the B.F. generated by $\mathcal{D}$.*

HINT: This is the Dynkin's form of a monotone class theorem which is expedient for certain applications. The proof as in Theorem 7 by replacing $\mathcal{F}_0$ and $\mathcal{G}$ with $\mathcal{D}$ and $\mathcal{A}$ respectively, viz., putting $\mathcal{F}_0 \longleftarrow \mathcal{D}$ and $\mathcal{G} \longleftarrow \mathcal{A}$.

Theorem 10. *Take $\Omega = \mathbb{R}^{n\times 1}$ or a separable metric space in Theorem 9 and let $\mathcal{D}$ be the class of all open sets. Let $\mathcal{H}$ be a class of real-valued functions on Ω satisfying the following conditions.*

(a) $1 \in \mathcal{H}$ *and* $\mathbb{1}_D \in \mathcal{H}$ *for each* $D \in \mathcal{D}$.

(b) $\mathcal{H}$ *is a vector space, namely: if* $f_1 \in \mathcal{H}, f_1 \in \mathcal{H}$ *and* $c_1 \in \mathbb{R}, c_2 \in \mathbb{R}$ *are any two real constants, then* $c_1 f_1 + c_2 f_2 \in \mathcal{H}$;

(c) $\mathcal{H}$ *is closed w.r.t.*[2] *increasing limits of positive functions, namely: if* $f_n \in \mathcal{H}, 0 \leq f_n \leq f_{n+1}$ *for all* n, *and* $f = \lim_{n\uparrow\infty} f_n < \infty$, *then* $f \in \mathcal{H}$.

[2] w.r.t.: with respect to

Then $\mathcal{H}$ contains all Borel measurable functions on Ω, namely all finite-valued functions measurable w.r.t. the topological Borel field (=the minimal B.F. containing all open sets of Ω).

HINT: Let $\mathcal{C} = \{E \subset \Omega : \mathbb{1}_E \in \mathcal{H}\}$, apply Theorem 9 to show that $\mathcal{C}$ contains the B.F. just defined. Each positive Borel measurable function is the limit of an increasing sequence of simple (finite-valued) functions.

Theorem 11. *Let $\mathcal{C}$ be a M.C. of subsets of $\mathbb{R}^{n\times 1}$ (or a separable metric space) containing all the open sets and closed sets, then $\mathcal{C} \supset \mathbb{R}^{n\times 1}$.*

HINT: In Theorem 10 taking the $\mathbb{R}^{n\times 1}$ as the topological Borel field, show that the minimal such class is a field.

1.3 Construction of Measure

1.3.1 What's Measure

Let Ω be an abstract space and $\mathcal{T}$ its Borel field, then $A \in \mathcal{T}$ means $A \subset \Omega$.

Definition 12 (Outer Measure)**.** *A function μ^* with domain $\mathcal{T}$ and range in $[0,\infty]$ is an* OUTER MEASURE *iff the following properties hold:*

❶ $\mu^*(\emptyset) = 0$;

❷ *(monotonicity) if $A_1 \subset A_2$, then $\mu^*(A_1) \le \mu^*(A_2)$;*

❸ *(subadditivity) if $\{A_j\}$ is a countable sequence of sets in $\mathcal{T}$, then*

$$\mu^*\left(\bigcup_j A_j\right) \le \sum_j \mu^*(A_j).$$

Definition 13 (Measure)**.** *Let $\mathcal{F}_0$ be a field in Ω. A function μ with domain $\mathcal{F}_0$ and range in $[0,\infty]$ is a* MEASURE *on $\mathcal{F}_0$ iff* ❶ *and the following property hold:*

❹ *(additivity) if $\{B_j\}$ is a countable sequence of disjoint sets in $\mathcal{F}_0$ and $\bigcup_j B_j \in \mathcal{F}_0$, then*

$$\mu\left(\bigcup_j B_j\right) = \sum_j \mu(B_j). \tag{1.4}$$

Let us show that the properties ❷ and ❸ for outer measure hold for a measure μ, provide all the sets involved belong to $\mathcal{F}_0$.

- If $A_1 \in \mathcal{F}_0, A_2 \in \mathcal{F}_0$, and $A_1 \subset A_2$, then $A_1^c A_2 \in \mathcal{F}_0$ because $\mathcal{F}_0$ is a field; $A_2 = A_1 \cup A_1^c A_2$ and so by ❹ and ❶:

  $$\mu(A_2) = \mu(A_1) + \mu(A_1^c A_2) \ge \mu(A_1) + 0 = \mu(A_1).$$

  Hence property ❷ holds.

- Next, if each $A_j \in \mathcal{F}_0$, and furthermore if $\bigcup_j A_j \in \mathcal{F}_0$ (this must be assumed for a countably infinite union because it is not implied by the definition of a field!), then

  $$\bigcup_j A_j = A_1 \cup A_1^c A_2 \cup A_1^c A_2^c A_3 \cup \cdots$$

and so by ❹, since each member of the disjoint union above belongs to $\mathcal{F}_0$:

$$\mu\left(\bigcup_j A_j\right) = \mu(A_1) + \mu(A_1^c A_2) + \mu(A_1^c A_2^c A_3) + \cdots$$
$$\leq \mu(A_1) + \mu(A_2) + \mu(A_3) + \cdots$$

be property ❷ just proved.

The symbol $\mathbb{N}$ denotes the sequence of natural numbers (strictly positive integers); when used as index set, it will frequently be omitted as understood. For instance, the index j used above ranges over $\mathbb{N}$ or a finite segment of $\mathbb{N}$.

Now let us suppose that the field $\mathcal{F}_0$ is a Borel field to be denoted by $\mathcal{F}$ and the $\mu : \mathcal{F} \to [0, \infty]$ is a measure. Then if $A_n \in \mathcal{F}$ for each $n \in \mathbb{N}$, the countable union $\bigcup\limits_n A_n$ and coubtable intersection $\bigcap\limits_n A_n$ both belong to $\mathcal{F}$. In this case we have the following fundamental properties.

❺ (increasing limit) if $A_n \subset A_{n+1}$ for all n and $A_n \uparrow A = \bigcup_n A_n$, then

$$\lim_n \uparrow \mu(A_n) = \mu(A).$$

❻ (decreasing limit) if $A_n \supset A_{n+1}$ for all n and $A_n \downarrow A = \bigcup_n A_n$, and for some n we have $\mu(A_n) < \infty$, then

$$\lim_n \downarrow \mu(A_n) = \mu(A).$$

The additional assumption in ❻ is essential. For a counterexample let $A_n = (n, \infty) \subset \mathbb{R}$, then $A_n \downarrow \emptyset$, but the measure (length!) of A_n is ∞ for all n, while $\emptyset$ surely must have measure 0. It can even be made discrete if we use the **counting measure** $|A|$ of natural numbers[3] in set A: let $A_n = \{n, n+1, n+2, \cdots, \cdots\}$ so that $|A_n| = \infty, \left|\bigcap\limits_n A_n\right| = 0$.

The measure μ is said to be **σ-finite** if there is a sequence of sets $A_n \in \mathcal{F}$ so that $\mu(A_n) < \infty$ and $\bigcup\limits_n A_n = \Omega$.

1.3.2 Extension of Measure

Beginning with a measure μ on a field $\mathcal{F}_0$, not a Borel field, we proceed to construct a measure on the Borel field $\mathcal{F}$ generated by $\mathcal{F}_0$, namely the minimal Borel field containing $\mathcal{F}_0$. This is called an extension of μ from $\mathcal{F}_0$ to $\mathcal{F}$, when the notation μ is maintained. Curiously, we do this by first constructing an an outer measure μ^* on the total Borel field $\mathcal{T}$ and then showing that μ^* is in truth a measure on a certain Borel field to be denoted by $\mathcal{F}^*$ that contains $\mathcal{F}_0$. Then of course $\mathcal{F}^*$ must contain the minimal $\mathcal{F}$, and so μ^* restricted to $\mathcal{F}$ is an extension of the original μ from $\mathcal{F}_0$ to $\mathcal{F}$. But we have obtained a further extension to $\mathcal{F}^*$ that is in general "larger" than $\mathcal{F}$ and possesses a further desirable property to be discussed.

Definition 14. *Given a measure μ on a field $\mathcal{F}_0$ in Ω, we define μ^* on $\mathcal{T}$ as follows, for any $A \in \mathcal{T}$:*

$$\mu^*(A) = \inf\left\{\sum_j \mu(B_j) : B_j \in \mathcal{F}_0 \textit{for all } j \textit{ and } \bigcup_j B_j \supset A\right\}. \tag{1.5}$$

[3]Note that another notation for the cardinality of a set A is $\#(A)$. Both $\#(A)$ and $|A|$ are frequently encountered in different references.

A countable (possibly finite) collection of sets $\{B_j\}$ satisfying the conditions indicated in (1.5) will be referred to below as a "covering" of $\{A\}$. The infimum inf taken over all such coverings exists because the single set Ω constitutes a covering of A, so that

$$0 \leq \mu^*(A) \leq \mu^*(\Omega) \leq \infty.$$

It is not trivial that $\mu^*(A) = \mu(A)$ if $A \in \mathcal{F}_0$, which is part of the next theorem.

Theorem 15. *The restriction of μ^* on $\mathcal{F}_0$ is μ, i.e. $\mu^*|_{\mathcal{F}_0} = \mu$, and μ^* on $\mathcal{T}$ is an outer measure.*

PROOF.

- Let $A \in \mathcal{F}_0$, then the single set A serves as a covering of A; hence $\mu^*(A) \leq \mu(A)$. For any covering $\{B_j\}$ of A, we have $AB_j \in \mathcal{F}_0$ and

$$\bigcup_j AB_j = A \in \mathcal{F}_0.$$

  Thus by property ❸ of μ on $\mathcal{F}_0$ followed by property ❷:

$$\mu(A) = \mu\left(\bigcup_j Ab_j\right) \leq \sum_j \mu(AB_j) \leq \sum_j \mu(B_j).$$

  Therefore

$$\mu(A) \leq \inf\left\{\sum_j \mu(B_j) : ...\right\}$$

  which shows that $\mu(A) \leq \mu^*(A)$ by (1.5). Thus $\mu^* = \mu$ on $\mathcal{F}_0$.

- To prove μ^* is an outer measure, the properties ❶ and ❷ are trivial. To prove ❸, let $\varepsilon > 0$. For each j, by the definition of $\mu^*(A_j)$, there exists a covering $\{B_{jk}\}$ of A_j such that

$$\sum_k \mu(B_{jk}) \leq \mu^*(A_j) + \frac{\varepsilon}{2^j}.$$

  The double sequence $\{B_{jk}\}$ is a covering of $\bigcup_j A_j$ such that

$$\sum_j \sum_k \mu(B_{jk}) \leq \sum_j \mu^*(A_j) + \varepsilon.$$

  Hence for any $\varepsilon > 0$:

$$\mu^*\left(\bigcup_j A_j\right) \leq \sum_j \mu^*(A_j) + \varepsilon$$

  that establishes ❸ for μ^*, since ε is arbitrarily small.

With the outer measureμ^*, a class of sets $\mathcal{F}^*$ is associated as follows.

Definition 16. *A set $A \subset \Omega$ belongs to $\mathcal{F}^*$ iff for every $Z \subset \Omega$ we have*

$$\mu^*(Z) = \mu^*(AZ) + \mu^*(A^cZ). \tag{1.6}$$

If in (1.6) we change "=" into "$\leq$", the resulting inequality holds by ❸; hence (1.6) is equivalent to the reverse inequality when "=" is changed into "$\geq$".

Theorem 17. *$\mathcal{F}^*$ is a Borel field and contains $\mathcal{F}_0$. On $\mathcal{F}^*$, μ^* is a measure.*

PROOF.

- Let $A \in \mathcal{F}_0$. For any $Z \subset \Omega$ and any $\varepsilon > 0$, there exists a covering $\{B_j\}$ of Z such that

$$\sum_j \mu(B_j) \leq \mu^*(Z) + \varepsilon. \tag{1.7}$$

Since $AB_j \in \mathcal{F}_0$, $\{AB_j\}$ is a covering of AZ; $\{A^cB_j\}$ is a covering of A^cZ; hence

$$\mu^*(AZ) \leq \sum_j \mu(AB_j), \quad \mu^*(A^cZ) \leq \sum_j \mu(A^cB_j). \tag{1.8}$$

Since μ is a measure on $\mathcal{F}_0$, we have for each j:

$$\mu(AB_j) + \mu(A^cB_j) = \mu(B_j). \tag{1.9}$$

It follows from (1.7), (1.8) and (1.9) that

$$\mu^*(AZ) + \mu^*(A^cZ) \leq \mu^*(Z) + \varepsilon.$$

Letting $\varepsilon \downarrow 0$ establishes the criterion (1.6) in its "$\geq$" form. Thus $A \in \mathcal{F}^*$, and we have proved that $\mathcal{F}_0 \subset \mathcal{F}^*$.

- To prove that $\mathcal{F}^*$ is a Borel field, it is trivial that it is closed under complementation because the criterion (1.6) is unaltered when A is changed into A^c. Next, to show that $\mathcal{F}^*$ is closed under union, let $A \in \mathcal{F}^*$ and $B \in \mathcal{F}^*$. Then for any $Z \subset \Omega$, we have by (1.6) with A replaced by B and Z replaced by ZA or ZA^c:

$$\begin{aligned}\mu^*(ZA) &= \mu^*(ZAB) + \mu^*(ZAB^c)\\ \mu^*(ZA^c) &= \mu^*(ZA^cB) + \mu^*(ZA^cB^c).\end{aligned}$$

Hence by (1.6) again as written:

$$\mu^*(Z) = \mu^*(ZAB) + \mu^*(ZAB^c) + \mu^*(ZA^cB) + \mu^*(ZA^cB^c).$$

Applying (1.6) with Z replaced by $Z(A \cup B)$, we have

$$\begin{aligned}\mu^*(Z(A \cup B)) &= \mu^*(Z(A \cup B)A) + \mu^*(Z(A \cup B)A^c)\\ &= \mu^*(ZA) + \mu^*(ZA^cB)\\ &= \mu^*(ZAB) + \mu^*(ZAB^c) + \mu^*(ZA^cB).\end{aligned}$$

Comparing the two proceeding equations, we see that

$$\mu^*(Z) = \mu^*(Z(A \cup B)) + \mu^*(Z(A \cup B)^c).$$

Hence $A \cup B \in \mathcal{F}^*$, and we have proved that $\mathcal{F}^*$ is a field.

- Now let $\{A_j\}$ be an infinite sequence of sets in $\mathcal{F}^*$; put

$$B_1 = A_1, \quad B_j = A_j \setminus \left(\bigcup_{i=1}^{j-1} A_i\right) = A_jA_{j-1}^cA_{j-2}^c \cdots A_2^cA_1^c, \qquad j \geq 2.$$

Then $\{B_j\}$ is a sequence of disjoint sets in $\mathcal{F}^*$ (because $\mathcal{F}^*$ is a field) and has the same union as $\{A_j\}$. For any $Z \subset \Omega$, we have for each $n \geq 1$:

$$\begin{aligned}\mu^*\left(Z\left[\bigcup_{j=1}^n B_j\right]\right) &= \mu^*\left(Z\left[\bigcup_{j=1}^n B_j\right]B_n\right) + \mu^*\left(Z\left[\bigcup_{j=1}^n B_j\right]B_n^c\right)\\ &= \mu^*(ZB_n) + \mu^*\left(Z\left[\bigcup_{j=1}^{n-1} B_j\right]\right)\end{aligned}$$

because $B_n \in \mathcal{F}^*$. It follows by induction on n that

$$\mu^*\left(Z\left[\bigcup_{j=1}^{n} B_j\right]\right) = \sum_{j=1}^{n} \mu^*(ZB_j) \tag{1.10}$$

Since $\bigcup_{j=1}^{n} B_j \in \mathcal{F}^*$, we have by (1.10) and the monotonicity of μ^*:

$$\begin{aligned}\mu^*(Z) &= \mu^*\left(Z\left[\bigcup_{j=1}^{n} B_j\right]\right) + \mu^*\left(Z\left[\bigcup_{j=1}^{n} B_j\right]^c\right) \\ &\geq \sum_{j=1}^{n} \mu^*(ZB_j) + \mu^*\left(\left[\bigcup_{j=1}^{\infty} B_j\right]^c\right)\end{aligned}$$

Letting $n \uparrow \infty$ and using property ❸ of μ^*, we obtain

$$\mu^*(Z) \geq \mu^*\left(Z\left[\bigcup_{j=1}^{\infty} B_j\right]\right) + \mu^*\left(Z\left[\bigcup_{j=1}^{\infty} B_j\right]^c\right)$$

that establishes $\bigcup_{j=1}^{\infty} B_j \in \mathcal{F}^*$. Thus $\mathcal{F}^*$ is a Borel field.

- Finally, let $\{B_j\}$ be a sequence of disjoint sets in $\mathcal{F}^*$. By the property ❷ of μ^* and (1.10) with $Z = \Omega$, we have

$$\mu^*\left(\bigcup_{j=1}^{\infty} B_j\right) \geq \limsup_n \mu^*\left(\bigcup_{j=1}^{n} B_j\right) = \lim_n \sum_{j=1}^{n} \mu^*(B_j) = \sum_{j=1}^{\infty} \mu^*(B_j).$$

Combined with the property ❸ of μ^*, we obtain the countable additivity of μ^* on $\mathcal{F}^*$, namely the property ❹ for a measure.

1.4 Characterization of Extensions

1.4.1 A

We have proved that

$$\mathcal{T} \supset \mathcal{F}^* \supset \mathcal{F} \supset \mathcal{F}_0,$$

where some of the "$\supset$" may turn out to be "$=$". Since we have extended the measure μ from $\mathcal{F}_0$ to $\mathcal{F}^*$ in Theorem 17, what for $\mathcal{F}$? The answer will appear in the sequel.

The $\langle \Omega, \mathcal{F}, \mu \rangle$ where $\mathcal{F}$ is a Borel field of subsets of Ω, and μ is a measure on $\mathcal{F}$, will be called a **measure space**. It is qualified by the adjective "finite" when $\mu(\Omega) < \infty$, and by the noun "probability" when $\mu(\Omega) = 1$.

A more general case is defined below.

Definition 18. *A measure μ on a field $\mathcal{F}_0$ (not necessarily Borel field) is said to be σ-finite iff there exists a sequence of sets $\{\Omega_n, n \in \mathbb{N}\}$ in $\mathcal{F}_0$ such that $\mu(\Omega_n) < \infty$ for each n, and $\bigcup_n \Omega_n = \Omega$. In this case the measure space $\langle \Omega, \mathcal{F}, \mu \rangle$, where $\mathcal{F}$ is the minimal Borel field containing $\mathcal{F}_0$, is said to be "σ-finite on $\mathcal{F}_0$".*

Theorem 19. *Let $\mathcal{F}_0$ be a field and $\mathcal{F}$ the Borel field generated by $\mathcal{F}_0$. Let μ_1 and μ_2 be two measures on $\mathcal{F}$ that agree on $\mathcal{F}_0$. If one of them, hence both are σ-finite on $\mathcal{F}_0$, then they agree on $\mathcal{F}$.*

PROOF.

- Let $\{\Omega_n\}$ be as in Definition 18. Define a class $\mathcal{C}$ of subsets of Ω as follows:

$$\mathcal{C} = \{A \subset \Omega : \mu_1(\Omega_n A) = \mu_2(\Omega_n A), \forall n \in \mathbb{N}\}$$

Since $\Omega_n \in \mathcal{F}_0$, for any $A \in \mathcal{F}_0$ we have $\Omega_n A \in \mathcal{F}_0$ for all n; hence $\mathcal{C} \supset \mathcal{F}_0$. Suppose $A_k \in \mathcal{C}, A_k \subset A_{k+1}$ for all $k \in \mathbb{N}$ and $A_k \uparrow A$. Then by property ❺ of μ_1 and μ_2 as measures on $\mathcal{F}$, we have for each n:

$$\mu_1(\Omega_n A) = \lim_n \uparrow \mu_1(\Omega_n A_k) = \lim_n \uparrow \mu_2(\Omega_n A_k) = \mu_2(\Omega_n A).$$

Thus $A \in \mathcal{C}$. Similarly by property ❻, and the hypothesis $\mu_1(\Omega_n) = \mu_2(\Omega_n) < \infty$, if $A_k \in \mathcal{C}$ and $A_k \downarrow A$, then $A \in \mathcal{C}$. Therefore $\mathcal{C}$ is closed under both increasing and decreasing limits; hence $\mathcal{C} \supset \mathcal{F}$ by Theorem 7. This implies for any $A \in \mathcal{F}$:

$$\mu_1(A) = \lim_n \uparrow \mu_1(\Omega_n A) = \lim_n \uparrow \mu_2(\Omega_n A) = \mu_2(A)$$

by property ❺ once again. Thus μ_1 and μ_2 agree on $\mathcal{F}$.

It follows from Theorem 19 that under the σ-finite assumption there, the outer measure μ^* in Theorem 17 restricted to the minimal Borel field $\mathcal{F}$ containing $\mathcal{F}_0$ is the unique extension of μ from $\mathcal{F}_0$ to $\mathcal{F}$. What about the more extensive extension to $\mathcal{F}^*$? We are going to prove that it is also unique when a further property is imposed on the extension. We begin by defining two classes of special sets in $\mathcal{F}$.

Definition 20. *Given the field $\mathcal{F}_0$ of sets in Ω, let $\mathcal{F}_0^{au}$ be the collection of all sets of the form $\bigcap_{m=1}^{\infty} \bigcup_{n=1}^{\infty} B_{mn}$ where each $B_{mn} \in \mathcal{F}_0$, and $\mathcal{F}_0^{ua}$ be the collection of all sets of the form $\bigcup_{m=1}^{\infty} \bigcap_{n=1}^{\infty} B_{mn}$ where each $B_{mn} \in \mathcal{F}_0$, i.e.,*

$$\mathcal{F}_0^{au} = \left\{ \bigcap_{m=1}^{\infty} \bigcup_{n=1}^{\infty} B_{mn} : B_{mn} \in \mathcal{F}_0 \right\}$$
$$\mathcal{F}_0^{ua} = \left\{ \bigcup_{m=1}^{\infty} \bigcap_{n=1}^{\infty} B_{mn} : B_{mn} \in \mathcal{F}_0 \right\}$$

where "au" corresponds to cap-cup $\cap\cup$ and "ua" corresponds to cup-cap $\cup\cap$.

Both these collections belong to $\mathcal{F}$ because the Borel field is closed under countable union and intersection, and these operations may be iterated, here twice only, for each collection. If $B \in \mathcal{F}_0$, then B belong to both $\mathcal{F}_0^{au}$ and $\mathcal{F}_0^{ua}$ because we can take $B_{mn} = B$. Finally, $A \in \mathcal{F}_0^{au}$ if and only if $A^c \in \mathcal{F}_0^{ua}$ because

$$\left[\bigcap_m \bigcup_n B_{mn}\right]^c = \bigcup_m \bigcap_n B_{mn}^c.$$

Theorem 21. *Let $A \in \mathcal{F}^*$. There exists $B \in \mathcal{F}_0^{au}$ such that*

$$A \subset B, \quad \mu^*(A) = \mu^*(B).$$

If μ is σ-finite on $\mathcal{F}_0$, then there exists $C \in \mathcal{F}_0^{ua}$ such that

$$C \subset A, \quad \mu^*(C) = \mu^*(A).$$

PROOF.

- For each m, there exists B_{mn} in $\mathcal{F}$ such that

$$A \subset \bigcup_n B_{mn}, \quad \sum_n \mu^*(B_{mn}) \leq \mu^*(A) + \frac{1}{m}.$$

Put

$$B_m = \bigcup_n B_{mn}, \quad B = \bigcap_m B_m;$$

then $A \subset B$ and $B \in \mathcal{F}_0^{au}$. We have

$$\mu^*(B) \leq \mu^*(B_m) \leq \sum_n \mu^*(B_{mn}) \leq \mu^*(A) + \frac{1}{m}.$$

Letting $m \uparrow \infty$ we see that $\mu^*(B) \leq \mu^*(A)$; hence $\mu^*(B) = \mu^*(A)$. The first assertion of the theorem is proved.

- To prove the second assertion, let Ω_n be as in Definition 18. Applying the first assertion to $\Omega_n A^c$, we have $B_n \in \mathcal{F}_0^{au}$ such that

$$\Omega_n A^c \subset B_n, \quad \mu^*(\Omega_n A^c) = \mu^*(B_n).$$

Hence we obtain

$$\Omega_n A^c \subset \Omega_n B_n, \quad \mu^*(\Omega_n A^c) = \mu^*(\Omega_n B_n).$$

Taking complements with respect to Ω_n, we have

$$\Omega_n A \supset \Omega_n B_n^c, \quad \mu^*(\Omega_n A) = \mu^*(\Omega_n) - \mu^*(\Omega_n A^c) = \mu^*(\Omega_n) - \mu^*(\Omega_n B_n) = \mu^*(\Omega_n B_n^c)$$

since $\mu^*(\Omega_n) < \infty$. Since $\Omega_n \in \mathcal{F}_0$ and $B_n^c \in \mathcal{F}_0^{ua}$, it is easy to verify that $\Omega_n B_n^c \in \mathcal{F}_0^{ua}$ by the distributive law for the intersection with a union. Put

$$C = \bigcup_n \Omega_n B_n^c,$$

it is trivial that $C \in \mathcal{F}_0^{ua}$ and

$$A = \bigcup_n \Omega_n A \supset C.$$

Consequently, we have

$$\mu^*(A) \geq \mu^*(C) \geq \liminf_n \mu^*(\Omega_n B_n^c) = \liminf_n \mu^*(\Omega_n A) = \mu^*(A),$$

the last equation owing to property ❺ of the measure $\mu*$. Thus $\mu^*(A) = \mu^*(C)$, and the assertion is proved.

The measure μ^* on $\mathcal{F}^*$ is constructed from the measure μ on the field $\mathcal{F}_0$. The restriction of μ^* to the minimal Borel field $\mathcal{F}$ containing $\mathcal{F}_0$ will henceforth be denoted by μ instead of μ^*.

1.4.2 Null Set and Measure Space

In a general measure space $\langle \Omega, \mathcal{G}, \nu \rangle$, let us denote by $\mathcal{N}(\mathcal{G}, \nu)$ the class of all sets A in $\mathcal{G}$ with $\nu(A) = 0$, i.e.,

$$\mathcal{N}(\mathcal{G}, \nu) = \{A \in \mathcal{G} : \nu(A) = 0\}.$$

They are called the **null** sets when $\mathcal{G}$ and ν are understood, or ν-null sets when $\mathcal{G}$ is understood.

Remark 22. *Beware that if $A \subset B$ and B is a null set, it does not follow that A is a null set because A may not be in $\mathcal{G}$!*

This remark introduces the following definition.

Definition 23. *The measure space $\langle \Omega, \mathcal{G}, \nu \rangle$ is called* ***complete*** *iff any subset of a null set is a null set.*

Theorem 24. *The following three collections of subsets of Ω are identical:*

(a) *$A \subset \Omega$ and the outer measure $\mu^*(A) = 0$;*

(b) *$A \in \mathcal{F}^*$ and $\mu^*(A) = 0$;*

(c) *$A \subset B$ where $B \in \mathcal{F}$ and $\mu(B) = 0$.*

The very collection is $\mathcal{N}(\mathcal{F}^*, \mu^*)$.

Proof.

- If $\mu^*(A) = 0$, we will prove $A \in \mathcal{F}^*$ by verifying the criterion 1.6. For any $Z \subset \Omega$, we have by properties ❶ and ❷ of μ^*:

$$0 \leq \mu^*(ZA) \leq \mu^*(A) = 0, \quad \mu^*(ZA^c) \leq \mu^*(Z);$$

  and consequently by property ❸:

$$\mu^*(Z) = \mu^*(ZA \cup ZA^c) \leq \mu^*(ZA) + \mu^*(ZA^c) \leq \mu^*(Z).$$

  Thus (1.6) is satisfied and we have proved that (a) and (b) are equivalent.

- Next, let $A \in \mathcal{F}^*$ and $\mu^*(A) = 0$. Then we have by the first assertion in Theorem 21 that there exists $B \in \mathcal{F}$ such that $A \subset B$ and $\mu^*(A) = \mu^*(B)$. Thus A satisfies (c). Conversely, if A satisfies (c), then by property ❷ of outer measure:

$$\mu^*(A) \leq \mu^*(B) = \mu(B) = 0,$$

  so (a) is true.

As consequence, any subset of a $\langle \mathcal{F}^*, \mu^* \rangle$-null set is a $\langle \mathcal{F}^*, \mu^* \rangle$-null set. This is the first assertion in the next theorem.

Theorem 25. *The measure space $\langle \Omega, \mathcal{F}^*, \mu^* \rangle$ is complete. Let $\langle \Omega, \mathcal{G}, \nu \rangle$ be a complete measure space; $\mathcal{G} \supset \mathcal{F}_0$ and $\nu = \mu$ on $\mathcal{F}_0$. If μ is σ-finite on $\mathcal{F}_0$ then $\mathcal{G} \supset \mathcal{F}^*$ and $\nu = \mu^*$ on $\mathcal{F}^*$.*

Proof.

- Let $A \in \mathcal{F}^*$, then by Theorem 21 there exists $B \in \mathcal{F}$ and $C \in \mathcal{F}$ such that

$$C \subset A \subset B, \quad \mu(C) = \mu^*(A) = \mu(B). \tag{1.11}$$

  Since $\nu = \mu$ on $\mathcal{F}_0$, we have by Theorem 19, $\nu = \mu$ on $\mathcal{F}$. Hence by (1.11) we have $\nu(B - C) = 0$. Since $A - C \subset B - C$ and $B - C \in \mathcal{G}$, and $\langle \Omega, \mathcal{G}, \nu \rangle$ is complete, we have $A - C \in \mathcal{G}$ and so $A = C \cup (A - C) \in \mathcal{G}$.

- Moreover, since C, A, and B belong to $\mathcal{G}$, it follows from (1.11) that

$$\mu(C) = \nu(C) \leq \nu(A) \leq \nu(B) = \mu(B)$$

  and consequently by (1.11) again $\nu(A) = \mu(A)$.

To summarize the gist of Theorems 21 and 25, if the measure μ on the field $\mathcal{F}_0$ is σ-finite on $\mathcal{F}_0$, then $\langle \mathcal{F}, \mu \rangle$ is its unique extension to σF, and $\langle \mathcal{F}^*, \mu^* \rangle$ is its minimal complete extension. Here one is tempted to change the notation μ to μ_0 on $\mathcal{F}_0$!

We will complete the picture by showing how to obtain $\langle \mathcal{F}^*, \mu^* \rangle$ from $\langle \mathcal{F}, \mu \rangle$, reversing the order of previous construction. Given the measure space $\langle \Omega, \mathcal{F}, \mu \rangle$, let us denote by $\mathcal{C}$ the collection of subsets of Ω as follows:

$$A \in \mathcal{C} \text{ iff there exists } B \in \mathcal{N}(\mathcal{F}, \mu) \text{ such that } A \subset B.$$

Clearly $\mathcal{C}$ has the "hereditary" property:

If A belongs to $\mathcal{C}$, then all subsets of A belong to $\mathcal{C}$; $\mathcal{C}$ is also closed under countable union.

Next we define the collection

$$\hat{\mathcal{F}} = \{A \subset \Omega : A = B - C, B \in \mathcal{F}, C \in \mathcal{F}\}, \tag{1.12}$$

where the symbol "$-$" denotes **strict difference** of sets, namely $B - C = BC^c$ for $C \subset B$. Finally we define a function $\hat{\mu}$ on $\hat{\mathcal{F}}$ as follows, for the A shown in (1.12):

$$\hat{\mu}(A) = \mu(B). \tag{1.13}$$

We will legitimize this definition and with the same stroke prove the monotonicity of $\hat{\mu}$. Suppose then

$$B_1 - C_1 \subset B_2 - C_2, \quad B_i \in \mathcal{F}, C_i \in \mathcal{C}, i = 1, 2. \tag{1.14}$$

Let $C_1 \subset D \in \mathcal{N}(\mathcal{F}, \mu)$. Then $B_1 \subset B_2 \cup D$ and so $\mu(B_1) \leq \mu(B_2 \cup) = \mu(B_2)$. When the $\subset$ in (1.14) is "=", we can interchange B_1 and B_2 to conclude that $\mu(B_1) = \mu(B_2)$, so that the definition (1.13) is legitimate.

Theorem 26. *$\hat{\mathcal{F}}$ is a Borel field and $\hat{\mu}$ is a measure on $\hat{\mathcal{F}}$.*

Proof.

- Let $A_n \in \hat{\mathcal{F}}$ and $n \in \mathbb{N}$, so that $A_n = B_n C_n^c$ as in (1.12). We have then

$$\bigcap_{n=1}^{\infty} A_n = \left(\bigcap_{n=1}^{\infty} B_n \right) \cap \left(\bigcup_{j=1}^{\infty} C_j \right)^c.$$

  Since the class $\mathcal{C}$ is closed under countable union, this shows that $\hat{\mathcal{F}}$ is closed under countable intersection. Next let $C \subset D, d \in \mathcal{N}(\mathcal{F}, \mu)$; then

$$\begin{aligned} A^c &= B^c \cup C = B^c \cup BC = B^c \cup [B(D - (D - C))] \\ &= (B^c \cup BD) - B(D - C). \end{aligned}$$

  Since $B(D - C) \subset D$, we have $B(D - C) \in \mathcal{C}$; hence the above shows that $\hat{\mathcal{F}}$ is also closed under complementation and therefore is a Borel field. Clearly $\hat{\mathcal{F}} \supset \mathcal{F}$ because we may take $C = \emptyset$ in (1.12).

- To prove $\hat{\mu}$ is countably additive on $\hat{\mathcal{F}}$, let $\{A_n\}$ be disjoint sets in $\hat{\mathcal{F}}$. Then

$$A_n = B_n - C_n, B_n \in \mathcal{F}, C_n \in \mathcal{C}.$$

  There exists D in $\mathcal{N}(\mathcal{G}, \mu)$ containing $\bigcup_{n=1}^{\infty} C_n$. Then $\{B_n - D\}$ are disjoint and

$$\bigcup_{n=1}^{\infty} (B_n - D) \subset \bigcup_{n=1}^{\infty} A_n \subset \bigcup_{n=1}^{\infty} B_n.$$

All these sets belong to $\hat{\mathcal{F}}$ and so by monotonicity of $\hat{\mu}$:

$$\hat{\mu}\left(\bigcup_n (B_n - D)\right) \leq \hat{\mu}\left(\bigcup_n A_n\right) \leq \hat{\mu}\left(\bigcup_n B_n\right).$$

Since $\hat{\mu} = \mu$ on $\mathcal{F}$, the first and third members above are equal to, respectively:

$$\mu\left(\bigcup_n (B_n - D)\right) = \sum_n \mu(B_n - D) = \sum_n \mu(B_n) = \sum_n \hat{\mu}(A_n);$$
$$\mu\left(\bigcup_n B_n\right) \leq \sum_n \mu(B_n) = \sum_n \hat{\mu}(A_n).$$

Therefore we have

$$\hat{\mu}\left(\bigcup_n A_n\right) = \sum_n \hat{\mu}(A_n).$$

Since $\hat{\mu}(\emptyset) = \hat{\mu}(\emptyset - \emptyset) = \mu(\emptyset) = 0$, $\hat{\mu}$ is a measure on $\hat{\mathcal{F}}$.

Corollary 27. *In truth:* $\hat{\mathcal{F}} = \mathcal{F}^*$ *and* $\hat{\mu} = \mu^*$.

PROOF.

- For any $A \in \mathcal{F}^*$, by the first part of Theorem 21, there exists $B \in \mathcal{F}$ such that

$$A = B - (B - A), \quad \mu^*(B - A) = 0.$$

  Hence by Theorem 24, $B - A \in \mathcal{C}$ and so $A \in \hat{\mathcal{F}}$ by (1.12). Thus $\mathcal{F}^* \subset \hat{\mathcal{F}}$.

- Since $\mathcal{F} \subset \mathcal{F}$ and $\mathcal{F} \in \mathcal{F}^*$ by Theorem 25, we have $\hat{\mathcal{F}} \subset \mathcal{F}^*$ by (1.12).

- It follows from the above that $\mu^*(A) = \mu(B) = \hat{\mu}(A)$. Hence $\mu^* = \hat{\mu}$ on $\hat{\mathcal{F}} = \mathcal{F}$.

The question arises naturally whether we can extend μ from $\mathcal{F}_0$ to $\mathcal{F}$ directly without the intervention of $\mathcal{F}^*$. This is indeed possible by a method of transfinite induction originally conceived by Borel[4]. It is technically lengthier than the method of outer measure expounded here.

1.4.3 An Example

Although the case of a countable space Ω can be treated in an obvious way, it is instructive to apply the general theory to see what happens.

Let $\Omega = \mathbb{N} \cup \omega$; $\mathcal{F}_0$ is the minimal field (not Borel field) containing each singleton n in $\mathbb{N}$, but not ω. Let N_f denote the collection of all finite subsets of $\mathbb{N}$; then $\mathcal{F}_0$ consists of N_f and the complements of members of N_f (w.r.t. Ω), the latter all containing ω. Let $0 \leq \mu(n) < \infty$ for all $n \in \mathbb{N}$; a measure μ is defined on $\mathcal{F}_0$ as follows:

$$\mu(A) = \sum_{n \in A} \mu(n) \text{ if } A \in N_f, \quad \mu(A^c) = \mu(\Omega) - \mu(A).$$

We must still define $\mu(\Omega)$. Observe that by the properties of a measure, we have $\mu(\Omega) \geq \sum_{n \in \mathbb{N}} \mu(n) = s$, say.

Now we use Definition 14 to determine the outer measure μ^*. It is easy to see that for any $A \subset \mathbb{N}$, we have

$$\mu^*(A) = \sum_{n \in A} \mu(n).$$

[4]Le Blanc and G. E. Fox: On the extension of measure by the method of Borel. *Canadian Journal of Mathematics*, 1956, pp. 516-523.

In particular $\mu*(\mathbb{N}) = s$. Next we have

$$\mu^*(\omega) = \inf_{A \in N_f} \mu(A^c) = \mu(\Omega) - \sum_{A \in N_f} \mu(A) = \mu(\Omega) - s$$

provided $s < \infty$; otherwise the inf above is ∞. Thus we have

$$\mu^*(\omega) = \begin{cases} \mu(\Omega) - s, & \text{if } \mu(\Omega) < \infty; \\ \infty, & \text{if } \mu(\Omega) = \infty. \end{cases}$$

It follows that for any $A \subset \Omega$:

$$\mu^*(A) = \sum_{n \in A} \mu^*(n)$$

where $\mu^*(n) = \mu(n)$ for $n \in \mathbb{N}$. Thus μ^* is a measure on $\mathcal{T}$, namely, $\mathcal{F}^* = \mathcal{T}$.

But it is obvious that $\mathcal{F} = \mathcal{T}$ since $\mathcal{F}$ contains $\mathbb{N}$ as countable union and so contains ω as complement. Hence $\mathcal{F} = \mathcal{F}^* = \mathcal{T}$.

If $\mu(\Omega) = \infty$ and $s = \infty$, the extension μ^* of μ to $\mathcal{T}$ is not unique, because we can define $\mu(\omega)$ to be any positive number and get an extension. Thus μ is not σ-finite on $\mathcal{F}_0$, by Theorem 19. But we can verify this directly when $\mu(\Omega) = \infty$, whether $s = \infty$ or $s < \infty$. Thus in the latter case, $\mu(\omega) = \infty$ is also the unique extension of μ from $\mathcal{F}_0$ to $\mathcal{T}$. This means that the condition of σ-finiteness on $\mathcal{F}_0$ is only a sufficient and not a necessary condition for the unique extension.

As a ramification of the example above, let $\Omega = \mathbb{N} \cup \omega_1 \cup \omega_2$, with two extra points adjoined to $\mathbb{N}$, but keep $\mathcal{F}_0$ as before. Then $\mathcal{F}(= \mathcal{F}^*)$ is strictly smaller than $\mathcal{T}$ because neither ω_1 nor ω_2 belongs to it. From Definition 14 we obtain

$$\mu^*(\omega_1 \cup \omega_2) = \mu^*(\omega_1) = \mu^*(\omega_2).$$

Thus μ^* is not even two-by-two additive on $\mathcal{T}$ unless the three quantities above are aero. The two points ω_1 and ω_2 form an inseparable couple. We leave it to the curious reader to wonder about other possibilities.

1.5 Measures in $\mathbb{R}$

1.5.1 Borel Field of $\mathbb{R}$

Let $\mathbb{R} = (-\infty, +\infty)$ be the set of real members, alias the real line, with its Euclidean topology. For $-\infty \le a < b \le +\infty$,

$$(a, b] = \{x \in \mathbb{R} : a < x \le b\} \tag{1.15}$$

is an interval of particular shape, namely open at left end and closed at right end. For $b = +\infty$, $(a, +\infty] = (a, +\infty)$ because $+\infty$ is not in $\mathbb{R}$. By choice of the particular shape, the complement of such an interval is the union of two intervals of the same shape:

$$(a, b]^c = (-\infty, a] \cup (b, +\infty].$$

When $a = b$, of course $(a, a] = \emptyset$ is the empty set. A finite or countably infinite number of such intervals may merge end to end into a single one as illustrated below:

$$(0, 2] = (0, 1] \cup (1, 2], \quad (0, 1] = \bigcup_{n=1}^{\infty} \left(\frac{1}{n+1}, \frac{1}{n} \right]. \tag{1.16}$$

Apart from this possibility, the representation of $(a, b]$ is unique.

The minimal Borel field containing all $(a, b]$ will be denoted by $\mathcal{B}(\mathbb{R})$ and called the **Borel field of** $\mathbb{R}$. Since a bounded open interval is the countable union of intervals like $(a, b]$, and

any open set in $\mathbb{R}$ is the countable union of (disjoint) bounded open intervals, the Borel field $\mathcal{B}(\mathbb{R})$ contains all open sets; hence by complementation it contains all closed sets, in particular all compact sets. Starting from one of these collections, forming countable union and countable intersection successively, a countable number of times, one can build up $\mathcal{B}(\mathbb{R})$ through a transfinite induction.

Now suppose a measure m has been defined on $\mathcal{B}(\mathbb{R})$, subject to the sole assumption that its value for a finite (alias bounded) interval be finite, namely if $-\infty < a < b < +\infty$, then

$$0 \leq m((a,b]) < \infty. \tag{1.17}$$

We associate a point function F on $\mathbb{R}$ with the set function m on $\mathcal{B}(\mathbb{R})$, as follows:

$$F(x) = \begin{cases} m((0,x]), & \text{for } x > 0; \\ 0, & \text{for } x = 0; \\ -m((x,0]), & \text{for } x < 0. \end{cases} \tag{1.18}$$

This function may be called the "generalized distribution" for m. We see that F is finite everywhere owing to (1.17), and

$$m((a,b]) = F(b) - F(a). \tag{1.19}$$

F is increasing (viz. nondecreasing) in $\mathbb{R}$ and so the limits

$$F(+\infty) = \lim_{x \to +\infty} F(x) \leq +\infty, \quad F(-\infty) = \lim_{x \to -\infty} F(x) \geq -\infty,$$

both exist. Next, F has unilateral limits everywhere, and is right-continuous:

$$F(x-) \leq F(x) = F(x+).$$

The right-continuity follows from the monotone limit properties ❺ and ❻ of m and the primary assumption (1.17). The measure of a single point x is given by

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

We shall denote a point and the set consisting of it (singleton) by the same symbol.

The simplest example of F is given by $F(x) \equiv x$. In this case F is continuous everywhere and (1.17) becomes

$$m((a,b]) = b - a.$$

We can replace $(a,b]$ above by $(a,b), [a,b)$ or $[a,b]$ because $m(x) = 0$ for each x. This measure is the **length** of the line-segment from a to b. It was in this classic case that the following extension was first conceived by Émile Borel (1871-1956).

1.5.2 Measure the Interval $(a,b]$

We are going to construct a measure on $\mathcal{B}(\mathbb{R})$ and a larger Borel field $\mathcal{B}^*(\mathbb{R})$ that fulfills the prescription (1.19) by following the methods proposed by H. Lebesgue and C. Carathédory.

- The first step is to determine the minimal field $\mathcal{B}_0$ containing all $(a,b]$. Since a field is closed under finite union, it must contain all sets of the form

$$B = \bigcup_{j=1}^{n} I_j, \quad I_j = (a_j, b_j], 1 \leq j \leq n, n \in \mathbb{R}. \tag{1.20}$$

  Without loss of generality, we may suppose the intervals I_j to be disjoint, by merging intersections as illustrated by

$$(1,3] \cup (2,4] = (1,4].$$

Then it is clear that the complement B^c is of the same form. The union of two sets like B is also of the same form. Thus the collection of all sets like B already forms a field and so it must be $\mathcal{B}_0$. Of course it contains (includes) the empty set $\emptyset = (a, a]$ and $\mathbb{R}$. However, it does not contain any (a, b) except $\mathbb{R}, [a, b), [a, b]$, or any single point!

- Next we define a measure m on $\mathcal{R}_0$ satisfying (1.19). Since the condition ❹ in Definition 13 requires it to be finitely additive, there is only one way: for the generic B in (1.20) with disjoint I_j we must put

$$m(B) = \sum_{j=1}^{n} m(I_j) = \sum_{j=1}^{n} [F(b_j) - F(a_j)]. \tag{1.21}$$

Having so defined m on $\mathcal{B}_0$, we must now prove that it satisfies the condition ❹ in Definition 13, in order to proclaim it to be a measure on $\mathcal{B}_0$. Namely, if $\{B_k : 1 \le k \le \ell \le \infty\}$ is a finite or countable sequence of disjoint sets in $\mathcal{B}_0$, we must prove

$$m\left(\bigcup_{k=1}^{\ell} B_k\right) = \sum_{k=1}^{\ell} m(B_k), \tag{1.22}$$

whenever ℓ is finite, and moreover when $\ell = \infty$ and the union $\bigcup_{k=1}^{\infty} B_k$ happens to be in $\mathcal{B}_0$.

The case for a finite ℓ is really clear. If each B_k is represented as in (1.20), then the disjoint union of a finite number of them is represented in a similar manner by pooling together all the disjoint I_j's from the B_k's. Then the equation (1.22) just means that a finite double array of numbers can be summed in two orders.

If that is so easy, what is the difficulty when $\ell = \infty$? It turns out, as Borel saw clearly, that the crux of the matter lies in the following fabulous "banality".

Lemma 28 (Borel)**.** *If $-\infty \le a < b \le +\infty$ and*

$$(a, b] = \bigcup_{j=1}^{\infty} (a_j, b_j], \tag{1.23}$$

where $a_j < b_j$ for each j, and the intervals $(a_j, b_j]$ are disjoint, then we have

$$F(b) - F(a) = \sum_{j=1}^{\infty} [F(b_j) - F(a_j)]. \tag{1.24}$$

Proof. The equation (1.24) can be established by two inequalities in opposite direction.

- The first inequality is easy by considering the first n terms in the disjoint union (1.23):

$$F(b) - F(a) \ge \sum_{j=1}^{n} [F(b_j) - F(a_j)].$$

As n goes to infinity we obtain (1.24) with "=" replaced by "$\ge$", i.e.,

$$F(b) - F(a) \ge \sum_{j=1}^{\infty} [F(b_j) - F(a_j)].$$

- The second inequality can be built with the help of Borel's covering theorem and it is necessary to alter the shape of the intervals $(a_j, b_j]$ to fit the picture in (1.3).

Let $-\infty < a < b < \infty$ and $\varepsilon > 0$. Choose a' in (a, b), and for each j choose $b'_j > b_j$ such that

$$F(a') - F(a) < \frac{\varepsilon}{2}, \quad F(b'_j) - F(b_j) < \frac{\varepsilon}{2^{j+1}}. \tag{1.25}$$

These choices are possible because F is right continuous; and now we have

$$[a', b] \subset \bigcup_{j=1}^{\infty} (a_j, b'_j)$$

as required in (1.3). Hence by Borel's theorem, there exists a finite $\ell(\varepsilon) \in \mathbb{N}$ such that

$$[a', b] \subset \bigcup_{j=1}^{\ell(\varepsilon)} (a_j, b'_j). \tag{1.26}$$

From this it follows "easily" that

$$F(b) - F(a') \leq \sum_{j=1}^{\ell(\varepsilon)} \left[F(b'_j) - F(a_j)\right]. \tag{1.27}$$

We will spell out the proof by induction on $\ell(\varepsilon)$.

- When $\ell(\varepsilon) = 1$, it is obvious.
- Suppose the assertion has been proved for $\ell(\varepsilon) - 1, \ell(\varepsilon) \geq 2$. From (1.26) as written, there is k, $1 \leq k \leq \ell(\varepsilon)$, such that $a_k < a' < b'_k$ and so

$$F(b'_k) - F(a') \leq F(b'_k) - F(a_k). \tag{1.28}$$

  If we intersect both sides of (26) with the complement of (a_k, b'_k), we obtain

$$[b'_k, b] \subset \bigcup_{j=1, j\neq k}^{\ell(\varepsilon)} (a_j, b'_j).$$

  Here the number of intervals on the right side is $\ell(\varepsilon) - 1$; hence by the induction hypothesis we have

$$F(b) - F(b'_k) \leq \sum_{j=1, j\neq k}^{\ell(\varepsilon)} \left[F(b'_j) - F(a_j)\right].$$

  Adding this to (1.28) we obtain (1.27), and the induction is complete.

It follows from (1.27) and (1.25) that

$$F(b) - F(a) \leq \sum_{j=1}^{\ell(\varepsilon)} [F(b_j) - F(a_j)] + \varepsilon.$$

Note that the $\ell = \ell(\varepsilon)$ above depends on ε. However, if we change ℓ to ∞ then the infinite series of course does not depend on ε. Therefore we can let $\varepsilon \to 0$ to obtain (1.24) when the "=" is changed to "$\leq$", namely the other half of Borel's lemma, for finite a and b. In other words,

$$F(b) - F(a) \leq \sum_{j=1}^{\infty} \left[F(b_j) - F(a_j)\right].$$

- It remains to treat the case $a = -\infty$ and/or $b = +\infty$. Let

$$(-\infty, b] \subset \bigcup_{j=1}^{\infty} (a_j, b_j]$$

  Then for any a in $(-\infty, b)$, (1.24) holds with "=" replaced by "$\leq$". Letting $a \to -\infty$ we obtain the desired result. The case $b = +\infty$ is similar. ■

Corollary 29. *Assume all I with subscripts denote intervals of the shape $(a, b]$, $\sum$ denotes union of disjoint sets. Let $B \in \mathcal{B}_0, B_j \in \mathcal{B}_0, j \in \mathbb{N}$. Suppose*

$$B = \sum_{j=1}^{\infty} B_j, \quad B_j = \sum_{k=1}^{n_j} I_{jk}$$

such that

$$\sum_{i=1}^{n} I_i = \sum_{j=1}^{\infty} \sum_{k=1}^{n_j} I_{jk}, \tag{1.29}$$

then

$$\sum_{i=1}^{n} m(I_i) = \sum_{j=1}^{\infty} \sum_{k=1}^{n_j} m(I_{jk}). \tag{1.30}$$

PROOF.

- For $n = 1$, (1.29) is of the form (1.23) since a countable set of sets can be ordered as a sequence. Hence (1.30) follows by Borel's lemma.
- In general, simple geometry shows that each I_i in (1.29) is the union of a subcollection of the $I'_{jk}s$. This is easier to see if we order the I_i's in algebraic order and, after merging where possible, separate them at nonzero distances. Consequently, (1.30) follows by adding n equations, each of which results from Borel's lemma. This completes the proof of the countable additivity of m on $\mathcal{B}_0$, namely (1.22) is true as stipulated there for $\ell = \infty$ as well as $\ell < \infty$.

1.5.3 From Abstract to Concrete

Table 1.3: Concepts of Measure Space

CONCEPT	GENERAL	SPECIAL
universal set	Ω	$\mathbb{R}$
total Borel Field	$\mathcal{T}$ or $\mathcal{T}(\Omega)$	$\mathcal{T}$ or $\mathcal{T}(\mathbb{R})$
(basis) field	$\mathcal{F}_0$	$\mathcal{B}_0$
B.F. generated from basis field	$\mathcal{F}$	$\mathcal{B}$
associated extensive B.F.	$\mathcal{F}^*$	$\mathcal{B}^*$
	$\mathcal{T}(\Omega) \supset \mathcal{F}^* \supset \mathcal{F} \supset \mathcal{F}_0$	$\mathcal{T}(\mathbb{R}) \supset \mathcal{B}^* \supset \mathcal{B} \supset \mathcal{B}_0$
	$\mathcal{F}^* \neq \mathcal{F}$	$\mathcal{B}^* \neq \mathcal{B}$
outer measure	μ^*	m^*
measure	μ	m
measure restricted	$\mu^*\|_{\mathcal{F}_0} = \mu$	$m^*\|_{\mathcal{B}_0} = m$
unique extension of measure	$\mu^*\|_{\mathcal{F}}$	$m^*\|_{\mathcal{B}}$
measure space	$\langle \Omega, \mathcal{F}_0, \mu \rangle$	$\langle \mathbb{R}, \mathcal{B}_0, m \rangle$
complete measure space	$\langle \Omega, \mathcal{F}^*, \mu^* \rangle$	$\langle \mathbb{R}, \mathcal{B}^*, m^* \rangle$

Note that there are some concepts and results frequently encountered:

- A member of $\mathcal{B}$ is called a **Borel set**.
- A member of $\mathcal{B}^*$ is usually called **Lebesgue-measurable**.
- There are incomparably more Lebesgue-measurable sets than Borel sets. These "special" sets are very important for fractal geometry.

- There are many sets in $\mathcal{B}^*$ but not in $\mathcal{B}$.
- The generalization to a generalized distribution function F is sometimes referred to as Borel-Lebesgue-Stieltjes.

1.6 Integral

Suppose we have a measure space $\langle \Omega, \mathcal{F}, \mu \rangle$.

1.6.1 Measurable Function

A function $f : \Omega \to \mathbb{R}^* = [-\infty, +\infty]$ is called $\mathcal{F}$-**measurable** iff

$$\forall c \in \mathbb{R}, \quad \{f \leq c\} = \{\omega \in \Omega : f(\omega) \leq c\} \in \mathcal{F}.$$

We write $f \in \mathcal{F}$ in this case. It follows that for each set $A \in \mathcal{B}$, namely a Borel set, we have

$$\{f \in A\} \in \mathcal{F}, \quad \{f = +\infty\} \in \mathcal{F}, \quad \{f = -\infty\} \in \mathcal{F}.$$

A function $f \in \mathcal{F}$ with range a countable set in $[0, +\infty]$ will be called a **basic function**. Let $\{a_j\}$ be its range (which may include "∞"), and $A_j = \{f = a_j\}$. Then the A_j's are disjoint sets with union Ω and

$$f = \sum_j a_j \mathbb{1}_{A_j}, \tag{1.31}$$

where the sum is over a countable set of j.

NOTATIONS.

- $\boxed{\mathcal{F}_+}$ — The class of positive $\mathcal{F}$-measurable functions will be denoted by $\mathcal{F}_+$, i.e.,

$$\mathcal{F}_+ = \{f \in \mathcal{F} : f \geq 0\}.$$

- $\boxed{f^+}$ — For any $f \in \mathcal{F}$ with range in $[-\infty, +\infty]$, we denote the positive part of f as

$$f^+ = \begin{cases} f, & \text{on } \{f \geq 0\}; \\ 0, & \text{on } \{f < 0\}. \end{cases}$$

  We find that $f^+ \in \mathcal{F}_+$.

- $\boxed{f^-}$ — For any $f \in \mathcal{F}$ with range in $[-\infty, +\infty]$, we denote the positive part of f as

$$f^- = \begin{cases} -f, & \text{on } \{f \leq 0\}; \\ 0, & \text{on } \{f > 0\}. \end{cases}$$

  Obviously, $f^+ \in \mathcal{F}_+$.

- $\boxed{|f|}$

$$f = f^+ - f^-, \quad |f| = f^+ + f^-.$$

- $\boxed{a.e.}$ — A set $A \subset \Omega$ is called a **null set** iff $A \in \mathcal{F}$ and $\mu(A) = 0$. A mathematical proposition $\mathcal{P}$ is said to hold **almost everywhere**, or **a.e.** iff there is a null set A such that $\mathcal{P}$ holds outside A, namely in $A^c = \Omega \setminus A$.

- $\boxed{C(f)}$ — A class of function f in which each function is equivalent to f, i.e.

$$C(f) = \left\{ g : g \overset{a.e.}{=} f, f \in \mathcal{F} \right\}$$

When $\langle \Omega, \mathcal{F}, \mu \rangle$ is a complete measure space, $g \in C(f)$ implies $g \in \mathcal{F}$. A member of $C(f)$ may be called a **version** of f, and may be substituted for f wherever a null set "does not count". A particularly version of f is the following **finite version**:

$$\bar{f} = \begin{cases} f, & \text{on } \{|f| < \infty\}; \\ 0, & \text{on } \{|f| = \infty\}. \end{cases}$$

where 0 may be replaced by some other number, e.g., by 1 in $\mathrm{E}(\log f)$. Note that in functional analysis, it is the class $C(f)$ rather than an individual f that is a member of L^1.

Definition 30. *For the basic function f in (1.31), its **integral** is defined to be*

$$\mathrm{E}(f) = \sum_j a_j \mu(A_j) \tag{1.32}$$

and is also denoted by

$$\int f \,\mathrm{d}\,\mu = \int_\Omega f(\omega) \mu(\mathrm{d}\,\omega).$$

If a term in (1.32) is $0 \cdot \infty$ or $\infty \cdot 0$, it is taken to be 0. In particular if $f \equiv 0$, then $\mathrm{E}(0) = 0$ even if $\mu(\Omega) = \infty$. If $A \in \mathcal{F}$ and $\mu(A) = 0$, then the basic function

$$\infty \cdot \mathbb{1}_A + 0 \cdot \mathbb{1}_{A^c}$$

has integral equal to

$$\infty \cdot 0 + 0 \cdot \mu(A^c) = 0.$$

We list some of the properties of the integral.

(i) Let $\{B_j\}$ be a countable set of disjoint sets in $\mathcal{F}$, with union Ω and $\{b_j\}$ arbitrary positive numbers or ∞, not necessarily distinct. Then the function

$$\sum_j b_j \mathbb{1}_{B_j} \tag{1.33}$$

is basic, and its integral is

$$\mathrm{E}\left(\sum_j b_j \mathbb{1}_{B_j} \right) = \int \left(\sum_j b_j \mathbb{1}_{B_j} \right) \mathrm{d}\,\mu = \sum_j b_j \mu(B_j).$$

(ii) If f and g are basic and $f \le g$, then

$$\mathrm{E}(f) \le \mathrm{E}(g).$$

In particular, if $\mathrm{E}(f) = +\infty$, then $\mathrm{E}(g) = +\infty$.

(iii) If f and g are basic functions, α and β and positive numbers, then $\alpha f + \beta g$ is basic and

$$\mathrm{E}(\alpha f + \beta g) = \alpha \,\mathrm{E}(f) + \beta \,\mathrm{E}(g).$$

(iv) Let $A \in \mathcal{F}$ and f be a basic function. Then the product $\mathbb{1}_A f$ is basic function and its integral will be denoted by

$$\mathrm{E}(A; f) = \int_A f(\omega) \mu(\mathrm{d}\,\omega) = \int_A f \,\mathrm{d}\,\mu. \tag{1.34}$$

Let $A_n \in \mathcal{F}, A_n \subset A_{n+1}$ for all n and $A = \bigcup_n A_n$, then we have

$$\lim_n \mathrm{E}(A_n; f) = \mathrm{E}(A; f). \tag{1.35}$$

PROOF.

Omitted.

■

Consider now an increasing sequence $\{f_n\}$ of basic functions, namely, $f_n \leq f_{n+1}$ for all n. Then $f = \lim_n \uparrow f_n$ exists and $f \in \mathcal{F}$, but of course f need not be basic; and its integral has yet to be defined. By property (ii), the numerical sequence $\mathrm{E}(f_n)$ is increasing and so $\lim_n \mathrm{E}(f_n)$ exists, possibly equal to $+\infty$. It is tempting to define $\mathrm{E}(f)$ to be that limit, but we need the following result to legitimize the idea.

Theorem 31. *Let $\{f_n\}$ and $\{g_n\}$ be two increasing sequences of basic functions such that*

$$\lim_n \uparrow f_n = \lim_n \uparrow g_n \tag{1.36}$$

(everywhere in Ω). Then we have

$$\lim_n \uparrow \mathrm{E}(f_n) = \lim_n \uparrow \mathrm{E}(g_n) \tag{1.37}$$

PROOF.

- Denote the common limit function in (1.36) by f and put

$$A = \{\omega \in \Omega : f(\omega) > 0\},$$

then $A \in \mathcal{F}$. Since $0 \leq g_n \leq f$, we have $\mathbb{1}_{A^c} g_n = 0$ identically; hence by property (iii):

$$\mathrm{E}(g_n) = \mathrm{E}(A; g_n) + \mathrm{E}(A^c; g_n) = \mathrm{E}(A; g_n). \tag{1.38}$$

Fix an n and put for each $k \in \mathbb{N}$:

$$A_k = \left\{\omega \in \Omega : f_k(\omega) > \frac{n-1}{n} g_n(\omega)\right\}.$$

Since $f_k \leq f_{k+1}$, we have $A_k \subset A_{k+1}$ for all k. We are going to prove that

$$\bigcup_{k=1}^{\infty} A_k = A. \tag{1.39}$$

  - If $\omega \in A_k$, then $f(\omega) \geq f_k(\omega) > \frac{n-1}{n} g_n(\omega) \geq 0$; hence $\omega \in A$.
  - On the other hand, if $\omega \in A$, then

$$\lim_k \uparrow f_k(\omega) = f(\omega) \geq g_n(\omega).$$

    and $f(\omega) > 0$; hence there exists an index k such that

$$f_k(\omega) > \frac{n-1}{n} g_n(\omega)$$

    and so $\omega \in A_k$. Thus (1.39) is proved.

By property (ii), since

$$f_k \geq \mathbb{1}_{A_k} f_k \geq \frac{n-1}{n} \cdot \mathbb{1}_{A_k} g_n$$

we have

$$\mathrm{E}(f_k) \geq \mathrm{E}(A_k; f_k) \geq \frac{n-1}{n} \mathrm{E}(A_k; g_n).$$

Letting $k \uparrow \infty$, we obtain by property (iv):

$$\lim_k \uparrow \mathrm{E}(f_k) \geq \frac{n-1}{n} \lim_k \mathrm{E}(A_k; g_n) = \frac{n-1}{n} \mathrm{E}(A; g_n) = \frac{n-1}{n} \mathrm{E}(g_n)$$

where the last equation is due to (1.38). Now let $n \uparrow \infty$ to obtain

$$\lim_k \uparrow \mathrm{E}(f_k) \geq \lim_n \uparrow \mathrm{E}(g_n).$$

Since $\{f_n\}$ and $\{g_n\}$ are interchangeable, (1.37) is proved.

Corollary 32. *Let f_n and f be basic functions such that $f_n \uparrow f$, then $\mathrm{E}(f_n) \uparrow \mathrm{E}(f)$.*

PROOF. Take $g_n = f$ for all n in the theorem. ■

1.6.2 Approximating Measurable Functions with Basic Functions

We define for any $f \in \mathcal{F}_+$ the approximating sequence $\left\{f^{(m)}\right\}_{m=0}^{\infty}$, by

$$f^{(m)}(\omega) = \frac{\lfloor 2^m f(\omega) \rfloor}{2^m} \tag{1.40}$$

where $\lfloor \cdot \rfloor$ means

$$\lfloor x \rfloor = n - 1, \quad \forall x \in (n-1, n], n \in \mathbb{Z}.$$

Each $f^{(m)}$ is a basic function with range in the set of binary numbers: $\{k/2^m\}$ where k is a nonnegative integer or ∞. We have $f^{(m)} \leq f^{(m+1)}$ for all m, by the magic property of bisection. Finally $f^{(m)} \uparrow$ owing to the left-continuity of the function $\lfloor x \rfloor$.

Definition 33. *For $f \in \mathcal{F}_+$, its integral is defined to be*

$$\mathrm{E}(f) = \lim_m \uparrow \mathrm{E}\left(f^{(m)}\right). \tag{1.41}$$

When f is basic, Definition 33 is consistent with Definition 30, by Corollary 32 to Theorem 31. The extension of property (ii) of integrals to $\mathcal{F}_+$ is trivial, because $f \leq g$ implies $f^{(m)} \leq g^{(m)}$. On the contrary, $(f+g)^{(m)}$ is not $f^{(m)} + g^{(m)}$ because of $\lfloor x + y \rfloor \neq \lfloor x \rfloor + \lfloor y \rfloor$, but since $f^{(m)} + g^{(m)} \uparrow (f+g)$, it follows from Theorem 31 that

$$\lim_m \uparrow \mathrm{E}\left(f^{(m)} + g^{(m)}\right) = \lim_m \uparrow \mathrm{E}\left([f+g]^{(m)}\right)$$

that yields property (iii) for $\mathcal{F}_+$, together with $\mathrm{E}\left(af^{(m)}\right) \uparrow a\,\mathrm{E}(f)$, for $a \geq 0$.

Property (iv) for $\mathcal{F}_+$ can be given in an equivalent form as follows:

(iv') For $f \in \mathcal{F}_+$, the function of sets defined on $\mathcal{F}$ by $\mathrm{E}(\,*\,; f)$

$$\mathrm{E}(\,*\,; f) : A \mapsto \mathrm{E}(A; f)$$

is a measure.

PROOF.

- We need only prove that if $A = \bigcup_n A_n$ where the A_n's are disjoint sets in $\mathcal{F}$, then

$$\mathrm{E}(A; f) = \sum_{n=1}^{\infty} \mathrm{E}(A_n; f).$$

For a basic f, this follows from properties (iii) and (iv).

- The extension to $\mathcal{F}_+$ can be done by the double limit theorem.

1.6.3 Property of Convergence

There are three fundamental theorems relating the convergence of functions with the convergence of their integrals. We begin with B. Levi's theorem on Monotone Convergence (1906), which is the extension of Corollary 32 to $\mathcal{F}_+$.

Theorem 34. *Let $\{f_n\}$ be an increasing sequence of functions in $\mathcal{F}_+$ with limit $f : f_n \uparrow f$. Then we have*

$$\lim_n \uparrow \mathrm{E}(f_n) = \mathrm{E}(f) \leq +\infty.$$

PROOF.

- We have $f \in \mathcal{F}_+$; hence by Definition 33, (1.41) holds. For each f_n, we have, using analogous notation:
$$\lim_m \uparrow \mathrm{E}\left(f_n^{(m)}\right) = \mathrm{E}(f_n). \tag{1.42}$$
Since $f_n \uparrow f$, the numbers $\lfloor 2^m f_n(\omega) \rfloor \uparrow \lfloor 2^m f(\omega) \rfloor$ as $n \uparrow \infty$, owing to the left continuity of $\lfloor x \rfloor$. Consequently by Corollary 32,
$$\lim_n \uparrow \mathrm{E}(f_n^{(m)}) = \mathrm{E}(f^{(m)}). \tag{1.43}$$
It follows that
$$\lim_m \uparrow \lim_n \uparrow \mathrm{E}(f_n^{(m)}) = \lim_m \uparrow \mathrm{E}(f^{(m)}) = \mathrm{E}(f).$$

- On the other hand, it follows from (1.42) that
$$\lim_m \uparrow \lim_n \uparrow \mathrm{E}(f_n^{(m)}) = \lim_m \uparrow \mathrm{E}(f_n).$$
Therefore the theorem is proved by the double limit lemma. ■

Theorem 35. *Let $f_n \in \mathcal{F}_+, n \in \mathbb{N}$. Suppose*

(a) $\lim_n f_n = 0$;

(b) $\mathrm{E}\left(\sup_n f_n\right) < \infty$.

Then we have
$$\lim_n \mathrm{E}(f_n) = 0. \tag{1.44}$$

PROOF.

- Put for $n \in \mathbb{N}$:
$$g_n = \sup_{k \geq n} f_k \tag{1.45}$$
Then $g_n \in \mathcal{F}_+$, and as $n \uparrow \infty, g_n \downarrow \limsup_n f_n = 0$ by (a); and $g_1 = \sup_n f_n$ so that $\mathrm{E}(g_1) < \infty$ by (b).

- Now consider the sequence $\{g_1 - g_n\}, n \in \mathbb{N}$. This is increasing with limit g_1. Hence by Theorem 34, we
$$\lim_n \uparrow \mathrm{E}(g_1 - g_n) = \mathrm{E}(g_1).$$
By property (iii) for $\mathcal{F}_+$,
$$\mathrm{E}(g_1 - g_n) + \mathrm{E}(g_n) = \mathrm{E}(g_1).$$
Substituting into the preceding relation and cancelling the finite $\mathrm{E}(g_1)$, we obtain $\mathrm{E}(g_n) \downarrow 0$. Since $0 \leq f_n \leq g_n$ by property (ii) for $\mathcal{F}_+$, (1.44) follows. ■

The following theorem, viz. Fatou's Lemma, has the virtue of "no assumptions" with the consequent one-sided conclusion, which is however often useful.

Theorem 36 (Fatou)**.** *Let $\{f_n\}$ be an arbitrary sequence of functions in f_+. Then we have*
$$\mathrm{E}\left(\liminf_n f_n\right) \leq \liminf_n \mathrm{E}(f_n). \tag{1.46}$$

PROOF.

- Put for $n \in \mathbb{N}$:

$$g_n = \inf_{k \geq n} f_k,$$

then

$$\liminf_n f_n = \lim_n \uparrow g_n.$$

Then by Theorem 34,

$$\mathrm{E}\left(\liminf_n f_n\right) = \lim_n \uparrow \mathrm{E}(g_n). \tag{1.47}$$

- Since $g_n \leq f_n$, we have $\mathrm{E}(g_n) \leq \mathrm{E}(f_n)$ and

$$\liminf_n \mathrm{E}(g_n) \leq \liminf_n \mathrm{E}(f_n).$$

The left member above is in truth the right member of (1.47); therefore (1.46) follows as a milder but neater conclusion.■

Remark 37. *From Theorem 36 to Theorem 34 and Theorem 35*

- *From Theorem 36 to Theorem 34:*
  - *If $f_n \uparrow f$, then (1.46) yields $\mathrm{E}(f) \leq \lim_n \uparrow \mathrm{E}(f_n)$. Since $f \geq f_n, \mathrm{E}(f) \geq \lim_n \uparrow \mathrm{E}(f_n)$; hence there is equality.*
- *From Theorem 36 to Theorem 35:*
  - *Using the notation in (1.45), we have $0 \leq g_1 - f_n \leq g_1$. Hence by condition (a) and (1.46),*

$$\begin{aligned} \mathrm{E}(g_1) &= \mathrm{E}\left[\liminf_n (g_1 - f_n)\right] \leq \liminf_n \left[\mathrm{E}(g_1) - \mathrm{E}(f_n)\right] \\ &= \mathrm{E}(g_1) - \limsup_n \mathrm{E}(f_n) \end{aligned}$$

that yields (1.44).

With the help of f^+ and f^-, by Definition 33 and property (iii), we can deduce that

$$\mathrm{E}(|f|) = \mathrm{E}(f^+) + \mathrm{E}(f^-). \tag{1.48}$$

By comparison, the integral for general $f \in \mathcal{F}$ can be defined as follows.

Definition 38. *For $f \in \mathcal{F}$, its integral is defined to be*

$$\mathrm{E}(f) = \mathrm{E}(f^+) - \mathrm{E}(f^-), \tag{1.49}$$

*provided the right side above is defined, namely not $\infty - \infty$. We say $f \in \mathcal{F}$ is **integral** or L^1-**integral**, for $f \in L^1$, if and only $\mathrm{E}(f^+)$ and $\mathrm{E}(f^-)$ are finite; in this case $\mathrm{E}(f)$ is a finite number.*

When $\mathrm{E}(f)$ exists but f is not integrable, then it must be equal to $+\infty$ or $-\infty$, by (1.49).

Theorem 39. *For the integrable functions, we have the following results:*

(i) *The function f in $\mathcal{F}$ is integrable if and only if $|f|$ is integrable; we have*

$$|\mathrm{E}(f)| \leq \mathrm{E}(|f|) \tag{1.50}$$

(ii) *For any $f \in \mathcal{F}$ and any null set A, we have*

$$\begin{aligned} \mathrm{E}(A; f) &= \int_A f \, \mathrm{d}\mu = 0; \\ \mathrm{E}(f) = \mathrm{E}(A^c; f) &= \int_{A^c} f \, \mathrm{d}\mu. \end{aligned} \tag{1.51}$$

(iii) *If $f \in L^1$, then the set $\{\omega \in \Omega : |f(\omega)| = \infty\}$ is a null set.*

(iv) *If $f \in L^1, g \in \mathcal{F}$, and $|g| \le |f|$ a.e., then $g \in L^1$.*

(v) *If $f \in \mathcal{F}, g \in \mathcal{F}$, and $g = f$ a.e., then $\mathrm{E}(g)$ exists if and only if $\mathrm{E}(f)$ exists, and then $\mathrm{E}(g) = \mathrm{E}(f)$.*

(vi) *If $\mu(\Omega) < \infty$, then any a.e. bounded $\mathcal{F}$-measurable function is integrable.*

PROOF.

- (i) is trivial from (1.48) and (1.49).
- (ii) follows from
$$\mathbb{1}_A |f| \le \mathbb{1}_A \cdot \infty$$
so that
$$0 \le \mathrm{E}(\mathbb{1}_A |f|) \le \mathrm{E}(\mathbb{1}_A \cdot \infty) = \mu(A) \cdot \infty = 0.$$
This implies (1.51).
- To prove (iii), let
$$A(n) = \{|f| \ge n\}$$
Then $A(n) \in \mathcal{F}$ and
$$m\mu(A(n)) = \mathrm{E}(A(n); n) \le \mathrm{E}(A(n); |f|) \le \mathrm{E}(|f|).$$
In consequence
$$\mu(A(n)) \le \frac{1}{n} \mathrm{E}(|f|). \tag{1.52}$$
Letting $n \uparrow \infty$, so that $A(n) \downarrow \{|f| = \infty\}$; since $\mu(A(1) \le \mathrm{E}(|f|) < \infty$, we have by property ❻ of measure μ:
$$\mu(\{|f| = \infty\}) = \lim_n \downarrow \mu(A(n)) = 0.$$
- To prove (iv), let $|g| \le |f|$ on A^c, where $\mu(A) = 0$. Then
$$|g| \le \mathbb{1}_A \cdot \infty + \mathbb{1}_{A^c} \cdot |f|$$
and consequently
$$\mathrm{E}(|g|) \le \mu(A) \cdot \infty + \mathrm{E}(A^c; |f|) \le 0 \cdot \infty + \mathrm{E}(|f|) = \mathrm{E}(|f|).$$
Hence $g \in L^1$ if $f \in L^1$.
- The proof of (v) is similar to that of (iv) and is omitted here.
- The assertion (vi) is a special case of (iv) since a constant is integrable when $\mu(\Omega) < \infty$.

Note that some results can be refined as below:

- Eq.(1.52) can be strengthened as follows:
$$\lim_n n\mu(A(n)) \le \lim_n \mathrm{E}(A(n); |f|) = 0. \tag{1.53}$$
This follows from property ❻ of the measure
$$A \to \mathrm{E}(A; |f|)$$
see property (iv) of the integral for $\mathcal{F}_+$.

- Property (ii) can be strengthened as follows:
  If $B_k \in \mathcal{F}$ and $\mu(B_k) \to$ as $k \to \infty$, then

$$\lim_k \mathrm{E}(B_k; f) = 0.$$

  Proof.

  – Without loss of generality, that $f \in \mathcal{F}_+$. We have then

$$\begin{aligned}\mathrm{E}(B_k; f) &= \mathrm{E}(B_k \cap A(n); f) + \mathrm{E}(B_k \cap A(n)^c; f) \\ &\leq \mathrm{E}(A(n); f) + \mathrm{E}(B_k)n.\end{aligned}$$

  Hence

$$\limsup_k \mathrm{E}(B_k; f) \leq \mathrm{E}(A(n); f)$$

  and the result follows by letting $n \to \infty$ and using (1.53).

- If $f \in L^1, g \in L^1$, and $f \leq g$ a.e., then

$$\mathrm{E}(f) \leq \mathrm{E}(g).$$

  Proof.

  – We have, except on a null set:

$$f^+ - f^- \leq g^+ - g^-$$

  but we cannot transpose terms that may be $+\infty$! Now substitute finite versions of f and g, viz. $\bar{f}$ and $\bar{g}$, and then transpose as

$$f^+ + g^- \leq g^+ + f^-.$$

  – Applying properties (ii) and (iii) for $\mathcal{F}_+$, we obtain

$$\mathrm{E}(f^+) + \mathrm{E}(g^-) \leq \mathrm{E}(g^+) + \mathrm{E}(f^-).$$

  By the assumptions of L^1, all the four quantities above are finite numbers. Transposing back we obtain the desired conclusion. ■

- If $f \in L^1, g \in L^1$, then $f + g \in L^1$, and

$$\mathrm{E}(f + g) = \mathrm{E}(f) + \mathrm{E}(g).$$

There is a more practical form for the Theorem 35, which is known as **Lebesgue's dominated convergence theorem**.

Theorem 40 (Lebesgue)**.** *For $f_n \in \mathcal{F}$, suppose: a) $\lim_n f_n = f$ a.e.; and b) there exists $\phi \in L^1$ such that for all n, $|f|_n \leq \phi$ a.e., then*

$$\lim_n \mathrm{E}(|f_n - f|) = 0.$$

Proof.

- Observe first that

$$\begin{aligned}\left|\lim_n f_n\right| &\leq \sup_n |f|_n \\ |f_n - f| &\leq |f|_n + |f| \leq 2 \sup_n |f_n|\,;\end{aligned}$$

provided the left members are defined. Since the union of a countable collection of null sets in a null set, under the hypothesis a) and b) there is a null set A such that on $\Omega - A$, we have $\sup_n |f_n| \leq \phi$ hence by Theorem 39 (iv), all $|f_n|$, $|f|$, $|f_n - f|$ are integrable, and therefore we can substitute their finite versions without affecting their integrals, and moreover $\lim_n |f_n - f| = 0$ on $\Omega - A$. (Remember that $f_n - f$ need not be defined before the substitutions!). By using Theorem 12 (ii) once more if need be, we obtain the conclusion from the positive version of Theorem 35. ■

- Moreover, when $\mu(\Omega) < \infty$, any constant M is integrable and may be used for ϕ; hence in this case the result is called **bounded convergence theorem**.

Curiously, the best known part of the theorem is the corollary below with a fixed $B \in \mathcal{F}$.

Corollary 41. *We have*

$$\lim_n \int_B f_n \,\mathrm{d}\,\mu = \int_B f \,\mathrm{d}\,\mu$$

uniformly in $B \in \mathcal{B}$.

Proof.

- This result is trivial from Theorem 40. Actually, in alternative notation:

$$|\mathrm{E}(B; f_n) - \mathrm{E}(B; f)| \leq \mathrm{E}(B; |f_n - f|) \leq \mathrm{E}(|f_n - f|).$$

- In the particular case where $B = \Omega$, the Corollary contains a number of useful results such as the integration term by term of power series or Fourier series. ■

1.7 Applications

1.7.1 Measure, Integral and Probability

Table 1.4: Notations and Interpretations for Sets and Events

Notation	Set Theory	Probability Theory
ω	element or point	outcome, sample point, elementary event
Ω	set of points	sample space; certain event
$\mathcal{F}$	σ-algebra of subsets	σ-algebra of events
$A \in \mathcal{F}$	set of points	event(if $\omega \in A$, we say the event A occurs)
$A^c, \bar{A}, \Omega \setminus A$	complement of A	event that A does not occur
$A \cup B$	union of A and B	event that either A or B occurs
$A \cap B, AB$	intersection of A and B	event that both A and B occur
$\emptyset$	empty set	impossible event
$A \cap B = \emptyset$	A and B are disjoint	events A and B are mutually exclusive
$A + B$	sum of sets, i.e. union of disjoint sets	event that one of two mutually events occurs
$A \setminus B, A - B$	difference of A and B	event that A occurs and B does not
$A \Delta B$	symmetric difference of sets i.e. $(A \setminus B) \cup (B \setminus A)$	event that A or B occurs, but not both
$\bigcup_{n=1}^{\infty} A_n$	union of the sets $A_1, A_2, \cdots$	event that at least one of $A_1, A_2, \cdots$ occurs
$\sum_{n=1}^{\infty} A_n$	sum, i.e. union of pairwise disjoint sets $A_1, A_2, \cdots$	event that one of the exclusive events $A_1, A_2, \cdots$ occurs
$\bigcap_{n=1}^{\infty} A_n$	intersection of $A_1, A_2, \cdots$	event that all the events $A_1, A_2, \cdots$ occur
$A_n \uparrow A$, $A = \lim_{n\to\infty} \uparrow A_n$	$A_1 \subseteq A_2 \subseteq \cdots$ and $A = \bigcup_{n=1}^{\infty} A_n$	the increasing sequence of events converges to event A
$A_n \downarrow A$, $A = \lim_{n\to\infty} \downarrow A_n$	$A_1 \supseteq A_2 \supseteq \cdots$ and $A = \bigcap_{n=1}^{\infty} A_n$	the decreasing sequence of events converges to event A
$\overline{\lim_{n\to\infty}} A_n$, $\limsup_{n\to\infty} A_n$, $\{A_n \text{ i.o.}\}$	the set $\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$	event that infinitely many of events $A_1, A_2, \cdots$ occur i.o. = infinitely often
$\underline{\lim_{n\to\infty}} A_n$, $\liminf_{n\to\infty} A_n$	the set $\bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k$	event that all the events $A_1, A_2, \cdots$ occur with possible exception of a finite number of them

1.7.2 Lebesgue Integral

A function $f : \mathbb{R} \to \mathbb{R}^* = [-\infty, +\infty]$ is called a **Borel function** iff $f \in \mathcal{B}$; it is called a **Lebesgue-measurable** function iff $f \in \mathcal{B}^*$.

The domain of definition of f, namely $\mathrm{Dom}(f)$, may be an arbitrary Borel set or Lebesgue measurable set D. This case is reduced to that for $D = \mathbb{R}$ by extending the value of f to be zero outside D, i.e.,

$$\hat{f} = \begin{cases} f, & \text{on } \mathrm{Dom}(f); \\ 0, & \text{outside } \mathrm{Dom}(f). \end{cases}$$

Thus $\hat{f}|_{\mathrm{Dom}(f)} = f$.

The integral of $f \in \mathbb{R}^*$ corresponding to the measure m^* constructed from F is denoted by

$$\mathrm{E}(f) = \int_{-\infty}^{\infty} f(x) \,\mathrm{d}\, F(x).$$

In case $F(x) \equiv x$, this is called the **Lebesgue integral** of f; in this case the usual notation is, for $A \in \mathcal{B}^*$:

$$\int_A f(x) \,\mathrm{d}\, x = \mathrm{E}(A; f).$$

1.7.3 Examples

Example 42. *Let I be a bounded interval in $\mathbb{R}$, $\{u_k\}$ a sequence of function on I; and for $x \in I$:*

$$s_n(x) = \sum_{k=1}^{n} u_k(x), \quad n \in \mathbb{N}.$$

Suppose the infinite series $\sum_k u_k(x)$ converges I; then in the notation:

$$\lim_{n\to\infty} \sum_{k=1}^{n} u_k(x) = \sum_{k=1}^{\infty} u_k(x) = s(x)$$

exists and is finite. Now suppose each u_k is Lebesgue-integral, then so is each s_n, by property (iii) of the integral; and

$$\int_I s_n(x) \,\mathrm{d}\, x = \sum_{k=1}^{n} \int_I u_k(x) \,\mathrm{d}\, x.$$

QUESTION*: Does the numerical series above converge? And if so the sum of integrals equal to the integral of the sum:*

$$\sum_{k=1}^{\infty} \int_I u_k(x) \,\mathrm{d}\, x \stackrel{?}{=} \int_I \sum_{k=1}^{\infty} u_k(x) \,\mathrm{d}\, x = \int_I s(x) \,\mathrm{d}\, x$$

This is the problem of integration term by term.

A very special but important case is when the interval $I = [a, b]$ is compact and the functions u_k are all continuous in I. If we assume that the series $\sum_{k=1}^{\infty} u_k(x)$ converges uniformly in I, then it follows from elementary analysis that the sequence of partial sums $\{s_n(x)\}$ is totally bounded, that is,

$$\sup_n \sup_x |s_n(x)| = \sup_x \sup_n |s_n(x)| < \infty.$$

Since $m(I) < \infty$, the bounded convergence theorem applies to yield

$$\lim_n \int_I s_n(x) \,\mathrm{d}\, x = \int_I \lim_n s_n(x) \,\mathrm{d}\, x.$$

The Taylor series of an analytic function always converges uniformly and absolutely in any compact subinterval of its interval of convergence. Thus the result above is fruitful.

Example 43. *Let $u_k \geq 0, u_k \in L^1$, then*

$$\int_a^b \left(\sum_{k=1}^{\infty} u_k \right) \mathrm{d}\,\mu = \sum_{k=1}^{\infty} \int_a^b u_k \,\mathrm{d}\,\mu. \tag{1.54}$$

Let $f_n = \sum_{k=1}^{n} u_k$, then $f_n \in L^1, f_n \uparrow f = \sum_{k=1}^{\infty} u_k$. Hence by monotone convergence

$$\mathrm{E}(f) = \lim_n \mathrm{E}(f_n)$$

this is (1.54).

When u_k is general the preceding result may be applied to $|u_k|$ to obtain

$$\int \left(\sum_{k=1}^{\infty} |u_k| \right) \mathrm{d}\,\mu = \sum_{k=1}^{\infty} \int |u_k| \,\mathrm{d}\,\mu.$$

If this is finite, then the same is true when $|u_k|$ is replaced by u_k^+ and u_k^-. It then follows by substraction that (1.54) is also true. This result of term-by-term integration may be regarded as a special case of the Fubini-Tonelli theorem, where one of the measures is the counting measure on $\mathbb{N}$.

Example 44. *Let $\langle I, \mathcal{B}^*, m \rangle$ be a measure space as in the preceding examples, but let $I = [a, b]$ be compact. Let f be a continuous function on I. We now reconsider the Riemann integration $\int_a^b f(x) \,\mathrm{d}\,x$ with the Borel-Lebesgue.*

Denote by $\mathcal{P}$ a partition of I as follows:

$$a = x_0 < x_1 < x_2 < \cdots < x_n = b;$$

and put

$$\lambda(\mathcal{P}) = \max_{1 \leq k \leq n} (x_k - x_{k-1}).$$

For each k, choose a point $\xi_k \in [x_{k-1}, x_k]$, and define a function $f_{\mathcal{P}}$ as follows:

$$f_{\mathcal{P}}(x) = \begin{cases} \xi_1, & \text{for } x \in [x_0, x_1]; \\ \xi_k, & \text{for } x \in (x_{k-1}, x_k], 2 \leq k \leq n. \end{cases}$$

Particular choices for ξ_k are

$$\xi_k = \arg \min_{x_{k-1} \leq x \leq x_k} f(x), \quad \text{or} \quad \xi_k = \arg \max_{x_{k-1} \leq x \leq x_k} f(x) \tag{1.55}$$

The $f_{\mathcal{P}}$ is called a **step function**; it is an approximation of f with the help of f, i.e.,

$$f(x) \sim \sum_{k=1}^{n} \xi_k \mathbb{1}_{[x_{k-1}, x_k]}$$

It is not basic since ξ_k's may be negative, but $f_{\mathcal{P}}^+$ and $f_{\mathcal{P}}^-$ are. Hence by Definition 30 and Definition 38, we have the **Riemann sum**

$$\mathrm{E}(f_{\mathcal{P}}) = \sum_{k=1}^{n} \xi_k (x_k - x_{k-1}).$$

When the ξ_k are chosen as in (1.55), the sum $\mathrm{E}(f_{\mathcal{P}})$ is called **lower sum** or **upper sum**.

Now, let $\{\mathcal{P}(n), n \in \mathbb{N}\}$ be a sequence of partitions such that $\lambda(\mathcal{P}(n)) \to 0$ as $n \to \infty$. Since f is continuous on a compact set, it is bounded. It follows that there is a constant C such that

$$\sup_{n \in \mathbb{N}} \sup_{x \in I} \left| f_{\mathcal{P}(n)}(x) \right| < C.$$

Since I is bounded, we can apply the bounded convergence theorem to conclude that

$$\lim_n \mathrm{E}(f_{\mathcal{P}(n)}) = \mathrm{E}(f).$$

The finite existence of the limit above signifies the Riemann-integrability of f, and the limit is then its **Riemann-integral** $\int_a^b f(x)\,\mathrm{d}\,x$. Thus we have proved that a continuous function on a compact interval is Riemann-integrable, and its Riemann-integral is equal to the Lebesgue integral.
Any bounded measurable function is integrable over any bounded measurable set.

For example, the function

$$\sin\frac{1}{x}, \quad x \in (0,1]$$

being bounded by 1 is integrable. But for the strict Riemannian point of view it has only an "improper" integral because $(0,1]$ is not close and the function is not continuous on $[0,1]$, indeed it is not definable there. Yet the limit

$$\lim_{\varepsilon\downarrow 0}\int_\varepsilon^1 \sin\frac{1}{x}\,\mathrm{d}\,x$$

exists and can be *defined* to be $\int_0^1 \sin\frac{1}{x}\,\mathrm{d}\,x$. As a matter of fact, the Riemann sums do converge despite the unceasing oscillation of f between 0 and 1 as $x \downarrow 0$.

Example 45. *The Riemann integral of a function on $(0,\infty)$ is called an "infinite integral" and is definable as follows:*

$$\int_0^\infty f(x)\,\mathrm{d}\,x = \lim_{n\to\infty}\int_0^n f(x)\,\mathrm{d}\,x$$

when the limit exists and is finite.

A famous example is

$$f(x) = \begin{cases} \frac{\sin x}{x}, & \text{for } x \in (0,\infty); \\ 1, & \text{for } x = 0. \end{cases} \tag{1.56}$$

This function is bounded by 1 and is continuous. A cute calculation yields the result

$$\lim_n \int_0^n \frac{\sin x}{x}\,\mathrm{d}\,x = \frac{\pi}{2}.$$

By contrast, the function $|f|$ is not Lebesgue-integrable. To show this, we use trigonometry:

$$\left(\frac{\sin x}{x}\right)^+ \geq \frac{1}{\sqrt{2}}\frac{1}{(2n+1)\pi} = C_n, \quad x \in A_n = \left(2n\pi + \frac{\pi}{4}, 2n\pi + \frac{3\pi}{4}\right).$$

Thus for $x > 0$:

$$\left(\frac{\sin x}{x}\right)^+ \geq \sum_{n=1}^\infty C_n \mathbb{1}_{A_n}(x).$$

The right member above is a basic function, with its integral

$$\sum_n C_n m(A_n) = \sum_n \frac{\pi}{\sqrt{2}(2n+1)2} = +\infty.$$

It follows that $\mathrm{E}(f^+) = +\infty$. Similarly $\mathrm{E}(f^-) = +\infty$; therefore by Definition 38 $\mathrm{E}(f)$ does not exist! This example is a splendid illustration of the following theorem.

Theorem 46. *Let $f \in \mathcal{B}$ and $f_n = f\mathbb{1}_{(0,n)}, n \in \mathbb{N}$. Then $f_n \in \mathcal{B}$ and $f_n \to f$ as $n \to \infty$. When the f_n's are "totally bounded", it **does not follow** that*

$$\lim_n \mathrm{E}(f_n) = \mathrm{E}(f) \tag{1.57}$$

*indeed $\mathrm{E}(f)$ may not exist. However, if we **add one condition**, the limit relation will hold. Actually,*

a) if $f \geq 0$, then (1.57) holds by Theorem 34;

b) if $\mathrm{E}(f)$ exists, in particular, $f \in L^1$, then (1.57) holds by Theorem 35.

Example 47. *The square of the function f in (1.56):*

$$|f(x)|^2 = \left(\frac{\sin x}{x}\right)^2, \quad x \in \Re$$

is integrable in the Lebesgue sense, and is also improperly integrable in the Riemann sense.

We have

$$|f(x)|^2 \leq \mathbb{1}_{(-1,+1)} + \mathbb{1}_{(-\infty,-1)\cup(+1,+\infty)} \frac{1}{x^2}$$

and the function on the right side is integrable, hence so is $|f|^2$.

Incredibly, we have

$$\int_0^\infty \left(\frac{\sin x}{x}\right)^2 \mathrm{d}\, x = \frac{\pi}{2} = (\text{R.I.}) \int_0^\infty \frac{\sin x}{x} \mathrm{d}\, x,$$

where we have inserted an "R.I." to warn against taking the second integral as a Lebesgue integral.

Example 48. *The notorious function*

$$\mathbb{1}_{\mathbb{Q}}(x) = \begin{cases} 1, & \text{for } x \in Q; \\ 0, & \text{for } x \notin Q. \end{cases}$$

where $\mathbb{Q}$ is the set of rational numbers in the unit interval $(0, 1)$.

The $\mathbb{1}_{\mathbb{Q}}$ is not Riemann-integrable. It is so totally discontinuous that the Riemannian way of approximating it, **horizontally** so to speak, fails utterly. But of course it is ludicrous even to consider this indicator function rather than the set $\mathbb{Q}$ itself. There was a historical reason for this folly: integration was regarded as the inverse operation to differentiation, so that to integrate was meant to "find the primitive" whose derivative is to be the integrand, for example $\int x \mathrm{d}\, x = \frac{1}{2}x^2 + C$. A primitive is called "indefinite integral", and $\int_1^2 x \mathrm{d}\, x$ e.g. is called a "definite integral". Thus the unsolvable problem was to find $\int_0^x \mathbb{1}_{\mathbb{Q}}(t) \mathrm{d}\, t, 0 < x < 1$.

The notion of measure as length, area, and volume is much more ancient than Newton's fluxion (derivative), not to mention the primitive measure of counting with fingers (and toes). The notion of "countable additivity" of a measure, although seemingly natural and facile, somehow did not take hold until Borel saw that

$$m(\mathbb{Q}) = \sum_{q \in \mathbb{Q}} m(q) = \sum_{q \in \mathbb{Q}} 0 = 0.$$

There can be no question that the "length" of a single point q is zero. Euclid gave it "zero dimension".

This is the beginning of MEASURE. An INTEGRAL is a weighted measure, as is obvious from Definition 30. The rest is approximation, **vertically** as in Definition 33, and convergence, as in all analysis.

1.8 Product Measure and Fubini-Tonelli Theorem

1.8.1 Product Measures

Let $\langle \Omega_1, \mathcal{F}_1, \mu_1 \rangle$ and $\langle \Omega_2, \mathcal{F}_2, \mu_2 \rangle$ be two σ-finite measure spaces. Let

$$\Omega = \Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \omega_k \in \Omega_k, k = 1, 2\} \tag{1.58}$$

$$\mathcal{S} = \{E \times G : E \in \mathcal{F}_1, G \in \mathcal{F}_2\} \tag{1.59}$$

$$\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2 \tag{1.60}$$

Sets in $\mathcal{F}$ are called **rectangles**. It is easy to see that $\mathcal{S}$ is a semi-algebra:

$$(E \times G) \cap (C \times D) = (E \cap C) \times (G \cap D)$$
$$(E \times G)^c = (E^c \times G) \cup (E \times G^c) \cup (E^c \times G^c)$$

$\mathcal{F}$ is a σ-algebra generated by $\mathcal{S}$.

Theorem 49. *There is a unique measure μ on $\mathcal{F}$ with*

$$\mu(E \times G) = \mu_1(E) \times \mu_2(G) \tag{1.61}$$

for each $E \times G$ in $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$. μ is often denoted by $\mu_1 \times \mu_2$.

PROOF.

- It is enough to show that if $\{E_i \times G_i\}$ is a partition of $E \times G$, i.e., $E \times G = \sum_i E_i \times G_i$, then

$$\mu(E \times G) = \sum_i \mu(E_i \times G_i).$$

For each $x \in E$, let $I(x) = \{i : x \in E_i\}$. $G = \sum_{i \in I(x)} G_i$, so

$$\mathbb{1}_E(x)\mu_2(G) = \sum_i \mathbb{1}_{E_i}(x)\mu_2(G_i)$$

Integrating with respect to μ_1 and we have

$$\mu_1(E)\mu_2(G) = \sum_i \mu_1(E_i)\mu_2(G_i).$$

This completes the proof. ■

Similarly, we can prove a proposition:

Proposition 50. *Let $\mathcal{E}_0 \subset \mathcal{E}$ and $\mathcal{G}_0 \subset \mathcal{G}$ be semi-algebras with $\sigma(\mathcal{E}_0) = \mathcal{E}$ and $\sigma(\mathcal{G}_0) = \mathcal{G}$. Given a measure μ_1 on $\mathcal{E}$ and a measure μ_2 on $\mathcal{G}$, there is a unique measure μ on $\mathcal{E} \times \mathcal{G}$ that has $\mu(E \times G) = \mu_1(E)\mu_2(G)$ for $E \in \mathcal{E}_0$ and $G \in \mathcal{G}_0$.*

The point of this proposition is that we can define Lebesgue measure λ on $\mathbb{R}^2$ by the requirement that

$$\forall (a,b] \in \mathcal{B}, (c,d] \in \mathcal{B}, \quad \lambda\left((a,b] \times (c,d]\right) = (b-a) \cdot (d-c)$$

Using Proposition 50 and induction, it follows that if $\langle \Omega_i, \mathcal{F}_i, \mu_i \rangle, i = 1, 2, \cdots, n$, are σ-finite measure spaces and $\Omega = \Omega_1 \times \cdots \times \Omega_n$, there is a unique measure μ on the σ-algebra $\mathcal{F}$ generated by sets of the form $A_1 \times \cdots \times A_n, A_i \in \mathcal{F}_i$, that has

$$\mu(A_1 \times \cdots \times A_n) = \prod_{i=1}^{n} \mu_i(A_i). \tag{1.62}$$

When $\langle \Omega_i, \mathcal{F}_i, \mu_i \rangle = \langle \mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda \rangle$ for all i, the result is Lebesgue measure on the Borel subsets of n dimensional Euclidean space $\mathbb{R}^n$.

Let

$$E_{\omega_1} = \{\omega_2 : (\omega_1, \omega_2) \in E\}$$

be the **cross-section** at ω_1. If $\omega_1 = x$, then E_{ω_1} will be denoted by E_x. Similarly,

$$E_{\omega_2} = \{\omega_1 : (\omega_1, \omega_2) \in E\}$$

be the **cross-section** at ω_2. If $\omega_2 = y$, then E_{ω_2} will be denoted by E_y.

Two technical things that need to be proved before we discuss the famous Fubini's Theorem

- When ω_1 is fixed, $\omega_2 \to f(\omega_1, \omega_2)$ is $\mathcal{F}_2$ measurable.
- $\omega_1 \to \int_{\Omega_2} f(\omega_1, \omega_2)\mu_2(\mathrm{d}\,\omega_2)$ is $\mathcal{F}_1$ measurable.

Lemma 51. *If $E \in \mathcal{F}$ then $E_{\omega_1} \in \mathcal{F}_2$.*

Proof. $(E^c)_{\omega_1} = (E_{\omega_1})^c$ and $\left(\bigcup_i E_i\right)_{\omega_1} = \bigcup_i (E_i)_{\omega_1}$, so if $\mathcal{E}$ is the collection of sets E for which $E_{\omega_1} \in \mathcal{F}_2$, then $\mathcal{E}$ is a σ-algebra. Since $\mathcal{E}$ contains the rectangles, the result follows.

Lemma 52. *If $E \in \mathcal{F}$ then $g(x) \equiv \mu_2(E_{\omega_1})$ is $\mathcal{F}_1$ measurable and*

$$\int_{\Omega_1} g\,\mathrm{d}\,\mu_1 = \mu(E).$$

Proof. Omitted.

Notice that it is not obvious that the collection of sets for which the conclusion is true is a σ-algebra since $\mu(E_1 \cup E_2) = \mu(E_1) + \mu(E_2) - \mu(E_1 \cap E_2)$.

1.8.2 Fubini's Theorem

Returning to the case in which $\langle \Omega, \mathcal{F}, \mu\rangle$ is the product of two measure spaces, $\langle \Omega_1, \mathcal{F}_1, \mu_1\rangle$ and $\langle \Omega_2, \mathcal{F}_2, \mu_2\rangle$, our key goal is to prove:

Theorem 53 (Fubini)**.** *If $f \geq 0$ or $\int |f|\,\mathrm{d}\,\mu < \infty$ then*

$$\int_{\Omega_1}\int_{\Omega_2} f(\omega_1, \omega_2)\mu_2(\mathrm{d}\,\omega_2)\mu_1(\mathrm{d}\,\omega_1) = \int_{\Omega_1\times\Omega_2} f\,\mathrm{d}\,\mu = \int_{\Omega_2}\int_{\Omega_1} f(\omega_1, \omega_2)\mu_1(\mathrm{d}\,\omega_1)\mu_2(\mathrm{d}\,\omega_2) \tag{1.63}$$

We will prove only the first equality, since the second one is similar.

Proof. The Fubini's theorem can be verified in four increasingly more general special cases.

Case-1 If $E \in \mathcal{F}$ and $f = \mathbb{1}_E$, then (1.63) follows from Lemma 52.

Case-2 Since each integral is linear in f, it follows that (1.63) holds for simple functions.

Case-3 Now if $f \geq 0$ and we let

$$f_n(x) = \min\left(\frac{\lfloor 2^n f(x)\rfloor}{2^n}, n\right) = \min(f^{(n)}, n),$$

then the f_n are simple and $f_n \uparrow f$, so it follows from the monotone convergence theorem that (1.63) holds for all $f \geq 0$.

Case-4 The general case now follows by writing $f = f^+ - f^-$ and applying Case-3 to f^+, f^-, and $|f|$. ■

To illustrate why the various hypothesis of Theorem 53 are needed, we will now give some examples where the conclusion fails.

Example 54. *Let $\Omega_1 = \Omega_2 = \{1, 2, \cdots\}$ with $\mathcal{F}_1 = \mathcal{F}_2 =$ all subsets and $\mu_1 = \mu_2 =$ counting measure. For $j \geq 1$, let $f(j, j) = 1$ and $f(j+1, j) = -1$, and let $f(j, k) = 0$ otherwise. We claim that*

$$\sum_m \sum_n f(m, n) = 1, \quad \text{but} \quad \sum_n \sum_m f(m, n) = 0.$$

A picture is worth several dozen words, see Figure 54. In words, if we sum the columns first, the first one gives us a 1 and the others 0, while if we sum the rows each one gives us a 0.

	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	0	0	0	$+1$	$\cdots$
$\uparrow$	0	0	$+1$	-1	$\cdots$
n	0	$+1$	-1	0	$\cdots$
	$+1$	-1	0	0	$\cdots$
		m	$\rightarrow$		

Figure 1.1: Array for $f(j,k)$ such that the sums are not interchangeable

Example 55. *Let $X=(0,1), Y=(0,\infty)$, both equipped with the Borel sets and Lebesgue measure. Let $f(x,y)=\mathrm{e}^{-xy}-2\mathrm{e}^{-2xy}$. Then we can find that*

$$\int_0^1\int_1^\infty f(x,y)\,\mathrm{d}\,y\,\mathrm{d}\,x=\int_0^1 x^{-1}(\mathrm{e}^{-x}-\mathrm{e}^{-2x})\,\mathrm{d}\,x>0,$$
$$\int_1^\infty\int_0^1 f(x,y)\,\mathrm{d}\,x\,\mathrm{d}\,y=\int_0^\infty y^{-1}(\mathrm{e}^{-2y}-\mathrm{e}^{-y})\,\mathrm{d}\,y<0.$$

Example 56. *Let $\Omega_1=(0,1)$ with $\mathcal{F}_1=$ Borel sets and $\mu_1=$ Lebesgue measure. Let $\Omega_2=(0,1)$ with $\mathcal{F}_2=$ all subsets and $\mu_2=$ counting measure. Let $f(\omega_1,\omega_2)=1$ if $\omega_1=\omega_2$ and 0 otherwise.*

$$\forall\omega_1,\quad \int_{\Omega_2} f(x,y)\mu_2(\mathrm{d}\,\omega_2)=1\Longrightarrow\int_{\Omega_1}\int_{\Omega_2} f(\omega_1,\omega_2)\mu_2(\mathrm{d}\,\omega_2)\mu_1(\mathrm{d}\,\omega_1)=1,$$
$$\forall\omega_2,\quad \int_{\Omega_1} f(x,y)\mu_1(\mathrm{d}\,\omega_1)=0\Longrightarrow\int_{\Omega_2}\int_{\Omega_1} f(\omega_1,\omega_2)\mu_1(\mathrm{d}\,\omega_2)\mu_2(\mathrm{d}\,\omega_1)=0.$$

Part II

Fundamentals

Chapter 2

Banach Spaces and Fixed-Point Theorems

> The role of functional analysis has been decisive exactly in connection with classical problems. Almost all problems are on the applications, where functional analysis enables one to focus on a specific set of concrete analytical tasks and organize material in a clear and transparent form so that you know what the difficulties are.
>
> Concrete and functional analysis exist today in an inextricable symbiosis. When someone writes down a system of axioms, no one is going to take them seriously, unless they arise from some intuitive body of concrete subject matter that you would really want to study, and about which you really want to find out something.
>
> Flex E. Browder, 1975

In a Banach space, the so-called norm

$$\begin{aligned} \|\cdot\| : \quad & \mathcal{V} \to [0, \infty), \\ & u \mapsto \|u\| . \end{aligned}$$

This generalizes the absolute value $|u|$ of a real number u. The norm can be used in order to define the **convergence**

$$\lim_{n\to\infty} u_n = u$$

by means of

$$\lim_{n\to\infty} \|u_n - u\| = 0.$$

As a standard example for a Banach space we will consider the space $C[a, b]$ which consists of all continuous functions $u : [a, b] \to \mathbb{R}$ along with the norm

$$\|u\| \triangleq \max_{x\in[a,b]} |u(x)| , \quad -\infty < a < b\infty.$$

Figure 2.1 shows the relations between Banach spaces and other important notions. Figure 2.1 tells us that each Banach space is also a normed space, etc. In this chapter, we will discuss two fundamental

fixed-point theorems of **Banach** and **Schauder**

along with applications to integral equations and ordinary differential equations, see Figure and Figure. Furthermore, the fundamental

implicit function theorem

is a simple consequence of the Banach fixed-point theorem.

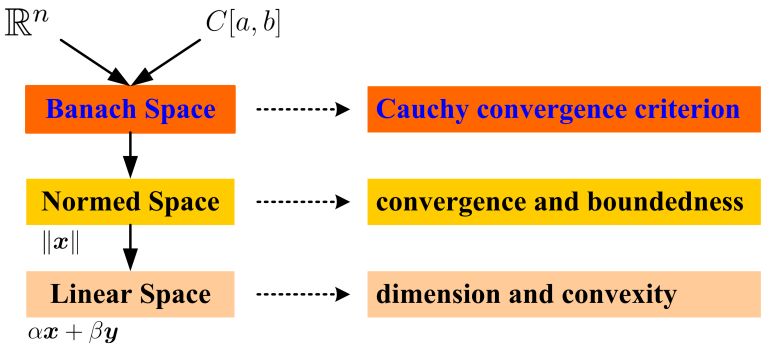
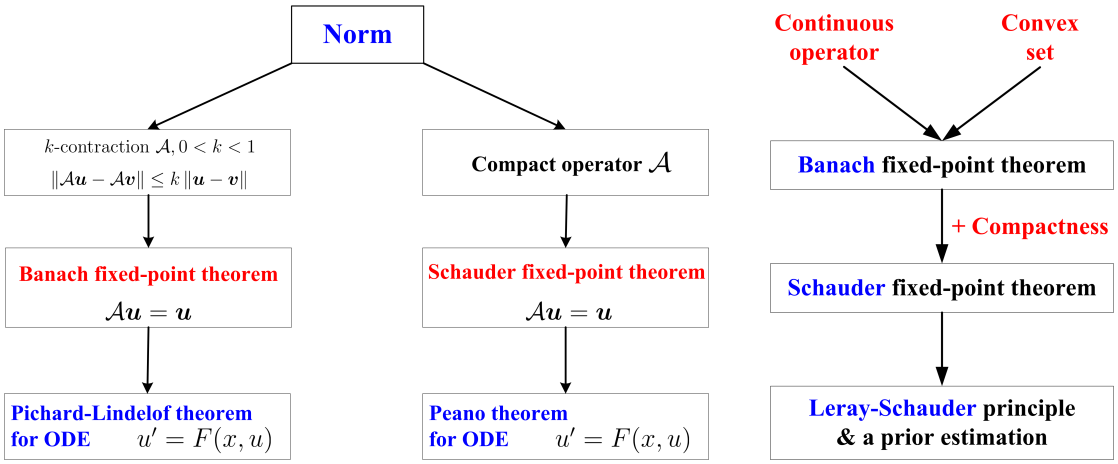


Figure 2.1: Banach space v.s. linear space



(a) Fixed-point with applications to ODE

(b) Fixed-point and Compactness

Figure 2.2: Fixed-point theorems in Banach space

2.1 Normed Spaces and convergence

Let $\mathbb{K}$ be a field of numbers and recall that $\mathbb{K}$ is $\mathbb{R}$ or $\mathbb{C}$.

Definition 57. *Let $\mathcal{X}$ be a linear space over $\mathbb{K}$. Then X is called a normed space over $\mathbb{K}$ iff there exists a norm $\|\cdot\|$ on $\mathcal{X}$, i.e., for all $\boldsymbol{u},\boldsymbol{v}\in\mathcal{X}$ and $\alpha\in\mathbb{K}$, the following are ture:*

① $\|\boldsymbol{u}\|\ge 0$ *(i.e. $\|\boldsymbol{u}\|$ is a nonnegative real number).*

② $\|\boldsymbol{u}\|$ *iff* $\boldsymbol{u}=0$.

③ $\|\alpha\boldsymbol{u}\|=|\alpha|\cdot\|\boldsymbol{u}\|$.

④ $\|\boldsymbol{u}+\boldsymbol{v}\|\le\|\boldsymbol{u}\|+\|\boldsymbol{v}\|$*(triangle inequality).*

A normed space over $\mathbb{K}=\mathbb{R}$ or $\mathbb{K}=\mathbb{C}$ is called a *real* or *complex* normed space, respectively. The number $\|\boldsymbol{u}-\boldsymbol{v}\|$ is called the distance between the two points $\boldsymbol{u}$ and $\boldsymbol{v}$. In particular, $\|\boldsymbol{u}\|$ is the distance between the two points $\boldsymbol{u}$ and the origin $\boldsymbol{v}=0$.

Example 58. *Let $\mathcal{X}=\mathbb{R}$. We set*

$$\|\boldsymbol{u}\|=\|u\mathbf{1}\|\triangleq|u|,\quad \forall\boldsymbol{u}=u\mathbf{1}\in\mathbb{R},$$

where $|u|$ denotes the absolute value of the real number u. Then, X becomes a real normed space.

Example 59. *Let $\mathcal{X} = \mathbb{C}$. We set*

$$\|\boldsymbol{u}\| \triangleq |u| = \sqrt{x^2 + y^2}, \quad \forall \boldsymbol{u} = x + \mathrm{i}y \in \mathbb{C}.$$

where $|u|$ denotes the absolute value of the complex number u. Then $\mathcal{X}$ becomes a complex normed space.

These two examples show that the norm generalizes the absolute value of numbers.

Proposition 60 (Generalized Triangle Inequality)**.** *Let $\mathcal{X}$ be normed space. Then, for all $\boldsymbol{u}, \boldsymbol{v} \in X$,*

$$|\|\boldsymbol{u}\| - \|\boldsymbol{v}\|| \leq \|\boldsymbol{u} \pm \boldsymbol{v}\| \leq \|\boldsymbol{u}\| + \|\boldsymbol{v}\|. \tag{2.1}$$

Proof. Omitted.

With the help of mathematical induction, we have the following inequality for a finite sum:

$$\left\| \sum_{j=1}^{N} \boldsymbol{x}_j \right\| \leq \sum_{j}^{n} \|\boldsymbol{x}_j\|.$$

Definition 61. *Let $\{\boldsymbol{u}_n\}_{n=1}^{\infty}$ be a sequence in the normed space $\mathcal{X}$, i.e., $\boldsymbol{u}_n \in \mathcal{X}$ for all n. We write*

$$\lim_{n\to\infty} \boldsymbol{u}_n = \boldsymbol{u} \tag{2.2}$$

iff $\lim_{n\to\infty} \|\boldsymbol{u}_n - \boldsymbol{u}\| = 0$.

We say that the sequence $(\boldsymbol{u}_n) = \{\boldsymbol{u}_n\}_{n=1}^{\infty}$ *converge*s to $\boldsymbol{u}$. Instead of (2.2), we also write $\boldsymbol{u}_n \to \boldsymbol{u}$ as $n \to \infty$. Intuitively, the *convergence* (2.2) means that the distance $\boldsymbol{u}_n - \boldsymbol{u}$ between points $\boldsymbol{u}_n$ and $\boldsymbol{u}$ goes to zero as $n \to \infty$.

Proposition 62. *Let $\mathcal{X}$ be a normed space over $\mathbb{K}$. Let $\boldsymbol{u}_n, \boldsymbol{v}_n, \boldsymbol{u}, \boldsymbol{v} \in \mathcal{X}$ and $\alpha_n, \alpha \in \mathbb{K}$ for all $n = 1, 2, \cdots$. Then the following are met:*

① *The limit point $\boldsymbol{u}$ in (2.2) is uniquely determined.*

② *If $\boldsymbol{u}_n \to \boldsymbol{u}$ as $n \to \infty$, then the sequence $\{\boldsymbol{u}_n\}$ is bounded, i.e., there exists a number $r \geq 0$ such that $\|\boldsymbol{u}_n\| \leq r$ for all n.*

③ *If $\boldsymbol{u}_n \to \boldsymbol{u}$ as $n \to \infty$, then*

$$\|\boldsymbol{u}_n\| \to \|\boldsymbol{u}\|, \quad \text{as } n \to \infty.$$

④ *If $\boldsymbol{u}_n \to \boldsymbol{u}$ and $\boldsymbol{v}_n \to \boldsymbol{v}$ as $n \to \infty$, then*

$$\boldsymbol{u}_n + \boldsymbol{v}_n \to \boldsymbol{u} + \boldsymbol{v}, \text{as } n \to \infty.$$

⑤ *If $\boldsymbol{u}_n \to \boldsymbol{u}$ and $\alpha_n \to \alpha$ as $n \to \infty$, then*

$$\alpha_n \boldsymbol{u}_n \to \alpha \boldsymbol{u} \quad \text{as } n \to \infty.$$

Obviously, this proposition is similar with the case of usual convergent sequence of numbers in analysis. So does the proof!

Definition 63 (Cauchy sequence)**.** *The sequence $\{\boldsymbol{u}_n\}$ in the normed space $\mathcal{X}$ is called a Cauchy sequence iff, for each $\varepsilon > 0$, there is a number $n_0(\varepsilon)$ such that*

$$\|\boldsymbol{u}_n - \boldsymbol{u}_m\| < \varepsilon, \quad \forall n, m \geq n_0(\varepsilon).$$

Note that the Cauchy sequence is also named with *fundamental sequence*.

Proposition 64. *In a normed space, each convergent sequence is Cauchy/fundamental.*

PROOF.

- Let $\boldsymbol{u}_n \to \boldsymbol{u}$ as $n \to \infty$. Hence $\|\boldsymbol{u}_n - \boldsymbol{u}\| \to 0$ as $n \to \infty$, i.e., for each $\varepsilon > 0$, there is a number $n_0(\varepsilon)$ such that
$$\|\boldsymbol{u}_n - \boldsymbol{u}\| < \frac{\varepsilon}{2}, \quad \forall n \geq n_0(\varepsilon).$$
This implies
$$\|\boldsymbol{u}_n - \boldsymbol{v}_m\| = \|(\boldsymbol{u}_n - \boldsymbol{u}) + (u - \boldsymbol{v}_m)\| \leq \|\boldsymbol{u}_n - \boldsymbol{u}\| + \|u - \boldsymbol{v}_m\| < \varepsilon \quad \forall n, m \geq n_0(\varepsilon).$$

QUESTION: Is a Cauchy sequence is a convergent one in general? No! We need the property of completeness, which depends on the implementation of the norm, to cope with this problem.

2.2 Banach Spaces and the Cauchy Convergence Criterion

Definition 65. *The normed space $\mathcal{X}$ is called a Banach space or complete normed space iff each Cauchy sequence is convergent.*

From Proposition 64 in the preceding section, we get the following so-called Cauchy convergent criterion:

In a Banach space, a sequence is convergent iff it is Cauchy/fundamental.

Example 66. *Let $N = 1, 2, \cdots$. The space $\mathcal{X} = \mathbb{K}^{N\times 1}$ is a Banach space over $\mathbb{K}$ with the norm*
$$\|\boldsymbol{x}\| = |\boldsymbol{x}|_\infty = \max_{1\leq j\leq N} |x_i|, \quad \boldsymbol{x} = [x_1, x_2, \cdots, x_N]^T.$$
Let $\boldsymbol{x}_n = [x_{11}, x_{2n}, \cdots, x_{Nn}]^T$. Then
$$\lim_{n\to\infty} \|\boldsymbol{x}_n - \boldsymbol{x}\| = 0, \quad \textit{iff } \lim_{n\to\infty} x_{kn} = x_k \quad \forall k = 1, 2, \cdots, N. \tag{2.3}$$
That is, the convergence $\boldsymbol{x}_n \to \boldsymbol{x}$ as $n \to \infty$ is equivalent to the convergence of the corresponding components.

PROOF.

- The inequality
$$|x_{kn} - x_k| \leq |\boldsymbol{x}_n - \boldsymbol{x}|_\infty = \max_{1\leq j\leq N} |x_{jn} - x_j|$$
implies statement (2.3). In fact, if $|\boldsymbol{x}_\infty - \boldsymbol{x}| \to 0$ as $n \to \infty$, then $x_{kn} \to x_k$ as $n \to \infty$ for all k, and the converse is also true.

- Now we prove that $|\cdot|_\infty$ is a norm.
  - Obviously,
  $$|\boldsymbol{x}|_\infty = 0, \quad \forall j \Longleftrightarrow \boldsymbol{x} = 0,$$
  and
  $$|\alpha\boldsymbol{x}|_\infty = \max_{1\leq j\leq N} |\alpha|\,|x_j| = |\alpha| \max_{1\leq j\leq N} |x_j| = |\alpha|\,|\boldsymbol{x}|_\infty.$$
  - Furthermore, the classical triangle inequality
  $$|x_j + y_j| \leq |x_j| + |y_j|, \forall x_i, y_j \in \mathbb{K}$$
  implies
  $$|\boldsymbol{x} + \boldsymbol{y}|_\infty = \max_{1\leq j\leq N} |x_j + y_j| \leq \max_{1\leq j\leq N} |x_j| + \max_{1\leq j\leq N} |y_j| = |\boldsymbol{x}|_\infty + |\boldsymbol{y}|_\infty.$$

- Finally, we show that $\mathcal{X}$ is a Banach space w.r.t. the norm $|\boldsymbol{x}|_\infty$. To this end, let $\{\boldsymbol{x}_n\}$ be a Cauchy sequence. Then

$$|x_{kn} - x_{km}| \le |\boldsymbol{x}_n - \boldsymbol{x}_m|_\infty < \varepsilon, \quad \forall n, m \ge n_0(\varepsilon).$$

  Thus, the sequence $\{\xi_{kn}\}$ is also Cauchy. The classical Cauchy convergence criterion implies the convergence

$$\lim_{n\to\infty} x_{kn} = x_k, \quad k = 1, \cdots, N.$$

  By (2.3), $\boldsymbol{x}_n \to \boldsymbol{x}$ as $n \to \infty$.

If we do not explicitly express the contrary, the the space $\mathbb{R}^N$ is equipped with the Euclidean norm $|\cdot|$. Note that in matrix analysis, we have the following kinds of important norms

- ℓ_1-norm or ℓ^1-norm

$$\|\boldsymbol{x}\|_1 = \sum_{j=1}^{N} |x_j| = |x_1| + \cdots + |x_N|.$$

- ℓ_2-norm or ℓ^2-norm

$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_{j=1}^{N} |x_j|^2} = \sqrt{|x_1|^2 + \cdots + |x_N|^2}.$$

- ℓ_∞-norm or ℓ^∞-norm

$$\|\boldsymbol{x}\|_\infty = \max_{1\le j\le N} |x_j| = \max(|x_1|, \cdots, |x_N|).$$

- ℓ_p-norm or ℓ^p-norm or Hölder norm

$$\|\boldsymbol{x}\|_p = \sqrt[p]{\sum_{j=1}^{N} |x_j|^p} = \sqrt[p]{|x_1|^p + \cdots + |x_N|^p}.$$

- Relation of $\|\boldsymbol{x}\|_p$ and $\|\boldsymbol{x}\|_\infty$

$$\|\boldsymbol{x}\|_\infty = \lim_{p\to\infty} \|\boldsymbol{x}\|_p = \lim_{p\to\infty} \sqrt[p]{\sum_{j=1}^{N} |x_j|^p}$$

Moreover, if we specify the metric tensor $\boldsymbol{G} = (g_{ij})$, then the ℓ_2-norm will be

$$\|\boldsymbol{x}\| = \sqrt{\sum_{i,j} g_{ij} x_i x_j} = \sqrt{\boldsymbol{x}^{\mathsf{T}} \boldsymbol{G} \boldsymbol{x}}.$$

Example 67. *Let $-\infty < a < b < \infty$. Then, $\mathcal{X} = C[a, b]$ is a real Banach space with the norm*

$$\|\boldsymbol{u}\| = \max_{x\in[a,b]} |u(x)|.$$

The convergence $\boldsymbol{u}_n = u_n(x) \to \boldsymbol{u} = u(x)$ in $\mathcal{X}$ as $n \to \infty$ means

$$\|\boldsymbol{u}_n - \boldsymbol{u}\| = \max_{x\in[a,b]} |u_n(x) - u(x)| \to 0, \quad \text{as } n \to \infty,$$

i.e., the sequence $\{u_n\}$ of continuous functions $\boldsymbol{u}_n : [a, b] \to \mathbb{R}$ converges uniformly on $[a, b]$ to the continuous function $\boldsymbol{u} : [a, b] \to \mathbb{R}$.

Proof.

- We first prove that $\|\cdot\|$ is a norm. Obviously,

$$\forall \alpha \in \mathbb{R}, \boldsymbol{u} \in \mathcal{X}, \quad \|\alpha \boldsymbol{u}\| = \max_{x\in[a,b]} |\alpha|\,|u(x)| = |\alpha| \max_{x\in[a,b]} |u(x)| = |\alpha|\,\|\boldsymbol{u}\|$$

and

$$\|\boldsymbol{u}\| = 0 \Longleftrightarrow \max_{x\in[a,b]} |u(x)| = 0 \Longleftrightarrow u(x) = 0 \text{ on } [a,b] \Longleftrightarrow u = 0 \text{ in } C[a,b].$$

Moreover, from $|u(x) + v(x)| \le |u(x)| + |v(x)|$ we get the triangle inequality

$$\|\boldsymbol{u} + \boldsymbol{v}\| \le \|\boldsymbol{u}\| + \|\boldsymbol{v}\|.$$

- Finally, we have to show that $\mathcal{X} = C[a,b]$ is a Banach space. Let $\{u_n(x)\}$ be a Cauchy sequence in $\mathcal{X}$, i.e.,

$$\|\boldsymbol{u}_n - \boldsymbol{u}_m\| = \max_{x\in[a,b]} |u_n(x) - u_m(x)| < \varepsilon, \quad \forall n, n \ge n_0(\varepsilon). \tag{2.4}$$

This implies the point-wise convergence

$$\forall x \in [a,b], \quad u_n(x) \to u(x) \quad \text{as } n \to \infty. \tag{2.5}$$

Letting $m \to \infty$ in (2.4), we obtain

$$\max_{x\in[a,b]} |u_n(x) - u(x)| \le \varepsilon \quad \forall n \ge n_0(\varepsilon).$$

Thus, the convergence in (2.5) is uniform on the interval $[a,b]$. By a classical result, this implies the continuity of the limit function $u : [a,b] \to \mathbb{R}$. Hence $\boldsymbol{u} \in \mathcal{X}$ and

$$\boldsymbol{u}_n \to \boldsymbol{u} \quad \text{as } n \to \infty. \qquad \blacksquare$$

Proposition 68. *Let $\{\boldsymbol{u}_n\}$ be a Cauchy sequence in the normed space X over $\mathbb{K}$, which has a convergent subsequence $\{\boldsymbol{u}_{n_j}\}$, that is,*

$$\boldsymbol{u}_{n_j} \to \boldsymbol{u} \quad \textit{in } \mathcal{X} \quad \textit{as } n \to \infty.$$

Then, the entire sequence converges to $\boldsymbol{u}$, i.e., $\boldsymbol{u}_n \to \boldsymbol{u}$ in $\mathcal{X}$ as $n \to \infty$.

Proof.

- Let $\varepsilon > 0$ be given. There is an $n_0(\varepsilon)$ such that

$$|\boldsymbol{u}_n - \boldsymbol{u}_m| < \varepsilon/2, \quad \forall n, m \ge n_0(\varepsilon).$$

- Since $\{\boldsymbol{u}_{n_j}\}$ converges to $\boldsymbol{u}$, there exists some fixed index m such that

$$\|\boldsymbol{u}_m - \boldsymbol{u}\| < \varepsilon/2, \quad m \ge n_0(\varepsilon).$$

By the triangle inequality,

$$\|\boldsymbol{u}_n - \boldsymbol{u}\| \le \|\boldsymbol{u}_n - \boldsymbol{u}_m\| + \|\boldsymbol{u}_m - \boldsymbol{u}\| \le \varepsilon, \quad n \ge n_0(\varepsilon).$$

Hence $\boldsymbol{u}_n \to \boldsymbol{u}$ as $n \to \infty$.

Corollary 69. *Suppose that*

$$\sum_{j=1}^{\infty} \|\boldsymbol{u}_{j+1} - \boldsymbol{u}_j\| < \infty,$$

where $\{\boldsymbol{u}_n\}$ is a sequence in a normed space $\mathcal{X}$ over $\mathbb{K}$. Then $\{\boldsymbol{u}_n\}$ is a Cauchy sequence in $\mathcal{X}$.

Proof.

- By the triangle inequality, for all $k \in 1, 2, \cdots$, we get

$$\|\boldsymbol{u}_n - \boldsymbol{u}_{n+k}\| \le \sum_{j=n}^{\infty} \|\boldsymbol{u}_{j+1} - \boldsymbol{u}_j\| \to 0 \quad \text{as } n \to \infty.$$

2.3 Product Spaces

Definition 70. *Let $\mathcal{X}_1, \cdots, \mathcal{X}_n$ be normed spaces over $\mathbb{K}$. The product space*

$$\mathcal{X}_1 \times \cdots \times \mathcal{X}_n$$

consists of all the n-tuples

$$(u_1, \cdots, u_n), \quad \text{where } u_k \in \mathcal{X}_k \text{ for } k = 1, \cdots, n.$$

For $\alpha \in \mathbb{K}$, we set

$$\alpha(u_1, \cdots, u_n) \triangleq (\alpha_1 u_1, \cdots, \alpha_n u_n),$$
$$(u_1, \cdots, u_n) + (v_1, \cdots, v_n) \triangleq (u_1 + v_1, \cdots, u_n + v_n),$$

and

$$\|(u_1, \cdots, u_n)\| \triangleq \sum_{k=1}^{n} \|u_k\|. \tag{2.6}$$

Then, $\mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ becomes a normed space over $\mathbb{K}$.

Proposition 71. *If $\mathcal{X}_1, \cdots, \mathcal{X}_n$ are Banach spaces, then so is the product space $\mathcal{X}_1 \times \cdots \times \mathcal{X}_n$.*

Proof.

- Let us consider the case where $n = 2$. The general case proceeds analogously.
- Suppose that the sequence of the points (u_n, v_n) is Cauchy in $\mathcal{X}_1 \times \mathcal{X}_2$. Then

$$\|(u_n, v_n) - (u_m, v_m)\| = \|u_n - u_m\| + \|v_n - v_m\| < \varepsilon, \quad \forall n, m \geq n_0(\varepsilon).$$

Thus, (u_n) and (v_n) are Cauchy in $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively. Hence

$$u_n \to u \quad \text{in} \quad \mathcal{X}_1 \quad \text{and} \quad v_n \to v \quad \text{in} \quad \mathcal{X}_2 \quad \text{as} \quad n \to \infty.$$

This implies

$$\|(u_n, v_n) - (u_m, v_m)\| = \|u_n - u_m\| + \|v_n - v_m\| \to 0, \quad \text{as} \quad n \to \infty,$$

that is, $(u_n, v_n) \to (u, v)$ in $\mathcal{X}_1 \times \mathcal{X}_2$ as $n \to \infty$. ■

By the same argument, it follows from (2.6) that

$$(u_1^{(k)}, \cdots, u_n^{(k)}) \to (u_1, \cdots, u_n) \quad \text{in} \quad \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \quad \text{as} \quad k \to \infty$$

iff all the components converge, that is,

$$u_m^{(k)} \to u_m \quad \text{in} \quad \mathcal{X}_m \quad \text{as} \quad k \to \infty$$

for all $m = 1, \cdots, n$.

2.4 Open and Closed Sets

Definition 72. *Let $\mathcal{X}$ be a normed space. For fixed $\boldsymbol{u}_0 \in \mathcal{X}$ and $\varepsilon > 0$, the set*

$$U(\boldsymbol{u}_0, \varepsilon) \triangleq \{\boldsymbol{u} \in \mathcal{X} : \|\boldsymbol{u} - \boldsymbol{u}_0\| < \varepsilon\}$$

is called an ε-neighborhood of the point $\boldsymbol{u}_0$.

By comparing with the classical analysis, we have

- The subset $M \subset \mathcal{X}$ is called *open* if, for each point $\boldsymbol{u}_0 \in M$, there is some ε-neighborhood $U(\boldsymbol{u}_0, \varepsilon)$ such that
$$U(\boldsymbol{u}_0, \varepsilon) \subset M.$$
- The subset M of $\mathcal{X}$ is called *closed* iff the set
$$M^c = \mathcal{X} \setminus M = \mathcal{X} - M = \{\boldsymbol{u} \in \mathcal{X} : \boldsymbol{u} \notin \mathcal{X}\}$$
is open.
- By an *open neighborhood* $U(\boldsymbol{u})$ of point $\boldsymbol{u}$, we understand an open subset of $\mathcal{X}$ containing $\boldsymbol{u}$.

Proposition 73. *Let $M \subset \mathcal{X}$, where $\mathcal{X}$ is a normed space. Then, the following are equivalent:*

① *M is closed.*

② *It follows from $\boldsymbol{u}_n \in M$ for all n and*
$$\boldsymbol{u}_n \to \boldsymbol{u} \quad \text{as } n \to \infty$$
that $u \in M$.

Proof.

- ①$\Longrightarrow$ ②. Let $\boldsymbol{u}_n \to \boldsymbol{u}$ as $n \to \infty$ and $\boldsymbol{u}_n \to M$ for all n. We have to show that $\boldsymbol{u} \in M$. If this is not true, then $\boldsymbol{u} \in \mathcal{X} - M$. Since the set $\mathcal{X} - M$ is open, there is some ε-neighborhood $U(\boldsymbol{u}, \varepsilon)$ such that
$$U(\boldsymbol{u}, \varepsilon) \subset \mathcal{X} - M.$$
From $\|\boldsymbol{u}_n - \boldsymbol{u}\| \to 0$ as $n \to \infty$ we get $\|\boldsymbol{u}_m - \boldsymbol{u}\| < \varepsilon$ for some index m, and hence
$$\boldsymbol{u}_m \in U(\boldsymbol{u}, \varepsilon) \subset \mathcal{X} - M,$$
i.e., $\boldsymbol{u}_m \in \mathcal{X} - M$. This contradicts $\boldsymbol{u}_m \in M$.
- ②$\Longrightarrow$ ①. Suppose that the set M is not closed, i.e., the set $\mathcal{X} - M$ is not open. Then, there exists a point
$$\boldsymbol{u} \in \mathcal{X} - M$$
such that *no* ε-neighborhood $U(\boldsymbol{u}, \varepsilon)$ is contained in the set $\mathcal{X} - M$. Thus, choosing $\varepsilon = \frac{1}{n}, n = 1, 2, \cdots$, we get a sequence $\{u_n\}$ such that
$$\boldsymbol{u}_n \in U\left(\boldsymbol{u}, \frac{1}{n}\right) \quad \text{and} \quad \boldsymbol{u}_n \in M \quad \forall n.$$
Hence
$$\|\boldsymbol{u}_n - \boldsymbol{u}\| \leq \frac{1}{n} \to 0, \quad \text{as} \quad n \to \infty.$$
By ②, $\boldsymbol{u} \in M$. This contradicts $\boldsymbol{u} \in \mathcal{X} - M$. ■

Example 74 (Ball)**.** *Let $\mathcal{X}$ be a normed space. For fixed $\boldsymbol{v} \in \mathcal{X}$ and fixed $r > 0$, define*
$$B(\boldsymbol{v}, r) = \{\boldsymbol{x} \in \mathcal{X} : \|\boldsymbol{x} - \boldsymbol{v}\| < r\}.$$
Then, $B(\boldsymbol{v}, r)$ is a closed ball of radius r around the point $\boldsymbol{v}$.

Proof.

- Let $\boldsymbol{x}_n \in B(\boldsymbol{v}, r)$ for all n, i.e.,
$$\|\boldsymbol{x}_n - \boldsymbol{v}\| \leq r, \quad \forall n.$$
If $\boldsymbol{x}_n \to \boldsymbol{x}$ as $n \to \infty$, then $\|\boldsymbol{x} - \boldsymbol{v}\| \leq r$, and hence $\boldsymbol{x} \in B$. ■

Note that the Ball $B(\boldsymbol{v}, r)$ is important for the signal analysis in the information theory and communication engineering since the action of noise on the signal points can be taken as shifting the signal by some statistical distance or geometrical distance, say σ or d.

2.5 Operators

2.5.1 Definition

Definition 75. *Let X and Y be sets. An* OPERATOR

$$\begin{aligned} A : X &\to Y \\ x &\mapsto y \end{aligned}$$

associates to each point x in X a point y in Y denoted by $y = Ax$.

The set

$$\mathrm{Dom}(A) = \{x \in X : Ax \in Y\}$$

is called the **domain of definition** of A. The set

$$\mathrm{Range}(A) = \{y \in Y : y = Ax, x \in \mathrm{Dom}(A)\}.$$

is called the **range** of A, and it can also be denoted by $A(M)$ such that $M = \mathrm{Dom}(A)$. Obviously we have

$$\mathrm{Dom}(A) \subset X, \quad \mathrm{Range}(A) \subset Y.$$

The operators are also called *functions* or *mappings*.
For operator $A : X \to Y$, if $Y = \mathbb{K}$, the A is called a **functional**.

2.5.2 Surjection, Injection and Bijection

- The operator $A : X \to Y$ is called *surjective* iff

$$\mathrm{Range}(A) = Y.$$

- The operator $A : X \to Y$ is called *injective* iff

$$Au = Av \Longrightarrow u = v.$$

- The operator $A : X \to Y$ is called *bijective* iff A is both surjective and injective.

2.5.3 Inverse

If the operator $A : X \to Y$ is bijective, then there exists the so-called inverse operator

$$A^{-1} : Y \to X$$

defined through

$$A^{-1}y = x \Longleftrightarrow Ax = y.$$

This definition makes sense, since for each given $y \in Y$, there exists exactly one $u \in \mathrm{Dom}(A)$ such that $Ax = y$. The set

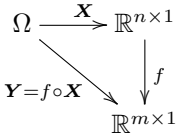
$$A^{-1}(N) = \{u \in \mathrm{Dom}(A) : Ax \in N\}$$

is called the **preimage** of the set N. Particularly, if $N = \{y\}$ is a set consists of only one element, the set

$$A^{-1}(y) = \{u \in \mathrm{Dom}(A) : y = Ax\}$$

is called a **fiber**.

Note that the concept of preimage is very important in measure theory and probability. Actually, for the probability space $\langle \Omega, \mathcal{F}, \Pr \rangle$, the measurable function (random vector) $\boldsymbol{X} : \Omega \to \mathbb{R}^{n\times 1}$, and the mapping $f : \mathbb{R}^{n\times 1} \to \mathbb{R}^{n\times 1}$, see the following diagram,



we have

$$\begin{aligned}\hat{\Pr}[\boldsymbol{Y} \in E] &= \hat{\Pr}[f \circ \boldsymbol{X} \in E] \\ &= \Pr[(f \circ \boldsymbol{X})^{-1}(E)] \\ &= \Pr[(\boldsymbol{X}^{-1} \circ f^{-1})(E)] \\ &= \Pr[\boldsymbol{X}^{-1}(f^{-1}(E))], \quad \forall E \subset \mathcal{B}(\mathbb{R}^{m\times 1}).\end{aligned}$$

2.5.4 Examples

Example 76. *Let $X = C^2(\mathbb{R})$, for each $x \in C^2(\mathbb{R})$, we define the operator L as*

$$\mathsf{L}(x) = a\frac{\mathrm{d}^2 x}{\mathrm{d}\, t^2} + b\frac{\mathrm{d}\, x}{\mathrm{d}\, t} + cx,$$

where $a, b, c \in \mathbb{R}$ are fixed constants.

This is the so-called *differential operator* for the *linear time-invariant* (LTI) second order system, which is widely in control theory and engineering.

Example 77. *Let $-\infty < a < b < \infty$, and let the function*

$$F : [a, b] \times [a, b] \times \mathbb{R} \to \mathbb{R}$$

be continuous. We set

$$(\mathsf{A}u)(x) \triangleq \int_a^x F(x, y, u(y))\, \mathrm{d}\, y, \quad \forall x \in [a, b],$$

and

$$(\mathsf{B}u)(x) \triangleq \int_a^b F(x, y, u(y))\, \mathrm{d}\, y, \quad \forall x \in [a, b].$$

Then, we obtain the two operators

$$\mathsf{A} : C[a, b] \to C[a, b], \quad \mathsf{B} : C[a, b] \to C[a, b].$$

In fact, it is a well-known classical result that the continuity of the function

$$u : [a, b] \to \mathbb{R}$$

implies the continuity of the two functions

$$\begin{aligned}\mathsf{A}u &: [a, b] \to \mathbb{R}, \\ \mathsf{B}v &: [a, b] \to \mathbb{R},\end{aligned}$$

i.e., $u \in C[a, b]$ implies both $\mathsf{A}u \in C[a, b]$ and $\mathsf{B}u \in C[a, b]$. The operators A and B are called *integral operator*s.

Example 78 (Sturm-Liouville)**.** *The Sturm-Liouville eigenvalue problem involves a general quadratic form*

$$Q : C^1[a,b] \to \mathbb{R}$$
$$\phi \mapsto \int_a^b \left[p(x) \left(\frac{\mathrm{d}\,\phi}{\mathrm{d}\,x} \right)^2 + q(x) \left(\phi(x) \right)^2 \right] \mathrm{d}\,x$$

where ϕ is restricted to functions that satisfy the boundary conditions

$$\phi(a) = 0, \quad \phi(b) = 0.$$

Obviously, Q is a non-linear functional since $Q(c_1\phi_1 + c_2\phi_2) \neq c_1 Q\phi_1 + c_2 Q\phi_2$ for any $c_1, c_2 \in \mathbb{K}$ and $\phi_1, \phi_2 \in C^1[a,b]$.

Example 79 (Sturm-Liouville ODE)**.** *Let $p : \mathbb{R} \to \mathbb{R}^+, \rho : \mathbb{R} \to \mathbb{R}^+, q : \mathbb{R} \to \mathbb{R}$ and*

$$\mathsf{L}\phi = \frac{\mathrm{d}}{\mathrm{d}\,x} \left[p(x) \frac{\mathrm{d}\,\phi}{\mathrm{d}\,x} \right] + \left[\lambda\rho(x) - q(x) \right] \phi = 0, \quad \phi \in C^1[a,b]$$

where λ is a parameter. The boundary condition for ϕ is

$$\phi(a) = 0, \quad \phi(b) = 0.$$

Then L is a linear differential operator.

Note that this ODE is very important for the physics and engineering problems, for illustration in the fields of electrodynamics, quantum mechanics, microwave and antennas, and so on.

2.6 Banach Fixed-Point Theorem and Iterative Method

The Banach fixed-point theorem represents a fundamental convergence theorem for a broad class of iteration methods.

Nonlinear functional analysis is the study of operators lacking the property of linearity. In this section, we consider nonlinear operator equations and their numerical solution. We begin the consideration of operator equations which take the form

$$u = \mathsf{A}u, \qquad u \in M \subset \mathcal{V} \tag{2.7}$$

Here M is a subset of a Banach space $\mathcal{V}$, and $\mathsf{A} : M \to M$. The solutions of this equation are called fixed points of the operator A, as they are left unchanged by A. The most important method for analyzing the solvability theory for such equations is the Banach fixed-point theorem.

2.6.1 Banach Fixed-point Theorem

Definition 80. *We say an operator $\mathsf{A} : M \subset \mathcal{V} \to \mathcal{V}$ is contractive with contractivity constant $\alpha \in [0,1)$, or α-contractive, if*

$$\|\mathsf{A}u - \mathsf{A}v\|_{\mathcal{V}} \leq \alpha \|u - v\|_{\mathcal{V}}, \quad \forall u, v \in M.$$

The operator A is called non-expansive if

$$\|\mathsf{A}u - \mathsf{A}v\|_{\mathcal{V}} \leq \|u - v\|_{\mathcal{V}}, \quad \forall u, v \in M,$$

and Lipschitz continuous if there exists a constant $L \geq 0$ such that

$$\|\mathsf{A}u - \mathsf{A}v\|_{\mathcal{V}} \leq L \|u - v\|_{\mathcal{V}}, \quad \forall u, v \in M.$$

We see the following implications:

$$\begin{aligned} \text{contractivity} &\Longrightarrow \text{non-expansiveness} \\ &\Longrightarrow \text{Lipschitz continuity} \\ &\Longrightarrow \text{continuity.} \end{aligned}$$

Theorem 81 (Banach Fixed-point Theorem)**.** *Assume that M is a nonempty closed set in a Banach space $\mathcal{V}$, and further, that $\mathsf{A} : M \to M$ is an α-contractive mapping, $0 \le \alpha < 1$. Then the following results hold:*

① *Existence and Uniqueness: There exists a unique $u \in M$ such that*

$$u = \mathsf{A}u$$

② *Convergence: For any $u_0 \in M$, the sequence $(u_n) \subset M$ defined by*

$$u_{n+1} = \mathsf{A}u_n, \quad n = 0, 1, 2, \cdots$$

converges to u:

$$\|u_n - u\|_{\mathcal{V}} \to 0 \quad as \quad n \to \infty.$$

③ *Error estimations of the iteration:*

- *A priori estimation:*
$$\|u_n - u\|_{\mathcal{V}} \le \frac{\alpha^n}{1-\alpha} \|u_0 - u_1\|_{\mathcal{V}}, \tag{2.8}$$
which can be used to determine the number of iterations needed to achieved a prescribed solution accuracy before actual computations take place.
- *A posteri estimation:*
$$\|u_n - u\|_{\mathcal{V}} \le \frac{\alpha}{1-\alpha} \|u_{n-1} - u_n\|_{\mathcal{V}}, \tag{2.9}$$
which gives a computable error bound once some numerical solutions are calculated.
- *Improvement via one iteration:*
$$\|u_n - u\|_{\mathcal{V}} \le \alpha \|u_{n-1} - u\|_{\mathcal{V}} \tag{2.10}$$

This theorem is called by a variety of names in the literature, with the *contractive mapping theorem* another popular choice. It is also called *Picard iteration* in settings related to differential equations. For simplicity, $\|\cdot\|_{\mathcal{V}}$ will be denoted by $\|\cdot\|$.

Proof

- Since $\mathsf{A} : M \to M$, the sequence (u_n) is well defined.
- Let us first prove that (u_n) is a Cauchy sequence. Using contractivity of the mapping A, we have
$$\|u_{n+1} - u_n\| \le \alpha \|u_n - u_{n-1}\| \le \cdots \le \alpha^n \|u_1 - u_0\|.$$
Then for any $m > n \ge 1$,
$$\begin{aligned} \|u_m - u_n\| &= \left\| \sum_{k=1}^{m-n-1} (u_{n+k+1} - u_{n+k}) \right\| \\ &\le \sum_{k=1}^{m-n-1} \|u_{n+k+1} - u_{n+k}\| \\ &\le \sum_{k=1}^{m-n-1} \alpha^{n+k} \|u_1 - u_0\| \\ &\le \frac{\alpha^n}{1-\alpha} \|u_1 - u_0\|. \end{aligned} \tag{2.11}$$

Since $\alpha \in [0,1), \|u_m - u_n\| \to 0$ ans $m, n \to \infty$. Thus (u_n) is a Cauchy sequence; and since M is a closed set in the Banach space $\mathcal{V}$, (u_n) has a limit $u \in M$. We take the limit $n \to \infty$ in $u_{n+1} = \mathsf{A}u_n$ to see that $u = \mathsf{A}u$ by the continuity of A, i.e., u is a fixed-point of A.

- Suppose $x_1, x_2 \in M$ are both fixed-point of A. Then from $x_1 = \mathsf{A}x_1$ and $x_2 = \mathsf{A}x_2$, we obtain

$$x_1 - x_2 = \mathsf{A}x_1 - \mathsf{A}x_2.$$

  Hence

$$\|x_1 - x_2\| = \|\mathsf{A}x_1 - \mathsf{A}x_2\| \leq \alpha \|x_1 - x_2\|$$

  which implies $\|x_1 - x_2\| = 0$ since $\alpha \in [0,1)$. So a fixed-point of a contractive mapping is unique.

- Now we prove the error estimates.
  - Letting $m \to \infty$ in (2.11), we get the estimate (2.8).
  - From $\|u_n - u\| = \|Tu_{n-1} - Tu\| \leq \alpha \|u_{n-1} - u\|$ we obtain the estimate (2.10).
  - From $\|u_n - u\| = \|Tu_{n-1} - Tu\| \leq \alpha \|u_{n-1} - u\|$ and $\|u_{n-1} - u\| \leq \|u_{n-1} - u_n\| + \|u_n - u\|$ we obtain the estimate (2.9). ■

Proposition 82. *Let $\mathsf{A} : \mathcal{V} \to \mathcal{V}$ be a contractive mapping. For every $y \in \mathcal{V}$, the equation*

$$u = \mathsf{A}u + y$$

has a unique solution, call it $u(y)$ and $u(y)$ is a continuous function of y.

As an application of the Banach fixed-point theorem, we consider the unique solvability of a nonlinear equation in a Hilbert space.

Theorem 83. *Let $\mathcal{V}$ be a Hilbert space. Assume that $\mathsf{A} : \mathcal{V} \to \mathcal{V}$ is strongly monotone and Lipschitz continuous, i.e., there exist two constants $c_1, c_2 > 0$ such that for any $u_1, u_2 \in \mathcal{V}$,*

$$\langle \mathsf{A}u_1 - \mathsf{A}u_2 | u_1 - u_2 \rangle \geq c_1 \|u_1 - u_2\|^2, \tag{2.12}$$

$$\|\mathsf{A}u_1 - \mathsf{A}u_2\| \leq c_2 \|u_1 - u_2\|. \tag{2.13}$$

Then for any $b \in \mathcal{V}$, there is a unique $u \in \mathcal{V}$ such that

$$\mathsf{A}u = b. \tag{2.14}$$

Moreover, the solution u depends Lipschitz continuisly on b: If $\mathsf{A}\hat{u}_1 = b_1$ and $\mathsf{A}\hat{u}_2 = b_2$, then

$$\|\hat{u}_1 - \hat{u}_2\| \leq \frac{1}{c_1} \|b_1 - b_2\|. \tag{2.15}$$

Proof

- The equation $\mathsf{A}u = b$ is equivalent to

$$u = u - \theta(\mathsf{A}u - b)$$

  for any $\theta \neq 0, \theta \in \mathbb{R}$. Define an operator $\mathsf{T}_\theta : \mathcal{V} \to \mathcal{V}$ by the formula

$$\mathsf{T}_\theta(u) \triangleq u - \theta[\mathsf{A}u - b].$$

- We now show that for $\theta > 0$ suficiently small, the operator T_θ is contractive. Write

$$\mathsf{T}_\theta(v_1) - \mathsf{T}_\theta(v_2) = (v_1 - v_2) - \theta(\mathsf{A}v_1 \mathsf{A}v_2).$$

Then

$$\|\mathsf{T}_\theta(v_1) - \mathsf{T}_\theta(v_2)\|^2 = \|v_1 - v_2\|^2 - 2\theta\langle \mathsf{A}v_1 - \mathsf{A}v_2 | v_1 - v_2\rangle + \theta^2 \|\mathsf{A}v_1 - \mathsf{A}v_2\|^2 .$$

Use the assumptions (2.12) and (2.13) to obtain

$$\|\mathsf{T}_\theta(v_1) - \mathsf{T}_\theta(v_2)\|^2 \le (1 - 2c_2\theta + c_1^2\theta^2) \|v_1 - v_2\|^2 .$$

For $\theta \in (0, 2c_2/c_1^2)$,

$$1 - 2c_2\theta + c_1^2\theta^2 < 1$$

and T_θ is a contraction. Then by the Banach fixed-point theorem, T_θ has a unique fixed-point $u \in \mathcal{V}$. Hence, the equation $\mathsf{A}u = b$ has a unique solution.

- Now we prove the Lipschitz continuity of the solution w.r.t. the right hand side. From $\mathsf{A}\hat{u}_1 = b_1$ and $\mathsf{A}\hat{u}_2 = b$, we obtain

$$\mathsf{A}\hat{u}_1 - \mathsf{A}\hat{u}_2 = b_1 - b_2.$$

Then

$$\langle \mathsf{A}u_1 - \mathsf{A}u_2 | u_1 - u_2\rangle = \langle b_1 - b_2 | u_1 - u_2\rangle.$$

Apply the assumption (2.12) and the Cauchy-Schwarz inequality, we have

$$c_1 \|u_1 - u_2\|^2 \le \|b_1 - b_2\| \|u_1 - u_2\| ,$$

which implies (2.13). ■.

The proof technique of Theorem 83 is useful when we prove the existence and uniqueness of solutions to some variational inequalities. The condition (2.12) relates to the degree of monotonicity of $\mathsf{A}u$ as u varies. For a real-valued function $\mathsf{A}u$ of a single real variable u, the constant c_1 can be chosen as the infimum of $\mathsf{A}'u$ over the domain of A, assuming this infimum is positive.

2.6.2 Applications to Iterative Methods

Nonlinear Algebraic Equations

Theorem 84. *Let $-\infty < a < b < \infty$ and $T : [a, b] \to [a, b]$ be a contractive function with contractivity constant $\alpha \in [0, 1)$. Then the following results hold:*

① *Existence and uniqueness: There exists a unique solution $x \in [a, b]$ to the equation $x = Tx$.*

② *Convergence of the iteration: For any $x_0 \in [a, b]$, the sequence $(x_n) \subset [a, b]$ defined by $x_{n+1} = T(x_n), n = 0, 1, 2, \cdots$ converges to x:*

$$x_n - x \to 0 \quad as \quad n \to \infty.$$

③ *Error estimates: The error bounds are*

$$|x_n - x| \le \frac{\alpha^n}{1 - \alpha} |x_0 - x_1|$$
$$|x_n - x| \le \frac{\alpha}{1 - x} |x_{n-1} - x_n| ,$$
$$|x_n - x| \le \alpha |x_{n-1} - x| .$$

The contractiveness of the function T is guaranteed from the assumption that

$$\|T\|_{C[a,b]} = \sup_{\substack{x\in[a,b]\\ x\neq 0}} \frac{|T(x)|}{|x|} = \sup_{x\in[a,b]} \|T'(x)\| < 1$$

Indeed, using the Mean Value Theorem, we then see that T is constractive with the contractivity constant $\alpha = \sup_{x\in[a,b]} |T'(x)|$.

Linear Algebraic Systems

Let $\boldsymbol{A} \in \mathrm{GL}(n,\mathbb{R}) \subset \mathbb{R}^{n\times n}$, then the linear system

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \quad \boldsymbol{x} \in \mathbb{R}^{n\times 1} \tag{2.16}$$

where $\boldsymbol{b} \in \mathbb{R}^{n\times 1}$ is given, has a unique solution.

① Let $\boldsymbol{A} = \boldsymbol{N} - \boldsymbol{M}$ such that $\left\|\boldsymbol{N}^{-1}\boldsymbol{M}\right\| \leq 1$, where $\|\cdot\|$ is some matrix norm operator norm, i.e., it is a norm induced by some vector norm

$$\|\boldsymbol{B}\| = \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{B}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}.$$

Thus the linear system is equivalent to

$$\begin{aligned} \boldsymbol{N}\boldsymbol{x} &= \boldsymbol{M}\boldsymbol{x} + \boldsymbol{b} \\ \boldsymbol{x} &= \boldsymbol{N}^{-1}\boldsymbol{M}\boldsymbol{x} + \boldsymbol{N}^{-1}\boldsymbol{b} \end{aligned}$$

with the correspongding iterative formula

$$\boldsymbol{x}_{n+1} = \boldsymbol{N}^{-1}\boldsymbol{M}\boldsymbol{x}_n + \boldsymbol{N}^{-1}\boldsymbol{b}$$

and contractive mapping

$$\boldsymbol{T}(\cdot) = \boldsymbol{N}^{-1}\boldsymbol{M}(\cdot).$$

② Let $\boldsymbol{A} = \boldsymbol{D} + \boldsymbol{L} + \boldsymbol{U}$, where $\boldsymbol{D}$ is the diagonal part of $\boldsymbol{A}$, $\boldsymbol{L}$ and $\boldsymbol{U}$ are the strict lower and upper triangular parts.

- If we take $\boldsymbol{N} = \boldsymbol{D}$, then we have

  $$\boldsymbol{D}\boldsymbol{x} = \boldsymbol{b} - (\boldsymbol{L} + \boldsymbol{U})\boldsymbol{x},$$

  and the corresponding iteration formula

  $$\boldsymbol{D}\boldsymbol{x}_{n+1} = \boldsymbol{b} - (\boldsymbol{L} + \boldsymbol{U})\boldsymbol{x}_n,$$

  which is the vector representation of the Jacobi method.
- If we take $\boldsymbol{N} = \boldsymbol{D} + \boldsymbol{L}$, then we obtain the Gauss-Seidel method

  $$(\boldsymbol{D} + \boldsymbol{L})\boldsymbol{x}_{n+1} = \boldsymbol{b} - \boldsymbol{U}\boldsymbol{x}_n,$$

- A more sophisticalted splitting is obtained by setting

  $$\boldsymbol{N} = \frac{1}{\omega}\boldsymbol{D} + \boldsymbol{L}, \quad \boldsymbol{M} = \frac{1-\omega}{\omega}\boldsymbol{D} - \boldsymbol{U},$$

  where $\omega \neq 0$ is an acceleration parameter. The corresponding iterative method with the (approximate) optimal choice ω is called the SOR method (*successive overrelaxation method*). For linear system arising in difference solution of some PDEs, there is a well-understood theory for the choice of an optimal value of ω; and with that optimal value, the iteration converges much more rapidly than does the original Gauss-Seidel method on which it is based.

Linear and Nonlinear Integratal Equations

We consider the *Urysohn integral equation*

$$u(x) = (\mathsf{A}u)(x) = \lambda \int_a^b F(x, y, u(y)) \,\mathrm{d}\, y = f(x), \quad x \in [a, b], \tag{2.17}$$

along with the iteration method

$$u_{n+1}(x) = (\mathsf{A}u_n)(x) = \lambda \int_a^b F(x, y, u_n(y)) \,\mathrm{d}\, y + f(x), \quad x \in [a, b], n = 0, 1, 2, \cdots \tag{2.18}$$

where $u_0(x) \equiv 0$ and $-\infty < a < b < \infty$.

Proposition 85. *Assume that*

① *The function $f : [a, b] \to \mathbb{R}$ is continuous.*

② *The function $F : [a, b] \times [a, b] \times \mathbb{R} \to \mathbb{R}$ is continuous, and the partial derivative*

$$F_u : [a, b] \times [a, b] \times \mathbb{R} \to \mathbb{R}$$

is also continuous.

③ *There is a positive number L such that the uniform Lipschitz condition holds, i.e.,*

$$|F_u(x, y, u)| \leq L, \quad \forall x, y \in [a, b], u(\cdot) \in \mathbb{R}.$$

④ *Let the real number λ be given such that $(b - a)\,|\lambda|\, L < 1$.*

⑤ *Set $\mathcal{X} \triangleq C[a, b]$ and $\|u\| = \max\limits_{x \in [a,b]} |u(x)|$.*

Then, the following hold true:

❶ *The original integral equation problem $u = \mathsf{A}u$ has a unique solotuin $u \in \mathcal{X}$.*

❷ *The sequence (u_n) constructed by $u_{n+1} = \mathsf{A}u_n$ converges to $u \in \mathcal{X}$.*

❸ *For all $n = 0, 1, 2, \cdots$ we get the following error estimates:*

$$\|u_n - \| \leq \frac{\alpha^n}{1 - \alpha} \|u_1\|,$$
$$\|u_{n+1} - u\| \leq \frac{\alpha}{1 - \alpha} \|u_{n+1} - u_n\|$$

where $\alpha \triangleq (b - a)\,|\lambda|\, L < 1$.

Proof

- For $\forall x, y \in [a, b]$ and $u(\cdot), v(\cdot) \in \mathbb{R}$, there exists a $w(\cdot) \in \mathbb{R}$ such that

$$|F(x, y, u) - F(x, y, v)| \leq |F_u(x, y, w)|\, u - v \leq L\,|u - v|$$

by the classical mean value theorem.

- Obviously, for the operator $\mathsf{A} : \mathcal{X} \to \mathcal{X}$ we have

$$\begin{aligned} \|\mathsf{A}u - \mathsf{A}v\| &= \max_{x \in [a,b]} |(\mathsf{A}u)(x) - (\mathsf{A}v)(x)| \\ &\leq |\lambda|\,(b - a) L \max_{x \in [a,b]} |u(x) - v(x)| \qquad = \alpha \|u - v\|, \quad \forall u, v \in \mathcal{X}. \end{aligned}$$

In consequence, A is α-contractive since $\alpha \in [0, 1)$.

- Let $M \triangleq \mathcal{X}$, the assertions follow from the Banach fixed-point theorem.

Example 1. (Linear Integral Equation). Let

$$F(x, y, u) \triangleq K(x, y)u, \tag{2.19}$$

and suppose that the function $K : [a, b] \times [a, b] \to \mathbb{R}$ is continuous. Then the assumptions of Proposition 85 are satisfied with the Lipschitz constant L

$$L = \max_{x,y\in[a,b]} |K(x, y)| .$$

Therefore, all the statements of Proposition 85 are ture for the integral equation (2.17) with (2.19). In the spacial case (2.19), theoriginal problem (85) is called a linear integral equation.

Example 2. (Nekrasov's Integral Equation).

$$\theta(x) = \lambda \int_0^\pi \ell(x, t) \frac{\sin \theta(t)}{1 + 3\lambda \int_0^t \sin \theta(s) \,\mathrm{d}\, x} \,\mathrm{d}\, t, \quad x \in [0, \pi] \tag{2.20}$$

where

$$\ell(x, t) = \frac{1}{\pi} \ln \frac{\sin \frac{x+t}{2}}{\frac{x-t}{2}}.$$

It is the nonzero solutions that are of interest. This arises in the study of the profile of water waves on liquid of infinite depth; and the equation involves interesting questions of solutions that bifurcate.

Example 3. (Nonlinear Volterra Integral Equations).

There are two kinds of Voterra integral equations:

- First kind
$$u(x) = \int_a^x F(x, y, u(y)) \,\mathrm{d}\, y, \quad x \in [a, b] \tag{2.21}$$
- Sencond kind
$$u(x) = f(x) + \int_a^x F(x, y, u(y)) \,\mathrm{d}\, y, \quad x \in [a, b] \tag{2.22}$$

Example 4. (Fourier Transform and Integral Equations). The Fourier transform is also related to integral equation: Suppose the function $\hat{f}(\omega)$ is known, find the function $f(t)$ such that

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{j}\omega t} f(t) \,\mathrm{d}\, t,$$

where the $K(\omega, t) = K(t, \omega)$ is symmetric and $|K(\omega, t)| \equiv 1$.

Ordinary Differential Equations in Banach Spaces

Let $\mathcal{V}$ be a Banach space and consider the initial value problem

$$\begin{cases} \dfrac{\mathrm{d}\, u}{\mathrm{d}\, t} = f(t, u(t)), & |t - t_0| < a, \\ u(t_0) = z. \end{cases} \tag{2.23}$$

Here $z \in \mathcal{V}$ and $f : [t_0 - a, t_0 + a] \times \mathcal{V} \to \mathcal{V}$ is continuous. For example, f could be an integral operator. This differential equaiton problem is equivalent to the integral equation

$$u(t) = z + \int_{t_0}^t f(s, u(s)) \,\mathrm{d}\, s, \quad |t - t_0| < a, \tag{2.24}$$

which is of the form $u = \mathcal{A}u$. This leads naturally to the fixed point iteration method

$$u_{n+1}(t) = z + \int_{t_0}^t f(s, u_n(s)) \,\mathrm{d}\, s, \quad |t - t_0| < a, n = 0, 1, 2, \cdots . \tag{2.25}$$

Denote, for $b > 0$,

$$Q_b \triangleq \{(t,u) \in \mathbb{R} \times \mathcal{V} : |t - t_0| \le a, \|u - z\| \le b\}.$$

We have the following existence and solvability theory for (2.23).

Theorem 86 (Generalized Picard-Lindel'of Theorem)**.** *Assume $f : Q_b \to \mathcal{V}$ is continuous and is uniformly Lipschitz continuous with respect to its second argument:*

$$\|f(t,u) - f(t,v)\| \le L\|u - v\|, \quad \forall (t,u),(t,v) \in Q_b,$$

where L is a constant independent of t. Let

$$M = \max_{(t,u)\in Q_b} \|f(t,u)\|$$

and

$$a_0 = \min\left\{a, \frac{b}{M}\right\}.$$

Then the initial value problem (2.23) has a unique continuously differentiable solution $u(\cdot)$ on $[t_0 - a_0, t_0 + a_0]$; and the iterative method (2.25) converges for any initial value u_0 for which $\|z - u_0\| < b$,

$$\max_{|t-t_0|\le a_0} \|u_n(t) - u(t)\| \to 0 \quad as \quad n \to \infty.$$

Moreover, with $\alpha = 1 - \mathrm{e}^{-La_0}$, the error

$$\ell_{err} = \max_{|t-t_0|\le a_0} \|u_n(t) - u(t)\| \, \mathrm{e}^{-L|t-t_0|}$$

is bounded by each of the following:

$$\begin{aligned}
\ell_{err} &\le \frac{\alpha^n}{1-\alpha} \max_{|t-t_0|\le a_0} \|u_1(t) - u_0(t)\| \, \mathrm{e}^{-L|t-t_0|}, \\
\ell_{err} &\le \frac{\alpha}{1-\alpha} \max_{|t-t_0|\le a_0} \|u_n(t) - u_{n-1}(t)\| \, \mathrm{e}^{-L|t-t_0|}, \\
\ell_{err} &\le \alpha \max_{|t-t_0|\le a_0} \|u_{n-1}(t) - u_0(t)\| \, \mathrm{e}^{-L|t-t_0|}.
\end{aligned}$$

Chapter 3

Hilbert Spaces

In Hilbert spaces, an inner product $\langle u|v\rangle$ is defined, allowing us to introduce the fundamental notion of orthogonality. Fig. 3.1 shows the relationship between Hilbert spaces and Banach spaces,

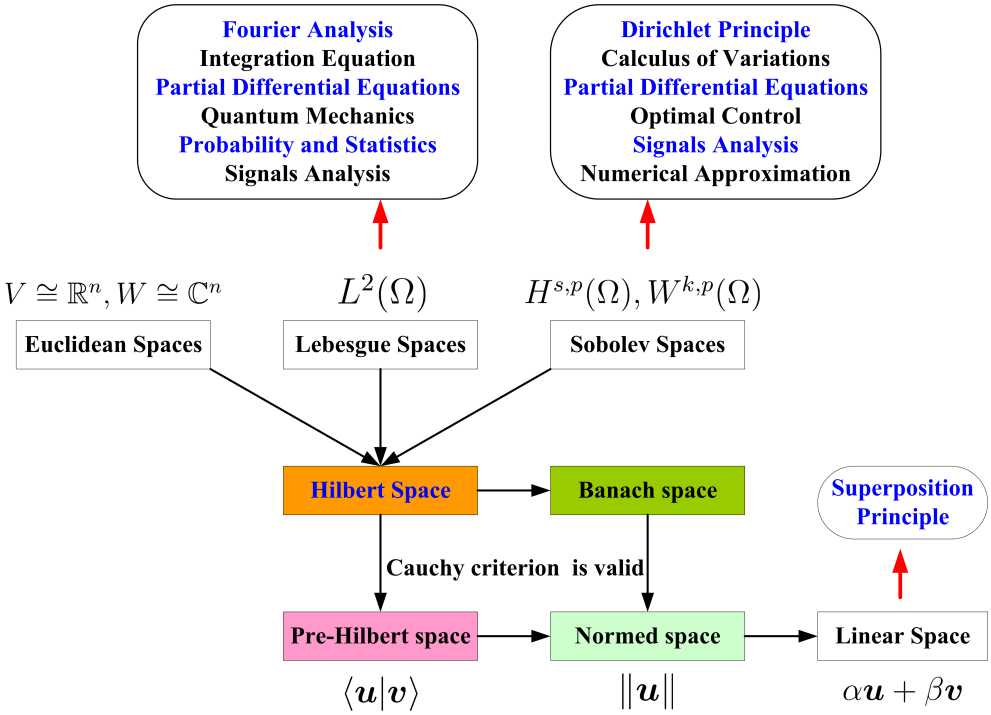


Figure 3.1: Hilbert Spaces and its Applications

and others, i.e., each Hilbert space is a Banach space, and so forth. With a view to applications, the most important Hilbert spaces are the Lebesgue spaces $L^2(\Omega)$ and the related Sobolev spaces $\mathrm{H}^m(\Omega)$ and $\mathrm{W}^{m,p}(\Omega)$. Roughly speaking, the Lebesgue over $\mathbb{F}$ ($\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$) consists of all functions

$$u : \Omega \subset \mathbb{F}^{n\times 1} \to \mathbb{F}$$

with

$$\int_\Omega |u(x)|^2 \,\mathrm{d}\,\mu < \infty.$$

The corresponding *inner product*

$$\langle u|v\rangle \triangleq \int_\Omega \overline{u(x)} v(x) \,\mathrm{d}\,\mu$$

generalizes the classic Euclidean inner product

$$\langle \boldsymbol{u} | \boldsymbol{v} \rangle = \sum_{j=1}^{n} \overline{u_j} v_j$$

on $\mathbb{R}^{n\times 1}$ and $\mathbb{C}^{n\times 1}$, where $\boldsymbol{u} = [u_1, \cdots, u_n]^{\mathsf{T}}$, $\boldsymbol{v} = [v_1, \cdots, v_n]^{\mathsf{T}}$ and the bar denotes complex conjugate. Observe that the integral $\int_\Omega \cdots \mathrm{d}\,\mu$ is to be understood in the sense of Lebesgue.

The theory of Hilbert spaces forces the use of the Lebesgue integral.

The deeper reason for this is the fact that in case of the classical Riemann integral the limiting relation

$$\lim_{n\to\infty} \int_\Omega u_n(x)\,\mathrm{d}\,x = \int_\Omega u(x)\,\mathrm{d}\,x$$

is only valid under very *restrictive* assumptions, in contrast to the Lebesgue integral. The Riemann integral leads only to pre-Hilbert spaces for which the fundamental Cauchy criterion is *not* valid.

The logical structure of this chapter is shown in Fig. 3.2

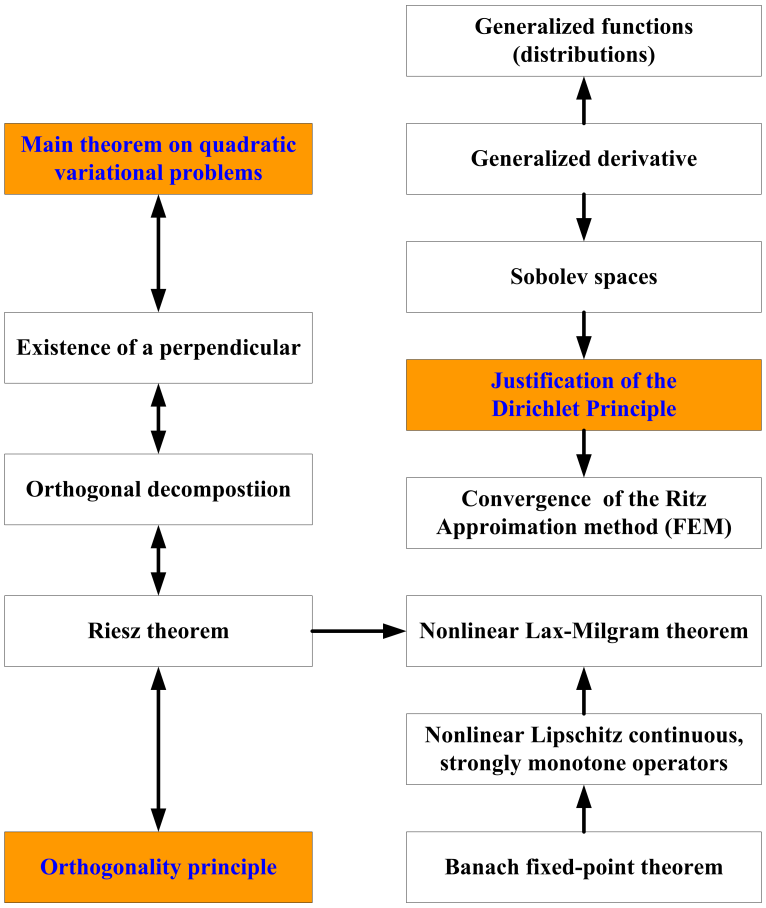


Figure 3.2: The Logic of Hilbert Spaces

3.1 Hilbert Spaces

Recall that $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$.

3.1.1 Fundamental Definitions

Definition 87. *Let $\mathcal{X}$ be a linear space over $\mathbb{F}$. An inner product on $\mathcal{X}$ is denoted by*

$$\begin{aligned}\langle\cdot|*\rangle : \mathcal{X}\times\mathcal{X} &\to \mathbb{F}\\ (u,v) &\mapsto \langle u|v\rangle\end{aligned}$$

such that the following hold for all $u,v,w\in\mathcal{X}$ and $\alpha,\beta\in\mathbb{F}$:

① $\langle u|u\rangle \geq 0$ *and* $\langle u|u\rangle = 0$ *iff* $u=0$*;*

② $\langle u|\alpha v+\beta w\rangle = \alpha\langle u|v\rangle + \beta\langle u|w\rangle$*;*

③ $\overline{\langle u|v\rangle} = \langle v|u\rangle$.

With the help of Definition 87, we have:

- An *inner product space* or *pre-Hilbert space* over $\mathbb{F}$ is a linear space $\mathcal{X}$ over $\mathbb{F}$ together with an inner product.
- It follows from ① and ② that

$$\langle \alpha u+\beta v|w\rangle = \overline{\alpha}\langle u|w\rangle + \overline{\beta}\langle v|w\rangle, \quad \forall u,v,w\in\mathbb{F}, \alpha,\beta\in\mathbb{F}. \tag{3.1}$$

- Let $u,v\in\mathcal{X}$, then u is called *orthogonal* or *perpendicular* to v iff

$$\langle u|v\rangle = 0. \tag{3.2}$$

Proposition 88 (Norm induced by inner product)**.** *Each inner product space $\mathcal{X}$ over $\mathbb{F}$ is also a normed space over $\mathbb{F}$ w.r.t. the norm*

$$\|u\| \triangleq \sqrt{\langle u|u\rangle}, \forall u\in\mathcal{X}. \tag{3.3}$$

Proposition 89 (Cauchy-Schwarz Inequality)**.** *Let $\mathcal{X}$ be an inner product space, then*

$$|\langle u|v\rangle| \leq \sqrt{\langle u|u\rangle}\cdot\sqrt{\langle v|v\rangle} = \|u\|\cdot\|v\| \tag{3.4}$$

From Proposition 88 we obtain the following:

All the notions and theorems for normed spaces remain valid for inner product spaces w.r.t. the norm $\|u\| = \sqrt{\langle u|u\rangle}$.

Particularly, the convergence

$$u_n \to u, \quad \text{as} \quad n\to\infty$$

in the inner product space $\mathcal{X}$ is to be understood in the following sense:

$$\|u_n - u\| \to 0, \quad \text{as} \quad n\to\infty.$$

Proposition 90. *Let $\mathcal{X}$ be an inner product, then the following hold true:*

- *The inner product is continuous, that is,*

$$u_n\to u \quad \textit{and} \quad v_n\to v \quad \textit{as} \quad n\to\infty$$

imply $\langle u_n|v_n\rangle \to \langle u|v\rangle$ as $n\to\infty$.

- *Let M be a dense subset of $\mathcal{X}$. If*

$$\langle u|u\rangle = 0, \quad \textit{for fixed } u\in\mathcal{X} \textit{ and all } v\in M,$$

then $u=0$.

Example — Product Rule: Let X be an inner product space, and let $u, v : U(s) \subset \mathbb{R} \to \mathcal{X}$ be two functions defined on an open neighborhood of $s \in \mathbb{R}$ that are differentiable at the point s. Then the function $g : t \mapsto \langle u(t)|v(t)\rangle$ is differentiable at s, where

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\langle u(t)|v(t)\rangle_s = \langle u'(s)|v(s)\rangle + \langle u(s)|v'(s)\rangle.$$

Definition 91 (Hilbert Space)**.** *A Hilbert space is a Banach space w.r.t. to the norm induced by inner product. In other words, a linear space $\mathcal{X}$ over $\mathbb{F}$ is a Hilbert space iff the following hold:*

① *there exists an inner product $\langle\cdot|*\rangle$ in $\mathcal{X}$, and*

② *each Cauchy sequence w.r.t. the norm $\|\cdot\| = \sqrt{\langle\cdot|\cdot\rangle}$ is convergent.*

If $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$, then $\mathcal{X}$ is called a real or complex Hilbert space, respectively.

Definition 92. *A Hilbert space over $\mathbb{C}$ with dimension 2 is called a qubit.*

Proposition 93. *Each finite-dimensional inner product space is a Hilbert space.*

This follows immediately from the fact that each finite-dimensional normed space is a Banach space.

Proposition 94. *Let $\mathcal{X}$ be a Hilbert space (resp., Banach space) over $\mathbb{F}$, and let $\mathcal{L}$ be a linear subspace of $\mathcal{X}$. Then the closure $\overline{\mathcal{L}}$ is also a Hilbert space (resp., Banach space) w.r.t. the restriction of the inner product (resp., norm) on $\mathcal{X}$ to $\overline{L}$.*

Proof.

- We first prove that $\overline{\mathcal{L}}$ is a linear space over $\mathbb{F}$. In fact, let $u, v \in \overline{\mathcal{L}}$ and $\alpha, \beta \in \mathbb{F}$. Then, there are sequences (u_n) and (v_n) in $\mathcal{L}$ such that

  $$u_n \to u, \quad v_n \to v \quad \text{in} \quad \mathcal{X} \quad \text{as} \quad n \to \infty.$$

  Letting $n \to \infty$, it follows from

  $$\alpha u_n + \beta v_n \in \mathcal{L}, \quad \forall n$$

  that $\alpha u + \beta v \in \overline{\mathcal{L}}$.

- Restrict the inner product (resp., norm) on $\mathcal{X}$ to the subset $\overline{\mathcal{L}}$ of $\mathcal{X}$. Then, $\overline{\mathcal{L}}$ is an inner product space (resp., normed space).

- Finally, let (u_n) be a Cauchy sequence in $\overline{\mathcal{L}}$. Then,

  $$u_n \to u \quad \text{in} \quad \mathcal{X} \quad \text{as} \quad n \to \infty.$$

  Since $\overline{\mathcal{L}}$ is closed, $u \in \overline{\mathcal{L}}$. Hence

  $$u_n \to u \quad \text{in} \quad \overline{\mathcal{L}} \quad \text{as} \quad n \to \infty. \qquad \blacksquare$$

3.1.2 Standard Examples

In the following, all integrals are to be understood in the sense of Lebesgue.

Example 1: $\mathbb{F}^{n\times 1}$

The space $\mathcal{X} \triangleq \mathbb{F}^{n\times 1}$ is an n-dim Hilbert space over $\mathbb{F}$ with the inner product

$$\langle \boldsymbol{x}|\boldsymbol{y}\rangle \triangleq \sum_{j=1}^{n} \overline{x_j} y_j, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{F}^{n\times 1}$$

where $\boldsymbol{x} = [x_1, \cdots, x_n]^{\mathsf{T}}$ and $\boldsymbol{y} = [y_1, \cdots, y_n]^{\mathsf{T}}$. The corresponding norm is given by

$$\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}|\boldsymbol{x}\rangle} = \sqrt{\sum_{j=1}^{n} |x_j|^2} \quad \forall \boldsymbol{x} \in \mathbb{F}^{n\times 1}.$$

Example 2: $L^2(a,b)$

Suppose that $-\infty \le a < b \le \infty$. Let $L^2(a,b)$ denote the set of all measurable functions

$$u : (a,b) \to \mathbb{R}$$

such that

$$\int_a^b |u(x)|^2 \,\mathrm{d}\,\mu < \infty.$$

Then

① $L^2(a,b)$ is a real Hilbert space w.r.t. the following inner product:

$$\langle u|v\rangle \triangleq \int_a^b u(x)v(x) \,\mathrm{d}\,\mu, \quad \forall u, v \in L^2(a,b).$$

② $\dim(L^2(a,b)) = \infty$

More precisely, we use the following identification principle: Two functions u and v correspond to the same element in the Hilbert space $L^2(a,b)$ iff

$$u(x) = v(x), \quad \text{for almost all } x \in (a,b).$$

Or equivalently,

$$u(x) = v(x), \quad \forall x \in (a,b) \setminus E, \mu(E) = 0.$$

or

$$u \overset{a.e.}{=} v \quad \text{in} \quad (a,b)$$

for simplicity.

Example 3: $L^2_{\mathbb{F}}(\Omega)$

Proposition 95 ($L^2_{\mathbb{F}}(\Omega)$)**.** *Let Ω be a nonempty measurable subset of $\mathbb{R}^{n\times 1}$, $n \ge 1$ (e.g., Ω is open or closed), and let $L^2_{\mathbb{F}}(\Omega)$ denote the set of all measurable functions*

$$u : \Omega \to \mathbb{F}$$

such that

$$\int_\Omega |u(x)|^2 \,\mathrm{d}\,\mu. \tag{3.5}$$

Then we have

① *$L^2_{\mathbb{F}}(\Omega)$ is a Hilbert space w.r.t. to the following inner product*

$$\langle u|v\rangle \triangleq \int_\Omega \overline{u}v \,\mathrm{d}\,\mu, \quad \forall u, v \in L^2_{\mathbb{F}}(\Omega). \tag{3.6}$$

For $u, v \in L^2_{\mathbb{F}}(\Omega)$, u and v correspond to the same element of the Hilbert space $L^2_{\mathbb{F}}(\Omega)$ iff $u \overset{a.e.}{=} v$ in Ω. For $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$, $L^2_{\mathbb{F}}(\Omega)$ is a real or complex Hilbert space, respectively.

② *If Ω is open, then $\dim(L^2_{\mathbb{F}}(\Omega)) = \infty$.*

For brevity of notation, we set

$$L^2(\Omega) \triangleq L^2_{\mathbb{F}}(\Omega)$$

if $\mathbb{F} = \mathbb{R}$. Note also that $L^2(a,b) = L^2(\Omega)$ with $\Omega = (a,b) =]a,b[$.

3.1.3 The Space $C_0^\infty(\Omega)$

The space $C_0^\infty(\Omega)$ plays a fundamental role in modern analysis.

The Space $C_0^\infty(\Omega)$ and Density in $L^2(\Omega)$

Definition 96. *Let Ω be a nonempty open set in $\mathbb{R}^{n\times 1}$, $n \geq 1$. Then*

(a) $C^k(\Omega)$ is the set of all real functions

$$u : \Omega \to \mathbb{R}$$

that have continuous partial derivatives of orders $m = 0, 1, \cdots, k$.

(b) $C^k(\Omega)$ is the set of all $u \in C^k(\Omega)$ for which all partial derivatives of order $m = 0, 1, \cdots, k$ can be extended continuously to the closure $\overline{\Omega}$ of Ω.

(c) If $u \in C^k(\Omega)$ (resp., $u \in C^k(\Omega)$) for all $k = 0, 1, 2, \cdots$, then we write $u \in C^\infty(\Omega)$ (resp., $u \in C^\infty(\overline{\Omega})$.

(d) $C_0^\infty(\Omega)$ is the set of all functions $u \in C^\infty(\Omega)$ that vanish outside a compact subset D of Ω that depends on u, i.e., $u(x) = 0$ for all $x \in \Omega \setminus D$.

Instead of $C^0(\Omega)$ (resp., $C^0(\overline{\Omega})$) we write briefly $C(\Omega)$ (resp., $C(\overline{\Omega})$). That is, $C(\Omega)$ consists of all continuous functions $u : \Omega \to \mathbb{R}$, and $C(\overline{\Omega})$ consists of all continuous functions $u : \overline{\Omega} \to \mathbb{R}$.

The set $C_0^\infty(\Omega)_{\mathbb{C}}$ consists of all functions $u : \Omega \to \mathbb{C}$ for which both the real and the imaginary parts of u belong to $C_0^\infty(\Omega)$. Similarly, we define $C^k(\Omega)_{\mathbb{C}}$, and so on.

If $u \in C^k(\Omega)$, then we say "u is C^k on Ω". In the 1-dim special case where $\Omega = (a,b)$, we write briefly

$$C^k(a,b) \triangleq C^k(\Omega), \quad C^k[a,b] \triangleq C^k(\overline{\Omega}),$$

and so forth.

Proposition 97. *Let Ω be a nonempty open set in $\mathbb{R}^{n\times 1}, n \geq 1$. Then, the following hold true:*

① *The set $C_0^\infty(\Omega)$ is dense in $L^2(\Omega)$.*

② *The set $C(\overline{\Omega})$ is dense in $L^2(\Omega)$.*

③ *The sets $C_0^\infty(\Omega)_{\mathbb{C}}$ and $C(\overline{\Omega})_{\mathbb{C}}$ are dense in $L^2_{\mathbb{C}}(\Omega)$.*

Corollary 98. *The spaces $L^2(\Omega)$ and $L^2_{\mathbb{C}}(\Omega)$ are separable.*

Example 4. Let $-\infty < a < b < \infty$. For all $u, v \in C[a,b]$, we define

$$\langle u|v\rangle \triangleq \int_a^b uv \,\mathrm{d}\,\mu. \tag{3.7}$$

One checks easily that this is an inner product on $C[a,b]$. The corresponding inner product space is denoted by $C_*[a,b]$. Obviously, the norm on $C_*[a,b]$, namely

$$\|u\| = \sqrt{\langle u|u\rangle} = \sqrt{\int_a^b |u|^2 \,\mathrm{d}\,\mu}, \quad \forall u \in C[a,b]$$

differs from the maximum norm $\max\limits_{x\in[a,b]} |u(x)|$ introduced previously. We can conclude that

The inner product space $C_*[a,b]$ is not a Hilbert space.

PROOF.

- Let $\mathcal{L} \triangleq C_*[a,b]$ and $\mathcal{X} \triangleq L^2(a,b)$. By Proposition 97, the linear subspace $\mathcal{L}$ is dense in $\mathcal{X}$, i.e., $\overline{\mathcal{L}} = \mathcal{X}$.
- If $\mathcal{L}$ were a Hilbert space, then $\mathcal{L}$ would be closed. Hence $\mathcal{L} = \overline{\mathcal{L}} = \mathcal{X}$. But this is impossible, since there are functions with $u \in \mathcal{X}$ and $u \in \mathcal{L}$. For example, this is true for

$$u(x) \triangleq \begin{cases} 1, & \text{if } a \le x \le c \text{ for fixed } c \in [a,b]; \\ 0, & \text{if } c < x \le b. \end{cases} \qquad \blacksquare \tag{3.8}$$

$C_0^\infty(\Omega)$ and the Variational Lemma

The variational lemma plays a fundamental role in the calculus of variations.

Lemma 99 (Variational Lemma)**.** *Let Ω be a nonempty open set in $\mathbb{R}^{n\times 1}, n \ge 1$. Then, it follows from $u \in L^2(\Omega)$ and*

$$\int_\Omega uv \,\mathrm{d}\,\mu = 0, \quad \forall v \in C_0^\infty(\Omega) \tag{3.9}$$

that $u(x) \overset{a.e.}{=} 0$ in Ω.

PROOF.

- Let $\mathcal{X} \triangleq L^2(\Omega)$. By 3.9,

  $$\langle u|v\rangle = 0, \quad \forall v \in C_0^\infty(\Omega).$$

  Since the set $C_0^\infty(\Omega)$ is dense in $\mathcal{X}$, we can take $v = u$. Thus

  $$\langle u|u\rangle = \int_\Omega |u|^2 \,\mathrm{d}\,\mu. \tag{3.10}$$

  Hence $u(x) \overset{a.e.}{=} 0$ in Ω.
- If u is continuous on Ω, then (3.10) implies $u(x) = 0$ for all $x \in \Omega$.

$C_0^\infty(\Omega)$ and Integration by Parts

The classic integration-by-parts formula reads as follows:

$$\int_a^b u'v \,\mathrm{d}\,x = uv\big|_a^b - \int_a^b uv' \,\mathrm{d}\,x, \tag{3.11}$$

with the "boundary integral"

$$uv\big|_a^b = u(b)v(b) - u(a)v(a).$$

In particular, if $v(a) = v(b) = 0$, then

$$\int_a^b u'v \,\mathrm{d}\,x = -\int_a^b uv' \,\mathrm{d}\,x. \tag{3.12}$$

Proposition 100. *Let $-\infty < a < b < \infty$. Then, the following are met:*

① *The integration-by-parts formula (3.11) holds for all*

$$u, v \in C^1[a,b].$$

② *Formula (3.12) holds for all*

$$u \in C^1(a,b), \quad v \in C_0^\infty(a,b).$$

Here, we set $C^1[a,b] \triangleq C^1(\overline{\Omega})$ and $C^1(a,b) \triangleq C^1(\Omega)$, where $\Omega = (a,b)$ and so on.

Proof.

- Ad ①. By the fundamental theorem of calculus,

$$\int_a^b (u'v + uv')\,\mathrm{d}\,x = \int_a^b (uv)'\,\mathrm{d}\,x = uv|_a^b\,.$$

- Ad ②. Since the function v vanishes in a neighborhood of the two boundary points $x = a$ and $x = b$, we can choose a subinterval $[c,d]$ of (a,b) such that $v \in C_0^\infty(c,d)$. Furthermore, $u \in C^1(a,b)$ implies $u \in C^1[c,d]$. Hence

$$\int_a^b (u'v + uv')\,\mathrm{d}\,x = \int_c^d (uv)'\,\mathrm{d}\,x = uv|_c^d = 0,$$

since $v(c) = v(d) = 0$. ■

The generalization of the integration-by-parts formula (3.11) to higher dimensions reads as follows:

$$\int_\Omega (\partial_j\, u)v\,\mathrm{d}\,\boldsymbol{x} = \int_{\partial\,\Omega} uvn_j\,\mathrm{d}\,S - \int_\Omega u\,\partial_j\, v\,\mathrm{d}\,\boldsymbol{x}, \quad j = 1, \cdots, N, \tag{3.13}$$

where $\boldsymbol{x} = [x_1, \cdots, x_N]^\mathsf{T}$, $\mathrm{d}\,\boldsymbol{x} = \mathrm{d}\,x_1 \wedge \cdots \wedge \mathrm{d}\,x_N$, and

$$\partial_j\, u \triangleq \frac{\partial u}{\partial x_j}.$$

In addition, the outer unit normal vector to the boundary $\partial\,\Omega$ is denoted by $\boldsymbol{n} = [n_1, \cdots, n_N]^\mathsf{T}$. For $N = 2$, the surface integral $\int\limits_{\partial\,\Omega} \cdots \mathrm{d}\,S$ is to be understood in the sense of $\int\limits_{\partial\,\Omega} \cdots \mathrm{d}\,s$, where s denotes arclength, and the boundary curve $\partial\,\Omega$ is oriented in such a way that the set Ω lies on the left-hand side of $\partial\,\Omega$.

In the special case where $v = 0$ on $\partial\,\Omega$, formula (3.13) passes over to

$$\int_\Omega (\partial_j\, u)v\,\mathrm{d}\,\boldsymbol{x} = -\int_\Omega u\,\partial_j\, v\,\mathrm{d}\,\boldsymbol{x}, \quad j = 1, \cdots, N. \tag{3.14}$$

Proposition 101 (Integration by parts)**.** *For $N = 1, 2, \cdots$, the following hold true*

① *Formula (3.13) holds for all*

$$u, v \in C^1(\overline{\Omega}),$$

provided Ω is a nonempty bounded open set in $\mathbb{R}^{N\times 1}$ that has a sufficiently smooth boundary.

② *Formula (3.14) holds for*

$$u \in C^1(\Omega) \quad \textit{and} \quad v \in C_0^\infty(\Omega),$$

provided Ω is a nonempty open set in $\mathbb{R}^{N\times 1}$.

The integration-by-parts formula (3.13) is the key to the modern theory of Partial Differential Equations (PDF) and to the modern calculus of variations.

3.2 Bilinear Form

Definition 102. *Let $\mathcal{X}$ be a normed space over $\mathbb{F}$. By a bounded bilinear form on $\mathcal{X}$ we understand a map*

$$\begin{aligned} a : \mathcal{X} \times \mathcal{X} &\to \mathbb{F} \\ (u, v) &\mapsto a(u, v) \end{aligned}$$

that has the following properties:

① *Bilinearity. For all $u, v, w \in \mathcal{X}$ and $\alpha, \beta \in \mathbb{F}$,*

$$\begin{aligned} a(\alpha u + \beta v, w) &= \alpha a(u, w) + \beta a(v, w) \\ a(w, \alpha u + \beta v) &= \alpha a(w, u) + \beta a(w, v) \end{aligned}$$

② *Boundedness. There is a constant $d > 0$ such that*

$$|a(u, v)| \leq d \|u\| \|v\|, \quad \forall u, v \in \mathcal{X}.$$

In addition,

- $a(\cdot, \cdot)$ is called *symmetric* iff

$$a(u, v) = a(v, u), \quad \forall u, v \in \mathcal{X};$$

- $a(\cdot, \cdot)$ is called *positive* iff

$$0 \leq a(u, u), \quad \forall u \in \mathcal{X};$$

- $a(\cdot, \cdot)$ is called *strongly positive* iff there is a constant $c > 0$ such that

$$c \|u\|^2 \leq a(u, u), \quad \forall u \in \mathcal{X}.$$

Proposition 103. *Let $a : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a bounded bilinear form on the normed space $\mathcal{X}$ over $\mathbb{F}$. Then*

$$u_n \to u \quad \text{and} \quad v_n \to v \quad \text{as} \quad n \to \infty$$

imply $a(u_n, v_n) \to a(u, v)$ as $n \to \infty$.

Proof.

- Since the sequence (v_n) is bounded, we get

$$\begin{aligned} |a(u_n, v_n) - a(u, v)| &= |a(u_n, v_n) - a(u_n, v) + a(u_n, v) - a(u, v)| \\ &= |a(u_n, v_n - v) + a(u_n - u, v)| \\ &\leq |a(u_n, v_n - v)| + |a(u_n - u, v)| \\ &\leq d \|u_n\| \|v_n - v\| + d \|u_n - u\| \|v\| \to 0, \quad \text{as} \quad n \to \infty. \blacksquare \end{aligned}$$

Obviously, for the bilinear form $a(u, u)$, we have the following *key identity*:

$$2a(u, u) + 2a(v, v) = a(u - v, u - v) + a(u + v, u + v), \quad \forall u, v \in \mathcal{X} \tag{3.15}$$

Proposition 104. *Let $\mathcal{X}$ be an inner product space over $\mathbb{F}$, then we have the so-called parallelogram identity*

$$2 \|u\|^2 + 2 \|v\|^2 = \|u + v\|^2 + \|u - v\|^2. \tag{3.16}$$

If we introduce the energetic inner product for the symmetric, bounded, strongly positive, bilinear form $a(\cdot,\cdot)$ as follows:

$$\langle u|v\rangle_E \triangleq a(u,v), \quad \forall u,v \in \mathcal{X}, \tag{3.17}$$

then the key identity can be written as

$$2\|u\|_E^2 + 2\|v\|_E^2 = \|u+v\|_E^2 + \|u-v\|_E^2, \tag{3.18}$$

where $\|u\|_E^2 \triangleq \langle u|u\rangle_E$. This is precisely the parallelogram identity w.r.t. the energetic inner product.

3.3 Main Theorem on Quadratic Variational Problems

We consider the minimal problem

$$u_{opt} = \arg\min_{u\in\mathcal{X}} \mathsf{F}(u) = \frac{1}{2}a(u,u) - b(u). \tag{3.19}$$

For example, the famous Dirichlet problem can be expressed as follows:

$$\begin{gathered}\min_u \mathsf{F}(u) = \frac{1}{2}\int_\Omega \sum_{j=1}^N (\partial_j u)^2 \,\mathrm{d}\boldsymbol{x} - \int_\Omega fu\,\mathrm{d}\boldsymbol{x} \\ \text{s.t.} \quad u = g, \quad \boldsymbol{x}\in\partial\Omega.\end{gathered} \tag{3.20}$$

which says that

$$a(u,u) = \frac{1}{2}\int_\Omega \sum_{j=1}^N \left|\frac{\partial u}{\partial x_j}\right|^2 \mathrm{d}\boldsymbol{x} = \frac{1}{2}\int_\Omega |\nabla u|^2\,\mathrm{d}\boldsymbol{x}, \quad b(u) = \int_\Omega fu\,\mathrm{d}\boldsymbol{x}.$$

Theorem 105 (Main Theorem on Quadratic Variational Problems)**.** *Suppose that*

(a) $a : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ *is a symmetric, bounded, strongly positive, bilinear form on the real Hilbert space* $\mathcal{X}$.

(b) $b : \mathcal{X} \to \mathbb{R}$ *is a linear continuous functional on* $\mathcal{X}$.

Then the following hold true:

① *The variational problem (3.19) has a unique solution.*

② *Problem (3.19) is equivalent to the following so-called variational equation:*

$$a(u,v) = b(v), \quad \text{for fixed } u\in\mathcal{X} \text{ and } v\in\mathcal{X}. \tag{3.21}$$

Proof.

By hypothesis, there are constants $c > 0$ and $d > 0$ such that

$$c\|u\|^2 \le a(u,u) \le d\|u\|^2, \quad \forall u\in\mathcal{X}. \tag{3.22}$$

Proposition 106.

Step-1: Equivalent equation. We show that (3.19) is equivalent to (3.21). To this end, we set

$$\mathsf{F}(u) \triangleq \frac{1}{2}a(u,u) - b(u), \quad u\in\mathcal{X}.$$

Moreover, for fixed $u, v \in \mathcal{X}$*, we set*

$$\phi(t) \triangleq F(u+tv), \quad \forall t\in\mathbb{R}.$$

Using the symmetry condition $a(u,v) = a(v,u)$, we obtain

$$\phi(t) = \frac{1}{2}t^2 a(u,v) + t[a(u,v) - b(v)] + \frac{1}{2}a(u,v) - b(u).$$

Note that $a(u,v) \geq c\|v\|^2 > 0$ for all $v \in \mathcal{X}$ with $n \neq 0$. Thus, the original problem (3.19),

$$\min_{u\in\mathcal{X}} \mathsf{F}(u)$$

has a solution u iff the real quadratic function $\phi = \phi(t)$ has a minimum at the point $t = 0$ for each fixed $v \in \mathcal{X}$, i.e.,

$$\phi'(0) = 0. \tag{3.23}$$

Equation (3.23) is identical to

$$a(u,v) - b(v) = 0, \quad \forall v \in \mathcal{X}.$$

This is (3.21).

Step-2: Uniqueness. Let u and w be solutions of the original problem $\min_{u\in\mathcal{X}} \mathsf{F}(u)$. By Step-1, for all $v \in \mathcal{X}$,

$$a(u,v) = b(v)$$
$$a(w,v) = b(v).$$

Let $v \triangleq u - w$, we get

$$c\|u-w\|^2 \leq a(u-w, u-w) = 0.$$

Hence $u = w$, i.e., the original problem (3.19) has at most one solution.

Step-3: Existence. Set

$$\alpha \triangleq \inf_{x\in\mathcal{X}} \mathsf{F}(u).$$

Since

$$\mathsf{F}(u) = \frac{1}{2}a(u,u) - b(u) \geq \frac{1}{2}c\|u\|^2 - \|b\|\,\|u\|,$$

we obtain $\mathsf{F}(u) \to +\infty$ if $\|u\| \to +\infty$. Hence $\alpha > -\infty$. By the definition of α, there is a sequence (u_n) such that

$$F(u_n) \to \alpha, \quad as \quad n \to \infty.$$

Obviously, with the help of the key identity we have:

$$2a(u_n,u_n) + 2a(u_m,u_m) = a(u_n - u_m, u_n - u_m) + a(u_n + u_m, u_n + u_m). \tag{3.24}$$

Thus

$$\begin{aligned} F(u_n) + F(u_m) &= \frac{1}{4}a(u_n - u_m, u_n - u_m) + 2F\left(\frac{u_n + u_m}{2}\right) \\ &\geq \frac{1}{4}c\|u_n - u_m\| + 2\alpha. \end{aligned}$$

Since $F(u_n) + F(u_m) \to 2\alpha$, it follows that (u_n) is a Cauchy sequence. Hence $F : \mathcal{X} \to \mathbb{R}$ is continuous,

$$F(u_n) \to \mathsf{F}(u), \quad as \quad n \to \infty.$$

This implies

$$\mathsf{F}(u) = \alpha,$$

i.e., the limit u is a solution of the original problem $\min_{u\in\mathcal{X}} \mathsf{F}(u)$.

We remark that

- Theorem 105 is equivalent to the perpendicular principle:
  In a Hilbert space, there exists a perpendicular from each point u to each given closed linear subspace $\mathcal{L}$.
- Dirichlet principle and orthogonality:
  The functional analytic justification of the Dirichlet principle is based on the idea of orthogonality.
- The theory of Hilbert spaces is the abstract and very efficient formulation of the idea of orthogonality.

It seems that these ideas have deep roots in our real world, since the Hilbert space theory is the right mathematical tool for describing quantum theory and signal theory.

3.4 Orthogonal Projection

We consider the minimum problem

$$\min_{v\in M} \|u-v\| = \min! \quad v \in M, \tag{3.25}$$

and make the following assumptions:

(H) Let M be a closed linear subspace of the real or complex Hilbert space $\mathcal{X}$, and let $u \in \mathcal{X}$ be given.

We seek the foot v of a perpendicular from the point u to the plane M. Let $M^\perp$ denote the orthogonal complement to M, that is, by definition,

$$M^\perp \triangleq \{w \in \mathcal{X} : \langle w|v\rangle = 0,\ \forall v \in M\}$$

Theorem 107 (Perpendicular Principle)**.** *Let M be a closed linear subspace of the real or complex Hilbert space $\mathcal{X}$, and let $u \in \mathcal{X}$ be given. Then, the minimum problem*

$$\hat{v} = \arg\min_{v\in M} \|u-v\|$$

has a unique solution $\hat{v}$, and $u-\hat{v} \in M^\perp$.

Proof

- Since
  $$\|u-v\|^2 = \langle u-v|u-v\rangle = \langle u|u\rangle - \langle u|v\rangle - \langle v|u\rangle + \langle v|v\rangle,$$
  problem (3.25) is equivalent to
  $$\min_{v\in M} \frac{1}{2}a(v,v) - b(v) \tag{3.26}$$
  where
  $$a(v,w) \triangleq \Re\langle v|w\rangle, \quad b(v) \triangleq \frac{1}{2}[\langle u|v\rangle + \langle v|u\rangle] = \Re\langle u|v\rangle.$$
  Note that $a(v,v) = \langle v|v\rangle$ for all $v \in \mathcal{X}$. By the Cauchy-Schwarz inequality,
  $$|a(v,w)| \le \|v\|\,\|w\|, \quad a(v,v) \ge \|v\|^2, \quad |b(v)| \le \|u\|\,\|v\|,$$
  for all $v, w \in \mathcal{X}$.
- We now consider the different cases for field $\mathbb{F}$.

– $\mathbb{F} = \mathbb{R}$. Let $\mathcal{X}$ be a real Hilbert space; then M is also a real Hilbert space. It follows from Theorem 105 that problem (3.26) has a unique solution v. Hence the original problem (3.25) has also a unique solution v.

– $\mathbb{F} = \mathbb{C}$. Let $\mathcal{X}$ be a complex Hilbert space. Then, $\mathcal{X}$ becomes a real Hilbert space with w.r.t. the new inner product

$$\langle v|w\rangle_* \triangleq \Re\langle v|w\rangle, \quad \forall v, w \in \mathcal{X}.$$

Again by Theorem 105, problem (3.26 has a unique solution v, and so the original problem (3.25) has also a unique solution v.

– General case. Let $\mathcal{X}$ be a Hilbert space over $\mathbb{F}$. We want to show that $u - v \in M^\perp$. Since v is a solution of (3.25), we get

$$\|u - v\|^2 \leq \|u - (v + \lambda w)\|^2, \quad w \in M, \lambda \in \mathbb{F}.$$

Hence

$$\langle u - v|u - v\rangle \leq \langle u - v|u - v\rangle - \lambda\langle u - v|w\rangle - \overline{\lambda}\langle w|u - v\rangle + |\lambda|^2 \langle w|w\rangle.$$

Suppose that $u - v \neq 0$ and $w \neq 0$. Letting $\lambda \triangleq \frac{\langle w|u-v\rangle}{\|w\|^2}$, we get $0 \leq -|\langle u - v|w\rangle|^2$, and hence

$$\langle u - v|w\rangle = 0, \quad \forall w \in M.$$

This remains true if $u - v = 0$. ■

Corollary 108 (Orthogonal Decomposition)**.** *If (H) holds, then there exists a unique decomposition of u of the form*

$$u = v + w, \quad v \in M, w \in M^\perp.$$

Proof.

- Let

$$u = v_1 + w_1, \quad v_1 \in M, w_1 \in M^\perp$$

be a second decomposition of u. Then

$$0 = (v - v_1) + (w - w_1), \quad v - v_1 \in M, w - w_1 \in M^\perp.$$

Hence $0 = \langle v - v_1|w - w_1\rangle = -\langle v_1 - v|v - v_1\rangle$, i.e., $v = v_1$ and $w = w_1$. ■

Corollary 109 (Pythagorean Theorem)**.** *If u is orthogonal to v, i.e., $\langle u|v\rangle = 0$, then*

$$\|u + v\|^2 = \|u\|^2 + \|v\|^2.$$

3.5 Linear Functionals and the Riesz Theorem

Theorem 110 (Riesz Theorem)**.** *Let $\mathcal{X}$ be a Hilbert space over $\mathbb{F}$, and let $\mathcal{X}^*$ denote the dual space of $\mathcal{X}$. Then, $f \in \mathcal{X}^*$ iff there is a $v \in \mathcal{X}$ such that*

$$f(u) = \langle v|u\rangle, \quad \forall u \in \mathcal{X}. \tag{3.27}$$

Here, the element v of $\mathcal{X}$ is uniquely determined by f. In addition,

$$\|f\| = \|v\|. \tag{3.28}$$

Proof.

- Step-1: Uniqueness of v. It follows from

$$\langle v|u\rangle = \langle v_1|u\rangle, \quad \forall u \in \mathcal{X}$$

that $\langle v - v_1|u\rangle = 0$ for $u = v - v)1$, and hence $v = v_1$.

- Step-2: Existence of v. Let $f \in \mathcal{X}^*$ with $f \neq 0$. The null space (kernel)

$$\mathrm{Ker}(f) \triangleq \{u \in \mathcal{X} : f(u) = 0\}$$

is closed linear subspace of $\mathcal{X}$.

  - In fact, if $f(u_n) = 0$ for all n and $u_n \to u$ as $n \to \infty$, then $f(u) = 0$, by the continuity of f. According to the orthogonal decomposition theorem, there exists an element

$$u_0 \in \mathrm{Ker}(f)^\perp \quad \text{with} \quad u_0 \neq 0.$$

    Otherwise we would have $\mathrm{Ker}(f)^\perp = \{0\}$, and hence $\mathrm{Ker}(f)^\perp = \mathcal{X}$. But this is impossible because of $f \neq 0$.

  - Since $u_0 \notin \mathrm{Ker}(f)$, we get $f(u_0) \neq 0$. Without any loss of generality we may assume that $f(u_0) = 1$. This implies

$$f(u - f(u)u_0) = 0 \quad \forall u \in \mathcal{X},$$

    i.e., $u - f(u_0)u_0 \in \mathrm{Ker}(f)$. Hence we obtain the orthogonal decomposition

$$u = w + f(u)u_0, \quad w \in \mathrm{Ker}(f), u_0 \in \mathrm{Ker}(f)^\perp. \tag{3.29}$$

    Inner multiplication by u_0 yields

$$\langle u_0|u\rangle = f(u)\langle u_0|u_0\rangle, \quad \forall u \in \mathcal{X}.$$

    This implies (3.27) with

$$v \triangleq \frac{u_0}{\langle u_0|u_0\rangle}.$$

  - If $f = 0$, then (3.27) holds with $v = 0$.

- Step-3: Conversely, if f is given through (3.27), then $f \in \mathcal{X}^*$. In fact, f is linear because

$$\begin{aligned} f(\alpha u + \beta w) &= \langle v|\alpha u + \beta w\rangle = \alpha\langle v|u\rangle + \beta\langle v|w\rangle \\ &= \alpha f(u) + \beta f(w), \quad \forall v, w \in \mathcal{X}, \alpha, \beta \in \mathbb{F} \end{aligned}$$

  Furthermore, the continuity of f follows from

$$|f(u)| = |\langle v|u\rangle| \leq |\langle v|u\rangle|, \quad \forall u \in \mathcal{X} \tag{3.30}$$

- Step-4: By (3.30), $\|f\| \leq \|v\|$. Furthermore, $f(v) = \langle v|v\rangle = \|v\|^2$. Hence

$$\|f\| = \sup_{\|u\|\leq 1} |f(u)| = \|v\|. \qquad \blacksquare$$

Equation (3.29) tells us the following fundamental geometrical fact:

Remark 111. *If f is a nonzero linear continuous functional on a Hilbert space, then the null space* $\mathrm{Ker}(f)$ *of f is a closed plane and its orthogonal complement* $\mathrm{Ker}(f)^\perp$ *has dimensional one, i.e.,*

$$\dim(\mathrm{Ker}(f)^\perp) = 1.$$

3.6 Duality Map

Definition 112. *Let $\mathcal{X}$ be a Hilbert space over $\mathcal{X}$. We define the duality map*

$$J : \mathcal{X} \to \mathcal{X}^*$$

of $\mathcal{X}$ by

$$J(v) \triangleq f,$$

where f is given by

$$f(u) = \langle v|u\rangle, \forall u \in \mathcal{X}.$$

In other words, we have

$$[J(v)](\cdot) = \langle v|\cdot\rangle.$$

Definition 113 (Pair Map)**.** *The pair map $\langle\cdot,\cdot\rangle$ defined by*

$$\begin{aligned}\langle\cdot,\cdot\rangle : \mathcal{X}^* \times \mathcal{X} &\to \mathbb{R} \\ (f,u) &\mapsto \langle f,u\rangle = f(u), \quad \forall f \in \mathcal{X}^*, u \in \mathcal{X}.\end{aligned}$$

is a bilinear form.

Using the notation of pair map $\langle f,u\rangle = f(u)$ for $f \in \mathcal{X}^*$ and $u \in \mathcal{X}$, this means

$$\langle J(v),u\rangle \triangleq \langle v|u\rangle, \forall u,v \in \mathcal{X}.$$

Proposition 114. *The duality map J is bijective, continuous, and norm preserving, i.e.,*

$$\|J(u)\| = \|u\|, \quad \forall u \in .$$

Moreover,

- *If $\mathcal{X}$ is a real Hilbert space, then J is linear, i.e.,*

$$J(\alpha v + \beta w) = \alpha J(v) + \beta J(w), \quad \forall \alpha,\beta \in \mathbb{C}, v,w \in \mathcal{X}.$$

- *If $\mathcal{X}$ is a complex Hilbert space, the J is conjugate-linear, i.e.,*

$$J(\alpha v + \beta w) = \overline{\alpha} J(v) + \overline{\beta} J(w), \quad \forall \alpha,\beta \in \mathbb{C}, v,w \in \mathcal{X}. \qquad \blacksquare$$

3.7 Duality for Quadratic Variational Problem

Along with the original minimum problem

$$\min_{u\in\mathcal{X}} F(u) = \frac{1}{2}a(u,u) - b(u), \tag{3.31}$$

let us consider the dual maximum problem

$$\max_{v\in v_0+\mathcal{Y}} F^*(v) \triangleq -a(v,v) \tag{3.32}$$

where

$$\mathcal{Y} \triangleq \{u \in \mathcal{Z} : a(u,w) = 0, \forall w \in \mathcal{X}\}$$

and $v_0 \in \mathcal{Z}$ is a fixed solution of the following equation:

$$a(v_0,w) = b(w), \quad \forall w \in \mathcal{X}.$$

We make the following assumptions:

(H1) Let $\mathcal{X}$ be a linear closed subspace of the real Hilbert space $\mathcal{Z}$.

(H2) Let $a : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a bounded, symmetric, positive bilinear form.

(H3) Let $a : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be strongly positive, i.e.,

$$c \|u\|_{\mathcal{X}}^2 \leq a(u,u) \quad \text{for all } u \in \mathcal{X}$$

where $c > 0$ is a constant.

(H4) The functional $b : \mathcal{X} \to \mathbb{R}$ is linear and continuous.

Theorem 115 (Duality)**.** *The original minimum problem (3.31) has a unique solution u_0. The element u_0 is also the unique solution of the dual maximum problem (3.32), and the extremal values of (3.31) and (3.32) are the same, i.e.,*

$$F(u_0) = F^*(u_0).$$

Moreover, we have

$$a(u-v, u-v) = F(u) - F^*(v), \quad \forall u \in \mathcal{X}, v \in v_0 + \mathcal{Y}. \tag{3.33}$$

Proof.

- By Theorem 105, problem (3.31) has a unique solution u_0 that satisfies the variational equation

  $$a(u, u_0) = b(u), \quad u \in \mathcal{X}.$$

  By construction of $\mathcal{Y}$,

  $$a(u, v - v_0) = 0, \quad \forall u \in \mathcal{X}, v \in v_0 + \mathcal{Y},$$

  and the choice of v_0 yields the key relation

  $$a(u,v) = a(u,v_0) = b(u), \quad \forall u \in \mathcal{X}, v \in v_0 + \mathcal{Y}.$$

  Hence

  $$\begin{aligned} 0 \leq a(u-v, u-v) &= a(u,u) - 2a(u,v) + a(v,v) \\ &= a(u,u) - 2b(u) + a(v,v) \\ &= F(u) - F^*(v), \quad \forall u \in \mathcal{X}, v \in v_0 + \mathcal{Y}. \end{aligned} \tag{3.34}$$

  This implies

  $$F^*(v) \leq F(u_0), \quad v \in v_0 + \mathcal{Y}.$$

- Furthermore, we shall show that $u_0 \in v_0 + \mathcal{Y}$ and

  $$F^*(u_0) = F(u_0). \tag{3.35}$$

  Thus, u_0 is a solution of the dual problem (3.32).

  - To prove (3.35) observe that

    $$a(u_0, w) = b(w) \quad \text{and} \quad a(v_0, w) = b(w), \quad \forall w \in \mathcal{X}.$$

    Hence $a(u_0 - v_0, w) = 0$ for all $w \in \mathcal{X}$, i.e., $u_0 - v_0 \in \mathcal{Y}$.
  - Furthermore, letting $u = v = u_0$ in (3.34), we get (3.35).

Corollary 116 (Error Estimates). *For all $u \in \mathcal{X}$ and $v \in v_0 + \mathcal{Y}$, we get*

$$F^*(v) \leq F(u_0) \leq F(u) \tag{3.36}$$

and

$$c\,\|u_0 - u\|_{\mathcal{Z}}^2 \leq a(u_0 - u, u_0 - u) \leq F(u) - F^*(v). \tag{3.37}$$

In numerical analysis, one computers u and v in Corollary 116 as solution of the *Ritz method* for (3.31) and (3.32), respectively. The Ritz method for the dual problem (3.32) is also called the Trefftz method.

Example. Let $\mathcal{X}$ be a linear closed subspace of the real Hilbert space $\mathcal{Z}$. For given $v_0 \in \mathcal{Z}$, we consider the minimum problem

$$\min_{u \in \mathcal{X}} F(u) \triangleq \|v_0 - u\|^2 - \|v_0\|^2 \tag{3.38}$$

together with the dual maximum problem

$$\max_{w \in \mathcal{X}^\perp} F^*(w) \triangleq -\|v_0 - w\|^2 \tag{3.39}$$

where $\|\cdot\|$ denotes the norm on $\mathcal{Z}$. Then, problem (3.38) has a unique solution u_0 and $u - u_0$ is the unique solution (3.38). Moreover,

$$F(u_0) = F^*(v_0 - u_0). \tag{3.40}$$

Relation (3.40) is identical to the Pythagorean theorem

$$\|v_0\|^2 = \|u_0\|^2 + \|v_0 - u_0\|^2, \quad u_0 \in \mathcal{X}, v_0 - u_0 \in \mathcal{X}^\perp.$$

Proof.

- Use Theorem 115 with

$$a(u, v) \triangleq \langle u|v\rangle_{\mathcal{Z}}, \quad \text{and} \quad b(u) \triangleq \langle v_0|u\rangle_{\mathcal{Z}}.$$

  Here, $\mathcal{Y} = \mathcal{X}^\perp$.

3.8 Linear Orthogonality Principle

Theorem 117 (Linear Orthogonality Principle). *The following three existence principles are mutually equivalent:*

① *The existence principle for quadratic minimum problems (Theorem 105).*

② *The perpendicular principle (Theorem 107).*

③ *The Riesz theorem (Theorem 110).*

These three principles represent variants of the linear orthogonality principle in Hilbert spaces.

Proof.

- In preceding sections, we have already proved that ① $\Longrightarrow$ ② $\Longrightarrow$ ③.
- ③$\Longrightarrow$①
  - We consider the minimum problem

$$\min_{u \in \mathcal{X}} \frac{1}{2} a(u, u) - b(u). \tag{3.41}$$

Consider the energetic inner product on $\mathcal{X}$ defined by

$$\langle u|v\rangle_E \triangleq a(u,v), \quad \forall u,v \in \mathcal{X}.$$

Then the energetic space $\mathcal{X}_E$ consists of the set $\mathcal{X}$ equipped with $\langle\cdot|\cdot\rangle_E$, and it is a Hilbert space.

– Moreover, there are positive constants α and β such that

$$\alpha \|u\|_E \leq \|u\|_X \leq \beta \|u\|_E, \quad \forall u \in \mathcal{X}.$$

By assumption, the linear functional $b : \mathcal{X} \to \mathbb{R}$ is continuous. Hence

$$|b(u)| \leq \|b\| \|u\|_{\mathcal{X}} \leq \beta \|b\| \|u\|_E, \quad \forall u \in \mathcal{X}.$$

That is, $b(\cdot)$ also represents a linear continuous functional on $\mathcal{X}_E$. By the Riesz theorem, there is a $v \in \mathcal{X}_E$ such that

$$b(u) = \langle v|u\rangle_E, \quad \forall u \in \mathcal{X}.$$

– Consequently, problem (3.41) can be written in the following form:

$$\min_{u\in\mathcal{X}_E} \frac{1}{2}\langle u|u\rangle_E - \langle v|u\rangle_E.$$

This is equivalent to the problem

$$\langle u-v|u-v\rangle_E \triangleq \langle u|u\rangle_E - 2\langle u|v\rangle_E + \langle v|v\rangle_E = \min, \quad u \in \mathcal{X}_E$$

which has the unique solution $u = v$.

3.9 Nonlinear Monotone Operators

We want to solve the nonlinear operator equation $Au = z, u \in \mathcal{X}$. To this end, we make the following assumptions:

(H1) The operator $A : \mathcal{X} \to \mathcal{X}$ is strongly monotone on the real Hilbert space, i.e., by definition, there is a constant $c > 0$ such that

$$\langle Au - Av|u - v\rangle \geq c \|u - v\|^2, \quad \forall u,v \in \mathcal{X}$$

(H2) The operator A is Lipschitz continuous, i.e., there is a constant $L > 0$ such that

$$\|Au - Av\| \leq \|L\| \|u - v\|, \quad \forall u,v \in \mathcal{X}.$$

Theorem 118. *For each given $z \in \mathcal{X}$, the nonlinear operator equation*

$$Au = z, \quad u \in \mathcal{X}, \tag{3.42}$$

where A satisfies the assumptions (H1) and (H2), then problem (3.42) has a unique solution u.

Proof. We will use the Banach fixed-point theorem. The idea of the proof is to replace the original equation (3.42) by the equivalent fixed-point theorem

$$u = Bu, \quad u \in \mathcal{X} \tag{3.43}$$

where

$$B \triangleq u - t(Au - z), \quad \text{for fixed real } t > 0.$$

- If $\mathcal{X} = \{0\}$, then the statement is trivial.
- Let $\mathcal{X} \neq \{0\}$. For all $u, v \in \mathcal{X}$,

$$\|Bu - Bv\|^2 = \|u - v\|^2 - 2t\langle Au - Av | u - v\rangle + t^2 \|Au - Av\|^2 \leq m \|u - v\|^2, \tag{3.44}$$

where

$$m \triangleq 1 - 2tc + t^2 L^2.$$

By (3.44), $m \geq 0$. If $t = 0$ or $t = \frac{2c}{L^2}$, then $m = 1$. This implies that

$$\alpha \triangleq \sqrt{m} < 1 \quad \forall t \in \left(0, \frac{2c}{L^2}\right).$$

Therefore,

$$\|Bu - Bv\| \leq \alpha \|u - v\|, \quad \forall u, v \in \mathcal{X},$$

i.e., the operator B is α-contractive for each $t \in (0, 2c/L^2)$.

- By the Banach fixed-point theorem, the problem (3.43) has a unique solution u.

3.10 Nonlinear Lax-Milgram Theorem

Our objective is to solve the equation

$$a(u, v) = b(v), \quad \text{for fixed } u \in \mathcal{X} \text{ and all } v \in \mathcal{X} \tag{3.45}$$

such that

(H1) Let $b : \mathcal{X} \to \mathbb{R}$ be a linear continuous functional on the real Hilbert space $\mathcal{X}$.

(H2) Let $a : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a function such that, for each $w \in \mathcal{X}$,

$$v \mapsto a(w, v)$$

represents a linear continuous functional on $\mathcal{X}$.

(H3) There are positive constants L and c such that, for all $u, v, w \in \mathcal{X}$,

$$c \|u - v\|^2 \leq a(u, u - v) - a(v, u - v)$$

and

$$|a(u, w) - a(v, w)| \leq L \|u - v\| \cdot \|w\|,$$

Theorem 119 (Nonlinear Lax-Milgram Theorem)**.** *Problem (3.45) has a unique solution.*

Theorem 119 can be regarded as a nonlinear orthogonality principle on Hilbert spaces.

PROOF.

- By (H2) and the Riesz theorem, for each $w \in \mathcal{X}$, there is an element called Aw such that

$$a(w, u) = \langle Aw | u\rangle, \quad \forall u \in \mathcal{X}.$$

This way we get an operator $A : \mathcal{X} \to \mathcal{X}$. It follows from (H3) that

$$c \|u - v\|^2 \leq \langle Au - Av | u - v\rangle, \quad u, v \in \mathcal{X},$$

i.e., A is *strongly monotone*.

- Furthermore,
$$|\langle Au - Av|w\rangle| \leq L \|u - v\| \|w\|, \quad \forall u, v, w \in \mathcal{X}.$$
Hence
$$\|Au - Av\| = \sup_{\|u\|\leq 1} |\langle Au - Av|w\rangle| \leq L \|u - v\|, \quad \forall u, v \in \mathcal{X}.$$

- Again, by the Riesz theorem, there is a $z \in \mathcal{X}$ such that
$$b(u) = \langle z|u\rangle, \quad \forall u \in \mathcal{X}.$$

- Consequently, the original problem (3.45) is equivalent to the operator equation
$$Au = z, \quad u \in \mathcal{X}. \tag{3.46}$$
It follows now form Theorem 118 that equation (3.46) has a unique solution u.

Remark. In the special case where $a : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is bilinear, bounded, and strongly positive, i.e.,
$$a(w, w) \geq c \|w\|^2, \quad \forall w \in \mathcal{X} \text{and fixed } c > 0,$$
assumptions (H2) and (H3) are satisfied. Then, Theorem 119 is called the linear Lax-Milgram theorem. If, in addition, $a(\cdot, \cdot)$ is symmetric, then problem (3.45) is identical to the variational equation from Theorem 105.

Chapter 4

Generalized Functions

4.1 History

4.1.1 Some Early History

In the mathematics of the nineteenth century, aspects of generalized function theory appeared, for example in the definition of the Green's function, in the Laplace transform, and in Riemann's theory of trigonometric series, which were not necessarily the Fourier series of an integrable function. These were disconnected aspects of mathematical analysis at the time.

The intensive use of the Laplace transform in engineering led to the heuristic use of symbolic methods, called operational calculus. Since justifications were given that used divergent series, these methods had a bad reputation from the point of view of pure mathematics. They are typical of later application of generalized function methods. An influential book on operational calculus was Oliver Heaviside's Electromagnetic Theory of 1899.

When the Lebesgue integral was introduced, there was for the first time a notion of generalized function central to mathematics. An integrable function, in Lebesgue's theory, is equivalent to any other which is the same almost everywhere. That means its value at a given point is (in a sense) not its most important feature. In functional analysis a clear formulation is given of the essential feature of an integrable function, namely the way it defines a linear functional on other functions. This allows a definition of weak derivative.

During the late 1920s and 1930s further steps were taken, basic to future work. The Dirac delta function was boldly defined by Paul Dirac (an aspect of his scientific formalism); this was to treat measures, thought of as densities (such as charge density) like honest functions. Sergei Sobolev, working in partial differential equation theory, defined the first adequate theory of generalized functions, from the mathematical point of view, in order to work with weak solutions of PDEs. Others proposing related theories at the time were Salomon Bochner and Kurt Friedrichs. Sobolev's work was further developed in an extended form by L. Schwartz.

4.1.2 Schwartz Distributions

The realization of such a concept that was to become accepted as definitive, for many purposes, was the theory of distributions, developed by Laurent Schwartz. It can be called a principled theory, based on duality theory for topological vector spaces. Its main rival, in applied mathematics, is to use sequences of smooth approximations (the 'James Lighthill' explanation), which is more ad hoc. This now enters the theory as mollifier theory.

This theory was very successful and is still widely used, but suffers from the main drawback that it allows only linear operations. In other words, distributions cannot be multiplied (except for

very special cases): unlike most classical function spaces, they are not an algebra. For example it is not meaningful to square the Dirac delta function. Work of Schwartz from around 1954 showed that this was an intrinsic difficulty.

Some solutions to the multiplication problem have been proposed. One is based on a very simple and intuitive definition a generalized function given by Yu. V. Egorov that allows arbitrary operations on, and between, generalized functions.

Another solution of the multiplication problem is dictated by the path integral formulation of quantum mechanics. Since this is required to be equivalent to the Schrödinger theory of quantum mechanics which is invariant under coordinate transformations, this property must be shared by path integrals. This fixes all products of generalized functions as shown by H. Kleinert and A. Chervyakov. The result is equivalent to what can be derived from dimensional regularization.

4.2 Test Function Space $C_0^\infty(\Omega)$

4.2.1 Terminology

In any discussion of functions of n variables, the term *multi-index* denotes an ordered n-tuple

$$\alpha = (\alpha_1, \cdots, \alpha_n) \tag{4.1}$$

of nonnegative integers $\alpha_i \in \{0, 1, 2, \cdots\}$. With each multi-index α is associated the differential operator

$$\begin{aligned} \mathrm{D}^\alpha &\triangleq \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_n}\right)^{\alpha_n} \\ &= \partial_{x_1}^{\alpha_1} \cdots \partial_{x_n}^{\alpha_n} \\ &= \partial_1^{\alpha_1} \cdots \partial_n^{\alpha_n} \end{aligned} \tag{4.2}$$

whose *order* is

$$|\alpha| \triangleq \sum_{j=1}^{n} \alpha_j = \alpha_1 + \cdots + \alpha_n. \tag{4.3}$$

If $|\alpha| = 0$, then $\mathrm{D}^\alpha f = f$.

Definition 120 ($C^\infty(\Omega)$ Space)**.** *A function $f : \Omega \subset \mathbb{R}^n \to \mathbb{F}$ is said to belong to $C^\infty(\Omega)$ if $\mathrm{D}^\alpha f \in C(\Omega)$ for every multi-index α, where Ω is an nonempty open set in $\mathbb{R}^n$.*

Definition 121 (Support Set)**.** *The support of a function f (on any topological space) is the closure of $\{x \in \mathrm{Dom}(f) : f(x) \neq 0\}$, that is,*

$$\mathrm{supp}(f) = \overline{\{x \in \mathrm{Dom}(f) : f(x) \neq 0\}}. \tag{4.4}$$

If K is a compact set in $\mathbb{R}^n$, then $\mathcal{D}_K$ denotes the space of all $f \in C^\infty(\mathbb{R}^n)$ whose support lies in K. If $K \subset \Omega$, then $\mathcal{D}_K$ may be identified with a subspace of $C^\infty(\Omega)$. We can define a topology on $C^\infty(\Omega)$ into a Fréchet space with the Heine-Borel property, such that $\mathcal{D}_K$ is a closed subspace of $C^\infty(\Omega)$ wherever $K \subset \Omega$. In fact, the proof is given as follows:

- Choose compact set $K_i (i = 1, 2, \cdots)$ such that K_i lies in the interior of K_{i+1} and $\Omega = \cup K_i$.
- Define semi-norms p_m on $C^\infty(\Omega)$, $m = 1, 2, \cdots$, by setting

$$p_m(f) \triangleq \sup_{\substack{x \in K_m \\ |\alpha| \leq m}} \left|\mathrm{D}^\alpha f(x)\right|. \tag{4.5}$$

  They define a metrizable locally convex on $C^\infty(\Omega)$. For each $x \in \Omega$, the functional $f \to f(x)$ is continuous in this topology.

- Since $\mathcal{D}_K$ is the intersection of the null spaces of these functionals, as x ranges over the complement of K, it follows that $\mathcal{D}_K$ is closed in $C^\infty(\Omega)$.
- A local base is given by the sets

$$V_m = \left\{ f \in C^\infty(\Omega) : p_m(f) < \frac{1}{m} \right\}, \quad m = 1, 2, 3, \cdots . \tag{4.6}$$

- If $\{f_i\}$ is a Cauchy sequence in $C^\infty(\Omega)$ and if N is fixed, then $f_i - f_j \in V_m$ if i and j are sufficiently large. Thus $|\mathrm{D}^\alpha f_i - \mathrm{D}^\alpha f_j| < \frac{1}{m}$ on K_m, if $|\alpha| \leq m$. It follows that each $\mathrm{D}^\alpha f_i$ converges (uniformly on compact subsets of Ω) to a function g_α. In particular, $f_i(x) \to g_0(x)$. It is now evident that $g_0 \in C^\infty(\Omega)$, that $g_\alpha = \mathrm{D}^\alpha g_0$, and that $f_i \to g$ in the topology of $C^\infty(\Omega)$.
- Thus $C^\infty(\Omega)$ is a Fréchet space. The same is true of each of its closed subspaces $\mathcal{D}_K$.

Definition 122 (Test Funciton Space $\mathcal{D}(\Omega)$)**.** *The union of the spaces $\mathcal{D}_K$, as K ranges over all compact subsets of Ω, is the test function space $\mathcal{D}(\Omega)$, that is,*

$$\mathcal{D}(\Omega) = \bigcup_K \mathcal{D}_K, \quad K \sqsubseteq \Omega, \tag{4.7}$$

where $K \sqsubseteq \Omega$ means that $K \subset \Omega$ and K is compact.

It is clear that $\mathcal{D}(\Omega)$ is a linear space, with respect to the usual definitions of addition and scalar multiplication of functions. Explicitly, $\phi \in \mathcal{D}(\Omega)$ iff $\phi \in C^\infty(\Omega)$ and $\mathrm{supp}(\phi)$ is a compact subset of Ω, i.e.,

$$\mathcal{D}(\Omega) = \{\phi \in C^\infty(\Omega) : \mathrm{supp}(\phi) \sqsubseteq \Omega\} . \tag{4.8}$$

Definition 123 (Norm in $\mathcal{D}(\Omega)$)**.** *The norm in $\mathcal{D}(\Omega)$ is defined as*

$$\|\phi\|_m = \sup_{\substack{x \in \Omega \\ |\alpha| \leq m}} |\mathrm{D}^\alpha \phi(x)|$$

for $\phi \in \mathcal{D}(\Omega)$ and $m = 0, 1, 2, \cdots$.

Definition 124. *$C_0^\infty(\Omega)$ is the set of all functions $u \in C_0^\infty(\Omega)$ that vanish outside a compact subset K of Ω that depends on u, i.e., $u(x) = 0$ for all $x \in \Omega - K$:*

$$C_0^\infty(\Omega) = \left\{ u : u(x)|_{\Omega - K} = 0, K \sqsubseteq \Omega \right\} .$$

Definition 125 (Convergence in $\mathcal{D}(\Omega)$)**.** *For the sequence $\{f_j\} \subset C^\infty(\Omega)$, if there exists a compact subset $K \subset \Omega$ such that*

$$\mathrm{supp}(f_j) \subset K, \quad j = 0, 1, 2, \cdots$$

and for any nonnegative integerm

$$\|f_j - f\|_m = \sup_{\substack{x \in K \\ |\alpha| \leq m}} |\mathrm{D}^\alpha f_j(x) - \mathrm{D}^\alpha f(x)| \to 0$$

as $j \to \infty$, then we say f_j converges to f on $C_0^\infty(\Omega)$.

4.2.2 Smoothing Function

Let

$$j(x) = \begin{cases} c_n \exp\left(-\frac{1}{1 - \|x\|^2} \right), & \|x\| < 1, \\ 0, & \|x\| \geq 1, \end{cases} \tag{4.9}$$

where

$$c_n \triangleq \left(\int_{\|x\|\le 1} \exp\left(-\frac{1}{1-\|x\|^2} \right) \mathrm{d}\, x \right)^{-1},$$

we call $j(x)$ a *smoothing function* or *mollifier*. Moreover, $j(x) \in C_0^\infty(\mathbb{R}^n)$ and satisfies the following properties:

- $j(x) \ge 0$, and $j(x) = 0$ for $\|x\| \ge 1$.
- $\int\limits_{\mathbb{R}^n} j(x)\, \mathrm{d}\, x = 1$.

With the help of $j(x)$, we can construct many new functions in $C_0^\infty(\mathbb{R}^n)$. For illustration, let

$$j_\varepsilon(x) = \frac{1}{\varepsilon^n} j\left(\frac{x}{\varepsilon}\right), \tag{4.10}$$

then $j_\varepsilon(x) \in C_0^\infty(\mathbb{R}^n)$ and

(a) $j_\varepsilon(x) \ge 0$, and $j_\varepsilon(x) = 0$ for $\|x\| \ge \Delta$;

(b) $\int\limits_{\mathbb{R}^n} j_\varepsilon(x)\, \mathrm{d}\, x = 1$.

Proposition 126. *Let K is a compact subset of Ω and u is a integrable function, for suficiently small ε, the function*

$$u_\varepsilon(x) \triangleq \int_\Omega u(y) j_\varepsilon(x-y)\, \mathrm{d}\, y = j_\varepsilon * u \tag{4.11}$$

belongs to $C_0^\infty(\Omega)$.

Theorem 127. *If $u \in C_0^k(\Omega)$, then*

$$\|u_\varepsilon - u\|_{C_0^k(\overline{\Omega})} \to 0 \quad as \quad \varepsilon \to 0.$$

Corollary 128. *For the test function space $C_0^\infty(\Omega)$, we have:*

① *$C_0^\infty(\Omega)$ is dense in $C_0^k(\Omega)$.*

② *If μ is an additive measure on Ω, then*

$$\int_\Omega \phi(x)\, \mathrm{d}\, \mu = 0, \quad \forall \phi \in C_0^\infty(\Omega)$$

implies

$$\int_\Omega \phi(x)\, \mathrm{d}\, \mu = 0, \quad \forall \phi \in C_0^0(\Omega).$$

4.3 Generalized Functions

4.3.1 $\mathcal{D}(\Omega)$ and $\mathcal{D}^*(\Omega)$

Definition 129. *The functional $f : \mathcal{D}(\Omega) \to \mathbb{R}$ is called a generalized function or distribution if the pair map $\langle f, \phi \rangle$ satisfies the following conditions:*

- *Linearity:*

$$\langle f, c_1\phi_1 + c_2\phi_2 \rangle = c_1 \langle f, \phi_1 \rangle + c_2 \langle f, \phi_2 \rangle, \quad \forall \phi_1, \phi_2 \in \mathcal{D}(\Omega), \forall c_1, c_2 \in \mathbb{R}$$

- *Continuous: For any sequence $\{\phi_n\} \subset \mathcal{D}(\Omega)$, if $\phi_n \to \phi$ on $\mathcal{D}(\Omega)$, then*

$$\langle f, \phi_n \rangle \to \langle f, \phi \rangle, \quad as \quad n \to \infty.$$

Usually, we use $\mathcal{D}^*(\Omega)$ to denote the set of all the generalized function defined on Ω. $\mathcal{D}^*(\Omega)$ is a linear space with respect to the linear operation as follows:

$$\langle \lambda_1 f_1 + \lambda_2 f_2, \phi \rangle \triangleq \lambda_1 \langle f_1, \phi \rangle + \lambda_2 \langle f_2, \phi \rangle, \quad \forall f \in \mathcal{D}^*(\Omega), \forall \lambda_1, \lambda_2 \in \mathbb{F}, \forall \phi \in C_0^\infty(\Omega). \tag{4.12}$$

Note that the notaitons appeared in French references may be different. The notation for the dual space in French references is like $\mathcal{X}'$ rather than $\mathcal{X}^*$, and similarly the notations for the elements/vectors of dual space are marked by the prime $(\cdot)'$ instead of the star notation $(\cdot)^*$. Hence you may see $x' \in \mathcal{X}'$ or $x^* \in \mathcal{X}^*$, $f \in \mathcal{D}'(\Omega)$ or $f \in \mathcal{D}^*(\Omega)$ in different references.

Definition 130 (Locally Integrable)**.** *Let $f : \Omega \to \mathbb{F}$, if for any compact set $K \subset \Omega$ we have*

$$\int_K |f(x)| \, \mathrm{d}\, \mu < \infty,$$

we say that f is locally integrable on Ω and denote it as $f \in L^1_{\mathrm{loc}}(\Omega)$.

Definition 131 (*Weak Convergence)**.** *The sequence $(f_n) \subset \mathcal{D}^*(\Omega)$ is called *-weakly convergent to $f \in \mathcal{D}^*(\Omega)$ if*

$$\langle f_n, \phi \rangle \to \langle f, \phi \rangle, \quad \forall \phi \in \mathcal{D}(\Omega),$$

which is denoted as $f_n \overset{*w}{\to} f$.

4.3.2 Dirac-δ Function

Define the linear functional as follows

$$\langle \delta, \phi \rangle = \phi(0), \quad \forall \phi \in \mathcal{D}(\Omega). \tag{4.13}$$

Obviously, δ is linear. When $\phi_j \to \phi$ on $\mathcal{D}(\Omega)$, we have

$$|\langle \delta, \phi_j \rangle - \langle \delta, \phi \rangle| = |\phi_j(0) - \phi(0)| \to 0, \quad \text{as} \quad j \to \infty.$$

Hence

$$\langle \delta, \phi_j \rangle \to \langle \delta, \phi \rangle, \quad j \to \infty.$$

Therefore δ is continuous on $\mathcal{D}(\Omega)$. Thus δ is a generalized function by definition.

There are some important results about the Dirac-δ function

① $\delta(x) \notin L^1_{\mathrm{loc}}(\Omega)$

② $\lim\limits_{\sigma \to 0} \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} = \delta(x - \mu)$.

③ $\langle \delta(x-a), \phi(x) \rangle = \phi(a), \forall \phi \in C_0^\infty(\Omega)$

Proof.

① We prove it by contradiction. If there exists $f \in L^1_{\mathrm{loc}}(\Omega)$ such that

$$\langle \delta, \phi \rangle = \int_\Omega f(x)\phi(x) \, \mathrm{d}\, x, \quad \forall \phi \in \mathcal{D}(\Omega).$$

For all $k \in \mathbb{N}$, let $\phi_k(x) = j(kx)$ in which

$$j(x) = \begin{cases} c_n \mathrm{e}^{\frac{1}{|x|^2 - 1}}, & |x| < 1; \\ 0, & |x| \geq 1; \end{cases}$$

where $c_n > 0$ is constant. On the other hand, we have

$$\left| \int_{\mathcal{B}_{1/k}} f(x)\phi_k(x) \, \mathrm{d}\, x \right| \leq c_n \int_{\mathcal{B}_{1/k}} |f(x)| \, \mathrm{d}\, x \to 0, \quad k \to \infty.$$

Therefore,

$$\langle \delta, \phi_k \rangle = \phi_k(0) = c_n \mathrm{e}^{-1} \to c_n \mathrm{e}^{-1}, \quad k \to \infty.$$

Thus $0 = c_n \mathrm{e}^{-1} \neq 0$. This is a contradiction.

② It is omitted here and left as exercise.

③ Let $\Omega = \mathbb{R}$, with the help of changing variable we have

$$\begin{aligned}\langle \delta(x-a), \phi(x) \rangle &= \int_{-\infty}^{\infty} \delta(x-a)\phi(x)\,\mathrm{d}\,x \\ &= \int_{-\infty}^{\infty} \delta(t)\phi(t+a)\,\mathrm{d}\,t \\ &= \phi(0+a) = \phi(a), \qquad \forall \phi \in C_0^\infty(\mathbb{R})\end{aligned}$$

4.4 Generalized Derivatives

4.4.1 Definition and Dual Operator

Theorem 132. *Let $L = \mathrm{D}^\alpha$ be the derivative operative of the generalized functions, then its dual operator is $L^* = (\mathrm{D}^\alpha)^* = (-1)^{|\alpha|}\,\mathrm{D}^\alpha$. In other words, we have*

$$\langle \mathrm{D}^\alpha T, \phi \rangle = \left\langle T, (-1)^{|\alpha|}\,\mathrm{D}^\alpha\,\phi \right\rangle, \qquad \forall T \in \mathcal{D}(\Omega), \forall \phi \in C_0^\infty(\Omega). \tag{4.14}$$

Proof

- We can get the result with the help of the integration-by-parts and the vanishing property of $\phi \in C_0^\infty(\Omega)$.
- By the definition of $\langle \mathrm{D}^\alpha T, \phi \rangle$, we have

$$\begin{aligned}\langle \mathrm{D}^\alpha T, \phi \rangle &= \int_\Omega (\mathrm{D}^\alpha T)\phi(x)\,\mathrm{d}\,x \\ &= \int_\Omega \mathrm{D}(\mathrm{D}^{\alpha-1} T)\phi(x)\,\mathrm{d}\,x \\ &= (\mathrm{D}^{\alpha-1} T)\phi(x)\big|_{\partial\,\Omega} - \int_\Omega (\mathrm{D}^{\alpha-1} T)(\mathrm{D}\,\phi(x))\,\mathrm{d}\,x \\ &= (-1)\cdot \int_\Omega (\mathrm{D}^{\alpha-1} T)(\mathrm{D}\,\phi(x))\,\mathrm{d}\,x \\ &= \cdots \\ &= (-1)^{|\alpha|} \int_\Omega T(\mathrm{D}^\alpha\,\phi(x))\,\mathrm{d}\,x \\ &= \int_\Omega T((-1)^{|\alpha|}\,\mathrm{D}^\alpha\,\phi(x))\,\mathrm{d}\,x \\ &= \left\langle T, (-1)^{|\alpha|}\,\mathrm{D}^\alpha\,\phi \right\rangle \\ &= \langle T, (\mathrm{D}^\alpha)^* \phi \rangle\end{aligned}$$

Hence $(\mathrm{D}^\alpha)^* = (-1)^{|\alpha|}\,\mathrm{D}^\alpha$.

4.4.2 Derivative of $|x|$

Consider the function $u : (-1, 1) \to \mathbb{R}$ with

$$u(x) \triangleq |x|, \quad \forall x \in (-1, 1).$$

Set

$$w(x) \triangleq \begin{cases} -1, & \text{for } -1 < x < 0 \\ c, & \text{for } x = 0 \\ +1, & \text{for } 0 < x < 1 \end{cases}$$

where c is a fixed, but otherwise arbitrary, real number. Then, the function w represents the generalized derivative of the function u on the interval $(-1, 1)$. We write

$$u' = w, \quad x \in (-1, 1).$$

Note that w is the classic derivative of u on both subinterval $(-1, 0)$ and $(0, 1)$, but the classic derivative of u does not exist at the point $x = 0$.

4.4.3 Heaviside Step Function

The Heaviside step function, or the unit step function, usually denoted by H (but more frequently U in electrical and electronics engineering), is a discontinuous function whose value is zero for negative argument and one for positive argument. It seldom matters what value is used for $H(0)$, since H is mostly used as a distribution. Some common choices can be seen below.

The function is used in the mathematics of control theory and signal processing to represent a signal that switches on at a specified time and stays switched on indefinitely. It is also used in structural mechanics together with the Dirac delta function to describe different types of structural loads. It was named after the English polymath Oliver Heaviside.

It is the cumulative distribution function of a random variable which is almost surely 0. (See constant random variable.)

The Heaviside function is the integral of the Dirac delta function: $H' = \delta$. This is sometimes written as

$$H(x) = \int_{-\infty}^{x} \delta(t)\,\mathrm{d}\,t, \qquad \delta(x) = H'(x) \tag{4.15}$$

although this expansion may not hold (or even make sense) for $x = 0$, depending on which formalism one uses to give meaning to integrals involving δ.

Proof

- For any $\phi \in C_0^\infty(\mathbb{R})$, we have

$$\begin{aligned} \langle \mathrm{D}\,H, \phi \rangle &= -\langle H, \mathrm{D}\,\phi \rangle \\ &= -\int_0^\infty \phi'(x)\,\mathrm{d}\,x = \phi(0) \\ &= \langle \delta, \phi \rangle \end{aligned}$$

Hence $\mathrm{D}\,H = \delta$.

4.4.4 Finite Jumps

Let $\Omega = \left\{ x : x \in \mathbb{R} - \bigcup_{j=1}^{k} x_j \right\}$, $f \in C^1(\Omega)$. Let $s_j \triangleq f(x_j + 0) - f(x_j - 0)$ be the jump amplitude at the point $x = x_j$, then

$$\begin{aligned}
\langle \mathrm{D}\, f, \phi \rangle &= -\langle f, \mathrm{D}\, \phi \rangle = -\int_{-\infty}^{\infty} f(x) \frac{\mathrm{d}\, \phi(x)}{\mathrm{d}\, x}\, \mathrm{d}\, x \\
&= \sum_{j=1}^{k} s_j \phi(x_j) + \int_{-\infty}^{\infty} \frac{\mathrm{d}\, f}{\mathrm{d}\, x} \phi(x)\, \mathrm{d}\, x \\
&= \left\langle \sum_{j=1}^{k} \delta(x - x_j), \phi(x) \right\rangle + \langle f', \phi \rangle
\end{aligned}$$

Therefore,

$$\mathrm{D}\, f(x) = f'(x) + \sum_{j=1}^{k} \delta(x - x_j). \tag{4.16}$$

Chapter 5

Soblev Spaces

The elements of the theory of Soblev spaces, a tool that, together with methods of functional analysis, provided for numerous successful attacks on the questions of existence and smoothness of solutions to many of the basic partial differential equations (PDFs).

5.1 Soblev Space $\mathrm{W}^{m,p}(\Omega)$

5.1.1 General Case

Definition 133. *Let $m \in \mathbb{N}$ and $p \in [1, +\infty]$. Define the set $\mathrm{W}^{m,p}(\Omega)$ as*

$$\mathrm{W}^{m,p}(\Omega) \triangleq \left\{ f \in \mathcal{D}'(\Omega) \cap L^p(\Omega) : \mathrm{D}^\alpha f \in L^p(\Omega), \forall \alpha, 0 \leq |\alpha| \leq m \right\}, \tag{5.1}$$

in which

$$L^p(\Omega) = \left\{ f : \int_\Omega |f|^p \,\mathrm{d}\, x < \infty \right\}, \forall p \in [1, \infty)$$

and

$$L^\infty(\Omega) = \left\{ f : \sup_{x \in \Omega} |f(x)| < \infty \right\}, \quad p = +\infty.$$

$\mathrm{W}^{m,p}(\Omega)$ is called the (m,p)-order Soblev space, which is a linear subspace of L^p with the norm specified by

$$\|f\|_{m,p} = \left(\sum_{|\alpha|=0}^{m} \int_\Omega |\mathrm{D}^\alpha f(x)|^p \,\mathrm{d}\, x \right)^{\frac{1}{p}} = \left[\sum_{|\alpha|=0}^{m} \|\mathrm{D}^\alpha f\|_{L^p(\Omega)}^p \right]^{\frac{1}{p}} \tag{5.2}$$

Soblev space provides powerful tools to measure the "smoothness" of a given funciton. If $f \in \mathrm{W}^{m,p}(\Omega)$, then its first generalized derivative $\mathrm{D}_j f$ belongs to $\mathrm{W}^{m-1,p}(\Omega)$, i.e.,

$$\mathrm{D}_j f \in \mathrm{W}^{m-1,p}(\Omega).$$

Actually, $\mathrm{W}^{m,p}(\Omega)$ includes all the functions in $L^p(\Omega)$ such that their $|\alpha|$ $(0 \leq |\alpha| \leq m)$ generalized derivatives still lie in $L^p(\Omega)$.

For illustrations, we can find some concrete cases:

- For $m = 0$, we have
$$\|f\|_{0,p} = \left(\int_\Omega |f(x)|^p \,\mathrm{d}\, x \right)^{\frac{1}{p}} = \|f\|_{L^p(\Omega)}$$
- For $m = 1$, we have
$$\|f\|_{1,p} = \left(\int_\Omega \left(|f(x)|^p + \sum_j |\mathrm{D}_j f|^p \right) \mathrm{d}\, x \right)^{\frac{1}{p}} = \left[\|f\|_{L^p(\Omega)}^p + \|\nabla f\|_{L^p(\Omega)}^p \right]^{\frac{1}{p}}$$

Theorem 134. *The* $\mathrm{W}^{m,p}(\Omega)$ *with the norm* $\|\cdot\|_{m,p}$ *is a Banach space.*

Proof.

- It is easy to verify that $\mathrm{W}^{m,p}(\Omega)$ is a linear space with the usual addition and scalar product, and $\|\cdot\|_{m,p}$ or $\|\cdot\|_{m,\infty}$ is a norm.
- Completeness.
  - Take a Cauchy sequence $(f_n) \subset \mathrm{W}^{m,p}(\Omega)$, then it is a Cauchy sequence in $L^2(\Omega)$ by definition. For any α such that $0 \leq |\alpha| \leq m$, the sequence $(\mathrm{D}^\alpha f_n)$ is also the Cauchy sequence of $L^p(\Omega)$. Since $L^p(\Omega)$ is complete, then there exist $f, g_k \in L^p(\Omega)$ such that
  $$\lim_{n\to\infty} \|f - f_n\|_{L^p(\Omega)} = 0, \quad \lim_{n\to\infty} \|g_k - \mathrm{D}^\alpha f_n\|_{L^p(\Omega)} = 0. \tag{5.3}$$
  - For any $n \in \mathbb{N}$ and α $(0 \leq |\alpha| \leq m)$, with the help of duality of D^α, we have
  $$\langle \mathrm{D}^\alpha f_n, \phi\rangle = (-1)^{|\alpha|} \langle f_n, \mathrm{D}^\alpha\rangle, \quad \forall \phi \in \mathcal{D}(\Omega).$$
  - Let $n \to \infty$, then
  $$\langle g_k, \phi\rangle = (-1)^{|\alpha|} \langle f, \mathrm{D}^\alpha \phi\rangle, \Delta quad \forall \phi \in \mathcal{D}(\Omega).$$
  Consequently,
  $$\mathrm{D}^\alpha f = g_k \in L^p(\Omega)$$
  and
  $$f \in \mathrm{W}^{m,p}(\Omega).$$
  Hence the completeness is proved. ■

5.1.2 Special Cases

Generally, we use the following deontions:

$$\mathrm{H}^0(\Omega) \triangleq L^2(\Omega) \tag{5.4}$$
$$\mathrm{W}_0^{m,p}(\Omega) \triangleq \text{the closure of } C_0^\infty(\Omega) \text{ in } \mathrm{W}^{m,p}(\Omega) \tag{5.5}$$
$$\mathrm{H}^m(\Omega) \triangleq \mathrm{W}^{m,2}(\Omega) = \left\{f : f \in L^2(\Omega), \mathrm{D}^\alpha f \in L^2(\Omega), |\alpha| \leq m\right\} \tag{5.6}$$
$$\mathrm{H}_0^m(\Omega) \triangleq \mathrm{W}_0^{m,2}(\Omega). \tag{5.7}$$

The space $\mathrm{H}^m(\Omega)$ is the most important Soblev space of interest since $p = 2$ relates the energy functional in many physics and engineering problems. The inner product in $\mathrm{H}^m(\Omega)$ is defined by

$$\langle f|g\rangle_{\mathrm{H}^m(\Omega)} \triangleq \int_\Omega \sum_{|\alpha|=0}^{m} \overline{\mathrm{D}^\alpha f(x)}\, \mathrm{D}^\alpha g(x)\, \mathrm{d}\, x. \tag{5.8}$$

Obviously, the norm induced by the inner product is

$$\|f\|_{\mathrm{H}^m(\Omega)} \triangleq \int_\Omega \sum_{|\alpha|=0}^{m} |\mathrm{D}^\alpha f(x)|^2\, \mathrm{d}\, x. \tag{5.9}$$

It is easy to verify that the following result:

Corollary 135. $\mathrm{H}^m(\Omega)$ *is a Hilbert space.*

5.2 The Soblev Space $\mathrm{H}^1(\Omega)$

Definition 136. *Let Ω be a nonempty open set in $\mathbb{R}^{N\times 1}, N \geq 1$. The Soblev space $\mathrm{W}^{1,2}(\Omega)$ consists precisely of all the funcitons*

$$u \in L^2(\Omega)$$

that have generalized derivatives

$$\mathrm{D}_j\, u \in L^2(\Omega), \quad \forall j = 1, \cdots, N.$$

Furthermore, for all $u, v \in \mathrm{H}^1(\Omega)$, we set

$$\langle u|v\rangle_{\mathrm{H}^1} \triangleq \int_\Omega \left(uv + \sum_{j=1}^{N} \mathrm{D}_j\, u\, \mathrm{D}_j\, v \right) \mathrm{d}\, x \tag{5.10}$$

and

$$\|u\|_{\mathrm{H}^1} = \sqrt{\langle u|u\rangle_{\mathrm{H}^1}} = \sqrt{\int_\Omega \left(u^2 + \sum_{j=1}^{N} (\mathrm{D}_j\, u)^2 \right) \mathrm{d}\, x} \tag{5.11}$$

Proposition 137. *The space $\mathrm{H}^1(\Omega)$ together with the inner product $\langle\cdot|\cdot\rangle_{\mathrm{H}^1}$ becomes a real Hilbert space, provided we identify two functions whose values differ only on a set of N-dimensional measure zero.*

Proof.

- Let $u \in \mathrm{H}^1(\Omega)$. From $\langle u|u\rangle_{\mathrm{H}^1} = 0$ we get $\int_\Omega u^2\, \mathrm{d}\, x = 0$, and hence $u(x) = 0$ for almost all $x \in \Omega$, i.e., u is the zero element. Hence $\langle\cdot|\cdot\rangle_{\mathrm{H}^1}$ is an inner product on $\mathrm{H}^1(\Omega)$, this H^1 is an inner product space.

- In order to prove that $\mathrm{H}^1(\Omega)$ is a Hilbert space, let (u_n) be a Cauchy sequence in $\mathrm{W}^{1,2}(\Omega)$, i.e.,

  $$\|u_n - u_m\|_{\mathrm{H}^1} < \varepsilon, \quad \forall n, m \geq n_0(\varepsilon).$$

  Hence (u_n) and $(\mathrm{D}_j\, u_n)$ are Cauchy sequences in $L^2(\Omega)$. Since $L^2(\Omega)$ is a Hilbert space, there are functions $u, w_j \in L^2(\Omega)$ such that, as $n \to \infty$,

  $$\begin{aligned} u_n &\to u \quad \text{in} \quad L^2(\Omega) \\ \mathrm{D}_j\, u &\to w_j \quad \text{in} \quad L^2(\Omega), \quad \forall j. \end{aligned} \tag{5.12}$$

  Letting $n \to \infty$, from

  $$\langle \mathrm{D}_j\, u_n, v\rangle = -\langle u_n, \mathrm{D}_j\, v\rangle$$

  we obtain

  $$\int_\Omega u\, \mathrm{D}_j\, v\, \mathrm{d}\, x = -\int_\Omega w_j v\, \mathrm{d}\, x, \quad \forall v \in C_0^\infty(\Omega), \tag{5.13}$$

  using the continuity of the inner product on the Hilbert space $L^2(\Omega)$. This equation shows that the function u has the generalized derivatives

  $$w_j = \mathrm{D}_j\, u \quad \text{on} \quad \Omega, \quad \forall j.$$

  Since $\mathrm{D}_j\, u \in L^2(\Omega)$ for all j, we get $u \in \mathrm{H}^1(\Omega)$.

- Finally, it follows from (5.12) that

  $$\|u_n - u\|_{\mathrm{H}^1} \to 0, \quad \text{as} \quad n \to \infty,$$

  i.e., $u_n \to u$ in $\mathrm{H}^1(\Omega)$ as $n \to \infty$. Hence $\mathrm{H}^1(\Omega)$ is a Hilbert space.

5.3 The Soblev Space $\mathrm{H}_0^m(\Omega)$

5.3.1 Fundamental Concepts and Results

Definition 138. *Let* $\mathrm{H}_0^1(\Omega)$ *denote the closure of the set* $C_0^\infty(\Omega)$ *in the Hilbert space* $\mathrm{H}^1(\Omega)$.

We will later discuss that it makes sense to say that all of the functions $u \in \mathrm{H}_0^1(\Omega)$ satisfy the boundary condition

$$u = 0 \quad \text{on} \quad \partial\,\Omega$$

in some generalized sense.

Proposition 139. *The space* $\mathrm{H}_0^1(\Omega)$ *is a real Hilbert space.*

Proof.

- Note that $C_0^\infty(\Omega)$ is a linear subspace of the Hilbert space $\mathrm{H}^1(\Omega)$.
- By Proposition 94, the proposition holds true.

In a special case where $N = 1$ and $\Omega = (a, b)$, let us briefly write

$$\mathrm{H}^1(a, b) \quad \text{and} \quad \mathrm{H}_0^1(a, b)$$

instead of $\mathrm{H}^1(\Omega)$ and $\mathrm{H}_0^1(\Omega)$, respectively.

5.3.2 Examples

Structure of $u \in \mathrm{H}^1(a, b)$

Let $-\infty < a < b < \infty$. If $u \in \mathrm{H}_0^1(a, b)$, then there exists a unique *continuous* function $v : [a, b] \to \mathbb{R}$ such that $u(x) = v(x)$ for almost all $x \in [a, b]$ and

$$v(a) = v(b) = 0.$$

In addition, we have the estimate

$$\|v\| = \max_{x\in[a,b]} |v(x)| \le \sqrt{b-a} \cdot \sqrt{\int_a^b \left(\frac{\mathrm{d}\,u}{\mathrm{d}\,x}\right)^2 \mathrm{d}\,x} \le \sqrt{b-a}\,\|u\|_{\mathrm{H}^1}\,.$$

since

$$\langle u|v\rangle_{\mathrm{H}^1} = \int_a^b (uv + u'v')\,\mathrm{d}\,x \quad \text{and} \quad \|u\|_{\mathrm{H}^1} = \sqrt{\int_a^b [u^2 + (u')^2]\,\mathrm{d}\,x}$$

for all $u, v \in \mathrm{H}^1(a, b)$.
Proof.

- Uniqueness of v. If two continuous function $v, w : [a, b] \to \mathbb{R}$ differ at a single point, then they also differ on a small interval J with means $\mathrm{E}\{J\} > 0$. Hence

$$\text{“}v(x) = w(x) \text{ for almost all } x \in (a, b)\text{”} \quad \Longrightarrow \quad v(x) = w(x) \text{ on } [a, b]$$

- Existence of v.
  - First let $w \in C_0^\infty(a, b)$. Then

$$w(x) = \int_a^x w'(y)\,\mathrm{d}\,y, \quad x \in [a, b].$$

By the Cauchy-Schwarz inequality, we can obtain

$$|w(x)| \le \int_a^b 1 \cdot w'(y)\,\mathrm{d}\,y \le \sqrt{\int_a^b \mathrm{d}\,y}\sqrt{\int_a^b |w'(y)|^2\,\mathrm{d}\,y}$$
$$\le \sqrt{b-a}\,\|w\|_{\mathrm{H}^1}, \quad \forall x \in [a,b].$$

– Now let $u \in \mathrm{H}_0^1(a,b)$. Then, there exists a sequence (v_n) in $C_0^\infty(a,b)$ such that

$$\|v_n - u\|_{\mathrm{H}^1} \to 0, \qquad n \to \infty.$$

Since (v_n) is a Cauchy sequence in $\mathrm{H}^1(a,b)$, it follows from

$$\max_{x\in[a,b]} |v_n(x) - v_m(x)| \le \sqrt{b-a}\,\|n_n - v_m\|_{\mathrm{H}^1}$$

that (v_n) is also a Cauchy sequence in the Banach space $C[a,b]$. Thus, there is a function $v \in C[a,b]$ such that

$$\max_{x\in[a,b]} |v_n(x) - v(x)| \to 0, \quad n \to \infty.$$

Since $v_n(a) = v_n(b) = 0$ for all n, this implies $v(a) = v(b) = 0$.

– Finally, it follows from

$$\int_a^b (v-u)^2\,\mathrm{d}\,x = \lim_{n\to\infty}\int_a^b (v_n-u)^2\,\mathrm{d}\,x \le \lim_{n\to\infty}\|v_n-u\|_{\mathrm{H}^1}^2 = 0$$

that $v(x) = u(x)$ for almost all $x \in [a,b]$, i.e., $v \overset{a.e.}{=} u$ in $[a,b]$. ■

This example shows that the function $u \in \mathrm{H}_0^1(a,b)$ posses a simple structure.

Piecewise Continuously Differentiable Functions

Let $-\infty < a < b < \infty$, and let the function

$$u : [a,b] \to \mathbb{R}$$

be continuous and piecewise continuously differentiable. Denote by S_c the set of points x where the classic derivative exists. Define the realo function

$$w(x) \triangleq \begin{cases} u'(x), & x \in S_c; \\ \text{arbitraryvalue}, & x \notin S_c. \end{cases}$$

More precisely, we assume the following:

(a) The function u is continuous on $[a,b]$.

(b) There exists a finite number of points a_j with

$$a = a_0 < a_1 < \cdots < a_n = n$$

such that, for all j, u is continuously differentiable on the open subintervals (a_j, a_{j+1}) and the derivative u' can be extended continuously to the closed subinterval $[a_j, a_{j+1}]$.

Then, the function u has the following properties:

① The function w is the generalized derivative of u on (a,b) , i.e., $w = u'$ on (a,b).

② $u \in \mathrm{H}^1(a,b)$.

③ $u \in \mathrm{H}_0^1(a,b)$ iff $u(a) = u(b) = 0$.

PROOF

- Ad ①. Divide the interval $[a,b]$ into the subintervals $[a_j, a_{j+1}]$ and use integration by parts.
- Ad ②. Since u is continuous and $w = u'$ is piecewise continuous and bounded, we get

$$\int_a^b u^2 \,\mathrm{d}\, x < \infty \quad \text{and} \quad \int_a^b [u']^2 \,\mathrm{d}\, x < \infty.$$

  Hence $u \in \mathrm{H}^1(a,b)$.
- Ad ③.
  - If $u \in \mathrm{H}_0^1(a,b)$, then $u(a) = u(b) = 0$ by the previous example. Conversely, let $u(a) = u(b) = 0$. Choose a number $\eta > 0$. By smoothing the function u at the corners, we obtain a function $v \in C^1[a,b]$ such that v vanishes in a neighborhood of the two boundary points $x = a$ and $x = b$ along with

$$\|u - v\|_{\mathrm{H}^1}^2 = \int_a^b [(u-v)^2 + (u'-v')^2] \,\mathrm{d}\, x < \eta^2.$$

    The idea of the construction of the function v related to u is pictured in Fig. xxxx.

FIGURESHERE

  - Choose the function j_ε as in Eq.(**??**). Letting

$$v_\varepsilon \triangleq j_\varepsilon * v = \int_a^b j_\varepsilon(x-y) v(y) \,\mathrm{d}\, y,$$

    then $v_\varepsilon \in C_0^\infty(a,b)$, for sufficiently small $\varepsilon > 0$, and

$$u_\varepsilon \to v \quad \text{in} \quad L^2(a,b) \qquad \text{as} \quad \varepsilon \to 0+.$$

    Differentiation and integration by parts yield

$$v_\varepsilon'(x) = \int_a^b j_\varepsilon'(x-y) v(y) \,\mathrm{d}\, y = \int_a^b j_\varepsilon(x-y) v'(y) \,\mathrm{d}\, y.$$

    Hence

$$v_\varepsilon' \to v' \qquad \text{in} \quad L^2(a,b) \quad \text{as} \quad \varepsilon \to 0+.$$

    Summarizing, we get

$$\|u - v_\varepsilon\|_{\mathrm{H}^1} \le \|u - v\|_{\mathrm{H}^1} + \|v - v_\varepsilon\|_{\mathrm{H}^1} < \eta$$

    for sufficiently small $\varepsilon > 0$.
  - Set $\varepsilon = \frac{1}{n}$ and $u_n \triangleq v_{1/n}$. Since $\eta > 0$ is arbitrary,

$$u_n \to u \quad \text{in} \quad \mathrm{H}^1(a,b) \qquad \text{as} \quad n \to \infty,$$

    where $u_n \in C_0^\infty(a,b)$ for all n. Hence $u \in \mathrm{H}_0^1(a,b)$. ■

5.4 Generalized Boundary Values

Definition 140. *Let Ω be a nonempty bounded open set in $\mathbb{R}^N, N \ge 1$. If $u \in \mathrm{H}_0^1(\Omega)$, then we say that the function u satisfies the boundary condition (B.C.)*

$$u(x) = 0, \quad x \in \partial\,\Omega \tag{5.14}$$

in the generalized sense.

Motivation of Eq.(5.14): A very formal motivation is based on the fact that the set $C_0^\infty(\Omega)$ is dense in $\mathrm{H}_0^1(\Omega)$ and the functions $u \in C_0^\infty(\Omega)$ vanish on a boundary strip of Ω, i.e., u satisfies condition (5.14) in the classic sense. A more convincing motivation is obtained as follows:

(a) Let $\Omega \subset \mathbb{R}^{N\times 1}$ with $N = 1$ and $\Omega = (a, b)$. Then Eq.(5.14) holds true in the "classic sense" by the first example in the previous section.

(b) Let $\Omega \subset \mathbb{R}^{N\times 1}$ with $n \geq 2$, and suppose that the boundary $\partial\,\Omega$ of the nonempty bounded open set Ω is sufficiently regular. Then it can be proved that

$$\int_{\partial\,\Omega} u^2 \,\mathrm{d}\, s \leq \text{const} \int_\Omega \left[u^2 + \sum_{j=1}^{N} (\partial_j\, u)^2 \right] \mathrm{d}\, x, \quad \forall u \in \mathrm{H}_0^1(\Omega). \tag{5.15}$$

This implies the following:

If $u \in \mathrm{H}^1(\Omega)$, then there exists a sequence (u_n) in $C_0^\infty(\Omega)$ such that $u_n \to u$ in $\mathrm{H}^1(\Omega)$ as $n \to \infty$.

By (5.15),

$$\int_{\partial\,\Omega} (u - u_n)^2 \,\mathrm{d}\, s \leq \text{const}\, \|u - u_n\|_{\mathrm{H}^1}^2 \to 0, \quad n \to \infty.$$

Since $u_n = 0$ on $\partial\,\Omega$, we get

$$\int_{\partial\,\Omega} u^2 \,\mathrm{d}\, s = 0,$$

and hence

$$u(x) \overset{a.e.}{=} 0, \quad x \in \partial\,\Omega,$$

in the sense of the surface measure on $\partial\,\Omega$.

5.5 Poincaré-Friedrichs Inequality

5.5.1 Example

We prove the following inequality

$$\int_a^b u^2 \,\mathrm{d}\, x \leq (b - a)^2 \int_a^b |u'|^2 \,\mathrm{d}\, x, \quad \forall u \in \mathrm{H}_0^1(a, b).$$

Proof.

- Step-1: Let $u \in C_0^\infty(a, b)$. Then

$$u(x) = \int_a^x u'(y) \,\mathrm{d}\, y, \quad x \in x[a, b].$$

By the Cauchy-Schwarz inequality , we have

$$|u(x)|^2 \leq \left(\int_a^b 1 \cdot |u'| \,\mathrm{d}\, y \right)^2 \leq \int_a^b \mathrm{d}\, y \int_a^b |u'|^2 \,\mathrm{d}\, 7,$$

and hence

$$\int_a^b u^2 \,\mathrm{d}\, x \leq (b - a)^2 \int_a^b |u'|^2 \,\mathrm{d}\, x.$$

- Step-2: Let $u \in \mathrm{H}_0^1(a,b)$. Then, there is a sequence (u_n) in $C_0^\infty(a,b)$ such that $\|u-u_n\|_{\mathrm{H}^1} \to 0$ as $n \to \infty$. Hence

$$u_n \to u \quad \text{in} \quad L^2(a,b)$$

and

$$u_n' \to u' \quad \text{in} \quad L^2(a,b) \quad \text{as} \quad n \to \infty.$$

By Step-1,

$$\int_a^b u_n^2 \,\mathrm{d}\, x \le (b-a)^2 \int_a^b |u_n'|^2 \,\mathrm{d}\, x, \quad \forall n.$$

Letting $n \to \infty$, this implies the spacial Poincaré-Fridrichs inequality,

$$\int_a^b u^2 \,\mathrm{d}\, x \le (b-a)^2 \int_a^b |u'|^2 \,\mathrm{d}\, x, \quad \forall u \in (a,b). \qquad \blacksquare.$$

5.5.2 Theorem

Proposition 141 (Poincaré-Friedrichs Inequality)**.** *Let Ω be a nonempty bounded open set in $\mathbb{R}^{N\times 1}, N = 1, 2, \cdots, \cdots$. Then there exists a constant $c(\Omega) > 0$ such that the inequality*

$$\begin{aligned} &\int_\Omega u^2 \,\mathrm{d}\, x \le c(\Omega) \int_\Omega \sum_{j=1}^N (\mathrm{D}_j\, u)^2 \,\mathrm{d}\, x, \quad \forall u \in \mathrm{H}_0^1(\Omega) \\ &\|u\|_{L^2(\Omega)} \le C(\Omega) \|\nabla u\|_{L^2(\Omega)}, \quad \forall u \in \mathrm{H}_0^1(\Omega) \end{aligned} \tag{5.16}$$

holds, where $C(\Omega) = \sqrt{c(\Omega)}$.

Poincaré-Friedrichs inequality shows that: if Ω is a bounded region, then the semi-norm

$$|u|_{\mathrm{H}^1(\Omega)} \triangleq \sqrt{\sum_{j=1}^n \|\mathrm{D}_j\, u\|_{L^2(\Omega)}^2}$$

defined on $\mathrm{H}^1(\Omega)$ is a norm that is equivalent to the norm $\|u\|_{\mathrm{H}^1(\Omega)}$.

5.6 Soblev Embedding Theorem and Negative Soblev Spaces

Question. For the space $\mathrm{W}^{m,p}(\Omega)$, if we fix the number p and n (the dimension of Ω), then whether the m should be large enough such that the functions in $\mathrm{W}^{m,p}(\Omega)$ are continuous?

5.6.1 Embedding

Definition 142. *Let both $\mathcal{X}$ and $\mathcal{Y}$ are normed linear spaces. If $\mathcal{X}$ is a linear subspace of $\mathcal{Y}$ and there exists a one-to-one and coninuous mapping (homeomorphism), then we call $\mathcal{X}$ is embedded into $\mathcal{Y}$, which is denoted as $\mathcal{X} \overset{\subsetneq}{\hookrightarrow} \mathcal{Y}$.*

If $\mathcal{X}$ in embedded into $\mathcal{Y}$, then there exists a constant $c > 0$ such that

$$\|\gamma x\|_{\mathcal{Y}} \le c \|x\|_{\mathcal{X}}, \quad \forall x \in \mathcal{X}, \tag{5.17}$$

where the c is called the embedding constant. Furthermore, if γ is a compact operator, the we call it compact embedding.

Example. Let $\Omega \subset \mathbb{R}^{n\times 1}$, then for $m \in \mathbb{N}$ we have

$$C^{m+1}(\overline{\Omega}) \overset{\subsetneq}{\hookrightarrow} C^m(\overline{\Omega}). \tag{5.18}$$

Actually, $C^{m+1}(\overline{\Omega})$ is a linear subspace of $C^m(\overline{\Omega})$ and for all $f \in C^{m+1}(\overline{\Omega})$. Let $g = \gamma f$, for any fixed $j \in \{1, \cdots, n\}$ we define

$$g(x) = (\gamma f)(x) = \int_0^{x_j} f(x_1, \cdots, x_{j-1}, \xi_j, x_{j+1}, \cdots, x_n)\, \mathrm{d}\, \xi_j, \quad \forall x \in \Omega.$$

Thus γ is an integral operator. Obviously, $g \in C^m(\overline{\Omega})$. Moreover, we can prove that γ is a one-to-one compact operator. Therefore, the embedding is a compact embedding.

Theorem 143 (Soblev Embedding Theorem)**.** *Let $\Omega \subset \mathbb{R}^{n\times 1}$ is bounded and simply connected, $1 \le k \le n$, S_k is a k-dim hyperplane in $\mathbb{R}^{n\times 1}$, i.e.,*

$$S^k = \left\{ x \in \mathbb{R}^{n\times 1} : \boldsymbol{A}x = b, \boldsymbol{A} \in \mathbb{R}^{n\times n}, \mathrm{Rank}(\boldsymbol{A}) = k \le n, b \in \mathbb{R}^{n\times 1} \right\},$$

and

$$\gamma^k = \Omega \cap S^k$$

is a k-dim regiin that is bounded and simpley connected. Let $j, m \in \mathbb{N}$, then there exist the following embeddings:

① *When $mp < n$ and $n - mp < k \le n$, if $p \le q \le \frac{kp}{n-mp}$, then*

$$\mathrm{W}^{m+j,p}(\Omega) \overset{\subsetneq}{\to} \mathrm{W}^{j,q}(\gamma^k)$$

② *When $mp = n$, for all $k \in \{1, \cdots, n\}$, if $p \le q \le \infty$, then*

$$\mathrm{W}^{m+j,p}(\Omega) \overset{\subsetneq}{\to} \mathrm{W}^{j,q}(\Gamma^k)$$

③ *When $mp > n$, if $1 \le p \le \infty$, then*

$$\mathrm{W}^{m+j,p} \overset{\subsetneq}{\to} \overline{C^j(\Omega)}$$

is a compact embedding.

We remark that

❶ for $mp > n$, each element in $\mathrm{W}^{m,p}(\Omega)$ is a continusous function.

❷ for all $n \in \mathbb{N}, \mathrm{W}^{m,p}(\Omega) \overset{\subsetneq}{\to} L^p(\Omega) = \mathrm{W}^{0,p}(\Omega)$.

❸ when $m \ge 1$ and $1 \le p \le \infty$, $\mathrm{W}^{m,p}(\Omega) \overset{\subsetneq}{\to} \mathrm{W}^{m-1,p}(\Omega)$ is a compact embedding.

5.6.2 Negative Soblev Space $\mathrm{W}^{-m,p}(\Omega)$ as a Dual Space

The dual space of $\mathrm{W}_0^{m,p}(\Omega)$ is

$$\mathrm{W}_0^{m,p}(\Omega)^* = \mathrm{W}_0^{-m,q}(\Omega), \quad m \in \mathbb{N}, \quad \frac{1}{p} + \frac{1}{q} = 1, \tag{5.19}$$

which is called $(-m, p)$-order *negative Soblev space*. The norm in $\mathrm{W}^{-m,p}(\Omega)$ is defined as

$$\|g\|_{\mathrm{W}^{-m,q}(\Omega)} = \sup_{\substack{f \in \mathrm{W}_0^{m,p}(\Omega) \\ f \neq 0}} \frac{|\langle g, f \rangle|}{\|f\|_{\mathrm{W}^{m,p}(\Omega)}} \tag{5.20}$$

according to the formula

$$\|\gamma\| = \sup_{\substack{x \in \mathcal{X} \\ x \neq 0}} \frac{|\langle \gamma, x \rangle|}{\|x\|}, \quad \forall \gamma \in \mathcal{X}^*.$$

For $p = 2$, the dual space of the Hilbert space $\mathrm{H}_0^m(\Omega) = \mathrm{W}_0^{m,2}(\Omega)$ is

$$\mathrm{H}_0^m(\Omega)^* = \mathrm{H}^{-m}(\Omega), \quad m \in \mathbb{N}. \tag{5.21}$$

What's the structure of a given element in the space $\mathrm{H}^{-m}(\Omega)$? The following representation answers this question perfectly.

Theorem 144 (Representation Theorem of Negative Soblev Space)**.** *When the integer $m \geq 1$, each element F of $\mathrm{H}^{-m}(\Omega)$ can be represented by*

$$\begin{aligned} F &= \sum_{\alpha=0}^{m} (-1)^{|\alpha|} \, \mathrm{D}^{\alpha} f_{\alpha} \\ &= \sum_{|\alpha|=0}^{m} (\mathrm{D}^{\alpha})^{*} f_{\alpha}, \quad f_{\alpha} \in L^2(\Omega). \end{aligned} \tag{5.22}$$

Proof.

- For $m = 0$, $\mathrm{H}_0^0(\Omega) = \mathrm{H}^0(\Omega) = L^2(\Omega)$, its dual space is $\mathrm{H}_0^{-0}(\Omega) = L^2(\Omega)$. By the Fréchet-Riesz theorem, for all $F \in \mathrm{H}^{-0}(\Omega)$, there exists $v \in L^2(\Omega)$ such that

$$F(u) = \int_{\Omega} v(x) u(x) \, \mathrm{d}\, x, \quad \forall u \in \mathrm{H}_0^0(\Omega) = L^2(\Omega).$$

- For $m \geq 1$, let

$$N(m) = |\{\alpha = (\alpha_1, \cdots, \alpha_n) \in \mathbb{N}^n : 0 \leq |\alpha| \leq m\}|$$

be the number of the multi-index α such that its order is not greater than m. The $\mathrm{H}_0^m(\Omega)$ can be regarded as a subspace $\mathcal{L}$ of the product space with continuous funcitonal F. With the help of the Hahn-Banach theorem, the F can be extended from $\mathcal{L}$ to $\prod_{j=1}^{N(m)} L^2(\Omega)$ and the norm $\|F\|$ can be kept. Again by the the Fréchet-Riesz theorem on the product space (of course it is a Hilbert space), there exists some $f_{\alpha} \in L^2(\Omega)$ such that

$$F(\{v_{\alpha}\}) = \sum_{|\alpha|=0}^{m} \int_{\Omega} f_{\alpha}(x) v_{\alpha}(x) \, \mathrm{d}\, x, \quad \{v_{\alpha}\} \in \prod_{j=1}^{N(m)} L^2(\Omega).$$

Particularly, put

$$\{v_{\alpha}\} = \{\mathrm{D}^{\alpha} u\}, \quad u \in C_0^{\infty}(\Omega),$$

then

$$\begin{aligned} F(u) = \langle F, u \rangle &= \sum_{|\alpha|=0}^{m} \int_{\Omega} f_{\alpha} \, \mathrm{D}^{\alpha} u(x) \, \mathrm{d}\, x \\ &= \sum_{|\alpha|=0}^{m} \langle f_{\alpha}, \mathrm{D}^{\alpha} u \rangle \\ &= \sum_{|\alpha|=0}^{m} (-1)^{|\alpha|} \langle \mathrm{D}^{\alpha} f_{\alpha}, u \rangle \\ &= \left\langle \sum_{|\alpha|=0}^{m} (-1)^{|\alpha|} \, \mathrm{D}^{\alpha} f_{\alpha}, u \right\rangle \end{aligned}$$

In consequence,

$$F = \sum_{|\alpha|=0}^{m} (-1)^{|\alpha|} \, \mathrm{D}^{\alpha} f_{\alpha} = \sum_{|\alpha|=0}^{m} (\mathrm{D}^{\alpha})^{*} f_{\alpha}.$$

- Conversely, if F can be represented by Eq.(5.22), then

$$\begin{aligned} |\langle F, u \rangle| &= \left| \sum_{|\alpha|=0}^{m} \langle f_{\alpha}, \mathrm{D}^{\alpha} u \rangle \right| = \left| \sum_{|\alpha|=0}^{m} \int_{\Omega} f_{\alpha} \, \mathrm{D}^{\alpha} u(x) \, \mathrm{d}\, x \right| \\ &\leq \sum_{|\alpha|=0}^{m} \|f_{\alpha}\|_{L^2(\Omega)} \cdot \|u\|_{L^2(\Omega)}, \quad \forall u \in C_0^{\infty}(\Omega). \end{aligned}$$

This implies that F is a bounded linear functional. Since $C_0^\infty(\Omega)$ is dense in $\mathrm{H}_0^m(\Omega)$, then

$$F \in (\mathrm{H}_0^m(\Omega)^* = \mathrm{H}^{-m}(\Omega).$$

Corollary 145. $\mathrm{H}^{-m}(\Omega)$ *is also a Hilbert space with the inner product defined by*

$$\langle F|G\rangle_{\mathrm{H}^{-m}(\Omega)} = \langle \mathsf{K}^{-1}F|\mathsf{K}^{-1}G\rangle_{\mathrm{H}^m(\Omega)}, \tag{5.23}$$

where $\mathsf{K} : \mathrm{H}_0^m(\Omega) \to \mathrm{H}^{-m}(\Omega)$ *is Frécher-Riesz mapping. Furthermore, the dense embedding relationConstruction of the trace operator*

$$\mathrm{H}_0^m(\Omega) \overset{\subsetneq}{\to} \mathrm{H}^m(\Omega) \overset{\subsetneq}{\to} \mathrm{H}^0(\Omega) = L^2(\Omega) \overset{\subsetneq}{\to} \mathrm{H}^{-m}(\Omega) \overset{\subsetneq}{\to} \prod_{j=1}^{N(m)} L^2(\Omega). \tag{5.24}$$

holds true.

5.7 Fractional Soblev Space and Trace Operator

5.7.1 Introduction

In mathematics, the concept of trace operator plays an important role in studying the existence and uniqueness of solutions to boundary value problems, that is, to partial differential equations with prescribed boundary conditions. The trace operator makes it possible to extend the notion of restriction of a function to the boundary of its domain to generalized functions in a Sobolev space.

Let Ω be a bounded open set in the Euclidean space $\mathbb{R}^{n\times 1}$ with C^1 boundary $\partial\,\Omega$. If u is a function that is C^1 (or even just continuous) on the closure $\overline{\Omega}$ of Ω, its function restriction is well-defined and continuous on $\partial\,\Omega$. If however, u is the solution to some partial differential equation, it is in general a weak solution, so it belongs to some Sobolev space. Such functions are defined only up to a set of measure zero, and since the boundary $\partial\,\Omega$ does have measure zero, any function in a Sobolev space can be completely redefined on the boundary without changing the function as an element in that space. It follows that simple function restriction cannot be used to meaningfully define what it means for a general solution to a partial differential equation to behave in a prescribed way on $\partial\,\Omega$.

The way out of this difficulty is the observation that while an element u in a Sobolev space may be ill-defined as a function, u can be nevertheless approximated by a sequence (u_k) of C^1 functions defined on $\overline{\Omega}$. Then, the restriction $u_{|\,\partial\,\Omega}$ of u to $\partial\,\Omega$ is defined as the limit of the sequence of restrictions $(u_{k|\,\partial\,\Omega})$.

Example. Generally a function $v \in \mathrm{H}^1(\Omega)$ (or $\mathrm{H}^m(\Omega)$) is not continuous, thus it may be impossible to determine its values on the bundary $\partial\,\Omega$.

For illustration, let $\Omega = \left\{(x_1, x_2) : 0 \le r = \sqrt{x_1^2 + x_2^2} \le R < 1\right\}$, and

$$v(r) = \left[\ln\frac{1}{r}\right]^k, \quad v'(r) = k\left[\ln\frac{1}{r}\right]^{k-1}\frac{1}{r} \tag{5.25}$$

then for $k < \frac{1}{2}, v \in \mathrm{H}^m(\Omega)$. In fact,

$$\|v\|^2_{L^2(\Omega)} = \int_\Omega |v|^2\,\mathrm{d}\,\mu = 2\pi\int_0^R \left[\ln\frac{1}{r}\right]^{2k} r\,\mathrm{d}\,r < +\infty, \quad k < 1,$$
$$|v|^2_{\mathrm{H}^1(\Omega)} = \int_\Omega |v'(r)|^2\,\mathrm{d}\,\mu = 2\pi\int_0^R \left[\ln\frac{1}{r}\right]^{2(k-1)} r^{-2}\,\mathrm{d}\,r < +\infty, \quad k < \frac{1}{2}.$$

Therefore, when $k \in (0, 1/2)$, $v \in \mathrm{H}^1(\Omega)$. However, there is a singular point at origin for v.

Only for $n = 1, \Omega = (a, b) \subset \mathbb{R}$, there exists some $f \in C[a, b]$ such $v \overset{a.e.}{=} f$ on Ω. In this case, there is no difficulty for determine the boundary $v(a)$ and $v(b)$. However, for $n \geq 2$ and $v \in \mathrm{H}^1(\Omega)$, it is not easy to determine the values on the boundary, i.e., $v|_{\partial\,\Omega}$. This is the most important reason to introduce the trace operator and trace theorem.

5.7.2 Construction of the Trace Operator in $\mathrm{W}^{1,p}(\Omega)$

Definition

To rigorously define the notion of restriction to a function in a Sobolev space, let $p \geq 1$ be a real number. Consider the linear operator

$$\gamma : C^1(\overline{\Omega}) \to L^p(\partial\,\Omega) \tag{5.26}$$

defined on the set of all C^1 functions on $\overline{\Omega}$ with values in the space $L^p(\partial\,\Omega)$, given by the formula

$$\gamma u = u_{|\,\partial\,\Omega} \tag{5.27}$$

The domain of γ is a subset of the Sobolev space $\mathrm{W}^{1,p}(\Omega)$. It can be proved that there exists a constant C depending only on Ω and p, such that

$$\|\gamma u\|_{L^p(\partial\,\Omega)} \leq C \|u\|_{\mathrm{W}^{1,p}(\Omega)}, \quad \forall u \in C^1(\overline{\Omega}). \tag{5.28}$$

Then, since the C^1 functions on $\overline{\Omega}$ are dense in $\mathrm{W}^{1,p}(\Omega)$, the operator γ admits a continuous extension

$$\gamma : \mathrm{W}^{1,p}(\Omega) \to L^p(\partial\,\Omega) \tag{5.29}$$

defined on the entire space $\mathrm{W}^{1,p}(\Omega)$. γ is called the trace operator. The restriction (or trace) $u_{|\,\partial\,\Omega}$ of a function u in $\mathrm{W}^{1,p}(\Omega)$ is then defined as γu.

This argument can be made more concrete as follows. Given a function u in $\mathrm{W}^{1,p}(\Omega)$, consider a sequence of functions (u_n) that are C^1 on $\overline{\Omega}$, with u_n converging to u in the norm of $\mathrm{W}^{1,p}(\Omega)$. Then, by the above inequality, the sequence $u_{n|\,\partial\,\Omega}$ will be convergent in $L^p(\partial\,\Omega)$. Define

$$u_{|\,\partial\,\Omega} = \lim_{n\to\infty} u_{n\,|\,\partial\,\Omega}. \tag{5.30}$$

It can be shown that this definition is independent of the sequence (u_n) approximating u.

Application in PDE

Consider the problem of solving Poisson's equation with zero boundary conditions:

$$\begin{cases} -\Delta u = f \text{ in } \Omega \\ u_{|\,\partial\,\Omega} = 0. \end{cases} \tag{5.31}$$

Here, f is a given continuous function on $\overline{\Omega}$.

With the help of the concept of trace, define the subspace $H_0^1(\Omega)$ to be all functions in the Sobolev space $\mathrm{W}^{1,2}(\Omega) = \mathrm{H}^1(\Omega)$ whose trace is zero. Then, the equation above can be given the weak formulation. Find u in $\mathrm{H}_0^1(\Omega)$ such that

$$\int_\Omega \nabla u(x) \cdot \nabla v(x)\,\mathrm{d}\,x = \int_\Omega f(x) v(x)\,\mathrm{d}\,x, \quad \forall v \in \mathrm{H}_0^1(\Omega). \tag{5.32}$$

Using the Lax – Milgram theorem one can then prove that this equation has precisely one solution, which implies that the original equation has precisely one weak solution.

One can employ similar ideas to prove the existence and uniqueness of solutions for more complicated partial differential equations and with other boundary conditions (such as Neumann and Robin), with the notion of trace playing an important role in all such problems.

5.7.3 Trace Theorem in $\mathrm{H}^1(\Omega)$

We intruduce the following notations

$$\mathbb{R}_+^{n\times 1} \triangleq \left\{x : x = (x', x_n) \in \mathbb{R}^{n\times 1}, x_n > 0\right\},$$
$$\Gamma \triangleq \partial\,\Omega \triangleq \left\{x : x = (x', 0), x' \in \mathbb{R}^{(n-1)\times 1}\right\},$$

Theorem 146. *$\mathcal{D}(\overline{\mathbb{R}_+^{n\times 1}})$ is dense in $\mathrm{H}^1(\overline{\mathbb{R}_+^{n\times 1}})$.*

Proof.

- Omitted here.

Theorem 147. *For all $u \in \mathcal{D}(\overline{\mathbb{R}_+^{n\times 1}})$, we have*

$$\|u(\cdot, 0)\|_{L^2(\mathbb{R}^{(n-1)\times 1})} \leq \|u\|_{\mathrm{H}^1(\mathbb{R}_+^{n\times 1})} \tag{5.33}$$

Proof.

- For $u \in \mathcal{D}(\overline{\mathbb{R}_+^{n\times 1}})$, we can deducet that

$$\begin{aligned}
[u(x', 0)]^2 &= -\int_0^\infty \frac{\partial}{\partial x_n}[u(x', x_n)]^2\,\mathrm{d}\,x_n \\
&= -2\int_0^\infty u(x', x_n)\frac{\partial}{\partial x_n}u(x', x_n)\,\mathrm{d}\,x_n \\
&\leq 2\left\{\int_0^\infty [u(x', x_n)]^2\,\mathrm{d}\,x_n\right\}^{1/2}\left\{\int_0^\infty \left[\frac{\partial}{\partial x_n}u(x', x_n)\right]^2\mathrm{d}\,x_n\right\} \\
&\leq \int_0^\infty \left\{\int_0^\infty [u(x', x_n)]^2 + \left[\frac{\partial}{\partial x_n}u(x', x_n)\right]^2\right\}\mathrm{d}\,x_n
\end{aligned}$$

By integrating the two sides w.r.t x', we immediately have

$$\int_{\mathbb{R}^{(n-1)\times 1}} [u(x', x_n)]^2\,\mathrm{d}\,x' \leq \|u\|^2_{L^2(\mathbb{R}^{n\times 1}} + \left\|\frac{\partial u}{\partial x_n}\right\|^2_{L^2(\mathbb{R}^{n\times 1}}.$$

Thus the Eq.(5.33) follows from the definition of the norm defined in $\mathrm{H}^1(\Omega)$.

Combining the Theorem 146 and Theorem 147 we find that:

① The map

$$\begin{aligned}
\gamma_0 : \mathcal{D}(\overline{\mathbb{R}_+^{n\times 1}}) &\to \mathcal{D}(\mathbb{R}^{(n-1)\times 1}) \\
u(\cdot, x_n) &\mapsto u(\cdot, 0)
\end{aligned}$$

can be continuously extended to a new linear and continuous mapping in $\mathrm{Hom}(\mathrm{H}^m(\mathbb{R}_+^{n\times 1}), L^2(\mathbb{R}^{(n-1)\times 1})$.

② When $\Omega = \mathbb{R}_+^{n\times 1}$, for all $u \in \mathrm{H}^1(\Omega)$, we can define its value on $\Gamma = \partial\,\Omega$ as a function in $L^2(\Gamma)$.

③ The result ② can be generalized to the case when $\Omega \subset \mathbb{R}^{n\times 1}$ is bounded, but the $\partial\,\Omega$ should be smooth enough.

Theorem 148 (Trace Theorem on $\mathrm{H}^1(\Omega)$)**.** *Let $\Omega \subset \mathbb{R}^{n\times 1}$ be open and bounded, its boundary $\partial\,\Omega$ is sufficiently smooth (first order regular), then $\mathcal{D}(\overline{\Omega})$ is dense in $\mathrm{H}^1(\Omega)$, and the mapping $\gamma_0 : \mathcal{D}(\overline{\Omega}) \to C^0(\partial\,\Omega)$ can be continuously extened to a continuous mapping in $\mathrm{Hom}(\mathrm{H}^1(\Omega) \to L^2(\partial\,\Omega)$, which is still denoted by γ_0 and called trace operator. For $v \in \mathrm{H}^1(\Omega)$, $\gamma_0 v$ is called the trace of v on $\partial\,\Omega$.*

The smoothness can be measure by m-th order regularity. We say that an open set Ω in $\mathbb{R}^{n\times 1}$ is m-th order regular if

❶ Ω is bounded;

❷ $\partial\,\Omega \in C^m$, i.e., $\partial\,\Omega$ is a $n-1$ dimensional C^m manifold and if it is observed locally, then Ω is only in a side of $\partial\,\Omega$.

5.7.4 Fractional Soblev Space and the General Trace Theorem

For any fraction $\sigma \in (0,1)$ and function $g : \mathbb{R}^{n\times 1} \to \mathbb{R}^{n\times 1}$, we define

$$\|g\|_{\sigma,\Omega}^2 \triangleq \int_\Omega \int_\Omega \frac{\|g(x)-g(y)\|}{\|x-y\|^{n+2\sigma}}\,\mathrm{d}\,x\,\mathrm{d}\,y \tag{5.34}$$

where $\|x\| = \sqrt{\sum_{j=1}^{n} x_j^2}$ denotes the Euclidean norm in $\mathbb{R}^{n\times 1}$.

For any $s \in \mathbb{R}^+$, let $s = \lfloor s \rfloor + \sigma, \lfloor s \rfloor \in \mathbb{Z}^+, \sigma \in [0,1)$. The fractional Soblev space is define by

$$\mathrm{H}^s(\Omega) \triangleq \left\{ f : \|f\|_{s,\Omega} < \infty \right\} \tag{5.35}$$

where

$$\|f\|_{s,\Omega}^2 \triangleq \|f\|_{\mathrm{H}^{\lfloor s \rfloor}(\Omega)}^2 + \sum_{|\alpha|=\lfloor s \rfloor} \|\mathrm{D}^\alpha f\|_{\sigma,\Omega}^2 \tag{5.36}$$

Similarly, for $s \in \mathbb{R}$, we can also define Soblev space.

Theorem 149. *For any positive number $s \in \mathbb{R}$, the fractional Soblev space $\mathrm{H}^s(\Omega)$ is a Hilbert space.*

Let $\partial\,\Omega$ be the boundary of Ω, we can define the Soblev space $\mathrm{H}^s(\partial\,\Omega)$ and get its dual space $\mathrm{H}^{-s}(\partial\,\Omega)$.

Definition 150. *Let $f \in C^m(\overline{\Omega})$ (and hence $f \in \mathrm{H}^m(\Omega)$). Put*

$$\gamma_j f \triangleq \left.\frac{\partial^j f}{\partial\, n^j}\right|_{\partial\,\Omega}, \quad j \in \{0,\cdots,m-1\} \tag{5.37}$$

where the outer normal vector $\boldsymbol{n} = (n_1,\cdots,n_j,\cdots)$ is defined almost everywhere in $\partial\,\Omega$. The operator γ_j is called the trace operator.

Theorem 151 (General Trace Theorem)**.** *For any $\phi \in L^2(\Omega)$, let*

$$\|\phi\| \triangleq \inf_{f\in \mathrm{H}^m(\Omega)} \left\{ \|f\|_{\mathrm{H}^m(\Omega)} : \phi = \gamma_j f \right\}, \quad j \in \{0,\cdots,m-1\}, \tag{5.38}$$

then $L^2(\partial\,\Omega)$ can be completed into a Hilbert space

$$\mathrm{H}^{m-j-\frac{1}{2}}(\partial\,\Omega), \quad j \in \{0,\cdots,m-1\}. \tag{5.39}$$

Hence

$$\gamma_j : \mathrm{H}^m(\Omega) \to \mathrm{H}^{m-j-\frac{1}{2}}(\partial\,\Omega)$$

is a continuous linear space. Therefore, there exists $C_j > 0$ such that

$$\|\gamma_j f\|_{m-j-1/2,\partial\,\Omega} \le C_j \|f\|_{\mathrm{H}^m(\Omega)}, \quad m \in \mathbb{N}, \forall f \in \mathrm{H}^m(\Omega).$$

Let $\boldsymbol{\gamma} = [\gamma_0, \gamma_1, \cdots, \gamma_{m-1}]^T$, then

$$\mathrm{Ker}(\boldsymbol{\gamma}) = \{f \in \mathrm{H}^m(\Omega) : \boldsymbol{\gamma} f = 0\} = \mathrm{H}_0^m(\Omega). \tag{5.40}$$

Since $\mathrm{H}_0^m(\Omega)$ is dense in $L^2(\Omega)$, we can deduce that $\mathrm{Ker}(\boldsymbol{\gamma})$ is dense in $L^2(\Omega)$.

Example. Let $\Omega \subset \mathbb{R}^{2\times 1}$ is an open, bounded and simply connected region, its boudary $\partial\,\Omega$ is smooth enough. Set

$$\begin{aligned}
\mathrm{H}^2(\Omega) &= \left\{ f \in L^2(\Omega) : f, \partial_x f, \partial_y f, \partial^2_{xx} f, \partial^2_{xy} f, \partial^2_{yy} f \in L^2(\Omega) \right\} \\
\mathrm{H}^2_0(\Omega) &= \left\{ f \in \mathrm{H}^2(\Omega) : f|_{\partial\,\Omega},\, \frac{\partial f}{\partial \boldsymbol{n}}\bigg|_{\partial\,\Omega} = 0 \right\},
\end{aligned}$$

where $\boldsymbol{n}$ is the outer normal vector. The trace operator $\boldsymbol{\gamma}$ is defined by

$$\boldsymbol{\gamma} f = \left[f|_{\partial\,\Omega},\, \frac{\partial f}{\partial \boldsymbol{n}}\bigg|_{\partial\,\Omega} \right]^{\mathsf{T}}.$$

Obviously, we have

$$\begin{aligned}
\mathrm{Ker}(\gamma) &= \{ f \in \mathrm{H}^m(\Omega) : \gamma f = \mathbf{0} \} \\
&= \left\{ f \in \mathrm{H}^2(\Omega) : f|_{\partial\,\Omega} = 0,\, \frac{\partial f}{\partial \boldsymbol{n}}\bigg|_{\partial\,\Omega} = 0 \right\} \\
&= \mathrm{H}^2_0(\Omega)
\end{aligned}$$

5.7.5 Trace Property

Let $\mathcal{H}$ is a Hilbert space satisfies the following conditions:

① there exists another Hilbert space $\mathcal{U}$ such that $\mathcal{H}$ is dense in $\mathcal{U}$, i.e.,

$$\mathcal{H} \subset \mathcal{U} = \mathcal{U}^* \subset \mathcal{H}^*; \tag{5.41}$$

② there exists a linear operator $\gamma : \mathcal{H} \to \partial\,\mathcal{H}$ such that $\partial\,\mathcal{H}$ is also a Hilbert space and its null space (kernal) $\mathrm{Ker}(\gamma) = \mathcal{H}_0$ is dense in $\mathcal{U}$, i.e.,

$$\mathrm{Ker}(\gamma) = \mathcal{H}_0 \subset \mathcal{H}, \quad \mathcal{H}_0 \subset \mathcal{U} = \mathcal{U}^* \subset \mathcal{H}^*. \tag{5.42}$$

Then we call that $\mathcal{H}$ has the trace property.

Example. The Soblev space $\mathrm{H}^m(\Omega)$ has the trace property. Actually, we have

$$\begin{aligned}
\mathcal{H} &= \mathrm{H}^m(\Omega) \\
\mathcal{U} &= \prod_{j=1}^{N(m)} L^2(\Omega) \\
\partial\,\mathcal{H} &= \prod_{j=1}^{N(m)} \mathrm{H}^{m-\frac{1}{2}}(\partial\,\Omega) \\
\mathcal{H}_0 &= \mathrm{H}^m_0(\Omega)
\end{aligned}$$

and $\gamma = [\gamma_0, \gamma_1, \cdots, \gamma_{m-1}]^{\mathsf{T}}$ is the trace operator.

Let $\mathcal{X}$ and $\mathcal{Y}$ be real Hilbert spaces with trace property, $a : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a bounded bilinear functional. Let $\hat{x} \in \mathcal{X}$ be fixed, then

$$\ell_{\hat{x}}(y) = a(\hat{x}, y), \quad y \in \mathcal{Y}_0 \tag{5.43}$$

is a linear and continuous functional on $\mathcal{Y}_0$. Put

$$\mathsf{A}\hat{x} \triangleq \ell_{\hat{x}}, \quad \forall \hat{x} \in \mathcal{X}, \tag{5.44}$$

then the operator $\mathsf{A} \in \mathrm{Hom}(\mathcal{X}, \mathcal{Y}_0^*)$ such that

$$a(\hat{x}, y) = \langle \mathsf{A}\hat{x}, y \rangle_{\mathcal{Y}}, \quad \forall \hat{x} \in \mathcal{X}, \forall y \in \mathcal{Y}_0, \tag{5.45}$$

is called the formal operator w.r.t. the bilinear operator $a(\cdot,\cdot)$. Similarly, let $\hat{y} \in \mathcal{Y}$ be fixed, then

$$s_{\hat{y}}(x) = a(x, \hat{y}), \forall x \in \mathcal{X}_0 \tag{5.46}$$

is a linear and continuous functional. Put

$$A^\dagger \hat{y} \triangleq s_{\hat{y}}, \quad \forall \hat{y} \in \mathcal{Y}, \tag{5.47}$$

then $A^\dagger \in \mathrm{Hom}(\mathcal{Y}, \mathcal{X}^*)$ such that

$$a(x, \hat{y}) = \left\langle x, A^\dagger \hat{y} \right\rangle, \quad \forall x \in \mathcal{X}_0, \forall \hat{y} \in \mathcal{Y}, \tag{5.48}$$

is called the formal adjoint operator of A.

Example. Let $\Omega \subset \mathbb{R}^{2\times 1}$ is an open, bounded and simply connected region, its boundary $\partial\,\Omega$ is smooth enough. Set

$$\mathcal{X} = \mathcal{Y} = \mathrm{H}^1(\Omega) = \left\{ f : f, f_x = \partial_x f, f_y = \partial_y f \in L^2(\Omega) \right\}.$$

Suppose $r(x,y), s(x,y), t(x,y)$ are functions which are sufficiently smooth (for example, they belong to $C^1(\Omega)$)and defined on Ω. We define

$$a(u,v) = \int_\Omega [r\nabla u \cdot \nabla v + sv\,\partial_x u + tv\,\partial_y u]\,\mathrm{d}\,x\,\mathrm{d}\,y$$

where $\nabla u = \mathrm{grad}(u) = [\partial_x u, \partial_y u]$ is the gradience, $\cdot$ denotes the scaler product 2-dim vector. Obviously, $a : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a bounded and bilinear functional.

For fixed $\hat{u} \in \mathcal{X}$ and for all $g \in \mathcal{Y}_0 = \mathrm{H}_0^1(\Omega) = \left\{ f \in \mathrm{H}^(\Omega) : f|_{\partial\,\Omega} = 0 \right\}$, according to the theory of PDE we have

$$\begin{aligned} a(\hat{u}, v) &= \int_\Omega v\,[-\nabla \cdot (r\nabla \hat{u}) + s\,\partial_x \hat{u} + t\,\partial_y \hat{u}]\,\mathrm{d}\,x\,\mathrm{d}\,y \\ &= \langle A\hat{u}, v \rangle, \qquad \forall v \in \mathcal{Y}_0 = \mathrm{H}_0^1(\Omega), \end{aligned}$$

where

$$A\hat{u} = -\nabla \cdot (r\nabla \hat{u}) + s\,\partial_x \hat{u} = t\,\partial_y \hat{v}$$

is a continuous funcitonal on $\mathcal{Y}_0$. Similarly, when $\hat{v}$ is fixed, the foramal dual operator

$$A^\dagger \hat{v} = -\nabla \cdot (r\nabla \hat{v}) - \partial_x(s\hat{v}) - \partial_y(t\hat{v}), \quad \forall \hat{v} \in \mathcal{X}_0 = \mathrm{H}_0^1(\Omega)$$

is a continuous linear functional on $\mathcal{X}_0 = \mathrm{H}_0^1(\Omega)$.

Theorem 152. *Let $\mathcal{X}$ and $\mathcal{Y}$ be two Hilbert spaces with trace property, and $\mathcal{U}, \mathcal{V}$ are their principle spaces respectively. $a(\cdot,\cdot) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a bounded and bilinear functional with the formal operator A and its formal adjoint $A^\dagger$. Set*

$$\begin{aligned} \mathcal{X}_A &= \{x \in \mathcal{X} : Ax \in \mathcal{V}\}, \\ \mathcal{Y}_{A^\dagger} &= \left\{ y \in \mathcal{Y} : A^\dagger y \in \mathcal{U} \right\}. \end{aligned} \tag{5.49}$$

Then there exist unique $\delta \in \mathrm{Hom}(\mathcal{X}_A, (\partial\,\mathcal{Y})^*)$ and $\delta^* \in \mathrm{Hom}(\mathcal{Y}_{A^\dagger}, (\partial\,\mathcal{X})^*)$ satisfying the following Green's formulas

$$a(x,y) = \langle y, Ax \rangle_{\mathcal{V}} + \left\langle \delta x, \gamma^\dagger y \right\rangle_{\partial\,\mathcal{X}}, \quad \forall x \in \mathcal{X}_A, \forall y \in \mathcal{Y}, \tag{5.50}$$

$$a(x,y) = \left\langle A^\dagger y, x \right\rangle_{\mathcal{U}} + \left\langle \gamma^\dagger y, \gamma x \right\rangle_{\partial\,\mathcal{Y}}, \quad \forall x \in \mathcal{X}, \forall y \in \mathcal{Y}_{A^\dagger}, \tag{5.51}$$

where $\gamma : \mathcal{X} \to \partial\,\mathcal{X}$ and $\gamma^\dagger : \mathcal{Y} \to \partial\,\mathcal{Y}$ are trace operators, $\langle \cdot,\cdot \rangle_{\partial\,\mathcal{X}}$ and $\langle \cdot,\cdot \rangle_{\partial\,\mathcal{Y}}$ represents the pair mappings defined on $(\partial\,\mathcal{X})^* \times \partial\,\mathcal{X}$ and $(\partial\,\mathcal{Y})^* \times \partial\,\mathcal{Y}$ respectively.

In the theory of partial differential equations, we have the following concepts:

- γ is called the Dirichlet operator w.r.t. A;
- δ is called the Neumann operator w.r.t. A;
- $\gamma^\dagger$ is called the Dirichlet operator w.r.t. $A^\dagger$;
- $\delta^\dagger$ is called the Neumann operator w.r.t. $A^\dagger$.

For $x \in \mathcal{X}_A$ and $y \in \mathcal{Y}_{A^\dagger}$, we can obtain the absract Green formula for $A \in \mathrm{Hom}(\mathcal{X}, \mathcal{Y}_0^*) \cap \mathrm{Hom}(\mathcal{X}_A, \mathcal{V})$ as follows

$$\left\langle A^\dagger y, x \right\rangle_{\mathcal{U}} = \langle y, Ax \rangle_{\mathcal{V}} + \left\langle \delta x, \gamma^\dagger y \right\rangle_{\partial \mathcal{Y}} - \left\langle \gamma^\dagger y, \gamma x \right\rangle_{\partial \mathcal{X}}, \quad \forall x \in \mathcal{X}_A, \forall y \in \mathcal{Y}_{A^\dagger} \tag{5.52}$$

in which

$$\begin{aligned} \Gamma : \mathcal{X}_A \times \mathcal{Y}_{A^\dagger} &\to \mathbb{R} \\ (x, y) &\mapsto \Gamma(x, y) = \left\langle \delta x, \gamma^\dagger y \right\rangle_{\partial \mathcal{Y}} - \left\langle \gamma^\dagger y, \gamma x \right\rangle_{\partial \mathcal{X}} \end{aligned} \tag{5.53}$$

is called the bilinear adjoint of A.

Chapter 6

Fourier Analysis

6.1 Fourier Series

In signals analysis and solving several model partial differential equations with the method of separation of variables, a natural question is whether a function can be represented by a trigonometric series. Indeed, J. Fourier used extensively the Fourier series in the study of the heat conduction, and he published his results in his famous work *Théorie Analytique de la Chaleur* in 1822.

6.1.1 Real Form

Let $f \in L^1(-\pi, \pi)$. Then its Fourier series is defined by

$$F(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n \cos(nx) + b_n \sin(nx)],$$

where the Fourier coefficients are defined by

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(nx) \, \mathrm{d}\, x, \quad n \geq 0,$$
$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(nx) \, \mathrm{d}\, x, \quad n \geq 1.$$

Generally, if $f \in L^1(-T/2, T/2)$, then its Fourier series is

$$F(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left[a_n \cos \frac{2n\pi x}{T} + b_n \sin \frac{2n\pi x}{T} \right], \tag{6.1}$$

$$a_n = \frac{2}{T} \int_{-T/2}^{T/2} f(x) \cos \frac{2n\pi x}{T} \, \mathrm{d}\, x, \quad n \geq 0, \tag{6.2}$$

$$b_n = \frac{2}{T} \int_{-T/2}^{T/2} f(x) \sin \frac{2n\pi x}{T} \, \mathrm{d}\, x, \quad n \geq 1. \tag{6.3}$$

6.1.2 Complex Form

By the Euler identity

$$\mathrm{e}^{\mathrm{j}\theta} = \cos\theta + \mathrm{j} \sin\theta,$$

we obtain

$$\cos\theta = \frac{\mathrm{e}^{\mathrm{j}\theta} + \mathrm{e}^{-\mathrm{j}\theta}}{2}, \quad \sin\theta = \frac{\mathrm{e}^{\mathrm{j}\theta} - \mathrm{e}^{-\mathrm{j}\theta}}{2\mathrm{j}}.$$

Ussing these formulas, we can rewrite $F(x)$ in the form

$$\begin{aligned} F(x) &= \sum_{n=-\infty}^{\infty} c_n \mathrm{e}^{\mathrm{j}\frac{2n\pi}{T}x} \\ &= \sum_{n=-\infty}^{\infty} c_n \mathrm{e}^{\mathrm{j}n\omega x}, \quad \omega = \frac{2\pi}{T}, \end{aligned} \tag{6.4}$$

where the Fourier coefficients are

$$c_n = \frac{1}{T}\int_{-T/2}^{T/2} f(x)\mathrm{e}^{-\mathrm{j}n\omega x}\,\mathrm{d}\,x, \quad \forall n \in \mathbb{Z}. \tag{6.5}$$

This is called the complex form of the Fourier series of $f \in L^1(-\pi,\pi)$. When f is real-valued, usually the real form of the Fourier series is used, and the corresponding Fourier coefficients are real. Nevertheless, for any L^1 function, real or complex valued, both forms of the Fourier series can be used. Obviously, we have the relaitons

$$a_n = c_n + c_{-n}, \quad b_n = \mathrm{j}(c_n - c_{-n}), \tag{6.6}$$

$$c_n = a_0 \qquad\qquad = 2c_0, \quad a_n = 2\Re(c_n), \quad b_n = -2\Im(c_n), \quad n = 1, 2, \cdots. \tag{6.7}$$

6.1.3 Sine and Cosine Series

Let $f \in L^1(T/2, -T/2)$ be an odd function, i.e., $f(-x) = -f(x)$ for $x \in [-T/2, T/2]$. Then its Fourier series reduces to a sine series:

$$F(x) = \sum_{n=0}^{\infty} b_n \sin(n\omega x), \quad \omega = \frac{2\pi}{T}, \tag{6.8}$$

$$b_n = \frac{4}{T}\int_{-T/2}^{T/2} f(x)\sin(n\omega x)\,\mathrm{d}\,x, \quad n \geq 1 \tag{6.9}$$

Similarly, suppose $f \in L^1(T/2, -T/2)$ be an even function, i.e., $f(-x) = f(x)$ for $x \in [-T/2, T/2]$. Then its Fourier series reduces to a cosine series:

$$F(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(n\omega x), \quad \omega = \frac{2\pi}{T}, \tag{6.10}$$

$$a_n = \frac{4}{T}\int_{-T/2}^{T/2} f(x)\cos(n\omega x)\,\mathrm{d}\,x, \quad n \geq 1 \tag{6.11}$$

Given a function $f \in L^1(0,\ell)$, we can develop a sine series for it. This is achieved as follows. First, we extend f to an odd fnction on $[-\ell, \ell]$:

$$f_{odd}(x) = \begin{cases} f(x), & 0 \leq x \leq \ell \\ -f(-x), & -\ell \leq x < 0. \end{cases}$$

Strictly speaking, f_{odd} is an odd function only if $f(0) = 0$. Nevertheless, even without this property, the Fourier series of f is a sine series since the coefficients of the Fourier series are computed from integrals and do not depend on the funciton value at any particular point (a set with null measure). Then, we use the sine series of f_{odd} to be that of f:

$$F(x) = \sum_{n=1}^{\infty} b_n \sin\frac{2n\pi x}{\ell}, \quad b_n = \frac{2}{\ell}\int_0^{\ell} f(x)\sin\frac{2n\pi x}{\ell}\,\mathrm{d}\,x, \quad n \geq 1.$$

We can develop a cosine series for the function $f \in L^1(0,\ell)$ as well. First, we extend f to an even fnction on $[-\ell, \ell]$:

$$f_{even}(x) = \begin{cases} f(x), & 0 \leq x \leq \ell \\ f(-x), & -\ell \leq x < 0. \end{cases}$$

Then, we use the cosine series of f_{even} to be that of f:

$$F(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\frac{2n\pi x}{\ell}, \quad a_n = \frac{2}{\ell}\int_0^{\ell} f(x)\cos\frac{2n\pi x}{\ell}\,\mathrm{d}\,x, \quad n \geq 0.$$

6.1.4 Convergence

Convergence in L^1

Theorem 153. *Assume f is a piecewise continuous, T-periodic function. Let $x \in [-T/2, T/2]$ (or $x \in \mathbb{R}$ due to the periodicity of f and its Fourier series F) be a point where the two one-sided derivatives $f'(x-)$ and $f'(x+)$ exist. Then*

$$F(x) = \frac{1}{2}[f(x-) + f(x+)].$$

In particular, if f is continuous at x, then

$$F(x) = f(x),$$

that is,

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\frac{2n\pi x}{T} + b_n \sin\frac{2n\pi x}{T}$$

where $\{a_n\}$ and $\{b_n\}$ are defined in (6.2) and (6.3).

We emphasize that even when the Fourier series $F(x)$ for a continuous function is convergent at a point, the limit does not need to be the function at the point.

Convergence in L^2

We always have convergence of the Fourier series in L^2 norm. For $f \in L^2(-T/2, T/2)$, let

$$S_m^f = \frac{a_0}{2} + \sum_{k=1}^{m} a_k \cos\frac{k2\pi x}{T} + b_k \sin\frac{k2\pi x}{T},$$

where a_k and b_k are given by (6.2) and (6.3), then

$$\left\|f - S_m^f\right\|_{L^2}^2 = \int_{-T/2}^{T/2} \left[f(x) - S_m^f(x)\right]^2 \mathrm{d}\,x \to 0, \quad \text{as} \quad m \to \infty.$$

Turn now to the issue of convergence in a general L^p-norm, we define the partial sum sequence S_m^f sequence with the coefficients $a_0, a_1, \cdots, a_m$ and $b_1, \cdots, b_m$. We have the following result.

Theorem 154. *Let $1 \leq p < \infty$. Then*

$$\left\|S_m^f - f\right\|_{L^p(-T/2,T/2)} \to 0 \quad \textit{as} \quad m \to \infty, \quad \forall f \in L^p(-T/2, T/2) \tag{6.12}$$

if and only if there exists a constant $C_p > 0$ such that

$$\left\|S_m^f\right\|_{L^p(-T/2,T/2)} \leq C_p \left\|f\right\|_{L^p(-T/2,T/2)}, \quad \forall m \geq 1, \forall f \in L^p(-T/2, T/2). \tag{6.13}$$

It can be shown that for $1 < p < \infty$, (6.13) holds, and so we have the convergence of the Fourier series in $L^p(-T/2, T/2)$ for any $L^p(-T/2, T/2)$ function. In particular, it is easy to verify (6.13) in the case $p = 2$. On the other hand, (6.13) does not hold for $p = 1$. Note that a consequence of the L^2-norm convergence of the Fourier series of $f \in L^2(-T/2, T/2)$ is the Parseval equality:

$$\begin{aligned} \frac{1}{T}\left\|f\right\|_{L^2(-T/2,T/2)}^2 &= \sum_{n=-\infty}^{\infty} |c_n|^2 = \frac{|a_0|^2}{4} + \frac{1}{2}\sum_{n=1}^{\infty}(|a_n|^2 + |b_n|^2), \\ \left\|f\right\|_{L^2(-T/2,T/2)}^2 &= \frac{1}{T}\int_{-T/2}^{T/2} |f(x)|^2 \,\mathrm{d}\,x \end{aligned} \tag{6.14}$$

6.2 Fourier Transform

6.2.1 Three Kinds of Equivalent Definitions

The Fourier transform can be viewed as a continuous form of the Fourier series. To introduce the Fourier transform, we consider the Fourier series of a function on the interval $[-T/2, T/2]$ and let $T \to \infty$. More precisely, let f be a smooth function with period 2ℓ. Then by the pointwise convergence theorem (the complex version), we have

$$f(x) = \sum_{n=-\infty}^{\infty} c_n \mathrm{e}^{\mathrm{j}2n\pi x/T},$$

where

$$c_n = \frac{1}{T}\int_{-T/2}^{T/2} f(t)\mathrm{e}^{-\mathrm{j}n2\pi t/T}\,\mathrm{d}\,t, \quad -\infty < n < \infty.$$

Thus,

$$f(x) = \sum_{n=-\infty}^{\infty}\left[\frac{1}{T}\int_{-T/2}^{T/2} f(t)\mathrm{e}^{-\mathrm{j}2n\pi t/T}\,\mathrm{d}\,t\right]\mathrm{e}^{\mathrm{j}2n\pi x/T}.$$

Let $\xi_n = \frac{2n\pi}{T}$, $\Delta\xi = \frac{2\pi}{T}$, and define

$$F_T(\xi) = \frac{1}{2\pi}\int_{-T/2}^{T/2} f(t)\mathrm{e}^{-\mathrm{j}\xi t}\,\mathrm{d}\,t.$$

Then

$$f(x) = \sum_{n=-\infty}^{\infty} F_T(\xi_n)\mathrm{e}^{\mathrm{j}\xi_n x}\Delta\xi.$$

For large T, this summation can be viewed as a Riemann sum. Taking the limit $T \to \infty$ and noting that $F_T(\xi)$ formally approaches

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)\mathrm{e}^{-\mathrm{j}\xi t}\,\mathrm{d}\,t$$

we would expect the identity

$$\begin{aligned} f(x) &\overset{①}{=} \int_{-\infty}^{\infty}\left[\frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)\mathrm{e}^{-\mathrm{j}\xi t}\,\mathrm{d}\,t\right]\mathrm{e}^{\mathrm{j}\xi x}\,\mathrm{d}\,\xi \\ &\overset{②}{=} \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left[\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(t)\mathrm{e}^{-\mathrm{j}\xi t}\,\mathrm{d}\,t\right]\mathrm{e}^{\mathrm{j}\xi x}\,\mathrm{d}\,\xi. \end{aligned}$$

It is thus natural to define the Fourier transformation of function g with the help of the previous identiy.

Unfortunately, there are three equivalent definitions for Fourier transformation in different references and/or science/technology fields.

- First definition: summetric form with the two coefficients $\frac{1}{\sqrt{2\pi}}$.

$$\boxed{\begin{aligned} G(\xi) = [\mathcal{F}(g)](x) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} g(x)\mathrm{e}^{-\mathrm{j}\xi x}\,\mathrm{d}\,x, \\ g(x) = [\mathcal{F}^{-1}(G)](\xi) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \hat{g}(\xi)\mathrm{e}^{\mathrm{j}\xi x}\,\mathrm{d}\,\xi. \end{aligned}} \tag{6.15}$$

- Second definition: un-summetric form with one coefficient $\frac{1}{2\pi}$.

$$\boxed{\begin{aligned} G(\xi) = [\mathcal{F}(g)](\xi) &= \int_{-\infty}^{\infty} g(x)\mathrm{e}^{-\mathrm{j}\xi x}\,\mathrm{d}\,x, \\ g(x) = [\mathcal{F}^{-1}(G)](x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{g}(\xi)\mathrm{e}^{\mathrm{j}\xi x}\,\mathrm{d}\,\xi \end{aligned}} \tag{6.16}$$

- Third definiton: symmetric form without any coefficient

$$\boxed{\begin{aligned} G(\xi) &= [\mathcal{F}(g)](\xi) = \int_{-\infty}^{\infty} g(x)\mathrm{e}^{-\mathrm{j}2\pi\xi x}\,\mathrm{d}\,x, \\ g(x) &= [\mathcal{F}^{-1}(G)](x) = \int_{-\infty}^{\infty} \hat{g}(\xi)\mathrm{e}^{\mathrm{j}2\pi\xi x}\,\mathrm{d}\,\xi \end{aligned}} \tag{6.17}$$

Furthermore, the $G(\cdot)$ may be denoted by $\hat{g}(\cdot)$, $[\mathcal{F}(g)](\cdot)$ is frequently denoted by $\mathcal{F}[g(\cdot)]$ or $\mathcal{F}\{g(\cdot)\}$, and $[\mathcal{F}^{-1}(G)](\cdot)$ is frequently denoted by $\mathcal{F}^{-1}[G(\cdot)]$ or $\mathcal{F}^{-1}\{G(\cdot)\}$.

The formal variables x-ξ have the following possible forms:

- (physics)Displacement-WaveNumber: x-k
- (probability theory) Fourier transform of probability density function: x-ω
- (signal analysis) Time-AngularFrequency: t-ω
- (signal analysis) Time-Frequency: t-f

6.2.2 Fourier Transform in $\mathbb{R}^n$ and $\mathcal{S}(\mathbb{R}^n)$

We will treat the Fourier transform over the general n-dimensional space $\mathbb{R}^n$. Extending (6.15) to the multi-variable case, we use the formula

$$[\mathcal{F}(g)](\boldsymbol{\xi}) = \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} g(\boldsymbol{x})\mathrm{e}^{-\mathrm{j}\boldsymbol{x}\cdot\boldsymbol{\xi}}\,\mathrm{d}\,\boldsymbol{x} \tag{6.18}$$

to define the Fourier transform of $g \in L^1(\mathbb{R}^n)$. For convenience, we will usually use $\hat{g}$ to denote the Fourier transform of g:

$$\hat{g} \triangleq \mathcal{F}(g).$$

It is easily seen that $\mathcal{F}$ is a linear and bounded from $L^1(\mathbb{R}^n)$ to $L^\infty(\mathbb{R}^n)$:

$$\begin{aligned} \mathcal{F}(\alpha_1 g_1 + \alpha_2 g_2) &= \alpha_1\,\mathcal{F}(g_1) + \alpha_2\,\mathcal{F}(g_2), \quad \forall g_1, g_2 \in L^1(\mathbb{R}^n), \alpha_1, \alpha_2 \in \mathbb{C}, \\ \|\mathcal{F}(g)\|_{L^\infty(\mathbb{R}^n)} &\le (2\pi)^{-n/2}\,\|g\|_{L^1(\mathbb{R}^n)}, \quad \forall g \in L^1(\mathbb{R}^n). \end{aligned}$$

Applying he Lebesgue dominated convergence theorem, we see that $\mathcal{F}(g) \in C(\mathbb{R}^n)$. Moreover, by Riemann-Lebesgue Lemma, we have

$$\hat{g}(\boldsymbol{\xi}) \to 0 \quad \text{as} \quad \|\boldsymbol{\xi}\| \to \infty.$$

When $\hat{g} \in L^1(\mathbb{R}^n)$, the Fourier inversion formula holds:

$$g(\boldsymbol{x}) \overset{a.e.}{=} (2\pi)^{-n/2} \int_{\mathbb{R}^n} \hat{g}(\boldsymbol{\xi})\mathrm{e}^{\mathrm{j}\boldsymbol{x}\cdot\boldsymbol{\xi}}\,\mathrm{d}\,\boldsymbol{\xi}, \quad \boldsymbol{x} \in \mathbb{R}^n. \tag{6.19}$$

The next step in the development of the theory would be to extend the definition of the Fourier transform from $L^1(\mathbb{R}^n)$ to $L^2(\mathbb{R}^n)$. Such an extension is achieved by a density argument based on the density of the space $L^1(\mathbb{R}^n) \cap \text{Ł}^2(\mathbb{R}^n)$ in $L^2(\mathbb{R}^n)$ and the identity

$$\|\hat{g}\|_{L^2(\mathbb{R}^n)} = \|g\|_{L^2(\mathbb{R}^n)}, \quad g \in L^1(\mathbb{R}^n) \cap L^2(\mathbb{R}^n). \tag{6.20}$$

Definition 155 (Schwartz Space)**.** *The space of test functions of rapid decay, known as the Schwartz space, $\mathcal{S}(\mathbb{R}^n)$, consists of smooth functions $\phi \in C^\infty(\mathbb{R}^n)$ such that for any multi-indices α and β,*

$$\boldsymbol{x}^\beta\,\mathrm{D}^\alpha\,\phi(\boldsymbol{x}) \to \qquad \text{as} \qquad \|\boldsymbol{x}\| \to \infty.$$

Given $\{\phi_1, \phi_2, \cdots\} \subset \mathcal{S}(\mathbb{R}^n)$ and $\phi \in \mathcal{S}(\mathbb{R}^n)$, we say ϕ_n converges to ϕ in $\mathcal{S}(\mathbb{R}^n)$ if for any multi-indices α and β,

$$\lim_{n\to\infty} \max_{\boldsymbol{x}\in\mathbb{R}^n} \left|\boldsymbol{x}^\beta\,\mathrm{D}^\alpha[\phi(\boldsymbol{x}) - \phi_n(\boldsymbol{x})]\right| = 0.$$

Recall that $C_0^\infty(\mathcal{S}(\mathbb{R}^n))$ denotes the space of all functions from $C^\infty(\mathbb{R}^n)$ that have compact support. Notice that algebraically, $C_0^\infty(\mathbb{R}^n) \subset \mathcal{S}(\mathbb{R}^n)$, but not conversely. For example, the function $\mathrm{e}^{-\|\boldsymbol{x}\|^2} \in \mathcal{S}(\mathbb{R}^n)$ but $\mathrm{e}^{-\|\boldsymbol{x}\|^2} \notin C_0^\infty(\mathbb{R}^n)$.

For any $g \in \mathcal{S}(\mathbb{R}^n)$, we use the formula (6.18) to define its Fourier Transform. We list below some properties of the Fourier transform:

❶ $\mathcal{F}$ is a linear operator.

$$\mathcal{F}(\alpha_1 g_1 + \alpha_2 g_2) = \alpha_1 \mathcal{F}(g_1) + \alpha_2 \mathcal{F}(g_2) \tag{6.21}$$

❷ Relation of norms

$$\|\mathcal{F}(g)\|_{L^\infty(\mathbb{R}^n)} \leq (2\pi)^{-n/2} \|g\|_{L^1(\mathbb{R}^n)} \tag{6.22}$$

❸ Differential operator is equivalent to $\mathrm{j}\boldsymbol{\xi}$.

$$[\mathcal{F}(\mathrm{D}^\alpha g)](\boldsymbol{\xi}) = (\mathrm{j}\boldsymbol{\xi})^\alpha [\mathcal{F}(g)](\boldsymbol{\xi}) \tag{6.23}$$

❹ Power and differential operator

$$\mathcal{F}[\boldsymbol{x}^\alpha g(\boldsymbol{x})] = \mathrm{j}^{|\alpha|} \mathrm{D}^\alpha \mathcal{F}[g(\boldsymbol{x})] \tag{6.24}$$

There are three crucial properties for the extension of the definition of the Fourier transform.

❺ $\mathcal{F}$ is continuous from $\mathcal{S}(\mathbb{R}^n)$ to $\mathcal{S}(\mathbb{R}^n)$.

❻ Duality property in $\mathcal{S}(\mathbb{R}^n)$.

$$\int_{\mathbb{R}^n} g(\boldsymbol{x}) \hat{h}(\boldsymbol{x}) \,\mathrm{d}\,\boldsymbol{x} = \int_{\mathbb{R}^n} \hat{g}(\boldsymbol{x}) h(\boldsymbol{x}) \,\mathrm{d}\,\boldsymbol{x}, \qquad g, h \in \mathcal{S}(\mathbb{R}^n). \tag{6.25}$$

Actually, this idendity can be proved by an application of the Fubini theorem on the function $g(\boldsymbol{x})h(\boldsymbol{y}) \in L^1(\mathbb{R}^n \times \mathbb{R}^n)$.

$$\begin{aligned}
\int_{\mathbb{R}^n} g(\boldsymbol{x}) \hat{h}(\boldsymbol{x}) \,\mathrm{d}\,\boldsymbol{x} &= \int_{\mathbb{R}^n} g(\boldsymbol{x}) (2\pi)^{-n/2} \int_{\mathbb{R}^n} h(\boldsymbol{y}) \mathrm{e}^{-\mathrm{j}\boldsymbol{x}\cdot\boldsymbol{y}} \,\mathrm{d}\,\boldsymbol{y} \,\mathrm{d}\,\boldsymbol{x} \\
&= (2\pi)^{-n/2} \int_{\mathbb{R}^n \times \mathbb{R}^n} g(\boldsymbol{x}) h(\boldsymbol{y}) \mathrm{e}^{-\mathrm{j}\boldsymbol{x}\cdot\boldsymbol{y}} \,\mathrm{d}\,\boldsymbol{x} \,\mathrm{d}\,\boldsymbol{y} \\
&= \int_{\mathbb{R}^n} \left[\int_{\mathbb{R}^n} (2\pi)^{-n/2} g(\boldsymbol{x}) \mathrm{e}^{-\mathrm{j}\boldsymbol{x}\cdot\boldsymbol{y}} \,\mathrm{d}\,\boldsymbol{x} \right] h(\boldsymbol{y}) \,\mathrm{d}\,\boldsymbol{y} \\
&= \int_{\mathbb{R}^n} \hat{g}(\boldsymbol{x}) h(\boldsymbol{x}) \,\mathrm{d}\,\boldsymbol{x}
\end{aligned}$$

❼ Inverse of $\mathcal{F}$ holds for every point if the function $g \in \mathcal{S}(\mathbb{R}^n)$.

$$g(\boldsymbol{x}) = (2\pi)^{-n/2} \int_{\mathbb{R}^n} \hat{g}(\boldsymbol{\xi}) \mathrm{e}^{\mathrm{j}\boldsymbol{x}\cdot\boldsymbol{\xi}} \,\mathrm{d}\,\boldsymbol{\xi}, \qquad g \in \mathcal{S}(\mathbb{R}^n). \tag{6.26}$$

Note that $\overset{a.e.}{=}$ is replaced by $=$.

6.2.3 Fourier Transform in $\mathcal{S}^*(\mathbb{R}^n)$

We will extend the definition of the Fourier transform to a much broader class of functions, the space of tempered distributions.

Definition 156. *The space of tempered distributions, $\mathcal{S}^*(\mathbb{R}^(\,)n)$, is the space of all the continuous linear functionals on $\mathcal{S}(\mathbb{R}^n)$.*

Note that a linear functional T on $\mathcal{S}(\mathbb{R}^n)$ is a tempered distribution if and only if

$$\phi_n \to \phi \quad \text{in} \quad \mathcal{S}(\mathbb{R}^n) \Longrightarrow T(\phi_n) = \langle T, \phi_n \rangle \to T(\phi) = \langle T, \phi \rangle$$

In the following, we will only consider those tempered distributions that are generated by functions. Then the action of T on ϕ will be written in the form of a duality pairing:

$$T(\phi) = \langle T, \phi \rangle .$$

As an example, any $g \in L^p(\mathbb{R}^n), 1 \le p \le \infty$, generates a tempered disribution

$$\mathcal{S}(\mathbb{R}^n) \ni \phi \mapsto \langle g, \phi \rangle = \int_{\mathbb{R}^n} g(\boldsymbol{x})\phi(\boldsymbol{x}) \, \mathrm{d}\, \boldsymbol{x}. \tag{6.27}$$

In this sense, $L^p(\mathbb{R}^n) \subset \mathcal{S}^*(\mathbb{R}^n)$.

Recalling the identity (6.25), we now define the Fourier transform on $tSn(n)$.

Definition 157. *Let $g \in \mathcal{S}^*(\mathbb{R}^n)$. Then its Fourier transform $\mathcal{F}\, g = \hat{g} \in \mathcal{S}^*(\mathbb{R}^g)$ is defined by the formula*

$$\boxed{\langle \hat{g}, \phi \rangle = \left\langle g, \hat{\phi} \right\rangle, \quad \phi \in \mathcal{S}(\mathbb{R}^n)} \tag{6.28}$$

It is left as an exercise to shown that $\hat{g}$ defined by (6.28) belongs to the space $\mathcal{S}^*(\mathbb{R}^n)$. Moreoever, when $g \in L^1(\mathbb{R}^n)$, the Fourier transform defined by Definition 157 coincides with the one given in (6.18). This can be verified by applying Fubini's theorem.

Notice that Definition 157 defines the Fourier transform for any $L^p(\mathbb{R}^n)$ function. We mainly use the Fourier transform on $L^2(\mathbb{R}^n)$ related spaces. It can be shown that for $g \in L^2(\mathbb{R}^n)$, its Fourier transform

$$\hat{f}(\boldsymbol{\xi}) = \lim_{R\to\infty} (2\pi)^{-n/2} \int_{\|\boldsymbol{x}\|<R} g(\boldsymbol{x}) \mathrm{e}^{-\mathrm{j}\boldsymbol{x}\cdot\boldsymbol{\xi}} \, \mathrm{d}\, \boldsymbol{x} \quad \text{in} \quad L^2(\mathbb{R}^n). \tag{6.29}$$

Also, we have the inversion formula

$$g(\boldsymbol{x}) = \lim_{R\to\infty} (2\pi)^{-n/2} \int_{\|\xi\|<R} \hat{g}(\boldsymbol{\xi}) \mathrm{e}^{\mathrm{j}\boldsymbol{x}\cdot\boldsymbol{\xi}} \, \mathrm{d}\, \boldsymbol{\xi} \quad \text{in} \quad L^2(\mathbb{R}^n). \tag{6.30}$$

We will simply write

$$\begin{aligned} \hat{g}(\boldsymbol{\xi}) &= (2\pi)^{-n/2} \int_{\mathbb{R}^n} g(\boldsymbol{x}) \mathrm{e}^{-\mathrm{j}\boldsymbol{x}\cdot\boldsymbol{\xi}} \, \mathrm{d}\, \boldsymbol{x}, \\ g(\boldsymbol{x}) &= (2\pi)^{-n/2} \int_{\mathbb{R}^n} \hat{g}(\boldsymbol{x}) \mathrm{e}^{\mathrm{j}\boldsymbol{x}\cdot\boldsymbol{\xi}} \, \mathrm{d}\, \boldsymbol{\xi}, \end{aligned} \tag{6.31}$$

enen when we only assume $g \in L^2(\mathbb{R}^n)$. Most properties of the Fourier transform on $\mathcal{S}(\mathbb{R}^n)$ carry over to that on $L^2(\mathbb{R}^n)$. For example, we still have the formulas (6.23) and (6.24), which play key roles in applying the Fourier transform in the study of differential equations, both ODEs and PDEs.

We now prove (6.23).

- Assume $g, \mathrm{D}^\alpha\, g \in L^2(\mathbb{R}^n)$. Then for any $\phi \in \mathcal{S}(\mathbb{R}^n)$, by Definition 157,

$$\langle \mathcal{F}(\mathrm{D}^\alpha\, g), \phi \rangle = \langle \mathrm{D}^\alpha\, g, \mathcal{F}(\phi) \rangle .$$

- Performing an integration by part,

$$\langle \mathcal{F}(\mathrm{D}^\alpha\, g), \phi \rangle = (-1)^{|\alpha|} \langle g, \mathrm{D}^\alpha\, \mathcal{F}(\phi) \rangle .$$

- For $\phi \in \mathcal{S}(\mathbb{R}^n)$, we can use (6.24). Then,

$$\begin{aligned}\langle \mathcal{F}(\mathrm{D}^\alpha g), \phi\rangle &= \int_{\mathbb{R}^n} g(\boldsymbol{x})\, \mathcal{F}[(\mathrm{j}\boldsymbol{\xi})^\alpha \phi(\boldsymbol{\xi})](\boldsymbol{x})\, \mathrm{d}\,\boldsymbol{x} \\ &= \int_{\mathbb{R}^n} [\mathcal{F}(g)(\boldsymbol{\xi})][(\mathrm{j}\boldsymbol{\xi})^\alpha \phi(\boldsymbol{\xi})]\, \mathrm{d}\,\boldsymbol{\xi}\end{aligned}$$

  where the second equation comes from (6.25).

Finally, we quote the Plancherel formula, which is the analogue for Fourier transform of Parseval's indentity for the Fourier series.

Theorem 158. *The Fourier transform operator $\mathcal{F} : L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ is an isometry, i.e.,*

$$\|\mathcal{F}(g)\|_{L^2(\mathbb{R}^n)} = \|g\|_{L^2(\mathbb{R}^n)}, \qquad \forall g \in L^2(\mathbb{R}^n). \tag{6.32}$$

The Plancherel identity (6.32) is equivalent to the identity

$$\begin{aligned}\langle g|h\rangle = \langle \mathcal{F}(g)|\, \mathcal{F}(h)\rangle, \quad g, h \in L^2(\mathbb{R}^n), \\ \int_{\mathbb{R}^n} \overline{g(\boldsymbol{x})} h(\boldsymbol{x})\, \mathrm{d}\,\boldsymbol{x} = \int_{\mathbb{R}^n} \overline{\mathcal{F}(g)(\boldsymbol{\xi})}\, \mathcal{F}(h)(\boldsymbol{\xi})\, \mathrm{d}\,\boldsymbol{\xi}, \quad g, h \in L^2(\mathbb{R}^n).\end{aligned} \tag{6.33}$$

Proof:

- Derive and use the identity

$$\int_{\mathbb{R}^n} \overline{g(\boldsymbol{x})} h(\boldsymbol{x})\, \mathrm{d}\,\boldsymbol{x} = \frac{1}{4}\left[\|g+h\|^2_{L^2(\mathbb{R}^n)} - \|g-h\|^2_{L^2(\mathbb{R}^n)}\right] + \frac{1}{4\mathrm{j}}\left[\|g-\mathrm{j}h\|^2_{L^2(\mathbb{R}^n)} - \|g+\mathrm{j}h\|^2_{L^2(\mathbb{R}^n)}\right]$$

  for any $g, h \in L^2(\mathbb{R}^n)$.

6.2.4 Dirac δ-function and its Fourier Transform

The generalized function $\delta(\boldsymbol{x})$ is determined as follows:

$$\begin{aligned}\langle \mathcal{F}(\delta), \phi\rangle = \langle \delta, \mathcal{F}(\phi)\rangle &= \int_{\mathbb{R}^n} \delta(\boldsymbol{x})\, \mathcal{F}(\phi)(\boldsymbol{x})\, \mathrm{d}\,\boldsymbol{x} \\ &= \mathcal{F}(\phi)(\mathbf{0}) = \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \phi(\boldsymbol{x}) \mathrm{e}^{-\mathrm{j}\boldsymbol{x}\cdot\mathbf{0}}\, \mathrm{d}\,\boldsymbol{x} \\ &= \int_{\mathbb{R}^n} (2\pi)^{-n/2} \cdot \phi(\boldsymbol{x})\, \mathrm{d}\,\boldsymbol{x} \\ &= \left\langle (2\pi)^{-n/2}, \mathcal{F}(\phi)\right\rangle, \qquad \forall \phi \in \mathcal{S}(\mathbb{R}^n).\end{aligned}$$

Therefore,

$$\mathcal{F}[\delta(\boldsymbol{x})] = (2\pi)^{-n/2}. \tag{6.34}$$

In practice, this process may be simplified as

$$\mathcal{F}(\delta)(\boldsymbol{\xi}) = (2\pi)^{-n/2} \int_{\mathbb{R}^n} \delta(\boldsymbol{x}) \mathrm{e}^{-\mathrm{j}\boldsymbol{x}\cdot\boldsymbol{\xi}}\, \mathrm{d}\,\boldsymbol{x} = (2\pi)^{-n/2} \mathrm{e}^{-\mathrm{j}\mathbf{0}\cdot\boldsymbol{\xi}} = (2\pi)^{-n/2}.$$

Thus the inverse transform is given by

$$\delta(\boldsymbol{x}) = (2\pi)^{-n/2} \int_{\mathbb{R}^n} (2\pi)^{-n/2} \mathrm{e}^{\mathrm{j}\boldsymbol{x}\cdot\boldsymbol{\xi}}\, \mathrm{d}\,\boldsymbol{\xi} = (2\pi)^{-n} \int_{\mathbb{R}^n} \mathrm{e}^{\mathrm{j}\boldsymbol{x}\cdot\boldsymbol{\xi}}\, \mathrm{d}\,\boldsymbol{\xi} \tag{6.35}$$

Or equivalently,

$$\boxed{\int_{\mathbb{R}^n} \mathrm{e}^{\pm\mathrm{j}\boldsymbol{x}\cdot\boldsymbol{\xi}}\, \mathrm{d}\,\boldsymbol{\xi} = (2\pi)^n \delta(\boldsymbol{x})} \tag{6.36}$$

since the $\delta(\boldsymbol{x})$ is symmetric w.r.t the origin.

Remark. Note that the first definition for the Fourier transform is used, otherwise the result will different. For example, we have

- With the first definition

$$\hat{g}(\boldsymbol{\xi}) = (2\pi)^{-n/2} \int g(\boldsymbol{x}) \mathrm{e}^{-\mathrm{j}\boldsymbol{x}\cdot\boldsymbol{\xi}} \,\mathrm{d}\,\boldsymbol{x}, \quad g(\boldsymbol{x}) = (2\pi)^{-n/2} \int \hat{g}(\boldsymbol{\xi}) \mathrm{e}^{-\mathrm{j}\boldsymbol{x}\cdot\boldsymbol{\xi}} \,\mathrm{d}\,\boldsymbol{\xi},$$

  we have

$$\hat{\delta} = (2\pi)^{-n/2}, \quad \hat{1} = (2\pi)^{n/2}\delta(\boldsymbol{\xi})$$

- With the second definition

$$\hat{g}(\boldsymbol{\xi}) = \int g(\boldsymbol{x}) \mathrm{e}^{-\mathrm{j}\boldsymbol{x}\cdot\boldsymbol{\xi}} \,\mathrm{d}\,\boldsymbol{x}, \quad g(\boldsymbol{x}) = (2\pi)^{-n} \int \hat{g}(\boldsymbol{\xi}) \mathrm{e}^{-\mathrm{j}\boldsymbol{x}\cdot\boldsymbol{\xi}} \,\mathrm{d}\,\boldsymbol{\xi},$$

  we have

$$\hat{\delta} = 1, \quad \hat{1} = (2\pi)^{n}\delta(\boldsymbol{\xi})$$

- With the third definition

$$\hat{g}(\boldsymbol{\xi}) = \int g(\boldsymbol{x}) \mathrm{e}^{-\mathrm{j}2\pi\boldsymbol{x}\cdot\boldsymbol{\xi}} \,\mathrm{d}\,\boldsymbol{x}, \quad g(\boldsymbol{x}) = \int \hat{g}(\boldsymbol{\xi}) \mathrm{e}^{-\mathrm{j}2\pi\boldsymbol{x}\cdot\boldsymbol{\xi}} \,\mathrm{d}\,\boldsymbol{\xi},$$

  we have

$$\hat{\delta} = 1, \quad \hat{1} = \delta(\boldsymbol{\xi})$$

6.3 Convolution

6.3.1 Definition

Definition 159 (Convolution)**.** *The convolution of two functions on $\mathbb{R}^n$ is defined by the formula*

$$(g * h)(\boldsymbol{x}) = \int_{\mathbb{R}^n} g(\boldsymbol{y}) h(\boldsymbol{x} - \boldsymbol{y}) \,\mathrm{d}\,\boldsymbol{y} \tag{6.37}$$

Usually, $(g * h)(\boldsymbol{x})$ may be denoted by $g * h(\boldsymbol{x})$ or $g(\boldsymbol{x}) * h(\boldsymbol{x})$.

Examples:

(a) Let $f = g = \mathbb{1}_{[0,1]}$. Then

$$\begin{aligned} f * g(x) &= \int_{\mathbb{R}} f(x-t) g(t) \,\mathrm{d}\,t = \int_0^1 \mathbb{1}_{[0,1]}(x-t) \,\mathrm{d}\,t = \mu([0,1] \cap [x-1, x]) \\ &= \begin{cases} 0, & \text{if } x \le 0, \\ x, & \text{if } 0 \le x \le 1, \\ 2-x, & \text{if } 1 \le x \le 2, \\ 0, & \text{if } x \ge 2. \end{cases} \end{aligned}$$

which is the "hat" or "triangle" function.

(b) Take $f \in L^1(\mathbb{R})$ and $g = \frac{1}{2h} \mathbb{1}_{[-h,h]}$ with $h > 0$. Then

$$f * g(x) = \frac{1}{2h} \int_{-h}^{+h} f(x-t) \,\mathrm{d}\,t = \frac{1}{2h} \int_{x-h}^{x+h} f(u) \,\mathrm{d}\,u,$$

which is the average of f on the interval $[x-h, x+h]$.

Lemma 160. *Let g and h be two functions for which $g * h$ exists. Then*

$$\operatorname{supp}(g * h) \subset \overline{\operatorname{supp}(g) + \operatorname{supp}(h)}.$$

6.3.2 Properties

Proposition 161. *For $g, h \in L^1(\mathbb{R})$, we have*

① *$g * h$ is defined almost everywhere and $g * h \in L^1(\mathbb{R})$.*

② *The convolution is a continuous bilinear operator from $L^1(\mathbb{R}) \times L^1(\mathbb{R})$ to $L^1(\mathbb{R})$ with*

$$\|g * h\|_{L^1(\mathbb{R})} \leq \|g\|_{L^1(\mathbb{R})} \cdot \|h\|_{L^1(\mathbb{R})}. \tag{6.38}$$

Proof

- Ad. ①: Since g and h are in $L^1(\mathbb{R})$, Fubini's theorem implies that the function $gh : (y, z) \mapsto g(y)h(z)$ is in $L^1(\mathbb{R}^2)$. By making the change of variables $y = x - t$ and $z = t$, we have

$$\iint_{\mathbb{R}\times\mathbb{R}} g(y)h(z)\,\mathrm{d}\,y\,\mathrm{d}\,z = \iint_{\mathbb{R}\times\mathbb{R}} g(x-t)h(t)\,\mathrm{d}\,x\,\mathrm{d}\,t.$$

  The function $x \mapsto \int_{\mathbb{R}} g(x-t)h(t)\,\mathrm{d}\,t$ is thus defined almost everywhere and belongs to $L^1(\mathbb{R})$, agian by Fubini's theorem.

- Ad. ②: To establish the inequality, we write

$$|g * h(x)| \leq int_{\mathbb{R}} |g(x-t)|\,|h(t)|\,\mathrm{d}\,t = |g| * |g|\,(x).$$

  Thus

$$\begin{aligned}\int_{\mathbb{R}} |g * h(x)|\,\mathrm{d}\,x &\leq \int_{\mathbb{R}} |g|\,|h|\,(x)\,\mathrm{d}\,x = \int_{\mathbb{R}} \mathrm{d}\,x \int_{\mathbb{R}} |g(x-t)|\,|h(t)|\,\mathrm{d}\,t \\ &= \int_{\mathbb{R}} |h| \left[\int_{\mathbb{R}} |g(x-t)|\,\mathrm{d}\,x\right] \mathrm{d}\,t = \|g\|_{L^1(\mathbb{R})} \cdot \|h\|_{L^1(\mathbb{R})}. \qquad \blacksquare\end{aligned}$$

Can the hypothesis of this last result be weakened? If g and h are in $L^1_{\mathrm{loc}}(\mathbb{R})$, the result is false (take $g = h = 1$). However, we have the following result.

Proposition 162. *Assume that $g \in L^1_{\mathrm{loc}}(\mathbb{R})$ and that $h \in L^1(\mathbb{R})$.*

- *If* $\mathrm{supp}(h)$ *is bounded, then $g * h$ exists* æ *and belongs to $L^1_{\mathrm{loc}}(\mathbb{R})$.*
- *If g is bounded, then $g * h$ exists for all x and belongs to $L^\infty(\mathbb{R})$.*

Proof

- h is zero æ outside some interval $[-a, a]$. Take x in a finite interval $[\alpha, \beta]$. For all $t \in [-a, a]$ and all $x \in [\alpha, \beta]$,

$$g(x-t)h(t) = \mathbb{1}_{[\alpha-a,\beta+a]}(x-t)g(x-t)h(t),$$

  and thus

$$g * h(x) = \int_{-a}^{+a} g(x-t)h(t)\,\mathrm{d}\,t = (\mathbb{1}_{[\alpha-a,\beta+a]}\,g) * h(x).$$

  $g * h$ coincides on $[\alpha, \beta]$ with the convolution of two functions in $L^1(\mathbb{R})$, so by Proposition 161 it is defined æ and is integrable. Thus $g * h$ is defined æ and is integrable on allcompact sets.

- If $g \in L^\infty(\mathbb{R})$, then

$$\left|\int_{\mathbb{R}} g(u)h(x-u)\,\mathrm{d}\,u\right| \leq \|g\|_{L^\infty(\mathbb{R})} \int_{\mathbb{R}} |h(x-u)|\,\mathrm{d}\,u = \|g\|_{L^\infty(\mathbb{R})}\,\|h\|_{L^1(\mathbb{R})} \tag{6.39}$$

  for all x, and $\|g * h\|_{L^\infty(\mathbb{R})} \leq \|g\|_{L^\infty(\mathbb{R})} \cdot \|h\|_{L^1(\mathbb{R})}$. $\blacksquare$

Similarly, we have the concolution in $L^p(\mathbb{R}^n)$.

Proposition 163. *Assume that $g \in L^p(\mathbb{R})$ and that $h \in L^q(\mathbb{R})$ (p and q conjugates, i.e., $\frac{1}{p} + \frac{1}{q}$). Then the following hold:*

- *$g * h$ is defined everywhere and is continuous and bouned on $\mathbb{R}$.*
- *$\|g * h\|_{L^\infty(\mathbb{R})} \leq \|g\|_{L^p(\mathbb{R})} \cdot \|h\|_{L^q(\mathbb{R})}$.*

When p, q are not conjugate, we have the following result:

Proposition 164. *If $g \in L^1(\mathbb{R})$ and that $h \in L^2(\mathbb{R})$, then the following hold:*

- *$g * h$ exists almost everywhere.*
- *$g * h \in L^2(\mathbb{R})$ and*

$$\|g * h\|_{L^2(\mathbb{R})} \leq \|g\|_{L^1(\mathbb{R})} \cdot \|h\|_{L^2(\mathbb{R})}.$$

6.3.3 Convolution and Fourier Transform

The Fourier transform is a homomorphism w.r.t the convolution and the usual product, i.e., for $g, h \in \mathcal{S}^*(\mathbb{R}^n)$, we have

$$g * h \in \mathcal{S}^*(\mathbb{R}^n)$$

and

$$\mathcal{F}(g * h) = \mathcal{F}(g) \cdot \mathcal{F}(h). \tag{6.40}$$

Let $h(\cdot) = \delta(\cdot)$, then

- For the first definition of Fourier Transform, with the help of the property

$$\mathcal{F}(\delta(\boldsymbol{x})) = (2\pi)^{-n/2},$$

we have

$$\mathcal{F}(g * \delta) = \mathcal{F}(g) \cdot \mathcal{F}(h) = (2\pi)^{-n/2}\, \mathcal{F}(g).$$

- For the second and third definitions of Fourier Transform, with the help of the property

$$\mathcal{F}(\delta(\boldsymbol{x})) = 1,$$

we have

$$\mathcal{F}(g * \delta) = \mathcal{F}(g) \cdot \mathcal{F}(h) = \mathcal{F}(g).$$

Therefore,

$$g * \delta = g, \quad \forall g$$

which implies that δ is the identity of the convolutional operations.

6.3.4 Uncertainty Principle

We will develop the relation exists between the location of a signal and the localization of its spectrum. We use the following notation

$$\hat{g}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} g(x) \mathrm{e}^{-\mathrm{j}x\xi}\, \mathrm{d}\, x$$
$$g(x) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} g(x) \mathrm{e}^{\mathrm{j}x\xi}\, \mathrm{d}\, \xi$$

Given a function $g : \mathbb{R} \to \mathbb{C}$ such that $g(x), xg(x)$ and $\xi\hat{g}(\xi)$ are in $L^2(\mathbb{R})$, we introduce the following definitions and notation:

- energy dispersion of g in time

$$\langle x^2 \rangle_g = \int_{\mathbb{R}} x^2 \left| g(x) \right|^2 \mathrm{d}\, x$$

- energy dispersion of g in frequency

$$\langle x^2 \rangle_{\hat{g}} = \int_{\mathbb{R}} \xi^2 \left| \hat{g}(\xi) \right|^2 \mathrm{d}\, \xi$$

- energy of g

$$\mathcal{E}_g = \|g\|_{L^2}^2 = \int_{\mathbb{R}} |f(x)|^2 \,\mathrm{d}\, x = \int_{\mathbb{R}} \left| \hat{f}(\xi) \right|^2 \mathrm{d}\, \xi = \|\hat{g}\|_{L^2}^2$$

The value σ_g, defined by

$$\sigma_g = \frac{\sqrt{\langle x^2 \rangle_g}}{\|g\|}$$

is called the *effective duration* of the signal g. $\sigma_{\hat{g}}$, defined by

$$\sigma_{\hat{g}} = \frac{\sqrt{\langle \xi^2 \rangle_{\hat{g}}}}{\|g\|}$$

is called the *effective bandwidth*.

The uncertainty principle is a relation between σ_g and $\sigma_{\hat{g}}$ which says that one cannot arbitrarily localize a signal in both time and frequency. This relation is

$$\sigma_g \cdot \sigma_{\hat{g}} \geq \frac{1}{2}, \tag{6.41}$$

which is the content of the following proposition.

Proposition 165. *Let $g : \mathbb{R} \to \mathbb{C}$ be a function in $C^1(\mathbb{R})$ such that $g(x), xg(x)$ and $\xi \hat{g}(\xi)$ are in $L^2(\mathbb{R})$. Then*

$$\langle x^2 \rangle_g \cdot \langle x^2 \rangle_{\hat{g}} \geq \frac{\|g\|^4}{4} \tag{6.42}$$

Proof

- We assume the following two results

  (i) $\lim\limits_{|x| \to \infty} x \left| f(x) \right|^2 = 0$

  (ii) $\hat{g'}(\xi) = \mathrm{j} \xi \hat{g}(\xi)$

  Thus we can obtain

$$\langle \xi^2 \rangle_{\hat{g}} = \int_{\mathbb{R}} \xi^2 \left| \hat{g}(\xi) \right|^2 \mathrm{d}\, \xi = \int_{\mathbb{R}} \left| \hat{g'}(\xi) \right|^2 \mathrm{d}\, \xi = \int_{\mathbb{R}} |g'(x)|^2 \,\mathrm{d}\, x$$

  On the other hand, $(f \cdot \overline{f})' = f' \cdot \overline{f} + f \cdot \overline{f}'$, and

$$\begin{aligned}
\left| \int_{\mathbb{R}} x[g(x)\overline{g(x)}]' \,\mathrm{d}\, x \right| &= \left| \int_{\mathbb{R}} x[g(x)\overline{g(x)}' + f'(x)\overline{g(x)}] \,\mathrm{d}\, x \right| \\
&\leq \int_{\mathbb{R}} \left| xg'(x)\overline{g(x)} \right| \mathrm{d}\, x + \int_{\mathbb{R}} \left| xg(x)\overline{g(x)}' \right| \mathrm{d}\, x \\
&= \sqrt{\int_{\mathbb{R}} \left| x\overline{g(x)} \right|^2 \mathrm{d}\, x} \cdot \sqrt{\int_{\mathbb{R}} |g'(x)|^2 \,\mathrm{d}\, x} + \sqrt{\int_{\mathbb{R}} |xg(x)|^2 \,\mathrm{d}\, x} \cdot \sqrt{\int_{\mathbb{R}} \left| \overline{g'(x)} \right|^2 \mathrm{d}\, x} \\
&= 2\sqrt{\int_{\mathbb{R}} |xg(x)|^2 \,\mathrm{d}\, x} \cdot \sqrt{\int_{\mathbb{R}} |g'(x)|^2 \,\mathrm{d}\, x} \\
&= 2\sqrt{\int_{\mathbb{R}} |x|^2 \left| g(x) \right|^2 \mathrm{d}\, x} \cdot \sqrt{\int_{\mathbb{R}} |\xi|^2 \left| \hat{g}(\xi) \right|^2 \mathrm{d}\, \xi} \\
&= 2\sqrt{\langle \xi^2 \rangle_{\hat{g}} \cdot \langle x^2 \rangle_g}.
\end{aligned}$$

But

$$\int_{\mathbb{R}} x[g(x)\overline{g(x)}]' \,\mathrm{d}\, x = \int_{\mathbb{R}} x\frac{\mathrm{d}\left[|f(x)|^2\right]}{\mathrm{d}\, x}\,\mathrm{d}\, x = \left[x\,|f(x)|^2\right]_{-\infty}^{\infty} - \int_{\mathbb{R}} |f(x)|^2\,\mathrm{d}\, x = -\,\|g\|_{L^2}^2\,,$$

since $\lim\limits_{|x|\to\infty} x\,|g(x)|^2 = 0$. Thus

$$\|g\|_{L^2}^4 \le 2\langle \xi^2\rangle_{\hat{g}} \cdot \langle x^2\rangle_g,$$

which implies that

$$\sigma_g \cdot \sigma_{\hat{g}} \ge \frac{1}{2}.$$

- Note that if we take the other definitions of Fourier Transform, the const $\frac{1}{2}$ may be replaced by other constants. ■

Proposition 166. *Let the effective bandwidth W be fixed. Then the Gaussian signal*

$$g(t) = \alpha \mathrm{e}^{-(2\pi W)^2 t^2}$$

minimizes the effective duration. In other words, $\sigma_g \cdot \sigma_{\hat{g}} = \frac{1}{2}$ if the function $g(t)$ is a Gaussian function.

6.4 Orthogonal Series

6.4.1 Notations

Assumption (H): Let $\mathcal{X}$ be a Hilbert space over $\mathbb{F} = \mathbb{R}, \mathbb{C}$, and let $\{\phi_0, \phi_1, \cdots\}$ be a finite or contable orthogonal system in $\mathcal{X}$, i.e., by definition

$$\langle \phi_k | \phi_m \rangle = \delta_{km} = \begin{cases} 1, & \text{for } k = m; \\ 0, & \text{for } k \neq m. \end{cases} \tag{6.43}$$

Our goal is to study the convergence of the so-called abstract Fourier series

$$u = \sum_{n=0}^{\infty} \langle \phi_n | u \rangle \phi_n. \tag{6.44}$$

We also set

$$s_m \triangleq \sum_{n=0}^{m} \langle \phi_n | u \rangle \phi_n. \tag{6.45}$$

The numbers $\langle \phi_n | u \rangle$ determined by inner product are called the *Fourier coefficients* of u.

6.4.2 Key Issues

Definition 167. *Assume (H). The finite orthogonormal system $\{\phi_0, \phi_1, \cdots, \phi_N\}$ is called complex in $\mathcal{X}$ iff*

$$u = \sum_{n=0}^{N} \langle \phi_n | u \rangle \phi_n, \qquad \forall \in u \in \mathcal{X}. \tag{6.46}$$

The countable orthonormal system $\{\phi_0, \phi_1, \cdots\}$ is called complete in $\mathcal{X}$ iff the infinite series (6.44) converges for all $u \in \mathcal{X}$, i.e.,

$$u = \lim_{m\to\infty} s_m, \qquad \forall u \in \mathcal{X}.$$

Proposition 168. *The finite orthonormal system $\{u_0, \cdots, u_N\}$ is complete in the Hilbert space $\mathcal{X}$ over $\mathbb{F}$ iff it is a basis of $\mathcal{X}$.*

PROOF

- Let $\{\phi_n\}_{n=0}^N$ be a basis in $\mathfrak{X}$. Then,

$$u = \sum_{n=0}^{N} c_n \phi_n, \quad \forall u \in$$

where the coefficients $c_0, \cdots, c_N \in \mathbb{F}$ depend on u. Using (6.43), we get

$$\langle u_k | u \rangle = \sum_{n=0}^{N} c_n \langle u_k | u_n \rangle = c_k, \qquad k = 0, 1, \cdots, N.$$

This implies (6.46), i.e., $\{\phi_n\}$ is complete.

- Conversely, let $\{\phi_n\}$ be a complete orthonormal system; them $\{\phi_n\}$ is a basis of $\mathfrak{X}$, by (6.46). In this connection, note that $\{\phi_0, \cdots, \phi_N\}$ is linearly independent, since $\sum_{n=0}^{N} c_n \phi_n = 0$ implies that $c_k = 0$ for all k. ■

Corollary 169. *Let $\{\phi_n\}$ be a countable orthonormal system in the Hilbert space over $\mathbb{F}$. Assume that the infinite series*

$$u = \sum_{n=0}^{\infty} c_n \phi_n, \quad \forall n, \ c_n \in \mathbb{F}$$

is convergent for some fixed $u \in \mathfrak{X}$. Then $c_n = \langle \phi_n | u \rangle$ for all n.

Proposition 170. *Assume (H). Let $\boldsymbol{c} = [c_0, c_1, \cdots, c_m]^T$, the unique solution of the optimization problem*

$$\boldsymbol{c}_{opt} = \arg \min_{\boldsymbol{c} \in \mathbb{R}^{(m+1)\times 1}} f(\boldsymbol{c}) = \left\| u - \sum_{k=0}^{m} c_k \phi_k \right\|^2 \tag{6.47}$$

is given through the Fourier coefficients

$$c_k = \langle \phi_k | u \rangle, k = 0, 1, \cdots, m.$$

PROOF

- By (6.43), we have

$$\begin{aligned} f(\boldsymbol{c}) &= \langle u - \sum_{n=0}^{m} c_n \phi_n | u - \sum_{k=0}^{m} c_k \phi_k \rangle \\ &= \langle u | u \rangle - \sum_{n=0}^{m} \overline{c_n} \langle \phi_n | u \rangle - \sum_{k=0}^{m} c_k \langle u | \phi_k \rangle + \sum_{n=0}^{m} |c_n|^2 . \end{aligned}$$

Hence

$$f(\boldsymbol{c}) = \|u\|^2 - \sum_{n=0}^{m} |\langle \phi_n | u \rangle|^2 + \sum_{n=0}^{m} |\langle \phi_n | u \rangle - c_n|^2 . \tag{6.48}$$

The smallest value of f is attained for $c_n = \langle \phi_n | u \rangle, n = 0, \cdots, m$.

In particular, it follows from $s_m = \sum_{n=0}^{m} \langle \phi_n | u \rangle$ that

$$\|u - s_m\|^2 \leq f(\boldsymbol{c}), \quad \forall \boldsymbol{c} \in \mathbb{F}^m, \ \forall m. \tag{6.49}$$

By (6.48),

$$\|u - s_m\|^2 = \|u\|^2 - \sum_{n=0}^{m} |\langle \phi_n | u \rangle|^2 , \quad \forall u \in \mathfrak{X}, \forall m. \tag{6.50}$$

Proposition 171 (Convergence Criterion)**.** *Let $\{\phi_n\}$ be a countable orthonormal system in the Hilbert space $\mathcal{X}$ over $\mathbb{F}$. Then, the series*

$$\sum_{n=0}^{\infty} c_n\phi_n, \quad c_n \in \mathbb{F} \quad \textit{for all } n,$$

is convergent iff the series $\sum_{n=0}^{\infty} |c_n|^2$ is convergent.

It follows from the convergence criterion and the Bessel inequality that for each $u \in \mathcal{X}$ the Fourier series is convergent, i.e., there is some $v \in \mathcal{X}$ such that

$$v = \sum_{n=0}^{\infty} \langle \phi_n | u \rangle \phi_n.$$

However, it is possible that $v \neq u$. But if the orthonormal system $\{\phi_n\}$ is complete, then $v = u$ for all $u \in \mathcal{X}$.

Theorem 172. *Let $\{\phi_n\}$ be a complete orthonormal system in the Hilbert space $\mathcal{X}$ over $\mathbb{F}$. Then, the following two conditions are equivalent:*

① *The system $\{\phi_n\}$ is complete in $\mathcal{X}$.*

② *The linear hull of $\{\phi_n\}$ is dense in $\mathcal{X}$.*

Corollary 173. *Let $\{\phi_n\}$ be a countable complete orthonormal system in the Hilbert space $\mathcal{X}$ over $\mathbb{F}$. Then, the following hold true:*

① *For all $u, v \in \mathcal{X}$,*

$$\langle u|v\rangle = \sum_{n=0}^{\infty} \overline{c_n(u)} c_n(v), \qquad \textit{the Parseval equation,} \tag{6.51}$$

where $v_n(w) = \langle \phi_n | w \rangle$.

② *For all $u \in \mathcal{X}$, the Bessel inequality is replaced with the so-called special Parseval equation*

$$\|u\|^2 = \sum_{n=0}^{\infty} |\langle \phi_n | u \rangle|^2 . \tag{6.52}$$

③ *If $\langle \phi_n | u \rangle = 0$ for all n and fixed $u \in \mathcal{X}$, then $u = 0$.*

6.4.3 Applications to Classic Fourier Series

Recall that the inner product in the Hilbert space $L_2(-\pi, \pi)$ is given through

$$\langle u|v\rangle = \int_{-\pi}^{\pi} u(x)v(x)\,\mathrm{d}\,x.$$

For all $x \in [-\pi, \pi]$, we set

$$\begin{aligned}
\phi_0(x) &\triangleq \frac{1}{\sqrt{2\pi}} \\
\phi_{2m-1}(x) &= \frac{1}{\sqrt{\pi}} \cos mx, \quad m = 1, 2, \cdots \\
\phi_{2m}(x) &= \frac{1}{\sqrt{\pi}} \sin mx, \quad m = 1, 2, \cdots
\end{aligned}$$

Proposition 174. *The set $\{\phi_0, \phi_1, \cdots, \}$ forms a complete orthogonal system in the Hilbert space $L^2(-\pi, \pi)$.*

This proposition telles us that for each $u \in L^2(-\pi, \pi)$ the Fourier series

$$u = \sum_{n=0}^{\infty} c_n \phi_n, \quad c_n \triangleq \langle \phi_n | u \rangle, \tag{6.53}$$

converges in $L^2(-\pi, \pi)$. This is identical to the classic Fourier series

$$u(x) = \frac{1}{2} a_0 + \sum_{k=1}^{\infty} a_k \cos kx + b_k \sin kx, \tag{6.54}$$

where

$$a_k \triangleq \frac{1}{\pi} \int_{-\pi}^{\pi} u(x) \cos kx \, \mathrm{d}\, x$$
$$b_k \triangleq \frac{1}{\pi} \int_{-\pi}^{\pi} u(x) \sin kx \, \mathrm{d}\, x, \quad k = 0, 1, 2, \cdots .$$

In fact, $\langle \phi_{2m} | u \rangle \phi_{2m}(x) = \frac{1}{\pi} \int\limits_{-\pi}^{\pi} u(x) \sin mx \, \mathrm{d}\, x, m = 0, 1, \cdots$ and so on.

Corollary 175. *For each $u \in L^2(-\pi, \pi)$, the classic Fourier series converges in $L^2(-\pi, pi)$, i.e.,*

$$\lim_{m \to \infty} \left\| u(x) - \left[\frac{a_0}{2} + \sum_{k=1}^{m} a_k \cos kx + b_k \sin kx \right] \right\|_{L^2(-\pi, \pi)} = 0.$$

Let $\mathcal{T}$ denote the set of all trigonometric polynomials, i.e., $p \in \mathcal{T}$ iff

$$p(x) \triangleq \sum_{n=0}^{m} \alpha_n \cos nx + \beta_n \sin nx,$$

where $m = 0, 1, \cdots$, and all the coefficients α_n, β_n are real numbers. It follows from the classical addition theorems for $\sin(\cdot)$ and $\cos(\cdot)$ that

$$p, q \in \mathcal{T} \quad \text{implies} \quad pq \in \mathcal{T}.$$

We also set

$$\|f\|_{C[a,b]} \triangleq \max_{x \in [a,b]} |f(x)| \, .Applicationsto$$

Lemma 176. *For each function $f \in C[-\pi, pi]$ with $f(-\pi) = f(\pi)$ and each $\varepsilon > 0$, there exists a function $p \in \mathcal{T}$ such that*

$$\|f - p\|_{C[-\pi, pi]} < \varepsilon.$$

Corollary 177. *The set $\mathcal{T}$ of trigonometric polynomials is dense in $L^2(-\pi, pi)$.*

6.4.4 Applications to Schmidt Orthogonalization Method

Proposition 178. *In each separable Hilbert space $\mathfrak{X}$ over $\mathbb{F}$ with $\mathfrak{X} \neq \{0\}$, there exists a complete orthogonormal system.*

Proof

- By assumption, there exists an at most countable set $\{v_0, v_1, \ldots\}$ that is dense in $\mathfrak{X}$. We may assume that $v_0 \neq 0$. Set
$$\phi_0 \triangleq \frac{v_0}{\|v_0\|}.$$
- Suppose that we have already constructed $\{\phi_0, \cdots, \phi_n\}$ forms an orthonormal system. Then, let
$$w_{n+1} \triangleq v_{n+1} - \sum_{k=0}^{n} \langle \phi_k | v_{n+1} \rangle \phi_k. \tag{6.55}$$

– If $w_{n+1} \neq 0$, then we set

$$\phi_{n+1} \triangleq \frac{w_{n+1}}{\|w_{n+1}\|}. \tag{6.56}$$

Thus, $\langle \phi_m | \phi_{n+1} \rangle = 0$ for $m = 0, \cdots, n$, and $\langle \phi_{n+1} | \phi_{n+1} \rangle = 1$.

– If $w_{n+1} = 0$, then we use v_{n+2}, and so forth. This way we obtain an orthogonormal system ϕ_m.

- By induction, it follows that all the v_m are finite linear combinations of the ϕ_n. Hence the linear hull of $\{\phi_n\}$ is dense in $\mathcal{X}$. If $\{\phi_n\}$ is countable, then it is complete by Theorem 172.
- If $\{\phi_n\}$ is finite, then $\operatorname{span}\{\phi_m\} = \mathcal{X}$, since each finite-dimensional linear subspace of a Hilbert space is closed. Thus $\{\phi_n\}$ is again complete. ■

The procedure of constructing the orthonormal basis proposed in the proof is called the Schmidt orthogonalization method.

Proposition 179. *We assume the following:*

① *Let $\{v_0, v_1, \cdots\}$ be a sequence in the Hilbert space $\mathcal{X}$ over $\mathbb{F}$ such that $v_0, v_1, \cdots, v_m$ are linearly independent for each $m = 0, 1, \cdots$.*

② *Let the linear space* $\operatorname{span} v_0, v_1, \cdots$ *be dense in $\mathcal{X}$.*

③ *Let $\{\phi_0, \phi_1, \cdots\}$ be a countable orthonormal system in $\mathcal{X}$ such that*

$$\phi_{n+1} = \alpha_{n+1} v_{n+1} + \sum_{k=0}^{n} \alpha_k v_k, \quad \alpha_{n+1} > 0, \tag{6.57}$$

$$\phi_0 = \alpha_0 v_0, \quad \alpha_0 > 0, \tag{6.58}$$

for all $n = 0, 1, \cdots$ and appropriate coefficients $\alpha_k \in \mathbb{F}, k = 0, 1, \cdots, n$.

Then, the follow hold true:

❶ *The system $\{\phi_n\}$ is obtained from $\{v_n\}$ by means of the Schmidt orthogonalization method.*

❷ *The system $\{\phi_n\}$ is complete in $\mathcal{X}$.*

Proof

- Ad ①. If follows from (6.57) that $\alpha_0 = \frac{1}{\|v_0\|}$, and hence

$$\phi_0 = \frac{v_0}{\|v_0\|}.$$

Let $n \geq 1$. By (6.58),

$$v_k \in \operatorname{span}\{\phi_0, \cdots, \phi_n\}, \quad k = 0, \cdots, n. \tag{6.59}$$

Thus

$$w_{n+1} = \alpha_{n+1} v_{n+1} + \sum_{m=0}^{n} \beta_m \phi_m,$$

where $\beta_0, \cdots, \beta_n \in \mathbb{F}$ are appropriate coefficients. Using $\langle \phi_k | \phi_{k+1} \rangle = 0$ for $k = 0, \cdots, n$, we get $\beta_k = -\alpha_{n+1} \langle \phi_k | v_{n+1} \rangle$. By (6.55),

$$\phi_{n+1} = \frac{w_{n+1}}{\|w_{n+1}\|}, \quad \text{and} \quad \phi_{n+1} \neq 0.$$

Thus, according to (6.56), ϕ_{n+1} corresponds to the Schmidt orthogonalization.

- Ad ②. Since $\operatorname{span} v_n$ is dense in $\mathcal{X}$, it follows from (6.59) that the set $\operatorname{span} v_n$ is also dense in $\mathcal{X}$. By Theorem 172, ϕ_n is complete. ■.

Corollary 180. *Let $\{v_0, v_1, \cdots\}$ be a sequence in the Hilbert space $\mathcal{X}$ over $\mathbb{F}$. Suppose that f*

$$u \in \mathcal{X}, \forall n, \quad \langle v_n | u \rangle = 0 \Longrightarrow u = 0. \tag{6.60}$$

Then, the set $\operatorname{span}\{v_0, v_1, \cdots\}$ *is dense in $\mathcal{X}$.*

Proof

- Let $S \triangleq \operatorname{span} v_0, v_1, \cdots$. By (6.60),
$$(\overline{S})^{\perp} = \{0\}.$$
- By the orthogonal decompistion theorem (Corollary 108) tells us that $\mathcal{X} = \overline{S}$).

6.5 Orthogonal Polynomials

6.5.1 Inner Product and Orthogonal Polynomials

Definition 181. *A non-negative function $w(x)$ defined on the interval $[a, b]$ (finite or infinite) such that*

- $\int_a^b x^n w(x)\,\mathrm{d}\,x$ *exists and it is finite for $n = 0, 1, 2, \cdots$;*
- *for non-negative function $f(x)$, if $\int_a^b f(x)w(x)\,\mathrm{d}\,x = 0$, then $f(x) \overset{a.e.}{=} 0$ on $[a, b]$.*

is called a weight function.

Actually, we can regard $w(x)\,\mathrm{d}\,x$ as

$$w(x)\,\mathrm{d}\,x = \mu(\mathrm{d}\,x) = \mathrm{d}\,\mu$$

where μ is a measure with finite value, and the integral $\int_a^b g(x)w(x)\,\mathrm{d}\,x$ as Lebesuge integral, i.e.

$$\int_a^b g(x)w(x)\,\mathrm{d}\,x = \int_a^b g(x)\,\mathrm{d}\,\mu$$

Let $f, g \in L^2(a, b)$ and the inner product is defined by

$$\langle f | g \rangle = \int_a^b f(x)g(x)\,\mathrm{d}\,\mu = \int_a^b f(x)g(x)w(x)\,\mathrm{d}\,x. \tag{6.61}$$

If $\langle f | g \rangle = 0$, we say that f and g are perpendicular with weight function $w(x)$, or $f \perp g$ for simplification.

If the function sequence $\{\phi_0, \phi_1, \cdots, \phi_n, \cdots\}$ satisfies the condtion

$$\langle \phi_i | \phi_j \rangle = \int_a^b \phi_i(x)\phi_j(x)\,\mathrm{d}\,\mu = \int_a^b \phi_i(x)\phi_j(x)w(x)\,\mathrm{d}\,x = A_i\delta_{ij}, \tag{6.62}$$

then $\{\phi_k\}$ are called orthogonal systems of functions with weight function $w(x)$ on $[a, b]$. Particularly, if each ϕ_k is a polynomial with degree k such that $\langle \phi_j | \phi_k \rangle = A_j\delta_{jk}$, then $\{\phi_k\}$ are called orthogonal polynomials with weight function $w(x)$.

For the spceified interval $[a, b]$ and weight function $w(x)$, the polunomials $\{1, x, x^2, \cdots\}$ can be orthogonalized with the Schmidt Procedure so as to get orthogonal polynomials $\{\phi_n\}$ as

$$\begin{aligned} &\phi_0(x) = 1, \\ &\phi_n(x) = x^n - \sum_{j=0}^{n} \frac{\langle x^n | \phi_j \rangle}{\langle \phi_j | \phi_j \rangle} \phi_j(x), \quad n = 1, 2, \cdots \end{aligned} \tag{6.63}$$

Furthermore, the orthogonal polynomials obtained have the following properties:

① $\phi_n(x)$ has the form $\phi_n(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$.

② For any n-th order polynomial $p(x)$ can be generated by $\{\phi_i(x)\}_{i=0}^n$, i.e.,

$$p(x) = a_0\phi_0(x) + a_1\phi_1(x) + \cdots + a_n\phi_n(x).$$

③ $\langle\phi_j|\phi_k\rangle = \delta_{jk} = 0$ for $k \neq j$.

④ Iterative relation

$$\phi_{n+1} = (x - \alpha_n)\phi_n(x) - \beta_n\phi_{n-1}(x), \quad n = 0, 1, 2, \cdots \tag{6.64}$$

where $\phi_0(x) = 1, \phi_{-1}(x) = 0$ and

$$\alpha_n = \frac{\langle x\phi_n(x)|\phi_n(x)\rangle}{\langle\phi_n(x)|\phi_n(x)\rangle}, \quad \beta_n = \frac{\langle\phi_n(x)|\phi_n(x)\rangle}{\langle\phi_{n-1}(x)|\phi_{n-1}(x)\rangle}$$

⑤ $\phi_n(x)$ $(n \geq 1)$ has n different roots in the interval $[a, b]$.

6.5.2 Legendre Polynomials

For $w(x) \equiv 1$ and $[a, b] = [-1, 1]$, the orthogonalization of $\{1, x, x^2, \cdots\}$ gives the Legendre polynomials.

$$P_0(x) = 1, \tag{6.65}$$

$$P_n(x) = \frac{1}{2^n n!}\frac{\mathrm{d}^n}{\mathrm{d}\,x^n}[(x^2 - 1)^n], \quad x \in [-1, 1] \tag{6.66}$$

$$= \frac{1}{2^n n!}\sum_{k=0}^{n}(-1)^{n-k}\binom{n}{k}\frac{(2k)!}{(2k-n)!}x^{2k-n} \tag{6.67}$$

Properties:

- Orthogonality

$$\langle P_n|P_k\rangle = \frac{2}{2n+1}\delta_{nk} \tag{6.68}$$

- Parity

$$P_n(-x) = (-1)^n P_n(x) \tag{6.69}$$

- Iterative relation

$$\begin{aligned}
P_0(x) &= 1, \\
P_1(x) &= x, \\
P_2(x) &= \frac{1}{2}(3x^2 - 1), \\
P_3(x) &= \frac{1}{2}(5x^3 - 3x), \\
P_4(x) &= \frac{1}{8}(35x^4 - 30x^2 + 3), \\
&\vdots \\
(n+1)P_{n+1}(x) &= (2n+1)xP_n(x) - nP_{n-1}(x), \quad n \geq 1
\end{aligned} \tag{6.70}$$

- $P_n(x)$ has n real roots in $[-1, 1]$.

- For $p(x) \in \mathcal{L}_n = \text{span}\left\{1, x, x^2, \cdots, x^n\right\}$ and $x \in [-1, 1]$,

$$\begin{aligned}\hat{P}_n(x) &= \arg\min_{p(x)\in\mathcal{L}_n} \|p(x) - 0\|_{C[-1,1]} \\ &= \frac{n!}{(2n)!}\frac{\mathrm{d}^n}{\mathrm{d}\,x^n}[(x^2-1)^n] = \frac{(n!)^2 2^n}{(2n)!}P_n(x)\end{aligned}$$

- Generating function

$$\frac{1}{\sqrt{1-2ux+u^2}} = \sum_{n=0}^{\infty} P_n(x)u^n \tag{6.71}$$

6.5.3 Cebyshev Polynomials

First Class Cebyshev Polynomials

For $w(x) = \frac{1}{\sqrt{1-x^2}}$ and $[a, b] = [-1, 1]$, the orthogonalization of $\left\{1, x, x^2, \cdots\right\}$ gives the Cebyshev polynomials.

$$T_n(x) = \cos(n \arccos x), \quad x \in [-1, 1] \tag{6.72}$$

Properties:

- Orthogonality

$$\langle T_n | T_k \rangle = \begin{cases} 0, & n \neq k \\ \frac{\pi}{2}, & n = k \neq 0, \\ \pi, & n = k = 0. \end{cases} \tag{6.73}$$

- Parity

$$T_n(-x) = (-1)^n T_n(x) \tag{6.74}$$

- Iterative relation

$$\begin{aligned} T_0(x) &= 1, \\ T_1(x) &= x, \\ T_2(x) &= 2x^2 - 1, \\ T_3(x) &= 4x^3 - 3x, \\ T_4(x) &= 8x^4 - 8x^2 + 1, \\ &\vdots \\ T_{n+1}(x) &= 2xT_n(x) - T_{n-1}(x), \quad n \geq 1 \end{aligned} \tag{6.75}$$

- $P_n(x)$ has n real roots in $[-1, 1]$

$$x_k = \cos\frac{(2k-1)\pi}{2n}, \quad k = 1, 2, \cdots, n.$$

- For $p(x) \in \mathcal{L}_n = \text{span}\left\{1, x, x^2, \cdots, x^n\right\}$ and $x \in [-1, 1]$, we have

$$\begin{aligned} 1 &= T_0 \\ x &= T_1 \\ x^2 &= \frac{1}{2}(T_0 + T_1) \\ x^3 &= \frac{1}{4}(3T_1 + T_3) \\ x^4 &= \frac{1}{8}(3T_0 + 4T_2 + T_4) \end{aligned}$$

- For $p(x) \in \mathcal{L}_n = \text{span}\left\{1, x, x^2, \cdots, x^n\right\}$ and $x \in [-1, 1]$,

$$\begin{aligned}\hat{T}_n(x) &= \arg \min_{p(x)\in\mathcal{L}_n} \|p(x) - 0\|_{C[-1,1]} \\ &= \frac{1}{2^{n-1}} T_n(x)\end{aligned}$$

- Generating function

$$\frac{1-u^2}{1-2ux+u^2} = \sum_{n=0}^{\infty} T_n(x)(2u)^n \tag{6.76}$$

Second Class Cebyshev Polynomials

For $w(x) = \sqrt{1-x^2}$ and $[a, b] = [-1, 1]$, the orthogonalization of $\left\{1, x, x^2, \cdots\right\}$ gives the second class Cebyshev polynomials.

$$U_n(x) = \frac{\sin[(n+1)\arccos x)]}{\sqrt{1-x^2}}, \quad x \in [-1, 1] \tag{6.77}$$

with

$$\langle U_n | U_k \rangle = \frac{\pi}{2}\delta_{nk}, \tag{6.78}$$

$$U_0(x) = 1, \tag{6.79}$$

$$U_1(x) = 2x, \tag{6.80}$$

$$U_{n+1}(x) = 2xU_n(x) - U_{n-1}(x), \quad n = 1, 2, \cdots \tag{6.81}$$

6.5.4 Jacobi Polynomials

For $w(x) = x^{q-1}(1-x)^{p-q}, q > 0, p - q > -1$, and $[a, b] = [0, 1]$, the orthogonalization of $\left\{1, x, x^2, \cdots\right\}$ gives the second class Cebyshev polynomials.

$$G_n(p, q, x) = \frac{x^{1-q}(1-x)^{q-p}}{q(q+1)\cdots(q+n-1)} \frac{\mathrm{d}^n}{\mathrm{d}\, x^n}\left[x^{q-1+n}(1-x)^{p-q+n}\right] \tag{6.82}$$

$$= \frac{(q-1)!}{(q+n-1)!} \sum_{s=0}^{n} \frac{(q+n+s-1)!}{(n-s)!s!(q+s-1)!} x^s \tag{6.83}$$

with

$$\langle G_n | G_k \rangle = \frac{n![(q-1)!]^2(p-q+n)!}{(q-1+n)!(p-1+n)!(p+2n)}\delta_{nk}, \tag{6.84}$$

$$\tag{6.85}$$

Generating function

$$G_n(p, q, x) \leftarrow \mathrm{F}\left(\left.\frac{\alpha, \beta}{\gamma}\right| x\right) \text{ with } \alpha = p + n, \beta = -n, \gamma = q$$

6.5.5 Laguerre Polynomials

For $w(x) = \mathrm{e}^{-x}$ and $[a, b] = [0, \infty)$, the orthogonalization of $\left\{1, x, x^2, \cdots\right\}$ gives the second class Cebyshev polynomials.

$$L_n(x) = \mathrm{e}^x \frac{\mathrm{d}^n}{\mathrm{d}\, x^n}(x^n \mathrm{e}^{-x}), \quad x \in [0, \infty) \tag{6.86}$$

with

$$\langle L_n|L_k\rangle = (n!)^2\delta_{nk}, \tag{6.87}$$
$$L_0(x) = 1, \tag{6.88}$$
$$L_1(x) = 1 - x, \tag{6.89}$$
$$L_{n+1}(x) = (1 + 2n - x)L_n(x) - n^2L_{n-1}(x), \quad n = 1, 2, \cdots \tag{6.90}$$

Generating function

$$\frac{\mathrm{e}^{-\frac{xu}{1-u}}}{1-u} = \sum_{n=0}^{\infty}\frac{L_n(x)}{n!}u^n \tag{6.91}$$

6.5.6 Hermite Polynomials

First Definition

For $w(x) = \mathrm{e}^{-x^2}$ and $[a, b] = (-\infty, \infty)$, the orthogonalization of $\{1, x, x^2, \cdots\}$ gives the second class Cebyshev polynomials.

$$H_n(x) = (-1)^n\mathrm{e}^{x^2}\frac{\mathrm{d}^n}{\mathrm{d}\,x^n}\mathrm{e}^{-x^2}, \quad x \in \mathbb{R} \tag{6.92}$$

with

$$\langle H_n|H_k\rangle = n!2^n\sqrt{\pi}\cdot\delta_{nk}, \tag{6.93}$$
$$H_0(x) = 1, \tag{6.94}$$
$$H_1(x) = 2x, \tag{6.95}$$
$$H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x), \quad n = 1, 2, \cdots \tag{6.96}$$

Generating function

$$\mathrm{e}^{2ux-u^2} = \sum_{n=0}^{\infty}\frac{H_n(x)}{n!}u^n \tag{6.97}$$

Second Definition

For $w(x) = \phi(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}$ and $[a, b] = (-\infty, \infty)$, the orthogonalization of $\{1, x, x^2, \cdots\}$ gives the second class Cebyshev polynomials.

$$H_n(x) = (-1)^n\frac{1}{\phi(x)}\frac{\mathrm{d}^n}{\mathrm{d}\,x^n}\phi(x), \quad x \in \mathbb{R} \tag{6.98}$$

with

$$\langle H_n|H_k\rangle = n!\cdot\delta_{nk}, \tag{6.99}$$
$$H_0(x) = 1, \tag{6.100}$$
$$H_1(x) = x, \tag{6.101}$$
$$H_2(x) = x^2 - 1, \tag{6.102}$$
$$H_3(x) = x^3 - 3x, \tag{6.103}$$
$$H_{n+1}(x) = xH_n(x) - H_n'(x), \quad n = 1, 2, \cdots \tag{6.104}$$
$$H_n'(x) = nH_{n-1}(x), \quad n = 1, 2, \cdots \tag{6.105}$$
$$H_{n+1}(x) = xH_n(x) - nH_{n-1}(x), \quad n = 1, 2, \cdots \tag{6.106}$$

6.5.7 Zernike Polynomials

See the Appendix A and Appendix B.

Chapter 7

Eigenvalue Problems

> The validity of theorems on eigenfunctions can be made plausible by the following observation made by Daniel Bernoulli (1700-1782). A mechanical system of n degrees of freedom posses exactly n eigensolutions. A membrance is, however, a system with an infinite number of degrees of freedom. This system will, therefore, have an infinite number of eigenoscillations.
>
> Arnold Sommerfeld, 1900

> A great master of mathematics passed away when Hilbert died in Göttingen on February 14, 1943, at the age of eighty-one. In retrospect, it seems that the era of mathematics upon which he impressed the seal of his spirit, and which is now sinking below the horizon, achieved a more perfect balance than has prevailed before or since, between the mastering of single concrete problems and the formation of general abstract concepts.
>
> Hermann Weyl, 1944

The objective of this chapter is to study the f eigenvalue problem on a given Hilbert space over some field, along with applicaitons to integral equations and boundary-value problems.

Definition 182. *Let $\mathcal{X}$ is a Hilbert space over $\mathbb{F}$ ($\mathbb{R}$ or $\mathbb{C}$). Each solution (u, λ) such that*

$$Au = \lambda u, \quad u \in \mathcal{X}, \lambda \in \mathbb{F}, u \neq 0 \tag{7.1}$$

is called an EIGENSOLUTION *of A, where u is called an* EIGENVECTOR *and λ is called an* EIGENVALUE *of A, respectively. The set of all the eivenvectors u that correspond to a fixed eigenvalue λ is called the* EIGENSPACE *of u, i.e.,*

$$\mathcal{L}_\lambda = \{u \in \mathcal{X} : Au = \lambda u\}.$$

By definition, the eigenvalue λ has **finite multiplicity** iff the corresponding eigenspace has finite dimension.

If the operator $A : \mathcal{X} \to \mathcal{X}$ is a linear compact symmetric operator, then we can show that such operators posses a *complete orthogonormal system of eigenvectors.*

7.1 Symmetric Operators

Definition 183. *The linear operator $A : \mathrm{Dom}(A) \subset \mathcal{X} \to \mathcal{X}$ on the Hilbert space $\mathcal{X}$ over $\mathbb{F}$ is called* SYMMETRIC *iff the domain of definition $\mathrm{Dom}(A)$ is dense in $\mathcal{X}$ and*

$$\langle Au | v \rangle = \langle u | Av \rangle, \quad \forall u, v \in \mathrm{Dom}(A).$$

Proposition 184. *Let $\mathsf{A} : \mathrm{Dom}(\mathsf{A}) \subset \mathcal{X} \to \mathcal{X}$ be a linear symmetric operator on the Hilbert space $\mathcal{X}$ over $\mathbb{F}$. Then*

① *$\langle \mathsf{A}u|v\rangle$ is real for all $u \in \mathrm{Dom}(\mathsf{A})$.*

② *All the eigenvalues of A are real.*

③ *Two eigenvalues of A with different eigenvalues are orthogonal.*

④ *Let $\{u_1, u_2, \cdots\}$ be a finite or countable complete orthonormal system of eigenvectors of A. Then the corresponding system $\{\lambda_1, \lambda_2, \cdots\}$ of eigenvalues contains all the eigenvalues of A.*

Proof

- Ad ①: Let $u \in \mathrm{Dom}(\mathsf{A})$, then

$$\langle \mathsf{A}u|u\rangle = \langle u|\mathsf{A}u\rangle = \overline{\langle \mathsf{A}u|u\rangle},$$

  which implies $\langle \mathsf{A}u|u\rangle \in \mathbb{R}$.

- Ad ②: By the definition of the eigenvalue, we have

$$\lambda\langle u|u\rangle = \langle u|\mathsf{A}u\rangle = \langle \mathsf{A}u|u\rangle = \overline{\lambda}\langle u|u\rangle$$

  with $\langle u|u\rangle \neq 0$. Hence $\lambda = \overline{\lambda}$. Thus λ is real.

- Ad ③: From $\mathsf{A}u = \lambda u$ and $\mathsf{A}v = \mu v$ along with $\lambda, \mu \in \mathbb{R}$ and $\lambda \neq \mu$ it follows that

$$(\lambda - \mu)\langle u|v\rangle = \langle \mathsf{A}u|v\rangle - \langle u|\mathsf{A}v\rangle = 0,$$

  and hence $\langle u|v\rangle = 0$.

- Ad ④: Since $\{u_n\}$ is complete, we have

$$u = \sum_n \langle u_n|u\rangle u_n, \quad \forall u \in \mathcal{X}. \tag{7.2}$$

  Let $\mathsf{A}u = \lambda u$ with $u \neq 0$ and $\lambda \neq \lambda_n$ for all n. For different eigenvectors u_n and u, by ③, $\langle u_n|u\rangle = 0$ for all n. Thus $u = 0$ by (7.2). This constractics $u \neq 0$. ■

Proposition 185. *Let $\mathsf{A} : \mathcal{X} \to \mathcal{X}$ be a linear continuous symmetric operator on the Hilbert space $\mathcal{X}$ over $\mathbb{F}$ with $\mathcal{X} \neq \{0\}$. Then*

$$\|\mathsf{A}\| = \sup_{\|u\|=1} |\langle \mathsf{A}u|u\rangle| .$$

Proof

- Set $\alpha \triangleq \sup_{\|u\|=1} |\langle \mathsf{A}u|u\rangle|$. Since A is linear,

$$|\langle \mathsf{A}v|v\rangle| \leq \alpha \|v\|^2, \quad \forall v \in \mathcal{X}.$$

  By the Cauchy-Schwarz inequality,

$$|\langle \mathsf{A}u|u\rangle| \leq \|\mathsf{A}u\|\, u \leq \|\mathsf{A}\|\, \|u\|^2, \quad \forall u \in \mathcal{X}.$$

  Hence $\alpha \leq \|A\|$.

- To prove that $\|A\| \leq \alpha$, we set

$$v_{\pm} \triangleq \lambda u \pm \lambda^{-1} \mathsf{A} u, \quad \lambda > 0.$$

It follows from $\langle \mathsf{A}^2 u | u \rangle = \langle \mathsf{A}u | \mathsf{A}u \rangle$ and $\|\mathsf{A}u\|^2 = \langle \mathsf{A}u | \mathsf{A}u \rangle$ that

$$\begin{aligned} \|\mathsf{A}u\|^2 &= \frac{1}{4} \left[\langle \mathsf{A}v_+ | v_+ \rangle - \langle \mathsf{A}v_- | v_- \rangle \right] \\ &\leq \frac{1}{4} \alpha (\|v_+\|^2 + \|v_-\|^2) \\ &= \frac{1}{2} \alpha (\lambda^2 \|u\|^2 + \lambda^{-2} \|\mathsf{A}u\|^2). \end{aligned}$$

Assume first that $\mathsf{A}u \neq 0$. Letting $\lambda^2 = \|\mathsf{A}u\|$ and $\|u\| = 1$, we find that $\|\mathsf{A}u\|^2 \leq \alpha \|\mathsf{A}u\|$ if $\|\mathsf{A}u\| \neq 0$. Hence

$$\|\mathsf{A}u\| \leq \alpha, \quad \forall u \in \mathcal{X} \quad \text{with} quad \|u\| = 1.$$

This implies $\|A\| \leq \alpha$. ■

7.2 Hilbert-Schmidt Theory

Theorem 186. *Let $\mathsf{A} : \mathcal{X} \to X$ be a linear compact sysmmetric operator on the separable Hilbert space $\mathcal{X}$ over $\mathbb{F}$ with $\mathcal{X} \neq \{0\}$. Then, the following hold true:*

① *The operator A has a complete orthonormal system of eigenvectors.*

② *All the eigenvalues of A are real, and each eigenvalue $\lambda \neq 0$ of A has finite multiplicity.*

③ *Two eigenvalues of A that correspond to different eigenvalues are orthogonal.*

④ *If the operator A has a countable set of eigenvalues (e.g. $\lambda = 0$ is not an eigenvalue of A and $\dim(\mathcal{X}) = \infty$), then the eigenvalues of A form a sequence (λ_n) such that*

$$\lambda_n \to 0 \quad as \quad n \to \infty.$$

Proof

❶ Additional assumption (A): Suppose that

$$\text{(A)} \quad \mathsf{A}u = 0 \quad \text{implies} \quad u = 0 \quad \text{and let} \quad \dim \mathcal{X} = \infty.$$

Since $\mathcal{X} \neq \{0\}$, this implies $\mathsf{A} \neq 0$ and hence $\|\mathsf{A}\| \neq 0$.

❷ Additional assumption (B): Suppose that

$$\text{(A)} \quad \mathsf{A}u = 0 \quad \text{implies} \quad u = 0 \quad \text{and let} \quad \dim \mathcal{X} = N \in \{1, 2, \cdots\}.$$

❸ We may assume that $\lambda = 0$ is an eigenvalue of A. Otherwise, we meet assumption (A) or (B). Set

$$\mathrm{Ker}(\mathsf{A}) \triangleq \{u \in \mathcal{X} : \mathsf{A}u = 0\}.$$

Then, $\mathrm{Ker}(\mathsf{A}) \neq \{0\}$. Since the operator A is continuous, $\mathrm{Ker}(\mathsf{A})$ is a closed linear subspace of $\mathcal{X}$. In fact, if $\mathsf{A}u_n = 0$ for all n and $u_n \to u$ as $n \to \infty$, then $\mathsf{A}u = 0$.

Since $\mathrm{Ker}(\mathsf{A})$ is a separable Hilbert space, there exists a complete orthonormal system $\{w_k\}$ in $\mathrm{Ker}(\mathsf{A})$. Hence

$$\mathsf{A}w_k = 0, \quad \forall k.$$

Moreover, if follows from the orthogonal decomposition that for each $u \in \mathcal{X}$ there exists the unique decomposition

$$u = w + z, \quad w \in \mathrm{Ker}(A), \quad z \in \mathrm{Ker}(A)^{\perp}. \tag{7.3}$$

Recall that

$$\mathrm{Ker}(A)^{\perp} \triangleq \{z \in \mathcal{X} : \langle z|w\rangle = 0, \quad \forall w \in \mathrm{Ker}(A)\}.$$

Thus, $\mathrm{Ker}(A)^{\perp}$ is a closed linear subspace of $\mathcal{X}$. We have these conditions:

(a) The operator A maps $\mathrm{Ker}(A)^{\perp}$ into $\mathrm{Ker}(A)^{\perp}$.

(b) If $Az = 0$ with $z \in \mathrm{Ker}(A)^{\perp}$, then $z = 0$.

Ad (a). Let $z \in \mathrm{Ker}(A)^{\perp}$. Then

$$\langle Az|w\rangle = \langle z|Aw\rangle = 0, \quad \forall w \in \mathrm{Ker}(A)^{\perp},$$

and hence $Az \in \mathrm{Ker}(A)^{\perp}$.

Ad (b). If $Az = 0$ with $z \in \mathrm{Ker}(A)^{\perp}$, then $z \in \mathrm{Ker}(A) \cap \mathrm{Ker}(A)^{\perp}$. By the uniqueness of the decomposition $u = w + z$, we have $z = 0$.

We now consider the restricted operator

$$A : \mathrm{Ker}(A)^{\perp} \to \mathrm{Ker}(A)^{\perp}.$$

By this way we get a complete orthogonal system $\{u_n\}$ of eigenvalues of A on $\mathrm{Ker}(A)^{\perp}$. Recall that $\{w_k\}$ forms a complete orthonormal system in $\mathrm{Ker}(A)$.

In (7.3), for each $u \in \mathcal{X}$,

$$u = w + z = \sum_k \langle w_k|w\rangle w_k + \sum_n \langle u_n|z\rangle u_n.$$

Since $w \in \mathrm{Ker}(A)$ and $z \in \mathrm{Ker}(A)^{\perp}$, we get $\langle w_k|z\rangle = 0$ for all k and $\langle u_n|w\rangle = 0$ for all n. Hence

$$u = \sum_k \langle w_k|u\rangle w_k + \sum_n \langle u_n|u\rangle u_n.$$

Consequently, $\{w_1, u_1, w_2, u_2, \cdots\}$ represents a complete orthonormal system of eigencevtors of A. ■.

7.3 Fredholm Alternative

Theorem 187. *Suppose $A : \mathcal{X} \to \mathcal{X}$ is linear, compact, and symmetric on the separable Hilbert space $\mathcal{X}$ over $\mathbb{F}$ with $\mathcal{X} = \{0\}$. We are given $b \in \mathcal{X}$ and the number $\lambda \in \mathcal{F}$ with $\lambda \neq 0$. For the nonhomegenious equation*

$$\lambda u - Au = b, \quad \forall u \in \mathcal{X}, \tag{7.4}$$

along with the homogenous problem

$$\lambda v - Av = 0, \quad v \in \mathcal{X}, \tag{7.5}$$

Eq.(7.4) has a solution iff for all solutions v of Eq.(7.5) we have

$$\langle b|v\rangle = 0. \tag{7.6}$$

Proof

- Suppose first that the operator A has a countable system of nonzero eigenvalues. Then there exists an orthonormal system $\{u_n\}$ in $\mathcal{X}$ such that

$$\mathsf{A}u = \sum_{n=1} \infty \lambda_n \langle u_n | u \rangle u_n, \tag{7.7}$$

along with $\mathsf{A}u_n = \lambda_n u_n$ for all n. The system $\{\lambda_n\}$ contains all the nonzero eigenvalues of A, and

$$\lambda_n \to 0, \quad \text{as} \quad n \to \infty. \tag{7.8}$$

Furthermore,

$$\langle u_n | \mathsf{A}u \rangle = \langle \mathsf{A}u_n | u \rangle = \lambda_n \langle u_n | u \rangle, \quad \forall u \in \mathcal{X}, \forall n. \tag{7.9}$$

- Case 1: Suppose that $\lambda \neq \lambda_n$ for all n, i.e., Eq.(7.5) only has the trivial solution $v = 0$. Then, equation (7.4) has at most one solution.

  – Let u be a solution of Eq.(7.4). Then

$$u = \lambda^{-1} \left[b + \sum_{n=1}^{\infty} \lambda_n \langle u_n | u \rangle u_n \right].$$

  By (7.4) and (7.9),

$$(\lambda - \lambda_n) \langle u_n | u \rangle = \langle u_n | (\lambda \mathbb{1} - \mathsf{A}) u \rangle = \langle u_n | b \rangle, \quad \forall n. \tag{7.10}$$

  Hence

$$u = \lambda^{-1} \left[b + \sum_{n=1}^{\infty} \alpha_n \langle u_n | u \rangle u_n \right], \quad \alpha_n \triangleq \frac{\lambda_n}{\lambda - \lambda_n}. \tag{7.11}$$

  – Conversely, it follows from the Bessel inequality

$$\sum_n |\langle u_n | b \rangle|^2 \leq \|b\|^2$$

  along with $\lambda_n \to 0$ as $n \to \infty$ that $\alpha_n \leq \text{const}$ for all n, and hence

$$\sum_n |\alpha_n \langle u_n | b \rangle|^2 \leq \text{const} \cdot \|b\|^2$$

  By the convergence criterion in Hilbert space (Proposition 171), the series (7.9) is convergent. It follows from (7.10) with $u = b$ and (7.11) that

$$\mathsf{A}u = \lambda^{-1} \left[\mathsf{A}b + \sum_{n=1}^{\infty} \alpha_n \lambda_n \langle u_n | b \rangle u_n \right] = \sum_{n=1}^{\infty} \lambda^{-1} \lambda_n (1 + \alpha_n) \langle u_n | b \rangle u_n$$

  and

$$\lambda u = b + \sum_{n=1}^{\infty} \alpha_n \langle u_n | b \rangle u_n.$$

  Hence

$$\lambda u - \mathsf{A}u = b,$$

  i.e., u is a solution of (7.6). Since the solution is unique, the inverse operator $(\lambda \mathbb{1} - \mathsf{A})^{-1} : \mathcal{X} \to \mathcal{X}$ exists. In addition, (7.11) tells us that

$$\begin{aligned} \|u\|^2 &= \left|\lambda^{-2}\right| \left[\|b\|^2 + \sum_{n=1}^{\infty} 2\alpha_n \left|\langle u_n | b \rangle\right|^2 + \alpha_n^2 \left|\langle u_n | b \rangle\right|^2 \right] \\ &\leq \left|\lambda^{-2}\right| \left[\|b\|^2 + \sum_{n=1}^{\infty} \text{const} \left|\langle u_n | b \rangle\right|^2 \right] \\ &\leq \text{const} \cdot \|b\|^2 . \end{aligned}$$

This implies

$$\|u\| \leq \text{const} \cdot \|b\|, \quad \forall b \in \mathcal{X}.$$

Hence the linear operator $(\lambda \mathbb{1} - \mathsf{A})^{-1} : \mathcal{X} \to \mathcal{X}$ is continuous.

- Case 2: Suppose λ is an eigenvalue of A, i.e., $\lambda = \lambda_m$ for some m. To simplify notation, let us assume that $\lambda = \lambda_1$. Then $\lambda_1 = \lambda_2 = \cdots = \lambda_N$ for some natural number N and $\lambda_n \neq \lambda_1$ for all $n > N$.
  - If u is a solution of (7.4), then it follows from (7.10) that

$$\langle u_n | b \rangle = 0, \quad \forall n = 1, \cdots, N. \tag{7.12}$$

    This is equivalent to condition (7.6). Hence

$$u = \lambda^{-1} \left[b + \sum_{n=N+1}^{\infty} \alpha_n \langle u_n | b \rangle u_n \right], \quad \alpha_n = \frac{\lambda_n}{\lambda - \lambda_n}. \tag{7.13}$$

  - Conversely, let condition (7.12) be fulfilled. As in Case 1, one checks easily that u from (7.13) satisfies $\lambda u - \mathsf{A}u = b$, i.e., u is a solution of (7.4).
- Finally, observe that if the operator A has only a finite number of nonzero eigenvalues λ_n, then the series from (7.11) and (7.13) reduce to finite sums $\sum_{n=1}^{N} \cdots$.

Corollary 188. *Suppose $\mathsf{A} : \mathcal{X} \to \mathcal{X}$ is linear, compact, and symmetric on the separable Hilbert space $\mathcal{X}$ over $\mathbb{F}$ with $\mathcal{X} = \{0\}$. We are given $b \in \mathcal{X}$ and $\lambda\mathbb{F}$ with $\lambda \neq 0$. Suppose that Eq.(7.4) has at most one solution. Then, the following hold true:*

① *There exists the linear continuous operator $(\lambda \mathbb{1} - \mathsf{A})^{-1} : \mathcal{X} \to \mathcal{X}$.*

② *Eq.(7.4) has the unique solution $u = (\lambda \mathbb{1} - \mathsf{A})^{-1} b$.*

This corollary tells us that for the original problem (7.4), uniqueness implies existence.

Proof

- If equation (7.4) has at most one solution, then

$$\lambda u - \mathsf{A}u = \lambda v - \mathsf{A}v \quad \text{implies} \quad u = v.$$

- By the linearity of A, this is equivalent to the fact that $\lambda w - \mathsf{A}w = 0$ implies $w = 0$. The assertion follows now from Case 1 in the preceding proof.

Corollary 189. *Suppose $\mathsf{A} : \mathcal{X} \to \mathcal{X}$ is linear, compact, and symmetric on the separable Hilbert space $\mathcal{X}$ over $\mathbb{C}$ with $\mathcal{X} = \{0\}$. Then, the following are met:*

① *If $\dim(\mathcal{X}) < \infty$, then the spectrum $\sigma(\mathsf{A})$ of the operator A consits precisely of all the eivenvalues of the operator A.*

② *If $\dim(\mathcal{X}) = \infty$, then the spectrum $\sigma(\mathsf{A})$ of the operator A consits precisely of A together with the point $\lambda = 0$.*

Proof

- Let $\lambda \neq \lambda_n$ for all n and $\lambda \neq 0$. By Corollary 188 ①, the point $\lambda \in \mathbb{C}$ belongs to the resolvent set of A.

- Ad ①. If $\lambda = 0$ is an eigenvalue of A, then $0 \in \sigma(\mathsf{A})$. If $\lambda = 0$ is not an eigenvalue of A, then $\mathsf{A}v = 0$ implies $v = 0$. By Theorem 187, the operator A has a complete orthonormal system $\{u_n\}$ of eigenvectors. Set
$$u \triangleq \sum_n \lambda^{-1} \langle u_n | b \rangle u_n.$$
Then, $\mathsf{A}u = b$ and $\|u\| \leq \text{const}\, \|b\|$. Thus, the operator $\mathsf{A}^{-1} : \mathcal{X} \to \mathcal{X}$ is continuous, i.e., $\lambda = 0$ belongs to the resolvent set of A.

- Ad ②. Suppose first that the operator A has only a finite number of nonzero eigenvalues. Since all these eigenvalues have finite multiplicity and $\dim(\mathcal{X}) = \infty$, it follows from Theorem 187 that there is some $v \neq 0$ for which $\mathsf{A}v = 0$, i.e., $\lambda = 0$ belongs to the spectrum $\sigma(\mathsf{A})$. Suppose now that the operator A has a countable set $\{\lambda_n\}$ of nonzero eigenvalues. Then $\lambda_n \to 0$ as $n \to \infty$. Since the spectrum $\sigma(\mathsf{A})$ is compact, the limit point $\lambda = 0$ belongs to $\sigma(\mathsf{A})$.

7.4 Applications to Integral Equations

7.4.1 Homogeneous Integral Equation

We want to study the following integral equation:

$$\int_a^b K(x,y)u(y)\,\mathrm{d}\,y = \lambda u(x), \quad x \in [a,b]. \tag{7.14}$$

Let $-\infty < a < b < \infty$. We are looking for eigensolutions $\lambda \in \mathbb{R}$ and $u \in L^2(a,b)$ with $u \neq 0$. To this end, we assume the following:

(H1) The function $K : [a,b] \times [a,b] \to \mathbb{R}$ is continuous.

(H2) The function K is symmetric, i.e.,

$$K(x,y) = K(y,x), \quad \forall x, y \in [a,b].$$

We set $\mathcal{X} = L^2(a,b)$ along with the inner product

$$\langle u | v \rangle = \int_a^b u(x)v(x)\,\mathrm{d}\,x$$

and define the integral operator

$$(\mathsf{A}u)(x) \triangleq \int_a^b K(x,y)u(y)\,\mathrm{d}\,y.$$

Then, the original equation (7.14) can be written in the following form:

$$\mathsf{A}u = \lambda u, \quad \lambda \in \mathbb{R}, \quad u \in \mathcal{X}, \quad u \neq 0. \tag{7.15}$$

Lemma 190. *Assume (H1) above and let $\mathcal{X} \triangleq L^2(a,b)$. Then, the following hold true:*

① *The operator $\mathsf{A} : \mathcal{X} \to \mathcal{X}$ is linear and compact.*

② *If $u \in \mathcal{X}$, then $\mathsf{A}u \in C[a,b]$.*

③ *If, in addition, (H2) holds, then the operator $\mathsf{A} : \mathcal{X} \to \mathcal{X}$ is symmetric.*

Proof

- Ad ① and ②. Let $u \in \mathcal{X}$, and let $\|u\|$ denote the norm of u in $\mathcal{X}$. By Cauchy-Schwartz inequality,

$$\int_a^b 1 \cdot |u(y)| \,\mathrm{d}\, y \le \sqrt{\int_a^b 1 \,\mathrm{d}\, y} \cdot \sqrt{\int_a^b |u(y)|^2 \,\mathrm{d}\, y} = \sqrt{b-a}\, \|u\| .$$

Set $v \triangleq \mathsf{A}u$, i.e.,

$$v(x) = \int_a^b K(x,y) u(y) \,\mathrm{d}\, y, \quad \forall x \in [a,b].$$

Since the set $[a,b] \times [a,b]$ is compact, the function K is uniformly continuous on $[a,b] \times [a,b]$. Thus, for each $\varepsilon > 0$, there is a $\delta > 0$ such that

$$x, z \in [a,b] \text{and} \, |x - z| < \delta \Longrightarrow \alpha \triangleq \max_{y \in [a,b]} |K(x,y) - K(z,y)| < \varepsilon.$$

Hence

$$|v(x) - v(z)| \le \alpha \int_a^b |u(y)| \,\mathrm{d}\, y \le \varepsilon \sqrt{b-a}\, \|u\| , \tag{7.16}$$

for all $x, z \in [a,b]$ with $|x - z| < \delta$. This proves the continuity of the function v on $[a,b]$.

Moreover, we also get

$$\max_{x \in [a,b]} |v(x)| \le \max_{x,y \in [a,b]} |K(x,y)| \int_a^b |u(y)| \,\mathrm{d}\, y \le \text{const}\, \|u\| . \tag{7.17}$$

This implies

$$\|\mathsf{A}u\| = \sqrt{\int_a^b |v(x)|^2 \,\mathrm{d}\, x} \le \text{const}\, \|u\| . \tag{7.18}$$

Obviously, the operator $\mathsf{A} : \mathcal{X} \to \mathcal{X}$ is *linear*. Consequently, relation (7.18) tells us that $\mathsf{A} : \mathcal{X} \to \mathcal{X}$ is *continuous*.

Let M be a bounded set in $\mathcal{X}$. Then it follows from (7.16) and (7.17) along with the Arzelà-Ascoli theorem that the set $\mathsf{A}(M)$ is relatively compact in $C[a,b]$.

Each relatively compact set in $C[a,b]$ is also relatively compact in $\mathcal{X} = L^2(a,b)$. In fact, if $v_n \to v$ in $C[a,b]$ as $n \to \infty$, then

$$\begin{aligned} \|v_n - v\| &= \sqrt{\int_a^b [v_n(x) - v(x)]^2 \,\mathrm{d}\, x} \\ &\le \max_{x \in [a,b]} |v_n(x) - v(x)| \sqrt{b-a} \to 0 \quad \text{as} \quad n \to \infty. \end{aligned}$$

Thus, the set $\mathsf{A}(M)$ is relatively compact in $\mathcal{X}$, and hence $\mathsf{A} : \mathcal{X} \to \mathcal{X}$ is compact.

- Ad ③. For all $u, v \in \mathcal{X}$,

$$\begin{aligned} \langle \mathsf{A}u | v \rangle &= \int_a^b \left[\int_a^b K(x,y) u(y) \,\mathrm{d}\, y \right] v(x) \,\mathrm{d}\, x \\ &= \int_a^b \left[\int_a^b K(x,y) v(x) \,\mathrm{d}\, x \right] u(y) \,\mathrm{d}\, y = \langle u | \mathsf{A}v \rangle \end{aligned}$$

since $K(x,y) = K(y,x)$ for all $x, y \in [a,b]$.

Proposition 191 (Eigensolutions)**.** *Under assumptions (H1) and (H2), the following hold true:*

① *The original equation (7.14) has a countable system of eigenfunctions $\{u_1, u_2, \cdots\}$, which forms a complete orthonormal system in the Hilbert space $L^2(a,b)$.*

② *Two eigenfunctions u and v of (7.14) which correspond to different eigenvalues are orthogonal in $L^2(a,b)$, i.e., $\langle u|v\rangle = 0$.*

③ *Each nonzero eigenvalue of (7.14) has finite multiplicity.*

④ *If the integral equation (7.14) has a countable number of eigenvalues (e.g., $\lambda = 0$ is not an eigenvalue of (7.14)), then all the nonzero eigenvalues of (7.14) form a sequence (λ_n) with $\lambda_n \to 0$.*

⑤ *For each eigenvalue $\lambda \neq 0$ of (7.14)), the eigenfunctions u are continuous on $[a,b]$.*

By ①, for each $u \in L^2(a,b)$, the Fourier series

$$u = \sum_{n=1}^{\infty} \langle u_n|u\rangle u_n \tag{7.19}$$

converges in $L^2(a,b)$, i.e.,

$$\lim_{m\to\infty} \int_a^b \left[u(x) - \sum_{n=1}^{m} \langle u_n|u\rangle u_n(x)\right]^2 \mathrm{d}\,x = 0.$$

Corollary 192 (Classical Convergence)**.** *Suppose that the function u allows the following representation:*

$$u(x) = \int_a^b K(x,y)v(y)\,\mathrm{d}\,y, \quad \forall x \in [a,b],$$

where $v \in L^2(a,b)$. Then, the Fourier series

$$u(x) = \sum_{n=1}^{\infty} \langle u_n|u\rangle u_n(x) \tag{7.20}$$

converges absolutely and uniformly on the interval $[a,b]$. In addition, we have $\langle u_n|u\rangle = 0$ if the eigenvector u_n corresponds to the eigenvalue $\lambda = 0$.

7.4.2 Non-homogeneous Integral Equation

We now consider the following nonhomogenous integral equation:

$$\int_a^b K(x,y)u(y)\,\mathrm{d}\,y - \lambda u(x) = h(x), \quad x \in [a,b]. \tag{7.21}$$

To this end, we need the corresponding homogeneous equation

$$\int_a^b K(x,y)u(y)\,\mathrm{d}\,y - \lambda u(x) = 0, \quad x \in [a,b]. \tag{7.22}$$

Proposition 193 (The Fredholm Alternative)**.** *Assume (H1) and (H2). Let the function $h \in L^2(a,b)$ and the real number $\lambda \neq 0$ be given. Then, the following hold ture:*

① *If λ is not an eigenvalue of the homogeneous integral equation (7.22), then the original equation (7.21) has a unique solution $u \in L^2(a,b)$.*

② *If λ is an eigenvalue of (7.21), then (7.22) has a solution $u \in L^2(a,b)$ iff*

$$\int_a^b h(x)v(x)\,\mathrm{d}\,x = 0,$$

for all the eigenfunctions v corresponding to λ.

③ *If $h \in C[a,b]$, then each solution u of (7.21) is continuous on $[a,b]$.*

This follows from Theorem 187 along with Lemma 190. Observe that equation (7.21) can be written in the following form:

$$\mathsf{A}u - \lambda u = h, \quad u \in \mathcal{X}, \quad \lambda \in \mathbb{R},$$

where $\mathcal{X} \triangleq L^2(a,b)$ and $\langle h|v\rangle \triangleq \int_a^b h(x)v(x)\,\mathrm{d}\,x$.

7.5 Applications to Boundary-Eigenvalue Problems

7.5.1 Homogeneous Equation

Let us consider the following boundary-eigenvalue problem:

$$\begin{cases} -\dfrac{\mathrm{d}^2 u}{\mathrm{d}\,x^2} = \mu u(x), \quad x \in (0,a), \\ u(0) = u(a) = 0, \quad u \in C^2[0,a], \mu \in \mathbb{R}. \end{cases} \tag{7.23}$$

This problem can be written in the following form:

$$\mathsf{A}u = \mu u, \quad \mu \in \mathbb{R}, \quad u \in \mathrm{Dom}(\mathsf{A}), \tag{7.24}$$

where $\mathsf{A}u \triangleq -\frac{\mathrm{d}^2 u}{\mathrm{d}\,x^2}$ and

$$\mathrm{Dom}(\mathsf{A}) \triangleq \left\{u \in C^2[0,a] : u(0) = u(a) = 0\right\}.$$

Let us also consider the following integral equation:

$$u(x) = \mu \int_0^a G(x,y)u(y)\,\mathrm{d}\,y, \quad x \in [0,a], u \in L^2(0,a), \tag{7.25}$$

where

$$G(x,y) \triangleq \begin{cases} \frac{(a-y)x}{a}, & \text{if } 0 \le x \le y \le a; \\ \frac{(a-x)y}{a}, & \text{if } 0 \le y < x \le a. \end{cases}$$

Obviously, the Green function G is continuous and symmetric on $[0,a] \times [0,a]$.

We set $\mathcal{X} \triangleq L^2(0,a)$ and

$$\langle u|v\rangle \triangleq \int_0^a u(x)v(x)\,\mathrm{d}\,x, \quad \forall u, v \in \mathcal{X}.$$

Lemma 194.

① *The linear operator $\mathsf{A} : \mathrm{Dom}(\mathsf{A}) \subset \mathcal{X} \to \mathcal{X}$ is symmetric.*

② *Two eigenfunctions u and v of (7.23) which correspond to different eigenvalues are orthogonal in $\mathcal{X}$, i.e., $\langle u|v\rangle = 0$.*

③ *Each eigenvalue μ of (7.23) is positive.*

④ *The original boundary-eigenvalue problem (7.23) is equivalent to the integral equation (7.25).*

Proof

- Ad ①. Integration by parts shows that, for all $u, v \in \mathrm{Dom}(\mathsf{A})$,

$$\begin{aligned} \langle \mathsf{A}u|v\rangle &= \int_0^a (-u'')v\,\mathrm{d}\,x = -\,u'v|_0^a + \int_0^a u'v'\,\mathrm{d}\,x \\ &= \int_0^a u'v'\,\mathrm{d}\,x = uv'|_0^a - \int_0^a uv''\,\mathrm{d}\,x \\ &= -\int_0^a uv''\,\mathrm{d}\,x = \langle u|\mathsf{A}v\rangle, \end{aligned} \tag{7.26}$$

since u and v vanish at the boundary points $x = 0$ and $x = a$.

- Ad ②. This follows from Proposition 184.
- Ad ③. Let $\mathsf{A}u = \mu u$, where $\mu \in \Re, u \in \mathrm{Dom}(\mathsf{A})$, and $u \neq 0$. By (7.26),

$$\mu\langle u|u\rangle = \langle \mathsf{A}u|u\rangle = \int_0^a [u']^2 \,\mathrm{d}\, x > 0.$$

  Hence $\mu > 0$.
- Ad ④. If u is a solution of (7.23), then u is also a solution of (7.25). Conversely, if u is a solution of (7.25), then it follows from Lemma 190 that $u \in C[0,a]$ and u is a solution of (7.23).

Proposition 195.

① *The original problem (7.13) has precisely the eigenvalues*

$$\mu_n = \Big(\frac{n\pi}{a}\Big)^2, \quad n = 1, 2, \cdots,$$

which are simple[1].

② *The normalized eigenfunction u_n to μ_n with $\langle u_n|\mu_n\rangle = 1$ is given by*

$$u_n(x) = \sqrt{\frac{2}{a}} \sin \frac{n\pi x}{a} \tag{7.27}$$

③ *For each function $u \in \mathrm{Dom}(\mathsf{A})$, the Fourier series*

$$u(x) = \sum_{n=1}^{\infty} \langle u_n|u\rangle u_n(x) \tag{7.28}$$

converges absolutely and uniformly on the interval $[0,a]$. The same is true for

$$u'(x) = \sum_{n=1}^{\infty} \langle u_n|u\rangle n_n'(x). \tag{7.29}$$

For each $u \in \mathrm{Dom}(\mathsf{A})$ with $u'' \in \mathrm{Dom}(\mathsf{A})$, the series

$$u''(x) = \sum_{n=1}^{\infty} \langle u_n|u\rangle n_n''(x). \tag{7.30}$$

converges absolutely and uniformly on the interval $[0,a]$.

④ *For each $u \in \mathcal{X}$, the Fourier series (7.28) converges in $\mathcal{X} \triangleq L^2(0,a)$, i.e., $\{u_1, u_2, \cdots\}$ forms a complete orthonormal system in $\mathcal{X}$.*

7.5.2 Non-homogeneous Equation

The set $C_0^\infty(0,a)$ is dense in $L^2(0,a)$. Let $v \in L^2(0,a)$, and let $\varepsilon > 0$ be given. Then, there is a function $u \in C_0^\infty(0,a)$ such that

$$\|v - u\| = \sqrt{\int_0^a [v(x) - u(x)]^2 \,\mathrm{d}\, x} < \varepsilon.$$

Let us now consider the following nonhomogeneous boundary-value problem:

$$\begin{cases} -\dfrac{\mathrm{d}^2 u}{\mathrm{d}\, x^2} = \mu u(x) + f(x), \quad x \in (0,a), \\ u(0) = u(a) = 0, \quad u \in C^2[0,a], \mu \in \mathbb{R}. \end{cases} \tag{7.31}$$

[1] Recall that eigenvalues of multiplicity one are called *simple*

Proposition 196 (The Fredholm Alternative)**.** *We are given the function $f \in C[0,a]$ and the real number μ. Then, the following hold true:*

① *If μ is not an eigenvalue of (7.23), then the original equation (7.31) has a unique solution u.*

② *If μ is an eigenvalue of (7.23), i.e., $\mu = \mu_n = \left[\frac{n\pi}{a}\right]^2$ for some $n = 1, 2, \cdots$, then equation (7.31) has a solution u iff the so-called non-resonance condition*

$$\int_0^a f(x)u_n(x)\,\mathrm{d}\,x = 0.$$

is satisfied.

7.6 Self-Adjoint Operators

7.6.1 Introduction

In this section, we want to study the following problems in a Hilbert space $\mathcal{X}$, where $A : \mathrm{Dom}(A) \subset \mathcal{X} \to \mathcal{X}$ is a linear symmetric operator that has additional properties to be discussed:

① Abstract boundary-value problem:

$$Au = f, \quad u \in \mathrm{Dom}(A). \tag{7.32}$$

② Abstract Dirichlet problem:

$$\min_{u\in\mathrm{Dom}(A)} \left[\frac{1}{2}\langle Au|u\rangle - \langle f|u\rangle\right] \tag{7.33}$$

which is equivalent to (7.33). This minimum problem is also equivalent to

$$\min_{u\in\mathcal{X}_E} \left[\frac{1}{2}\langle u|u\rangle_E - \langle f|u\rangle\right] \tag{7.34}$$

where $\mathcal{X}_E$ denotes the so-called energetic space of the operator A.

③ Abstract boundary-eigentic problem:

$$Au - \mu u = f, \quad u \in \mathrm{Dom}(A), \mu \in \mathbb{R}. \tag{7.35}$$

④ Abstract heat equation:

$$\begin{cases} u'(t) + Au(t) = 0, \quad t \geq 0, \\ u(0) = u_0, \end{cases} \tag{7.36}$$

with the solution

$$u(t) = \mathrm{e}^{-At}u_0, \quad \text{for all } t \geq 0,$$

which corresponds to the semigroup $\left\{\mathrm{e}^{-At} : t \geq 0\right\}$.

⑤ Abstract wave equation:

$$\begin{cases} u''(t) + Au(t) = 0, \quad t \geq 0, \\ u(0) = u_0, u'(0) = u_1, \end{cases} \tag{7.37}$$

with the solution

$$u(t) = \left[\cos\left(tA^{1/2}\right)\right] u_0 + A^{-1/2}\left[\sin\left(tA^{1/2}\right)\right] u_1.$$

⑥ Abstract Schrödinger equation:

$$\begin{cases} \mathrm{i}u'(t) = Au(t), \quad t \in \mathbb{R}, \\ u(0) = u_0, \end{cases} \tag{7.38}$$

with the solution

$$u(t) = \mathrm{e}^{-\mathrm{i}At}u_0,$$

which corresponds to the one-parameter group $\{\mathrm{e}^{-\mathrm{i}At} : t \in \mathbb{R}\}$.

Problems ①-⑥ allow important applications to the partial differential equations of mathematical physics.

For example, this concerns boundary-value problems and boundary-eigenvalue problems for the Laplace equation or the Poisson equation, the heat equation, the wave equation, and the Schródiinger equation in quantum mechanics. Such applications of the abstract theory will be considered. In particular, in applications to elasticity the "energetic space" $\mathcal{X}_E$ corresponds to "states" u of the elastic body which has finite energy, and

$$\frac{1}{2}\langle u|u\rangle_E = \text{elastic energy in the state } u,$$
$$\langle f|u\rangle = \text{work of outer forces.}$$

Thus, the variational problem (7.34) corresponds to the *principle of minimal potential energy.*

Observe that the "solutions" in ④-⑥ correspond to classic solutions of ordinary differential equations if we assume that A is a real number and $u = u(t)$ is a real or complex function. The *beauty of functional analysis* consists in the fact that

The classic formula remain true for operator equations if we define operator functions

$$A \mapsto F(A)$$

in an approriate way.

The SIMPLEST METHOD for constructing such operator funcitons is the following.

Suppose that the operator A has a *complete orthonormal system* $\{u_1, u_2, \cdots\}$ of eigenvectors with the corresponding eigenvalues $\{\lambda_1, \lambda_2, \cdots\}$, i.e., $Au_n = \lambda_n u_n$ for all n. Then

$$u = \sum_{n=1}^{\infty} \langle u_n|u\rangle u_n, \quad \forall u \in \mathcal{X}.$$

This yields

$$Au = \sum_{n=1}^{\infty} \langle u_n|Au\rangle u_n = \sum_{n=1}^{\infty} \lambda_n \langle u_n|u\rangle u_n, \quad \forall u \in \mathcal{X} \tag{7.39}$$

Since the symmetry of A implies

$$\langle u_n|Au\rangle = \langle Au_n|u\rangle = \lambda_n\langle u_n|u\rangle.$$

Formula (7.39) motivates the following definition:

$$F(A)u \triangleq \sum_{n=1}^{\infty} F(\lambda_n)\langle u_n|u\rangle u_n. \tag{7.40}$$

In quantum mechanics, this express will be denoted by

$$F(A) \triangleq \sum_{n=1}^{\infty} F(\lambda_n)|u_n\rangle\langle u_n| = \sum_{n=1}^{\infty} F(\lambda_n)|n\rangle\langle n|.$$

with the help of Dirac notation. It is quite natural to define the domain of definition $\mathrm{Dom}(F(A)$ of the operator $F(A)$ as follows:

$$u \in \mathrm{Dom}(F(A)) \quad \text{iff the series (7.40) converges.}$$

By the convergence criterion for abstract/generalized Fourier series, we get

$$u \in \mathrm{Dom}(F(A)) \quad \text{iff} \quad \sum_{n=1}^{\infty} |F(\lambda_n)\langle u_n|u\rangle|^2 < \infty.$$

If we apply this to (7.39), then we obtain

$$u \in \mathrm{Dom}(A) \Longleftrightarrow \sum_{n=1}^{\infty} |\lambda_n \langle u_n|u\rangle|^2 < \infty. \tag{7.41}$$

It turns out that

If the linear operator $A : \mathrm{Dom}(A) \subset \mathcal{X} \to \mathcal{X}$ is self-adjoint, then condition (7.41) is satisfied.

This way the functional calculus leads to self-adjoint operators in a natural way.

In mathematical physics one encounters the following situation. For example, the *classic boundary-value problem* for the Poisson equation

$$\begin{cases} -\triangle u = f, & x \in \Omega, \\ u = 0, & x \in \partial\,\Omega. \end{cases} \tag{7.42}$$

can be written by

$$Bu = f, u \in \mathrm{Dom}(B), \tag{7.43}$$

where we set $Bu \triangleq -\triangle u$ and

$$\mathrm{Dom}(B) \triangleq \left\{u \in C^2(\overline{\Omega}) : u(x) = 0, x \in \partial\,\Omega\right\}.$$

Here G is a nonempty bounded open set in $\mathbb{R}^n$. Letting $\mathcal{X} \triangleq L^2(\Omega)$, we get the linear symmetric operator

$$B : \mathrm{Dom}(B) \subset \mathcal{X} \to \mathcal{X}.$$

However, if $n \geq 2$, then the operator B is not *surjective*. More precisely:

There are functions $f \in C(\overline{\Omega})$ for which equation (7.43) has no solution.

This is identical to the fact that problem (7.42) has not always a classic solution if $f \in C(\overline{\Omega})$.

The idea of Friedrichs was to extend the operator B to a self-adjoint operator

$$A : \mathrm{Dom}(A) \subset \mathcal{X} \to \mathcal{X},$$

i.e., we have

$$Bu = Au \quad \forall u \in \mathrm{Dom}(B) \subset \mathrm{Dom}(A).$$

The operator A is called the *Friedrichs extension* of the original operator B. It turns out that the equation

$$Au = f, \quad u \in \mathrm{Dom}(A) \tag{7.44}$$

has a *unique solution* u for each given $f \in \mathcal{X}$. This solution u of (7.44) can be regarded as

a generalized solution to the classic problem (7.44).

In terms of the expansion (7.39), the situation is as follows. If Ω has a sufficiently smooth boundary, then the symmetric operator $B : \mathrm{Dom}(B) \subset \mathcal{X} \to \mathcal{X}$ has an orthonormal system $\{u_1, u_2, \cdots\}$ of classic eigenfunctions, which is complete in the Hilbert space $\mathcal{X} = L^2(\Omega)$. These eigenfunctions u_n correspond to eigensolutions of the following classic eigenvalue problem:

$$-\triangle u_n(x) = \lambda_n u_n(x), \quad x \in \Omega$$
$$u_n(x) = 0, \quad x \in \partial\,\Omega.$$

Using the same argument as for (7.39), we obtain that

$$Bu = \sum_{n=1}^{\infty} \lambda_n \langle u_n | u \rangle u_n, \quad \forall u \in \mathrm{Dom}(B). \tag{7.45}$$

However, this series also converges for points $u \in \mathcal{X}$ that do not live in $\mathrm{Dom}(B)$. Naturally enough, the *Friedrichs extension* A of the original operator B is given through formulas (7.39) and (7.41), i.e.,

$$Au = \sum_{n=1}^{\infty} \lambda_n \langle u_n | u \rangle u_n, \quad \forall u \in \mathrm{Dom}(A), \tag{7.46}$$

where

$$u \in \mathrm{Dom}(A) \Longleftrightarrow \sum_{n=1}^{\infty} \lambda_n \langle u_n | u \rangle u_n \text{is convergent} \Longleftrightarrow \sum_{n=1}^{\infty} |\lambda_n \langle u_n | u \rangle|^2 < \infty.$$

The preceding considerations motivate the appearance of the Friedrichs extension in a quite natural way. However, the general theory of the Friedrichs extension is independent of Fourier series expansions. The basic idea is the following. We are given a linear, symmetric, *strongly monotone* operator $B : \mathrm{Dom}(B) \subset \mathcal{X} \to \mathcal{X}$ on the real Hilbert space $\mathcal{X}$, i.e.,

$$\langle Bu | u \rangle \geq c \, \|u\|^2 \quad \text{for all} u \in \mathrm{Dom}(B) \text{and fixed} c > 0.$$

We first construct the so-called energetic extension

$$B_E : \mathcal{X}_E \to \mathcal{X}_E^*$$

of the operator B, where

$$\mathrm{Dom}(B) \subset \mathcal{X}_E^* \subset \mathcal{X} \subset \mathcal{X}_E^*, \tag{7.47}$$

and B_E is the *duality map* of $\mathcal{X}_E$. Then, the *Friedrichs extension*

$$A : \mathrm{Dom}(A) \subset \mathcal{X} \to \mathcal{X}$$

of B is an appropriate *restriction* of B_E, namely, we get

$$Au \triangleq B_E u, \quad \forall u \in \mathrm{Dom}(A),$$

where $\mathrm{Dom}(A) \triangleq \{u \in \mathcal{X}_E : B_E u \in \mathcal{X}\}$. This construction guarantees automatically that the Friedrichs extension $A : \mathrm{Dom}(A) \subset \mathcal{X} \to \mathcal{X}$ is *bijective*, since the duality map B_E is bijective. That is, the equation $Au = f, u \in \mathrm{Dom}(A)$, has a unique solution for each given $f \in \mathcal{X}$.

For brevity we write

$$B \subset A \subset B_E, \tag{7.48}$$

i.e., A is an extension of B and in turn, B_E is an extension of A.

In terms of Fourier series, the energetic space $\mathcal{X}_E$ of the symmetric operator $B : \mathrm{Dom}(B) \subset \mathcal{X} \to \mathcal{X}$ from (**??**) is given through

$$\mathcal{X}_E = \left\{ u \in \mathcal{X} : \sum_{n=1}^{\infty} \lambda_n \, |\langle u_n | u \rangle|^2 < \infty \right\}.$$

If the operator B corresponds to the classic boundary-value problem (7.43), then

$$\mathcal{X}_E = \mathrm{W}_0^{1,2}(\Omega),$$

i.e., the energetic space is a *Soblev space*.

The *compactness* of the embedding

$$\mathrm{W}_0^{1,2} \subset L^2(\Omega)$$

plays a fundamental role. This is the famous *Rellich compactness theorem*. In fact, this compact embedding guarantees that the Laplacian with zero boundary conditions possesses a *complete* orthonormal system of eigenfunctions in the Hilbert space $L^2(\Omega)$, i.e., series (**??**) is convergent. This result will be critically used in order to solve the heat equation and wave equations. Our approach justifies the classic Fourier method of physicists.

The Friedrichs extension represents the functional analytic core of mathematical physics.

This approach is closely related to the fundamental physical concept of *energy*. In quantum mechanics, physical states correspond to unit vectors in a Hilbert space and the physical quantities (e.g., energy, momentum, and so on) correspond to self-adjoit operators.

Moreover, we shall show that

Self-adjoint operators are closely related to both orthogonality and generalized derivatives.

7.6.2 Extensions and Embeddings

Definition 197 (Extension of Operator)**.** *Let $A : \mathrm{Dom}(A) \subset \mathcal{X} \to \mathcal{Y}$ and $B : \mathrm{Dom}(B) \subset \mathcal{X} \to \mathcal{Y}$ be operators, where $\mathcal{X}$ and $\mathcal{Y}$ are linear spaces over $\mathbb{F}$. We write*

$$B \subset A$$

if and only if

$$A = Bu, \quad \forall u \in \mathrm{Dom}(B), \mathrm{Dom}(B) \subset \mathrm{Dom}(A).$$

In this case, we say that the operator A is an EXTENSTION *of the operator B.*

Obviously, $A = B$ iff $B \subset A$ and $A \subset B$.

Definition 198 (Embedding)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be normed spaces over $\mathbb{F}$.*

① *We say that the embedding "$\mathcal{X} \subset \mathcal{Y}$" is* CONTINUOUS *iff there exists an operator*

$$j : \mathcal{X} \to \mathcal{Y} \tag{7.49}$$

that is linear, CONTINUOUS*, and injective.*

② *The embedding "$\mathcal{X} \subset \mathcal{Y}$" is called* COMPACT *iff the operator $j : \mathcal{X} \to \mathcal{Y}$ is linear,* COMPACT*, and injective.*

Let $\mathcal{X}$ be a subset of $\mathcal{Y}$, i.e., $\mathcal{X} \subset \mathcal{Y}$. Then we set $j(u) \triangleq u$ for all $u \in \mathcal{X}$. In terms of sequences, we have the following:

(a) The embedding $\mathcal{X} \subset \mathcal{Y}$ is CONTINUOUS iff, as $n \to \infty$,

$$u_n \to u \quad \text{in } \mathcal{X} \Longrightarrow u_n \to u \quad \text{in } \mathcal{Y}.$$

(b) The embedding $\mathcal{X} \subset \mathcal{Y}$ is COMPACT if it is continuous and each bounded sequence (u_n) has a subsequence that converges in $\mathcal{Y}$, i.e., as $n' \to \infty$,

$$u_{n'} \to v \quad \text{in} \quad \mathcal{Y}.$$

In the general case of Definition 198, we may identify u with $j(u)$. This makes sense since $j : \mathcal{X} \to \mathcal{Y}$ is injective. In this sense, we may regard the space $\mathcal{X}$ as a subset of $\mathcal{Y}$, and we may write $\mathcal{X} \subset \mathcal{Y}$ instead of "$\mathcal{X} \subset \mathcal{Y}$" for brevity.

Example. Let Ω be a nonempty bounded open set in $\mathbb{R}^n$, $n \geq 1$. Then, the following hold true:

(i) The embedding $C(\overline{\Omega}) \subset L^2(\Omega)$ is continuous.

(ii) The embedding $\mathrm{W}_0^{1,2}(\Omega) \subset L^2(\Omega)$ is compact.

This is the prototype for *embedding theorems*, which play a fundamental role in modern analysis.

7.6.3 Self-Adjoint Operators

While working with operators $A : \mathrm{Dom}(A) \subset \mathcal{X} \to \mathcal{X}$ that are not defined on the total space, observe carefully the specific form of the domain of definition $\mathrm{Dom}(A)$ of A.

A. Definitions and Proposiitons

The definition of the adjoint operator $A^\dagger$ is based on the following formula

$$\boxed{\langle Au|v\rangle = \langle u|A^\dagger v\rangle, \quad \forall u \in \mathrm{Dom}(A) \subset \mathcal{X}, \forall v \in \mathrm{Dom}(A^\dagger) \subset \mathcal{X}.} \tag{7.50}$$

The self-adjoint operator is also called Hermitian operator.

Definition 199. *Let $A : \mathrm{Dom}(A) \subset \mathcal{X} \to \mathcal{X}$ be a linear operator, where $\mathrm{Dom}(A)$ is* DENSE *in the Hilbert space $\mathcal{X}$ over $\mathbb{F}$. By definition,*

$$v \in \mathrm{Dom}(A^\dagger) \subset \mathcal{X}$$

iff there exists an element $w \in \mathcal{X}$ such that

$$\langle Au|v\rangle = \langle u|w\rangle, \quad \forall u \in \mathrm{Dom}(A).$$

Furthermore, we set $A^\dagger v \triangleq w$. This way we obtain the ADJOINT OPERATOR

$$A^\dagger : \mathrm{Dom}(A^\dagger) \subset \mathcal{X} \to \mathcal{X}.$$

We have to show that this definition makes sense. In fact, suppose that relation (7.50) also holds if we replace w with w_1. Then

$$\langle u|w - w_1\rangle = 0, \quad \forall u \in \mathrm{Dom}(A).$$

Since $\mathrm{Dom}(A)$ is dense in $\mathcal{X}$, we get $w = w_1$.

Proposition 200. *Let $A : \mathrm{Dom}(A) \subset \mathcal{X} \to \mathcal{X}$ and $B : \mathrm{Dom}(B) \in \mathcal{X} \to \mathcal{X}$ be linear operators, where $\mathrm{Dom}(A)$ and $\mathrm{Dom}(B)$ are dense in the Hilbert space $\mathcal{X}$ over $\mathcal{F}$. Then, the following hold true:*

① *The adjoint operator $A^\dagger : \mathrm{Dom}(A^\dagger) \subset \mathcal{X} \to \mathcal{X}$ is linear.*

② *For each $\alpha \in \mathbb{F}$,*

$$(\alpha A)^\dagger = \overline{\alpha} A^\dagger. \tag{7.51}$$

③ $A \subset B$ *implies* $B^\dagger \subset A^\dagger$.

Consequently, if $\mathrm{Dom}(A^\dagger)$ *is* DENSE *in* $\mathcal{X}$*, then the operator* $(A^\dagger)^\dagger$ *exists. In this case, we set*

$$A^{\dagger\dagger} \triangleq (A^\dagger)^\dagger.$$

PROOF

- Ad ①. Let $\alpha_1, \alpha_2 \in \mathbb{F}$. If $\langle Au|v_j\rangle = \langle u|w_j\rangle$ for all $u \in \mathrm{Dom}(A), j = 1, 2$, then

  $$\langle Au|\alpha_1 v_1 + \alpha_2 v_2\rangle = \alpha_1\langle Au|v_1\rangle + \alpha_2\langle Au|v_2\rangle = \langle u|\alpha_1 w_1 + \alpha_2 w_2\rangle, \quad \forall u \in \mathrm{Dom}(A).$$

  Thus, $v_j \in \mathrm{Dom}(A^\dagger)$ for $j = 1, 2$ implies $\alpha_1 v_1 + \alpha_2 v_2 \in \mathrm{Dom}(A^\dagger)$ and

  $$A^\dagger(\alpha_1 v_1 + \alpha_2 v_2) = \alpha_1 A^\dagger v_1 + \alpha_2 A^\dagger v_2.$$

- Ad ②. For $\alpha = 0$, the assertation is trivial. If $\alpha \neq 0$, then relation (7.51) follows from

  $$\langle Au|v\rangle = \langle u|w\rangle \Longleftrightarrow \langle \alpha Au|v\rangle = \langle u|\overline{\alpha}w\rangle.$$

- Ad ③. Let $A \subset B$. It follows from

  $$\langle Bu|v\rangle = \langle u|B^\dagger v\rangle, \quad \forall u \in \mathrm{Dom}(B), v \in \mathrm{Dom}(B^\dagger)$$

  that $\langle Au|v\rangle = \langle u|B^\dagger v\rangle$ for all $u \in \mathrm{Dom}(A)$, and hence $A^\dagger v = B^\dagger v$ for all $v \in \mathrm{Dom}(B)$. This implies $B^\dagger \subset A^\dagger$.

Definition 201. *Let* $A : \mathrm{Dom}(A) \subset \mathcal{X} \to \mathcal{X}$ *be a linear operator, where* $\mathrm{Dom}(A)$ *is dense in the Hilbert space* $\mathcal{X}$ *over* $\mathbb{F}$.

① A *is called* SYMMETRIC *iff* $A \subset A^\dagger$*, i.e.,*

$$\forall u, v \in \mathrm{Dom}(A), \quad \langle Au|v\rangle = \langle u|Av\rangle.$$

② A *is called* SELF-ADJOINT *iff* $A = A^\dagger$.

③ A *is called* SKEW-SYMMETRIC *iff* $A \subset -A^\dagger$*, i.e.,*

$$\forall u, v \in \mathrm{Dom}(A), \quad \langle Au|v\rangle = -\langle u|Av\rangle.$$

④ A *is called* SKEW-ADJOINT *iff* $A = -A^\dagger$.

Proposition 202. *Let the operator* $A : \mathcal{X} \to \mathcal{X}$ *be linear and continuous on the Hilbert space* $\mathcal{X}$ *over* $\mathbb{F}$*. Then, the adjoint operator*

$$A^\dagger : \mathcal{X} \to \mathcal{X}$$

is also linear and continuous. In addition, $\|A\| = \left\|A^\dagger\right\|$*. Moreover,* $A^{\dagger\dagger} = A$.

PROOF

- Let $v \in \mathcal{X}$. Set

  $$f(u) \triangleq \langle v|Au\rangle, \quad \forall u \in \mathcal{X}.$$

  By the Cauchy-Schwarz inequality,

  $$|f(u)| \leq \|Au\| \cdot \|v\| \leq \|A\| \cdot \|u\| \cdot \|v\|, \quad \forall u \in \mathcal{X}.$$

  Hence the linear functional $f : \mathcal{X} \to \mathbb{F}$ is continuous with $\|f\| \leq \|A\|\,\|v\|$.

- By the Riesz representation theorem, there exists an element $w \in \mathcal{X}$ such that

$$f(u) = \langle w|u\rangle, \quad u \in \mathcal{X},$$

and $\|w\| = \|f\|$. Hence $\langle Au|v\rangle = \langle u|w\rangle$ for all $u \in \mathcal{X}$. This implies

$$A^\dagger v = w,$$

and $\left\|A^\dagger v\right\| \leq \|A\| \, \|v\|$ for all $v \in \mathcal{X}$. Therefore,

$$\left\|A^\dagger\right\| \leq \|A\| . \tag{7.52}$$

- It follows from $\langle Au|v\rangle = \langle u|A^\dagger v\rangle$ that

$$\langle A^\dagger v|u\rangle = \langle v|Au\rangle, \quad \forall u, v \in \mathcal{X}.$$

Hence $(A^\dagger)^\dagger = A$. Replacing A with $A^\dagger$ in (7.52), we get $\|A\| = \left\|(A^\dagger)^\dagger\right\| \leq \left\|A^\dagger\right\|$. This implies $\|A\| = \left\|A^\dagger\right\|$. ■

Proposition 203. *Let $A : \mathrm{Dom}(A) \subset \mathcal{X} \to \mathcal{X}$ be a linear operator on the Hilbert space $\mathcal{X}$ over $\mathbb{F}$ over $\mathbb{F}$ such that $\mathrm{Dom}(A)$ is dense in $\mathcal{X}$. Then, the following hold true:*

① *A is self-adjoint iff A is symmetric and*

$$\langle Au|v\rangle = \langle u|w\rangle, \quad \textit{for all } u \in \mathrm{Dom}(A) \textit{ and fixed } v, w \in \mathcal{X} \tag{7.53}$$

implies that $v \in \mathrm{Dom}(A)$ and $w = Av$.

② *A is skew-adjoint iff A is skew-symmetric and*

$$\langle Au|v\rangle = -\langle u|w\rangle, \quad \textit{for all } u \in \mathrm{Dom}(A) \textit{ and fixed } v, w \in \mathcal{X} \tag{7.54}$$

implies that $v \in \mathrm{Dom}(A)$ and $w = Av$.

③ *Let $\mathbb{F} = \mathbb{C}$ and $\alpha \in \mathbb{R}$ with $\alpha \neq 0$. Then,*

$$A \textit{ is skew-symmetric} \iff \mathrm{i}\alpha A \textit{ is symmetric}$$
$$A \textit{ is skew-adjoint} \iff \mathrm{i}\alpha A \textit{ is self-adjoint}$$

where $\mathrm{i} = \sqrt{-1}$.

Proof

- Ad ①. Observe that $A = A^\dagger$ iff $A \subset A^\dagger$ and $A^\dagger \subset A$.
- Ad ②. Use $A = -A^\dagger$ iff $A \subset -A$ and $-A^\dagger \subset A$.
- Ad ③. Since $(\mathrm{i}\alpha A)^\dagger = -\mathrm{i}\alpha A^\dagger$, we get

$$A \subset -A^\dagger \iff \mathrm{i}\alpha A \subset (\mathrm{i}\alpha A)^\dagger$$

and

$$A = -A^\dagger \iff \mathrm{i}\alpha A = (\mathrm{i}\alpha A)^\dagger.$$

Corollary 204 (Maximally Self-Adjoint and Skew-Adjoint)**.**

① *Each self-adjoint linear operator $A : \mathrm{Dom}(A) \subset \mathcal{X} \to \mathcal{X}$ on the Hilbert space $\mathcal{X}$ over $\mathbb{F}$ is maximally symmetric, i.e., by definition, if we have*

$$A \subset S$$

② *Each skew-adjoint linear operator* $A : \mathrm{Dom}(A) \subset \mathcal{X} \to \mathcal{X}$ *is maximally skew-symmetric.*

PROOF

- Ad ①. It follows from $A \subset S$ that $S^\dagger \subset A^\dagger$. Since $A = A^\dagger$ and $S \subset S^\dagger$, we get $S \subset A$. Thus, $S = A$.
- Ad ②. If $A \subset S$ with $A = -A^\dagger$ and $S \subset -S^\dagger$, then $S^\dagger \subset A^\dagger$, and hence $S \subset A$. Thus $A = S$.

B. Examples

Standard Example (Integral Operator). Let $K : [a,b] \times [a,b] \to \mathbb{R}$ be a continuous function, where $-\infty < a < b < \infty$. Define

$$(Au)(x) \triangleq \int_a^b K(x,y)u(y)\,\mathrm{d}\,y, \quad \forall x \in [a,b], \tag{7.55}$$

and set $\mathcal{X} \triangleq L^2(a,b)$. Then, the following are met:

① The operator $A : \mathcal{X} \to \mathcal{X}$ is linear and compact.

② The ADJOINT operator $A^\dagger : \mathcal{X} \to \mathcal{X}$ is given through

$$(A^\dagger u)(x) = \int_a^b K(y,x)u(y)\,\mathrm{d}\,y, \quad \forall x \in [a,b]. \tag{7.56}$$

The operator $A^\dagger : \mathcal{X} \to \mathcal{X}$ is linear and compact.

③ If K is symmetric, i.e., $K(x,y) = K(y,x)$ for all $x, y \in [a,b]$, then the operator $A : \mathcal{X} \to \mathcal{X}$ is SELF-ADJOINT.

PROOF

① Ad ①. This follows from Lemma 190.

② We are given $v \in \mathcal{X}$. Set

$$w(x) \triangleq \int_a^b K(y,x)v(y)\,\mathrm{d}\,y, \quad \forall x \in [a,b].$$

As in the proof of Lemma 190, it follows from the Tonelli Theorem that

$$\begin{aligned}\langle Au|v\rangle &= \int_a^b \left[\int_a^b K(x,y)u(y)\,\mathrm{d}\,y\right] v(x)\,\mathrm{d}\,x \\ &= \int_a^b \left[\int_a^b K(x,y)v(x)\,\mathrm{d}\,x\right] u(y)\,\mathrm{d}\,y \\ &= \int_a^b w(y)u(y)\,\mathrm{d}\,y = \langle u|w\rangle, \quad \forall u \in \mathcal{X}.\end{aligned}$$

Hence $A^\dagger v = w$. This yields (7.56). By (7.56) and Lemma 190, the operator $A^\dagger : \mathcal{X} \to \mathcal{X}$ is linear and compact.

④ Ad ④. If K is symmetric, then $A = A^\dagger$, by (7.55) and (7.56). ■.

Standard Example (Differential Operator). Let $\mathcal{X} =\triangleq L^2(\mathbb{R})$. Define

$$(Au)(x) \triangleq \mathrm{D}\,u(x) = u'(x), \quad \forall x \in \mathbb{R},$$

where $\mathrm{Dom}(A) = \{u \in \mathcal{X} : u' \in \mathcal{X}\}$. Here, the derivative u' is to be understood in the generalized sense. Then, the following hold true:

① The operator $A : \mathrm{Dom}(A) \subset \mathcal{X} \to \mathcal{X}$ is SKEW-ADJOINT.

② For each $\alpha \in \mathbb{R}$, the operator $\mathrm{i}\alpha A$ is SELF-ADJOINT.

On the contratry, if we consider the classic differential operator

$$(Bu)(x) \triangleq u'(x), \quad \forall x \in \mathbb{R},$$

where $\mathrm{Dom}(B) \triangleq \{u \in C^1(\mathbb{R}) : u, u' \in \mathcal{X}\}$. Then

③ The operator $B : \mathrm{Dom}(B) \subset \mathcal{X} \to \mathcal{X}$ is skew-symmetric but not skew-adjoint.

④ For each $\alpha \in \mathbb{R}$ with $\alpha \neq 0$, the operator $\mathrm{i}\alpha B$ is symmetric but not self-adjoint.

PROOF

- Ad ①. We have $u \in \mathrm{Dom}(A)$ iff $u \in \mathcal{X}$ and there is a function $w \in \mathcal{X}$ such that

$$\int_{\mathbb{R}} u(x)\psi'(x)\,\mathrm{d}\,x = -\int_{\mathbb{R}} w(x)\psi(x)\,\mathrm{d}\,x, \quad \forall \psi \in C_0^\infty(\mathbb{R}). \tag{7.57}$$

  If this holds true, then we get $w = u'$ in the generalized sense.

  – Step 1: Approximation. Let $v \in \mathrm{Dom}(A)$. We want to show that there exists a sequence $\{v_n\} \subset C^\infty(\mathbb{R})$ such that, as $n \to \infty$,

$$v_n \to v \quad \text{in} \quad L^2(\mathbb{R}), \qquad v_n' \to v' \quad \text{in} \quad L^2(\mathbb{R}).$$

  We set

$$v_n(x) \triangleq \int_{\mathbb{R}} j_{\frac{1}{n}}(x-y)v(y)\,\mathrm{d}\,y, \quad n = 1, 2, \cdots$$

  where $j_\varepsilon(x) = \frac{1}{\varepsilon^n} j\left(\frac{x}{\varepsilon}\right)$ is the smoothing function defined by equatons (4.9) and (4.10). Then, $v_n \in C_0^\infty(\mathbb{R})$ for all n, and we get $v_n \to v$ in $L^2(\mathbb{R})$ as $n \to \infty$.
  Since for fixed $x \in \mathbb{R}$, the function $y \mapsto j_{\frac{1}{n}}(x-y)$ belongs to the space $C_0^\infty(\mathbb{R})$, it follows from the definition of the generalized derivative that

$$\begin{aligned} v_n'(x) &= \int_{\mathbb{R}} \left[\frac{\mathrm{d}}{\mathrm{d}\,x} j_{\frac{1}{n}}(x-y)\right] v(y)\,\mathrm{d}\,y \\ &= -\int_{\mathbb{R}} \left[\frac{\mathrm{d}}{\mathrm{d}\,y} j_{\frac{1}{n}}(x-y)\right] v(y)\,\mathrm{d}\,y \\ &= \int_{\mathbb{R}} j_{\frac{1}{n}}(x-y)v'(y)\,\mathrm{d}\,y. \end{aligned}$$

  Hence $v_n' \to v'$ in $L^2(\mathbb{R})$ as $n \to \mathbb{R}$.
  Since the function $j_{\frac{1}{n}}$ is real, we also get

$$\overline{v_n} \to \overline{v} \quad \text{and} \quad \overline{(v_n')} \to \overline{(v')} \quad \text{in} \quad L^2(\mathbb{R}) \quad \text{as} \quad n \to \mathbb{R}. \tag{7.58}$$

  – Step 2: We want to show that the operator $A = \mathrm{D}$ is skew-symmetric. Replacing v with $\overline{v_n}$ in (7.57) and letting $n \to \infty$, we get

$$\int_{\mathbb{R}} u(x)\overline{v'(x)}\,\mathrm{d}\,x = -\int_{\mathbb{R}} u'(x)\overline{v(x)}\,\mathrm{d}\,x, \quad \forall u, v \in \mathrm{Dom}(A). \tag{7.59}$$

  Hence

$$\langle Av|u\rangle = -\langle v|Au\rangle, \quad \forall u, v \in \mathrm{Dom}(A).$$

– Step 3: We prove that the operator A is skew-adjoint. In fact, it follows from

$$\langle \mathsf{A}v|u\rangle = -\langle v|w\rangle, \quad \text{for all } v \in \mathrm{Dom}(\mathsf{A}) \text{ and fixed } u, w \in \mathcal{X}$$

that

$$\int_{\mathbb{R}} \overline{v'} u \,\mathrm{d}\, x = -\int_{\mathbb{R}} \overline{v} w \,\mathrm{d}\, x, \quad \forall v \in C_0^\infty(\mathbb{R}).$$

Thus, we get $\mathrm{D}\, u = u' = w$ in the generalized sense, i.e., $w = \mathsf{A}u$. By Propersition 203, A is skew-adjoint.

- Ad ②. This follows from Proposition 203③. ■
- Ad ③. By Step 2, B is skew-symmetric. Set

$$u(x) \triangleq |x|\,, \quad w(x) \triangleq \begin{cases} u'(x), & \text{if } x \neq 0. \\ 0, & \text{if } x = 0. \end{cases}$$

In the generalized sense, we have $w = u'$.
However, $u \in \mathrm{Dom}(\mathsf{B})$ since u is not in C^1. Since

$$\mathsf{A}u = w,$$

the operator A is a proper skew-symmetric extension of B. Thus, B is not skew-adjoint by Corollary 204.

- Ad ④. This follows from ① and Proposition 203③.

Stantdard Example (The multiplication operator). Let $\mathcal{X} \triangleq L^2(\mathbb{R})$. Define

$$(\mathsf{M}u)(x) \triangleq xu(x), \quad \forall x \in \mathbb{R},$$

where $\mathrm{Dom}(\mathsf{M}) \triangleq \{u \in \mathcal{X} : \mathsf{M}u \in \mathcal{X}\}$. Then, the operator $\mathsf{M} : \mathrm{Dom}(\mathsf{M}) \subset \mathcal{X} \to \mathcal{X}$ is self-adjoint.
PROOF

- For all $u, v \in \mathrm{Dom}(\mathsf{M})$,

$$\langle \mathsf{M}u|v\rangle = \int_{\mathbb{R}} \overline{xu(x)} v(x) \,\mathrm{d}\, x = \int_{\mathbb{R}} \overline{u(x)}\,[xv(x)] \,\mathrm{d}\, x = \langle u|\mathsf{M}v\rangle.$$

Hence M is symmetric.

- Moreover, it follows from

$$\langle \mathsf{M}u|v\rangle = \langle u|v\rangle, \quad \forall u \in \mathrm{Dom}(\mathsf{M}) \text{and fixed } v, w \in \mathcal{X}$$

that

$$\int_{\mathbb{R}} \overline{u(x)}\,[xv(x)] \,\mathrm{d}\, x = \int_{\mathbb{R}} \overline{u(x)} w(x) \,\mathrm{d}\, x, \quad \forall u \in C_0^\infty(\mathbb{R}).$$

Hence $w(x) = xv(x)$ for almost all $x \in \mathbb{R}$, i.e., $w = \mathsf{M}v$. Thus, the multiplication operator M is self-adjoint by Proposition 203. ■

Remarks

- This example shows that the notation of the generalized derivative is quite natural from the operator theory viewpoint.
- For $\mathrm{i}\alpha\mathsf{A} = -\mathrm{i}\hbar\frac{\mathrm{d}}{\mathrm{d}\, x}$, it is the momentum operator in quantum mechanics. The more general case is $\mathsf{p} = -\mathrm{i}\hbar\nabla$, which is an Hermitian operator and self-adjoit.
- The multiplication operator $\mathsf{M} = x$, or more generally, $\mathsf{M} = \mathsf{x} = \boldsymbol{x}$, is the postion operator in quantum mechanics.

C. Self-adjoint Operators and Orthogonal Projections

Let $\mathcal{L}$ be a closed linear subspace of the Hilbert space $\mathcal{X}$ over $\mathbb{F}$. For each $u \in \mathcal{X}$, there exists the unique decomposition

$$u = v + w, \quad v \in \mathcal{L}, w \in \mathcal{L}^{\perp}. \tag{7.60}$$

Definition 205. *For the orthogonal decomposition $u = v + w, v \in \mathcal{L}, w \in \mathcal{L}^{\perp}$ in the Hilbert space $\mathcal{X}$ such that linear subspace $\mathcal{L} \subset \mathcal{X}$. The operator*

$$P u \triangleq v$$

is called the ORTHOGONAL PROJECTION *from $\mathcal{X}$ to $\mathcal{L}$.*

Proposition 206. *Let $\mathcal{X}$ be a Hilbert space over $\mathbb{F}$. Then*

① *The orthogonal projection $P : \mathcal{X} \to \mathcal{L}$ from $\mathcal{X}$ onto the closed linear subspace $\mathcal{L}$ of $\mathcal{X}$ is linear, continuous and self-adjoint and $P^2 = P$ (i.e., indempotent). If $\mathcal{L} \neq \{0\}$, then $\|P\| = 1$.*

② *Conversely, let $P : \mathcal{X} \to \mathcal{X}$ be a linear continuous self-adjoint operator with $P^2 = P$. Then, P is the orthogonal projection from $\mathcal{X}$ onto the closed linear subspace $P(\mathcal{X})$.*

PROOF

- Ad ①.
  - By the Pythogorean theorem, we have
    $$\|u\|^2 = \|v\|^2 + \|w\|^2 .$$
    Hence $\|Pu\| = \|v\| \leq \|u\|$ for all $u \in \mathcal{X}$. Moreover, if $u \in \mathcal{L}$, then $Pu = u$. Thus $\|P\| = 1$ according to the definition of $\|P\|$.
  - Let
    $$u_j = v_j + w_j, \quad v_j \in \mathcal{L}, w_j \in \mathcal{L}^{\perp}, j = 1, 2.$$
    Then, $\langle v_j | w_j \rangle = 0$ for $j, k = 1, 2$. Hence
    $$\langle v_1 | u_2 \rangle = \langle u_1 | v_2 \rangle = \langle v_1 | v_2 \rangle.$$
    This implies
    $$\langle Pu_1 | u_2 \rangle = \langle v_1 | u_2 \rangle = \langle v_1 | v_2 \rangle = \langle u_1 | Pu_2 \rangle, \quad \forall u_1, u_2 \in \mathcal{X}.$$
    Hence $P = P^{\dagger}$, i.e., P is SELF-ADJOINT.
  - If $v \in \mathcal{L}$, then $v = v + 0$, where $v \in \mathcal{L}, 0 \in \mathcal{L}^{\perp}$. Hence $Pv = v$. By the unique decomposition (7.60),
    $$P^2 u = Pv = v = Pu, \quad \forall u \in \mathcal{X}.$$
    Therefore, $P^2 = P$.
- Ad ②. Set $\mathcal{L} = P(\mathcal{X})$. Since P is linear, $\mathcal{L}$ is a linear subspace of $\mathcal{X}$. It follows from $P^2 = P$ that $\mathcal{L}$ is closed. In fact, let (v_n) be a sequence in $\mathcal{L}$ such that $v_n \to v$ as $n \to \infty$. Then, $v_n = Pu_n$ for some u_n and $Pv_n = P(Pu_n) = (P^2)u_n = Pu_n = v_n$. Hence
  $$v = \lim_{n \to \infty} v_n = \lim_{n \to \infty} Pv_n = Pv,$$
  i.e., $v \in \mathcal{L}$. Furthermore, since P is self-adjoint and $P^2 = P$, we get
  $$\begin{aligned} \langle Px | (\mathbb{1} - P) y \rangle &= \langle Px | y \rangle - \langle Px | Py \rangle \\ &= \langle Px | y \rangle - \langle P^2 x | y \rangle = 0, \quad \forall x, y \in \mathcal{L}. \end{aligned}$$

Hence $(\mathbb{1}-P)u \in \mathcal{L}^\perp$ for all $u \in \mathcal{X}$. Thus, it follows from

$$u = Pu + (\mathbb{1}-P)u, \quad Pu \in \mathcal{L}, \quad P_\perp = (\mathbb{1}-P)u \in \mathcal{L}^\perp,$$

which shows that P is the orthogonal projection from $\mathcal{X}$ onto $\mathcal{L}$ and $P_\perp$ is the orthogonal projection from $\mathcal{X}$ onto $\mathcal{L}^\perp$

Proposition 207. *Let $A : \mathrm{Dom}(A) \subset \mathcal{X} \to \mathcal{X}$ be a linear symmetric operator on the Hilbert space $\mathcal{X}$, where $\mathrm{Range}(A)$ is dense in $\mathcal{X}$. Then,*

$$(A^{-1})^\dagger = (A^\dagger)^{-1},$$

where all the appearing inverse and adjoint operator exist. If, in addition, A is self-adjoint, then so is A^{-1}.

PROOF

- The operator A is injective. In fact, $Au = 0$ implies

  $$\langle u|Av\rangle = \langle Au|v\rangle = 0, \quad \forall v \in \mathrm{Dom}(A).$$

  Since $\mathrm{Range}(A)$ is dense in $\mathcal{X}$, $u = 0$.

- The operator $A^\dagger$ is also injective. In fact, if $A^\dagger u = 0$, then it follows from

  $$\langle Av|u\rangle = \langle v|A^\dagger u\rangle = 0, \quad \forall v \in \mathrm{Dom}(A)$$

  that $u = 0$.

- Conversely, the inverse operators A^{-1} and $(A^\dagger)^{-1}$ exist. Since $\mathrm{Dom}(A^{-1}) = \mathrm{Range}(A)$ and $\mathrm{Range}(A)$ is dense in $\mathcal{X}$, the adjoint operator $(A^{-1})^\dagger$ exists.

- Set $B \triangleq (A\text{PROOF}^{-1})^\dagger$. We have

  $$\langle u|v\rangle = \langle A^{-1}Au|v\rangle = \langle Au|Bv\rangle, \quad \forall u \in \mathrm{Dom}(A), v \in \mathrm{Dom}(B), \tag{7.61}$$

  and

  $$\langle z|w\rangle = \langle AA^{-1}z|w\rangle = \langle A^{-1}z|A^\dagger w\rangle, \quad \forall w \in \mathrm{Dom}(A^\dagger), z \in \mathrm{Dom}(A^{-1}). \tag{7.62}$$

  It follows from (7.61) that $Bv \in \mathrm{Dom}(A^\dagger)$ and

  $$\langle u|v\rangle = \langle u|A^\dagger Bv\rangle, \quad \forall u \in \mathrm{Dom}(A), v \in \mathrm{Dom}(B).$$

  Since $\mathrm{Dom}(A)$ is dense in $\mathcal{X}$, this implies

  $$A^\dagger Bv = v, \quad \forall v \in \mathrm{Dom}(B). \tag{7.63}$$

  Analogously, it follows from (7.62) that

  $$BA^\dagger w = w, \quad \forall w \in \mathrm{Dom}(A^\dagger). \tag{7.64}$$

  Hence $A^\dagger B = \mathbb{1}$ and $BA^\dagger = \mathbb{1}$, which implies that $B = (A^\dagger)^{-1}$ and $(A^{-1})^\dagger = (A^\dagger)^{-1}$.

- If A is self-adjoint, then $A = A^\dagger$, and hence $(A^{-1})^\dagger = (A^\dagger)^{-1} = A^{-1}$, i.e., A^{-1} is also self-adjoint. ■

Proposition 208. *Let $A : \mathrm{Dom}(A) \subset \mathcal{X} \to \mathcal{X}$ be a linear operator where $\mathrm{Dom}(A)$ is dense in the Hilbert space $\mathcal{X}$ over $\mathcal{F}$. Suppose that there exists a sequence $(u_n) \subset \mathrm{Dom}(A^\dagger)$ such that*

$$u_n \to u \quad \textit{and} \quad A^\dagger u_n \to v \quad \textit{in} \quad \mathcal{X} \quad \textit{as} \quad n \to \infty.$$

Then, $u \in \mathrm{Dom}(A^\dagger)$ and $A^\dagger u = v$.

Proof

- Letting $n \to \infty$, it follows from

$$\langle Aw|u_n\rangle = \langle w|A^\dagger u_n\rangle, \quad \forall w \in \mathrm{Dom}(A)$$

that $\langle Aw|u\rangle = \langle w|v\rangle$ for all $w \in \mathrm{Dom}(A)$. Hence $A^\dagger u = v$. ■

- This implies that if A is self-adjoint or skew-adjoint, then

$$u_n \to u \quad \text{and} \quad Au_n \to v \quad \text{in} \quad \mathcal{X} \quad \text{as} \quad n \to \infty$$

imply $u \in \mathrm{Dom}(A)$ and $Au = v$.

Proposition 209. *For a linear operator $U : \mathcal{X} \to \mathcal{X}$ on the Hilbert space $\mathcal{X}$ over $\mathbb{F}$, the following four conditions are mutaullu equivalent:*

① *U is unitary, i.e., U is surjective and $\langle Ux|Uy\rangle = \langle x|y\rangle$ for all $x, y \in \mathcal{X}$.*

② *$UU^\dagger = U^\dagger U = \mathbb{1}$.*

③ *U is bijective and $U^{-1} = U^\dagger$.*

④ *U is surjective and $\|Ux\| = \|x\|$ for all $x \in \mathcal{X}$.*

Proof

- ①⇒ ②. It follows from $\langle Ux|Uy\rangle = \langle x|y\rangle$ for all $x, y \in \mathcal{X}$ that

$$U^\dagger(Ux) = x, \quad \forall x \in \mathcal{X},$$

i.e., $U^\dagger U = \mathbb{1}$. Hence $UU^\dagger UU^{-1}x = x$ for all $x \in \mathcal{X}$, i.e., $UU^\dagger = \mathbb{1}$.

- ②⇒③. This is trivial.

- ②⇒①. From $UU^\dagger = \mathbb{1}$ if follows that $\mathrm{Dom}(U^\dagger) = \mathcal{X}$, and $U^\dagger U = \mathbb{1}$ implies

$$\langle Ux|Uy\rangle = \langle x|U^\dagger Uy\rangle = \langle x|y\rangle, \quad \forall x, y \in \mathcal{X}.$$

- ①⇒④. Observe that the inner product can be expressed by norms. For example, if $\mathcal{X}$ is a real Hilbert space, then

$$\langle x|y\rangle = \frac{1}{4}(\|x+y\|^2 - \|x-y\|^2), \quad \forall x, y \in \mathcal{X}.$$

Hence $\|Ux\| = \|x\|$ for all $x \in \mathcal{X}$ is equivalent to $\langle Ux|Uy\rangle = \langle x|y\rangle$ for all $x, y \in \mathcal{X}$.

7.7 Compact Operator

7.7.1 Definition and Properties

Definition 210 (Compact Operator)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ are normed linear space, $\mathsf{T} : \mathcal{X} \to \mathcal{Y}$ is a linear operator. For any bouned sequence $\{u_n\} \subset \mathcal{X}$, the sequence $\{\mathsf{T}u_n\}$ contains a convergent subsequence, then T is called a compact operator, or equivalently T is compact.*

Example 1 Suppose that $K(\cdot,\cdot) : [a,b] \times [a,b] \to \mathbb{F}$ is continuous, $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$, then the linear integral operator $\mathsf{A} : C[a,b] \to C[a,b]$ defined by

$$(\mathsf{A}\phi)(x) = \int_a^b K(x,y)\phi(y)\,\mathrm{d}\,y, \quad x \in [a,b]$$

is compact. Furthermore, if $K(\cdot,\cdot) \in L^2([a,b] \times [a,b])$, i.e.,

$$\int_a^b \int_a^b |K(x,y)|^2\,\mathrm{d}\,x\,\mathrm{d}\,y < \infty,$$

the operator $\mathsf{A} : L^2[a,b] \to L^2[a,b]$ is also compact.

Example 2 The operator $\mathsf{D} = -\mathrm{i}\frac{\mathrm{d}}{\mathrm{d}\,x}$ acts on $L^2[0,1]$. For the bounded sequence $\left\{u_n = \mathrm{e}^{\mathrm{i}nx}\right\} \subset L^2[0,1]$, the sequence $\left\{\mathsf{D}u_n = n\mathrm{e}^{\mathrm{i}nx}\right\}$ has no convergent subsequences. Actually, there are many non compact operators. Particularly, the identity operator is not compact.

There are some significant properties for compact operators

① *All of the linear operators defined on finite dimensional linear space are compact.*
In other words, each element $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ are compact. Therefore, the concept of compact operator is not necessary in finite dimensional vector space.

② *The sum of finite numbers of compact operators is also compact.*

③ *Any operator with finite rank is compact.*
Actually, an operator with finite rank is a sum of finite numbers of compact operator. Thus ③ is equivalent to ①.

④ *The composition of two compact operators is also compact.*

⑤ *The composition of a compact operator and a bounded compact operator is compact.*

⑥ *The adjoint operator of a compact operator is also compact.*

⑦ *The limit of a sequence of compact operators is also compact.*

Theorem 211 (Approximation with Compact Operators)**.** *Let $\mathcal{X}, \mathcal{Y}$ are normed linear space and $\mathsf{K} : \mathcal{X} \to \mathcal{Y}$. Suppose $\mathsf{K}_n : \mathcal{X} \to \mathcal{Y}$ is compact for any n and*

$$\lim_{n\to\infty} \|\mathsf{K}_n - \mathsf{K}\| = 0,$$

then K is compact.

Proof

- Let $\{f_n\} \subset \mathcal{X}$ be a bounded sequence such that $\|f_n\| \le C$. Since K_1 is compact, we can pick up a bounded subsequence $\left\{f_{n_1}^{(1)}\right\}$ such that $\left\{\mathsf{K}_1 f_{n_1}^{(1)}\right\}$ is convergent. Similarly, since K_2 is compact, we can pick up a bounded subsequence $\left\{f_{n_2}^{(2)}\right\}$ such that $\left\{\mathsf{K}_2 f_{n_2}^{(2)}\right\}$ is convergent. This process can be repeated any time. Since K_i is compact, we can pick up a bounded subsequence $\left\{f_{n_i}^{(i)}\right\}$ such that $\left\{\mathsf{K}_i f_{n_i}^{(i)}\right\}$ is convergent. Thus we can obtain a bounded sequence $\{\phi_m\}$ such that $\|\phi_m\| \le C$ and for any K_i, the sequence $\{\mathsf{K}_i\phi_m\}$ converges.

- We now consider

$$K\phi_n - K\phi_m = K\phi_n - K_i\phi_n + K_i\phi_n - K_i\phi_m + K_i\phi_m + K_i\phi_m - K\phi_m.$$

With the help of triangle inequality we immediately have

$$\begin{aligned}\|K\phi_n - K\phi_m\| &= \|K\phi_n - K_i\phi_n + K_i\phi_n - K_i\phi_m + K_i\phi_m + K_i\phi_m - K\phi_m\| \\ &\leq \|K\phi_n - K_i\phi_n\| + \|K_i(\phi_n - \phi_m)\| + \|K_i\phi_m - K\phi_m\| \\ &\leq \|K - K_i\|\,\|\phi_n\| + \|K_i(\phi_n - \phi_m)\| + \|K_i - K\|\,\|\phi_m\| \\ &\leq \|K - K_i\|\,\|\phi_n\| + \|K_i(\phi_n - \phi_m)\| + \|K_i - K\|\,\|\phi_m\| \\ &\leq 2C\,\|K - K_i\| + \|K_i(\phi_n - \phi_m)\|\end{aligned}$$

- $\|K_i - K\| \to 0$ implies that for any $\varepsilon > 0$ there exists a sufficnetly large i such that

$$\|K_i - K\| < \frac{\varepsilon}{4C}.$$

At the same time, since $\{K_i\phi_n\}$ converges, we have

$$\|K_i(\phi_n - \phi_m)\| < \frac{\varepsilon}{2}$$

for sufficiently large n and m. Hence,

$$\|K\phi_n - K\phi_m\| \leq 2C\,\|K - K_i\| + \|K_i(\phi_n - \phi_m)\| < 2C\frac{\varepsilon}{4C} + \frac{\varepsilon}{2} = \varepsilon$$

Therefore, the sequence $\{K\phi_n\}$ is convergent, which implies that the limit of the sequence $\{K_i\}$ is compact. ■

Theorem 211 tells us that if an operator can be approximated with a sequence of compact operators, it will be compact. The inverse also holds true: Any compact operator can be approximated with a sequence of compact operators.

7.7.2 Decomposition of Compact Operators

Theorem 212. *Let $\mathcal{X}$ be a Hilbert space with basis $\{\phi_n\}_n$ and $\mathcal{L} = \operatorname{span}\phi_0, \phi_1, \cdots, \phi_n$ be a subspace of $\mathcal{X}$.. Assume K is a compact operator acts on $\mathcal{X}$ and $P_n : \mathcal{X} \to \mathcal{L}$ is a projection operator. Let $K_n = P_nKP_n$, then $\|K_n - K\| \to 0$ as $n \to \infty$.*

Proof

- The operator $K - K_n$ can be decomposed as

$$\begin{aligned}K - K_n &= K - P_nKP_n \\ &= K - P_nK + P_nK - P_nKP_n \\ &= (\mathbb{1} - P_n)K + P_nK(\mathbb{1} - P_n)\end{aligned} \tag{7.65}$$

Thus we need to show that the righthand side in the previous equation should approach to zero as $n \to \infty$.

- Assume $\{f_n\}$ is a bounded sequence in $\mathcal{X}$ such that $\|f_n\| \leq C$. Then

$$\|(K - K_n)f_n\| \leq \|(\mathbb{1} - P_n)Kf_n\| + \|P_nK(\mathbb{1} - P_n)f_n\| \tag{7.66}$$

We will prove that the righthand side of (7.66) approaches to zero as $n \to \infty$.

- Let $g_n = \frac{f_n}{\|f_n\|}$, then $\{g_n\}$ is bounded and $\|g_n\| = 1$. Then (7.66) is equivalent to the following

$$\begin{aligned} \|(K - K_n)g_n\| &\leq \|(\mathbb{1} - P_n)Kg_n\| + \|P_n\| \, \|K(\mathbb{1} - P_n)g_n\| \\ &= \|(\mathbb{1} - P_n)Kg_n\| + \|K(\mathbb{1} - P_n)g_n\| \end{aligned} \tag{7.67}$$

- Since the operator $\mathbb{1} - P_n$ is bounded, then the sequence $\{h_n = (\mathbb{1} - P_n)g_n\}$ is also bounded. K is compact, which implies both $\{Kg_n\}$ and $\{Kh_n\}$ have convergent subsequences, say $\{K\hat{g}_n\}$ and $\left\{K\hat{h}_n\right\}$ with limits ξ and η in $\mathcal{L}$ respectively.

- For the subsequence $\{K\hat{g}_n\}$, we now estimate the value of $\|(\mathbb{1} - P_n)K\hat{g}_n\| + \|K(\mathbb{1} - P_n)\hat{g}_n\|$.

  – Firstly,

  $$(\mathbb{1} - P_n)K\hat{g}_n = (\mathbb{1} - P_n)(K\hat{g}_n - EP - thm - ComOPDecom\xi) + (\xi - P_n\xi).$$

  Since $K\hat{g}_n \to \xi$ and $P_n\xi \to \xi$, then $\|K\hat{g}_n - \xi\| \to 0$ and $\|P_n\xi - \xi\| \to 0$. Therefore,

  $$\|(\mathbb{1} - P_n)K\hat{g}_n\| \leq \|\mathbb{1} - P_n\| \, \|K\hat{g}_n - \xi\| + \|\xi - P_n\xi\| \to 0.$$

  – Secondly, with the help of $K(\mathbb{1} - P_n)\hat{g}_n \to \eta$, $(\mathbb{1} - P)^\dagger = \mathbb{1} - P$ and the Schwartz inequaltity, we can obtain

  $$\left|\langle(\mathbb{1} - P_n)K^\dagger\eta|\hat{g}_n\rangle\right| \leq \left\|(\mathbb{1} - P_n)K^\dagger\eta\right\| \cdot \|\hat{g}_n\| = \left\|(\mathbb{1} - P_n)K^\dagger\eta\right\|$$

  such that

  $$\begin{aligned} \|\eta\|^2 = \langle\eta|\eta\rangle &= \lim_{n\to\infty} |\langle\eta|K(\mathbb{1} - P_n)\hat{g}_n\rangle| \\ &= \lim_{n\to\infty} \left|\langle(\mathbb{1} - P_n)K^\dagger\eta|\hat{g}_n\rangle\right| \\ &\leq \lim_{n\to\infty} \left|\langle(\mathbb{1} - P_n)K^\dagger\eta|\hat{g}_n\rangle\right| \\ &\leq \lim_{n\to\infty} \left\|(\mathbb{1} - P_n)K^\dagger\eta\right\| \to 0 \end{aligned} \tag{7.68}$$

  Because $K^\dagger\eta$ is a fixed abstract vector and $\lim_{n\to\infty} P_n = \mathbb{1}$, thus $\left\|(\mathbb{1} - P_n)K^\dagger\eta\right\| \to 0$ as $n \to \infty$. Equation (7.68) implies that η must be the zero vector as $n \to \infty$.

  Consequently,

  $$\|K(\mathbb{1} - P_n)\hat{g}_n\| \to 0, \tag{7.69}$$

  which implies that

  $$\|(K - K_n)\hat{g}_n\| \to 0. \tag{7.70}$$

- We now prove by contradiction. Suppose $K - K_n \nrightarrow 0$, then there exists a constant $a > 0$ and integer $n_0 > 0$ such that

  $$\|(K - K_n)g_n\| \geq a \, \|g_n\|, \quad \forall n > n_0. \tag{7.71}$$

  Since $\{\hat{g}_n\}$ is a subsequence of $\{g_n\}$, then (7.70) and (7.71) are contradictive. This completes the proof. ■.

Theorem 212 shows that for any compact operator K, it can be expressed by

$$K = K_n + (K - K_n) \tag{7.72}$$

such that

$$K \leq \|K_n\| + \|K - K_n\|$$

where $K_n = P_nKP_n$ is an operator with finite rank. When n is sufficiently large, $\|K_n - K\| \to 0$, i.e., $K_n - K$ is sufficiently small, which means that $K_n - K$ is a small operator. In concequence, any compact operator can be decomposed into the sum of an operator K_n with finite rank and a small operator $K - K_n$. This property is important for solving integral equations.

7.7.3 Compact Opertors and Integral Equations

The typical compact operator is the integral oprator defined by

$$(Af)(x) = \int_a^b K(x,y)f(y)\,\mathrm{d}\,y, \quad x \in [a,b]$$

where $K(x,y)$ is known as the kernal. This kind of compact operator, involved with integral equations, is of particular significance in mathematical physics and engineering sicience such as electrodynamics, microwaves, quantum mechanics and signals analysis.

A. Introduction

An integral equation is an equation in which an unknown function appears under an integral sign and the problem of solving the equation is to determine that function. Some problems of mathematical physics lead directly to integral equations, and other problems, which lead first to ordinary or partial differential equations, can be handled more expenditiously by converting them to integral equations. Historicaly, some famous mathematicians presented some isolated problems involving integral equations occurred long before the subject acquired a distinct status and methodology.

① Laplace, 1782.

$$f(x) = \int_{-\infty}^{\infty} \mathrm{e}^{-xt} g(t)\,\mathrm{d}\,t \tag{7.73}$$

$f(x)$ is called the Laplace transform of $g(t)$.

② Poisson discovered the expression for $g(t)$ involved in the Laplace transform in 1823, namely,

$$g(t) = \frac{1}{2\pi\mathrm{i}} \int_{a-\mathrm{i}\infty}^{a+\mathrm{i}\infty} \mathrm{e}^{xt} f(x)\,\mathrm{d}\,x \tag{7.74}$$

for large enough a.

③ The noteworthy results that really belong to the history of integral equations stems from Fourier's famous 1811 paper on the theory of heat. Here one finds

$$f(x) = \int_0^{\infty} \cos(xt)u(t)\,\mathrm{d}\,t \tag{7.75}$$

and the inversion formula

$$u(t) = \frac{2}{\pi} \int_0^{\infty} \cos(xt)f(x)\,\mathrm{d}\,x. \tag{7.76}$$

④ The first conscious direct use and solution of an integral equation go back to Abel in 1823. He proposed the following integral equation

$$f(x) = \int_0^x \frac{v'(\xi)}{\sqrt{x-\xi}}\,\mathrm{d}\,\xi$$

and obtained the solution

$$v(\xi) = \int_0^{\xi} \frac{f(x)}{\sqrt{\xi - x}}\,\mathrm{d}\,x.$$

Actually Abel undertook to solve the more general problems

$$f(x) = \int_a^x \frac{u(\xi)}{(x-\xi)^{\lambda}}\,\mathrm{d}\,\xi, \quad \lambda \in (0,1) \tag{7.77}$$

and obtained

$$u(z) = \frac{\sin\lambda\pi}{\pi} \frac{\mathrm{d}}{\mathrm{d}\,z} \int_a^z \frac{f(x)}{(z-x)^{1-\lambda}}\,\mathrm{d}\,x \tag{7.78}$$

⑤ Liouville, who worked independently of Abel, solved special integral equations from 1832 on. A more significant step by Liouville was to show how the solution of certain differential equations can be obtained by solviing integral equations. The ODE to be solved is

$$y'' + [\rho^2 - \sigma(x)]y = 0, \quad x \in [a, b] \tag{7.79}$$

where ρ is a parameter. Let $u(x)$ be the particular solution that satisfies the initial condition

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

This function will also be a solution of the nonhomogeneous equation

$$y'' + \rho^2 y = \sigma(x)u(x). \tag{7.80}$$

Then by a basic result on ODE,

$$u(x) = \cos\rho(x - a) + \frac{1}{\rho}\int_a^x \sigma(\xi)\sin\rho(x - \xi)u(\xi)\,\mathrm{d}\,\xi. \tag{7.81}$$

Liouville obtained the solution be a method of successive substitutions attributed to Car G. Neumann, whose work *Untersuchungen über das logarithmische und Newton'sche Potential* (1877) came thirty years later.

⑥ The integral equaitons treated by Abel and Liouville are of basic types. Abel's is of the form

$$f(x) = \int_a^x K(x, \xi)u(\xi)\,\mathrm{d}\,\xi, \tag{7.82}$$

and Liouville's of the form

$$u(x) = f(x) + \int_a^x K(x, \xi)u(\xi)\,\mathrm{d}\,\xi. \tag{7.83}$$

In both of these $f(x)$ and $K(x, \xi)$ are known, and $u(\xi)$ is the function to be determined. The terminlogy used today, introduced by Hilbert, referes to these equations as the first and second kind, respectively, and $K(x, \xi)$ is called the *kernal*. As stated, they are also referred to as Volterra's equations, whereas when the upper limit is a fixed number b, they are called Fredholm's equations.

⑦ By the middle of the nineteenth century the chief interest in integral equations centerd around the solution of the boundary-value problem associated with the potential equation

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \quad (x, y) \in \Omega\mathbb{R}^2, \quad \partial\,\Omega = C. \tag{7.84}$$

The equation holds in a given plane area Ω that is bounded by some curve C, i.e., $\partial\,\Omega = C$. If the boundary value of u is some function $f(s)$ given as a function of arc lenfths s along C, then a solution of this problem can be represented by

$$u(x, y) = \frac{1}{2\pi}\int_C \rho(s)\log\frac{1}{r(s; x, y)}, \tag{7.85}$$

wherein $r(s; x, y)$ is the distance from a point s to any point (x, y) in the inerior or boundary and $\rho(s)$ is an unknown function satisfying for $s = (x, y)$ on C

$$f(s) = \frac{1}{2\pi}\int_C \rho(t)\log\frac{1}{r(t; x, y)}\,\mathrm{d}\,t. \tag{7.86}$$

This is an integral equation of the first kind for $\rho(t)$. Alternatively, if one takes as a solution of (7.84) with the same boundary condition

$$v(x, y) = \frac{1}{2\pi}\int_C \phi(s)\frac{\partial}{\partial n}\left[\log\frac{1}{r(s; x, y)}\right]\mathrm{d}\,s, \tag{7.87}$$

where $\frac{\partial}{\partial n}$ denotes the normal derivative to the boundary, then $\phi(s)$ must satisfy the integral equation

$$f(s) = \frac{1}{2}\phi(s) + \frac{1}{2\pi}\int_C \phi(t)\frac{\partial}{\partial n}\left[\log\frac{1}{r(s;x,y)}\right]\mathrm{d}\,t, \tag{7.88}$$

an integral equation of the second kind. These equations were solved by Neumann for convex area in his *Untersuchungen* and later publications.

⑧ The equation

$$\Delta u + \lambda u = f(x,y) \tag{7.89}$$

arises in the study of wave motion when the time dependence of the corresponding hyperbolic equation

$$\Delta u - \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} = f(x,y),$$

usually taken to be $\mathrm{e}^{-\mathrm{i}\omega t}$, is eliminated. Poincaré in 1894 considered the innomogeneous case (7.89) with complex λ. He was able to produce a function, meromorphic in λ, which represented a unique solution of (7.89) for any λ which is not an eigenvalue, and whose residues produce eigenfunctions for the homogeneous case, that is, when $f = 0$. Poincaré in 1896 considered the equation

$$u(x) + \lambda\int_a^b K(x,y)u(y)\,\mathrm{d}\,y = f(x),$$

which he derived from (7.89), and affirmed that the solution is meromorphic function of λ. This result was established by Fredholm in a paper.

The conversion of differential equations to integral equations, which is ilustrated by the above examples, became a major technique for solving initial- and boundary-value problems of ODEs and PDEs, and was the strongest impetus for the study of integral equations.

B. Volterra's and Fredholm's Integral Equations

Definition 213. *The linear equations of the unknown function* $f : [a,b] \to \mathbb{R}$

$$\int_a^x K(x,y)f(y)\,\mathrm{d}\,y + g(x) = 0 \tag{7.90}$$

$$\lambda\int_a^x K(x,y)f(y)\,\mathrm{d}\,y + g(x) = f(x) \tag{7.91}$$

$$\lambda\int_a^x K(x,y)f(y)\,\mathrm{d}\,y + g(x) = q(x)f(x) \tag{7.92}$$

are called the first, second, and third kinds of Volterra's (linear) integral equations respectively. The Volterra's nonlinear integral equations are of the form

$$f(x) = g(x) = \int_a^x H(x,y,f(y))\,\mathrm{d}\,y. \tag{7.93}$$

Vito Volterra (1860-1940), who succeeded Beltrami as professor of mathematical physics at Rome, is the first of the founders of a general theory of integral equations. He wrote papers on the subject from 1884 on and principal ones in 1896 and 1897.

If $K(x,y)$ is continuous, Volterra's first kind integral equations can be transformed into the second kind ones. Actually, taking the derivative of the left side of (7.95) results

$$f(x) + \int_a^x \frac{K_x(x,y)}{K(x,x)}f(y)\,\mathrm{d}\,y = \frac{g'(x)}{K(x,x)} \tag{7.94}$$

where $K_x(x,y) = \frac{\partial}{\partial x}K(x,y)$.

Definition 214. *The linear equations of the unknown function* $f : [a, b] \to \mathbb{R}$

$$\int_a^b K(x, y)f(y)\,\mathrm{d}\,y + g(x) = 0 \tag{7.95}$$

$$\lambda \int_a^b K(x, y)f(y)\,\mathrm{d}\,y + g(x) = f(x) \tag{7.96}$$

$$\lambda \int_a^b K(x, y)f(y)\,\mathrm{d}\,y + g(x) = q(x)f(x) \tag{7.97}$$

are called the first, second, and third kinds of Fredholm's integral equations respectively.

Actually the Volterra equations are special cases, respectively, of Fredholms's because one can always take

$$K(x, \xi) = 0, \quad \xi > x$$

and then regard the Volterra equations as Fredholm equations. The special case of the equation of the second kind in which $f(x) \equiv 0$ is called the homogeneous equation.

C. Classification of Integral Kernals

There are some special and improtant kinds of kernals:

❶ If the kernal $K(x, y)$ is continuous on $(x, y) \in [a, b] \times [a, b]$, then it is called a continuous kernal.

❷ If $K(x, y) \in L^2([a, b] \times [a, b])$, i.e.,

$$\int_a^b \mathrm{d}\,x \int_a^b |K(x, y)|^2\,\mathrm{d}\,y < \infty,$$

then $K(x, y)$ is called the square integral kernal, or simply L^2-kernal.

❸ If the complex conjugate of $K(x, y)$ satisfies

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

then it is called Hermitian kernal or conjugate kernal. Usually, it is denoted as

$$\overline{K}(y, x) = K^\dagger(x, y).$$

Furthermore, if $K(x, y)$ is real, the conjugate kernal satisfies $K(x, y) = K(y, x)$, and it is called symmetric kernal.

❹ If the kernal can be decomposed as

$$K(x, y) = \sum_{i=1}^{n} \phi_i(x)\overline{\chi}_i(y), \tag{7.98}$$

then it is called degerate kernal or separabel kernal, and the n is called the rank of the kernal. Particularly, for $n = 1$, we have

$$K(x, y) = \phi(x)\overline{\chi}(y). \tag{7.99}$$

❺ The kernal such that

$$K(x, y) = K(\pm(x - y))$$

is called the convolutional kernal. The corresponding integral equation is called the convolutional integral equation.

D. Integral Equations and Differential Equations

The initial-value problem

$$\begin{cases} y'' + p(x)y' + q(x)y = f(x) \\ y(a) = \alpha, \quad y'(a) = \beta \end{cases} \tag{7.100}$$

is equivalent to

$$y(x) = g(x) + \lambda \int_a^x K(x,t)y(t)\,\mathrm{d}\,t \tag{7.101}$$

where

$$\begin{aligned} K(x,t) &= (t-x)[q(t) - p'(t)] - p(t) \\ g(x) &= \int_a^x (x-t)f(t)\,\mathrm{d}\,t + [p(a)\alpha + \beta](x-a) + \alpha. \end{aligned}$$

The boundary-value problem

$$\frac{\mathrm{d}^2 f}{\mathrm{d}\,x^2} = \phi(x, f(x)), \quad x \in [a,b] \tag{7.102}$$

is equivalent to

$$f(x) = A + Bx + \int_a^x (x-u)\phi(u, f(u))\,\mathrm{d}\,u \tag{7.103}$$

where A and B can be determined by the boundary conditions.

E. Fredholm's Homogeneous Integral Equations

Definition 215. *For the second kind of Fredholm's homogeneous integral equation*

$$\begin{aligned} f(x) &= \lambda(\mathsf{A}f)(x) \\ &= \lambda \int_a^b K(x,y)f(y)\,\mathrm{d}\,y, \end{aligned} \tag{7.104}$$

if the equation has nontrivial solution $f(x)$ for λ, then λ is called the eigenvalue belonging to the kernal $K(x,y)$, and the corresponding $f(x)$ is called the eigenfunctions belonging to λ.

Definition 216. *Let $\lambda = \lambda_i$ is an eigenvalue of $K(x,y)$, it has m linear independent eigenfunctions $f_{i,1}(x), f_{i,2}(x), \cdots, f_{i,m}(x)$ such that*

$$f_{i,k}(x) = \lambda_i(\mathsf{A}f_{i,k})(x) = \lambda_i \int_a^b K(x,y)f_{i,k}(y)\,\mathrm{d}\,y, \quad k = 1, 2, \cdots, m$$

then the λ_i is called m-degenerate, and m is called the rank of the eigenvalue λ_i or degree of degeration.

Theorem 217 (Fredholm Theorem)**.** *For the second kind of Fredholm's homogeneous integral equation $f = \lambda\mathsf{A}f$, we have the following results:*

① *For any finite domain in the λ-plane, there are finite eigenvalues for the kernal $K(x,y)$.*

② *Each eigenvalue has at least one eigenfunctions, and the linear independent eigenfunctions for the same eigenvalue λ span a finite dimensional linear space $\mathcal{V}_\lambda$.*

③ *If λ is the eigenvalue of $K(x,y)$, then $\overline{\lambda}$ is the eigenvalue of $\overline{K}(y,x)$, and $\dim(\mathcal{V}_\lambda) = \dim(\mathcal{V}_{\overline{\lambda}})$.*

Theorem 218 (Hermitian Kernal)**.** *For the Hermitian kernal $K(x,y)$, if $K(x,y) \in L^2$, then there exists at least one eigenvalue λ for the linear interal equation $\lambda\mathsf{A}f = f$.*

If the kernal is not Hermitian, this theorem may not hold true.

Theorem 219. *For the second kind of Fredholm's homogeneous integral equation $f = \lambda \mathsf{A} f$, if the kernal K is Hermitian, then*

- *Each eigenvalue λ is real.*
- *If $f = \lambda \mathsf{A} f$ and $g = \mu \mathsf{A} g$ such that $\mu \neq \lambda$, then $\langle f | g \rangle = 0$.*

For the Kernal $K(x, y)$ and the eigenvalue euqation

$$f_{i,k} = \lambda_i \mathsf{A} f_{i,k}, \quad i = 1, 2, \cdots,$$

we can sort the eigenvalues with decreasing order

$$|\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \geq \cdots.$$

The normalized eigenfuctions can be denoted as $\{\phi_k\}$ such that $\phi_k = \lambda_k \mathsf{A} \phi_k$ (Note that the degenreate eigenvalues are used). Thus we have eigenpairs $\{\lambda_k, \phi_k\}$, we call $\{\lambda_k\}$ and $\{\phi_k\}$ the eigensystem of $K(x, y)$. Generally, we have

$$\sum_{k=1}^{\infty} \frac{1}{|\lambda_k|^2} \leq \int_a^b \int_a^b |K(x, y)|^2 \, \mathrm{d}\, x \, \mathrm{d}\, y. \tag{7.105}$$

The reader will find that this is the altenative version of the Bessels's inequality.

Theorem 220 (Expansion of Integral Kernal)**.** *If the kernal $K(x, y) \in L^2$ and is Hermitian, $\{\lambda_i\}$ and $\{\phi_i\}$ are the eigenvalues and eigenfuncitons of K, i.e., $\phi_i = \lambda_i \mathsf{A} \phi$. Then the kernal can be expanded as*

$$K(x, y) = \sum_{i=1}^{\infty} \frac{1}{\lambda_i} \phi_i(x) \overline{\phi_i(y)}. \tag{7.106}$$

Proof

- Let the expansion of K be
$$K(x, y) = \sum_{i=1}^{\infty} c_i(y) \phi_i(x),$$
then the coefficient $c_i(y)$ is
$$\begin{aligned} c_i(y) &= \langle \phi_i(x) | K(x, y) \rangle = \int_a^b \overline{\phi_i(x)} K(x, y) \, \mathrm{d}\, x \\ &= \overline{\int_a^b \phi_i(x) K^\dagger(x, y) \, \mathrm{d}\, x} \\ &= \frac{1}{\lambda_i} \overline{\phi_i(y)} \end{aligned}$$
since $\lambda_i = \overline{\lambda_i}$ and $K^\dagger(x, y) = K(y, x)$.
- Therefore,
$$\sum_{i=1}^{\infty} \frac{1}{\lambda_i} \phi_i(x) \overline{\phi_i(y)}.$$

Obviously, this is the concrete example of the decomposition theorem of abstract compact operator.

Theorem 221. *For the Hermitian kernal $K(x, y)$,*

if the $K(x, y)$ is nondegenrate, there are countbable infinite eigenvalues;

the $K(x, y)$ is degenrate if and only if it has finite eigenvalues.

F. Iterative Solution

For the Fredholm's second kind of integral equation, we have the following theorem:

Theorem 222. *The linear integral equation*

$$f(x) = g(x) + \lambda(\mathsf{A}f)(x) = g(x) + \int_a^b K(x,y)f(y)\,\mathrm{d}\,y$$

has a unique solution

$$\begin{aligned} f &= \sum_{n=0}^{\infty} \lambda^n \mathsf{A}^n g \\ &= g(x) + \lambda \int_a^b K(x,y)g(y)\,\mathrm{d}\,y \\ &\quad + \lambda^2 \int_a^b \int_a^b \mathrm{d}\,\xi_1\,\mathrm{d}\,\xi_2 K(x,\xi_1)K(x,\xi_2)g(\xi_1)g(\xi_2) \\ &\quad + \lambda^3 \int_a^b \int_a^b \int_a^b \mathrm{d}\,\xi_1\,\mathrm{d}\,\xi_2\,\mathrm{d}\,\xi_3 K(x,\xi_1)K(x,\xi_2)K(x,\xi_3)g(\xi_1)g(\xi_2)g(\xi_3) \\ &\quad + \cdots \end{aligned} \tag{7.107}$$

if $\|\lambda\mathsf{A}\| < 1$.

Generally, for the operator equation in Banach space $\mathfrak{X}$

$$f = g + \lambda\mathsf{A}f, \quad \forall g \in \mathfrak{X}, \tag{7.108}$$

where only f is unknown, if the operator T defined by

$$\mathsf{T}f = g + \lambda\mathsf{A}f$$

is contractive, i.e., $\|\lambda\mathsf{A}\| < 1$, then the Banach's fixed-point theorem implies that the iterative sequence

$$f_{n+1} = \mathsf{T}f_n = g + \lambda\mathsf{A}f_n, \quad n = 1, 2, \cdots \tag{7.109}$$

with initial condition $f_0 = g$ has a unique solution

$$f = \lim_{n\to\infty} f_n = (\mathbb{1} - \lambda\mathsf{A})^{-1} g.$$

Chapter 8

Hahn-Banach Theorem and Optimization Problem

The Hahn-Banach theorem, proved independently by Hahn in 1926 and by Banach in 1929, is the most important theorem about the structure of linear continuous functionals on normed space. In terms of geometry, the Hahn-Banach theorem guarantees the separation of convex sets in normed spaces by hyperplanes. Figure 8.1 describes a number of important consequences of the Hahn-Banach theorem.

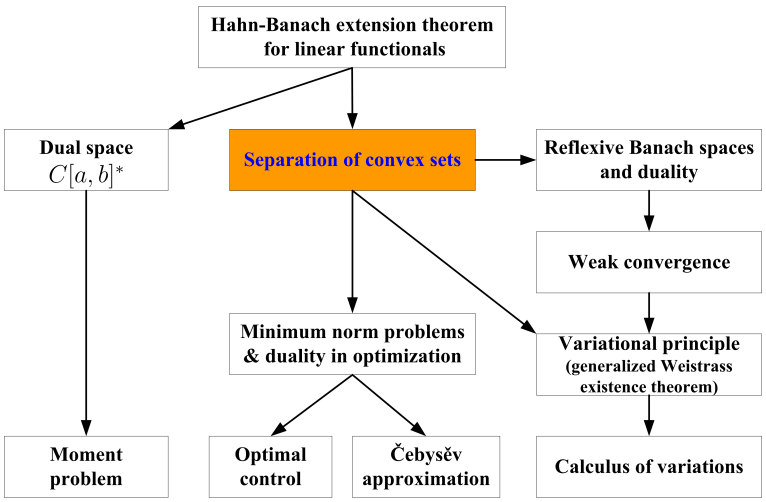


Figure 8.1: Hahn-Banach theorem and its applications

The Hahn-Banach theorem represents a fundamental existence principle in linear functional analysis that allows the solution of variational problems without using any compactness.

8.1 Hahn-Banach Theorem

Theorem 223 (The Hahn-Banach theorem for linear spaces)**.** *We assume that*

① $\mathcal{L}$ *is linear subspace of the real linear space* $\mathcal{X}$.

② $p : \mathcal{X} \to \mathbb{R}$ *is a sublinear functional, that is, for all* $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{X}$ *and all* $\alpha \geq 0$,

$$p(\boldsymbol{u} + \boldsymbol{v}) \leq p(\boldsymbol{u}) + p(\boldsymbol{v}) \quad \textit{and} \quad p(\alpha \boldsymbol{u}) = \alpha p(\boldsymbol{u}).$$

③ $F : \mathcal{L} \to \mathbb{R}$ *is a linear functional such that*

$$F(\boldsymbol{u}) \le p(\boldsymbol{v}), \quad \forall \boldsymbol{u} \in \mathcal{L}. \tag{8.1}$$

Then, F can be extended to a linear functional $f : \mathcal{X} \to \mathbb{R}$ such that

$$f(\boldsymbol{u}) \le p(\boldsymbol{u}), \quad \forall \boldsymbol{u} \in \mathcal{X}.$$

Proof.

- Step-1. We first prove the statement in the special case where

$$\mathcal{X} = \mathcal{L} + \operatorname{span}\{\boldsymbol{v}\}, \quad \text{with fixed } \boldsymbol{v} \notin \mathcal{L}.$$

  To this end, we set

$$f(\boldsymbol{u} + \lambda \boldsymbol{v}) \triangleq F(\boldsymbol{u}) + c\lambda, \quad \forall \boldsymbol{u} \in \mathcal{L}, \lambda \in \mathbb{R},$$

  where c is a fixed real number that satisfies the following condition:

$$\sup_{\boldsymbol{u} \in \mathcal{L}} (F(\boldsymbol{u}) - p(\boldsymbol{u} - \boldsymbol{v})) \le c \le \inf_{\boldsymbol{w} \in \mathcal{L}} (p(\boldsymbol{w} + \boldsymbol{v}) - F(\boldsymbol{w})). \tag{8.2}$$

  We have to show that such a number c exists.

  – In fact, for all $\boldsymbol{u}, \boldsymbol{w} \in \mathcal{L}$, we get

$$\begin{aligned} F(\boldsymbol{u}) + F(\boldsymbol{v}) &= F(\boldsymbol{u} + \boldsymbol{v}) \le p(\boldsymbol{u} + \boldsymbol{w}) \\ &= p(\boldsymbol{u} - \boldsymbol{v} + \boldsymbol{w} + \boldsymbol{v}) \le p(\boldsymbol{u} - \boldsymbol{v}) + p(\boldsymbol{w} + \boldsymbol{v}), \end{aligned}$$

    and hence

$$F(\boldsymbol{u}) - p(\boldsymbol{u} - \boldsymbol{v}) \le p(\boldsymbol{w} + \boldsymbol{v}) - F(\boldsymbol{w}), \quad \forall \boldsymbol{u}, \boldsymbol{w} \in \mathcal{L}.$$

    This proves (8.2).

  Obviously, the functional $f : \mathcal{X} \to \mathbb{R}$ is linear. Thus, it remains to show that

$$F(\boldsymbol{u}) + c\lambda \le p(\boldsymbol{u} + \lambda \boldsymbol{v}), \quad \forall \boldsymbol{u} \in \mathcal{L}, \lambda \in \mathbb{R}. \tag{8.3}$$

  – In fact, this is true for $\lambda = 0$. Let $\lambda > 0$. By (8.2),

$$c \le p(\lambda^{-1}\boldsymbol{u} + \boldsymbol{v}) - F(\boldsymbol{\lambda u}) = \lambda^{-1}(p(\boldsymbol{u} + \lambda \boldsymbol{v}) - F(\boldsymbol{u})).$$

    This is (8.3). In the case where $\lambda < 0$, it follows from (8.2) that

$$c \ge F(-\lambda^{-1}\boldsymbol{u}) - p(-\lambda^{-1}\boldsymbol{u} - \boldsymbol{v}) = -\lambda^{-1}[F(\boldsymbol{u}) - p(\boldsymbol{u} + \lambda \boldsymbol{v})].$$

    and again we get (8.3).

- Step-2. Induction. Suppose that there exists a sequence $\{\mathcal{L}_n\}$ of linear subspaces of $\mathcal{X}$ such that $\mathcal{L} = \mathcal{L}_1 \subset \mathcal{L}_2 \subset \mathcal{L}_3 \subset \cdots$ along with

$$\mathcal{X} = \bigcup_n \mathcal{L}_n,$$

  where

$$\mathcal{L}_{n+1} = \mathcal{L}_n + \operatorname{span}\{\boldsymbol{v}_n\} \quad \text{for some fixed } \boldsymbol{v}_n \in \mathcal{X} \text{ and } \boldsymbol{v}_n \notin \mathcal{L}_n,$$

  and for all n. Using Step-1, a simple induction argument shows that F can be extended to $\mathcal{L}_n$ for all n. This yields the desired extension f on F.

- Step-3. If the situation from Step-2 is not at hand, then we can use the Zorn Lemma. To this end, let $\mathcal{C}$ denote the set of all the linear functionals

$$g : \mathrm{Dom}(g) \subset \mathcal{X} \to \mathbb{K}$$

that are an extension of F such that

$$g(\boldsymbol{u}) \leq p(\boldsymbol{u}) \quad \forall \boldsymbol{u} \in \mathrm{Dom}(g).$$

In other words, we have

$$\mathcal{C} \triangleq \{g \mid g : \mathrm{Dom}(g) \subset \mathcal{X} \to \mathbb{K} \text{ and } \forall \boldsymbol{u} \in \mathrm{Dom}(g), g(\boldsymbol{u}) \leq p(\boldsymbol{u}).\}$$

We write

$$g \leq h \quad \text{iff } h : \mathrm{Dom}(h) \to \mathbb{K} \text{ is an extension of } g : \mathrm{Dom}(g) \to \mathbb{K}.$$

This way $\mathcal{C}$ becomes an ordered set. Let $\mathcal{T}$ be a totally ordered subset of $\mathcal{C}$, that is, $g, h \in \mathcal{T}$ implies

$$g \leq h \quad \text{or} h \leq g.$$

Then, there exists an upper bound $b \in \mathcal{C}$ for $\mathcal{T}$, that is,

$$g \leq b, \quad \forall g \in \mathcal{T}.$$

  - To show this, let $\mathrm{Dom}(b)$ be the union of all the sets $\mathrm{Dom}(g)$ with $g \in \mathcal{T}$ and define

$$b(\boldsymbol{u}) \triangleq g(\boldsymbol{u}) \quad \text{on} \quad \mathrm{Dom}(g).$$

    Since $\mathcal{T}$ is totally ordered, the linear functional $b : \mathrm{Dom}(b) \to \mathbb{K}$ is well defined, and $b(\boldsymbol{u}) \leq p(\boldsymbol{u})$ for all $\boldsymbol{u} \in \mathrm{Dom}(b)$.

By the Zorn lemma, there is a maximal element f in $\mathcal{C}$. That is, the linear functional $f : \mathrm{Dom}(f) \subset \mathcal{X} \to \mathbb{K}$ has no proper extension in the sense of $\mathcal{C}$. This implies $\mathrm{Dom}(f) = \mathcal{X}$ and $f(\boldsymbol{u}) \leq p(\boldsymbol{u})$ for all $\boldsymbol{u} \in \mathcal{X}$. In fact, suppose that $\mathrm{Dom}(f) \neq \mathcal{X}$. Then, there exists an extension of f in the sense of $\mathcal{C}$, by Step-1. This contradicts the maximality of f. Thus, $f : \mathcal{X} \to \mathbb{R}$ is the desired extension of F. ■

Theorem 224 (Hahn-Banach theorem for normed spaces)**.** *We assume that*

① $\mathcal{L}$ *is linear subspace of the real linear space* $\mathcal{X}$.

② $F : \mathcal{L} \to \mathbb{R}$ *is a linear functional such that*

$$|F(\boldsymbol{u})| \leq \alpha \|\boldsymbol{u}\|, \quad \text{for fixed } \alpha \geq 0 \text{ and } \forall \boldsymbol{u} \in \mathcal{L}. \tag{8.4}$$

Then, F *can be extended to a linear functional* $f : \mathcal{X} \to \mathbb{R}$ *such that*

$$|f(\boldsymbol{u})| \leq \alpha \|\boldsymbol{u}\|, \quad \forall \boldsymbol{u} \in \mathcal{X}.$$

Proof.

- Step-1. Let $\mathbb{K} = \mathbb{R}$. Set

$$p(\boldsymbol{u}) = \alpha \|\boldsymbol{u}\|, \quad \forall \boldsymbol{u} \in \mathcal{X}.$$

By Theorem 223, the fundamental F can be extended to a linear functional $f : \mathcal{X} \to \mathbb{R}$ such that

$$f(\boldsymbol{u}) \leq \alpha \|\boldsymbol{u}\|, \quad \forall \boldsymbol{u} \in \mathcal{X}.$$

Since $f(\pm\boldsymbol{u}) = \pm f(\boldsymbol{u})$, we get $|f(\boldsymbol{u})| \leq \alpha \|\boldsymbol{u}\|$. Thus, f is continuous.

- Step-2. Let $\mathbb{K} = \mathbb{C}$. Define

$$H(\boldsymbol{u}) \triangleq \Re[F(\boldsymbol{u})], \quad \forall \boldsymbol{u} \in \mathcal{L}.$$

Then,

$$\begin{aligned} F(\boldsymbol{u}) &= \Re[F(\boldsymbol{u})] + \mathrm{i}\Im[F(\boldsymbol{u})] = \Re[F(\boldsymbol{u})] - \mathrm{i}\Re[\mathrm{i}\boldsymbol{u}], \\ &= H(\boldsymbol{u}) - \mathrm{i}H(\mathrm{i}\boldsymbol{u}), \quad \boldsymbol{u} \in \mathcal{L}, \end{aligned}$$

and

$$|H(\boldsymbol{u})| \leq \alpha \|\boldsymbol{u}\| \quad \forall \boldsymbol{u} \in \mathcal{L}.$$

If we regard $\mathcal{X}$ as a real normed space, then it follows from Step-1 that there exists a linear continuous functional $h : \mathcal{X} \to \mathbb{R}$ such that $h(\boldsymbol{u}) = H(\boldsymbol{u})$ for all $\boldsymbol{u} \in \mathcal{L}$ and

$$|h(\boldsymbol{u})| \leq \alpha \|\boldsymbol{u}\|, \quad \forall \boldsymbol{u} \in \mathcal{X}.$$

Define

$$f(\boldsymbol{u}) \triangleq h(\boldsymbol{u}) - \mathrm{i}h(\mathrm{i}\boldsymbol{u}), \quad \forall \boldsymbol{u} \in \mathcal{X}.$$

Hence $h(\boldsymbol{u}) = \Re[f(\boldsymbol{u})]$. We want to show that f is the desired functional.

Obviously, $f : \mathcal{X} \to \mathbb{C}$ is an extension of F. Moreover, f is linear. This follows from

$$f(\mathrm{i}\boldsymbol{u}) = h(\mathrm{i}\boldsymbol{u}) - \mathrm{i}h(-\boldsymbol{u}) = \mathrm{i}f(\mathrm{i}\boldsymbol{u}), \quad \boldsymbol{u} \in \mathcal{X},$$

and from the linearity of h w.r.t. $\mathbb{R}$. Finally, we have to show that

$$|f(\boldsymbol{u})| \leq \alpha \|\boldsymbol{u}\|, \forall \boldsymbol{u} \in \mathcal{X}.$$

In fact, for each $\boldsymbol{u} \in \mathcal{X}$, we get $f(\boldsymbol{u}) = r\mathrm{e}^{\mathrm{i}\beta}$ with $r \geq 0$. Hence

$$\begin{aligned} |f(\boldsymbol{u})| = r &= \Re[\mathrm{e}^{-\mathrm{i}\beta} f(\boldsymbol{u})] = \Re[f(\mathrm{e}^{-\mathrm{i}\beta}\boldsymbol{u})] \\ &= h(\mathrm{e}^{-\mathrm{i}\beta}\boldsymbol{u}) \leq \alpha \left\|\mathrm{e}^{-\mathrm{i}\beta}\boldsymbol{u}\right\| = \alpha \|\boldsymbol{u}\|. \quad \blacksquare \end{aligned}$$

Example 225. *Let X be a normed space over $\mathbb{K}$. Then, for each given $\boldsymbol{u}_0 \in \mathcal{X}$ with $\boldsymbol{u}_0 \neq \boldsymbol{0}$, there exists a functional $f \in \mathcal{X}^*$ such that*

$$f(\boldsymbol{u}_0) = \|\boldsymbol{u}_0\| \quad \textit{and} \quad \|f\| = 1.$$

Proof.

Set $L \triangleq \operatorname{span}\{\boldsymbol{u}_0\}$ and

$$F(\boldsymbol{u}) \triangleq \lambda \|\boldsymbol{u}_0\|, \quad \forall \boldsymbol{u} = \lambda \boldsymbol{u}_0 \in \mathcal{L}.$$

Obviously, $|F(\boldsymbol{u})| = \|\boldsymbol{u}\|$ for all $\boldsymbol{u} \in \mathcal{L}$. By Theorem 224, there exists a functional $f \in \mathcal{X}^*$ such that $f(\boldsymbol{u}) = F(\boldsymbol{u})$ on $\mathcal{L}$ and

$$|f(\boldsymbol{u})| \leq \|\boldsymbol{u}\|, \quad \forall \boldsymbol{u} \in \mathcal{X}.$$

Hence $\|f\| = 1$.

Corollary 226. *Let $\mathcal{X}$ be a normed space over $\mathbb{K}$. Then, for all $\boldsymbol{u}_0 \in \mathcal{X}$,*

$$\|\boldsymbol{u}\|_0 = \max_{f \in \mathcal{X}^*, \|f\| \leq 1} |f(\boldsymbol{u}_0)|.$$

Proof.

Since $|f|(\boldsymbol{u}_0) \leq \|f\| \|\boldsymbol{u}_0\|$ for all $f \in \mathcal{X}^*$, the assertion follows from Example 225.

Corollary 227. *Let $\mathcal{X}$ be a normed space over $\mathbb{K}$. Then, we have*

$$\forall f \in \mathcal{X}^*, \quad f(\boldsymbol{u}) = 0 \Longrightarrow \boldsymbol{u} = 0.$$

This is an immediate consequence of Example 225.

8.2 Applications to the Separation of Convex Sets

Definition 228. *By a* ***closed hyperplane*** H *in the real normed space* X*, we understood a set*

$$H \triangleq \{\boldsymbol{u} \in \mathcal{X} : f(\boldsymbol{u}) = \alpha\},$$

where $f : \mathcal{X} \to \mathbb{R}$ *is a linear continuous functional and* α *is a fixed real number. We also define the* ***half-spaces*** $H_{\leq}$ *and* $H_{>}$ *of* H *through*

$$H_{\leq} \triangleq \{\boldsymbol{u} \in \mathcal{X} : f(\boldsymbol{u}) \leq \alpha\} \quad \textit{and} \quad H_{>} \triangleq \{\boldsymbol{u} \in \mathcal{X} : f(\boldsymbol{u}) > \alpha\}.$$

Let A and B be two subsets of $\mathcal{X}$. Then, we say that the closed hyperplane H strictly *separates* the sets A and B iff

$$A \subset H_{\leq} \quad \text{and} \quad B \subset H_{>}.$$

Furthermore, we say that the closed hyperplane H separates the sets A and B iff $A \subset H_{\leq}$ and $B \subset H_{>}$.

Proposition 229. *Let* $\mathcal{L}$ *be a linear subspace of the normed space* X *over* $\mathbb{K}$*. Then, for each point* $\boldsymbol{u}_0 \in \mathcal{X}$ *with*

$$\operatorname{dist}(\boldsymbol{u}_0, \mathcal{L}) > 0,$$

there exists a linear continuous functional $f : \mathcal{X} \to \mathbb{K}$ *such that*

$$f(\boldsymbol{u}) = 0, \quad \forall \boldsymbol{u} \in \mathcal{L},$$

along with $\|f\| = 1$ *and* $f(\boldsymbol{u}_0) = \operatorname{dist}(\boldsymbol{u}_0, \mathcal{L})$.

Recall that

$$\operatorname{dist}(\boldsymbol{u}_0, \mathcal{L}) \triangleq \inf_{\boldsymbol{v} \in \mathcal{L}} \|\boldsymbol{u}_0 - \boldsymbol{v}\|. \tag{8.5}$$

If $\mathcal{X}$ is a real normed space, then this means that the closed hyperplane

$$H \triangleq \{\boldsymbol{u} \in \mathcal{X} : f(\boldsymbol{u}) = 0\}$$

separates strictly the linear subspace $\mathcal{L}$ and the point $\boldsymbol{u}_0$, where $\mathcal{L} \subset H$.

Theorem 230. *Let* M *be a nonempty closed convex subset of a normed space* $\mathcal{X}$ *over* $\mathbb{K}$*, and let* $\boldsymbol{u}_0$ *be a point of* $\mathcal{X}$ *with* $\boldsymbol{u}_0 \notin M$*. Then, there exists a linear continuous functional* $f : \mathcal{X} \to \mathbb{K}$ *such that*

$$\forall \boldsymbol{u} \in M, \quad \Re[f(\boldsymbol{u})] \leq 1 \quad \textit{and} \quad \Re[f(\boldsymbol{u}_0)] > 1.$$

In terms of geometry, this theorem tells us the following: Let $\mathcal{X}$ be a real normed space, and set

$$H \triangleq \{\boldsymbol{u} \in \mathcal{X} : f(\boldsymbol{u}) = 1\},$$

then the closed hyperplane H separates the set M can the point $\boldsymbol{u}_0$.

8.3 The Dual Space of $C[a, b]^*$

Proposition 231. *Let* $-\infty < a < b < \infty$*. Then,* $F \in C[a, b]^*$ *iff there exists a function* $\rho : [a, b] \to \mathbb{R}$ *of bounded variation such that*

$$F(u) = \int_a^b u(x) \, \mathrm{d}\, \rho(x), \quad \forall u \in C[a, b]. \tag{8.6}$$

In a addition, $\|F\| = \mathrm{V}(\rho)$*, where* $\mathrm{V}(\rho)$ *denotes the total variation of* ρ.

The integral (8.6) represents a *Stieltjes* integral.

Example 232. *Let $w : [a, b] \to \infty$ be a continuous function, where $-\infty < a < b < \infty$. Set*

$$F(u) \triangleq \int_a^b u(x)w(x)\,\mathrm{d}\,x, \quad \forall u \in C[a, b].$$

Then, $F \in C[a, b]^*$ and

$$\|F\| = \int_a^b |w(x)|\,\mathrm{d}\,x.$$

8.4 Applications to The Moment Problem

8.4.1 The Finite Moment Problem

Let $-\infty < a < b < \infty$. We are given the real numbers $\mu_0, \mu_1, \cdots, \mu_N$ for fixed $N \geq 0$. We are looking for a function $\rho : [a, b] \to \mathbb{R}$ of bounded variation such that

$$\int_a^b x^k\,\mathrm{d}\,\rho(x) = \mu_k, \quad \forall k \in \{0, \cdots, N\}. \tag{8.7}$$

In terms of physics, we are looking for a charge ρ that has the prescribed moments $\mu_k, k = 0, \cdots, N$. In particular, μ_0 is equal to the total charge on the interval $[a, b]$.

Proposition 233. *The finite moment problem has always a solution.*

8.4.2 The Moment Problem

Let $-\infty < a < b < \infty$. We are given the real numbers $\mu_0, \mu_1, \cdots$. We are looking for a function $\rho : [a, b] \to \mathbb{R}$ of bounded variation such that

$$\int_a^b x^k\,\mathrm{d}\,\rho(x) = \mu_k, \quad \forall k \in \{0, 1, \cdots\} \tag{8.8}$$

Proposition 234. *The moment problem has a solution iff there is a constant $c > 0$ such that*

$$\left|\sum_{k=0}^{N} a_k\mu_k\right| \leq c \max_{x\in[a,b]} \left|\sum_{k=0}^{N} a_k x^k\right|, \qquad \forall a_k \in \mathbb{R}, \forall N \in \{0, 1, 2, \cdots\}$$

8.5 Minimum Norm Problems and Duality Theory

Along with the **primal problem**

$$\inf_{u\in\mathcal{L}} \|u - u_0\| = \alpha, \tag{8.9}$$

let us consider the **dual problem**

$$\sup_{u^*\in\mathcal{L}^\perp} \langle u^*, u_0\rangle = \beta \quad \text{s.t.} \quad \|u^*\| \leq 1 \tag{8.10}$$

where

$$\mathcal{L}^\perp \triangleq \{u^* \in \mathcal{X}^* : \langle u^*, u_0\rangle = 0, \forall u \in \mathcal{X}\}$$

is the orthogonal-complementary space of $\mathcal{L}$.

Theorem 235 (Minimum norm problem on the normed space $\mathcal{X}$)**.** *Let $\mathcal{L}$ be a linear subspace of the real normed space $\mathcal{X}$. We are given $u_0 \in \mathcal{X}$. Then the following conditions hold:*

① *Extremal values: $\alpha = \beta$.*

② *Dual problem: The dual problem (8.10) has a solution* u^*.

③ *Primal problem: Let* u^* *be a fixed solution of the dual problem (8.9). Then, the point* $u \in \mathcal{L}$ *is a solution of the primal problem (8.9) iff*

$$\langle u^*, u_0 - u \rangle = \|u - u_0\| . \tag{8.11}$$

Corollary 236. *If* $\dim(\mathcal{L}) < \infty$*, then the primal problem (8.9) always has a solution.*

Let $v \in \mathcal{L}$ and $v^* \in \mathcal{L}^\perp$ with $\|v^*\| \leq 1$. Then, from ① we obtain the two-sided *error estimate* for the minimal value α:

$$\langle v^*, u_0 \rangle \leq \alpha \leq \|v - u_0\|$$

Remark 237. *Let* $\dim(\mathcal{L}) = \infty$*, where* L *is a closed linear subspace of the real reflexive Banach space* $\mathcal{X}$ *(e.g.* X *is a real Hilbert space), and let* $u_0 \in \mathcal{X}$ *be given. Then, the primal problem (8.9) has a solution.*

In contrast to (8.9), we now consider the **modified primal problem**

$$\inf_{u^* \in \mathcal{L}^\perp} \|u^* - u_0^*\| = \alpha, \tag{8.12}$$

along with the dual problem

$$\sup_{u \in \mathcal{L}} \langle u_0^*, u \rangle = \beta \quad \text{s.t.} \quad \|u\| \leq 1. \tag{8.13}$$

Thus the primal problem (8.12) refers to the dual space $\mathcal{X}^*$, whereas the dual problem (8.13) refers to the original space $\mathcal{X}$.

Theorem 238 (Minimum norm problem on the dual space $\mathcal{X}^*$)**.** *Let* $\mathcal{L}$ *be a linear subspace of the real normed space* $\mathcal{X}$*. We are given* $u_0^* \in \mathcal{X}^*$*. Then the following conditions hold:*

① *Extremal values:* $\alpha = \beta$.

② *Primal problem: The primal problem (8.12) has a solution* u^*.

③ *Dual problem: Let* u^* *be a fixed solution of the primal problem (8.12). Then, the point* $u \in \mathcal{L}$ *with* $\|u\| \leq 1$ *is a solution of the dual problem (8.13) iff*

$$\langle u_0^* - u^*, u \rangle = \|u_0^* - u^*\| . \tag{8.14}$$

Let $-\infty < a \leq c \leq b < \infty$. Set

$$\delta_c(u) \triangleq u(c), \quad \forall u \in C[a, b].$$

Obviously, $\delta_c \in C[a, b]^*$ and $\|\delta_c\| = 1$.

Lemma 239. *Let* $u^* \in C[a, b]^*$ *be such that* $\|u^*\| \neq 0$*. Suppose that*

$$\langle u^*, u \rangle = \|u^*\| \, u \quad \textit{where} \quad \|u\| = \max_{x \in [a,b]} |u(x)| ,$$

and $u : [a, b] \to \mathbb{R}$ *is a continuous function such that* $|u(x)|$ *achieves its maximum at precisely* N *points of* $[a, b]$ *denoted by* $x_1, \cdots, x_N$*. Then, there exist real numbers* $\alpha_1, \cdots, \alpha_N$ *such that*

$$u^* = \alpha_1 \delta_{x_1} + \cdots + \alpha_N \delta_{x_N} ,$$

and $|\alpha_1| + \cdots + |\alpha_N| = \|u^*\|$.

8.6 Applications to Cebysev Approximation

8.6.1 Cebysev approximation of the function with polynomial

For the given continuous function $u_0 : [a, b] \to \mathbb{R}$ on the compact interval $[a, b]$, let us consider the following approximation problem:

$$\max_{x\in[a,b]} |u_0(x) - u(x)| = \min!, \quad u \in \mathcal{L} \tag{8.15}$$

where $\mathcal{L}$ denotes the set of all real polynomials $p(x)$ of degree $\partial(p(x)) \leq N$, for fixed $N \geq 1$. Problem (8.15) correspons to the so-called Cebysev approximation of the function u_0 by polynomials.

Proposition 240. *Problem (8.15) has a solution. If u is solution of (8.15), then*

$$|u_0(x) - u(x)|$$

achieves its maximum at least $N + 2$ points of $[a, b]$.

Proof.

- Set $\mathcal{X} \triangleq C[a, b]$ and $\|v\| = \max_{x\in[a,b]} |v(x)|$. Then, the original problem (8.15) can be written in the form
$$\|u_0 - u\| = \min!, \quad u \in \mathcal{L}. \tag{8.16}$$
Since $\dim(\mathcal{L}) < \infty$, this problem has a solution, by Corollary 236.

- We may assume that $u_0 \notin \mathcal{L}$. Otherwise, the statement is trivial. Let u be a solution of (8.16). Then, $\|u_0 - u\| > 0$. By the duality theory from Theorem 235, there exists a function $u^* \in C[a, b]^*$ such that
$$\langle u^*, u_0 - u\rangle = \|u_0 - u\| \tag{8.17}$$
along with $\|u^*\| = 1$ and
$$\langle u^*, p\rangle = 0, \quad \forall p \in \mathcal{L}. \tag{8.18}$$

- Suppose that $|u_0(x) - u(x)|$ achieves its maximum on $[a, b]$ at precisely the points $x_1, \cdots, x_M$, where $1 \leq M \leq N + 2$. It follows from (8.17) and Lemma 239 that there are real numbers $\alpha_1, \cdots, \alpha_M$ with $|\alpha_1| + \cdots + |\alpha_M| = 1$ such that
$$u^* = \alpha_1\delta_{x_1} + \cdots + \alpha_M\delta_{x_M}.$$
Assume that $\alpha_M \neq 0$. Choose a real polynomial p of degree N such that
$$p(x_1) = p(x_2) = \cdots = p(x_{M-1}) = 0 \quad \text{and} \quad p(x_M) \neq 0.$$
This is possible, since $M - 1 \leq N$. Then, $p \in \mathcal{L}$ and $\langle u^*, p\rangle \neq 0$, contradicting (8.18).■

8.6.2 Uniqueness of the Cebysev Approximation

Set $\mathcal{X} = C[a, b]$, where $-\infty < a < b < \infty$ and

$$\|u - v\| \triangleq \max_{x\in[a,b]} |u(x) - v(x)|.$$

Let $\mathcal{L}$ be a finite-dimensional linear subspace of X with $\dim(\mathcal{L}) = N+1$. By definition, $\mathcal{L}$ satisfies the *Haar condition* iff each nonzero function $v : [a, b] \to \mathbb{R}$ from $\mathcal{L}$ has at most N zeros. For given $u \in \mathcal{X}$, the approximation problem

$$\hat{v} = \arg\min_{v\in\mathcal{L}} \|u - v\| \tag{8.19}$$

has a solution. In addition, the following can be shown.

① If $\mathcal{L}$ satisfies the Harr condition, then the solution $\hat{v}$ of (8.19) is unique.

② Suppose that $\mathcal{L}$ satisfies the Harr condition. Let $u \notin \mathcal{L}$, and let $v \in \mathcal{L}$ be a given function. Suppose that there is a finite set of points $a \le t_1 < t_2 < \cdots < t_{N+1} \le b$ such that

$$u(t_j) - v(t_j), \quad j \in \{1, 2, \cdots, N+3\},$$

attains alternatively the values $\|u - v\|$ and $-\|u - v\|$ at consecutive poiints t_j.

Then, $\hat{v}$ is unique solution of (8.19).

8.7 Applications to the Optimal Control

We want to study the motion of a vertically ascending rocket that reaches a given altitude h with minimum fuel expenditure.

The motion $x = x(t)$ of the rocket is governed by the equation

$$\begin{aligned} m\ddot{x}(t) &= F(t) - mg, \quad t \in (0, T), \\ x(0) &= \dot{x}(0) = 0, \quad x(T) = h, \end{aligned} \tag{8.20}$$

where m is the mass of the rocket, mg is the force of gravity, and $F(t)$ is the rocket force. We neglect the loss of mass by the burning of fuel. To simplify notation, we choose physical units with $m = g = 1$.

Let us measure the minimal fuel expenditure during the time interval $[0, T]$ through the integral

$$\int_0^T |F(t)| \, \mathrm{d}\, t$$

over the rocket force F. First let $T > 0$ be fixed. Then, the minimal fuel expenditure $\alpha(T)$ during the time interval $[0, T]$ is given by a solution of the folowing minimum problem:

$$\hat{F} = \arg\min_F \int_0^T |F(t)| \, \mathrm{d}\, t, \quad \alpha(T) = \int_0^T \left|\hat{F}(t)\right| \mathrm{d}\, t \tag{8.21}$$

where we vary over all integrable funcitons $F : [0, T] \to \mathbb{R}$. We now choose the final time $\alpha(T)$ in such a way that $\alpha(T)$ becomes minimal, that is

$$\alpha(T) = \min! \tag{8.22}$$

Integration of (8.20) yields

$$x(t) = \int_0^t (t - \tau) F(\tau) \, \mathrm{d}\, \tau - \frac{1}{2} t^2,$$

and hence

$$h = \int_0^T (T - \tau) F(\tau) \, \mathrm{d}\, \tau - \frac{1}{2} T^2. \tag{8.23}$$

In summary, for a given altitude $h > 0$, we have to determine the optimal thrust program $F(\cdot)$ and the final time T as a solution of problems (8.21) throught (8.22).

This formulation has the following shortcoming. If we consider only classical force function F, then an impulse at time t of the form

$$F = \delta(t)$$

is excluded. However, we expect that such thrust programs may be of importance. For this reason, let us consider the following generalized problem for functionals:

❶ For a given altitude h and fixed final time $T > 0$, we are looking for a solution F of the following minimal problem:

$$\min \|F\| = \alpha(T), \quad F \in C[0,T]^*, \tag{8.24}$$

along with the side condition

$$h = F(w) - \frac{1}{2}T^2, \quad w = T - t. \tag{8.25}$$

❷ We determine the final time T in such a way that

$$\alpha(T) = \min!$$

Observe that condition (8.24) generalize (8.21). In fact, if the functional $f \in C[0,T]^*$ has the following special form

$$f(u) = \int_0^T u(t)F(t)\,\mathrm{d}\,t, \quad \forall u \in C[0,T]$$

where the fixed function $F : [0,T] \to \mathbb{R}$ is continuous, then

$$\|f\| = \int_0^T |F(t)|\,\mathrm{d}\,t.$$

Proposition 241. *Problem ❶+❷ has the following solution*

$$f = T\delta_0 \quad \textit{and} \quad T = \sqrt{2h},$$

with the minimal "fuel expenditure" $\|f\| = T$.

This solution corresponds to an impulse at the initial time $t = 0$. Proposition 241 shows that, in control theory, it is quite natural to use minimum problems with respect to functionals.

Proof.

- Solution of ❶. Let $\mathcal{X} = C[a,b]$ and $\mathcal{L} = \operatorname{span} w$. By the Hahn-Banach theorem, there exists a functional $f_0 \in C[a,b]^*$ such that

  $$f_0(w) = h + \frac{T^2}{2}.$$

  Then, the condition (8.25) says that $(f_0 - f)(w) = 0$, i.e., $(f_0 - f) \in \mathcal{L}^\perp$. Consequently, problem ❶ is equivalent to the primal problem:

  $$\min_{(f_0 - f) \in \mathcal{L}^\perp} \|(f_0 - f) - f_0\| = \alpha(T). \tag{8.26}$$

  By Theorem 238, the *dual problem* reads as follows:

  $$\sup_{u \in \operatorname{span}\{w\}} f_0(u) = \alpha(T) \quad \text{s.t.} \quad \|u\| \le 1. \tag{8.27}$$

  Let us solve (8.26) and (8.27).

  – Observe that the dual problem (8.27) is one-dimensional. Since $\|w\| = \max\limits_{t \in [0,T]} |w(t)| = T$, (8.27) has the solution

  $$u = T^{-1}w.$$

  Hence

  $$\alpha(T) = f_0(T^{-1}w) = T^{-1}h + 2^{-1}T.$$

  Explicitly,

  $$u(t) = T^{-1}(T - t), \quad \forall t \in [0,T].$$

– By Theorem 238, the primal problem (8.26) has a solution $f_0 - f \in \mathcal{L}^\perp$. Hence

$$\|f\| = \alpha(T) \quad \text{and} f_0(w) = f(w).$$

Since $u = T^{-1}w$, the functional $f \in C[a,b]^*$ satisfies the equation $f(u) = f_0(u) = \alpha(T)$, i.e.,

$$f(u) = \|f\| \cdot \|u\|, \tag{8.28}$$

because $\|u\| = 1$. Since the functional $u(\cdot)$ achieves its maximum on $[0,T]$ precisely at the point $t = 0$, it follows from (8.28) and Lemma 239 that

$$F = \beta\delta_0$$

for some real number β with $|\beta| = \|f\|$. Since $\|\delta_0\| = 1$, this implies $f = \pm\|f\|\,\delta_0$. From $f(w) = f_0(w) > 0$ and $\delta_0(w) = w(0) > 0$, we get

$$f = \|f\|\,\delta_0, \quad \text{i.e.} \quad f = \alpha(T)\delta_0.$$

- Solution of problem ❷. It follows from $\alpha'(T) = -T^{-2}h + 2^{-1} = 0$ that the problem

$$\alpha(T) = \min!$$

has the solution $T = (2h)^{\frac{1}{2}}$. Hence $\alpha(T) = T^{-1}h + 2^{-1}T = \sqrt{2h} = T$.

Chapter 9

Principles of Linear Functional Analysis

Linear functional analysis is based on two important principles, viz., the HAHN-BANACH theorem, and the BAIRE THEOREM. The most important consequences of the Baire theorem are the following (See Fig. 9.1):

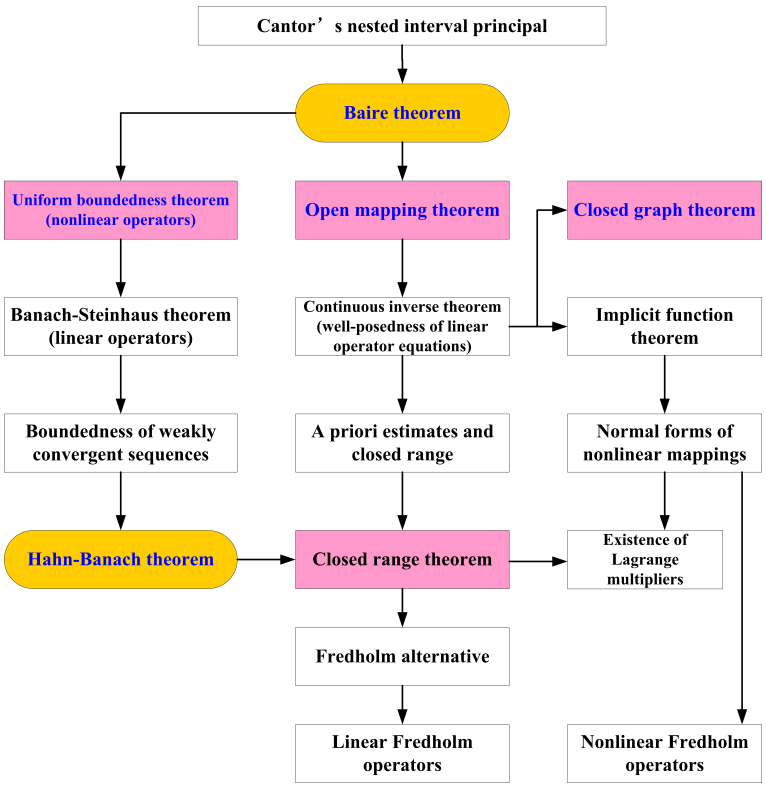


Figure 9.1: Baire Theorem and its relatives

(a) the UNIFORM BOUNDEDNESS theorem;

(b) the OPEN MAPPING theorem;

(c) the CLOSED GRAPH theorem; and

(d) the CLOSED RANGE theorem

These fundamental results were proved by Banach in the late 1920s. The prototype of the Baire theorem was proved by Baire in 1899 before the creation of functional analysis.

9.1 Baire Theorem

9.1.1 Category Sets and Nested Interval Principle

We now give some basic specifications as follows:

Definition 242. *Let M be a subset of a normed space $\mathcal{X}$ over $\mathbb{F}$. Then*

(i) M is called NOWHERE DENSE *in $\mathcal{X}$ iff*

$$\mathrm{Int}(\overline{M}) = \emptyset,$$

that is, the closure $\overline{M}$ of M does not contain any interior points.

(ii) M is said to be of the FIRST CATEGORY *in $\mathcal{X}$ iff M is the countable union of nowhere dense subsets M_n of X, that is,*

$$M = \bigcup_{n=1}^{\infty} M_n.$$

Sets of the first category are also called MEAGER.

(iii) M is said to be of the SECONDARY CATEGORY *in $\mathcal{X}$ iff M is not of the first category. Such sets are also called* FAT.

For illustration, we have

- Each finite set $\{x_1, \cdots, x_n\}$ in $\mathbb{R}$ is nowhere dense in $\mathbb{R}$.
- Each at most countable subset of $\mathbb{R}$ is of the first category in $\mathbb{R}$.
- The set of rational numbers $\mathbb{Q}$ is of the first category in $\mathbb{R}$.
- Each nonempty open subset of $\mathbb{R}$ (e.g., $\mathbb{R}$ itself) is of the second category in $\mathbb{R}$.

Proposition 243 (Cantor's nested interval principle)**.** *Let $M_1 \supset M_2 \supset \cdots$ be a sequence of nonempty closed subsets of M_n of a Banach space $\mathcal{X}$ such that*

$$\lim_{n\to\infty} \mathrm{diam}(M_n) = 0. \tag{9.1}$$

Then there exists a unique point u with $u \in M_n$ for all n.

PROOF.

- Existence. Choose a point $u_n \in M_n$ for each n. By (9.1), the sequence (u_n) is Cauchy, and hence there is a point u such that $u_n \to u$ as $n \to \infty$. Since $u_n \in M_k$ for all $n \geq k$ and the set M_k is closed, $u \in M_k$ for each k.
- Uniqueness. Let $u, v \in M_n$ for all n. By (9.1), $\|u - v\|$ is arbitrarily small. Hence $u = v$. ■

9.1.2 Second Category sets and Baire theorem

Theorem 244 (The Baire theorem)**.** *Each nonempty open subset U of a Banach space $\mathcal{X}$ over $\mathbb{F}$ (e.g., $U = \mathcal{X}$) is of the second category in $\mathcal{X}$.*

PROOF.

- If U were not of the second category, then U would be of the first category. Then there would exist a family $\{M_n\}$ of sets in $\mathcal{X}$ such that

$$U = \bigcup_{n=1}^{\infty} M_n \quad \text{and} \quad \text{Int}(\overline{M}) = \emptyset, \quad \forall n.$$

  Let us introduce the closed ball

$$\mathcal{B}_r(a) \triangleq \{u \in \mathcal{X} : \|u - a\| \leq r\}$$

  of radius $r > 0$. First choose a point $a \in U$. Since the set U is open,

$$\mathcal{B}_r(a) \subset U \quad \text{for some} \quad r > 0.$$

  Since $\text{Int}(\overline{M_1}) = \emptyset$, there exists a point $a_1 \in \text{Int}\,\mathcal{B}_r(a)$ such that $\text{dist}(a_1, \overline{M_1}) > 0$. Thus, there is a number r_1 with $0 < r_1 < \frac{r_0}{2}$ such that

$$\overline{M_1} \cap \mathcal{B}_{r_1}(a_1) = \emptyset.$$

  Otherwise, $\text{dist}(b, \overline{M_1}) = 0$ for all $b \in \text{Int}(\mathcal{B}_r(a))$. Since $\overline{M_1}$ is closed, this implies $b \in \overline{M_1}$ for all $b \in \text{Int}(\mathcal{B}_r(a))$, and hence $\text{Int}(\overline{M_1}) \neq \emptyset$. This is a contradiction.

- Continuing this argument, we obtain a sequence of balls

$$\mathcal{B}_r(a) \supset \mathcal{B}_{r_1}(a_1) \supset \mathcal{B}_{r_2}(a_2) \cdots \quad \lim_{n\to\infty} r_n = 0 \tag{9.2}$$

  such that

$$\overline{M_n} \cap \mathcal{B}_{r_n}(a_n) = \emptyset, \quad \forall n = 1, 2, \cdots. \tag{9.3}$$

  It follows from (9.2) and the NESTED INTERVAL PRINCIPLE that there exists a point u with $u \in \mathcal{B}_{r_n}(a_n)$ for all n. By (9.3), $u \notin M_n$ for all n. This is a contradiction to

$$u \in \mathcal{B}_r(a) \subset U = \bigcup_{n=1}^{\infty} M_n. \qquad \blacksquare$$

9.1.3 Existence of Nondifferentiable Continuous Functions

Existence Principle

Let M be a subset of a Banach space $\mathcal{X}$, and let M be of the first category in $\mathcal{X}$. Then, there exists a point $u \in \mathcal{X}$ such that

$$u \notin M.$$

Moreover, the set $\mathcal{X} - M$ is of the second category in $\mathcal{X}$.

Proposition 245 (Weierstrass function)**.** *There exists a nondifferentiable continuous function $f : [0, 1] \to \mathbb{R}$. (Actually, this kind of functions are related with fractals.)*

PROOF.

- By the Baire theorem (Theorem 244), $\mathcal{X}$ is of the second category. Since $\mathcal{X} = M \cup (\mathcal{X} - M)$ and M is of the first category, the set $\mathcal{X} - M$ must be of the second category. Note that the union of two sets of the first category yields a set of the first category. $\blacksquare$

Proposition 246. *Let the set*

$$M = \{f \in C[0,1] : \exists x_* \in [0,1] \text{ such that the right-hand derivative } f'(x_*+0) \text{ exists}\},$$

then M is of the first category in $C[0,1]$.

This implies that the set $C[0,1] - M$ is of the second category in the Banach space $C[0,1]$. Hence Proposition 246 implies Proposition 245. Roughly speaking, Proposition 246 tells us that "most" continuous functions $f : [0,1] \to \mathbb{R}$ are nondifferentiable. In 1806 Ampère tried to prove that "each continuous function is differentiable." More than fifty years later, Weierstrass showed that such a statement is wrong.

PROOF.

- Let M_n denote the set of all functions $f \in C[0,1]$ such that there exists a point $x_* \in [0,1]$ with
$$|f(x_*+h) - f(x_*)| \le nh, \quad \forall h \in [0,1] \text{ with } x_*+h \le 1. \tag{9.4}$$
If $f \in M$, then $f'(x_*+0)$ exists and f is continuous on $[0,1]$. Thus, $f \in M_n$ for some n, and hence
$$M \subset \bigcup_{n=1}^{\infty} M_n.$$
We have to show that each set M_n is nowhere dense in $C[0,1]$. Then M is of the first category in $C[0,1]$.
  - We first prove that M_n is closed. To this end, let (f_k) be a sequence in M_n such that $f_k \in M_n$ for all $k = 1, 2, \cdots$. Then there exists points x_k such that
$$|f(x_k+h) - f(x_k)| \le nh, \quad \forall h \in [0,1] \text{ with } x_k+h \le 1, \quad k = 1, 2, \cdots \tag{9.5}$$
Since $x_k \in [0,1]$ for all k, there is a subsequence, again denoted by (x_k), such that $x_k \to x_*$ as $k \to \infty$. Letting $k \to \infty$ in (9.5), we have[1]
$$|f(x_*+h) - f(x_*)| \le nh, \quad \forall h \in [0,1] \text{ with } x_*+h \le 1.$$
Hence $f \in M_n$, i.e., M_n is closed.
  - We now show that $\mathrm{Int}(M_n) = \emptyset$. Let $f \in M_n$, $\forall \varepsilon > 0$, there exists a piecewise linear, continuous function $g : [0,1] \to \mathbb{R}$ such that
$$\|f - g\| \triangleq \max_{x \in [0,1]} |f(x) - g(x)| < \varepsilon$$
and $|g'(x+0)| > n$ for all $x \in [0,1]$. This implies $g \notin M_n$. Hence f is not an interior point of M_n. ■.

9.2 Uniform Boundedness Theorem

9.2.1 Theory

Theorem 247 (Uniform Boundedness Theorem)**.** *Let $\mathcal{F}$ be a nonempty set of continuous maps*

$$F : \mathcal{X} \to \mathcal{Y}$$

where $\mathcal{X}$ is a Banach space over $\mathbb{F}$ and $\mathcal{Y}$ is a normed space over $\mathbb{F}$. Suppose that

$$\sup_{F \in \mathcal{F}} \|Fu\| < \infty, \quad \forall u \in \mathcal{X}.$$

[1] This limit exists, since $f_k(x) \to f(x)$ as $k \to \infty$ uniformly on $[0,1]$ and f is uniformly continuous on $[0,1]$.

Then there exists a closed ball $\mathcal{B}$ in $\mathcal{X}$ of positive radius such that

$$\sup_{u\in\mathcal{B}}\left(\sup_{F\in\mathcal{F}}\|Fu\|\right)<\infty.$$

PROOF.

- Define the set M_k as
$$M_k \triangleq \bigcap_{F\in\mathcal{F}}\{u\in\mathcal{X}:\|Fu\|\le k\}.$$
Obviously,
$$\mathcal{X}=\bigcup_{n=1}^{\infty}M_n.$$
Since F is continuous, the set M_k is closed.[2]

- By the Baire theorem (Theorem 244), $\mathrm{Int}(M_k)\neq\emptyset$ for some k. Hence the set M_k contains a closed ball $\mathcal{B}$ of positive radius. Then, by the definition of M_k,
$$\sup_{u\in\mathcal{B}}\left(\sup_{F\in\mathcal{F}}\|Fu\|\right)\le k. \qquad \blacksquare$$

Corollary 248 (Banach-Steinhaus Theorem)**.** *Let $\mathcal{L}$ be a nonempty set of linear continuous operators*

$$L:\mathcal{X}\to\mathcal{Y},$$

where $\mathcal{X}$ is a Banach space over $\mathbb{F}$ and $\mathcal{Y}$ is a normed space over $\mathbb{F}$. Suppose that

$$\sup_{L\in\mathcal{L}}\|Lu\|<\infty,\quad \forall u\in\mathcal{X},$$

then $\sup_{L\in\mathcal{L}}\|L\|<\infty$.

PROOF.

- By Theorem 244, there exists a closed ball $\mathcal{B}$ of positive radius in $\mathcal{X}$ such that
$$\sup_{x\in\mathcal{B}}\left(\sup_{L\in\mathcal{L}}\|Fx\|\right)<\infty. \tag{9.6}$$
Since L is linear, we get
$$\|Lr(u-u_0)\|\le r\|Lu\|+r\|Lu_0\|,\quad \forall r>0,u_0\in\mathcal{X}.$$
Thus, relation (9.6) remains true if B denotes the closed unit ball. Therefore,
$$\sup_{L\in\mathcal{L}}\|L\|=\sup_{L\in\mathcal{L}}\left(\sup_{\|u\|\le 1}\|Lu\|\right)<\infty. \qquad \blacksquare$$

Proposition 249. *Let (L_n) be a sequence of linear continuous operators*

$$L_n:\mathcal{X}\to\mathcal{Y},$$

where $\mathcal{X}$ is a Banach space over $\mathbb{F}$ and $\mathcal{Y}$ is a normed space over $\mathbb{F}$. Then the following two conditions are equivalent:

[2]Actually, it is obvious that the set $\{u\in\mathcal{X}:\|Fu\|\le k\}$ is closed. Furthermore, observe that the intersection of an arbitrary number of closed sets is again closed.

① *There exists a linear continuous operator* $L : \mathcal{X} \to \mathcal{Y}$ *such that*

$$Lu = \lim_{n\to\infty} L_n u, \quad \forall u \in \mathcal{X}.$$

② *There is a dense subset* D *of* $\mathcal{X}$ *such that* $\lim_{n\to\infty} L_n u$ *exists for all* $u \in D$, *and* $\sup_n \|L_n\| < \infty$.

Proof.

- ①$\Longrightarrow$ ②. This follows from the Banach-Steinhaus theorem (Corollary 248).
- ②$\Longrightarrow$ ①. Let $u \in \mathcal{X}$. Then, for each $\varepsilon > 0$, there exists a point $v \in D$ such that

$$\|u - v\| < \varepsilon.$$

Since $(L_n v)$ is Cauchy,

$$\|L_n v - L_m v\| < \varepsilon, \quad \forall n, m \geq n_0(\varepsilon).$$

Consequently

$$\begin{aligned}\|L_n u - L_m u\| &= \|(L_n u - L_n v) + (L_n v - L_m v) + (L_m v - L_m u)\| \\ &\leq \|L_n u - L_n v\| + \|L_n v - L_m v\| + \|L_m v - L_m u\| \\ &\leq 2\left(\sup_n \|L_n\|\right)\|u - v\| + \varepsilon, \quad \forall n, m \geq n_0(\varepsilon).\end{aligned}$$

Thus, the sequence $(L_n u)$ is Cauchy and is hence convergent. Define

$$Lu \triangleq \lim_{n\to\infty} L_n u.$$

Obviously, the operator $L : \mathcal{X} \to \mathcal{Y}$ is linear. Moreover,

$$\|Lu\| \leq \left(\sup_n \|L_n\|\right)\|u\|.$$

In other words, L is also continuous. ■

Weak Convergence

Let (u_n) be a sequence in the normed space $\mathcal{X}$ over $\mathbb{F}$. Then the following two conditions are equivalent:

(i) $u_n \to u$ as $n \to \infty$.

(ii) The sequence $(\|u_n\|)$ is bounded, and there is a dense subset D of $\mathcal{X}^*$ such that

$$\langle f, u_n\rangle \to \langle f, u\rangle \quad \text{as } n \to \infty \text{ for all } f \in D.$$

Actually, we can set $L_n f \triangleq \langle f, u_n\rangle = f(u_n)$ for all $f \in \mathcal{X}^*$ and fixed n. Since

$$|L_n f| \leq \|f\|\,\|u_n\|, \quad \forall f \in \mathcal{X}^*,$$

the operator $L_n : \mathcal{X}^* \to \mathbb{F}$ is linear and continuous. By Corollary 226 we have

$$\|L_n\| = \|u_n\| = \max_{f\in\mathcal{X}^*, \|f\|\leq 1} |\langle f, u\rangle|, \quad \forall n.$$

The assertion follows now from Proposition 250. Note that $\mathcal{X}^*$ is a Banach space.

9.2.2 Cubature Formulas and Numerical Integration

Let $-\infty < a = x_0 < x_1 < \cdots < x_n = b < \infty$, where $\{x_k\}$ are points which constructs a partition of the interval $[a, b]$. By a cubature formula, we understand a formula of the following form:

$$\int_a^b u(x)\,\mathrm{d}\,x = \mathsf{L}_n u + r_n(u), \tag{9.7}$$

where

$$\mathsf{L}_n u \triangleq \sum_{k=0}^{n} c_k u(x_k), \quad n = 0, 1, 2, \cdots,$$

and $r_n(u)$ denotes the remainder. Our problem is to choose the real numbers c_k in such a way that we obtain a convergent cubature formula for the given function u:

$$\lim_{n\to\infty} r_n(u) = 0.$$

Proposition 250. *The following two conditions are equivalent:*

- *The cubature formula (9.7) is convergent for all continuous functions $u : [a, b] \to \mathbb{R}$.*
- *The cubature formula (9.7) is convergent for all polynomials u and*

$$\sup_{n\geq 1} \sum_{k=0}^{n} |c_k| < \infty. \tag{9.8}$$

Proof.

- Let $\mathcal{X} = C[a, b]$, and set

$$\mathsf{L}u \triangleq \int_a^b u(x)\,\mathrm{d}\,x.$$

  Then, the operators $\mathsf{L}, \mathsf{L}_n : C[a, b] \to \mathbb{R}$ are linear and continuous. Moreover,

$$\|\mathsf{L}_n\| = \sum_{k=0}^{n} |c_k|, \quad n = 1, 2, \cdots. \tag{9.9}$$

  – To prove this, let $u \in C[a, b]$. Fix the number n. Then

$$|\mathsf{L}_n u| \leq \sum_{k=0}^{n} |c_k| \max_{x\in[a,b]} |u(x)| = \sum_{k=0}^{n} |c_k| \, \|u\|.$$

  – Furthermore, let us construct a piecewise linear, continuous function $w : [a, b] \to \mathbb{R}$ by prescribing the values

$$w(x_k) \triangleq \operatorname{sign}(c_k), \quad k = 0, \cdots, n$$

  at all the node points x_k. Then

$$|\mathsf{L}_n w| = \left| \sum_{k=0}^{n} c_k \operatorname{sign}(c_k) \right| = \sum_{k=0}^{n} |c_k| \, \|w\|,$$

  since $\|w\| = 1$. This yields (9.9).

- By the Weierstrass approximation theorem, the set of polynomials is dense in the Banach space $C[a, b]$, Therefore, the assertion follows from the Banach-Steinhaus theorem.

Corollary 251. *Suppose that all the numbers c_k are nonnegative and that the cubature formula is exact for the function $u \equiv 1$; then condition (9.8) is satisfied.*

In fact, letting $u(x) \equiv 1$ in (9.7), we get

$$\sum_{k=0}^{n} c_k = \int_a^b \mathrm{d}\,x.$$

Trapezoid Formula

Let $x_k \triangleq \frac{k(b-a)}{n}, k = 0, 1, \cdots$, then

$$L_n u \triangleq \frac{b-a}{n}\left[\frac{u(b)+u(a)}{2} + u(x_1) + \cdots + u(x_{n-1})\right] \tag{9.10}$$

is called the trapezoid formula, where $n = 1, 2, \cdots$. Moreover, we have:

- $\forall u \in C^2[a, b]$, we get the following error estimates:

$$|r_n(u)| \leq \frac{(b-a)^3}{12n^2} \max_{x\in[a,b]} |u''(x)|, \quad n = 1, 2, \cdots \tag{9.11}$$

- $\forall u \in C[a, b]$, the trapezoid formula converges as $n \to \infty$.

Proof.

- We set $\alpha = x_k$ and $\beta = x_{k+1}$. Let

$$r \triangleq \int_\alpha^\beta u(x)\,\mathrm{d}\,x - \frac{\beta}{\alpha}[u(\alpha) + u(\beta)]$$

For given y with $\alpha < y < \beta$, define the linear function

$$p(x) \triangleq u(\alpha) + (x-\alpha)\frac{u(\beta)-u(\alpha)}{\beta-\alpha}$$

and set

$$\rho(x) \triangleq u(x) - p(x) - \frac{u(y)-p(y)}{(y-\alpha)(y-\beta)}(x-\alpha)(x-\beta). \tag{9.12}$$

Then, $\rho(\alpha) = \rho(y) = \rho(\beta) = 0$. By the mean value theorem, this implies the existence of numbers ξ and η with $\alpha < \xi < y < \eta < \beta$ such that

$$\rho'(\xi) = \rho'(\eta) = 0.$$

Again by the mean value theorem, there is a number ζ with $\xi < \zeta < \eta$ such that

$$\rho''(\zeta) = 0.$$

According to (9.12), this implies

$$u''(\zeta) - 2\frac{u(y)-p(y)}{(y-\alpha)(y-\beta)} = 0,$$

and hence

$$u(x) - p(x) = \frac{u''(\zeta(x))}{2}(x-\alpha)(x-\beta) \quad \forall x \in [\alpha, beta].$$

Integration yields

$$\left|\int_\alpha^\beta u(x)\,\mathrm{d}\,x - \int_\alpha^\beta p(x)\,\mathrm{d}\,x\right| \leq \frac{1}{2}\max_{x\in[\alpha,\beta]}|u''(x)| \int_\alpha^\beta (x-\alpha)(x-\beta)\,\mathrm{d}\,x.$$

Hence

$$\left|\int_\alpha^\beta u(x)\,\mathrm{d}\,x - \frac{1}{2}(\beta-\alpha)[u(\alpha)+u(\beta)]\right| \leq \max_{x\in[a,b]}|u''(x)|\,\frac{(\beta-\alpha)^3}{12}.$$

Recall that $\alpha = x_k = \frac{k(b-a)}{n}$ and $\beta = x_{k+1} = \frac{(k+1)(b-a)}{n}$; then summation over k yields (9.11).

- If u is a polynomial, then $r_n(u) \to 0$ as $n \to \infty$ by (9.11). Moreover, the trapezoid formula (9.10) is exact for $u \equiv 1$, again by (9.11).

- The assertion (ii) is a special case of Proposition 250 and Corollary 251. ■

9.3 Open Mapping Theorem

Theorem 252 (Banach's Open Mapping Theorem)**.** *Let $A : \mathcal{X} \to \mathcal{Y}$ be a linear continuous operator, where $\mathcal{X}$ and $\mathcal{Y}$ are Banach spaces over $\mathbb{F}$. Then, the following conditions are equivalent:*

① *A is surjective.*

② *A is open, that is, A maps open sets onto open sets.*

PROOF.

- ①$\Longrightarrow$②. Let us introduce the open ball $\mathcal{B}_R \triangleq \{u \in X : \|u\| < R\}$.
  - Step-1. Since A is surjective,
  $$\mathcal{Y} = \bigcup_{n=1}^{\infty} \overline{A(\mathcal{B}_n)}. \tag{9.13}$$
  By the Baire theorem, there is some index m such that the closure $\overline{\mathcal{B}_m}$ is not nowhere dense. Thus, there is a point $w \in \mathcal{Y}$ such that
  $$w \in \mathrm{Int}(\overline{A(\mathcal{B}_m)}).$$
  Since A is surjective, there exists some point $u \in \mathcal{X}$ such that $w = Au$. In consequence,
  $$0 \in \mathrm{Int}\, \overline{A(\mathcal{B}_m - u)}.$$
  Finally, choose the number $r > 0$ so large that $\mathcal{B}_m - u \subset \mathcal{B}_r$. Then
  $$0 \in \mathrm{Int}\, \overline{A(\mathcal{B}_r)}. \tag{9.14}$$
  - Step-2. Let us prove the stronger result that
  $$0 \in \mathrm{Int}\, A(\mathcal{B}_r) \tag{9.15}$$
  Condition (9.14) means that there is some number $r > 0$ such that
  $$\|v\| \le r \quad \text{with} \quad \text{implies} \quad v \in \overline{A(\mathcal{B}_r)}.$$
  In particular, this implies the following
  $$\forall v \in \mathcal{Y} \text{ with } \|v\| < r, \quad \exists u \in \mathcal{B}_r \text{ such that } \|v - Au\| < \frac{r}{2}. \tag{9.16}$$
  To prove (9.15) it is sufficient to show that, for each $v \in \mathcal{Y}$ with $\|v\| < r$, there is some point $u \in \mathcal{B}_{3R}$ such that
  $$v = Au \tag{9.17}$$
  In fact, this means that $0 \in \mathrm{Int}(A(\mathcal{B}_{3R}))$, and hence we get (9.15), by the linearity of A. Let $v \in \mathcal{Y}$ be given with $\|v\| < r$. Using (9.16), we construct a sequence (u_n) in the ball $\mathcal{B}_r$ such that $v_0 \triangleq v$ and
  $$\|2(v_n - Au_n)\| < r, \quad v_{n+1} = 2(v_n - Au_n), \quad n = 0, 1, \cdots.$$
  Hence
  $$2^{-n-1} v_{n+1} = 2^{-n} v_n - A(2^{-n} u_n), \quad n = 0, 1, \cdots.$$
  This implies
  $$A\left(\sum_{n=0}^{m} 2^{-n} u_n\right) = v_0 - 2^{-m-1} v_{m+1}. \tag{9.18}$$

Since $\sum_{n=0}^{m} \|2^{-n}u_n\| \le \sum_{n=0}^{m} 2^{-n}R \le 2R$, the series

$$u \triangleq \sum_{n=0}^{\infty} 2^{-n}u_n$$

is convergent. Hence $\|u\| < 3R$. Letting $m \to \infty$ in (9.18), we get (9.17).

– Step-3: Let U be an open subset of $\mathcal{X}$, and let $u \in U$. Then there is some $r > 0$ such that

$$u + r\mathcal{B}_R \subset U.$$

Using the linearity of the operator A and (9.15), we obtain

$$Au \in \mathrm{Int}(A(u + r\mathcal{B}_R)),$$

and hence $Au \in \mathrm{Int}(A(U))$. Thus, the set $A(U)$ is open.

- ②$\Longrightarrow$①. Since A is open, the set $A(\mathcal{X})$ contains an interior point. This implies $A(\mathcal{X}) = \mathcal{Y}$, by the linearity of A.

Proposition 253 (Banach's Continuous Inverse Theorem)**.** *Let $A : \mathcal{X} \to \mathcal{Y}$ be a linear continuous operator, where $\mathcal{X}$ and $\mathcal{Y}$ are Banach spaces over $\mathbb{F}$. If the inverse operator*

$$A^{-1} : \mathcal{Y} \to \mathcal{X}$$

exists, then it is continuous.

Proof.

- By the open mapping theorem (Theorem 252), the operator A is open. Thus, if the set W is open in $\mathcal{X}$, then $A(W)$ is open in $\mathcal{Y}$. A general result about continuous maps on topological spaces tells us that this implies the continuity of A^{-1}.

- A direct proof goes like this. Set $\mathcal{B}_\varepsilon = \{u \in \mathcal{X} : \|u\| < \varepsilon\}$. Since A^{-1} is linear, it is sufficient to prove that A^{-1} is continuous at the point $v = 0$. In fact, for each given $\varepsilon > 0$, the set $A(\mathcal{B}_\varepsilon)$ is open, since A is open. Hence $0 \in \mathrm{Int}(A(\mathcal{B}_\varepsilon))$ because $A(0) = 0$. Thus, there is some number $\delta(\varepsilon) > 0$ such that $\|Au\| < \delta(\varepsilon)$ implies $u \in \mathcal{B}_\varepsilon$, that is,

$$\|Au\| < \delta(\varepsilon) \quad \text{implies} \quad \|u\| < \varepsilon.$$

Hence $\|v\| < \delta(\varepsilon)$ implies $\left\|A^{-1}v\right\| < \varepsilon$. This means that the operator A^{-1} is continuous at $v = 0$. ■

The following corollary reprfesents an important reformulation of Proposition 253 in terms of the operator equation

$$Au = v, \quad u \in \mathcal{X}. \tag{9.19}$$

Corollary 254 (Well-posedness Principle)**.** *Let $A : \mathcal{X} \to \mathcal{X}$ be a linear continuous operator, where $\mathcal{X}$ and $\mathcal{Y}$ are Banach spaces over $\mathbb{F}$. Then the following two condinitions are equivalent:*

(i) Eq. (9.19) is well posed, that is, by definition, for each given $v \in \mathcal{Y}$, Eq. (9.19) has a unique solution u, which depends continuously on v.

(ii) For each $v \in \mathcal{Y}$, Eq. (9.19) has a solution u, and $Aw = 0$ implies $w = 0$.

9.4 Closed Graph Theorem

9.4.1 Theory

Definition 255. *Let $\mathcal{X}$ and $\mathcal{Y}$ be normed spaces over $\mathbb{F}$. By the graph* $\operatorname{Graph}(A)$ *of the operator*

$$A : \operatorname{Dom}(A) \subset \mathcal{X} \to \mathcal{Y},$$

we mean the subset

$$\operatorname{Graph}(A) \triangleq \{(u, Au) : u \in \operatorname{Dom}(A)\}$$

of the product space $\mathcal{X} \times \mathcal{Y}$.

The operator A is called graph-closed iff $\operatorname{Graph}(A)$ is closed in $\mathcal{X}$ in $\mathcal{X} \times \mathcal{Y}$. This means that for each sequence (u_n) in the set $\operatorname{Dom}(A)$ it follows from

$$u_n \to u \quad \text{in} \quad \mathcal{X} \quad \text{as} \quad n \to \infty \tag{9.20}$$

and

$$Au_n \to v \quad \text{in} \quad \mathcal{Y} \quad \text{as} \quad n \to \infty$$

that $u \in \operatorname{Dom}(A)$ and $v = Au$.

Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}$. Then the following are true:

❶ The function $A : [1, 2] \to \mathbb{R}$ pictured in Fig. ? is continuous and graph-closed in $\mathbb{R} \times \mathbb{R}$.

❷ The function $A : [0, 1] \to \mathbb{R}$ pictured in Fig. ? is continuous but is not graph-closed.

❸ The function $A : \mathbb{R} \to \mathbb{R}$ pictured in Fig. ? is not continuous but is graph-closed.

It follows from (9.20) that each continuous operator $A : \mathcal{X} \to \mathcal{Y}$ is also graph-closed. The converse is **not** always true, by ❸. However, the following theorem tells us that the situation is nice in the linear case.

Theorem 256 (Banach's Closed Graph theorem)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces over $\mathbb{F}$. Then, each graph-closed linear operator $A : \mathcal{X} \to \mathcal{Y}$ is continuous.*

Proof.

- Let us define the following two linear continuous operators,

  $$P_1 : \operatorname{Graph}(A) \to \mathcal{X}$$
  $$P_2 : \operatorname{Graph}(A) \to \mathcal{Y}$$

  through

  $$P_1(u, Au) \triangleq u$$
  $$P_2(u, Au) \triangleq Au$$

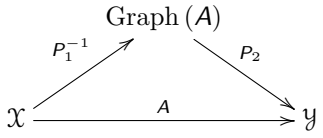
  for all $u \in \mathcal{X}$. Obviously,

  $$P_1(u, Au) = 0$$

  implies $u = 0$ and $Au = 0$. Thus, the operator P_1 is bijective. Since A is graph-closed, $\operatorname{Graph}(A)$ is a closed linear subspace of the Banach space $\mathcal{X} \times \mathcal{Y}$. Hence $\operatorname{Graph}(A)$ is also a Banach space. By the conditinuous inverse theorem (Proposition 253), the inverse operator

  $$P_1^{-1} : \mathcal{X} \to \operatorname{Graph}(A)$$

is continuous. Obviously, the diagram



is commutative (i.e. $A = P_2 P_1^{-1}$). Therefore, A is continuous. ■

Standard Example

Let $A : \mathcal{X} \to \mathcal{X}$ be linear self-adjoint operator on the Hilbert space $\mathcal{X}$ over $\mathbb{F}$. Then A is continuous.

PROOF.

- Let $u_n \to u$ and $Au_n \to v$ in $\mathcal{X}$ as $n \to \mathcal{X}$. It follows from

$$\langle Au_n | w \rangle = \langle u_n | Aw \rangle, \quad \forall w \in \mathcal{X}$$

that

$$\langle v | w \rangle = \langle u | Aw \rangle = \langle Au | w \rangle, \quad \forall w \in \mathcal{X},$$

that is, $Au = v$. Thus, A is graph-closed, and hence continuous, by Theorem 256.

9.4.2 Applications to Factor Spaces

Let $\mathcal{L}$ be a linear subspace of the linear space $\mathcal{X}$ over $\mathbb{F}$. For all $u, v \in \mathcal{K}$, we define

$$n \equiv v \ (\mod \mathcal{L}) \quad \text{iff } u - v \in \mathcal{L} \tag{9.21}$$

This is equivalent relation. In fact, for all $u, v, w, z \in \mathcal{X}$ and $\alpha \in \mathbb{F}$, we have the following:

$$u \equiv u \ (\mod \mathcal{L})$$
$$u \equiv v \ (\mod \mathcal{L})$$

$lspaceLongrightarrowv \equiv u \ (\mod \mathcal{L})$;

$$u \equiv v \ (\mod \mathcal{L}), \quad v \equiv w \ (\mod \mathcal{L}) \Longrightarrow u \equiv w \ (\mod \mathcal{L}). \tag{9.22}$$

This equivalence relation is compatible with the linear structure of $\mathcal{L}$:

$$\begin{aligned} &u \equiv v \ (\mod \mathcal{L}) \Longrightarrow \alpha u = \alpha v \ (\mod \mathcal{L}); \\ &u \equiv w \ (\mod \mathcal{L}), v \equiv z \ (\mod \mathcal{L}) \Longrightarrow u + v \equiv w + z \ (\mod \mathcal{L}). \end{aligned} \tag{9.23}$$

Definition 257. *The factor space $\mathcal{X}/\mathcal{L}$ consists of all the equivalence classes $[u]$ with respect to (9.21), that is,*

$$v \in [u] \quad \textit{iff} \quad u \equiv v \ (\mod \mathcal{L}).$$

Explicitly, this means that

$$[u] = u + \mathcal{L}.$$

The elements v of the class $[u]$ are called the *representatives* of $[u]$. Obviously,

$$[u] = [v] \Longleftrightarrow u \equiv v \ (\mod \mathcal{L}). \tag{9.24}$$

If we introduce the linear operations

$$\begin{aligned} \alpha[u] &\equiv [\alpha u], \\ [u] + [v] &\equiv [u + v], \end{aligned} \tag{9.25}$$

the factor space $\mathcal{X}/\mathcal{L}$ becomes a *linear space*. The operations in (9.25) are well defined, namely, they are *independent* of the chosen representatives. This follows from (**??**) and (9.24). For example, if $[u] = [v]$, then $u \equiv v \pmod{\mathcal{L}}$, and hence $\alpha u \equiv \alpha v \pmod{\mathcal{L}}$, that is, $[\alpha u] = [\alpha v]$.

In other words, the factor space $\mathcal{X}/\mathcal{L}$ consists of all the different sets

$$u + \mathcal{L}, \quad \text{where} \quad u \in \mathcal{X},$$

and the linear operations on $\mathcal{X}/\mathcal{L}$ are given through

$$(u + \mathcal{L}) + (v + \mathcal{L}) = (u + v) + \mathcal{L},$$
$$\alpha(u + \mathcal{L}) = \alpha u + \mathcal{L},$$

which corresponds to the usual operations $A + B$ and αA for subsets A and B of linear spaces.

Proposition 258. *Let $\mathcal{L}$ be a closed linear subspace of the normed space $\mathcal{X}$ over $\mathbb{F}$. Then the following are true:*

① *The factor space $\mathcal{X}/\mathcal{L}$ becomes a normed space over $\mathbb{F}$ w.r.t. the norm*

$$\|[u]\| = \inf_{v \in [u]} \|v\|. \tag{9.26}$$

Since $[u] = u + \mathcal{L}$, we get

$$\|[u]\| = \operatorname{dist}(0, u + \mathcal{L}) = \operatorname{dist}(u, \mathcal{L}).$$

② *If $\mathcal{X}$ is Banach space, then so is $\mathcal{X}/\mathcal{L}$.*

Proof.

- Ad ①.
  - We first show that
    $$\|[u]\| = 0 \Longleftrightarrow [u] = 0.$$
    This is identical to
    $$\|[u]\| = 0 \Longleftrightarrow u \in \mathcal{L}.$$
    In fact, if $u \in \mathcal{L}$, then $[u] = \mathcal{L}$. Thus $0 \in [u]$ and $\|[u]\| = 0$ by (9.26). Conversely, let $\|[u]\| = 0$. Since $\mathcal{L}$ is closed, so is the set $[u] = u + \mathcal{L}$. By (9.26), $0 \in u + \mathcal{L}$. Hence $u \in \mathcal{L}$.
  - Let $\alpha \in \mathbb{F}$. Since $\|\alpha v\| = |\alpha| \, \|v\|$, we have
    $$\|\alpha[u]\| = \inf_{w \in [u]} \|\alpha w\| = |\alpha| \inf_{w \in [u]} \|w\| = |\alpha| \, \|[u]\|.$$
  - Finally, it follows from $\|w_1 + w_2\| \le \|w_1\| + \|w_2\|$ that
    $$\begin{aligned} \|[u] + [v]\| &= \inf_{w_1 \in [u], w_2 \in [v]} \|w_1 + w_2\| \\ &\le \inf_{w_1 \in [u]} \|w_1\| + \inf_{w_2 \in [v]} \|w_2\| \\ &= \|[u]\| + \|[v]\|. \end{aligned}$$

- Ad ②.
  - It follows from (9.26) that each class $[u]$ contains a point v such that
    $$\|v\| \le 2 \, \|[u]\|. \tag{9.27}$$

– Now let $([u_n])$ be a Cauchy sequence in $\mathcal{X}/\mathcal{L}$. Using a simple induction argument based on (9.27), we obtain a sequence (v_n) in $\mathcal{X}$ such that $v_n \in [u_n]$ and

$$\|v_n - v_{n+1}\| \leq 2\,\|[u_n] - [u_{n+1}\|\,, \quad \forall n. \tag{9.28}$$

First suppose that

$$\|[u_n] - [u_{n+1}]\| \leq 2^{-n}, \quad \forall n. \tag{9.29}$$

It follows from (9.28) and the triangle inequality that

$$\|v_{n+m} - v_n\| \leq 2^{-n}(1 + 2^{-1} + 2^{-2} + \cdots),$$

that is, (v_n) is Cauchy in $\mathcal{X}$. Since X is a Banach space, we have

$$v_n \to v \quad \text{in} \quad \mathcal{X} \quad \text{as} \quad n \to \infty.$$

By (9.26),

$$\|[u_n] - [v]\| \leq \|v_n - v\| \to 0, \quad \text{as} \quad n \to \infty,$$

Hence $([u_n])$ is convergent in $\mathcal{X}/\mathcal{L}$.

– In the general case, there exists a subsequence, again denoted by $([u_n])$, such that (9.29) holds from Proposition 68.

Example. Let $\mathcal{X} = \mathbb{R}^{2\times 1}$ with the Euclidiean norm $\|\cdot\|$. In Fig.?, the factor space $\mathcal{X}/\mathcal{L}$ consists of all the straight lines $[u] = u + \mathcal{L}$ parallel to $\mathcal{L}$, and the norm $\|u\|$ is equal to the distance from the origin to the straight line $[u]$.

Definition 259. *Let $\mathcal{L}$ be a linear subspace of the linear space $\mathcal{X}$ over $\mathbb{F}$. Then, the canonical mapping*

$$\pi : \mathcal{X} \to \mathcal{X}/\mathcal{L}$$

is defined through

$$\pi(u) \triangleq [u], \quad \forall u \in \mathcal{X},$$

where $[u] = u + \mathcal{L}$.

Proposition 260. *If $\mathcal{L}$ is a closed linear subspace of the normed space $\mathcal{X}$ over K, then the canonical mapping $\pi : \mathcal{X} \to \mathcal{X}/\mathcal{L}$ is linear, continuous, and surjective.*

Proof.

- For all $u \in \mathcal{X}$, $\|\pi(u)\| = \|[u]\| \leq \|u\|$. ■

Let $A : \mathcal{X} \to \mathcal{Y}$ be a *linear continuous operator*, where $\mathcal{X}$ and $\mathcal{Y}$ are *Banach spaces* over $\mathbb{F}$. We define the operator

$$[A] : X/\operatorname{Ker}(A) \to \operatorname{Range}(A) \tag{9.30}$$

through

$$[A][u] \triangleq Au.$$

This definition is independent of the selected representative. Actually, let $[u] = [v]$, then $u - v \in \operatorname{Ker}(A)$, viz., $A(u - v) = 0$, and hence $Au = Av$.

Proposition 261. *Let the range* $\operatorname{Range}(A)$ *of the operator A be closed.*

① *The operator $[A]$ from (9.30) is a linear homeomorphism.*

② *There exists a number $c > 0$ such that*

$$c \cdot \operatorname{dist}(u, \operatorname{Ker}(A)) \leq \|Au\|\,, \quad \forall u \in \mathcal{X}. \tag{9.31}$$

PROOF.

- Ad ①.
  - The null space $\operatorname{Ker}(A) = \{u \in \mathcal{X} : Au = 0\}$ is closed. In fact, if
    $$Au_n = 0 \quad \text{and} \quad u_n \to u \quad \text{as} \quad n \to \infty,$$
    then $Au = 0$. Thus, $\mathcal{X}/\operatorname{Ker}(A)$ is a Banach space. Obviously, the operator $[A]$ is linear. Since
    $$\|[A][u]\| = \|Av\| \le \|A\| \, \|v\|, \quad \forall v \in [u],$$
    we have $\|[A][u]\| \le \|A\| \, \|u\|$, and thus $[A]$ is continuous.
  - Furthermore, the operator $[A]$ is bijective. In fact, if $[A][u] = 0$, then $u \in \operatorname{Ker}(A)$, and hence $[u] = 0$.
  - Since $\operatorname{Range}(A)$ is a closed linear subspace of the Banach space $\mathcal{Y}$, the range $\operatorname{Range}(A)$ is also a Banach space. The continuous inverse theorem tells us that the inverse operator $[A]^{-1} : \operatorname{Range}(A) \to \mathcal{X}/\operatorname{Ker}(A)$ is continuous.
- Ad ②.
  - By ①, there is a constant $d > 0$ such that
    $$\big\|[A]^{-1}[u]\big\| \le d \, \|[u]\|, \quad \forall [u] \in \mathcal{X}/\operatorname{Ker}(A).$$
    Consequently,
    $$\|[v]\| \le d \, \|[A][v]\|, \quad [v] \in \mathcal{X}/\operatorname{Ker}(A).$$
    This is (9.31) with $c = d^{-1}$.

9.4.3 Applications to Direct Sums and Projections

Projections

Definition 262. *Let $\mathcal{X}$ be a linear space over $\mathbb{F}$, and let $\mathcal{L}_1$ and $\mathcal{L}_2$ be linear subspaces of $\mathcal{X}$.*

- *We write*
  $$\mathcal{X} = \mathcal{L}_1 \oplus \mathcal{L}_2 \tag{9.32}$$
  iff each $u \in \mathcal{X}$ allows the following unique representation:
  $$u = u_1 + u_2, \quad u_1 \in \mathcal{L}_1, u_2 \in \mathcal{L}_2 \tag{9.33}$$
  We say that $\mathcal{X}$ is the direct sum of $\mathcal{L}_1$ and $\mathcal{L}_2$, and that $\mathcal{L}_2$ is an algebraic complement of $\mathcal{L}_1$ in $\mathcal{X}$.

The operator $P : \mathcal{X} \to \mathcal{X}$ is called an algebraic projection iff P is linear and $P^2 = P$.

If $\mathcal{X}$ is a normed space, then the operator $P : \mathcal{X} \to \mathcal{X}$ is called a continuous projection iff P is a continuous algebraic projection.

Obviously,
$$\mathcal{X} = \mathcal{L}_1 \oplus \mathcal{L}_2 \quad \text{iff} \quad \mathcal{X} = \mathcal{L}_2 \oplus \mathcal{L}_1.$$
Moreover, let $\mathcal{X} = \mathcal{L}_1 \oplus \mathcal{L}_2$. Then
$$u \in \mathcal{L}_1 \cap \mathcal{L}_2 \quad \text{implies} \quad u = 0.$$
This follows from $u = u + 0 = 0 + u$ and from the uniqueness of the decomposition in (9.33).

Using the Zorn lemma, we can deduce that

Each linear subspace $\mathcal{L}_1$ of the linear space $\mathcal{X}$ has an algebraic complement $\mathcal{L}_2$ in $\mathcal{X}$.

Proposition 263. *Let $\mathcal{X}$ be a linear space. Then the following statements hold true:*

① *Suppose that $\mathcal{X} = \mathcal{L}_1 \oplus \mathcal{L}_2$. If we set*

$$P u \triangleq u_1$$

in (9.33), then $P : \mathcal{X} \to \mathcal{X}$ is an algebraic projection onto the linear subspace $\mathcal{L}_1$. Moreover,

$$\begin{aligned} \mathcal{L}_1 &= P(\mathcal{X}) \\ \mathcal{L}_2 &= (\mathbb{1} - P)(\mathcal{X}) = \mathrm{Ker}(P). \end{aligned} \tag{9.34}$$

We call P the projection onto $\mathcal{L}_1$ along $\mathcal{L}_2$, and call $P_\perp \triangleq \mathbb{1} - P$ the orthogonal complement of P.

② *Conversely, if $P : \mathcal{X} \to \mathcal{X}$ is an algebraic projection, then $\mathcal{X} = \mathcal{L}_1 \oplus \mathcal{L}_2$ with (9.34).*

Proof.

- Ad ①.
  - Since the decomposition in (9.33) is unique, and since
  $$u_1 = u_1 + 0, \quad u_1 \in \mathcal{L}_1, 0 \in \mathcal{L}_2,$$
  we obtain $P u_1 = u_1$, and hence $P^2 u = P u_1 = u_1 = P u$. That is, $P^2 = P$.
  - By (9.33), $u_2 = u - u_1 = (\mathbb{1} - P)u$. Hence $\mathcal{L}_2 = (\mathbb{1} - P)(\mathcal{X})$. Finally, it follows from (9.33) that
  $$P u = 0 \Longleftrightarrow u \in \mathcal{L}_2,$$
  that is, $\mathrm{Ker}(P) = \mathcal{L}_2$.
- Ad ②.
  - Let $u \in \mathcal{X}$. Setting $u_1 \triangleq P u$ and $u_2 \triangleq (\mathbb{1} - P)u$, we obtain
  $$u = u_1 + u_2, \quad u_1 \in \mathcal{L}_1, u_2 \in \mathcal{L}_2,$$
  by (9.34). This decomposition is *unique*. In fact, let
  $$u = v_1 + v_2, \quad v_1 \in \mathcal{L}_1, v_2 \in \mathcal{L}_2.$$
  By (9.34), we get $v_1 = P v$ and $v_2 = (\mathbb{1} - P)w$ for some $v, w \in \mathcal{X}$. Since $P^2 = P$, this implies $P v_1 = v_1$ and $P v_2 = 0$. Hence
  $$u_1 = P u = P v_1 + P v_2 = P v_1 = v_1.$$
  This yields $u_1 = v_1$ and $u_2 = v_2$.

Definition 264 (Direct sum and topology)**.**

① *The direct sum $\mathcal{X} = \mathcal{L}_1 \oplus \mathcal{L}_2$ is called a topological direct sum iff the corresponding projection $P : \mathcal{X} \to \mathcal{X}$ onto $\mathcal{L}_1$ along $\mathcal{L}_2$ is continuous. Then we say that $\mathcal{L}_2$ is a topological compelment of $\mathcal{L}_1$ in $\mathcal{X}$.*

② *The linear subspace $\mathcal{L}_1$ splits the normed space $\mathcal{X}$ iff $\mathcal{L}_1$ has a topological complement in $\mathcal{X}$.*

Example. Let $\mathcal{X} = \mathbb{R}^2$. Then

$$\mathbb{R}^2 = \mathcal{L}_1 \oplus \mathcal{L}_2, \tag{9.35}$$

where $\mathcal{L}_1$ and $\mathcal{L}_2$ denote the two straight lines pictured in Fig. ?. The projection $P : \mathcal{X} \to \mathcal{X}$ onto $\mathcal{L}_1$ parallel to $\mathcal{L}_2$. Since P is continuous, (9.35) represents a topological direct sum. Moreover, $\mathcal{L}_1$ and $\mathcal{L}_2$ splits $\mathbb{R}^2$.

Proposition 265. *Let $\mathcal{L}$ be a linear subspace of the normed space $\mathcal{X}$ over $\mathbb{F}$. Then*

❶ *$\mathcal{L}$ splits $\mathcal{X}$ iff there exists a continuous projection $P : \mathcal{X} \to \mathcal{X}$ onto $\mathcal{L}$.*

❷ *If $\mathcal{L}$ splits $\mathcal{X}$, then $\mathcal{L}$ is closed.*[3]

Proof.

- Ad ❶: This follows from Proposotion 273.
- Ad ❷: By ❶, we have $\mathcal{L} = P\mathcal{X}$, where the projection $P : \mathcal{X} \to \mathcal{L}$ is continuous. Let (u_n) be a sequence in $\mathcal{L}$ such that $u_n \to u$ as $n \to \infty$. Letting $n \to \infty$, it follows from

  $$u_n = Pu_n, \quad \forall n$$

  that $u = Pu$, and therefore $u \in \mathcal{L}$. Thus, $\mathcal{L}$ is closed.

Proposition 266. *Let $\mathcal{X} = \mathcal{L}_1 \oplus \mathcal{L}_2$ be a direct sum, where $\mathcal{L}_1$ and $\mathcal{L}_2$ are linear subspaces of the Banach space $\mathcal{X}$. Then the following two conditions are equivalent:*

① *$\mathcal{X} = \mathcal{L}_1 \oplus \mathcal{L}_2$ represents a topological direct sum.*

② *Both $\mathcal{L}_1$ and $\mathcal{L}_2$ are closed.*

Proof.

- ①$\Longrightarrow$②.
  - Both $\mathcal{L}_1$ and $\mathcal{L}_2$ split $\mathcal{X}$, and so $\mathcal{L}_1$ and $\mathcal{L}_2$ are closed.
- ②$\Longrightarrow$①.
  - Let $P : \mathcal{X} \to \mathcal{X}$ be the algebraic projection onto $\mathcal{L}_1$ along $\mathcal{L}_2$. We have to show that P is continuous. To this end, let (u_n) be a sequence in $\mathcal{X}$. Then

    $$u_n = u_{1n} + u_{2n}, \quad u_{1n} \in \mathcal{L}_1, u_{2n} \in \mathcal{L}_2. \tag{9.36}$$

    Hence $u_{1n} = Pu_n$. Suppose that

    $$u_n \to u, \quad \text{and} \quad Pu_n = v, \quad \text{as} \quad n \to \infty.$$

    Letting $n \to \infty$ in (9.36), we get

    $$u = v + w,$$

    where $u_{2n} \to w$ as $n \to \infty$. Since $\mathcal{L}_1$ and $\mathcal{L}_2$ are closed, we have $v \in \mathcal{L}_1$. Thus, $v = Pu$. Consequently, the operator P is graph-closed. The closed graph theorem tells us that P is *continuous*.

[3] Unfortunately, the converse of ❷ is not true.

Codimension

Definition 267. *Let $\mathcal{L}$ be a linear subspace of the linear space $\mathcal{X}$ over $\mathbb{F}$. Then, the codimension of $\mathcal{L}$ in $\mathcal{X}$ is defined as the dimension of the factor space $\mathcal{X}/\mathcal{L}$, denoted as*

$$\operatorname{codim}(\mathcal{L}) \triangleq \dim(\mathcal{X}/\mathcal{L}).$$

Obviously, if $\mathcal{L} = \mathcal{X}$, then $\mathcal{X}/\mathcal{L} = \{0\}$, and hence $\operatorname{codim}(\mathcal{X}) = 0$. The following proposition explains the intuitive meaning of $\operatorname{codim}(\mathcal{L})$.

Proposition 268. *Let $\mathcal{L}$ be a linear subspace of the linear space $\mathcal{X}$ over $\mathbb{F}$. Then the following statements hold true:*

① *There exists a linear subspace $\mathcal{M}$ of $\mathcal{X}$ such that*

$$\mathcal{X} = \mathcal{L} \oplus \mathcal{M}. \tag{9.37}$$

② *If $\mathcal{M}$ is any linear subspace of $\mathcal{X}$ such that (9.37) holds, then*

$$\operatorname{codim}(\mathcal{L}) = \dim(\mathcal{M}).$$

③ *From (9.37) we get*

$$\dim(\mathcal{X}) = \dim(\mathcal{L}) + \dim(\mathcal{M}),$$

and hence

$$\dim(\mathcal{X}) = \dim(\mathcal{L}) + \operatorname{codim}(\mathcal{L}).$$

It follows from ③ that if $\mathcal{X} = \mathcal{L} \oplus \mathcal{M}$ and $\dim(\mathcal{X}) < \infty$, then

$$\operatorname{codim}(\mathcal{L}) = \dim(\mathcal{X}) - \dim(\mathcal{L}). \tag{9.38}$$

Proof.

- Ad ①.
  - Let $\mathcal{C}$ be the class of all the linear operators

    $$P : \operatorname{Dom}(P) \subset \mathcal{X} \to \mathcal{L}$$

    such that $\mathcal{L} \subset \operatorname{Dom}(P)$ and $Pu = u$ on $\mathcal{L}$. We write

    $$P_1 \leq P_2 \quad \text{iff} \quad P_2 \text{ is an extension of } P_1.$$

    By the Zorn lemma, $\mathcal{C}$ contains a maximal element P_0. Then $\operatorname{Dom}(\mathcal{P}_0) = \mathcal{X}$. Otherwise, there would exist a point $u_0 \in \mathcal{X} - \operatorname{Dom}(P_0)$. Set $N \triangleq \operatorname{Dom}(P_0) + \operatorname{span}\{u_0\}$, and define the operator $P : N \to \mathcal{L}$ through

    $$P(u + \alpha u_0) \triangleq P_0(u), \quad u \in \operatorname{Dom}(P_0), \alpha \in \mathbb{F}.$$

    Then P is a proper extension of P_0, contradicting the maximality of P_0.
  - In addition, we get

    $$P_0^2 = P_0.$$

    Actually, for each $v \in \mathcal{X}$, it follows from $P_0 v \in \mathcal{L}$ that $P_0(P_0 v) = P_0 v$, by the construction of $\mathcal{C}$. Therefore, the operator $P_0 : \mathcal{X} \to \mathcal{X}$ is an algebraic projection onto $\mathcal{L}$.

– Letting $\mathcal{M} \triangleq (\mathbb{1} - P_0)(\mathcal{X})$, we obtain

$$\mathcal{X} = P_0(\mathcal{X}) \oplus (\mathbb{1} - P_0)(\mathcal{X}) = \mathcal{X} \oplus \mathcal{M}.$$

- Ad ②.

  – For each $u \in \mathcal{X}$, we have

  $$u = v + w, \quad v \in \mathcal{L}, w \in \mathcal{M}.$$

  – Define the map $\phi : \mathcal{M} \to \mathcal{X}/\mathcal{L}$ by

  $$\phi(w) \triangleq [w], \quad \forall w \in \mathcal{M}.$$

  Then ϕ is linear and surjective. Moreover, $\phi(w) = 0$ with $w \in \mathcal{M}$ implies $w \in \mathcal{L}$. It follows from $w \in \mathcal{L} \cap \mathcal{M}$ and $\mathcal{X} = \mathcal{L} \oplus \mathcal{M}$ that $w = 0$. Thus, ϕ is a bijection. This yields

  $$\dim(\mathcal{M}) = \dim(\mathcal{X}/\mathcal{L}),$$

  and hence $\dim(\mathcal{M}) = \operatorname{codim}(\mathcal{L})$.

- Ad ③.

  – By ②, it is sufficient to prove that $\mathcal{X} = \mathcal{L} \oplus \mathcal{M}$ implies

  $$\dim(\mathcal{X}) = \dim(\mathcal{L}) + \dim(\mathcal{M}). \tag{9.39}$$

  – First let $\dim(\mathcal{L}) = \infty$. Then $\mathcal{L} \subset \mathcal{X}$ implies $\dim(\mathcal{X}) = \infty$. Analogously, $\dim(\mathcal{M}) = \infty$ yields $\dim(\mathcal{X}) = \infty$.

  – Next suppose that $\dim(\mathcal{L}) < \infty$ and $\dim(\mathcal{M}) < \infty$. Then (9.39) follows from the fact that the union of a basis in $\mathcal{L}$ and a basis in $\mathcal{M}$ represents a basis in $\mathcal{X}$.

Example. For an m-dimensional linear subspace $\mathcal{L}$ of $\mathbb{R}^N$, $N \geq 1$, we get

$$\operatorname{codim}(\mathcal{L}) = N - m. \tag{9.40}$$

For example, if $\mathcal{L}$ is a plane through the origin in $\mathbb{R}^3$, then

$$\dim(\mathcal{L}) = 2, \quad \operatorname{codim}(\mathcal{L}) = 1.$$

Corollary 269. *Let $\mathcal{L}$ be a linear subspace of the linear space $\mathcal{X}$ over $\mathbb{F}$. Suppose that $u_1, \cdots, u_m$ are linearly independent elements of $\mathcal{X}$ such that*

$$\mathcal{L} \cap \operatorname{span}\{u_1, \cdots, u_m\} = \{0\}. \tag{9.41}$$

Then $m \leq \operatorname{codim}(\mathcal{L})$.

Proof.

- It follows from (9.41) that $[u_1], \cdots, [u_m]$ are linearly independent elements of $\mathcal{X}/\mathcal{L}$. Hence $m \leq \dim(\mathcal{X}/\mathcal{L})$. ■

Lemma 270. *Let $A : \mathcal{X} \to \mathcal{Y}$ be a continuous operator, where $\mathcal{X}$ and $\mathcal{Y}$ are normed spaces over $\mathbb{F}$. Let $\mathcal{W} \subset \mathcal{Y}$. The following conditions hold:*

① *If $\mathcal{W}$ is open, then so is $A^{-1}(\mathcal{W})$.*

② *If $\mathcal{W}$ is closed, then so is $A^{-1}(\mathcal{W})$.*

PROOF. This is a special case of a more general result about continuous maps on topological spaces. A direct proof resembles the following.

- Ad ①: Let $\mathcal{W}$ be open, and let $u_0 \in \mathsf{A}^{-1}(\mathcal{W})$. For each $\varepsilon > 0$, there is a $\delta(\varepsilon) > 0$ such that

$$\|u - u_0\| < \delta(\varepsilon) \quad \text{implies} \quad \|\mathsf{A}u - \mathsf{A}u_0\| < \varepsilon,$$

  by the continuity of A. If we choose the number ε sufficiently small, then

$$\|u - u_0\| < \delta(\varepsilon) \quad \text{implies} \quad \mathsf{A}u \in \mathcal{W},$$

  and hence u_0 is an interior point of $\mathsf{A}^{-1}(\mathcal{W})$. Thus, $\mathcal{W}$ is open.

- Ad ②: Use ① and the fact that the complements of closed sets are open. ■

Corollary 271. *Let $\mathcal{L}$ be a closed linear subspace of the Banach space $\mathcal{X}$ over $\mathbb{F}$ with* $\operatorname{codim}(\mathcal{L}) < \infty$*, and let $\mathcal{S}$ be a linear subspace of $\mathcal{X}$ such that*

$$\mathcal{L} \subset \mathcal{S} \subset \mathcal{X}.$$

Then, $\mathcal{S}$ is closed and $\operatorname{codim}(\mathcal{S}) < \infty$.

PROOF.

- Let us consider the canonical mapping

$$\pi : \mathcal{X} \to \mathcal{X}/\mathcal{L}.$$

  Recall that $\pi(u) \triangleq [u] = u + \mathcal{L}$ for all $u \in \mathcal{X}$. The restriction of π to $\mathcal{S}$ is given by

$$\pi : \mathcal{S} \to \mathcal{S}/\mathcal{L}$$

  is therefore also closed, by Lemma 270.

- Since $\mathcal{L} \subset \mathcal{S}$, it follows from

$$\alpha_1 u_1 + \cdots + \alpha_m u_m \equiv 0 \ (\mod \mathcal{L})$$

  that

$$\alpha_1 u_1 + \cdots + \alpha_m u_m \equiv 0 \ (\mod \mathcal{S})$$

  where $\alpha_1, \cdots, \alpha_m \in \mathbb{F}$. Therefore, if $u_1, \cdots, u_m$ are liknearly independent $\mod \mathcal{S}$, then they are also linearly independent $\mod \mathcal{L}$. Hence

$$\dim(\mathcal{X}/\mathcal{L}) \leq \dim(\mathcal{X}/\mathcal{L}).$$

  This yields $\operatorname{codim}(\mathcal{S}) \leq \operatorname{codim}(\mathcal{L})$.

9.4.4 Linear Operator Equations

Let us consider the linear operator equation

$$\mathsf{A}u = b, \quad u \in \mathcal{X}. \tag{9.42}$$

Proposition 272. *Suppose that the operator $\mathsf{A} : \mathcal{X} \to \mathcal{X}$ is linear, where $\mathcal{X}$ and $\mathcal{Y}$ are linear spaces over $\mathbb{F}$. Let $\mathcal{L}$ be any fixed algebraic complement of the null space* $\operatorname{Ker}(\mathsf{A})$*, namely, $\mathcal{L}$ is a linear subspace of $\mathcal{X}$ such that*

$$\mathcal{X} = \operatorname{Ker}(\mathsf{A}) \oplus \mathcal{L}. \tag{9.43}$$

Then the following statements are true:

① *The restriction*

$$A : \mathcal{L} \to \text{Range}(A) \tag{9.44}$$

is linear and bijective. Hence

$$\text{codim}(\text{Ker}(A)) = \dim(\text{Range}(A)). \tag{9.45}$$

② *In addition, suppose that* $\mathcal{X}$ *and* $\mathcal{Y}$ *are Banach spaces,* $\mathcal{L}$ *and* Range(A) *are closed, and the operator* $A : \mathcal{X} \to \mathcal{Y}$ *is continuous. Then the operator from (9.44) is a linear homeomorphism.*

PROOF.

- Ad ①: It follows from $Au = 0$ with $u \in \mathcal{L}$ that $u \in \text{Ker}(A) \cap \mathcal{L}$. Hence $u = 0$, by (9.43).
- Ad ②: This follows from the continuous inverse theorem.

Recall that $\text{Range}(A) = A(\mathcal{X})$. The number of $\dim(\text{Range}(A))$ is called the *rank* of A. We denote this as

$$\text{Rank}(A) \triangleq \dim(\text{Range}(A)).$$

Suppose that $\dim(\mathcal{X}) < \infty$ and $\dim(\mathcal{Y}) < \infty$. Let

$$B : \mathcal{L} \to \text{Range}(A)$$

denote the restriction of the operator $A : \mathcal{X} \to \text{Range}(A)$ to the linear subspace $\mathcal{L}$ of $\mathcal{X}$. Then, for each given $b \in \mathcal{Y}$, the solution set of the original equation (9.42) is given through

$$B^{-1}b + \text{Ker}(A),$$

where $\dim(\text{Ker}(A)) = \dim(\mathcal{X}) - \text{Rank}(A)$, by (9.45)

Proposition 273. *Let* $f_1, \cdots, f_n$, $f : \mathcal{X} \to \mathbb{F}$ *be linear functionals on the linear space* $\mathcal{X}$ *over* $\mathbb{F}$. *Suppose that each solution* $u \in \mathcal{X}$ *of the system*

$$f_j(u) = 0, \quad j = 1, \cdots, n,$$

is also a solution of the equation

$$f(u) = 0.$$

Then there exists numbers $\alpha_1, \cdots, \alpha_n \in \mathbb{F}$ *such that*

$$f = \alpha_1 f_1 + \cdots + \alpha_n f_n.$$

PROOF. We may assume that $f_1, \cdots, f_n$ are linearly independent. The proof proceeds by induction.

- Step-1: We prove the statement for $n = 1$. Since $f_1 \neq 0$, there exists a point $u_1 \in \mathcal{X}$ such that $f_1(u_1) \neq 0$. Replacing u_1 with βu_1, if necessary, we get

  $$f_1(u_1) = 1.$$

  Set

  $$v \triangleq u - f_1(u)u_1.$$

  Then $f_1(v) = 0$, and hence $f(v) = 0$, by hypothesis. This implies

  $$0 = f(u) - f_1(u)f(u_1), \quad \forall u \in \mathcal{X},$$

  that is, $f = \alpha f_1$ for some $\alpha \in \mathbb{F}$.

- Step-2: We prove the statement for $n = 2$. Since f_2 is linearly independent of f_1, there exists a point $u_2 \in \mathcal{X}$ such that
$$f_1(u_2) = 0 \quad \text{and} \quad f_2(u_2) \neq 0,$$
for $n = 1$ and $f = f_2$. Analogously, there exists a point $u_1 \in \mathcal{X}$ such that
$$f_2(u_1) = 0 \quad \text{and} \quad f_1(u_1) \neq 0.$$
We may assume that
$$f_1(u_1) = f_2(u_2) = 1.$$
Set
$$v \triangleq u - f_1(u)u_1 - f_2(u)u_2.$$
Then, $f_1(v) = f_2(v) = 0$, and hence $f(v) = 0$ by hypothesis. This implies
$$0 = f(u) - f_1(u)f(u_1) - f_2(u)f(u_2), \quad \forall u \in \mathcal{X},$$
that is, $f = \alpha_1 f_1 + \alpha_2 f_2$, where $\alpha_j \triangleq f(u_j)$.
- Step-3: If the assertion is true for n, then a similar argument as in Step-2 shows that the statement is also true for $n + 1$.

9.4.5 Biorthogonal Systems and Splitting Subspaces

Definition 274. *Let $\mathcal{X}$ be a normed space over $\mathbb{F}$. By an $\mathcal{X}$-biorthogonal system $\left\{u_j, u_j^*\right\}_{j=1}^n$, we understand a system of points $u_1, \cdots, u_n \in \mathcal{X}$ and functionals $u_1^*, \cdots, u_n^* \in \mathcal{X}^*$ such that*
$$\langle u_i^*, u_j \rangle = \delta_{ij}, \quad i, j = 1, \cdots, n.$$

Proposition 275. *Let $\mathcal{X}$ be a normed space over $\mathbb{F}$.*

① *Each system $u_1, \cdots, u_n \in \mathcal{X}$ of linearly independent points can be extended to an $\mathcal{X}$-biorthogonal system.*

② *Each system $u_1^*, \cdots, u_n^* \in \mathcal{X}^*$ of linearly independent functionals can be extended to an $\mathcal{X}$-biorthogonal system.*

Proof.

- Ad ①: Let $\mathcal{X} = \operatorname{span} u_1, \cdots, u_n$. Define the linear functional $u_j^* : \mathcal{L} \to \mathbb{F}$ by
$$\left\langle u_i^*, \sum_{j=1}^n \alpha_j u_j \right\rangle \triangleq \alpha_i, \quad i = 1, \cdots, n.$$
By the Hahn-Banach theorem, u_i^* can be extended to a linear continuous functional $u_i^* : \mathcal{X} \to \mathbb{F}$.
- Ad ②: Set $f_j \triangleq u_j^*$. Then the exitence of points $u_1, \cdots, u_n$ with $\langle u_j^*, u_i \rangle = \delta_{ji}$ follows as in the proof of Proposition 273. ■

Example. Let $\mathcal{L}$ be a linear subspace of the Banach space $\mathcal{X}$ over $\mathbb{F}$. Then $\mathcal{L}$ splits $\mathcal{X}$ if one of the following three conditions is met:

- $\mathcal{L}$ is a closed linear subspace of the Hilbert space $\mathcal{X}$.
- $\dim(\mathcal{X}) < \infty$.
- $\mathcal{L}$ is closed and $\operatorname{codim}(\mathcal{L}) < \infty$.

9.4.6 Pseudo-Orthogonal Complements

Definition 276 (Pseudo-Orthogonal Complements)**.** *Let $\mathcal{L}$ be a linear subspace of the normed space $\mathcal{X}$ over $\mathbb{F}$ and let $\mathcal{M}$ be a linear subspace of $\mathcal{X}^*$.*

- *The set*
$$\mathcal{L}^{\perp} \triangleq \{u^* \in \mathcal{X}^* : \langle u^*, u\rangle = 0, \forall u \in \mathcal{L}\} \tag{9.46}$$
is called the pseudo-orthogonal complement to $\mathcal{L}$.

- *Then we set*
$${}^{\perp}\mathcal{M} \triangleq \{u \in \mathcal{X} : \langle u^*, u\rangle = 0, \forall u^* \in \mathcal{M}\}.$$
is called the pseudo-orthogonal complement to $\mathcal{M}$.

These notions generalize orthogonal compelments $\mathcal{L}^{\perp}$ in Hilbert spaces.

Proposition 277. *$\mathcal{L}^{\perp}$ and ${}^{\perp}\mathcal{M}$ are closed linear subspaces of $\mathcal{X}$ and $\mathcal{X}^*$, respectively.*

Proof.

- Suppose that $u_n^* \in \mathcal{L}^{\perp}$ for all n and
$$u_n^* \to u^* \quad \text{in} \quad \mathcal{X}^* \quad \text{as} \quad n \to \infty.$$
If we let $n \to \infty$, it follows from $\langle u_n^*, v\rangle = 0$ for all n and $v \in \mathcal{L}$ that $\langle u^*, v\rangle = 0$ for all $v \in \mathcal{L}$, and hence $u^* \in \mathcal{L}^{\perp}$.

- Suppose that $u_n \in^{\perp} \mathcal{M}$ for all n and
$$u_n \to u \quad \text{in} \quad \mathcal{X} \quad \text{as} \quad n \to \infty.$$
If we let $n \to \infty$, it follows from $\langle u^*, u_n\rangle = 0$ for all n and all $u^* \in \mathcal{M}$ that $\langle u^*, u\rangle = 0$ for all $u^* \in \mathcal{M}$, and hence $u \in^{\perp} \mathcal{M}$. ■

Proposition 278. *Let $\mathcal{L}$ be a linear subspace of the normed space $\mathcal{X}$ over $\mathbb{F}$, then*
$$\overline{\mathcal{L}} =^{\perp} (\mathcal{L}^{\perp}).$$

Proof.

- By (9.46), $(\overline{\mathcal{L}})^{\perp} = \mathcal{L}^{\perp}$. Therefore, it is sufficient to prove that
$$M =^{\perp} (\mathcal{M}^{\perp}),$$
where $\mathcal{M}$ is a closed linear subspace of $\mathcal{X}$. By the definition of pseudo-orthogonal complement (Definition 276), we have
$$u \in^{\perp} (\mathcal{M}^{\perp}) \quad \text{iff} \quad \langle u^*, u\rangle = 0 \quad \forall u^* \mathcal{M}^{\perp}.$$
Thus $\mathcal{M} \subset^{\perp} (\mathcal{M}^{\perp})$.

- Conversely, we want to show that ${}^{\perp}(\mathcal{M}^{\perp}) \subset \mathcal{M}$. Let $v \in^{\perp} (\mathcal{M}^{\perp})$ and suppose that $v \notin \mathcal{M}$. By Proposition 229, it follows from the Hahn-Banach theorem that there exists a functional $u^* \in \mathcal{X}^*$ such that
$$u^* = 0 \quad \text{on} \quad \mathcal{M} \qquad \langle u^*, u\rangle \neq 0$$
Hence $u^* \in \mathcal{M}^{\perp}$ and $v \notin^{\perp} (\mathcal{M}^{\perp})$. This is a contradiction. ■

Proposition 279. *Let $\mathcal{X}$ be a normed space over $\mathbb{F}$. Then*

① *If $\mathcal{L}$ is a finite-dimensional linear subspace of $\mathcal{X}$, then*

$$\operatorname{codim}(\mathcal{L}^{\perp}) = \dim(\mathcal{L}) \quad \textit{in} \quad \mathcal{X}^*.$$

② *If $\mathcal{M}$ is a finite-dimensional linear subspace of $\mathcal{X}^*$, then*

$$\operatorname{codim}({}^{\perp}\mathcal{M}) = \dim(\mathcal{M}) \quad \textit{in} \quad \mathcal{X}.$$

③ *If $\mathcal{L}$ is a closed linear subspace of $\mathcal{X}$ such that $\mathcal{L}^{\perp}$ is finite-dimensional, then*

$$\operatorname{codim}(\mathcal{L}) = \dim(\mathcal{L}^{\perp}) \quad \textit{in} \quad \mathcal{X}.$$

PROOF.

- Ad ①.
  - If $\mathcal{L} = \{0\}$, then $\mathcal{L}^{\perp} = \mathcal{X}^*$, and hence $\operatorname{codim}(\mathcal{L}^{\perp})) = 0$.
  - Suppose now that $\dim(\mathcal{L}) = n$, where $n > 0$. Let $\{u_1, \cdots, u_n\}$ be a basis of $\mathcal{L}$. Extend this to an $\mathcal{X}$-biorthogonal system $\left\{u_j, u_j^*\right\}$. Define the continuous projection operator $P : \mathcal{X} \to \mathcal{X}$ through
  $$Pu^* \triangleq u^* - \sum_{j=1}^{n} \langle u^*, u_j \rangle \, u_j^*, \quad \forall u^* \in \mathcal{X}^*.$$
  Obviously, $Pu^* = u^*$ iff $\langle u^*, u_j \rangle = 0$ for all j (i.e., $u^* \in \mathcal{L}^{\perp}$). Thus, $P(\mathcal{X}^*) = \mathcal{L}^{\perp}$. Therefore $\operatorname{codim}(\mathcal{L}^{\perp}) = \dim(\mathbb{1} - P)(\mathcal{X}^*) = n$, by Proposition 268 along with $\mathcal{X}^* = P(\mathcal{X}^*) \oplus (\mathbb{1} - P)(\mathcal{X}^*)$.
- Ad ②: Use a similar argument as in the proof of ①.
- Ad ③: By Proposition 278, $\mathcal{L} =^{\perp} (\mathcal{L}^{\perp})$. It follows from ② that $\operatorname{codim}(\mathcal{L}) = \dim(\mathcal{L}^{\perp})$.

9.5 Dual Operators

9.5.1 A, A^* and $A^{\dagger}$

The theory of linear operator equations in Banach spaces is essentially based on the concept of *duality*. To this end, we need dual operators.

The *key* relation for dual operators is given through

$$\langle A^* u^*, u \rangle = \langle u^*, Au \rangle, \quad \forall u \in \mathcal{X}, u^* \in \mathcal{X}^*. \tag{9.47}$$

Proposition 280. *Let*

$$A : \mathcal{X} \to \mathcal{Y}$$

be a linear continuous operator, where $\mathcal{X}$ and $\mathcal{Y}$ are normed spaces over $\mathbb{F}$. Then there exists precisely one linear operator

$$A^* : \mathcal{Y}^* \to \mathcal{X}^*$$

such that relation (9.47) holds. In addition, A^* is continuous.

The operator A^* is called the *transposed* or *dual operator* to A. We can show that in finite-dimensional Hilbert spaces, the transposed operator A^* and the adjoint operator $A^{\dagger}$ correspond to the transposed matrix and the adjoint matrix, respectively. Particularly, if $\mathcal{X} = \mathbb{C}^{n \times 1}$, then $A^* = \boldsymbol{A}^{\mathsf{H}}$ and $A^{\dagger} = \boldsymbol{A}^{\dagger}$.

PROOF.

- EXISTENCE. Let $u^* \in \mathcal{Y}^*$ be given. Set

$$f(u) \triangleq \langle u^*, Au \rangle, \quad \forall u \in \mathcal{X}.$$

Then

$$|f(u)| \le \|u^*\| \, \|Au\| \le \|u^*\| \, \|A\| \, \|u\|, \quad \forall u \in \mathcal{X}. \tag{9.48}$$

So $f : \mathcal{X} \to \mathbb{F}$ is a linear continuous functional, namely, $f \in \mathcal{X}^*$. Define

$$A^* u^* \triangleq f.$$

Obviously, $\langle A^* u^*, u \rangle = \langle u^*, Au \rangle$ for all $u \in \mathcal{X}$. Thus we obtain the linear operator $A^* : \mathcal{Y}^* \to \mathcal{X}^*$. By (9.48),

$$\|A^* u^*\| = \|f\| \le \|A\| \, \|u^*\|, \quad \forall u^* \in \mathcal{Y}^*,$$

and hence A^* is continuous.

- UNIQUENESS. Let $u^* \in \mathcal{Y}^*$ and $v^* \in \mathcal{X}^*$ be given. Suppose that

$$\langle v^*, u^* \rangle = \langle u^*, Au \rangle, \quad \forall u \in \mathcal{X}.$$

It follows from (9.47) that $\langle v^* - A^* u^*, u \rangle = 0$ for all $u \in \mathcal{X}$, and hence $v^* = A^* u^*$. ■

Proposition 281. *Let $A : \mathcal{X} \to \mathcal{X}$ be a linear continuous operator on the Hilbert space $\mathcal{X}$ over $\mathbb{F}$. Then, the following diagram is commutative:*

$$\begin{array}{ccc} \mathcal{X} & \xrightarrow{A^\dagger} & \mathcal{X} \\ {\scriptstyle J}\downarrow & & \downarrow{\scriptstyle J} \\ \mathcal{X}^* & \xrightarrow[A^*]{} & \mathcal{X}^* \end{array}$$

Here, J denotes the duality map of $\mathcal{X}$. Explicitly,

$$A^\dagger = J^{-1} A^* J.$$

PROOF.

- Let $u, v \in \mathcal{X}$, by the definition map of J we have

$$\langle Ju, v \rangle = \langle u | v \rangle.$$

Hence

$$\begin{aligned} \langle J^{-1} A^* J u | v \rangle &= \langle J \underline{J^{-1}} A^* J u, v \rangle = \langle \underline{J J^{-1}} A^* J u, v \rangle \\ &= \langle A^* J u, v \rangle = \langle J u, A v \rangle \\ &= \langle u | A v \rangle = \langle A^\dagger u | v \rangle. \end{aligned}$$ ■

9.5.2 Matrix-Vector Equation

Let $\mathcal{X} \neq \{0\}$ be a finite-dimensional Hilbert space over $\mathbb{F}$ with the orthogonal basis $\{e_1, \cdots, e_n\}$ (e.g, $\mathcal{X} = \mathbb{K}^{n \times 1}, n \ge 1$). Then, for each $u, b \in \mathcal{X}$, we have the representations

$$u = \sum_{j=1}^{n} \xi_j e_j, \quad b = \sum_{j=1}^{n} \beta_j e_j,$$

where $\xi_j, \beta_j \in \mathbb{F}$ for all j. Let

$$A : \mathcal{X} \to \mathcal{Y}$$

be a linear operator. We are given $b \in \mathcal{X}$ and $b^* \in \mathcal{X}^*$. Then the following relations between operator equations and matrix equations hold true:

- The original equation

$$\mathsf{A}u = b, \quad u \in \mathcal{X} \tag{9.49}$$

corresponds to the matrix equation

$$\boldsymbol{A}\boldsymbol{\xi} = \boldsymbol{b}, \quad \boldsymbol{\xi} \in \mathbb{K}^{n\times 1} \tag{9.50}$$

where we set

$$\boldsymbol{A} \triangleq \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}, \quad \boldsymbol{\xi} \triangleq \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix}, \quad \boldsymbol{\beta} \triangleq \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix},$$

and the matrix elements a_{jk} of the operator A are given by

$$\mathsf{A}e_k = \sum_{j=1}^{n} a_{jk} e_j, \quad k = 1, \cdots, n. \tag{9.51}$$

- The adjoint equation

$$\mathsf{A}^\dagger u = b, \quad u \in \mathcal{X} \tag{9.52}$$

and the dual equation

$$\mathsf{A}^* u^* = b^*, \quad u^* \in \mathcal{X}^* \tag{9.53}$$

corresponds to the matrix equations

$$\boldsymbol{A}^\dagger \boldsymbol{\xi} = \boldsymbol{\beta} \tag{9.54}$$

and

$$\boldsymbol{A}^* \boldsymbol{\xi}^* = \boldsymbol{\beta}^*, \tag{9.55}$$

respectively. Here $\boldsymbol{A}^\dagger$ and $\boldsymbol{A}^*$ denote the adjoint matrix and the transposed matrix to $\boldsymbol{A}$ respectively. Explicitly,

$$\boldsymbol{A}^\dagger = (a_{ij}^\dagger) = (\bar{a}_{ji}) \equiv \boldsymbol{A}^{\mathsf{H}}$$
$$\boldsymbol{A}^* = (a_{ij}^*) = (a_{ji}) \equiv \boldsymbol{A}^{\mathsf{T}}$$

for $i, j \in \{1, \cdots, n\}$. The bar denotes the conjugate complex number. In addition,

$$u^* = \sum_{j=1}^{n} \xi_j^* e_n^*, \quad b^* = \sum_{j=1}^{n} \beta_j^* e_j^*,$$

where $\xi_j^* = \xi_j, \beta_j^* = \beta_j \in \mathbb{F}$ for all $j \in \{1, \cdots, n\}$, and $\{e_1^*, \cdots, e_n^*\}$ is the basis of the dual space $\mathcal{X}^*$ such that $\langle e_i^*, e_j \rangle = \delta_{ij}$, where δ_{ij} is the Kronecker symbols, i.e., its value is 1 for $i = j$ and 0 for $i \neq j$.

- If $\mathcal{X}$ is a real space, i.e., $\mathbb{F} = \mathbb{R}$, then $\boldsymbol{A}^\dagger = \boldsymbol{A}^*$. With more famimilar form, we have $\boldsymbol{A}^{\mathsf{H}} = \boldsymbol{A}^{\mathsf{T}}$.

9.6 Dual Functor

9.6.1 Fundamentals

Definition 282. *Let*

$$\mathsf{A} : \mathcal{X} \to \mathcal{Y} \tag{9.56}$$

be a linear continuous operator, where $\mathcal{X}$ and $\mathcal{Y}$ are normed spaces over $\mathbb{F}$. The duality functor $\mathcal{D}$ assigns to (9.56) the dual operator

$$\mathsf{A}^* : \mathcal{Y}^* \to \mathcal{X}^*. \tag{9.57}$$

Proposition 283. *Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ be normed spaces over $\mathbb{F}$, and let $A : \mathcal{X} \to \mathcal{Y}$ and $B : \mathcal{Y} \to \mathcal{Z}$ be linear continuous operators. Then, the duality functor $\mathcal{D}$ is contravariant, viz., $\mathcal{D}$ assigns to the sequence*

$$\mathcal{X} \xrightarrow{\ A\ } \mathcal{Y} \xrightarrow{\ B\ } \mathcal{Z}$$

the following sequence:

$$\mathcal{X}^* \xleftarrow{\ A^*\ } \mathcal{Y} \xleftarrow{\ B^*\ } \mathcal{Z}^*$$

PROOF.

- We have to show that
$$(BA)^* = A^* B^*.$$
- This follows immediately from
$$\big\langle (BA)^* v^*, u \big\rangle = \langle v^*, BAu \rangle = \langle B^* v^*, Au \rangle = \langle A^* B^* v^*, u \rangle\,, \quad \forall u \in \mathcal{X}, v^* \in \mathcal{Z}^*. \qquad \blacksquare$$

Corollary 284. *If the operator $A : \mathcal{X} \to \mathcal{Y}$ is linear, continuous, and bijective, then so is the dual operator $A^* : \mathcal{Y}^* \to \mathcal{X}^*$. Moreover, we get*

$$(A^*)^{-1} = (A^{-1})^*. \tag{9.58}$$

Thus we can denote $(A^*)^{-1}$ by $A^{-\star}$ since the inverse and dual operations are commutative.

PROOF.

- Let $\mathbb{1}_{\mathcal{X}}$ denote the identity operator on $\mathcal{X}$. It follows from
$$A^{-1}A = \mathbb{1}_{\mathcal{X}} \quad \text{and} \quad AA^{-1} = \mathbb{1}_{\mathcal{Y}}$$
that
$$A^*(A^{-1})^* = \mathbb{1}_{\mathcal{X}^*} \quad \text{and} \quad (A^{-1})^* A^* = \mathbb{1}_{\mathcal{Y}^*},$$
- Since $\mathbb{1}_{\mathcal{X}}^* = \mathbb{1}_{\mathcal{X}^*}$ and $\mathbb{1}_{\mathcal{Y}}^* = \mathbb{1}_{\mathcal{Y}^*}$, we have $(A^*)^{-1} = (A^{-1})^*$. $\blacksquare$

Let $\mathcal{X}$ be a Banach space over $\mathbb{F}$. If we set

$$j_{\mathcal{X}}(u)(f) \triangleq \langle f, u \rangle\,, \quad \forall u \in \mathcal{X}, f \in \mathcal{X}^*,$$

then the linear continuous operator $j_{\mathcal{X}} : \mathcal{X} \to \mathcal{X}^{**}$ preserves the norm, that is, $\|j_{\mathcal{X}}(u)\| = \|u\|$ for all $u \in \mathcal{X}$. Set

$$A^{\star\star} \triangleq (A^*)^*.$$

Proposition 285. *Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces over $\mathbb{F}$.*

① *The following diagram*

$$\begin{array}{ccc} \mathcal{X} & \xrightarrow{\ j_{\mathcal{X}}\ } & \mathcal{X}^{**} \\ {\scriptstyle A}\downarrow & & \downarrow{\scriptstyle A^{\star\star}} \\ \mathcal{Y} & \xrightarrow[\ j_{\mathcal{Y}}\ ]{} & \mathcal{Y}^{**} \end{array}$$

is commutative for all operators $A \in \mathrm{L}(\mathcal{X}, \mathcal{Y})$.

② *The duality functor $\mathcal{D}$ is norm-preserving, that is,*

$$\|A^*\| = \|A\|\,, \quad \forall A \in \mathrm{L}(\mathcal{X}, \mathcal{Y}).$$

③ *The duality functor $\mathcal{D}$ is compact, that is, if $A \in \mathrm{L}(\mathcal{X}, \mathcal{Y})$ is compact, then so is $A^* \in \mathrm{L}(\mathcal{Y}^*, \mathcal{X}^*)$.*

9.6.2 Exactness of the Dual Functor

> Riemann has shown us that proofs are better achieved through ideas than through long calculations.
>
> —David Hilbert—

Concepts of Exactness

The *language* of exact sequences plays a fundamental role in modern mathematics (e.g., in algebraic topology, differential geometry, and computational conformal geometry). We want to show that this language allows us to give elegant proofs in linear operator theory, see Fig. 9.2.

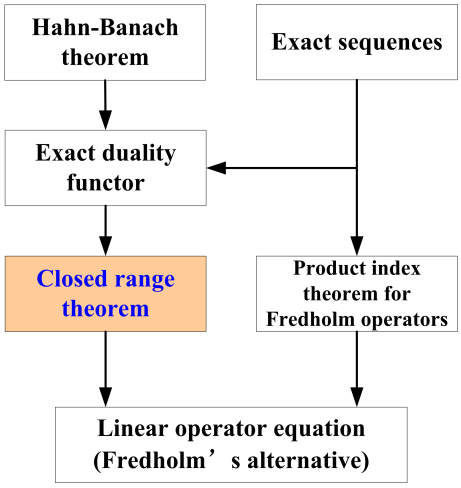


Figure 9.2: Exact sequences and linear operators

Definition 286. *Let $\mathcal{X}_1, \cdots, \mathcal{X}_n$ be linear spaces over $\mathbb{F}$, and let $A_j : \mathcal{X}_j \to \mathcal{X}_{j+1}, j = 1, \cdots, n-1$, be linear operators. Then the sequence*

$$\mathcal{X}_1 \xrightarrow{A_1} \mathcal{X}_2 \xrightarrow{A_2} \mathcal{X}_3 \xrightarrow{A_3} \cdots \xrightarrow{A_{n-2}} \mathcal{X}_{n-1} \xrightarrow{A_{n-1}} \mathcal{X}_n \tag{9.59}$$

is called exact iff $\text{Range}(A_j) = \text{Ker}(A_{j+1})$ *for all* $j = 1, \cdots, n-2$.

The sequence (9.59) is called an *exact Banach sequence* iff it is exact and all the operators

$$A_j : \mathcal{X}_j \to \mathcal{X}_{j+1}, \quad j = 1, \cdots, n-1$$

are linear and continuous, where $\mathcal{X}_1, \cdots, \mathcal{X}_n$ are Banach spaces over $\mathbb{F}$ and the range $\text{Range}(A_{n-1})$ is closed.[4]

In particular, the exactness of

$$\mathcal{X} \xrightarrow{A} \mathcal{Y} \xrightarrow{B} \mathcal{Z}$$

means that $\text{Range}(A) = \text{Ker}(B)$.

Example. Let $A : \mathcal{X} \to \mathcal{Y}$ be a linear operator, where $\mathcal{X}$ and $\mathcal{Y}$ are linear spaces over $\mathbb{F}$. Then

① A is *injective* iff the sequence

$$0 \longrightarrow \mathcal{X} \xrightarrow{A} \mathcal{Y} \tag{9.60}$$

is exact.

[4]This implies that all the ranges $\text{Range}(A_j), j = 1, \cdots, n-1$, are closed. In fact, we have $\text{Range}(A_j) = \text{Ker}(A_{j+1})$, and the null space $\text{Ker}(A_{j+1})$ is closed for all $j = 1, \cdots, n-2$, since A_{j+1} is continuous.

② A is *surjective* iff the sequence

$$\mathcal{X} \xrightarrow{A} \mathcal{Y} \longrightarrow 0 \tag{9.61}$$

is exact.

③ A is bijective iff the sequence

$$0 \longrightarrow \mathcal{X} \xrightarrow{A} \mathcal{Y} \longrightarrow 0 \tag{9.62}$$

is exact.

Here, $0 \longrightarrow \mathcal{X}$ and $\mathcal{Y} \longrightarrow 0$ denote the trivial maps $0 \mapsto 0$ and $u \mapsto 0$, respectively.

PROOF.

- Ad ①: The exactness of (9.60) means that $\mathrm{Ker}(A) = \{0\}$.
- Ad ②: The exactness of (9.61) means that $\mathrm{Range}(A) = \mathcal{Y}$.
- Ad ③: The exactness of (9.62) is equivalent to the exactness of (9.60) and (9.61).

Proposition 287. *Let*

$$0 \longrightarrow \mathcal{X} \xrightarrow{A} \mathcal{Y} \xrightarrow{B} \mathcal{Z} \longrightarrow 0$$

be an exact sequence, where $\mathcal{X}, \mathcal{Y}$, and $\mathcal{Z}$ are finite-dimensional linear spaces over $\mathbb{F}$. Then

$$\dim(\mathcal{X}) - \dim(\mathcal{Y}) + \dim(\mathcal{Z}) = 0.$$

PROOF.

- The operator A is injective. Hence $\dim(\mathrm{Range}(A)) = \dim(\mathcal{X})$.
- Let $\mathcal{W}$ denote an algebraic complement of $\mathrm{Ker}(B)$ in $\mathcal{Y}$:

  $$\mathcal{Y} = \mathrm{Ker}(B) \oplus \mathcal{W}. \tag{9.63}$$

  The operator B is surjective. Thus, the restriction $B : \mathcal{W} \to \mathcal{Z}$ is bijective, and hence $\dim(\mathcal{Z}) = \dim(\mathcal{W})$.
- Since $\mathrm{Ker}(B) = \mathrm{Range}(A)$, it follows from (9.63) that

  $$\dim(\mathcal{Y}) = \dim(\mathrm{Range}(A)) + \dim(\mathcal{W}) = \dim(\mathcal{X}) + \dim(\mathcal{Z}). \quad \blacksquare$$

Proposition 288. *The duality functor $\mathcal{D}$ is exact, i.e., $\mathcal{D}$ sends exact Banach sequences to exact sequences.*

This theorem tells us that the closedness of the range $\mathrm{Range}(A_1)$ in (9.59) implies $\mathrm{Range}(A_1^*) = \mathrm{Ker}(A_1)^\perp$; thus the range $\mathrm{Range}(A_1^*)$ is also closed. Consequently, we get the following stronger result: the duality functor $\mathcal{D}$ sends exact Banach sequences to exact Banach sequences. PROOF.

- Step-1: Let us first consider the short exact Banach sequence

  $$\mathcal{X} \xrightarrow{A} \mathcal{Y} \xrightarrow{B} \mathcal{Z}$$

  That is, $\mathrm{Range}(A) = \mathrm{Ker}(B)$, and $\mathrm{Range}(B)$ is a closed linear subspace of $\mathcal{Z}$. We have to show that

  $$\mathcal{X}^* \xleftarrow{A^*} \mathcal{Y}^* \xleftarrow{B^*} \mathcal{Z}^*$$

  is an exact sequence, that is, $\mathrm{Range}(B^*) = \mathrm{Ker}(A^*)$.

– Since $\text{Range}(A) = \text{Ker}(B)$, we get

$$BA = 0,$$

and hence $A^*B^* = (BA)^* = 0$. This implies

– Conversely, we now show that $\text{Ker}(A^*) \subset \text{Range}(B^*)$. To this end, choose $u^* \in \text{Ker}(A^*)$. Hence $u^* \in \mathcal{Y}^*$ and

$$\langle u^*, Au\rangle = \langle A^*u^*, u\rangle = 0, \quad \forall u \in \mathcal{X}.$$

This yields $u^*(v) \triangleq \langle u^*, v\rangle = 0$ for all $v \in \text{Range}(A)$. Define

$$[u^*](v + \text{Range}(A)) \triangleq u^*(v), \quad \forall v \in \mathcal{Y}.$$

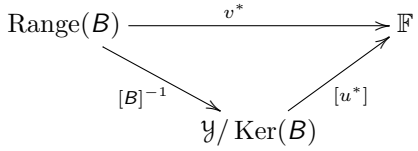
It follows as in the proof of Proposition 261 that the linear functional

$$[u^*] : \mathcal{Y}/\,\text{Range}(A) \to \mathbb{F}$$

is *continuous*. Letting $[B](v + \text{Ker}(B)) \triangleq Bv$ for all $v \in \mathcal{Y}$, we get the linear homeomorphism

$$[B] : \mathcal{Y}/\,\text{Ker}(B) \to \text{Range}(B),$$

by Proposition 261. Observe that the range $\text{Range}(B)$ is closed. The decisive trick of our proof consists in introducing the linear functional v^* through the commutative diagram



that is, we set

$$v^* \triangleq [u^*][b]^{-1}. \tag{9.64}$$

Recall that $\text{Range}(A) = \text{Ker}(B)$. The functional v^* is continuous on $\text{Range}(B)$. Hence

$$|v^*(w)| \leq \text{const}\,\|w\|, \quad \forall w \in \text{Range}(B),$$

where $\text{Range}(B) \subset \mathcal{Z}$. By the Hahn-Banach theorem (Theorem 224), there exists a linear continuous extension

$$v^* : \mathcal{Z} \to \mathbb{F}.$$

Relation (9.64) tells us that, for all $v \in \mathcal{Y}$,

$$\begin{aligned} v^*(Bv) &= [u^*][B]^{-1}(Bv) \\ &= [u^*](v + \text{Ker}(B)) = [u^*](v + \text{Range}(A)) = u^*(v). \end{aligned}$$

This yields

$$\langle v^*, Bv\rangle = \langle u^*, v\rangle, \quad \forall v \in \mathcal{Y},$$

and hence $u^* = B^*v^*$, which means $u^* \in \text{Range}(B^*)$. Therefore,

$$\text{Ker}(A^*) \subset \text{Range}(B^*).$$

- Step-2: The general case can easily be reduced to Step-1. In fact, the sequence in (9.59) is an exact Banach sequence iff all the possible short sequences

$$\mathcal{X}_j \xrightarrow{A_j} \mathcal{X}_{j+1} \xrightarrow{A_{j+1}} \mathcal{X}_{j+2}$$

are exact Banach sequences for $j = 1, \cdots, n-2$. ■

Exact Sequences and Embedding Map

Example 1. Let $\mathcal{X}$ be a closed linear subspace of the Banach space $\mathcal{Y}$ over $\mathbb{F}$, and let $j : \mathcal{X} \to \mathcal{Y}$ denote the trivial embedding map defined through $j(u) \triangleq u$ for all $u \in \mathcal{X}$. Then j is *injective*, i.e., the sequence

$$0 \longrightarrow \mathcal{X} \xrightarrow{\ j\ } \mathcal{Y}$$

is an exact Banach sequence. By Proposition 288, the dual sequence

$$0 \longleftarrow \mathcal{X}^* \xleftarrow{\ j^*\ } \mathcal{Y}^*$$

is also exact (i.e., the dual operator j^* is *surjective*). Moreover, $\mathrm{Ker}(j^*) = \mathcal{X}^\perp$.
PROOF.

- For all $u \in \mathcal{X}$ and $u^* \in \mathcal{Y}^*$,
$$\langle j^*(u^*), u\rangle_{\mathcal{X}} = \langle u^*, j(u)\rangle_{\mathcal{Y}}\,.$$
Therefore, the functional $j^*(u^*)$ represents the restriction of the functional $u^* : \mathcal{Y} \to \mathbb{F}$ to the subspace $\mathcal{X}$. Obviously,
$$j^*(u^*) = 0 \quad \text{iff} \quad u^* = 0 \quad \text{on} \quad \mathcal{X}$$
(i.e., $u^* \in \mathcal{X}^\perp$). Hence $\mathrm{Ker}(j^*) = \mathcal{X}^\perp$. ■

Exact Sequence and Projection

Example 2. Let $\mathcal{X}$ be a closed subspace of the Banach space $\mathcal{Y}$ over $\mathbb{F}$, and let

$$\pi : \mathcal{X} \to \mathcal{Y}/\mathcal{X}$$

be the canonical mapping defined by $\pi(u) \triangleq u + \mathcal{X}$ for all $u \in \mathcal{X}$. Obviously, $\mathrm{Ker}(\pi) = \mathcal{X}$. Since π is linear, continuous, and surjective, the sequence

$$\mathcal{Y} \xrightarrow{\ \pi\ } \mathcal{Y}/\mathcal{X} \longrightarrow 0$$

is exact. By Example 1,

$$0 \longrightarrow \mathcal{X} \xrightarrow{\ j\ } \mathcal{Y} \xrightarrow{\ \pi\ } \mathcal{Y}/\mathcal{X} \longrightarrow 0$$

is an exact Banach sequence. It follows from Proposition 288 that the dual sequence

$$0 \longleftarrow \mathcal{X}^* \xleftarrow{\ j^*\ } \mathcal{Y}^* \xleftarrow{\ \pi^*\ } (\mathcal{Y}/\mathcal{X})^* \longleftarrow 0$$

is exact. Hence the dual operator π^* is *injective*, and $\mathrm{Range}(\pi^*) = \mathrm{Ker}(j^*) = \mathcal{X}^\perp$.

9.6.3 Closed Range Theorem and Fredholm Alternatives

The following result represents the most important theorem on linear operator equations.

Theorem 289 (Banach's Closed Range Theorem)**.** *Let $A : \mathcal{X} \to \mathcal{Y}$ be a linear continuous operator where $\mathcal{X}$ and $\mathcal{Y}$ are Banach spaces over $\mathbb{F}$. Then the following three conditions are equivalent:*

① *Fredholm alternative.*

$$\mathrm{Range}(A) =^\perp \mathrm{Ker}(A^*), \quad \mathrm{Range}(A^*) = \mathrm{Ker}(A)^\perp.$$

② *Closed range:* $\mathrm{Range}(A)$ *is closed.*

③ *A priori estimate: There is a constant $c > 0$ such that*

$$c \cdot \operatorname{dist}(u, \operatorname{Ker}(A)) \le \|Au\|, \quad \forall u \in \mathcal{X}. \tag{9.65}$$

PROOF.

- ①$\Longrightarrow$②. By Proposition 277, the set ${}^{\perp}\operatorname{Ker}(A^*)$ is closed.
- ②$\Longrightarrow$①. Let $\operatorname{Range}(A)$ be closed. According to the examples in the previous subsection, we have

  $$0 \longrightarrow \operatorname{Ker}(A) \xrightarrow{\ j\ } \mathcal{X} \xrightarrow{\ A\ } \mathcal{Y} \xrightarrow{\ \pi\ } \mathcal{Y}/\operatorname{Range} A \longrightarrow 0$$

  represents an exact Banach sequence. By Proposition 288, the dual sequence

  $$0 \longleftarrow \operatorname{Ker}(A)^* \xleftarrow{\ j^*\ } \mathcal{X}^* \xleftarrow{\ A^*\ } \mathcal{Y}^* \xleftarrow{\ \pi^*\ } \mathcal{Y}/\operatorname{Range} A^* \longleftarrow 0$$

  is exact. This implies

  $$\operatorname{Ker}(A^*) = \operatorname{Range}(\pi^*), \quad \operatorname{Range}(A^*) = \operatorname{Ker}(j^*).$$

  By the examples in the previous subsection, we get

  $$\operatorname{Range}(\pi^*) = \operatorname{Range}(A)^{\perp}, \quad \operatorname{Ker}(j^*) = \operatorname{Ker}(A)^{\perp}.$$

  Since $\operatorname{Range}(A)$ is closed, it follows from Proposition 278 that

  $$\operatorname{Range}(A) = \overline{\operatorname{Range}(A)} =^{\perp} (\operatorname{Range}(A))^{\perp}.$$

  Hence

  $$\operatorname{Range}(A^*) = \operatorname{Ker}(A)^{\perp}, \quad \operatorname{Range}(A) =^{\perp} \operatorname{Ker}(A^*).$$

- ②$\Longrightarrow$③. This is Proposition 261.
- ③$\Longrightarrow$②. First let $\operatorname{Ker}(A) = \{0\}$. Then

  $$c \cdot \|u\| \le \|Au\|, \quad \forall u \in \mathcal{X}$$

  This implies that $\operatorname{Range}(A)$ is closed.

  - In fact, if $Au_n \to v$ as $n \to \infty$, then (Au_n) is Cauchy, and $c\|u_n - u_m\| \le \|Au_n - Au_m\|$ shows that (u_n) is also Cauchy. Hence $u_n \to u$ as $n \to \infty$, that is, $Au = v$.
  - If $\operatorname{Ker}(A) \neq \{0\}$, then we use the operator

    $$[A] : \mathcal{X}/\operatorname{Ker}(A) \to \mathcal{Y}$$

    from Proposition 261. Recall that $[A][u] \triangleq Au$ for all $u \in \mathcal{X}$. Thus, the a priori estimate in (9.65) is equivalent to

    $$c\,\|[u]\| \le \|[A][u]\|, \quad \forall [u] \in \mathcal{X}/\operatorname{Ker}(A).$$

    The same argument as the preceding one shows that $\operatorname{Range}([A])$ is closed. Since $\operatorname{Range}(A) = \operatorname{Range}([A])$, the range $\operatorname{Range}(A)$ is also closed. ■

In terms of the operator equation

$$Au = b, \quad u \in \mathcal{X}, \tag{9.66}$$

and its dual equation

$$A^*u^* = b^*, \quad u^* \in \mathcal{Y}^*, \tag{9.67}$$

The Banach's closed range theorem (Theorem 289) means the following:

Let the range $\operatorname{Range}(A)$ be closed, then:

(a) For given $b \in \mathcal{Y}$, the original equation (9.66) has a solution iff

$$\langle u^*, b\rangle = 0 \tag{9.68}$$

for all solution u^* of the homogeneous dual equation (9.67), i.e., $A^*u^* = 0$.

(b) Conversely, for given $b^* \in \mathcal{X}^*$, the dual equation (9.67) has a solution iff

$$\langle b^*, u\rangle = 0 \tag{9.69}$$

for all solutions u of the homogeneous original equation (9.66), i.e., $Au = 0$.

Observe that condition (9.68) is quite natural. In fact, if $Au = b$ and $A^*u^* = 0$, then

$$\langle u^*, b\rangle = \langle u^*, Au\rangle = \langle A^*u^*, u\rangle = \langle 0, u\rangle = 0. \tag{9.70}$$

Thus (9.68) represents a simple *necessary* solvability condition or (9.66). The closed range theorem tells us that this condition is also a *sufficient* solvability condition provided that the range $\mathrm{Range}(A)$ is *closed*.

Furthermore, if $A^*u^* = b^*$ and $Au = 0$, then

$$\langle b^*, u\rangle = \langle A^*u^*, u\rangle = \langle u^*, Au\rangle = \langle u^*, 0\rangle = 0. \tag{9.71}$$

This is the solvability condition (9.69) for the dual equation (9.67).

If $\mathcal{X}$ and $\mathcal{Y}$ are finite-dimensional spaces, then $\mathrm{Range}(A)$ is closed automatically. In this case, statements (a) and (b) correspond to classic results on finite linear systems.

Corollary 290 (Closed Range Theorem for Hilbert Spaces)**.** *Let $A : \mathcal{X} \to \mathcal{X}$ be linear continuous operator on the Hilbert space $\mathcal{X}$ over $\mathbb{F}$. Then the following two conditions are equivalent:*

- $\mathrm{Range}(A) = \mathrm{Ker}(A^*)^\perp$ *and* $\mathrm{Range}(A^*) = \mathrm{Ker}(A)^\perp$.
- $\mathrm{Range}(A)$ *is closed.*

Proof.

- By Proposition 281, we have
$$A^\dagger = J^{-1}A^*J,$$
and $\langle Ju, v\rangle = \langle u|v\rangle$ for all $u, v \in \mathcal{X}$. The assertion follows from Theorem 289.

In terms of the operator equation

$$Au = b, \quad u \in \mathcal{X},$$

and its *adjoint* equation

$$A^\dagger v = c, \quad v \in \mathcal{X}, \tag{9.72}$$

this means the following facts:

Let the range $\mathrm{Range}(A)$ be closed, then we have

(a) For given $b \in \mathcal{X}$, the original equation (9.66) has a solution iff

$$\langle v|b\rangle = 0$$

for all solutions v of the homogeneous adjoint equation (9.72), i.e., $A^\dagger v = 0$.

(b) For given $c \in \mathcal{X}$, the adjoint equation (9.72) has a solution iff

$$\langle c|u\rangle = 0$$

for all solutions u of the homogeneous original equation (9.66), i.e., $Au = 0$.

In fact, by Theorem 289, the original equation (9.66) has a solution iff

$$\langle u^*, b\rangle = 0, \quad \text{for all } u^* \text{with} \quad A^*u^* = 0.$$

If we let $v \triangleq J^{-1}u^*$, this is equivalent to

$$\langle v|b\rangle = 0 \quad \text{for all } v \text{ with} \quad J^{-1}A^*Jv = 0,$$

that is, $b \in \mathrm{Ker}(A^\dagger)^\perp$. Hence $\mathrm{Range}(A) = \mathrm{Ker}(A^\dagger)^\perp$.

Moreover, the adjoint equation (9.72) can be written as

$$J^{-1}A^*(Jv) = J^{-1}c.$$

By Theorem 289, this equation has a solution iff

$$\left\langle J^{-1}c, u\right\rangle = 0, \quad \text{for all } u \text{ with } Au = 0.$$

This is equivalent to

$$\langle c|u\rangle = 0, \quad \text{for all } u \text{ with } Au = 0.$$

(i.e., $c \in \mathrm{Ker}(A)^\perp$). Hence $\mathrm{Range}(A^\dagger) = \mathrm{Ker}(A)^\perp$.

Example 1

Let $A : \mathcal{X} \to \mathcal{Y}$ be a linear continuous operator, where $\mathcal{X}$ and $\mathcal{Y}$ are Banach spaces over $\mathbb{F}$. If $\mathrm{codim}(\mathrm{Range}(A)) < \infty$, then the range $\mathrm{Range}(A)$ is closed.

Example 2

Let $A : \mathcal{X} \to \mathcal{Y}$ be a linear continuous operator, where $\mathcal{X}$ and $\mathcal{Y}$ are Banach spaces over $\mathbb{F}$. Then the following two conditions are equivalent:

- A priori estimate. There is a constant $c > 0$ such that

$$c\,\|u\| \le \|Au\|\,, \forall u \in \mathcal{X}.$$

- The range $\mathrm{Range}(A)$ is closed and $Au = 0$ implies $u = 0$.

Example 3

Let $A : \mathcal{X} \to \mathcal{Y}$ be a linear continuous operator, where $\mathcal{X}$ and $\mathcal{Y}$ are Banach spaces over $\mathbb{F}$. Furthermore, let $\mathcal{Z}$ be a Banach space over $\mathbb{F}$ such that the embedding

$$\mathcal{X} \subset \mathcal{Z}$$

is *compact*. Then the following two statements are equivalent:

- A priori estimate. There is a constant $c > 0$ such that

$$c\,\|u\|_{\mathcal{X}} \le \|Au\|_{\mathcal{Y}} + \|u\|_{\mathcal{Z}}\,, \forall u \in \mathcal{X}.$$

- The range $\mathrm{Range}(A)$ is closed and $\dim(\mathrm{Ker}(A)) < \infty$.

This result plays an important role in the theory of elliptic-type linear partial differential equations.

Part III

Unifying the Theory and Practices

Chapter 10

Exercises and Problems

10.1 Preliminaries

10.1.1 Integrations and Functionals

For $f \in C[a,b]$, let

$$J(f) \triangleq \int_a^b f(x)\,\mathrm{d}\,x.$$

① Verify that $J : C[a,b] \to \mathbb{R}$ is a linear functional.

② Design a program to compute the $J(f)$ with C/C++ (or MATLAB, etc.) and verify its correctness with concrete examples.

10.1.2 Norm in $C[a,b]$

The norm in $C[a,b]$ is defined by

$$\|f\| \triangleq \max_{x \in C[a,b]} |f(x)|\,.$$

① For $f(x) = -x^2 - 2x + 3$ and $[a,b] = [0.5, 3]$, find $\|f\|$.

② For $f(x) = -2x^3 + 21x^2 - 60x + 50$ and $[a,b] = [0.5, 3.5]$, find $\|f\|$.

③ How to find $\|f\|$ with one-dimensional searching method ? (Optimization)

10.1.3 Mahalanobis Distance

The Mahalanobis distance was proposed by P. C. Mahalanobis in 1936 for the purpose of applications in statistics. For the two sample vector $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{n\times 1}$ and the covariance matrix $\boldsymbol{C}$, the Mahalanobis distance is defined by

$$D_M(\boldsymbol{x}, \boldsymbol{y}) \triangleq \sqrt{(\boldsymbol{x}-\boldsymbol{y})^{\mathsf{T}} \boldsymbol{C}^{-1} (\boldsymbol{x}-\boldsymbol{y})}$$

① Show that if $\boldsymbol{C} = \mathbb{1}$, i.e., $\boldsymbol{C}$ is the identity of $\mathbb{R}^{n\times n}$, then the Mahalanobis distance is just the Euclidean distance.

② For $n = 2$, $\boldsymbol{x} = [1,3]^{\mathsf{T}}, \boldsymbol{y} = [-2,4]^{\mathsf{T}}$ and $\boldsymbol{C} = \begin{bmatrix} 5 & 1 \\ 1 & 5 \end{bmatrix}$, compute $D_M(\boldsymbol{x}, \boldsymbol{y})$.

③ Let $\boldsymbol{A} \in O(n, \mathbb{R}) = \left\{ \boldsymbol{P} \in \mathbb{R}^{n\times n} : \boldsymbol{P}^{\mathsf{T}}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}^{\mathsf{T}} = \mathbb{1} \right\}$. Is $D_M(\boldsymbol{x}, \boldsymbol{y}) = D_M(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{A}\boldsymbol{y})$? Or equivalently, is the Mahalanobis distance invariant under the act of the orthogonal transformation group?

10.1.4 Threshold Distance and Histograms

Suppose $d(\cdot,\cdot)$ is a metric (distance) on a linear normed space $\mathcal{X}$ over $\mathbb{F}$, and t is a positive constant, let

$$d_t(x,y) \triangleq \min(t, d(x,y)), \quad \forall x,y \in \mathcal{X}$$

show that $d_t(\cdot,\cdot)$ is a metric.

Remark. The metric d_t has the following properties:

(1) It corresponds to the way humans perceive distances

(2) It is robust to outlier noise and quantization effects

(3) It is a kind of metric.

(4) Thresholding the ground distance improves both accuracy and speed in practical computation process.

For more details, see the reference:

- O. Pele,et al., Fast and Robust Earth Mover's Distances, In ICCV 2009.
- `http://www.cs.huji.ac.il/~ofirpele/FastEMD/`

10.2 Banach Spaces

10.2.1 Iterative Method for Linear Systems

For linear algebraic equation

$$\boldsymbol{x} = \boldsymbol{C}\boldsymbol{x} + \boldsymbol{b}, \quad \boldsymbol{C} = (c_{ij})_{n\times n} \in \mathbb{R}^{n\times n}, \boldsymbol{x} \in \mathbb{R}^{1\times}, \boldsymbol{b} \in \mathbb{R}^{n\times 1},$$

let

$$\|\boldsymbol{A}\| \triangleq \max_i \sum_{j=1}^{n} |a_{ij}|, \quad \forall \boldsymbol{A} = (a_{ij})_{m\times n} \in \mathbb{R}^{m\times n}.$$

① Show that $\|\cdot\|$ is a norm defined on $\mathbb{R}^{m\times n}$ ($m,n \geq 1$ are fixed integers).

② Prove that if $\|C\| < 1$, then the equation $\boldsymbol{x} = \boldsymbol{C}\boldsymbol{x} + \boldsymbol{b}$ have unique solution.

③ Set

$$\boldsymbol{C} = \begin{bmatrix} \frac{1}{6} & \frac{-1}{6} & \frac{1}{7} \\ \frac{1}{6} & \frac{1}{5} & \frac{1}{3} \\ \frac{-1}{6} & \frac{1}{3} & \frac{-1}{4} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}.$$

- Find the numerical solution with MATLAB.
- Find the numerical solution via the iterative method with the following iteration formula

$$\boldsymbol{x}_{k+1} = T\boldsymbol{x}_k \triangleq \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{b}$$

and estimate the error.

④ Suppose the linear systems

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \quad \boldsymbol{A} \in \mathbb{R}^{n\times n}, \boldsymbol{x}, \boldsymbol{b} \in \mathbb{R}^{n\times 1}$$

has a unique solution, i.e., $\mathrm{Rank}(\boldsymbol{A}) = n$.

- Please program with C/C++ (or MATLAB, Java, Python) to find the numerical solution with the SOR method.
- Verify the correctness, efficiency and robustness of your program with concrete examples.

10.2.2 Integration Equation

For the following integration equation:

$$\phi(x) = 1 + \frac{1}{10}\int_0^1 K(x,t)\phi(t)\,\mathrm{d}\,t$$

where

$$K(x,t) = \begin{cases} x, & 0 \le x < t; \\ t, & t \le x \le 1. \end{cases}$$

① Find an approximation solution $\hat{\phi}$ such that the error $\left\|\phi - \hat{\phi}\right\| \le 10^{-4}$.

② Suppose the integration equation

$$\phi(x) = g(x) + \lambda\int_a^b K(x,t)\phi(t)\,\mathrm{d}\,t$$

has a unique solution $\phi(x)$ on $[a,b]$, design a program to find the numerical solution $\{\langle x_k, \phi(x_k)\rangle\}_{k=0}^n$ with sufficient large integer n, where $\{x_k\}$ are nodes such that

$$a = x_0 < x_1 < \cdots < x_k < \cdots < x_n = b.$$

③ Verify the correctness of your program with the help of ①.

10.2.3 Newton's Method for Roots

For $f : \mathbb{R} \to \mathbb{R}$ such that $f \in C^2(\mathbb{R})$, if $\hat{x}$ is the unique root of $f(x) = 0$ in $[a,b]$. Let

$$x_{n+1} = \mathsf{T}x_n = x_n - \frac{f(x_n)}{f'(x_n)},$$

where (x_n) is a sequence in some neighbourhood of $\hat{x}$.

① Prove that T is contractive.

② Make a program with C/C++ to find the root $\hat{x}$.

③ For a fixed positive a and $f(x) = x^2 - a$, design an iterative formula for computing $\sqrt{a}$. Particulary, for $a = 3$, find the numerical solution with the program obtained in ②.

Consider the following area-March number relation encountered in aerodynamics

$$\left(\frac{A}{A^*}\right)^2 = \frac{1}{M^2}\left[1 + \frac{\gamma-1}{2}M^2\right]^{(\gamma+1)/(\gamma-1)},$$

where $\gamma = \frac{c_p}{c_v} = 1.4$ is a constant for the air, A is the area of the cross section of the nozzle, A^* is the area at the throat, $\frac{A}{A^*}$ is the ratio of area.

(1) For the given ratio $\frac{A}{A^*}$, how many solutions for the March number are there?

(2) Design an iterative formula

$$M_{k+1} = \mathsf{T}M_k$$

for the March number M such that T is contractive.

(3) For $\frac{A}{A^*} = 2$, find the March number with the program obtained in ②.

10.3 Hilbert Spaces

10.3.1 Simple Identities

Let $\mathcal{X}$ be a pre-Hilbert space (i.e., inner product space) over $\mathbb{F}$ with the inner product $\langle\cdot|\cdot\rangle$. Show that the following hold true:

① If $\mathbb{F} = \mathbb{R}$, then

$$4\langle u|v\rangle = \|u+v\|^2 - \|u-v\|^2, \quad \forall u, v \in \mathcal{X}.$$

② If $\mathbb{F} = \mathbb{C}$, then

$$4\langle u|v\rangle = \|u+v\|^2 - \|u-v\|^2 - \mathrm{i}\,\|u+\mathrm{i}v\|^2 + \mathrm{i}\,\|u-\mathrm{i}v\|^2, \quad \forall u, v \in \mathcal{X}.$$

③ Appolonius' identity. If $\mathbb{F} = \mathbb{R}, \mathbb{C}$, then

$$\|w-u\|^2 + \|w-v\|^2 = \frac{1}{2}\|u-v\|^2 + 2\left\|w - \frac{u+v}{2}\right\|^2, \quad \forall u, v, w \in \mathcal{X}.$$

10.3.2 The Role of the Parallelogram Identity

Let $\mathcal{X}$ be a normed space over $\mathbb{F}$. Show that $\mathbb{F}$ is an inner product space iff the parallelogram identity holds, i.e.,

$$2\|u\|^2 + 2\|v\|^2 = \|u-v\|^2 + \|u-v\|^2, \quad \forall u, v \in \mathcal{X}.$$

Hint: Use the results obtained in the previous "simple identities" ① and ②.

10.3.3 Least Squares Approach for Over-determined Linear Systems

Consider the linear systems with noise

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} + \boldsymbol{v}, \quad \boldsymbol{A} \in \mathbb{R}^{m\times n}, \boldsymbol{x} \in \mathbb{R}^{n\times 1}, \boldsymbol{b} \in \mathbb{R}^{m\times 1}, m \geq n$$

where $\mathrm{Rank}(\boldsymbol{A}) = n$ and $\boldsymbol{v}$ is the noise vector such that $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\,\mathbb{1})$. Let

$$\boldsymbol{P} = \boldsymbol{A}(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathsf{T}},$$

① Show that $\boldsymbol{P}$ is a projection.

② Solve the following problem

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^{n\times 1}} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2$$

where $\|\cdot\|_2$ denotes the Euclidean norm.

③ Give a geometric interpretation of the solution in ② with the help of ①.

10.3.4 Best Squares Approximation

Let $\phi_0, \phi_1, \cdots, \phi_n \in C[a,b]$ be linearly independent and

$$\mathcal{V}_n = \mathrm{span}\,\{\phi_0, \phi_1, \cdots, \phi_n\} \subset C[a,b].$$

Define the norm $\|\cdot\|_2$ as

$$\|f-g\|_2 \triangleq \sqrt{\int_a^b |f(x)-g(x)|^2\,\mathrm{d}\,x}, \quad \forall f, g \in C[a,b].$$

For $f \in C[a,b]$, the solution of the problem

$$\hat{f} = \arg \inf_{u \in \mathcal{V}_n} \|f - u\|_2$$

is called the best squares approximation of f.

We now take $f(x) = \mathrm{e}^x, x \in [-1,1]$, $\phi_0(x) = 1, \phi_1(x) = x, \phi_2(x) = x^2$.

① For $\mathcal{V}_1 = \operatorname{span}\{\phi_0, \phi_1\} = \operatorname{span}\{1, x\}$, find the best square approximation $\hat{f}_1$.

② For $\mathcal{V}_2 = \operatorname{span}\{\phi_0, \phi_1, \phi_2\} = \operatorname{span}\{1, x, x^2\}$, find the best square approximation $\hat{f}_2$.

③ Compare the solutions $\hat{f}_1$ and $\hat{f}_2$ with the exact function $f(x) = \mathrm{e}^x$.

10.3.5 The Ritz Method

Problem 1. The variational problem

$$\min_{u \in C^2[0,\pi]} F(u), \quad \text{s.t.} \quad u(0) = u(\pi) = 0,$$

where

$$F(u) = \int_0^\pi \left(\frac{1}{2} \left[\frac{\mathrm{d}\, u}{\mathrm{d}\, x} \right]^2 - u \cos x \right) \mathrm{d}\, x,$$

is equivalent to the boundary-value problem

$$u''(x) + \cos x = 0, \quad x \in [0,\pi], \quad u(0) = u(\pi) = 0,$$

which has a unique solution u. Explicitly,

$$u(x) = \cos x + \frac{2x}{\pi} - 1.$$

Use the Ritz method in order to compute an approximate solution u_{2n} of $\min\limits_{u \in C^2[0,\pi]} F(u)$ with the constraint $u(0) = u(\pi) = 0$, by making the sum

$$u_{2n}(x) = \sum_{k=1}^{2n} c_k \sin kx.$$

Determine the coefficients $c_1, \cdots, c_{2n}$. Show that the sequence (u_{2n}) converges uniformly on $[0,\pi]$ to the solution exact u.

Problem 2. Suppose $y = y(x) : [0,1] \to \mathbb{R}$ and let

$$J(y) = \int_0^1 \left[\left(\frac{\mathrm{d}\, y}{\mathrm{d}\, x} \right)^2 - y^2 + 4xy \right] \mathrm{d}\, x.$$

Solve the optimization problem:

$$y_{opt} = \arg \min_{y \in C[0,1]} J(y) \quad \text{s.t.} \quad y(0) = y(1) = 0$$

① Find the analytical solution.

② Find a numerical solution with the form $y_n = \sum\limits_{k=0}^{n} a_k x^k (1-x)$. Take $n = 1$ and $n = 2$ respectively.

③ Compare the analytical solution with the numerical results.

10.4 Generalized Functions

10.5 Soblev Spaces

10.6 Fourier Analysis

10.6.1 Fourier Series and Funciton Representation

10.6.2 Calculate the Fourier Transform

With the help of the second definiton of Fourier Transform

$$F(\omega)=\int_{-\infty}^{\infty}f(t)\mathrm{e}^{-\mathrm{j}\omega t}\,\mathrm{d}\,t,\qquad f(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}f(t)\mathrm{e}^{\mathrm{j}\omega t}\,\mathrm{d}\,\omega,$$

determine the $F(\omega)=\mathcal{F}[f(t)]$ and plot the graphs of the functions:

① Gate function (rectangular pulse)

$$f(t)=\begin{cases}E, & |t|\le\frac{\tau}{2}\quad(\tau>0)\\ 0, & \text{otherwise.}\end{cases}\tag{10.1}$$

② Exponential decay function

$$f(t)=\begin{cases}\mathrm{e}^{-\alpha t}, & t\ge 0\quad(\alpha>0)\\ 0, & \text{otherwise.}\end{cases}\tag{10.2}$$

③ Trianglular function

$$f(t)=\begin{cases}\frac{2A}{\tau}\left(\frac{\tau}{2}+t\right), & -\frac{\tau}{2}\le t<0,\\ \frac{2A}{\tau}\left(\frac{\tau}{2}-t\right), & 0\le t\le\frac{\tau}{2},\\ 0, & \text{otherwise.}\end{cases}\tag{10.3}$$

④ Gaussian function

$$f(t)=\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(t-\mu)^2}{2\sigma^2}},\quad \sigma>0,\mu\in\mathbb{R}\tag{10.4}$$

⑤ Rectangular RF impulse

$$f(t)=\begin{cases}E\cos\omega_0 t, & |t|\le\frac{\tau}{2}\\ 0, & \text{otherwise.}\end{cases}\tag{10.5}$$

⑥ Periodic impulses

$$f(t)=\sum_{n=-\infty}^{\infty}\delta(t-nT)\tag{10.6}$$

⑦ Unit step function

$$f(t)=U(t)=\begin{cases}1, & t\ge 0\\ 0, & \text{otherwise.}\end{cases}\tag{10.7}$$

⑧ Cosine and sine

$$f(t)=a\sin\omega_0 t+b\cos\omega_0 t\tag{10.8}$$

⑨ Sign function

$$\operatorname{sign}(t)=\begin{cases}1, & t>0\\ 0, & t=0\\ -1, & t<0.\end{cases}\tag{10.9}$$

10.6.3 Heat Equation and Fourier Transform

Consider the PDE

$$\begin{cases} \dfrac{\partial u}{\partial t} = D\dfrac{\partial^2 u}{\partial x^2}, & (x,t) \in \mathbb{R} \times [0, +\infty) \\ u(x,0) = g(x), & x \in \mathbb{R}, t = 0. \end{cases} \tag{10.10}$$

① Show that if $\lim_{(x,t)\to(x_0,0+)} u(x,t) = f(x_0)$, then we have

$$u(x,t) = \frac{1}{\sqrt{4\pi D t}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{(x-y)^2}{4Dt}} g(y)\,\mathrm{d}\,y, \quad t > 0.$$

② Let

$$H(x-y,t) \triangleq \frac{1}{\sqrt{4\pi D t}} \mathrm{e}^{-\frac{(x-y)^2}{4Dt}}, \quad t > 0$$

Verify that

$$\frac{\partial H}{\partial t} = D\frac{\partial^2 H}{\partial x^2}.$$

Note that $H(x-y,t)$ is called the fundamental solution (thermal kernal) of the heat equation.

10.6.4 Signals and Systems

Linear System

Nonlinear System

The voltage-current relation of some nonelinear device is specified by

$$i = b_0 + b_1(v - V_0) + b_2(v - V_0)^2 + b_3(v - V_0)^3.$$

Suppose that the voltage across the device is

$$v = V_0 + V_{1m}\cos\omega_1 t + V_{2m}\cos\omega_2 t, \quad \omega_1 > \omega_2.$$

Find out the frequency spectrum of i and interpret the results. Hint:

$$\begin{aligned} i = b_0 &+ \frac{1}{2}b_2 V_{1m}^2 + \frac{1}{2}b_2 V_{2m}^2 \\ &+ (\cdots)\cos\omega_1 t \\ &+ (\cdots)\cos\omega_2 t \\ &+ (\cdots)\cos 2\omega_1 t + (\cdots)\cos 2\omega_2 t \\ &+ (\cdots)\cos(\omega_1+\omega_2)t + (\cdots)\cos(\omega_1-\omega_2)t \\ &+ (\cdots)\cos 3\omega_1 t + (\cdots)\cos 3\omega_2 t \\ &+ (\cdots)\cos(2\omega_1+\omega_2)t \\ &+ (\cdots)\cos(2\omega_1-\omega_2)t \\ &+ (\cdots)\cos(\omega_1+2\omega_2)t \\ &+ (\cdots)\cos(\omega_1-2\omega_2)t \end{aligned}$$

10.7 Eigenvalue Problem

10.7.1 Oscillators and Schrödinger Equation

The 1-dim stationary Schrödinger equation in quantum mechanics can be expressed by the eigenvalue problem

$$\mathsf{H}\psi = E\psi,$$

where ψ is the wave function (quantum state function) such that

$$\int_{-\infty}^{\infty} |\psi(x)|^2 \,\mathrm{d}\, x = 1,$$

and the E is the energy eigenvalue. Define the Hamiltonian (energy operator) H as

$$(\mathsf{H}\psi)(x) \triangleq \left[-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}\, x^2} + \frac{1}{2}m\omega^2 x^2\right]\psi(x)$$

where m is the mass of a particle, $\hbar$ is the reduced Planck's constant, ω is the angular frequency. The eigenvectors of

$$\mathsf{H}\psi = E\psi$$

span a Hilbert space $\mathcal{L} = \mathrm{span}\,\{\psi_0, \psi_1, \psi_2, \cdots\}$ with infinite dimension.

① Let $\alpha = \sqrt{\frac{m\omega}{\hbar}}, \beta = \sqrt{\frac{1}{m\hbar\omega}}$. Find the dimension of α and β.

② Let $\xi = \alpha x, \mathsf{A} = \frac{1}{2}\xi^2 - \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}\,\xi^2}, \lambda = \frac{E}{\hbar\omega}$, show that $\mathsf{H}\psi = E\psi$ can be rewritten by

$$\mathsf{A}\psi(\xi) = \lambda\psi(\xi)$$

and equivalently

$$\frac{\mathrm{d}^2}{\mathrm{d}\,\xi^2}\psi(\xi) - (\xi^2 - 2\lambda)\psi(\xi) = 0.$$

③ Let $\psi(\xi) = H(\xi)\mathrm{e}^{-\frac{\xi^2}{2}}$, show that

$$\frac{\mathrm{d}^2}{\mathrm{d}\,\xi^2}H(\xi) - 2\xi\frac{\mathrm{d}}{\mathrm{d}\,\xi}H(\xi) + (2\lambda - 1)H(\xi) = 0.$$

④ Find the corresponding eigenvalue λ such that $|H(\xi)|$ is finite for any finite ξ. Hint: Suppose that $H(\xi) = \xi^\rho \sum_{k=0}^{\infty} a_k \xi^k$ and find out the ρ and a_k.

⑤ Let

$$H_n(z) = (-1)^n \mathrm{e}^{z^2}\frac{\mathrm{d}^n}{\mathrm{d}\, z^n}\mathrm{e}^{-z^2}, \quad n = 0, 1, 2, \cdots.$$

Show that

- $H_0(z) = 1, H_1(z) = 2z, H_2(z) = 4z^2 - 2.$
- $H_n'(z) = 2nH_{n-1}(z)$, for $n \geq 1$.
- $H_n''(z) - 2zH_n'(z) + 2nH_n(z) = 0.$
- $H_n(z) = 2zH_{n-1}(z) - H_{n-1}'(z)$, for $n \geq 1$.
- $H_n(z) = 2zH_{n-1}(z) - 2(n-1)H_{n-2}(z)$, for $n \geq 2$.

⑥ With the help of H_n, the n-th eigenvector ψ of H can be expressed by

$$\psi_n(x) = \left(\frac{\alpha^2}{\pi}\right)^{\frac{1}{4}}\frac{1}{\sqrt{2^n n!}}\mathrm{e}^{-\xi^2/2}H_n(\xi) = \left(\frac{\alpha^2}{\pi}\right)^{\frac{1}{4}}\frac{1}{\sqrt{2^n n!}}\mathrm{e}^{-\alpha^2 x^2/2}H_n(\alpha x), \quad n = 0, 1, 2, \cdots$$

The states described by $\psi_0, \psi_1, \psi_2, \cdots$ are called the ground state, the first excited state, the second excited state, ..., respectively.

⑦ Let $\hat{x} = x, \hat{p} = -\mathrm{i}\hbar\frac{\mathrm{d}}{\mathrm{d}\,x}$. Define the annihilation operator a_- and creation operator a_+ as

$$a_- = \frac{1}{\sqrt{2}}\left(\xi + \frac{\mathrm{d}}{\mathrm{d}\,\xi}\right) = \frac{1}{\sqrt{2}}\left(\alpha\hat{x} + \mathrm{i}\beta\hat{p}\right),$$
$$a_+ = \frac{1}{\sqrt{2}}\left(\xi - \frac{\mathrm{d}}{\mathrm{d}\,\xi}\right) = \frac{1}{\sqrt{2}}\left(\alpha\xi - \mathrm{i}\beta\hat{p}\right).$$

Show that

- $H = \hbar\omega\left(a_+a_- + \frac{1}{2}\right)$
- For $n \geq 1$, $a_-\psi_n(x) = \sqrt{n}\psi_{n-1}(x)$.
- For $n = 0, 1, 2, \cdots$, $a_+\psi_n(x) = \sqrt{n+1}\psi_{n+1}(x)$.

⑧ Find the expression of $\psi_n(x)$ by using a_+ and $\psi_0(x)$.

⑨ Find the eigenvector of $a_\pm$, i.e., solve the eigen equations

$$a_\pm\phi(x) = \lambda_\pm\phi(x), \quad \phi \in \mathcal{L}.$$

The state ϕ is called the coherent state in quantum optics. Hint: Expand $\phi(x)$ via Fourier series with the help of the basis $\{\psi_n\}_{n=0}^{\infty}$.

In quantum mechanics, an observable is automatically associated with a Hermitian linear operator: Let F be an observable, and λ_n the possible values that can be the result of measuring F. We can define the mean value and standard variance of measuring the observable F as

$$< F >= \langle\psi|F\psi\rangle, \quad \sigma_F = \sqrt{< F^2 > - < F >^2}$$

when the state of the quantum system is ψ.

❶ For the ground state ψ_0,

- Calculate $\sigma_{\hat{x}}, \sigma_{\hat{p}}$
- Verify that $\sigma_{\hat{x}} \cdot \sigma_{\hat{p}} \geq \frac{\hbar}{2}$.

❷ Calculate $< a_\pm >= \langle\phi|a_\pm\phi\rangle$.

Observe the iterative relations appeared in ⑤ and design a C/C++ computer program to compute $H_n(z)$. Please verify your program and evaluate its performance.
Remark: The $H_n(z)$ are called Hermitian polynomials and the $H_n(z)\mathrm{e}^{-z^2/2}$ are called Gauss-Hermite functions.

10.7.2 Stochastic Processes and Eigenvalue Problems

Let $K_x(t,u) = \mathrm{E}\{(x(t) - \mathrm{E}\{x(t)\})(x(u) - \mathrm{E}\{x(u)\})\}$ be the covariance function of random process $x(t)$, then $K_x(t,u)$ is symmetric and nonnegative definite. Define

$$(\mathsf{A}\phi)(t) = \int_{T_i}^{T_f} K_x(t,u)\phi(u)\,\mathrm{d}\,u, \quad \forall t,u \in [T_i, T_f]$$

Consider the following eigenvalue problem

$$\mathsf{A}\phi = \lambda\phi.$$

① For the Wiener process, we can show that

$$K_x(t,u) = \sigma^2 \min(t,u)$$

where σ^2 is positive constant. Find the discrete eigenvalues λ_n and the corresponding eigenvectors $\phi_n(t)$.

② For the bandlimited stochastic signal $x(t)$, its bandlimited spectra is

$$S_x(\omega) = \begin{cases} \frac{\pi P}{\alpha}, & |\omega| \le \alpha; \\ 0, & |\omega| > \alpha, \end{cases}$$

where $\alpha = 2\pi W$, and α, P, W are constants. The corresponding covariance function is

$$K_x(t,u) = P\frac{\sin\alpha(t-u)}{\alpha(t-u)}.$$

The integral equation of interest becomes

$$\int_{-T/2}^{T/2} P\frac{\sin\alpha(t-u)}{\alpha(t-u)}\phi(u)\,\mathrm{d}\,u = \lambda\phi(t),$$

where the initial time is $T_i = -\frac{T}{2}$ and the final time is $T_f = \frac{T}{2}$. For fixed $c = \frac{\alpha T}{2} = \pi WT$,

- For $2WT = 2.55$ and $2WT = 5.10$, find the eigenvalues λ_n. Hint: for $n = 0, 1, \cdots, 5$, the values of λ_n are

Table 10.1: $2WT = 2.55$ and $2WT = 5.10$

n	λ_n	Remark	λ_n	Remark
0	$0.996\frac{P}{2W}$	$2WT = 2.55$	$1.000\frac{P}{2W}$	$2WT = 5.10$
1	$0.912\frac{P}{2W}$	$2WT = 2.55$	$0.999\frac{P}{2W}$	$2WT = 5.10$
2	$0.519\frac{P}{2W}$	$2WT = 2.55$	$0.997\frac{P}{2W}$	$2WT = 5.10$
3	$0.110\frac{P}{2W}$	$2WT = 2.55$	$0.961\frac{P}{2W}$	$2WT = 5.10$
4	$0.009\frac{P}{2W}$	Ignorable	$0.748\frac{P}{2W}$	$2WT = 5.10$
5	$0.0004\frac{P}{2W}$	Ignorable	$0.321\frac{P}{2W}$	$2WT = 5.10$
6		Ignorable	$0.061\frac{P}{2W}$	$2WT = 5.10$
7		Ignorable	$0.006\frac{P}{2W}$	Ignorable
8		Ignorable	$0.0004\frac{P}{2W}$	Ignorable

- Check that for $n > 2WT + 1$, the values of λ_n rapidly approach zero. Note that $N = \lfloor 2WT \rfloor$ is important in information theory, which is introduced by Claude E. Shannon.

– Check the total energy in the remaining eigenvalues, for

$$\sum_{n=0}^{\infty} \lambda_n = \int_{-T/2}^{T/2} K_x(t,t)\,\mathrm{d}\,t = PT.$$

③ Is A a compact operator?

④ Design a C/C++ program to find the numerical solution to $A\phi = \lambda\phi$ when $K_x(t,u), T_f, T_i$ are given. Vefify and evaluate your methods and results obtained. Hint: Convert the integral equation into the corresponding matrix equation $\boldsymbol{A\phi} = \lambda\boldsymbol{\phi}$.

Appendix A

Factorials, Polynomials and Hypergeometric Series

A.1 Symbols for Factorial

A.1.1 Knuth k-product

$$
\begin{aligned}
w^{[k]} &= w(w+1)(w+2)\cdots(w+k-1) \\
w_{[k]} &= w(w-1)(w-2)\cdots(w-k+1) \\
w^{[k]} &= (w+k-1)_{[k]} \\
(-w)_{[k]} &= (-1)^k w^{[k]} = (-w)(-w-1)(-w-2)\cdots(-w-k+1)
\end{aligned}
$$

The symbol $w^{[k]}$ is introduced by Donald E. Knuth (It is Knuth who created the TEX !) [6]. In mathematical and physics literatures, $w^{[k]}$ may be replaced by $w^{\overline{k}}$ and $w_{[k]}$ may be replaced by $w^{\underline{k}}$.

MATLAB procedure for computing $w^{[k]}$: See `zhy_maths_prodku`

MATLAB procedure for computing $w_{[k]}$: See `zhy_maths_prodkd`

A.1.2 Factorial

$$\Gamma(n+1) = n! = 1 \cdot 2 \cdot \cdots \cdot (n-1) \cdot n = 1^{[n]} = n_{[n]}$$

where $\Gamma(\cdot)$ is the (complete) Γ-function

$$\Gamma(x) = \int_0^\infty \mathrm{e}^{-t} t^{x-1}\,\mathrm{d}\,t = \lim_{a\to\infty} \gamma(x,a) = \lim_{a\to\infty} \int_0^a \mathrm{e}^{-t} t^{x-1}\,\mathrm{d}\,t$$

Another notations related with factorial:

$$
\begin{aligned}
n!! &= n(n-2)(n-4)\cdots \\
n!!! &= n(n-3)(n-6)\cdots
\end{aligned}
$$

A.1.3 Binomial Coefficients

$$\begin{aligned}
\binom{n}{k} &= \frac{n!}{k!(n-k)!} = C_n^k, \quad k \in \{0, 1, 2, \cdots, n\} \\
\binom{\alpha}{k} &= \frac{\alpha \cdot (\alpha-1) \cdots (\alpha-k+1)}{k!} = \frac{\alpha_{[k]}}{k!} = \frac{\alpha_{[k]}}{1^{[k]}}, \quad \forall\ \alpha \in \mathbb{R} \\
\binom{-r}{k} &= (-1)^k \cdot \binom{r+k-1}{k}, \quad \forall\ r \in \mathbb{R} \\
(x+y)^n &= \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}, \quad \forall x, y \in R \text{ and } xy = yx \\
(1+z)^\alpha &= \sum_{k=0}^{\infty} \binom{\alpha}{k} z^k, \quad \forall \alpha \in \mathbb{R}, z \in \mathbb{C}, |z| \le 1
\end{aligned}$$

MATLAB procedure for computing $\binom{w}{k}$: See `zhy_maths_binom`

A.1.4 Miltinomial Coefficients

$$\begin{aligned}
\binom{k_1+k_2+k_3+\cdots+k_r}{k_1, k_2, k_3, \cdots, k_r} &= \frac{(k_1+k_2+k_3+\cdots+k_r)!}{k_1!k_2!k_3!\cdots k_r!} \\
(x_1+x_2+\cdots+x_r)^n &= \sum_{\substack{k_1,k_2,\cdots,k_r \\ k_1+k_2+\cdots+k_r=n}} \binom{n}{k_1, k_2, \cdots, k_r} x_1^{k_1} x_2^{k_2} \cdots x_r^{k_r} \\
\binom{k_1+k_2+k_3}{k_1, k_2, k_3} &= \frac{(k_1+k_2+k_3)!}{k_1!k_2!k_3!} = \binom{k_1+k_2+k_3}{k_1} \cdot \binom{k_2+k_3}{k_2} = \cdots \\
(x+y+z)^n &= \sum_{\substack{k_1,k_2,k_3 \\ k_1+k_2+k_3=n}} \binom{n}{k_1, k_2, k_3} x^{k_1} y^{k_2} z^{k_3}, \quad \forall x, y, z \in \mathbb{C}
\end{aligned}$$

A.2 Polynomials

A.2.1 Polynomials and Ring

Let $\mathcal{P}[x]$ denote the set of all the polynomials with coefficients in field $\mathbb{R}$, i.e.,

$$\mathcal{P}[x] = \left\{ f(x) = \sum_k a_k x^k \ : \ a_k \in \mathbb{R} \right\}$$

Define the multiplication and addition as follows

$$\begin{aligned}
\forall f(x) = \sum_k a_k x^k \in \mathcal{P}[x], g(x) &= \sum_j b_j x^j \in \mathcal{P}[x] \\
f(x) + g(x) &= (a_0+b_0) + (a_1+b_1)x + (a_2+b_2)x^2 + \cdots + (a_i+b_i)x^i + \cdots \\
f(x) \cdot g(x) &= \left(\sum_k a_k x^k \right) \cdot \left(\sum_j b_j x^j \right) \\
&= \sum_k \sum_j a_k b_j x^{k+m} \\
&= (a_0+b_0) + (a_1 b_0 + a_0 b_1)x + (a_2 b_0 + a_1 b_1 + a_0 b_2)x^2 + \cdots
\end{aligned}$$

Then the triplet $\langle \mathcal{P}[x], +, \cdot \rangle$ must be a ring.

If we let $\boldsymbol{a} = (a_0, a_1, a_2, \cdots, \cdots)$ and $\boldsymbol{b} = (b_0, b_1, \cdots, b_2, \cdots)$ be two infinite sequence determined by $f(x)$ and $g(x)$ respectively, then we have the following bijections

$$\begin{aligned} f(x) &\longleftrightarrow \boldsymbol{a} = (a_0, a_1, a_2, \cdots, a_k, \cdots) \\ g(x) &\longleftrightarrow \boldsymbol{b} = (b_0, b_1, b_2, \cdots, b_k, \cdots) \\ f(x) + g(x) &\longleftrightarrow \boldsymbol{a} + \boldsymbol{b} = (a_0 + b_0, a_1 + b_1, a_2 + b_2, \cdots) \\ f(x) \cdot g(x) &\longleftrightarrow \boldsymbol{a} \circledast \boldsymbol{b} = (a_0 b_0, a_1 b_0 + a_0 b_1, a_2 b_0 + a_1 b_1 + a_0 b_2, \cdots) \end{aligned}$$

Attention, for the degrees of two polynomials and their product, we have

$$\deg(f(x) \cdot g(x)) = \deg(f(x)) + \deg(g(x))$$

A.2.2 Polynomials and Vector Space

Let $\mathcal{P}_n[x]$ be the set of one-variable polynomials whose degrees are less than or equal to n, i.e.

$$\mathcal{P}_n[x] = \left\{ f(x) = \sum_{k=0}^{n} a_k x^k \ : \ a_k \in \mathbb{R} \right\} \tag{A.1}$$

and define the scalar multiplication and addition as follows

$$\begin{aligned} \cdot \ : \ \mathbb{R} \times \mathcal{P}_n[x] &\longrightarrow \mathcal{P}_n[x] \\ \left(\lambda, \sum_{k=0}^{n} a_k x^k\right) &\mapsto \lambda \sum_{k=0}^{n} a_k x^k = \sum_{k=0}^{n} \lambda a_k x^k \\ + \ : \ \mathcal{P}_n[x] \times \mathcal{P}_n[x] &\longrightarrow \mathcal{P}_n[x] \\ \left(\sum_{k=0}^{n} a_k x^k, \sum_{j=0}^{n} b_j x^j\right) &\mapsto \sum_{k=0}^{n} a_k x^k + \sum_{j=0}^{n} b_j x^j = \sum_{k=0}^{n} (a_k + b_k) x^k \end{aligned}$$

then the triplet $\langle \mathcal{P}_n[x], +, \cdot \rangle$ must be a vector space.

Similarly, for the set of two-variables polynomials

$$\mathcal{P}_n[x, y] = \left\{ f(x, y) = \sum_{\substack{k,j \\ 0 \le k+j \le n}} a_k^j x^k y^j \ : \ a_{kj} \in \mathbb{R} \right\} \tag{A.2}$$

and two polynomials $f, g \in \mathcal{P}_n[x, y]$,

$$\begin{aligned} f(x, y) &= a_n^0 x^n + a_{n-1}^1 x^{n-1} y + \cdots a_{n-k}^k x^{n-k} y^k + \cdots + a_1^{n-1} x y^{n-1} + a_0^n y^n \\ g(x, y) &= b_n^0 x^n + b_{n-1}^1 x^{n-1} y + \cdots b_{n-k}^k x^{n-k} y^k + \cdots + b_1^{n-1} x y^{n-1} + b_0^n y^n \end{aligned}$$

we can define the addition and multiplication as follows

$$\begin{aligned} + \ : \ \mathcal{P}_n[x, y] \times \mathcal{P}_n[x, y] &\longrightarrow \mathcal{P}_n[x, y] \\ (f(x, y), g(x, y)) &\mapsto f(x, y) + g(x, y) \\ f(x, y) + g(x, y) &= (a_n^0 + b_n^0) x^n + (a_{n-1}^1 + b_{n-1}^1) x^{n-1} y + \\ &\qquad \cdots + (a_{n-k}^k + b_{n-k}^k) x^{n-k} y^k + \cdots + (a_1^{n-1} + b_1^{n-1}) x y^{n-1} + (a_0^n + b_0^n) y^n \\ \cdot \ : \ \mathbb{R} \times \mathcal{P}_n[x, y] &\longrightarrow \mathcal{P}_n[x, y] \\ (\lambda, f(x, y)) &\mapsto \lambda f(x, y) \\ \lambda \cdot f(x, y) &= \lambda a_n^0 x^n + \lambda a_{n-1}^1 x^{n-1} y + \cdots \lambda a_{n-k}^k x^{n-k} y^k + \cdots + \lambda a_1^{n-1} x y^{n-1} + \lambda a_0^n y^n \end{aligned}$$

then the triplet $\langle \mathcal{P}_n[x, y], +, \cdot \rangle$ must be a vector space.

In general, $\langle \mathcal{P}_n[x_1, \cdots, x_m], +, \cdot \rangle$ is also a vector space.

A.3 Hypergeometric Series

A.3.1 Definition

The hypergeometric series is very useful in mathematical physics and engineering fields [8].

Definition 291 (Gauss)**.**

$$\begin{aligned} {}_2F_1(\alpha,\beta;\gamma|x) &= 1+\frac{\alpha\cdot\beta}{\gamma}\cdot\frac{x}{1!}+\frac{\alpha(\alpha+1)\cdot\beta(\beta+1)}{\gamma(\gamma+1)}\cdot\frac{x^2}{2!}+\cdots \\ &= 1+\frac{\alpha^{[1]}\beta^{[1]}}{\gamma^{[1]}}\cdot\frac{x}{1!}+\frac{\alpha^{[2]}\beta^{[2]}}{\gamma^{[2]}}\cdot\frac{x^2}{2!}+\cdots+\frac{\alpha^{[k]}\beta^{[k]}}{\gamma^{[k]}}\cdot\frac{x^k}{k!}+\cdots \end{aligned} \tag{A.3}$$

Why it is named hypergeometrical series? For the coefficient of x^k, we have

$$a_k=\frac{\alpha^{[k]}\beta^{[k]}}{\gamma^{[k]}}\cdot\frac{1}{k!}=\frac{\alpha^{[k]}\beta^{[k]}}{\gamma^{[k]}1^{[k]}},$$

Therefore,

- If $a_k\equiv 1$, then the series is the geometrical series.
- If $a_k=\frac{1}{k!}$, namely $\alpha^{[k]}\beta^{[k]}\equiv\gamma^{[k]}$, then we will get the exponent function e^x.

A more convenient notation for the hypergeometric series is introduced by Knuth [7], viz.

$$\mathrm{F}\left(\left.\frac{\alpha,\beta}{\gamma}\right|x\right)=1+\frac{\alpha^{[1]}\beta^{[1]}}{\gamma^{[1]}}\cdot\frac{x}{1!}+\frac{\alpha^{[2]}\beta^{[2]}}{\gamma^{[2]}}\cdot\frac{x^2}{2!}+\cdots+\frac{\alpha^{[k]}\beta^{[k]}}{\gamma^{[k]}}\cdot\frac{x^k}{k!}+\cdots \tag{A.4}$$

Obviously, it can be extended to the more general case, i.e.,

$$\mathrm{F}\left(\left.\frac{\alpha_1,\alpha_2,\cdots,\alpha_r}{\gamma_1,\gamma_2,\cdots,\gamma_s}\right|x\right)=\sum_{k=0}^{\infty}\frac{{\alpha_1}^{[k]}\cdots{\alpha_r}^{[k]}}{{\gamma_1}^{[k]}\cdots{\gamma_s}^{[k]}}\cdot\frac{x^k}{k!} \tag{A.5}$$

A.3.2 Relation with other Special functions

The hypergeometrical series is relationed with Legendre, Jacobi and Tchebycheff Polynomials closely.

$$\begin{aligned} P_n(x) &= \mathrm{F}\left(\left.\frac{-n,n+1}{1}\right|x\right)=1+\frac{(-n)(n+1)}{1!}\cdot\frac{x}{1!}+\frac{(-n)(-n+1)(n+1)(n+2)}{2!}\cdot\frac{x^2}{2!}+\cdots \\ G_n(p,q,x) &= \mathrm{F}\left(\left.\frac{p+n,-n}{q}\right|x\right)=1+\frac{(p+n)(-n)}{q}\cdot\frac{x}{1!}+\frac{(p+n)(p+n+1)(-n)(-n+1)}{q(q+1)}\cdot\frac{x^2}{2!}+\cdots \\ P_n(x) &= G_n\left(1,1,\frac{1-x}{2}\right)={}_2F_1\left(n+1,-n;1\left|\frac{1-x}{2}\right.\right) \\ T_n(x) &= \frac{1}{2^{n-1}}G_n\left(0,\frac{1}{2},\frac{1-x}{2}\right)=\frac{1}{2^{n-1}}\cdot\mathrm{F}\left(\left.\frac{n,-n}{1/2}\right|\frac{1-x}{2}\right) \end{aligned} \tag{A.6}$$

Appendix B

Zernike Polynomials

B.1 A Global View

In mathematics, the Zernike polynomials [1, 2, 4, 5] are a sequence of polynomials that are orthogonal on the unit disk. Named after Frits Zernike, they play an important role in beam optics and optical design[3]. The Zernike polynomials, which are orthogonal over a circular pupil, represent balanced aberrations that yield minimum variance. For small aberrations, a minimum of aberration variance yields a maximum of Strehl ratio.

There are even and odd Zernike polynomials. The even ones are defined as

$$\mathrm{Z}_n^m(\rho,\theta) = \mathrm{R}_n^m(\rho)\cos(m\theta), \quad j = j(n,m) \text{ is even.}$$

and the odd ones as

$$\mathrm{Z}_n^{-m}(\rho,\theta) = \mathrm{R}_n^m(\rho)\sin(m\theta), \quad j = j(n,m) \text{ is odd.}$$

where:

- m and n are nonnegative integers such that $n \geq m$ and $n - m$ is even;
- θ is the azimuthal angle;
- ρ is the radial distance such that $0 \leq \rho \leq 1$;
- j is the order of the Zernike polynomials introduced by Noll and $j = 1, 2, 3, \cdots$;
- $\mathrm{R}_n^m(\rho)$ is the radial polynomials.

B.2 Some General Considerations for Zernike Polynomials

It is not difficult to show that there exists an infinity of complete sets of polynomials in two real variables x, y which are orthogonal for the interior of the unit circle, i.e. which satisfy the orthogonality condition [1]

$$\begin{aligned}\langle V_\alpha | V_\beta \rangle &\triangleq \iint\limits_{\|\boldsymbol{x}\|\leq 1} \overline{V}_\alpha(\boldsymbol{x}) V_\beta(\boldsymbol{x})\,\mathrm{d}^2\,\boldsymbol{x} \\ &= \iint\limits_{x^2+y^2\leq 1} \overline{V}_\alpha(x,y) V_\beta(x,y)\,\mathrm{d}\,x\,\mathrm{d}\,y \\ &= A_{\alpha\beta}\delta_{\alpha\beta}\end{aligned} \tag{B.1}$$

Here V_α and V_β denote two typical polynomials of the binary polynomial ring $P[x, y]$, $\star$ denotes the complex conjugates, δ is the Kronecker symbol, and $A_{\alpha\beta}$ are normalization constants. The circle polynomials of Zernike are distinguished from the other sets by certain simple invariance properties which can best be explained from group theoretical considerations.

B.2.1 Rotation invariant

Invariant in form with respect to rotations of axes about the origin.

$$SO(2,\mathbb{R}) \ni \boldsymbol{U}(\theta) : \boldsymbol{x} \mapsto \boldsymbol{x}' \qquad \begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{B.2}$$

This implies that

$$V(x,y) = G(\theta)V(x',y') \quad \text{or} \quad V(\boldsymbol{x}) = G(\theta)V(\boldsymbol{x}') \tag{B.3}$$

where $G(\theta)$ is a continuous function with period 2π of the angle of rotation θ, and $G(0) = 1$.

Now the application of two successive rotations through angles θ_1 and θ_2 is equivalent to a single rotation through an angle $\theta_1 + \theta_2$. Thus G is an isomorphism

$$G(\theta_1)G(\theta_2) = G(\theta_1 + \theta_2) \tag{B.4}$$

How to solve it?

$$\begin{aligned}
&G(0)\cdot G(0) = G(0) \Longrightarrow G(0) = 1 \\
&G(\theta+\Delta\theta)G(\theta) = G(\Delta\theta) \\
&\Longrightarrow \frac{G(\theta+\Delta\theta)-G(\theta)}{\Delta\theta} = \frac{G(\theta)[G(\Delta\theta)-1]}{\Delta\theta} = G(\theta)\cdot\frac{G(\Delta\theta)-G(0)}{\Delta\theta-0} \\
&\Longrightarrow \lim_{\Delta\theta\to 0}\frac{G(\theta+\Delta\theta)-G(\theta)}{\Delta\theta} = G(\theta)\lim_{\Delta\theta\to 0}\frac{G(\Delta\theta)-G(0)}{\Delta\theta} \\
&\Longrightarrow \frac{\mathrm{d}\,G(\theta)}{\mathrm{d}\,\theta} = G(\theta)\cdot G'(0) \\
&\Longleftrightarrow G(\theta) = G(0)\cdot \mathrm{e}^{G'(0)\theta} = \mathrm{e}^{a\theta}, \quad a = G'(0)
\end{aligned}$$

The general solution with period 2π is

$$G(\theta) = \mathrm{e}^{\mathrm{i}\ell\theta}, \quad a = \mathrm{i}\ell, \ell \in \mathbb{Z} \tag{B.5}$$

On substituting from (B.5) into (B.3), setting $x' = r, y' = 0$, and using (B.2) it follows that V must be of the form

$$V(\rho\cos\theta, \rho\sin\theta) = R(\rho)\mathrm{e}^{\mathrm{i}\ell\theta} \tag{B.6}$$

where $\rho = \|\boldsymbol{x}\| = \sqrt{x^2+y^2}, R(\rho) = V(\rho, 0)$ is a function of ρ alone.

Suppose V is a polynomial of degree n in the variable $x = \rho\cos\theta, y = \rho\sin\theta$, i.e., $\deg(V) = n$; it then follows from (B.6) that:

- $R(\rho)$ is a polynomial in r of degree n and contains no power of of ρ of degree lower that $|\ell|$.
- $R(\rho)$ is evidently an even or an odd polynomial according as ℓ is even or odd, i.e. $n - |\ell|$ is even.

B.2.2 Radial Polynomials and Zernike Polynomials

The set of the **Zernike circle polynomials** is distinguished from all other such sets by the property that it **contains a polynomial for each pair of the permissible values of n (degree) and ℓ (angular dependence)**, i.e. for integral values of n and ℓ, such that $n \geq 0$, $\ell \in \mathbb{Z}$, $n \geq |\ell|$, and $n - |\ell|$ is even. We denote a typical polynomial of this set by [1][4]

$$\boxed{\mathrm{Z}_n^\ell(\rho\cos\theta, \rho\sin\theta) = \mathrm{R}_n^\ell(\rho)\mathrm{e}^{\mathrm{i}\ell\theta}, \quad n \in \{0,1,2,\cdots\}, |\ell| \leq n, \text{ and } n-|\ell| \text{ is even.}} \tag{B.7}$$

It follows from (B.1) and (7) that

$$\begin{aligned}
\langle \mathrm{Z}_\alpha \,|\, \mathrm{Z}_\beta \rangle = \langle \mathrm{Z}_n^\ell \,|\, \mathrm{Z}_n^\ell \rangle &\triangleq \iint\limits_{\|\boldsymbol{x}\|\leq 1} \overline{\mathrm{Z}}_\alpha(x,y)\,\mathrm{Z}_\beta(x,y)\,\mathrm{d}\,x\,\mathrm{d}\,y \\
&= \int_0^{2\pi} \mathrm{d}\,\theta \int_0^1 \rho\,\mathrm{d}\,\rho \cdot \mathrm{R}_n^\ell\,\mathrm{e}^{-\mathrm{i}\ell\theta} \cdot \mathrm{R}_{n'}^\ell\,\mathrm{e}^{\mathrm{i}\ell\theta} \\
&= 2\pi \int_0^1 \rho\,\mathrm{R}_n^\ell(\rho) \cdot \mathrm{R}_{n'}^\ell(\rho)\,\mathrm{d}\,\rho \\
&= \delta_{nn'} \cdot a_n^\ell \cdot 2\pi \\
&= A_n^\ell \delta_{nn'}, \qquad a_n^\ell = \frac{A_n^\ell}{2\pi}
\end{aligned}$$

The radial polynomials $\mathrm{R}_n^\ell(\rho)$ satisfy the relation

$$\langle n\ell | n'\ell \rangle \triangleq \int_0^1 \mathrm{R}_n^\ell(\rho) \cdot \mathrm{R}_{n'}^\ell(\rho)\rho\,\mathrm{d}\,\rho = a_n^\ell \delta_{nn'} \tag{B.8}$$

For any given value ℓ, the lower index n can only take the values $|\ell|, |\ell|+2, |\ell|+4, \cdots$. The corresponding sequence $\mathrm{R}_{|\ell|}^\ell(\rho), \mathrm{R}_{|\ell|+2}^\ell(\rho), \mathrm{R}_{|\ell|+4}^\ell(\rho), \cdots$ may be obtained by orthogonalizing the powers

$$\rho^{|\ell|}, \rho^{|\ell|+2}, \rho^{|\ell|+4}, \cdots \tag{B.9}$$

with the weighting factor ρ over the interval $0 \leq \rho \leq 1$. Moreover, since only the absolute values of ℓ occur in (B.9),

$$\mathrm{R}_n^{-\ell}(\rho) = \beta_n^\ell\,\mathrm{R}_n^\ell(\rho), \tag{B.10}$$

where β_n^ℓ is a const depending only on the normalization of the two polynomials $\mathrm{R}_n^{-\ell}$ and R_n^ℓ. In particular we may normalize in such a way that

$$\beta_n^\ell = 1, \quad \forall\, \ell \in \mathbb{Z}, \forall\, n \in \{0,1,2,\cdots\}$$

and then

$$\mathrm{Z}_n^{\pm m}(\rho\cos\theta, \rho\sin\theta) = \mathrm{R}_n^m(\rho)\mathrm{e}^{\pm\mathrm{i}m\theta}, \qquad m = |\ell| \tag{B.11}$$

Theorem 292. *The set of the circle polynomials contains $\frac{1}{2}(n+1)(n+2)$ linearly independent polynomials of degree less than or equal to n. Hence every monomial $x^i y^j$ ($i \geq 0, j \geq 0$ integers) and, consequently every polynomial in x, y may be expressed as a linear combination of a finite number of the circle polynomials Z_n^ℓ. Moreover, the set of polynomials is complete.*

B.3 Explicit Expression for Zernike Polynomials

B.3.1 Radial Function

Let $k = \frac{1}{2}(n-m)$, then [1]

$$\begin{aligned}
\mathrm{R}_n^{\pm m}(\rho) &= \frac{1}{G_k(m+1, m+1, 1)} \cdot \rho^m \cdot G_k(m+1, m+1, \rho^2) \\
\mathrm{R}_n^m(\rho) &= (-1)^m\,\mathrm{R}_n^m(-\rho) \\
\mathrm{R}_n^{\pm m}(1) &= 1
\end{aligned} \tag{B.12}$$

For the Jacobi polynomials, we have [1]

$$\frac{\left[z-1+\sqrt{1-2z(1-2\rho^2)+z^2}\right]^m}{(2z\rho^2)^m\sqrt{1-2z(1-2\rho^2)+z^2}}=\sum_{s=0}^{\infty}\binom{m+s}{s}G_s(m+1,m+1,\rho^2)z^s. \tag{B.13}$$

Consequently, for $\rho=1$, we have

$$\frac{1}{1+z}=\sum_{s=0}^{\infty}\binom{m+s}{s}G_s(m+1,m+1,1)z^s$$

thus

$$G_s(m+1,m+1,1)=\frac{(-1)^s}{\binom{m+s}{s}}=(-1)^s\frac{m!s!}{(m+s)!} \tag{B.14}$$

Substitute (B.13) and (B.14) into (B.12), we can obtain the explicit expression for the radial polynomials $\mathrm{R}_n^{\pm m}(\rho)$ by (**??**). This important result is:

$$\begin{aligned}\mathrm{R}_n^{\pm m}(\rho)&=\frac{1}{(\frac{n-m}{2})!\rho^m}\left[\frac{\mathrm{d}}{\mathrm{d}(\rho^2)}\right]^{\frac{n-m}{2}}\left[(\rho^2)^{\frac{n+m}{2}}(\rho^2-1)^{\frac{n-m}{2}}\right]\\&=\sum_{s=0}^{\frac{1}{2}(n-m)}(-1)^s\cdot\frac{(n-s)!}{s!\,(\frac{n+m}{2}-s)!\,(\frac{n-m}{2}-s)!}\cdot\rho^{n-2s}\\&=\sum_{s=0}^{k}(-1)^s\cdot\binom{n-s}{s,k-s,n-k-s}\cdot\rho^{n-2s},\quad k=\frac{1}{2}(n-m)\end{aligned} \tag{B.15}$$

When programming with computer language such as C/C++ or MATLAB, we should to avoid such a case that the factorial $r!$ is very large. We can use the following identity:

$$\begin{aligned}\frac{(n-s)!}{s!(k-s)!(n-k-s)!}&=\binom{n-s}{s,k-s,n-k-s}\\&=\binom{n-s}{s}\binom{n-2s}{k}=\frac{(n-s)_{[s]}}{s!}\cdot\frac{(n-2s)_{[k]}}{k!}\\&=\binom{n-s}{k-s}\binom{n-k}{s}=\frac{(n-s)_{[k-s]}}{(k-s)!}\cdot\frac{(n-k)_{[s]}}{s!}\\&=\binom{n-s}{n-k-s}\binom{k}{s}=\frac{(n-s)_{[n-k-s]}}{(n-k-s)!}\cdot\frac{k_{[s]}}{s!}\end{aligned}$$

B.3.2 Zernike Polynomials

Given the expression for $\mathrm{R}_n^{\pm m}(\rho)$, the Zernike polynomials can be written by

$$\begin{aligned}\mathrm{Z}_n^{\pm m}(\rho,\theta)&=\mathrm{R}_n^{\pm m}(\rho)\mathrm{e}^{\pm\mathrm{i}m\theta}\\&=\mathrm{e}^{\pm\mathrm{i}m\theta}\cdot\sum_{s=0}^{k}(-1)^s\cdot\binom{n-s}{s,k-s,n-k-s}\cdot\rho^{n-2s},\quad k=\frac{1}{2}(n-m)\end{aligned} \tag{B.16}$$

B.3.3 Orthogonal Property

For radial polynomials, we have

$$\begin{aligned}\langle\mathrm{R}_n^{\ell}\,|\,\mathrm{R}_{n'}^{\ell}\rangle&=\int_0^1\mathrm{R}_n^{\ell}(\rho)\,\mathrm{R}_{n'}^{\ell}\,\rho\,\mathrm{d}\,\rho\\&=\frac{1}{2(n+1)}\delta_{nn'}\end{aligned} \tag{B.17}$$

For Zernike polynomials, we have

$$\boxed{\begin{aligned}\langle \mathrm{Z}_n^\ell \,|\, \mathrm{Z}_{n'}^{\ell'} \rangle &= \int_{x^2+y^2\le 1} \overline{\mathrm{Z}_n^\ell}(\rho,\theta)\, \mathrm{Z}_{n'}^{\ell'}(\rho,\theta)\rho\,\mathrm{d}\,\rho\,\mathrm{d}\,\theta \\ &= \int_{x^2+y^2\le 1} \mathrm{R}_n^\ell(\rho)\mathrm{e}^{-\mathrm{i}\ell\theta}\,\mathrm{R}_{n'}^{\ell'}\,\mathrm{e}^{\mathrm{i}\ell'\theta}\rho\,\mathrm{d}\,\rho\,\mathrm{d}\,\theta \\ &= \int_0^{2\pi} \mathrm{e}^{\mathrm{i}(\ell'-\ell)\theta}\,\mathrm{d}\,\theta \cdot \int_0^1 \mathrm{R}_n^\ell(\rho)\,\mathrm{R}_{n'}^{\ell}\,\rho\,\mathrm{d}\,\rho \\ &= 2\pi\delta_{\ell\ell'}\cdot\frac{1}{2(n+1)}\delta_{nn'} \\ &= \frac{\pi}{n+1}\cdot\delta_{nn'}\delta_{\ell\ell'}\end{aligned}} \tag{B.18}$$

B.3.4 Relation with Legendre Polynomials

The radial polynomials have the generating function

$$\frac{\left[1+z-\sqrt{1-2z(1-2\rho^2)+z^2}\right]^m}{(2z\rho)^m\sqrt{1-2z(1-2\rho^2)+z^2}} = \sum_{s=0}^{\infty} z^s\,\mathrm{R}_{m+2s}^{\pm m}(\rho) \tag{B.19}$$

When $m=0$, the left-hand side reduces to the generating function for the Legendre polynomials of argument $2\rho^2-1$, so that

$$\boxed{\mathrm{R}_{2n}^0(\rho) = P_n(2\rho^2-1).} \tag{B.20}$$

The recurrence relation of Legendre function gives that

$$\begin{aligned}\mathrm{R}_{2s}^0(\rho)\,\mathrm{R}_4^0(\rho) &= a_s\,\mathrm{R}_{2s+4}^0(\rho) + b_s\,\mathrm{R}_{2s}^0(\rho) + c_s\,\mathrm{R}_{2s-4}^0(\rho) \\ a_s &= \frac{3}{2}\cdot\frac{(s+2)(s+1)}{(2s+3)(2s+1)} \\ b_s &= \frac{(s+1)s}{(2s+3)(2s-1)} \\ c_s &= \frac{3}{2}\cdot\frac{s(s-1)}{(2s+1)(2s-1)}\end{aligned}$$

B.3.5 Relation with Bessel Function

$$\int_0^1 \mathrm{R}_n^m(\rho)J_m(v\rho)\rho\,\mathrm{d}\,\rho = (-1)^{\frac{n-m}{2}}\cdot\frac{J_{n+1}(v)}{v} \tag{B.21}$$

where J_n is a Bessel function of the first kind with order n [1][4].

B.3.6 Real Zernike Polynomials

For the complex function $\mathrm{Z}_n^{\pm m}(\rho,\theta) = \mathrm{R}_n^m(\rho)\mathrm{e}^{\pm\mathrm{i}m\theta}$, we can take its real part and imaginary part as follows [1]

$$\begin{aligned}U_n^m &= \Re(\mathrm{Z}_n^m(\rho,\theta)) = \mathrm{R}_n^m(\rho)\cos m\theta \\ U_n^{-m} &= \Im(\mathrm{Z}_n^m(\rho,\theta)) = \mathrm{R}_n^m(\rho)\sin m\theta\end{aligned} \tag{B.22}$$

U_n^m is important because the wave distortions are symmetrical about the meridional plane $\theta=0$ and consequently the aberration function is an even function of θ.

B.3.7 MATLAB Code for $\mathrm{R}_n^m(\rho)$ and $\mathrm{Z}_n^m(\rho,\theta)$

MATLAB procedures

- computing $\mathrm{R}_n^m(\rho)$: See `zhy_optik_ZernikePolRnm`
- information about $\mathrm{Z}_n^m(\rho,\theta)$: See `zhy_optik_GntZernikeFun`
- listing the Zernike functions on the screen : See `demo_ZernikeDisplay`
- printing the $\mathrm{Z}_n^m(\rho,\theta)$ into a TXT file: See `zhy_optik_ZernikePrint2Txt`
- computing $\mathrm{Z}_n^m(\rho,\theta)$: See `zhy_optik_ZernikeZnmBPO`, `zhy_optik_ZernikeZnmZMX`

B.4 Zernike Functions in ZEMAX

B.4.1 Expressions of Zernike Functions

In ZEMAX, the Zernike functions are denoted as [3][5]

$$\begin{aligned}
\mathrm{Z}_j(\rho,\theta) &= \begin{cases} \sqrt{2n+2}\cdot \mathrm{R}_n^{|m|}(\rho)\cdot\cos(|m|\theta), & |m|\neq 0, j \text{ is even;} \\ \sqrt{2n+2}\cdot \mathrm{R}_n^{|m|}(\rho)\cdot\sin(|m|\theta), & |m|\neq 0, j \text{ is odd;} \\ \sqrt{n+1}\cdot \mathrm{R}_n^{|m|}(\rho), & m=0. \end{cases} \\
\langle \mathrm{Z}_j \,|\, \mathrm{Z}_k\rangle &= \int_{\|\boldsymbol{x}\|\le 1} \overline{\mathrm{Z}}_j(\boldsymbol{x})\,\mathrm{Z}_k(\boldsymbol{x})\,\mathrm{d}^2\,\boldsymbol{x}, \quad \boldsymbol{x}=(x,y)=(\rho\cos\theta,\rho\sin\theta) \\
&= \int_0^1\int_0^{2\pi} \overline{\mathrm{Z}}_j(\rho,\theta)\,\mathrm{Z}_k(\rho,\theta)\rho\,\mathrm{d}\,\rho\,\mathrm{d}\,\theta \\
&= \pi\delta_{jk}
\end{aligned} \tag{B.23}$$

Here the ZEMAX order j is determined by the pair $< n, m >$.

$$j(n,m) = \frac{n(n+1)}{2} + r \tag{B.24}$$

where r is the index of $v_r^n = m$ and the sequence $v^n = \{v_r^n | r = 1, 2, \cdots, n+1\}$ is defined by the following formula:

$$v^n = \begin{cases} \langle 0, -2, 2, -4, 4, \cdots, -n, n\rangle, & \text{for } n \text{ is even;} \\ \langle -1, 1, -3, 3, \cdots, -n, n\rangle, & \text{for } n \text{ is odd.} \end{cases} \tag{B.25}$$

Once the sequence v^n is generated, r can be determined by a searching algorithm. With pseudocode, we have

$$r = SearchAlgorithm(v^n, v_r^n \text{ EQ } m) \tag{B.26}$$

The MATLAB procedure for determining j from $< n, m >$ is `zhy_optik_ZernikeOrdNM2J`.

We now take a look at the other MATLAB procedures listed above.

- `zhy_optik_GntZernikeFun` generates the expression for the Zernike function which contains the radial function and angular function with given pair $< n, m >$. Furthermore, both the notations in Born's Principles of Optics (BPO) and ZEMAX software are supported. For example, for $n = 4, m = 2$, we have:

```
>> C = zhy_optik_GntZernikeFun(4,2);
>> disp(C)
     PowerOrder: [4 2]
     PowerCoeff: [4 -3]
```

```
RadialFunc: '4 * r.^4 - 3 * r.^2'
NormaCoeff: '10^(1/2)'
AngFunBorn: 'exp(i*2*A)'
ZemaxOrder: 13
ZemaxFormu: '(4 * r.^4 - 3 * r.^2)*sin(2*A)'
```

For $n = 5, m = -3$, we have:

```
>> C = zhy_optik_GntZernikeFun(5,-3);
>> disp(C)
     PowerOrder: [5 3]
     PowerCoeff: [5 -4]
     RadialFunc: '5 * r.^5 - 4 * r.^3'
     NormaCoeff: '12^(1/2)'
     AngFunBorn: 'exp(-i*3*A)'
     ZemaxOrder: 18
     ZemaxFormu: '(5 * r.^5 - 4 * r.^3)*cos(3*A)'
```

The symbol expression `ZemaxFormu` is the same as that in ZEMAX.

- `zhy_optik_ZernikeZnmZMX` computes the value of Zernike function $\mathrm{Z}_j(\rho,\theta)$ with ZEMAX notation where $j = j(n,m)$ is determined by `zhy_optik_ZernikeOrdNM2J`.
- `zhy_optik_ZernikeZnmBPO` computes the value of Zernike function $\mathrm{Z}_n^m(\rho,\theta)$ with BPO notation.
- `zhy_optik_ZernikePolRnm` computes the value of Zernike radial polynomials $\mathrm{R}_n^m(\rho,\theta)$.

B.4.2 Physical Interpretations and Aberrations

The parameters j, n, m and $k = \frac{n-|m|}{2}$ are important for classifying the aberrations in optical design.

(1) j distinguishes the x and y directions.

- j is even $\longleftrightarrow$ x direction
- j is odd $\longleftrightarrow$ y direction

(2) n determines the maximum order of $\mathrm{R}_n^m(\rho)$ and the form and symmetry of the corresponding surface are effected greatly. When n is odd, the spherical aberrations disappears.

- $n = 0, \mathrm{Z}_0^0(\rho,\theta) \equiv 1$
- $n = 1$, linear function. There is no aberration.
- $n = 2, m = 0$. Power
- $n^2 + m^2 > 4$. Nonlinear aberrations!

(3) k names the nonlinear order of aberrations.

$$k = \frac{n-|m|}{2} = \left|\left\{\rho^{|m|}, \rho^{|m|+2}, \cdots, \rho^{n-2}, \rho^n\right\}\right|$$

Noth that

- Spherical aberration: $k \geq 2$ since $m = 0, n \in \{4, 6, 8, \cdots\}$.
- Comma aberration: $k \geq 1$ since $m = \pm 1, n \in \{3, 5, 7, \cdots\}$
- Astigmatism aberration: $k \geq 0$ since $m = \pm 2, n \in \{2, 4, 6, \cdots\}$
- Trefoil aberration: $k \geq 0$ since $m = \pm 3, n \in \{3, 5, 7, \cdots\}$
- Tetrafoil aberration: $k \geq 0$ since $m = \pm 4, n \in \{4, 6, 8, \cdots\}$

How to name the aberration: OrdinalNumAdj + AberrType

- First/Primary: $k = k_0, k_0 \in \{0, 1, 2\}$
- Secondary: $k = k_0 + 1$
- Tertiary: $k = k_0 + 2$
- Quatenary: $k = k_0 + 3$
- $\cdots$

(4) m distinguishes the aberrations $\longleftrightarrow \left\{\rho^{|m|}, \rho^{|m|+2}, \cdots, \rho^{n-2}, \rho^n\right\}$

- $m = 0$: Spherical, see MATLAB procedure `demo_AbberSpherical`
  - $k = 2$: Primary Spherical
  - $k = 3$: Secondary Spherical
  - $k = 4$: Tertiary Spherical
  - $k = 5$: Quatenary Spherical
- $|m| = 1$: Comma, see MATLAB procedure `demo_AbberComma`
  - $k = 1$: Comma
  - $k = 2$: Secondary Comma
  - $k = 3$: Tertiary Comma
  - $k = 4$: Quatenary Comma
- $|m| = 2$: Astigmatism, see MATLAB procedure `demo_AbberAstigmatism`
  - $k = 0$: Astigmatism
  - $k = 1$: Secondary Astigmatism
  - $k = 2$: Tertiary Astigmatism
  - $k = 3$: Quatenary Astigmatism
- $|m| = 3$ Trefoil, see MATLAB procedure `demo_AbberTrefoil`
  - $k = 0$: Trefoil
  - $k = 1$: Secondary Trefoil
  - $k = 2$: Tertiary Trefoil
  - $k = 3$: Quatenary Trefoil
- $|m| = 4$ Tetrafoil, see MATLAB procedure `demo_AbberTetrafoil`
  - $k = 0$: Tetrafoil
  - $k = 1$: Secondary Tetrafoil
  - $k = 2$: Tertiary Tetrafoil
  - $k = 3$: Quatenary Tetrafoil

B.4.3 Conversion between $< n, m >$ and j.

Problem 293 (NM2J). *Given the pair $< n, m >$, how to determine j?*

MATLAB procedure for computing $j = j(n, m)$: See `zhy_optik_ZernikeOrdNM2J`

Problem 294 (J2NM). *Given the ZEMAX order j, how to determine the pair $< n, m >$?*

In practice, a simple way to cope with problem is by the Look-Up-Table approach. However, we can find a close formula such that $< n, m >= F(j)$ since the mapping

$$\psi : \ j \mapsto < n, m >$$

is a bijection. By (B.24), we have

$$\frac{n(n+1)}{2} < j \leq \frac{n(n+1)}{2} + n + 1. \tag{B.27}$$

Consequently,

$$-1 + \frac{-1+\sqrt{8j+1}}{2} \leq n < \frac{-1+\sqrt{8j+1}}{2} \tag{B.28}$$

This implies that

$$n = \left\lceil \frac{-3+\sqrt{8j+1}}{2} \right\rceil, \tag{B.29}$$

After knowing the n, the index r in v^n can be determined by

$$r = j - \frac{n(n+1)}{2}.$$

Then m can be picked up from the sequence v^n since

$$v_r^n = m.$$

MATLAB procedure: See `zhy_optik_ZernikeOrdJ2NM`

Table B.1: Zernike Functions $\mathrm{Z}_j(\rho,\theta)$ and $\mathrm{Z}_n^m(\rho,\theta)$

j	n	m	N_n^m	Expression $\mathrm{Z}_j(\rho,\theta)$	Physical Interpretation
1	0	0	$\sqrt{1}$	1	Piston or Bias
2	1	-1	$\sqrt{4}$	$\rho\cos(\theta)$	Tilt x
3	1	1	$\sqrt{4}$	$\rho\sin(\theta)$	Tile y
4	2	0	$\sqrt{3}$	$2\rho^2-1$	Power/Defocus
5	2	-2	$\sqrt{6}$	$\rho^2\sin(2\theta)$	Astigmatism y
6	2	2	$\sqrt{6}$	$\rho^2\cos(2\theta)$	Astigmatism x
7	3	-1	$\sqrt{8}$	$(3\rho^3-2\rho)\sin(\theta)$	Comma y
8	3	1	$\sqrt{8}$	$(3\rho^3-2\rho)\cos(\theta)$	Comma x
9	3	-3	$\sqrt{8}$	$\rho^3\sin(3\theta)$	Trefoil y
10	3	3	$\sqrt{8}$	$\rho^3\cos(3\theta)$	Trefoil x
11	4	0	$\sqrt{5}$	$6\rho^4-6\rho^2+1$	Primary Spherical
12	4	-2	$\sqrt{10}$	$(4\rho^4-3\rho^2)\cos(2\theta)$	Secondary Astigmatism x
13	4	2	$\sqrt{10}$	$(4\rho^4-3\rho^2)\sin(2\theta)$	Secondary Astigmatism y
14	4	-4	$\sqrt{10}$	$\rho^4\cos(4\theta)$	Tetrafoil x
15	4	4	$\sqrt{10}$	$\rho^4\sin(4\theta)$	Tetrafoil y
16	5	-1	$\sqrt{12}$	$(10\rho^5-12\rho^3+3\rho)\cos(\theta)$	Secondary Comma x
17	5	1	$\sqrt{12}$	$(10\rho^5-12\rho^3+3\rho)\sin(\theta)$	Secondary Comma y
18	5	-3	$\sqrt{12}$	$(5\rho^5-4\rho^3)\cos(3\theta)$	Secondary Trefoil x
19	5	3	$\sqrt{12}$	$(5\rho^5-4\rho^3)\sin(3\theta)$	Secondary Trefoil y
20	5	-5	$\sqrt{12}$	$\rho^5\cos(5\theta)$	Pentafoil x
21	5	5	$\sqrt{12}$	$\rho^5\sin(5\theta)$	Pentafoil y
22	6	0	$\sqrt{7}$	$20\rho^6-30\rho^4+12\rho^2-1$	Secondary Spherical
23	6	-2	$\sqrt{14}$	$(15\rho^6-20\rho^4+6\rho^2)\sin(2\theta)$	Tertiary Astigmatism y
24	6	2	$\sqrt{14}$	$(15\rho^6-20\rho^4+6\rho^2)\cos(2\theta)$	Tertiary Astigmatism x
25	6	-4	$\sqrt{14}$	$(6\rho^6-5\rho^4)\sin(4\theta)$	Secondary Tetrafoil y
26	6	4	$\sqrt{14}$	$(6\rho^6-5\rho^4)\cos(4\theta)$	Secondary Tetrafoil x
27	6	-6	$\sqrt{14}$	$\rho^6\sin(6\theta)$	
28	6	6	$\sqrt{14}$	$\rho^6\cos(6\theta)$	
29	7	-1	$\sqrt{16}$	$(35\rho^7-60\rho^5+30\rho^3-4\rho^1)\sin(\theta)$	Tertiary Coma y
30	7	1	$\sqrt{16}$	$(35\rho^7-60\rho^5+30\rho^3-4\rho^1)\cos(\theta)$	Tertiary Coma x
31	7	-3	$\sqrt{16}$	$(21\rho^7-30\rho^5+10\rho^3)\sin(3\theta)$	Tertiary Trefoil y
32	7	3	$\sqrt{16}$	$(21\rho^7-30\rho^5+10\rho^3)\cos(3\theta)$	Tertiary Trefoil x
33	7	-5	$\sqrt{16}$	$(7\rho^7-6\rho^5)\sin(5\theta)$	
34	7	5	$\sqrt{16}$	$(7\rho^7-6\rho^5)\cos(5\theta)$	
35	7	-7	$\sqrt{16}$	$\rho^7\sin(7\theta)$	
36	7	7	$\sqrt{16}$	$\rho^7\cos(7\theta)$	
37	8	0	$\sqrt{9}$	$70\rho^8-140\rho^6+90\rho^4-20\rho^2+1$	Tertiary Spherical
38	8	-2	$\sqrt{18}$	$(56\rho^8-105\rho^6+60\rho^4-10\rho^2)\cos(2\theta)$	Quatenary Astigmatism x
39	8	2	$\sqrt{18}$	$(56\rho^8-105\rho^6+60\rho^4-10\rho^2)\sin(2\theta)$	Quatenary Astigmatism y
40	8	-4	$\sqrt{18}$	$(28\rho^8-42\rho^6+15\rho^4)\cos(4\theta)$	Tertiary Tetrafoil

Bibliography

[1] Max Born, Emil Wolf, *Principles of Optics*, 7th edition, London, Cambridge University Press, 1999

[2] Daniel Malacara, Optical Shop Testing,Wiley-Interscience, A John Wiley & Sons, Inc. Publication, 2007

[3] ZEMAX Development Corporation, *Zemax Optical Design Program User's Guide*, 2008, `www.zemax.com`

[4] A. J. E. M. Janssen, *Zernike circle polynomials and infinite integrals involving the product of Bessel functions*, available online: `http://arXiv:1007.0667v1` [math-ph], Jun 5, 2010.

[5] Richard J. Mathar, *Zernike Basis to Cartesian Transformations*, available online: `http://arXiv:0809.2368v1` [math-ph], Sept. 13, 2008

[6] Donald E. Knuth, *The Art of Computer Programming: Fundamental Algorithms*, Vol. 1, Addison-Wesley Professional, 3rd edition, 1997

[7] Ronald L. Graham, Donald E. Knuth, Oren Patashnik, Concrete Mathematics: A Foundation for Computer Science, 2nd edition, Addison Wesley, 1994

[8] Richard Courant and David Hilbert, *Methods of Mathematical Physics*, Vol.1, Wiley Classics Edition Published in 1989

Appendix C

Fourier Transformation in Signals Analysis

C.1 Definition

The Fourier transform of a function/signal $g(t)$ is defined as

$$G(f) = \int_{-\infty}^{\infty} g(t)\mathrm{e}^{-\mathrm{j}2\pi ft}\,\mathrm{d}\,t \longleftrightarrow g(t) = \int_{-\infty}^{\infty} G(f)\mathrm{e}^{+\mathrm{j}2\pi ft}\,\mathrm{d}\,f \tag{C.1}$$

or

$$G(\omega) = \int_{-\infty}^{\infty} g(t)\mathrm{e}^{-\mathrm{j}\omega t}\,\mathrm{d}\,t \longleftrightarrow g(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} G(\omega)\mathrm{e}^{+\mathrm{j}\omega t}\,\mathrm{d}\,\omega \tag{C.2}$$

C.2 Properties

Table C.1: Summary of properties of the Fourier transformation

No.	Property	Mathematical Description
1	Linearity	$ag_1(t) + bg_2(t) \longleftrightarrow aG_1(f) + bG_2(f)$
2	Time scaling	$g(at) \longleftrightarrow \frac{1}{\lvert a\rvert}G(f/a)$
3	Duality	If $g(t) \longleftrightarrow G(f)$, then $G(t) \longleftrightarrow g(-f)$
4	Time shifting	$g(t - t_0) \longleftrightarrow G(f)\exp(-\mathrm{j}2\pi ft_0)$
5	Frequency shifting	$\exp(\mathrm{j}2\pi f_0 t)g(t) \longleftrightarrow G(f - f_0)$
6	Area under $g(t)$	$\int_{-\infty}^{\infty} g(t)\,\mathrm{d}\,t = G(0)$
7	Area under $G(f)$	$g(0) = \int_{-\infty}^{\infty} G(f)\,\mathrm{d}\,f$
8	Differentiation in the time domain	$\frac{\mathrm{d}\,g(t)}{\mathrm{d}\,t} \longleftrightarrow \mathrm{j}2\pi fG(f)$
9	Integration in the time domain	$\int_{-\infty}^{t} g(\tau)\,\mathrm{d}\,\tau \longleftrightarrow \frac{1}{\mathrm{j}2\pi f}G(f) + \frac{G(0)}{2}\delta(f)$
10	Conjugate functions	If $g(t) \longleftrightarrow G(f)$, then $g^*(t) \longleftrightarrow G^*(-f)$
11	Multiplication in the time domain	$g_1(t)g_2(t) \longleftrightarrow G_1(f) \star G_2(f)$
12	Convolution in the time domain	$g_1(t) \star g_2(t) \longleftrightarrow G_1(f)G_2(f)$

C.3 Fourier-transform pairs

We now define the rectangular impulse of unit amplitude and unit duration centered on the origin as follows [1]

$$\text{rect}(t) = \begin{cases} 1, & |t| \leq \frac{1}{2}; \\ 0, & |t| > \frac{1}{2}. \end{cases} \tag{C.3}$$

The Fourier transform of $\text{rect}(t)$ is a sinc funciton which is defined as

$$\text{sinc}(f) = \frac{\sin(\pi f)}{\pi f}. \tag{C.4}$$

A general rectangular impulse with amplitude A and duration T which is centered at a is

$$\text{rect}\left(\frac{t-t_0}{T}\right) = \begin{cases} A, & |t-t_0| \leq \frac{T}{2}; \\ 0, & |t-t_0| > \frac{T}{2}. \end{cases} \tag{C.5}$$

Furthermore,

$$\text{rect}\left(\frac{t}{T}\right) \star \text{rect}\left(\frac{t}{T}\right) = \begin{cases} 1-\frac{t}{T}, & |t| \leq T; \\ 0, & |t| > T. \end{cases} \tag{C.6}$$

is a symmetric triangle impulse with duration $2T$ and centered at zero.

Table C.2: Fourier-transform pairs

No.	Name	Time function $g(t)$	Fourier Transform $G(f)$
1	rectangular impulse	$\text{rect}\left(\frac{t}{T}\right)$	$T\,\text{sinc}(fT)$
2	sine counting function	$\text{sinc}(2Wt)$	$\frac{1}{2W}\,\text{rect}(\frac{f}{2W})$
3	single-side exponent	$\exp(-at)U(t), \quad a>0$	$\frac{1}{a+\text{j}2\pi f}$
4	double-side exponent	$\exp(-a\,\lvert t\rvert), \quad a>0$	$\frac{2a}{a^2+(2\pi f)^2}$
5	Gaussian impulse	$\exp(-\pi t^2)$	$\exp(-\pi f^2)$
6	triangle impulse	$\text{rect}\left(\frac{t}{T}\right) \star \text{rect}\left(\frac{t}{T}\right)$	$T\,\text{sinc}^2(fT)$
7	Dirac-δ function (ideal impulse)	$\delta(t)$	1
8	unit direct current	1	$\delta(f)$
9	ideal delayed impulse	$\delta(t-t_0)$	$\exp(-\text{j}2\pi f t_0)$
10	phase shifting function	$\exp(\text{j}2\pi f_c t)$	$\delta(f-f_c)$
11	cosine	$\cos(2\pi f_c t)$	$\frac{1}{2}[\delta(f-f_c)+\delta(f+f_c)]$
12	sine	$\sin(2\pi f_c t)$	$\frac{1}{2\text{j}}[\delta(f-f_c)-\delta(f+f_c)]$
13	sign function	$\text{sign}(t)$	$\frac{1}{\text{j}\pi f}$
14	Hilbert transform kernel	$\frac{1}{\pi t}$	$-\text{j}\,\text{sign}(f)$
15	unit step function	$U(t)$	$\frac{1}{2}\delta(f)+\frac{1}{\text{j}2\pi f}$
16	Dirac comb	$\sum_{k=-\infty}^{\infty}\delta(t-kT)$	$\frac{1}{T}\sum_{n=-\infty}^{\infty}\delta\left(f-\frac{n}{T}\right)$
17	Modulation	$g(t)\cos(2\pi f_c t)$	$\frac{1}{2}[G(f-f_c)+G(f+f_c)]$

[1] Sometimes, the notation for $\text{rect}(t)$ may be replaced by $\Pi(t)$.