Singular value decomposition of matrices and compact linear operators
Lecture notes, Fall 2026
(Version: June 17, 2026)

Pu-Zhao Kow

DEPARTMENT OF MATHEMATICAL SCIENCES, NATIONAL CHENGCHI UNIVERSITY
Email address: pzkow@g.nccu.edu.tw

Preface

We will use this lecture note as the main teaching material for the course *From linear algebra to functional analysis* (701853001, 751786001), during Fall 2026 (115-1). The lecture note may updated during the course. I was inspired by the monograph [**Tre26**], which is intended for a student who, while not yet very familiar with abstract reasoning. I would like to point out that, there is another linear algebra textbook [**Axl24**] that is written by or for an algebraist, which is not the target audience for this course. For future reference, the reader may also consult [**Ber09**], which contains a comprehensive collection of results, identities, and formulas from linear algebra. For the functional analysis, I will mainly follow my existing lecture note [**Kow25**], which was inspired by Mikko Salo's lecture note. We will also refer some facts in the monograph [**Bre11**]. In this course, we will mainly focus on compact linear operator, which behaves similar to matrices.

Title. From linear algebra to functional analysis (Fall 2026, 3 credits)

Lectures. Thursday (13:10–16:00), September 7, 2026 ~ December 25, 2026.

Language. Chinese and English. Materials will be prepared in English.

Instructor. Pu-Zhao Kow (Email: `pzkow@g.nccu.edu.tw`)

Office hour. Thursday (16:10–17:00)

Completion. Homework Assignments 60%, Midterm exam 20%, Final exam 20%

Contents

Chapter 1. Representation of elements by linear combinations 1
1.1. Finite dimensional case 1
1.2. Infinite dimensional case (strong topology) 7
1.3. An example of orthonormal basis: Fourier series 16
1.4. Additional properties of Fourier series 23
1.5. Infinite dimensional subspace: Sobolev embedding theorems 26

Chapter 2. Characteristics of linear operators: Singular value decomposition 37
2.1. Determinants 37
2.2. Spectral theorem for general matrices 38
2.3. Schmidt decomposition and singular values: finite dimensional case 43
2.4. Schmidt decomposition and singular values for compact linear operators 45

Chapter 3. Linearization 50
3.1. Limits and continuity 50
3.2. Linearization of functions and Landweber iteration: finite dimensional case 51
3.3. Basic Properties of Derivatives 54
3.4. Linearization of operators and Landweber iteration: infinite dimensional case 57

Bibliography 59

CHAPTER 1

Representation of elements by linear combinations

1.1. Finite dimensional case

In practice, it is more convenient to work with the complex number $\mathbb{C}$ rather than the real number $\mathbb{R}$. We will go through linear algebra on $\mathbb{C}$. We denote the space $\mathbb{C}^m$ consists of all columns of size m, that is,

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{pmatrix}$$

whose entries are complex numbers $\mathbb{C}$. Addition and multiplication are defined entrywise:

$$\alpha \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{pmatrix} + \beta \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{pmatrix} = \begin{pmatrix} \alpha v_1 + \beta w_1 \\ \alpha v_2 + \beta w_2 \\ \vdots \\ \alpha v_m + \beta w_m \end{pmatrix}$$

for all $\alpha, \beta, v_1, \cdots, v_m, w_1, \cdots, w_m \in \mathbb{C}$.

We denote the space $\mathbb{C}^{m\times n}$ consists of all $m \times n$ matrices with entries in $\mathbb{C}$. In particular, we identify $\mathbb{C}^{m\times 1} \cong \mathbb{C}^m$. Each $A \in \mathbb{C}^{m\times n}$ can be denoted as

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix},$$

of we slightly abuse the notation by writing $A = (A_{ij})$. Similarly, addition and multiplication are defined entrywise:

$$\alpha \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix} + \beta \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1n} \\ B_{21} & B_{22} & \cdots & B_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ B_{m1} & B_{m2} & \cdots & B_{mn} \end{pmatrix} = \begin{pmatrix} \alpha A_{11} + \beta B_{11} & \alpha A_{12} + \beta B_{12} & \cdots & \alpha A_{1n} + \beta B_{1n} \\ \alpha A_{21} + \beta B_{21} & \alpha A_{22} + \beta B_{22} & \cdots & \alpha A_{2n} + \beta B_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha A_{m1} + \beta B_{m1} & \alpha A_{m2} + \beta B_{m2} & \cdots & \alpha A_{mn} + \beta B_{mn} \end{pmatrix}$$

for all $\alpha, \beta, A_{11}, \cdots, A_{mn}, B_{11}, \cdots, B_{mn} \in \mathbb{C}$.

Given $A \in \mathbb{C}^{m\times n}$, its transpose $A^{\intercal} \in \mathbb{C}^{n\times m}$ and its adjoint (also known as conjugate transpose or "Hermitian") $A^* \in \mathbb{C}^{n\times m}$ are defined by

$$(A^{\intercal})_{ji} = A_{ij} \quad (A^*)_{ji} = \overline{A_{ij}} \quad \text{for all } i = 1,\cdots,m \text{ and } j = 1,\cdots,n,$$

where $\bar{z} \in \mathbb{C}$ is the conjugate of $z \in \mathbb{C}$.

A $\mathbb{C}$-linear combination of $A^{(1)},\cdots,A^{(p)} \in \mathbb{C}^{m\times n}$ is a sum of form

$$\sum_{k=1}^{p} \alpha_k A^{(k)} := \alpha_1 A^{(1)} + \cdots + \alpha_p A^{(p)}.$$

DEFINITION 1.1.1. A system of vectors $\{A^{(1)},\cdots,A^{(p)}\} \subset \mathbb{C}^{m\times n}$ is said to be *linear independent*, if

$$\sum_{k=1}^{p} \alpha_k A^{(k)} = 0 \quad \text{if and only if } \alpha_1 = \cdots = \alpha_p = 0.$$

We denote the p-dimensional subspace

$$\operatorname{span}\{A^{(1)},\cdots,A^{(p)}\} := \left\{ \sum_{k=1}^{p} \alpha_k A^{(k)} : \alpha_1,\cdots,\alpha_k \in \mathbb{C} \right\},$$

called the span of $\{A^{(1)},\cdots,A^{(p)}\} \subset \mathbb{C}^{m\times n}$. A system of vectors $\{A^{(1)},\cdots,A^{(p)}\} \subset \mathbb{C}^{m\times n}$ is called a *basis*, if any $A \in \mathbb{C}^{m\times n}$ admits a *unique* representation

$$A = \sum_{k=1}^{p} \alpha_k A^{(k)}.$$

EXAMPLE 1.1.2. For each $i = 1,\cdots,m$ and $j = 1,\cdots,n$, we define $\mathbf{e}_{ij} \in \mathbb{C}^{m\times n}$ by

$$(\mathbf{e}_{ij})_{k\ell} = \begin{cases} 1 & \text{if } k = i \text{ and } \ell = j, \\ 0 & \text{otherwise.} \end{cases}$$

It is straightforward to check that $\{\mathbf{e}_{ij} : i = 1,\cdots,m, j = 1,\cdots,n\}$ forms a basis of $\mathbb{C}^{m\times n}$.

EXERCISE 1.1.3. Show that, if $\{A^{(1)},\cdots,A^{(p)}\}$ forms a basis of $\mathbb{C}^{m\times n}$, then $\{A^{(1)},\cdots,A^{(p)}\}$ is linear independent and $p = mn$.

Next, we can talk about the geometry of $\mathbb{C}^{m\times n}$. We define the scalar product $(\cdot,\cdot) \equiv (\cdot,\cdot)_{\mathbb{C}^{m\times n}}$ on $\mathbb{C}^{m\times n}$ by

$$(A,B) \equiv (A,B)_{\mathbb{C}^{m\times n}} := \sum_{j=1}^{n} \sum_{i=1}^{m} A_{ij} \overline{B_{ij}} \quad \text{for all } A,B \in \mathbb{C}^{m\times n}, \tag{1.1.1}$$

and the corresponding Euclidean norm $\|\cdot\| \equiv \|\cdot\|_{\mathbb{C}^{m\times n}}$ (aka Frobenius norm) is given by

$$\|A\| \equiv \|A\|_{\mathbb{C}^{m\times n}} = \sqrt{(A,A)_{\mathbb{C}^{m\times n}}} \quad \text{for all } A \in \mathbb{C}^{m\times n}. \tag{1.1.2}$$

Each nonzero matrix A can be written in terms of *polar coordinate* $A = \|A\|\hat{A}$ for some $\hat{A}$ with $\|\hat{A}\| = 1$. We say that A is orthogonal to B, denoted by $A \perp B$, if $(A,B)_{\mathbb{C}^{m\times n}} = 0$. Now it is natural to introduce the following definition:

DEFINITION 1.1.4. Given a basis $\{A^{(1)},\cdots,A^{(mn)}\}$ of $\mathbb{C}^{m\times n}$.

(1) It is said to be *orthogonal* if $(A^{(k)},A^{(\ell)})_{\mathbb{C}^{m\times n}} = 0$ for all $k \neq \ell$.
(2) It is said to be *orthonormal* if $(A^{(k)},A^{(\ell)})_{\mathbb{C}^{m\times n}} = \delta_{k\ell}$ for all $k,\ell \in \{1.\cdots,mn\}$, where $\delta_{k\ell}$ is the Kronecker delta.

EXERCISE 1.1.5. Show that, if $\{A^{(1)},\cdots,A^{(mn)}\} \subset \mathbb{C}^{m\times n}$ is linear independent, then it forms a basis of $\mathbb{C}^{m\times n}$. (**Hint.** Let $A,B \in \mathbb{C}^{m\times n}$, show that $A-(A,\hat{B})\hat{B}$ is orthogonal to B with $\hat{B} := B/\|B\|_{\mathbb{C}^{m\times n}}$)

EXERCISE 1.1.6 (Gram-Schmidt). Given a basis of $\{A^{(1)},\cdots,A^{(mn)}\} \subset \mathbb{C}^{m\times n}$, construct an orthonormal basis of $\mathbb{C}^{m\times n}$ accordingly. (**Hint.** Let $A,B \in \mathbb{C}^{m\times n}$, show that $A-(A,\hat{B})\hat{B}$ is orthogonal to B with $\hat{B} := B/\|B\|_{\mathbb{C}^{m\times n}}$)

We now consider the abstract mapping $T : \mathbb{C}^n \to \mathbb{C}^m$. A naive method to study the mapping T is simply test it by using all $\mathbf{v} \in \mathbb{C}^n$, which is obviously not practical. Intuitively, if T has some special structure, then we need not to test all $\mathbf{v} \in \mathbb{C}^n$. In this course, we are in particular interested in the case when $T : \mathbb{C}^n \to \mathbb{C}^m$ is linear, that is,

$$T(\alpha\mathbf{v}+\beta\mathbf{w}) = \alpha T(\mathbf{v})+\beta T(\mathbf{w}) \quad \text{for all } \alpha,\beta \in \mathbb{C} \text{ and for all } \mathbf{v},\mathbf{w} \in \mathbb{C}^n.$$

Let's say we have a basis $\{\mathbf{v}^{(1)},\cdots,\mathbf{v}^{(n)}\}$ of $\mathbb{C}^n$, then each $\mathbf{v} \in \mathbb{C}^n$ can be uniquely represented as

$$\mathbf{v} = \sum_{j=1}^{n} c_j \mathbf{v}^{(j)},$$

and therefore

$$T(\mathbf{v}) = \sum_{j=1}^{n} c_j T(\mathbf{v}^{(j)}). \tag{1.1.3}$$

In other words, it is suffice to measure $T(\mathbf{v}^{(1)}),\cdots,T(\mathbf{v}^{(n)})$, and thus for each $\mathbf{v} \in \mathbb{C}^n$ the result $T(\mathbf{v})$ can be reconstructed by the formula (1.1.3).

We first measure $T(\mathbf{v}^{(1)})$. Fix any basis $\{\tilde{\mathbf{v}}^{(1)},\cdots,\tilde{\mathbf{v}}^{(m)}\} \subset \mathbb{C}^m$, the measurement $T(\mathbf{v}^{(1)})$ can be uniquely represented as

$$T(\mathbf{v}^{(1)}) = \sum_{i=1}^{m} A_{i1}\tilde{\mathbf{v}}^{(i)}.$$

Similarly, after measuring $T(\mathbf{v}^{(j)})$, since $\{\tilde{\mathbf{v}}^{(1)},\cdots,\tilde{\mathbf{v}}^{(m)}\} \subset \mathbb{C}^m$ is a basis, then the measurement $T(\mathbf{v}^{(j)})$ can be uniquely represented as

$$T(\mathbf{v}^{(j)}) = \sum_{i=1}^{m} A_{ij}\tilde{\mathbf{v}}^{(i)}. \tag{1.1.4}$$

Note that the above procedure is valid for all $j = 1, \cdots, n$, each of them produce a unique set of complex numbers $\{A_{1j}, \cdots, A_{mj}\}$. We now combine (1.1.3) and (1.1.4) to see that

$$T(\mathbf{v}) = \sum_{j=1}^{n} c_j T(\mathbf{v}^{(j)}) = \sum_{j=1}^{n} c_j \sum_{i=1}^{m} A_{ij} \tilde{\mathbf{v}}^{(i)} = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} A_{ij} c_j \right) \tilde{\mathbf{v}}^{(i)}. \tag{1.1.5}$$

In other words, each T can be uniquely represented by a matrix

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix},$$

and this representation is independent of the choice of $\mathbf{v}$ (i.e., $c_1, \cdots, c_n$). Encoding $c_1, \cdots, c_n$ using the vector $\mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}$, we simply denote the vector $A\mathbf{c} \equiv \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}$ by the *matrix multiplication* given by

$$(A\mathbf{c})_i := \sum_{j=1}^{n} A_{ij} c_j \quad \text{for all } i = 1, \cdots, m, \tag{1.1.6}$$

so that we can simplify the representation in (1.1.5) as

$$T(\mathbf{v}) = \sum_{i=1}^{m} (A\mathbf{c})_i \tilde{\mathbf{v}}^{(i)} \quad \text{provided } \mathbf{v} = \sum_{j=1}^{n} c_j \mathbf{v}^{(j)} \text{ and } \mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}. \tag{1.1.7}$$

We then summarized the above discussions in the following theorem:

THEOREM 1.1.7. *Fix any bases* $\{\mathbf{v}^{(1)}, \cdots, \mathbf{v}^{(n)}\} \subset \mathbb{C}^n$ *and* $\{\tilde{\mathbf{v}}^{(1)}, \cdots, \tilde{\mathbf{v}}^{(m)}\} \subset \mathbb{C}^m$. *Then there is a one-to-one correspondence between*

$$\{\textit{linear operator } T : \mathbb{C}^n \to \mathbb{C}^m\} \longleftrightarrow \mathbb{C}^{m \times n}$$

via the formula (1.1.7) *with matrix multiplication* (1.1.4).

REMARK 1.1.8. In terms of the matrix multiplication (1.1.6), it is interesting to point out that $(\mathbf{v}, \mathbf{w})_{\mathbb{C}^n} = \sum_{i=1}^{n} v_i \overline{w_i} = \mathbf{w}^* \mathbf{v}$ and $\|\mathbf{v}\|_{\mathbb{C}^n} = \sqrt{\mathbf{v}^* \mathbf{v}}$.

REMARK 1.1.9. If we fix an orthonormal basis $\{\mathbf{v}^{(1)}, \cdots, \mathbf{v}^{(n)}\} \subset \mathbb{C}^n$, then $\mathbf{v} = \sum_{j=1}^{n} c_j \mathbf{v}^{(j)}$ with $c_j = (\mathbf{v}, \mathbf{v}^{(j)})_{\mathbb{C}^n}$. If we fix an *orthonormal* basis $\{\tilde{\mathbf{v}}^{(1)}, \cdots, \tilde{\mathbf{v}}^{(m)}\} \subset \mathbb{C}^m$, then (1.1.7) can be represented as

$$\sum_{j=1}^{n} c_j \left(T(\mathbf{v}^{(j)}), \tilde{\mathbf{v}}^{(i)} \right)_{\mathbb{C}^m} = \left(T(\mathbf{v}), \tilde{\mathbf{v}}^{(i)} \right)_{\mathbb{C}^m} = (A\mathbf{c})_i = \sum_{j=1}^{n} A_{ij} c_j.$$

By arbitrariness of $c_1, \cdots, c_n \in \mathbb{C}$, we then conclude that

$$A_{ij} = \left(T(\mathbf{v}^{(j)}), \tilde{\mathbf{v}}^{(i)} \right)_{\mathbb{C}^m} \quad \text{for all } i = 1, \cdots, m \text{ and for all } j = 1, \cdots, n.$$

We now consider another linear mapping $\tilde{T} : \mathbb{C}^m \to \mathbb{C}^p$ and we now study the composition $\tilde{T} \circ T : \mathbb{C}^n \to \mathbb{C}^p$ defined by

$$(\tilde{T} \circ T)(\mathbf{v}) := \tilde{T}(T(\mathbf{v})) \quad \text{for all } \mathbf{v} \in \mathbb{C}^m.$$

It is straight forward to check that $\tilde{T} \circ T : \mathbb{C}^n \to \mathbb{C}^p$ is also linear. We now choose any basis $\{\mathbf{w}^{(1)}, \cdots, \mathbf{w}^{(p)}\} \subset \mathbb{C}^p$, using (1.1.7) we see that

$$\tilde{T}(\tilde{\mathbf{v}}) = \sum_{\ell=1}^{p} (\tilde{A}\mathbf{d})_\ell \mathbf{w}^{(\ell)} \quad \text{provided } \tilde{\mathbf{v}} = \sum_{i=1}^{m} d_i \tilde{\mathbf{v}}^{(i)} \text{ and } \mathbf{d} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_m \end{pmatrix}$$

We now choose $\tilde{\mathbf{v}} = T(\mathbf{v})$ and combine with (1.1.7) to see that

$$\tilde{T}(T(\mathbf{v})) = \sum_{\ell=1}^{p} (\tilde{A}\mathbf{d})_\ell \mathbf{w}^{(\ell)} \quad \text{provided } \overbrace{\sum_{i=1}^{m} (A\mathbf{c})_i \tilde{\mathbf{v}}^{(i)} = \sum_{i=1}^{m} d_i \tilde{\mathbf{v}}^{(i)}}^{\text{if and only if } A\mathbf{c}=\mathbf{d}} \text{ and } \mathbf{v} = \sum_{j=1}^{n} c_j \mathbf{v}^{(j)} \text{ and } \mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix},$$

that is,

$$\tilde{T}(T(\mathbf{v})) = \sum_{\ell=1}^{p} (\tilde{A}(A\mathbf{c}))_\ell \mathbf{w}^{(\ell)} \quad \text{provided } \mathbf{v} = \sum_{j=1}^{n} c_j \mathbf{v}^{(j)} \text{ and } \mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}$$

Using (1.1.6) twice, we see that

$$(\tilde{A}(A\mathbf{c}))_\ell = \sum_{i=1}^{m} \tilde{A}_{\ell i} (A\mathbf{c})_i = \sum_{i=1}^{m} \tilde{A}_{\ell i} \sum_{j=1}^{n} A_{ij} c_j = \sum_{j=1}^{n} \left(\sum_{i=1}^{m} \tilde{A}_{\ell i} A_{ij} \right) c_j \quad \text{for all } \ell = 1, \cdots, p.$$

By arbitrariness of $\mathbf{c}$, this immediately suggest the following definition of matrix multiplication $\tilde{A}A \in \mathbb{C}^{p \times n}$ to make sure the above discussions are consistent:

$$(\tilde{A}A)_{\ell j} := \sum_{i=1}^{m} \tilde{A}_{\ell i} A_{ij} \quad \text{for all } \ell = 1, \cdots, p \text{ and for all } j = 1, \cdots, n. \tag{1.1.8}$$

Note that (1.1.6) is a special case of (1.1.8). What we have shown that:

THEOREM 1.1.10. *The composition of two linear operators is uniquely represented by the matrix product of their corresponding matrix representations, as defined in* (1.1.8).

COROLLARY 1.1.11. *The composition of N linear operators is uniquely represented by the matrix product of their corresponding matrix representations, as defined in*

$$(A^{(1)}A^{(2)}\cdots A^{(N)})_{ij} := \sum_{k_1,\cdots,k_{N-1}} A^{(1)}_{ik_1}A^{(2)}_{k_1k_2}A^{(3)}_{k_2k_3}\cdots A^{(N)}_{k_{N-1}j}.$$

Let $\mathbf{v}\in\mathbb{C}^m$ and $\mathbf{w}\in\mathbb{C}^n$, we define the juxtaposition (aka tensor product)

$$\mathbf{v}\otimes\mathbf{w} := \mathbf{v}\mathbf{w}^* \in \mathbb{C}^{m\times n}, \tag{1.1.9}$$

which is a very convenient notation, especially while dealing with system of PDEs such as the system of elasticity (aka Lamé system). In particular, if we consider square matrix, by writing the identity matrix Id as

$$\mathrm{Id}_{ij} := \delta_{ij} \quad \text{for all } i,j\in\{1,\cdots,n\}, \tag{1.1.10}$$

and if we consider a unit vector $\hat{\mathbf{v}}$ (that is, $\|\hat{\mathbf{v}}\|=1$), we see that

$$\mathrm{Id}-\hat{\mathbf{v}}\otimes\hat{\mathbf{v}}$$

is a projection, in the sense that

$$\begin{aligned}(\mathrm{Id}-\hat{\mathbf{v}}\otimes\hat{\mathbf{v}})(\mathrm{Id}-\hat{\mathbf{v}}\otimes\hat{\mathbf{v}}) &= \mathrm{Id}-\hat{\mathbf{v}}\otimes\hat{\mathbf{v}}-\hat{\mathbf{v}}\otimes\hat{\mathbf{v}}+(\hat{\mathbf{v}}\otimes\hat{\mathbf{v}})(\hat{\mathbf{v}}\otimes\hat{\mathbf{v}})\\ &= \mathrm{Id}-\hat{\mathbf{v}}\otimes\hat{\mathbf{v}}-\hat{\mathbf{v}}\otimes\hat{\mathbf{v}}+\hat{\mathbf{v}}\overbrace{(\hat{\mathbf{v}}^*\hat{\mathbf{v}})}^{=\|\hat{\mathbf{v}}\|=1}\hat{\mathbf{v}}^*\\ &= \mathrm{Id}-\hat{\mathbf{v}}\otimes\hat{\mathbf{v}}.\end{aligned}$$

For any vector $\mathbf{w}\in\mathbb{C}^n$, it is interesting to see that $(\mathrm{Id}-\hat{\mathbf{v}}\otimes\hat{\mathbf{v}})\mathbf{w}$ is perpendicular to $\hat{\mathbf{v}}$:

$$\begin{aligned}((\mathrm{Id}-\hat{\mathbf{v}}\otimes\hat{\mathbf{v}})\mathbf{w},\hat{\mathbf{v}})_{\mathbb{C}^n} &= \hat{\mathbf{v}}^*(\mathrm{Id}-\hat{\mathbf{v}}\otimes\hat{\mathbf{v}})\mathbf{w} \quad (\text{cf. Remark 1.1.8})\\ &= (\hat{\mathbf{v}}^*-\overbrace{(\hat{\mathbf{v}}^*\hat{\mathbf{v}})}^{=\|\hat{\mathbf{v}}\|=1}\hat{\mathbf{v}}^*)\mathbf{w}=0.\end{aligned}$$

Note that the Gram-Schmidt process (cf. Exercise 1.1.6) can be reformulated in terms of (1.1.9).

The trace of $A\in\mathbb{C}^{n\times n}$ is defined by

$$\mathrm{tr}\,(A) := \sum_{j=1}^n A_{jj}.$$

EXERCISE 1.1.12. Let $A,B\in\mathbb{C}^{n\times n}$, show that

(a) $\|A\| = (\mathrm{tr}\,(A^*A))^{1/2}$,
(b) $\|A+B\|\le\|A\|+\|B\|$,
(c) $\|AB\|\le\|A\|\|B\|$.

1.2. Infinite dimensional case (strong topology)

It is natural to extend the preceding discussion to the infinite-dimensional setting, which arises frequently in practical applications. A fundamental example is the Fourier series, which will be introduced in this section. We shall carefully explain the similarities and differences between the finite-dimensional theory (of linear algebra) and the infinite-dimensional theory (of functional analysis).

DEFINITION 1.2.1. An abstract space X is called a $\mathbb{C}$-linear space if

$$au+bv \in X \text{ for all } u,v \in X \text{ and for all } a,b \in \mathbb{C}.$$

For example, the space of polynomials is a linear space because every linear combination of polynomials is again a polynomial. Given any *finitely* many elements $u_1,\cdots,u_m$, their linear span is defined by

$$\operatorname{span}\{u_1,\cdots,u_m\} := \left\{ \sum_{k=1}^{m} \alpha_k u_k : \alpha_1,\cdots,\alpha_k \in \mathbb{C} \right\}.$$

The linear span has essentially the same algebraic structure as $\mathbb{C}^m$. If there exist linearly independent elements $u_1,\cdots,u_m \in X$ such that $X = \operatorname{span}\{u_1,\cdots,u_m\}$, then X is said to be *finite dimensional*, and its dimension is defined by $\dim X = m$. Otherwise, X is called *infinite dimensional*.

In many applications, one encounters infinite dimensional spaces. A typical example is the space $L^2(\Omega)$, where $\Omega \subset \mathbb{R}^n$ is a bounded open set. Roughly speaking, for given any $u \in L^2(\Omega)$, one would represent u as an infinite series (via partial sum)

$$u = \sum_{k} \alpha_k u_k \tag{1.2.1}$$

with respect to a suitable basis $\{u_k\}$. To make such representative meaningful, we must introduce a rigorous notion of convergence for infinite sums of functions. This will be one of the main goals of the present section. We begin with the following definition:

DEFINITION 1.2.2. Given a linear space X over $\mathbb{C}$. A real valued function $\|\cdot\| : X \to \mathbb{R}$ is called a norm if the following properties are satisfied:

(1) **Positive definiteness.** $\|u\| \geq 0$ for all $u \in X$ and there exists a unique element $0 \in X$ such that $\|u\| = 0$ if and only if $u = 0$.
(2) **Absolute homogeneity.** $\|\alpha u\| = |\alpha|\|u\|$ for all $\alpha \in \mathbb{C}$ and $u \in X$.
(3) **Subadditivity/Triangle inequality.** $\|u+v\| \leq \|u\| + \|v\|$ for all $u,v \in X$.

The pair $(X,\|\cdot\|)$ is called the *normed space*.

For finite dimensional case, the following holds true (this is the main reason why we don't mention topology in linear algebra course):

THEOREM 1.2.3 ([Rud91, Theorem 1.21]). *If Y is a linear space with* $\dim(Y) = n \in \mathbb{N}$*, then* $Y \cong \mathbb{C}^n$ *and Y is topological complete (i.e. topological closed).*

EXERCISE 1.2.4. All norms on finite-dimensional vector spaces V are equivalent, that is, if $\|\cdot\|_1$ and $\|\cdot\|_2$ are norms on V, then there exists a constant $c > 0$ such that

$$c\|u\|_1 \le \|u\|_2 \le c^{-1}\|u\|_1 \quad \text{for all } u \in V.$$

Unfortunately, the infinite-dimensional analogues of both Theorem 1.2.3 and Exercise 1.2.4 are generally false. Consequently, unlike in finite-dimensional linear algebra, convergence in an infinite-dimensional space cannot be discussed without specifying the underlying topology. Roughly speaking, whenever we say that a sequence converges, we must always ask "in what sense does it converge". Therefore we need the following notion:

DEFINITION 1.2.5. Given a normed space $(X, \|\cdot\|)$. A sequence $\{u_k\}_{k\in\mathbb{N}} \subset X$ is said to be Cauchy if the following holds true: Given any $\varepsilon > 0$, there exists $N = N(\varepsilon) > 0$ such that

$$\|u_k - u_\ell\| < \varepsilon \quad \text{for all integers } k, \ell \ge N.$$

A normed space $(X, \|\cdot\|)$ is said to be *completed* (or *Banach*) if the following holds true: Given any Cauchy sequence $\{u_k\}_{k\in\mathbb{N}} \subset X$, there exists a unique $u \in X$ such that

$$u_k \to u \text{ strongly in } X, \text{ that is, } \lim_{k\to\infty} \|u_k - u\| = 0. \tag{1.2.2}$$

REMARK 1.2.6. Unlike the finite dimensional setting, in the infinite dimensional setting one must carefully specify the underlying topology. In particular, there are notions of convergence other than the strong convergence defined in (1.2.2), such as the weak convergence and the weak-$\star$ convergence. These concepts play a fundamental role in functional analysis and its applications. However, since these lecture notes are intended for students who are still developing familiarity with abstract mathematical reasoning, we shall not pursue these topics further at this stage.

EXERCISE 1.2.7. Construct a sequence of functions $\{f_k : [0,1] \to \mathbb{C}\}_{k\in\mathbb{N}}$ such that f_k converges strongly in $L^1(0,1)$, but not converges strongly in $L^\infty(0,1)$.

DEFINITION 1.2.8. Let Ω be an open set in $\mathbb{R}^n$. a sequence of functions $\{f_k : \Omega \to \mathbb{C}\}_{k\in\mathbb{N}}$ is said to be converge uniformly to $f : \Omega \to \mathbb{C}$ if $f_k \to f$ strongly in $L^\infty(\Omega)$.

REMARK 1.2.9. It is well known that a sequence of continuous functions converges uniformly must have continuous limit. In other words, $(C(\Omega), \|\cdot\|_{L^\infty(\Omega)})$ is a Banach space.

We now describe the "angle" of an abstract object (such as functions/operators...):

DEFINITION 1.2.10. Let H be a vector space. We say that a mapping $(\cdot,\cdot) : H \times H \to \mathbb{C}$ is a *sesquilinear form* if

(1) $(\cdot,u)$ **is linear for each fixed** $u \in H$, that is, $(\alpha_1 u_1 + \alpha_2 u_2, u) = \alpha_1(u_1,u) + \alpha_2(u_2,u)$ for all $u_1,u_2 \in H$ and for all $\alpha_1,\alpha_2 \in \mathbb{C}$.
(2) $(v,\cdot)$ **is sesquilinear for each fixed** $v \in H$, that is $(v,\alpha_1 u_1 + \alpha_2 u_2) = \overline{\alpha_1}(v,u_1) + \overline{\alpha_2}(v,u_2)$ for all $u_1,u_2 \in H$ and for all $\alpha_1,\alpha_2 \in \mathbb{C}$.

A *scalar product* or *inner product* is a bilinear form $(\cdot,\cdot) : H \times H \to \mathbb{C}$ such that

(1) **Positive definiteness.** $(u,u) \geq 0$ for all $u \in H$ and $(u,u) = 0$ iff $u = 0$;
(2) **Conjugate symmetry.** $(u,v) = \overline{(v,u)}$ for all $u,v \in H$; and

In this case, we call the pair $(H,(\cdot,\cdot))$ an *inner product space*.

REMARK 1.2.11. In the infinite-dimensional setting, the term "inner product" can sometimes be ambiguous, as some mathematicians denote it as the "duality pairing" between a space and its dual. To avoid possible confusion and to remain consistent with the terminology commonly used in linear algebra, we shall use the term scalar product throughout these notes when referring to an inner product.

EXERCISE 1.2.12. Show the following Cauchy-Schwartz inequality:

$$|(u,v)| \leq (u,u)^{\frac{1}{2}}(v,v)^{\frac{1}{2}} \quad \text{for all } u,v \in H.$$

In addition, show that the function $\|\cdot\|$ defined by $\|u\| := (u,u)^{\frac{1}{2}}$ for all $u \in H$ is a norm, which satisfies the parallelogram law:

$$\left\|\frac{u+v}{2}\right\|^2 + \left\|\frac{u-v}{2}\right\|^2 = \frac{1}{2}(\|u\|^2 + \|v\|^2) \quad \text{for all } u,v \in H. \tag{1.2.3}$$

DEFINITION 1.2.13. Given an inner product space $(H,(\cdot,\cdot))$, and induce a norm $\|\cdot\|$ as above. If the *normed* space $(H,\|\cdot\|)$ is complete, then we called H a *Hilbert space*.

Such a framework is frequently used in spaces of functions, hence the name "functional analysis". Before exhibiting a first example, let us quickly briefing the Lebesgue measure and Lebesgue integral, following [**WZ15**]. The volume of the n-dimensional rectangles $R := [a_1,b_1] \times \cdots \times [a_n,b_n]$ is given by $|R| := \prod_{j=1}^n (b_j - a_j)$. Given an arbitrary set $E \subset \mathbb{R}^n$, a natural way to approximate its "volume" is to cover it by *countably many* rectangles, possibly overlapping, as follows:

$$E \subset \bigcup_{j\in\mathbb{N}} R_j,$$

and thus its "volume" should have an upper bound $\sum_{j\in\mathbb{N}} |R_j|$. Let $\mathfrak{R}$ be the collection of all possible rectangular cover $\{R_j\}_{j\in\mathbb{N}}$ as described above. Then, the Lebesgue (outer) measure of E is simply defined by

$$|E| := \inf_{\{R_j\}_{j\in\mathbb{N}} \in \mathfrak{R}} \sum_{j\in\mathbb{N}} |R_j|.$$

EXERCISE 1.2.14. If $\{E_k\}_{k\in\mathbb{N}}$ is a countable collection of sets, show that

$$\left|\bigcup_{k\in\mathbb{N}} E_k\right| \leq \sum_{k\in\mathbb{N}} |E_k|. \tag{1.2.4}$$

A natural question would be, if such collection $\{E_k\}_{k\in\mathbb{N}}$ is disjoint, could we obtain the equality in (1.2.4)? Indeed, yes, provided that E_k are *measurable* in the following sense:

DEFINITION 1.2.15 (Lebesgue measurable). A subset E of $\mathbb{R}^n$ is said to be (Lebesgue) *measurable*, if given $\varepsilon > 0$, there exists an open set U such that

$$E \subset U \text{ and } |U \setminus E| < \varepsilon.$$

This is a mild technical assumption that is readily satisfied in practical applications. We shall not pursue this matter further in this note. For convenience, we use the convention $[0,+\infty] := [0,+\infty) \cup \{+\infty\}$ and $[-\infty,+\infty] := \mathbb{R} \cup \{\pm\infty\}$. The abstract elements $\pm\infty$ can be defined via the *axiom of choices* so that $[-\infty,+\infty]$ forms a well-ordered set with "minimum" $-\infty$ and "maximum" $+\infty$. Now it is natural to introduce the definition of Lebesgue intgeral of nonnegative functions:

DEFINITION 1.2.16. The *Lebesgue integral* of a *nonnegative* function $f : E \to [0,+\infty]$ on E, denoted as

$$\int_E f \quad \text{or} \quad \int_E f \,\mathrm{dm} \quad \text{or} \quad \int_E f(x)\,\mathrm{d}x,$$

is defined as the $\mathbb{R}^{n+1}$-Lebesgue measure of the region under f over E, given by

$$\left\{ (x,y) \in \mathbb{R}^n \times \mathbb{R} : x \in E, \begin{array}{l} 0 \leq y \leq f(x) \text{ if } f(x) < +\infty \\ 0 \leq y < +\infty \text{ if } f(x) = +\infty \end{array} \right\}. \tag{1.2.5}$$

Accordingly, the Lebesgue integral of a (signed) function $g : E \to [-\infty,+\infty]$ on E is defined by

$$\int_E g := \int_E g_+ - \int_E g_-,$$

where $g_+ = \max\{g,0\}$ and $g_- = -\min\{g,0\} = \max\{-g,0\}$ are defined pointwisely. We say that $g \in L^1(E)$ if both $\int_E g_+ < +\infty$ and $\int_E g_- < +\infty$.

The only requirement to make sure Definition 1.2.16 is well defined is the set (1.2.5) need to be measurable in the sense of Definition 1.2.15. If this is the case, we say that the *function is measurable*. From this perspective, the definition of the Riemann integral is somewhat unnatural, since there is no intrinsic reason to approximate (1.2.5) exclusively by rectangles whose "bases" lie in $E \times \{0\}$. Therefore, it is somewhat surprising to me that so much time is still devoted to teaching the Riemann integral in introductory calculus courses.

EXERCISE 1.2.17. Show that, if a nonnegative function $f : E \to \mathbb{R} \cup \{+\infty\}$ satisfy $\int_E f < +\infty$, then $f < +\infty$ almost everywhere in E, that is, the set $\{x \in E : f(x) = +\infty\}$ has measure zero.

EXAMPLE 1.2.18. Let Ω be any open set in $\mathbb{R}^n$, and we define $L^2(\Omega) := \{u : \Omega \to \mathbb{R} \text{ with } \int_\Omega |u(x)|^2 \,\mathrm{d}x < \infty\}$. The mapping (defined using Lebesgue integral)

$$(\cdot,\cdot)_{L^2(\Omega)} : L^2(\Omega) \times L^2(\Omega) \to \mathbb{R}$$
$$(u,v)_{L^2(\Omega)} := \int_\Omega u(x)\overline{v(x)}\,\mathrm{d}x \quad \text{for all } u,v \in L^2(\Omega)$$

is a scalar product, and we denote $\|u\|_{L^2(\Omega)} = \sqrt{(u,u)_{L^2(\Omega)}}$ for all $u \in L^2(\Omega)$ the corresponding norm. In fact, $L^2(\Omega)$ is complete with respect to the norm $\|\cdot\|_{L^2(\Omega)}$, see e.g. [**WZ15**].

EXERCISE 1.2.19 ([**Bre11**, Exercise 5.1]). Let $(H, \|\cdot\|)$ be a normed space. Suppose that the norm $\|\cdot\|$ satisfies the parallelogram law (1.2.3). Define

$$(u,v) := \frac{1}{4}\sum_{k=1}^{4} \mathrm{i}^k \|u + \mathrm{i}^k v\| \quad \text{for all } u,v \in H.$$

(1) Check that $(u,u) = \|u\|^2$, $(u,v) = \overline{(v,u)}$, $(-u,v) = -(u,v)$ and $(2u,v) = 2(u,v)$, as well as $(\mathbf{i}u,v) = \mathbf{i}(u,v)$, for all $u,v \in H$.
(2) Prove that $(u+v,w) = (u,w) + (v,w)$ for all $u,v,w \in H$. [Hint: compute $\|u_1+u_3\|^2 + \|u_2\|^2$ and $\|u_2+u_3\|^2 + \|u_1\|^2$ as well as $\|u_1-u_3\|^2 + \|u_2\|^2$ and $\|u_2-u_3\|^2 + \|u_1\|^2$ using the parallelogram law (1.2.3)]
(3) Prove that $(\lambda u, v) = \lambda(u,v)$ for all $\lambda \in \mathbb{C}$ and for all $u,v \in H$. [Hint: Consider first the case $\lambda \in \mathbb{N}$, then $\lambda \in \mathbb{Q}$, then $\lambda \in \mathbb{R}$ and finally $\lambda \in \mathbb{C}$]
(4) Conclude that $(\cdot,\cdot)$ is a scalar product on H.

EXERCISE 1.2.20 (L^p is not a Hilbert space for $p \neq 2$ [**Bre11**, Exercise 5.2]). Let Ω be any open set in $\mathbb{R}^n$. For each $1 \le p \le +\infty$, we define

$$\|u\|_{L^p(\Omega)} := \begin{cases} \left(\int_\Omega |u|^p\right)^{1/p} & \text{if } 1 \le p < +\infty, \\ \sup_\Omega |u| & \text{if } p = +\infty. \end{cases}$$

Show that $\|\cdot\|_{L^p(\Omega)}$ is a norm and it satisfies the parallelogram law (1.2.3) if and only if $p = 2$. [Hint: Use functions with disjoint supports]

It is important to mention the following theorem.

THEOREM 1.2.21 (Riesz-Fischer). *Let Ω be an open set. For each $1 \le p \le +\infty$, the linear space $L^p(\Omega) := \{f : \Omega \to \mathbb{R} \text{ with } \|u\|_{L^p(\Omega)} < \infty\}$ is Banach.*

It is important to note that Theorem 1.2.21 holds true only when the space $L^p(\Omega)$ is defined using Lebesgue integral. For example, consider the nonnegative function $u : [0,1] \to [0,+\infty]$ defined by

$$u(x) := \begin{cases} x^{-1/2} & 0 < x \le 1, \\ 0 & x = 0. \end{cases}$$

This function is not Riemann integrable on $[0,1]$ (although it can be interpreted as an improper Riemann integral). However, for each $N \in \mathbb{N}$, the truncated functions

$$u_N(x) := \begin{cases} x^{-1/2} & 1/N < x \le 1, \\ 0 & 0 \le x \le 1/N, \end{cases}$$

are Riemann integrable. Moreover, the sequence $\{u_N\}$ is a Cauchy sequence with respect to the norm $\|\cdot\|_{L^1(0,1)}$, while its pointwise limit is the function u. This example illustrates that the class of Riemann integrable functions is not an appropriate for limiting processes, and thus is not well-suited as a framework for modern analysis.

We now collect some properties of L^1 spaces in the following theorem.

LEMMA 1.2.22 (Monotone convergence theorem, Beppo Levi). *Let Ω be an open set in $\mathbb{R}^n$, and let $\{f_k\}_{k\in\mathbb{N}}$ be a sequence of functions in $L^1(\Omega)$ satisfying*

$$f_1(x) \le f_2(x) \le \cdots \le f_k(x) \le f_{k+1}(x) \le \cdots \quad \textit{for a.e. } x \in \Omega, \quad \sup_{k\in\mathbb{N}} \int_\Omega f_k < \infty,$$

then $f_k(x)$ converges a.e. on Ω to a finite limit, which we denote by $f(x)$. In addition, such limit function $f(x)$ belongs to $L^1(\Omega)$ and satisfies

$$f_k \to f \textit{ in } L^1(\Omega), \quad \textit{that is,} \quad \lim_{k\to\infty} \|f_k - f\|_{L^1(\Omega)} = 0.$$

LEMMA 1.2.23 (Lebesgue dominated convergence theorem [**Bre11**, Theorem 4.2]). *Let Ω be an open set in $\mathbb{R}^n$ and let $\{f_k\}_{k\in\mathbb{N}} \subset L^1(\Omega)$ be a sequence of functions satisfying*

(1) *$f_k(x) \to f(x)$ a.e. in Ω;*
(2) *there is a function $g \in L^1(\Omega)$ such that $|f_k(x)| \le g(x)$ a.e. in Ω for all $k \in \mathbb{N}$.*

Then $f \in L^1(\Omega)$ and $f_k \to f$ in $L^1(\Omega)$.

LEMMA 1.2.24 (Fatou's lemma). *Let Ω be an open set in $\mathbb{R}^n$ and let $\{f_k\}_{k\in\mathbb{N}} \subset L^1(\Omega)$ be a sequence of functions satisfying*

$$f_k(x) \ge 0 \textit{ for a.e. } x \in \Omega \textit{ and for all } k \in \mathbb{N}, \quad \sup_{k\in\mathbb{N}} \int_\Omega f_k < \infty.$$

We set $f(x) := \liminf_{k\to\infty} f_k(x) \le +\infty$. Then $f \in L^1(\Omega)$ and

$$\int_\Omega f(x)\,\mathrm{d}x \le \liminf_{k\to\infty} \int_\Omega f_k(x)\,\mathrm{d}x.$$

It is also important to mention the following fact:

THEOREM 1.2.25 (Fubini's theorem). *Let Ω_1 be an open set in $\mathbb{R}^n$ and let Ω_2 be an open set in $\mathbb{R}^m$, and let $F : \Omega_1 \times \Omega_2 \to \mathbb{R}$ be a (measurable) function.*

(1) *If* $F \geq 0$ *a.e. in* $\Omega_1 \times \Omega_2$*, then*

$$\int_{\Omega_2}\int_{\Omega_1} F(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_{\Omega_1}\int_{\Omega_2} F(x,y)\,\mathrm{d}y\,\mathrm{d}x. \tag{1.2.6}$$

(2) *If* $F \in L^1(\Omega_1 \times \Omega_2)$*, i.e.* $\int_{\Omega_2}\int_{\Omega_1} |F(x,y)|\,\mathrm{d}x\,\mathrm{d}y < \infty$*, then* (1.2.6) *also holds.*

We now collect some elementary properties of L^p spaces in the following theorem.

THEOREM 1.2.26 ([**Bre11**, Theorems 4.6–4.8]). *Let* Ω *be an open set in* $\mathbb{R}^n$*. Assume that* $f \in L^p(\Omega)$ *and* $g \in L^{p'}(\Omega)$ *with* $1 \leq p \leq \infty$ *and* $\frac{1}{p'} + \frac{1}{p} = 1$*. Then* $fg \in L^1(\Omega)$ *and the following* Hölder's inequality *holds:*

$$\int_\Omega |f(x)g(x)|\,\mathrm{d}x \leq \|f\|_{L^p(\Omega)}\|g\|_{L^{p'}(\Omega)} \tag{1.2.7}$$

and the equality in (1.2.7) *holds when there exists* $c \in \mathbb{R}$ *such that* $|g(x)| = c|f(x)|^{p-1}$ *for a.e.* $x \in \Omega$*. In addition, the function* $\|\cdot\|_{L^p(\Omega)} : L^p(\Omega) \to \mathbb{R}$ *defines a norm, and* $(L^p(\Omega), \|\cdot\|_{L^p(\Omega)})$ *is a Banach space for all* $1 \leq p \leq \infty$ *(this is well-known as* Fischer-Riesz Theorem).

EXERCISE 1.2.27. For each $1 < p < \infty$ and $\frac{1}{p'} + \frac{1}{p} = 1$, show the following inequality:

$$ab \leq \frac{1}{p}a^p + \frac{1}{p'}b^{p'} \quad \text{for all } a \geq 0 \text{ and } b \geq 0.$$

Use this to conclude the Hölder's inequality (1.2.7). [Hint: One way to show this is using the concavity of the logarithmic function on $(0,\infty)$.]

EXERCISE 1.2.28. Let $f \in L^p(\Omega)$, $g \in L^q(\Omega)$ and $h \in L^r(\Omega)$ for some $1 \leq p,q,r \leq \infty$ with $\frac{1}{p} + \frac{1}{q} + \frac{1}{r} = 1$. Then $fgh \in L^1(\Omega)$ and

$$\int_\Omega |f(x)g(x)h(x)|\,\mathrm{d}x \leq \|f\|_{L^p(\Omega)}\|g\|_{L^q(\Omega)}\|h\|_{L^r(\Omega)}.$$

EXERCISE 1.2.29. Show that $\|\cdot\|_{L^p(\Omega)} : L^p(\Omega) \to \mathbb{R}$ defines a norm. [Hint: Use the convexity of the mapping $t \mapsto t^p$ for each $1 \leq p < \infty$]

EXERCISE 1.2.30. Given any $f \in L^p(\Omega)$, show that

$$\|f\|_{L^p(\Omega)} = \sup_{\|g\|_{L^{p'}(\Omega)}=1} \int_\Omega f(x)g(x)\,\mathrm{d}x \tag{1.2.8}$$

as well as

$$\|f\|_{L^p(\Omega)} = \sup_{\|g\|_{L^{p'}(\Omega)}=1} \int_\Omega |f(x)g(x)|\,\mathrm{d}x \tag{1.2.9}$$

EXERCISE 1.2.31 (Minkowski's integral inequality). In fact, (1.2.8) holds true for any (measurable) f (not necessary in $L^p(\Omega)$). Using this fact to show

$$\left(\int_{\Omega_2}\left|\int_{\Omega_1} F(x,y)\,\mathrm{d}x\right|^p \mathrm{d}y\right)^{\frac{1}{p}} \leq \int_{\Omega_1}\left(\int_{\Omega_2} |F(x,y)|^p\,\mathrm{d}y\right)^{\frac{1}{p}} \mathrm{d}x.$$

EXERCISE 1.2.32. Deduce that if $f \in L^p(\Omega)\cap L^q(\Omega)$ with $1 \leq p \leq \infty$ and $1 \leq q \leq \infty$, then $f \in L^r(\Omega)$ for every r between p and q. More precisely, write

$$\frac{1}{r} = \frac{\alpha}{p} + \frac{1-\alpha}{q} \quad \text{with } 0 \leq \alpha \leq 1$$

and prove that

$$\|f\|_{L^r(\Omega)} \leq \|f\|_{L^p(\Omega)}^{\alpha}\|f\|_{L^q(\Omega)}^{1-\alpha}.$$

Let $1 \leq p < +\infty$, let $E \subset \mathbb{Z}$ and define $\ell^p(E) := \left\{\{a_k\}_{k\in E} : a_k \in \mathbb{C}, \|\{a_k\}_{k\in E}\|_{\ell^p(E)} < +\infty\right\}$, where[1]

$$\|\{a_k\}_{k\in E}\|_{\ell^p(E)} := \left(\sum_{k\in E} |a_k|^p\right)^{1/p}.$$

Show that $\|\cdot\|_{\ell^p(E)}$ is a norm and it satisfies the parallelogram law (1.2.3) if and only if $p = 2$.

REMARK 1.2.33. It is interesting to note that the corresponding scalar product of $\ell^2(E)$ is

$$(\{a_k\}_{k\in E},\{b_k\}_{k\in E})_{\ell^2(E)} = \sum_{k\in E} a_k\overline{b_k},$$

which is comparable to (1.1.1) and Remark 1.1.8.

We now clarify the following concept:

DEFINITION 1.2.34. Given an inner product space $(H,(\cdot,\cdot))$ a set $\{u_\alpha\}_{\alpha\in\Lambda}$ is said to be linearly independent if the following holds true for arbitrary *finite* set $I \subset \Lambda$:

$$\sum_{\alpha\in I} c_\alpha u_\alpha = 0 \quad \text{if and only if} \quad c_\alpha = 0 \text{ for all } \alpha \in I.$$

We also denote its span by

$$\operatorname{span}\{u_\alpha\}_{\alpha\in\Lambda} := \left\{\sum_{\alpha\in I} c_\alpha u_\alpha : c_\alpha \in \mathbb{C}, I \text{ is finite}\right\}.$$

Moreover,

(1) The linearly independent set $\{u_\alpha\}_{\alpha\in\Lambda}$ is said to be *orthogonal* if $(u_\alpha,u_\beta) = 0$ for all $\alpha \neq \beta$.
(2) The linearly independent set $\{u_\alpha\}_{\alpha\in\Lambda}$ is said to be *orthonormal* if $(u_\alpha,u_\beta) = \delta_{\alpha\beta}$ for all $\alpha,\beta \in \Lambda$.

[1]The series converges absolutely, hence, we may interchange the order of summation (a discrete version of Fubini's theorem).

This means that linear independence can only be checked for *finitely many elements* at a time. Given an inner product space $(H,(\cdot,\cdot))$ and an orthonormal set $\{\phi_k\}_{k\in\mathbb{N}}$ which is countable. It is easy to check that

$$\|u\|^2 \geq \sum_{k\in\mathbb{N}} |(u,\phi_k)|^2 \quad \text{for all } u\in H.$$

We now exhibit criteria to check whether $\{\phi_k\}_{k\in\mathbb{N}}$ forms an orthonormal basis. In particular, the following general result on separable Hilbert space can be proved using the Hahn-Banach theorem [**Bre11**, Corollary 1.8]:

PROPOSITION 1.2.35. *Let $\big(H,(\cdot,\cdot)\big)$ be a separable*[2] *Hilbert space, and let $\{\phi_k\}_{k\in\mathbb{N}}$ be an orthonormal subset of H. Then the following are equivalent:*

(1) *$\{\phi_k\}_{k\in\mathbb{N}}$ is an orthonormal (Hilbert) basis, that is, each element $u\in H$ can be approximated by its projection $\sum_{k=1}^N (u,\phi_k)\phi_k$ in the sense of Definition 1.2.5.*
(2) ***Parseval identity.*** *The following equality holds:*

$$\|u\|^2 = \sum_{k\in\mathbb{N}} |(u,\phi_k)|^2 \quad \textit{for all } u\in H.$$

(3) ***Linearly independence.*** *If $u\in H$ and $(u,\phi_k)=0$ for all $k\in\mathbb{N}$, then $u\equiv 0$.*

The finite dimensional analogue of Proposition 1.2.35 is typically covered in elementary linear algebra courses. The main difference between the finite dimensional and infinite dimensional cases is that, in the latter, it is no longer possible to formulate a criterion in terms of the cardinality of $\{\phi_k\}$. For example, a natural orthonormal basis for $\ell^2(\mathbb{N})$ would be the sequences

$$\begin{aligned} \mathbf{e}_1 &= (1,0,0,\cdots) \\ \mathbf{e}_2 &= (0,1,0,\cdots) \\ &\ \vdots \end{aligned}$$

Obviously, if we taken away $\mathbf{e}_1$, the orthonormal set $\{\mathbf{e}_k\}_{k\geq 2}$ is no longer a basis of $\ell^2(\mathbb{N})$, despite $\{\mathbf{e}_k\}_{k\geq 2}$ and $\{\mathbf{e}_k\}_{k\geq 1}$ have same cardinality.

We now consider the abstract mapping $T: X\to Y$ from Hilbert space $(X,(\cdot,\cdot)_X)$ to Hilbert space $(Y,(\cdot,\cdot)_Y)$. Similar as above, a naive method to study the mapping T is simply test it by using all $u\in X$, which is obviously not practical. Intuitively, if T has some special structure, then we need not to test all $u\in X$. In this course, we are in particular interested in the case when $T: X\to Y$ is linear, that is,

$$T(\alpha v+\beta w) = \alpha T(v)+\beta T(w) \quad \text{for all } \alpha,\beta\in\mathbb{C} \text{ and for all } v,w\in X.$$

[2]This means that there exists a countable subset $H_0\subset H$ such that $\overline{H_0}=H$.

Let's say we have an orthonormal basis $\{\phi_j\}_{j\in\mathbb{N}}$ of X, then each $u \in X$ can be uniquely represented as

$$u = \sum_{j\in\mathbb{N}} c_j \phi_j \quad \text{(in the sense of Definition 1.2.5)}$$

and therefore

$$T(u) = \sum_{j\in\mathbb{N}} c_j T(\phi_j) \quad \text{(in the sense of Definition 1.2.5).}$$

In other words, it is suffice to measure $\{T(\phi_j)\}_{j\in\mathbb{N}}$, and thus for each u the result $T(u)$ can be reconstructed according to the above identity. We now fix an orthonormal basis $\{\psi_i\}_{i\in\mathbb{N}}$ of Y, then the measurement $T(\phi_j)$ can be uniquely represented as

$$T(\phi_j) = \sum_{i\in\mathbb{N}} \big(T(\phi_j), \psi_i\big)_Y \, \psi_i \quad \text{(in the sense of Definition 1.2.5).}$$

This shows that T is uniquely determined by the coefficients

$$\big(T(\phi_j), \psi_i\big)_Y \quad \text{for all } i, j \in \mathbb{N}$$

of an "infinite dimensional matrix", which is analogue to Remark 1.1.9.

1.3. An example of orthonormal basis: Fourier series

We begin with the construction of a orthonormal basis of $L^2(0,\pi)$. We begin with some computations to motivate the construction.

For odd function, we would like to consider the sequence $\{\phi_k\}_{k\in\mathbb{N}}$ defined by

$$\phi_k(x) = \sqrt{\frac{2}{\pi}} \sin(kx) \quad \text{for } k = 1, 2, \cdots.$$

In particular, we compute that

$$\begin{aligned} \frac{2}{\pi}\int_0^\pi \sin(kx)\sin(kx)\,dx &= \frac{1}{\pi}\int_0^\pi \big(1-\cos(2kx)\big)\,dx \\ &= 1 - \frac{1}{2k}\sin(2kx)\Big|_{x=0}^{\pi} = 1 \end{aligned}$$

and for each $k_1 \neq k_2$ we have

$$\begin{aligned} &\frac{2}{\pi}\int_0^\pi \sin(k_1 x)\sin(k_2 x)\,dx \\ &= \frac{1}{\pi}\int_0^\pi \big(\cos((k_1-k_2)x) - \cos((k_1+k_2)x)\big)\,dx \\ &= \frac{1}{\pi}\left(\frac{1}{k_1-k_2}\sin((k_1-k_2)x) - \frac{1}{k_1+k_2}\sin((k_1+k_2)x)\right)\Big|_{x=0}^{x=\pi} = 0, \end{aligned}$$

showing that $\{\phi_k\}_{k\in\mathbb{N}}$ is an orthonormal set in $L^2(0,\pi)$. For some suitable f, we are interested to the series

$$f(x) = \sum_{k=1}^{\infty} a_k \sin(kx) \quad \text{with} \quad a_k = \frac{2}{\pi}\int_0^{\pi} \sin(kx) f(x)\,dx. \tag{1.3.1}$$

The expansion (1.3.1) is called the *Fourier sine expansion*.

For even function, we would like to consider the sequence $\{\tilde{\phi}_k\}_{k\in\mathbb{N}\cup\{0\}}$ defined by

$$\tilde{\phi}_k(x) = \sqrt{\frac{2}{\pi}}\cos(kx) \quad \text{for } k = 0,1,2,\cdots.$$

Similar computation shows that $\{\tilde{\phi}_k\}_{k\in\mathbb{N}\cup\{0\}}$ is also an orthonormal set in $L^2(0,\pi)$. For some suitable f, we are interested to the series

$$f(x) = \frac{1}{2}b_0 + \sum_{k=1}^{\infty} b_k \cos(kx) \quad \text{with} \quad b_k = \frac{2}{\pi}\int_0^{\pi} \cos(kx) f(x)\,dx. \tag{1.3.2}$$

The expansion (1.3.2) is called the *Fourier cosine expansion*.

Given any function $f:\mathbb{R}\to\mathbb{R}$, we can decompose it into the sum of even function and odd function by the following simple observation:

$$f(x) = \overbrace{\frac{f(x)-f(-x)}{2}}^{\text{odd function}} + \overbrace{\frac{f(x)+f(-x)}{2}}^{\text{even function}}. \tag{1.3.3}$$

Moreover, the decomposition (1.3.3) is unique: If $f(x) = f_{\rm odd}(x) + f_{\rm even}(x)$ for some odd function $f_{\rm odd}$ and even function $f_{\rm even}$, then from (1.3.3) we can write

$$\overbrace{f_{\rm odd}(x) - \frac{f(x)-f(-x)}{2}}^{\text{odd function}} = \overbrace{\frac{f(x)+f(-x)}{2} - f_{\rm even}(x)}^{\text{even function}} \quad \text{for all } x\in\mathbb{R}.$$

This implies

$$f_{\rm odd}(x) = \frac{f(x)-f(-x)}{2} \quad \text{and} \quad f_{\rm even}(x) = \frac{f(x)+f(-x)}{2}.$$

We see that $\sin(kx)$ are odd functions, while $\cos(kx)$ are even functions. Therefore, \for some suitable f, we are interested to the Fourier series

$$f(x) = \frac{1}{2}B_0 + \sum_{k=1}^{\infty}(\overbrace{A_k\sin(kx)}^{\text{odd part}} + \overbrace{B_k\cos(kx)}^{\text{even part}}) \quad \text{for } x\in(-\pi,\pi) \tag{1.3.4}$$

in a suitable sense. Using similar computations, the coefficients are necessarily given by

$$\begin{aligned} A_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(kx)\,dx \quad \text{for } k = 1,2,\cdots, \\ B_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(kx)\,dx \quad \text{for } k = 0,1,2,\cdots. \end{aligned}$$

Up to this point, we note that these computations involve many trigonometric identities, making them rather cumbersome.

Since $e^{i\theta} = \cos\theta + \mathrm{i}\sin\theta$ for all $\theta \in \mathbb{R}$, we may alternatively consider the series

$$f(x) = \sum_{k=-\infty}^{\infty} c_k e^{\mathrm{i}kx} \quad \text{for } x \in (-\pi, \pi) \tag{1.3.5}$$

with $c_k \in \mathbb{C}$. We can write (1.3.5) as

$$\begin{aligned}
f(x) = \sum_{k=-\infty}^{\infty} c_k e^{\mathrm{i}kx} &= c_0 + \sum_{k=1}^{\infty} (c_k e^{\mathrm{i}kx} + c_{-k} e^{-\mathrm{i}kx}) \\
&= c_0 + \sum_{k=1}^{\infty} (\Re c_k + \mathrm{i}\Im c_k)\big(\cos(kx) + \mathrm{i}\sin(kx)\big) \\
&\quad + \sum_{k=1}^{\infty} (\Re c_{-k} + \mathrm{i}\Im c_{-k})\big(\cos(kx) - \mathrm{i}\sin(kx))\big) \\
&= c_0 + \sum_{k=1}^{\infty} \big(\Re c_k \cos(kx) - \Im c_k \sin(kx)\big) + \mathrm{i}\big(\Re c_k \sin(kx) + \Im c_k \cos(kx)\big) \\
&\quad + \sum_{k=1}^{\infty} \big(\Re c_{-k} \cos(kx) + \Im c_{-k} \sin(kx)\big) + \mathrm{i}\big(-\Re c_{-k} \sin(kx) + \Im c_{-k} \cos(kx)\big) \\
&= c_0 + \sum_{k=1}^{\infty} \big(-\Im c_k + \Im c_{-k} + i(\Re c_k - \Re c_{-k})\big) \sin(kx) \\
&+ \sum_{k=1}^{\infty} \big(\Re c_k + \Re c_{-k} + \mathrm{i}(\Im c_k + \Im c_{-k})\big) \cos(kx) \\
&= c_0 + \sum_{k=1}^{\infty} \mathrm{i}(c_k - c_{-k}) \sin(kx) + \sum_{k=1}^{\infty} (c_k + c_{-k}) \cos(kx).
\end{aligned}$$

Therefore from (1.3.4) (the coefficient are unique) we have

$$c_0 = \frac{1}{2} B_0, \quad c_k + c_{-k} = B_k \text{ and } \mathrm{i}(c_k - c_{-k}) = A_k \quad \text{for all } k \in \mathbb{N}.$$

Equivalently,

$$c_0 = \frac{1}{2} B_0, \quad c_k = \frac{1}{2}(B_k - \mathrm{i}A_k) \text{ and } c_{-k} = \frac{1}{2}(B_k + \mathrm{i}A_k) \quad \text{for all } k \in \mathbb{N}. \tag{1.3.6}$$

In particular, (1.3.6) is equivalent to

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-\mathrm{i}kx} \, dx \quad \text{for all } k \in \mathbb{Z}.$$

REMARK 1.3.1. If f is real-valued, then

$$\overline{c_{-k}} = \overline{\frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{\mathrm{i}kx} \, dx} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-\mathrm{i}kx} \, dx = c_k \quad \text{for all } k \in \mathbb{Z}.$$

Conversely, if $\overline{c_{-k}} = c_k$ for all $k \in \mathbb{Z}$, then from (1.3.5) we have

$$\overline{f(x)} = \sum_{k=-\infty}^{\infty} \overline{c_k e^{\mathrm{i}kx}} = \sum_{k=-\infty}^{\infty} \overline{c_k} e^{-\mathrm{i}kx} = \sum_{k=-\infty}^{\infty} \overline{c_{-k}} e^{\mathrm{i}kx} = \sum_{k=-\infty}^{\infty} c_k e^{\mathrm{i}kx} = f(x),$$

that is, f is real-valued.

Unlike trigonometric functions, it is easy work check that $\{e^{\mathrm{i}kx}\}_{k\in\mathbb{Z}}$ forms an orthogonal set in $L^2(-\pi,\pi)$. This also allows us to extend the ideas to the multivariable setting with almost no additional effort. If $f : \mathbb{R}^n \to \mathbb{C}$ is of 2π-periodic in each variable, we want to represent it by the Fourier series

$$f(x) = \sum_{k\in\mathbb{Z}^n} c_k e^{\mathrm{i}k\cdot x} = \sum_{k\in\mathbb{Z}^n} c_k e^{\mathrm{i}k_1x_1} \cdots e^{\mathrm{i}k_2x_2} \quad \text{for all } x = (x_1,\cdots,x_n) \in \mathbb{R}^n \tag{1.3.7}$$

in some suitable sense. We now consider the cube $Q = [-\pi,\pi]^n$, and normalize the inner product on $L^2(Q)$ by

$$(f,g) \equiv (f,g)_{L^2(Q)} := \frac{1}{|Q|}\int_Q f\overline{g}\,dx \equiv \fint_Q f\overline{g}\,dx \quad \text{for all } f,g \in L^2(Q).$$

with $|Q| = (2\pi)^n$.

LEMMA 1.3.2. *The countable set $\{e^{\mathrm{i}k\cdot x}\}_{k\in\mathbb{Z}^n}$ is an orthonormal subset of $L^2(Q)$.*

PROOF. For each $k,\ell \in \mathbb{Z}^n$, using Fubini's theorem we see that

$$\begin{aligned}(e^{\mathrm{i}k\cdot x}, e^{\mathrm{i}\ell\cdot x}) &= (2\pi)^{-n}\int_Q e^{\mathrm{i}(k-\ell)\cdot x}\,dx \\ &= (2\pi)^{-n}\prod_{j=1}^n \int_{-\pi}^{\pi} e^{\mathrm{i}(k_j-\ell_j)x_j}\,dx_j = \begin{cases} 1 & ,k=\ell, \\ 0 & ,k\neq\ell, \end{cases}\end{aligned}$$

which conclude the lemma. □

It remains to prove the following proposition.

PROPOSITION 1.3.3. $\{e^{\mathrm{i}k\cdot x}\}_{k\in\mathbb{Z}^n}$ *a* complete orthonormal basis *of* $L^2(Q)$.

REMARK 1.3.4. Given any $f \in L^2(Q)$ be such that $(f, e^{\mathrm{i}k\cdot x}) = 0$ for all $k \in \mathbb{Z}^n$. If we can prove $f = 0$ a.e. in Q, using Proposition 1.2.35 we conclude Proposition 1.3.3. Since $e^{\mathrm{i}k\cdot x} = e^{\mathrm{i}k_1x_1} \cdots e^{\mathrm{i}k_nx_n}$, We see that

$$0 = \int_{[-\pi,\pi]^n} f e^{-\mathrm{i}k\cdot x}\,dx = \int_{-\pi}^{\pi} \left(\int_{[-\pi,\pi]^{n-1}} f(x) e^{-\mathrm{i}k'\cdot x'}\,dx'\right) e^{-\mathrm{i}k_1x_1}\,dx_1 \quad \text{for all } k \in \mathbb{Z}^n, \tag{1.3.8}$$

with $k' = (k_2,\cdots,k_n)$ and $x' = (x_2,\cdots,x_n)$. If we can show Proposition 1.3.3 for the case when $n = 1$, using Proposition 1.2.35 we know that

$$0 = \int_{[-\pi,\pi]^{n-1}} f(x) e^{-\mathrm{i}k'\cdot x'}\,dx' \quad \text{for all } k' \in \mathbb{Z}^{n-1}. \tag{1.3.9}$$

Repeating the arguments that proving from (1.3.8) to (1.3.9), we conclude that $f = 0$ a.e. in Q.

Before showing Proposition 1.3.3 for $n = 1$, we first prove a consequence of Proposition 1.3.3, which is the main result of this section.

THEOREM 1.3.5 (Fourier series of L^2 functions). *If $f \in L^2(Q)$, then one has the Fourier series*

$$f(x) = \sum_{k\in\mathbb{Z}^n} \hat{f}(k) e^{\mathrm{i}k\cdot x} \text{ converges in } L^2(Q), \tag{1.3.10}$$

with the Fourier coefficients

$$\hat{f}(k) = (f, e^{\mathrm{i}k\cdot x}) = \fint_Q f(x) e^{-\mathrm{i}k\cdot x}\,dx. \tag{1.3.11}$$

One has the Parseval identity

$$\|f\|_{L^2(Q)}^2 = \sum_{k\in\mathbb{Z}^n} |\hat{f}(k)|^2.$$

Conversely, if $c = (c_k) \in \ell^2(\mathbb{Z}^n)$, then the series $\sum_{k\in\mathbb{Z}^n} c_k e^{\mathrm{i}k\cdot x}$ converges in $L^2(Q)$ to some $f \in L^2(Q)$ and it is necessarily $c_k = \hat{f}(k)$.

REMARK 1.3.6. Here we denote by $\ell^2(\mathbb{Z}^n)$ the space of the complex sequences $c = (c_k)_{k\in\mathbb{Z}^n}$ with norm

$$\|c\|_{\ell^2(\mathbb{Z}^n)} = \left(\sum_{k\in\mathbb{Z}^n} |c_k|^2\right)^{\frac{1}{2}}.$$

Theorem 1.3.5 says that there is a 1-1 corresponding between the elements in $L^2(Q)$ with the elements in $\ell^2(\mathbb{Z}^n)$. In other words, (1.3.11) can be viewed as the *discrete Fourier transform*, and the *inverse discrete Fourier transform* is given by the formula (1.3.10).

PROOF OF THEOREM 1.3.5. The first part of Theorem 1.3.5 is an immediate consequence of Proposition 1.3.3. For the converse, if $(c_k) \in \ell^2(\mathbb{Z}^n)$, then we see that the partial sum

$$S_N := \sum_{|k|\le N} c_k e^{\mathrm{i}k\cdot x}$$

is a Cauchy sequence in $L^2(Q)$. Since $L^2(Q)$ is complete, then we know that the series $\sum_{k\in\mathbb{Z}^n} c_k e^{\mathrm{i}k\cdot x}$ converges in $L^2(Q)$ to some $f \in L^2(Q)$. For each $N \ge k$, we also see that

$$|c_k - \hat{f}(k)| = |(S_N - f, e^{\mathrm{i}kx})| \le C_n \|e^{\mathrm{i}kx}\|_{L^2(Q)} \|S_N - f\|_{L^2(Q)} \to 0 \quad \text{as } N \to \infty,$$

which conclude $c_k = \hat{f}(k)$. □

COROLLARY 1.3.7 (sine series). *The sequence $\{\phi_k\}_{k\in\mathbb{N}}$ defined by*

$$\phi_k(x) = \sqrt{\frac{2}{\pi}} \sin(kx) \quad \textit{for } k = 1, 2, \cdots$$

is an orthogonal basis of $L^2(0,\pi)$.

COROLLARY 1.3.8 (cosine series). *The sequence* $\{\tilde{\phi}_k\}_{k\in\mathbb{N}\cup\{0\}}$ *defined by*

$$\tilde{\phi}_k(x) = \sqrt{\frac{2}{\pi}}\cos(kx) \quad \text{for } k = 0,1,2,\cdots.$$

is an orthogonal basis of $L^2(0,\pi)$.

It remains to prove Proposition 1.3.3 (for the case when $n=1$). To this end, we introduce the following special kernel:

DEFINITION 1.3.9. A sequence $\{Q_N(x)\}_{N\in\mathbb{N}}$ of 2π-period continuous functions on the real line is called an *approximate identity* if

(1) $Q_N \geq 0$ for all $N \in \mathbb{N}$,
(2) $⨍_{-\pi}^{\pi} Q_N(x)\,dx \equiv (2\pi)^{-1}\int_{-\pi}^{\pi} Q_N(x)\,dx = 1$ for all $N\in\mathbb{N}$, and
(3) for each $0<\varepsilon<\pi$ one has $\lim_{N\to\infty}\sup_{\varepsilon\leq|x|\leq\pi} Q_N(x) = 0$.

We now prove the existence of such function satisfies Definition 1.3.9.

LEMMA 1.3.10. *The sequence*

$$Q_N(x) := c_N\left(\frac{1+\cos x}{2}\right)^N \quad \text{with} \quad c_N = 2\pi\left(\int_{-\pi}^{\pi}\left(\frac{1+\cos x}{2}\right)^N dx\right)^{-1}$$

is an approximate identity.

PROOF. It is easy to see that $Q_N \geq 0$ and $(2\pi)^{-1}\int_{-\pi}^{\pi} Q_N(x)\,dx = 1$ for all $N\in\mathbb{N}$. We estimate the constant c_N as followings:

$$\begin{aligned} 1 &= \frac{c_N}{2\pi}\int_{-\pi}^{\pi}\left(\frac{1+\cos x}{2}\right)^N dx = \frac{c_N}{\pi}\int_0^{\pi}\left(\frac{1+\cos x}{2}\right)^N dx \\ &\geq \frac{c_N}{\pi}\int_0^{\pi}\left(\frac{1+\cos x}{2}\right)^N \sin x\,dx \\ &= \frac{c_N}{\pi}\int_{-1}^{1}\left(\frac{1+t}{2}\right)^N dt = \frac{2c_N}{\pi}\int_0^1 s^N\,ds = \frac{2c_N}{\pi(N+1)}. \end{aligned}$$

Thus for each $0<\varepsilon<\pi$ we have

$$0 \leq \sup_{\varepsilon\leq|x|\leq\pi} Q_N(x) \leq Q_N(\varepsilon) = c_N\left(\frac{1+\cos\varepsilon}{2}\right)^N \leq \frac{\pi(N+1)}{2}\left(\frac{1+\cos\varepsilon}{2}\right)^N \to 0 \quad \text{as } N\to\infty,$$

because $0 < \frac{1+\cos\varepsilon}{2} < 1$. □

Let f and g be two 2π-period functions. Then we formally define the convolution $f*g$ by

$$(f*g)(x) := \frac{1}{2\pi}\int_{-\pi}^{\pi} f(y)g(x-y)\,dy \equiv \frac{1}{2\pi}\int_{-\pi}^{\pi} g(y)f(x-y)\,dy, \tag{1.3.12}$$

which is also a 2π-periodic function (the second identity holds only for 2π-period functions). The following lemma explains the naming of Definition 1.3.9.

LEMMA 1.3.11. *Let Q_N be an approximate identity as in Definition 1.3.9 and let f be a 2π-periodic function. If f is continuous, then*

$$\lim_{N\to\infty} Q_N * f = f \quad \text{converges in } L^\infty(-\pi,\pi).$$

If $f \in L^p(-\pi,\pi)$ for some $1 \le p < \infty$, then

$$\lim_{N\to\infty} Q_N * f = f \quad \text{converges in } L^p(-\pi,\pi).$$

PROOF OF LEMMA 1.3.11. We first observe that

$$(Q_N * f - f)(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} Q_N(y)\big(f(x-y) - f(x)\big)\,dy.$$

Case 1: f is a continuous 2π-periodic function. Given any $\varepsilon > 0$, there exists $\delta(\varepsilon) > 0$ such that

$$\sup_{|y|\le\delta(\varepsilon)} \big|f(x-y) - f(x)\big| \le \varepsilon \quad \text{for all } x \in \mathbb{R}$$

and

$$\sup_{\delta(\varepsilon)\le|x|\le\pi} Q_N(x) \le \varepsilon \quad \text{for all sufficiently large } N. \tag{1.3.13}$$

Then for all sufficiently large N we estimate

$$\begin{aligned}
|(Q_N * f - f)(x)| &\le \frac{1}{2\pi}\left(\int_{|y|\le\delta(\varepsilon)} + \int_{\delta(\varepsilon)\le|y|\le\pi}\right) Q_N(y)\big|f(x-y) - f(x)\big|\,dy \\
&\le \frac{\varepsilon}{2\pi}\Bigg(\overbrace{\int_{|y|\le\delta(\varepsilon)} Q_N(y)\,dy}^{\le 2\pi} + \overbrace{\int_{\delta(\varepsilon)\le|y|\le\pi} \big|f(x-y) - f(x)\big|\,dy}^{\le 4\pi\|f\|_{L^\infty(\mathbb{R})}}\Bigg) \\
&\le \varepsilon(1 + 2\|f\|_{L^\infty(\mathbb{R})}),
\end{aligned}$$

which gives

$$\limsup_{N\to\infty} |(Q_N * f - f)(x)| \le \varepsilon(1 + 2\|f\|_{L^\infty(\mathbb{R})}).$$

By arbitrariness of $\varepsilon > 0$, we conclude the first part of the lemma.

Case 2: $f \in L^p(-\pi,\pi)$ for some $1 \le p < \infty$. Using the Minkowski's inequality (Exercise 1.2.31), we estimate

$$\begin{aligned}
\|Q_N * f - f\|_{L^p(-\pi,\pi)} &\le \frac{1}{2\pi}\left(\int_{-\pi}^{\pi}\left|\int_{-\pi}^{\pi} Q_N(y)\big(f(x-y) - f(x)\big)\,dy\right|^p dx\right)^{\frac{1}{p}} \\
&\le \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\int_{-\pi}^{\pi} \big|Q_N(y)\big(f(x-y) - f(x)\big)\big|^p dx\right)^{\frac{1}{p}} dy \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} Q_N(y)\|f(\cdot - y) - f\|_{L^p(-\pi,\pi)}\,dy.
\end{aligned}$$

Since $f \in L^p(-\pi,\pi)$, by approximate it by $C_c^\infty(-\pi,\pi)$ functions, given any $\varepsilon > 0$, there exists $\delta(\varepsilon) > 0$ such that

$$\sup_{|y|\leq\delta(\varepsilon)} \|f(\cdot - y) - f\|_{L^p(-\pi,\pi)} \leq \varepsilon.$$

Again using (1.3.13), then for all sufficiently large N we estimate

$$\begin{aligned}
\|Q_N * f - f\|_{L^p(-\pi,\pi)} &\leq \frac{1}{2\pi}\left(\int_{|y|\leq\delta(\varepsilon)} + \int_{\delta(\varepsilon)\leq|y|\leq\pi}\right) Q_N(y)\|f(\cdot - y) - f\|_{L^p(-\pi,\pi)}\,dy \\
&\leq \frac{\varepsilon}{2\pi}\left(\overbrace{\int_{|y|\leq\delta(\varepsilon)} Q_N(y)\,dy}^{\leq 2\pi} + \overbrace{\int_{\delta(\varepsilon)\leq|y|\leq\pi} \|f(\cdot - y) - f\|_{L^p(-\pi,\pi)}\,dy}^{\leq 4\pi\|f\|_{L^p(-\pi,\pi)}}\right) \\
&\leq \varepsilon(1 + 2\|f\|_{L^p(\mathbb{R})}),
\end{aligned}$$

and we prove the second part of the lemma similar as in first part. □

We are now ready to give a direct proof to Proposition 1.3.3.

PROOF OF PROPOSITION 1.3.3 USING DEFINITION 1.3.9. As mentioned in Remark 1.3.4, it is suffice to prove the case when $n = 1$. Let $f \in L^2(-\pi,\pi)$ be such that

$$(f, e^{ikx}) = 0 \quad \text{for all } k \in \mathbb{Z}. \tag{1.3.14}$$

Using Proposition 1.2.35, it is suffice to show $f \equiv 0$ a.e. Using Lemma 1.3.10, in particular from (1.3.14) we have $Q_N * f = 0$ for all $N \in \mathbb{N}$. Since $Q_N * f \to f$ as $N \to \infty$ in $L^2(-\pi,\pi)$, we conclude our result. □

1.4. Additional properties of Fourier series

In light of the discussion preceding Definition 1.2.5, it is natural to investigate other modes of convergence for Fourier series. We briefly state some additional results below without proof. For a more comprehensive treatment of Fourier series and its convergence properties, the reader is referred to my lecture notes [**Kow25**].

We first exhibit the pointwise convergence of the Fourier series:

THEOREM 1.4.1. *Let f be a 2π-periodic function in $\mathbb{R}$ which is of bounded variation. The Fourier coefficients are given by*

$$\hat{f}(k) = (f, e^{ikx}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-ikx}\,dx = ⨍_{-\pi}^{\pi} f(x)e^{-ikx}\,dx.$$

Then the (1-dimensional) Fourier series $\sum_{k\in\mathbb{Z}} \hat{f}(k)e^{ikx}$ of f converges to

$$\frac{1}{2}\big(f(x+) + f(x-)\big) \text{ pointwsely,}$$

where $f(x\pm) := \lim_{\theta\to 0_+} f(x \pm \theta)$.

Theorem 1.4.1 is a consequence of the Dirichlet-Jordan test [**Zyg02**, Theorem (8.1)(i), Section 8, Chapter II][3]. Rather explaining the precise definition of the term "bounded variation", we note that a piecewise continuous function also has bounded variation. From Theorem 1.4.1, we know that if f has a jump, then the Fourier series of f never converges to f uniformly. The following theorem, which is a consequence of the Dirichlet-Jordan test [**Zyg02**, Theorem (8.1)(ii), Section 8, Chapter II], shows that if the Fourier series of a continuous function converges uniformly:

THEOREM 1.4.2. *Let f be a continuous 2π-periodic function. The Fourier coefficients are given by*

$$\hat{f}(k) = (f, e^{ikx}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-ikx}\,dx = \fint_{-\pi}^{\pi} f(x)e^{-ikx}\,dx.$$

Then the (1-dimensional) Fourier series $\sum_{k\in\mathbb{Z}} \hat{f}(k)e^{ikx}$ of f converges to f uniformly in $\mathbb{R}$.

EXERCISE 1.4.3. Prove Theorem 1.4.2 for the special case when $f \in C^1(\mathbb{R})$. (Hint: Compute the Fourier coefficients of f'.)

Despite the partial sum $S_m f$ converges pointwisely to the piecewise continuous function f, the partial sum $S_m f$ produces large peaks around the jump of f, which overshoot and undershoot the function's actual values. This approximation error approaches a limit of about 9%. This phenomenon is called the *Gibbs-Wilbraham phenomenon*, and we refer the details to the survey paper [**HH79, Onn82**]. The Gibbs phenomenon is not limited to Fourier expansion. In fact many researchers have reported the phenomenon in many different eigenfunction expansions, including those based on the Chebyshev and Legendre polynomials, which are special cases of the Gegenbauer polynomials [**GS97**]. In other words, given the Gegenbauer expansion coefficients of a piecewise analytic function, one faces the same convergence problem.

We now state the following theorem (without proof, which can be found in [**Onn82**][4]):

THEOREM 1.4.4 (Gibbs-Wilbraham, Theorem F of [**HH79**]). *Let f be the function given in Theorem 1.4.1. Let $\mathscr{D}$ be the set of discontinuities of f. For each $x \in \mathscr{D}$, let ℓ_x be the vertical line segment with*

$$\textit{length } \frac{2}{\pi}\mathrm{Si}(\pi)|f(x+) - f(x-)| \textit{ centered at } \frac{1}{2}\big(f(x+) - f(x-)\big).$$

Let $\mathscr{G}(g)$ be the graph of g. Then we have

$$\lim_{m\to\infty} \mathscr{G}(S_m f) = \mathscr{G}(f) \cup \bigcup_{x\in\mathscr{D}} \ell_x \quad \textit{(limit as a set).}$$

REMARK 1.4.5. Note that

$$\frac{2}{\pi}\int_0^{\pi} \frac{\sin\theta}{\theta}\,d\theta = \frac{2}{\pi}\mathrm{Si}(\pi) = 1 + 2\big(\overbrace{0.0894\cdots}^{\substack{\text{about } 9\% \\ \text{overshoot}}}\big).$$

[3]I would like to thank Aleksi Harmoinen for pointing this out to me.

[4]I would like to thank Aleksi Harmoinen for pointing this out to me.

Sometimes $\mathrm{Si}(\pi)$ is known as the *Gibbs-Wilbraham constant*. In general, there are many jumps in signal, therefore this 9% overshoot actually causing significant noise in computation.

According to [**GS97**], the first attempt to resolve the Gibbs phenomenon was made by the Hungarian mathematician Lipót Fejér (or Leopold Fejér): In 1990 he discovered that the Cesàro means of the partial sums converge uniformly. Before explaining how to resolve the Gibbs phenomenon, let us begin our discussions from the following simple observation:

LEMMA 1.4.6. *Let $\{c_m\}_{m=0,1,2,\cdots}$ be a sequence of complex numbers. Suppose that it converges to a limit $c \in \mathbb{C}$, then so its Cesàro sum:*

$$\lim_{N\to\infty} \frac{1}{N+1} \sum_{m=0}^{N} c_n = \ell.$$

PROOF. Given any $\varepsilon > 0$, there exists a sufficiently large $N_0 \in \mathbb{N}$ such that

$$|a_m - \ell| \leq \varepsilon \quad \text{for all } m > N_0.$$

For each $N > N_0$, we write

$$\begin{aligned}
\left| \frac{1}{N+1} \sum_{m=0}^{N} c_m - \ell \right| &\leq \frac{1}{N} \sum_{m=0}^{N} |a_m - \ell| = \frac{1}{N} \sum_{m=0}^{N_0} |a_m - \ell| + \frac{1}{N} \sum_{m=N_0+1}^{N} |a_m - \ell| \\
&\leq \frac{1}{N}\left((N_0+1) \sup_{0\leq m\leq N_0} |a_m - \ell| \right) + \overbrace{\frac{N-N_0}{N}}^{\leq 1} \overbrace{\sup_{N_0<m\leq N} |a_m - \ell|}^{\leq \varepsilon} \\
&\leq \frac{1}{N}\left((N_0+1) \sup_{0\leq m\leq N_0} |a_m - \ell| \right) + \varepsilon,
\end{aligned}$$

which implies

$$\limsup_{N\to\infty} \left| \frac{1}{N+1} \sum_{m=0}^{N} c_m - \ell \right| \leq \varepsilon.$$

Our lemma follows from the arbitrariness of $\varepsilon > 0$. □

EXAMPLE 1.4.7 (Grandi's series). Let $a_m = (-1)^m$ for $m \geq 0$. Hence $\{a_m\}_{m=0}^{\infty}$ is the sequence $1, -1, 1, -1, \cdots$. Clearly the partial sum $S_m = \sum_{k=0}^{m} a_k$ does not converges, and in particular

$$\{S_m\}_{m=0}^{\infty} = \{1, 0, 1, 0, \cdots\}.$$

We see that the Cesàro sums are given by

$$\sigma_N = \frac{1}{N+1} \sum_{m=0}^{N} S_m = \begin{cases} \frac{M}{2M-1} & \text{if } N = 2M-1 \text{ is odd}, \\ \frac{1}{2} & \text{if } N = 2M \text{ is even}, \end{cases}$$

which implies

$$\lim_{N\to\infty} \sigma_N = \frac{1}{2}.$$

The above observation suggests that, instead of the partial sums $S_m f$, we consider the Cesàro sums

$$\sigma_N f(x) = \frac{1}{N+1}\sum_{m=0}^{N} S_m f(x).$$

This can be written in convolution form as

$$\sigma_N f(x) = \frac{1}{N+1}\sum_{m=0}^{N} (D_m * f)(x) = (F_N * f)(x),$$

where F_N is the *Fejér kernel*:

$$\begin{aligned} F_N(x) &= \frac{1}{N+1}\sum_{m=0}^{N} \frac{e^{i(m+\frac{1}{2})x} - e^{-i(m+\frac{1}{2})x}}{e^{i\frac{1}{2}x} - e^{-i\frac{1}{2}x}} \\ &= \frac{1}{N+1}\frac{e^{i\frac{1}{2}x}\left(\frac{e^{i(N+1)x}-1}{e^{ix}-1}\right) - e^{-i\frac{1}{2}x}\left(\frac{e^{-i(N+1)x}-1}{e^{-ix}-1}\right)}{e^{i\frac{1}{2}x} - e^{-i\frac{1}{2}x}} \\ &= \frac{1}{N+1}\frac{e^{i(N+1)x} - 1 + e^{-i(N+1)x} - 1}{(e^{i\frac{1}{2}x} - e^{-i\frac{1}{2}x})^2} \\ &= \frac{1}{N+1}\frac{\sin^2(\frac{N+1}{2}x)}{\sin^2(\frac{1}{2}x)}. \end{aligned}$$

EXERCISE 1.4.8. Verify that the Fejér kernel is an approximate identity as in Definition 1.3.9.

Our previous sections concerning how the partial sum $S_m f \equiv D_m * f$ converges to f in different senses. Since the Dirichlet kernel D_m takes negative values, it is not an approximate identity (Definition 1.3.9). However, using the summation method, we obtain an approximate kernel. In other words, the Cesàro sums "regularize" the kernel. Therefore using Lemma 1.3.11 we conclude the following theorem.

THEOREM 1.4.9 (Cesàro summability of Fourier series). *Let f be a 2π-periodic function in $\mathbb{R}$. If $f \in L^p(-\pi,\pi)$ for some $1 \le p < \infty$, then*

$$\lim_{N\to\infty} \|\sigma_N f - f\|_{L^p(-\pi,\pi)} = 0.$$

If f is continuous, then

$$\lim_{N\to\infty} \|\sigma_N f - f\|_{L^\infty(-\pi,\pi)} = 0. \tag{1.4.1}$$

Finally, we see that the uniform convergence (1.4.1) indicates that the lack of Gibbs-Wilbraham phenomena by considering the Cesàro sums $\sigma_N f$ of a continuous function f.

1.5. Infinite dimensional subspace: Sobolev embedding theorems

We begin by recalling some basic results from calculus. We now collect some definitions and facts about differentiation. For a one variable function u, its derivative at $t \in \mathbb{R}$ is (at least formally) defined by

$$u'(t) := \lim_{h\to 0} \frac{u(t+h)-u(t)}{h}.$$

Let e_j be the j^{th}-column of the $n\times n$ identity matrix, and the partial derivatives are (formally) defined by

$$\partial_j u(x) := \lim_{h\to 0} \frac{u(x+he_j)-u(x)}{h}.$$

We recall the following fundamental theorem:

LEMMA 1.5.1 (see e.g. [**Apo74**, Theorem 12.13]). *If both partial derivatives $\partial_{i_1} u$ and $\partial_{i_2} u$ exist* near *a point x_0 and if both $\partial_{i_1}\partial_{i_2} u$ and $\partial_{i_2}\partial_{i_1} u$ are continuous* at x_0, *then*

$$\partial_{i_1}\partial_{i_2} u(x_0) = \partial_{i_2}\partial_{i_1} u(x_0).$$

Therefore, we usually denote $\partial_{i_1 i_2} = \partial_{i_1}\partial_{i_2} = \partial_{i_2}\partial_{i_1}$, regardless the order of partial derivatives. We also exhibit some notations which are helpful to express higher order derivatives. For each multi-index $\alpha = (\alpha_1,\cdots,\alpha_n)$ with non-negative integers α_j, we define

$$\mathrm{supp}\,(\alpha) := \left\{ j \in \{1,\cdots,n\} : \alpha_j \neq 0 \right\}, \quad |\alpha| := \alpha_1 + \alpha_2 + \cdots + \alpha_n \equiv \sum_{j=1}^{n} \alpha_j,$$

as well as

$$\partial^\alpha := \prod_{j\in\mathrm{supp}\,(\alpha)} \partial_j^{\alpha_j} \quad \text{with the convention} \quad \partial^{(0,\cdots,0)} := \mathrm{Id},$$

and *the partial derivatives in ∂^α are pairwise commute*. In view of Lemma 1.5.1, for each open set Ω and a non-negative integer k, we define the spaces

$$\begin{aligned}
C^k(\Omega) &:= \{u : \Omega \to \mathbb{C} : \partial^\alpha u \text{ is continuous for all } \alpha \text{ with } |\alpha| \le k\}, \\
C^k(\overline{\Omega}) &:= \left\{ u|_{\overline{\Omega}} : u \in C^k(U) \text{ for some open set } U \supset \overline{\Omega} \right\}, \\
C^k_c(\Omega) &:= \left\{ u \in C^k(\Omega) : \mathrm{supp}\,(u) \subset \Omega \text{ is compact} \right\}.
\end{aligned}$$

By considering the zero extension, one also sees that

$$C^k_c(\Omega) = \left\{ u \in C^k(\mathbb{R}^n) : \mathrm{supp}\,(u) \subset \Omega \text{ is compact} \right\}. \tag{1.5.1}$$

Similarly, we also write

$$\begin{aligned}
C^\infty_c(\Omega) &:= \{u \in C^\infty(\Omega) : \mathrm{supp}\,(u) \subset \Omega \text{ is compact}\} \\
&= \{u \in C^\infty(\mathbb{R}^n) : \mathrm{supp}\,(u) \subset \Omega \text{ is compact}\}.
\end{aligned}$$

The following density result is fundamental:

LEMMA 1.5.2. *Let Ω be an open set in $\mathbb{R}^n$. Then $C^\infty_c(\Omega)$ is dense in $L^p(\Omega)$ for any $1 \le p < \infty$, that is, given any $f \in L^p(\Omega)$, there exists a sequence of functions $\{f_k\}_{k\in\mathbb{N}}$ in $C^\infty_c(\Omega)$ such that $f_k \to f$ in $L^p(\Omega)$.*

EXERCISE 1.5.3. Show that Lemma 1.5.2 does not hold true when $p = \infty$.

We now state the following proposition, which serves as the most important ingredient in this course:

PROPOSITION 1.5.4 (Divergence theorem, see e.g. [**Str08**, Appendix A.3]). *Let Ω be a bounded domain in $\mathbb{R}^n$ with a piecewise-C^1 boundary $\partial\Omega$. Let $\nu = (\nu_1, \cdots, \nu_n)$ be the unit outward normal vector on $\partial\Omega$, then*

$$\int_\Omega \partial_i u \, \mathrm{d}x = \int_{\partial\Omega} \nu_i u \, \mathrm{d}S_x \tag{1.5.2}$$

for all $u \in C^1(\overline{\Omega})$, where $\mathrm{d}S_x$ is the surface element (can be characterized in terms of Hausdorff measure) on $\partial\Omega$.

EXERCISE 1.5.5. Suppose that all assumptions in Proposition 1.5.4 holds. Show that

$$\int_\Omega \nabla \cdot u \, \mathrm{d}x = \int_{\partial\Omega} \nu \cdot u \, \mathrm{d}S_x \tag{1.5.3}$$

for all $u = (u_1, \cdots, u_n) \in (C^1(\overline{\Omega}))^n$, where $\nabla \cdot u(x) := \partial_1 f_1(x) + \cdots + \partial_n f_n(x)$ is called the *divergence* of f.

Using product rule, it is easy to see that

$$\int_{\partial\Omega} \nu_i uv \, \mathrm{d}S_x = \int_\Omega \partial_i(uv) \, \mathrm{d}x = \int_\Omega \partial_i u(x) v(x) \, \mathrm{d}x + \int_\Omega u(x) \partial_i v(x) \, \mathrm{d}x.$$

Hence we immediately reach the following useful corollary:

COROLLARY 1.5.6 (Integration by parts). *Suppose that all assumptions in Proposition 1.5.4 holds, then*

$$\int_\Omega \partial_i u(x) v(x) \, \mathrm{d}x = \int_{\partial\Omega} \nu_i(x) u(x) v(x) \, \mathrm{d}S_x - \int_\Omega u(x) \partial_i v(x) \, \mathrm{d}x$$

for all $u, v \in C^1(\overline{\Omega})$.

REMARK. We now restrict ourselves when $n = 1$. When $\Omega = (a,b)$, the above identity simply reads

$$\int_a^b u'(x) v(x) \, \mathrm{d}x = u(x)v(x)|_{x=a}^{x=b} - \int_a^b u(x) v'(x) \, \mathrm{d}x$$

which is just the usual integration by parts. In fact, each open set Ω can be written as union of countably many disjoint open intervals [**WZ15**], and therefore the above formula can be extended for arbitrary open sets in $\mathbb{R}^1$. The fundamental theorem of calculus is simply a special case of divergence theorem. Indeed Corollary 1.5.6 can be extended for general bounded Lipschitz domains and in weak sense (see Theorem 1.5.25 below).

EXERCISE 1.5.7 (Green's theorem as a special case of divergence theorem). Suppose that all assumptions in Proposition 1.5.4 holds with $n=2$. The Green's theorem stated that

$$\int_{\Omega} (\partial_x q(x,y) - \partial_y p(x,y)) \,\mathrm{d}x\,\mathrm{d}y = \int_{\partial D} (p\,\mathrm{d}x + q\,\mathrm{d}y),$$

for all $p,q \in C^1(\overline{\Omega})$, where the right-hand-side is the line integral defined by

$$\int_{\partial D} (p\,\mathrm{d}x + q\,\mathrm{d}y) := \int_{\partial D} (p(\gamma(s)), q(\gamma(s))) \cdot \boldsymbol{t}(s)\,\mathrm{d}s$$

where $\gamma(s)$ is the arc-lengh parametrization of ∂D and $\boldsymbol{t}(s)$ is the unit tangent vector field (usually chosen to be counterclockwise oriented). Show that Green's theorem is a special case of the divergence theorem.

The main theme of this section is the study of several important L^p-based function spaces. Because the objects under consideration are functions, it is natural to begin by examining their differentiability properties. Recall that an L^p function is defined only up to sets of measure zero. In other words, two functions that differ only on a measure zero set represent the same element of L^p. Consequently, pointwise notions such as differentiability, as introduced in elementary calculus, are not directly applicable in this setting. This motivates the introduction of the concept of a weak derivative, which generalizes the classical notion of differentiation taught in calculus. Let Ω be any open set in $\mathbb{R}^n$. By using integration by parts formula, one has

$$\int_{\Omega} (\partial_j u)\varphi\,\mathrm{d}x = -\int_{\Omega} u\partial_j\varphi\,\mathrm{d}x \quad \text{for all } u,\varphi \in C_c^1(\Omega),$$

where $C_c^k(\Omega)$ is given in (1.5.1). In fact, we have:

EXERCISE 1.5.8. Show that

$$\int_{\Omega} (\partial^{\alpha} u)\varphi\,\mathrm{d}x = (-1)^{|\alpha|}\int_{\Omega} f\partial^{\alpha}\varphi\,\mathrm{d}x \quad \text{for all } \varphi \in C_c^{\infty}(\Omega)$$

for all multi-indices α.

Therefore, it is quite natural to consider the following definition (and we can interpret the PDE using the following weak derivatives):

DEFINITION 1.5.9 (Weak derivatives). We define the locally-L^1 space by

$$L^1_{\mathrm{loc}}(\Omega) := \left\{ u \text{ defined on } \Omega : \|u\|_{L^1(K)} := \int_K |u(x)|\mathrm{d}x < \infty \text{ for all compact set } K \subset \Omega \right\}.$$

A function $v \in L^1_{\mathrm{loc}}(\Omega)$ (if exists) is called a *weak derivative* of $u \in L^1_{\mathrm{loc}}(\Omega)$ (of order α) if

$$\int_{\Omega} v\varphi\,\mathrm{d}x = (-1)^{|\alpha|}\int_{\Omega} u\partial^{\alpha}\varphi\,\mathrm{d}x \quad \text{for all } \varphi \in C_c^{\infty}(\Omega).$$

and we often denote v as $\partial^{\alpha} u$.

The well-definedness of the weak derivatives (i.e. each function $v \in L^1_{\rm loc}(\Omega)$ produced from $u \in L^1_{\rm loc}(\Omega)$ must be unique) is guaranteed by the following lemma:

LEMMA 1.5.10 (Uniqueness of weak derivatives [Mit18, Theorem 1.3]). *If $v \in L^1_{\rm loc}(\Omega)$ satisfying $v = 0$ in Ω in distribution sense, i.e.*

$$\int_\Omega v\varphi \,\mathrm{d}x = 0 \quad \text{for all } \varphi \in C_c^\infty(\Omega),$$

then $v = 0$ a.e. in Ω.

REMARK 1.5.11. The converse of Lemma 1.5.10 is trivial. Here and after, we shall omit the notation "a.e." if there is no any ambiguity. The above lemma only guarantee uniqueness, but not existence.

We now give a first example of weak derivatives.

EXAMPLE 1.5.12. We consider the Heaviside function

$$H(x) := \begin{cases} 1 & \text{for all } x > 0, \\ 0 & \text{for all } x \le 0. \end{cases} \tag{1.5.4}$$

It is easy to see that $H \in L^1_{\rm loc}(\mathbb{R})$. We define

$$f(x) := \begin{cases} x & \text{for all } x > 0, \\ 0 & \text{for all } x \le 0. \end{cases}$$

One can verify that

$$-\int_{\mathbb{R}} f(x)\varphi'(x)\,\mathrm{d}x = -\int_0^\infty x\varphi'(x)\,\mathrm{d}x = -x\varphi(x)|_{x=0}^{x\to\infty} + \int_0^\infty \varphi(x)\,\mathrm{d}x = \int_{\mathbb{R}} H(x)\varphi(x)\,\mathrm{d}x,$$

which shows that the Heaviside function (1.5.4) is the weak derivative of f (of order 1), and we simply denote $f' = H$.

However, not all $L^1_{\rm loc}(\Omega)$ function admits weak derivatives:

EXAMPLE 1.5.13. We now show that the weak derivative of the Heaviside function H given in (1.5.4) of order 1 does not exist. Suppose the contrary, that H has a weak derivative of order 1, says $g \in L^1_{\rm loc}(\mathbb{R})$. By Definition 1.5.9, we see that

$$\int_{-\infty}^\infty g(x)\varphi(x)\,\mathrm{d}x = -\int_{-\infty}^\infty H(x)\varphi'(x)\,\mathrm{d}x = -\int_0^\infty \varphi'(x)\,\mathrm{d}x = \varphi(0) \tag{1.5.5}$$

for all $\varphi \in C_c^\infty(\mathbb{R})$. Hence we know that

$$\int_{-\infty}^\infty g(x)\varphi(x)\,\mathrm{d}x = 0 \quad \text{for all } \varphi \in C_c^\infty(\mathbb{R}\setminus\{0\}).$$

By using Lemma 1.5.10 with $\mathbb{R}\setminus\{0\}$, we conclude that $g = 0$ a.e. in $\mathbb{R}$, and from (1.5.5) we have $\varphi(0) = 0$ for all $\varphi \in C_c^\infty(\mathbb{R})$, which is an obvious contradiction.

With weak derivative at hand, we now ready to introduce the following linear space.

DEFINITION 1.5.14. Let $\Omega \subset \mathbb{R}^n$ be an open set and let $p \in \mathbb{R}$ with $1 \leq p \leq \infty$. For each $m \in \mathbb{N}$, we define the Sobolev spaces

$$W^{m,p}(\Omega) = \{u \in L^p(\Omega) : \partial^\alpha u \in L^p(\Omega) \text{ for all } \alpha \text{ with } |\alpha| \leq m\}.$$

In fact, $W^{m,p}(\Omega)$ is a normed space with respect to the norm $\|\cdot\|_{W^{m,p}(\Omega)}$ given by

$$\|u\|_{W^{m,p}(\Omega)} = \left(\sum_{|\alpha|\leq m} \|\partial^\alpha u\|_{L^p(\Omega)}^p\right)^{\frac{1}{p}} \quad \text{for } 1 \leq p < \infty,$$
$$\|u\|_{W^{m,\infty}(\Omega)} = \max_{|\alpha|\leq m} \|\partial^\alpha u\|_{L^\infty(\Omega)}.$$

The elements in $W^{m,p}(\Omega)$ also called the *Sobolev functions*.

LEMMA 1.5.15 ([**AH09**, Theorem 7.2.3]). *Let $1 \leq p \leq \infty$, let $m \in \mathbb{N}$ and let Ω be an open set in $\mathbb{R}^n$. Then the Sobolev space $W^{m,p}(\Omega)$ is Banach.*

Similar to L^p-functions, Sobolev functions also can be approximated by smooth functions.

LEMMA 1.5.16 ([**AH09**, Theorem 7.3.1]). *Let Ω be an open set in $\mathbb{R}^n$, let $1 \leq p < \infty$ and let $m \in \mathbb{N}$. Given any $u \in W^{m,p}(\Omega)$, there exists a sequence $\{u_k\}_{k\in\mathbb{N}}$ in $C^\infty(\Omega) \cap W^{m,p}(\Omega)$ such that*

$$u_n \to u \text{ in } W^{m,p}(\Omega), \quad \text{that is,} \quad \lim_{k\to\infty} \|u_k - u\|_{W^{m,p}(\Omega)} = 0.$$

Note that in this lemma, the approximation functions $\{u_k\}$ are smooth *only in the interior* of Ω. To have the smoothness *up to the boundary* of the approximation sequence, we need to make a smoothness assumption on the boundary $\partial\Omega$. For each $k \in \mathbb{N} \cup \{\infty\}$, we define

$$C^k(\overline{\Omega}) := \left\{u|_{\overline{\Omega}} : u \in C^k(U) \text{ for some open set } U \supset \overline{\Omega}\right\}.$$

LEMMA 1.5.17 ([**AH09**, Theorem 7.3.2]). *Let Ω be a bounded Lipschitz domain in $\mathbb{R}^n$, let $1 \leq p < \infty$ and let $m \in \mathbb{N}$. Given any $u \in W^{m,p}(\Omega)$, there exists a sequence $\{u_k\}_{k\in\mathbb{N}}$ in $C^\infty(\overline{\Omega})$ such that*

$$u_n \to u \text{ in } W^{m,p}(\Omega), \quad \text{that is,} \quad \lim_{k\to\infty} \|u_k - u\|_{W^{m,p}(\Omega)} = 0.$$

Before introducing the Sobolev embeddings, we first introducing the following concept:

DEFINITION 1.5.18. Let X and Y be two Banach spaces. We say that the space X is *continuous embedded* in Y if

$$\|u\|_Y \leq c\|u\|_X \quad \text{for all } u \in X. \tag{1.5.6}$$

We say that the space X is *compactly embedded* in Y if (1.5.6) holds and each bounded sequence in X has a convergent subsequence in Y.

DEFINITION 1.5.19. Let X and Y be two Banach spaces. A linear operator $T : X \to Y$ is said to be *bounded* if there exists a constant $C > 0$ such that

$$\|Tu\|_Y \leq C\|u\|_X \quad \text{for all } u \in X.$$

A bounded linear operator $T : X \to Y$ is said to be compact if for any bounded sequence $\{x_j\}_{j\in\mathbb{N}}$ in X, the sequence $\{Tx_j\}_{j\in\mathbb{N}}$ contains a convergent subsequence in Y. The set of all compact operators from X to Y is denoted as $\mathscr{K}(X,Y)$. If $X = Y$, we simply write $\mathscr{K}(X) = \mathscr{K}(X,X)$.

The following fact is useful to check whether the bounded linear operator is compact or not.

LEMMA 1.5.20 ([**Bre11**, Proposition 6.3]). *Let X, Y and Z be three Banach spaces. Let $\mathscr{L}(X,Y)$ denotes the set of bounded linear operator from X to Y.*

(1) *If $T \in \mathscr{L}(X,Y)$ and $S \in \mathscr{K}(Y,Z)$, then $S \circ T \in \mathscr{K}(X,Z)$.*
(2) *If $T \in \mathscr{K}(X,Y)$ and $S \in \mathscr{L}(Y,Z)$, then $S \circ T \in \mathscr{K}(X,Z)$.*

We usually use the following form of Lemma 1.5.20: If $Y \Subset Z$, i.e. the embedding $\iota : Y \to Z$ is compact, then by identifying $T \cong \iota \circ T$, one sees that $T \in \mathscr{L}(X,Y)$ implies $T \in \mathscr{K}(X,Z)$. Similarly, if $X \Subset Y$, one sees that $S \in \mathscr{L}(Y,Z)$ implies $S \in \mathscr{K}(X,Z)$. Many authors (including myself) simply denote $X \subset Y$ if the Banach space X is continuous embedded in another Banach space Y, despite that X is not necessarily a subset of Y. We will also denote $X \Subset Y$ if X is compactly embedded in Y. Here and after (including the next theorem), we will use these notations without mentioning explicitly. Let $\lfloor x \rfloor$ denotes the integer part of x, and we have the following theorem:

THEOREM 1.5.21 (Sobolev embedding theorems [**AH09**, Theorem 7.3.7 and Theorem 7.3.8]). *Let Ω be a bounded Lipschitz domain in $\mathbb{R}^n$. Then the following statements are valid:*

(a) *If $k < \frac{n}{p}$, then $W^{k,p}(\Omega) \Subset L^q(\Omega)$ for any $q < p^*$ and $W^{k,p}(\Omega) \subset L^q(\Omega)$ when $q \leq p^*$, where $\frac{1}{p^*} = \frac{1}{p} - \frac{k}{n}$.*
(b) *If $k = \frac{n}{p}$, then $W^{k,p}(\Omega) \Subset L^q(\Omega)$ for any $q < \infty$.*
(c) *If $k > \frac{n}{p}$, then*

$$W^{k,p}(\Omega) \Subset C^{k-\lfloor \frac{n}{p} \rfloor -1,\beta}(\Omega) \quad \textit{for all } \beta \in \left[0, \left\lfloor \frac{n}{p} \right\rfloor + 1 - \frac{n}{p}\right) \tag{1.5.7}$$
$$W^{k,p}(\Omega) \subset C^{k-\lfloor \frac{n}{p} \rfloor -1,\beta}(\Omega) \quad \textit{with } \beta = \begin{cases} \lfloor \frac{n}{p} \rfloor + 1 - \frac{n}{p} & \textit{if } \frac{n}{p} \notin \mathbb{Z}, \\ \textit{any positive number } < 1 & \textit{if } \frac{n}{p} \in \mathbb{Z}. \end{cases}$$

REMARK 1.5.22. Theorem 1.5.21 is also valid for $W^{k,p}$-spaces with $k \in \mathbb{R}$, see e.g. [**AH09, McL00**] for precise definitions. Here we will cover these topics in this lecture note. Part (c) of Theorem 1.5.21 in particular gives some sufficient condition in terms of weak derivatives to guarantee the well-definedness of the strong/classical derivatives. It is also interesting to note that (1.5.7) can be viewed as a quantitative version of the Arzela-Ascoli theorem.

It is important to mention that the proof of Theorem 1.5.21 is based on the existence of the bounded linear extension operator

$$E : W^{k,p}(\Omega) \to W^{k,p}(\mathbb{R}^n)$$

for any nonnegative integer k and for any $1 \leq p \leq \infty$. In fact, the operator norm of the extension operator can be eplicitly given:

THEOREM 1.5.23 ([**Bur99**, Theorem 3.4]). *Let Ω be a bounded Lipschitz domain in $\mathbb{R}^n$. There exists a constant $C = C(\Omega) > 1$ such that*

$$(C^{-1}k)^k \leq \inf_E \|E\|_{W^{k,p}(\Omega)\to W^{k,p}(\mathbb{R}^n)} \leq (Ck)^k$$

for all $k \in \mathbb{N}$ and for all $1 \leq p \leq \infty$, where the infimum is taken over all extension operator $E : W^{k,p}(\Omega) \to W^{k,p}(\mathbb{R}^n)$.

REMARK 1.5.24. Here we emphasize that the constant C in Theorem 1.5.23 is independent of both k and p.

In fact, the integration by parts also holds true for weak derivatives (which generalized Corollary 1.5.6):

THEOREM 1.5.25 (Integration by parts [**EG15**, Theorem 1 in Section 4.3]). *Let Ω be a bounded Lipschitz domain in $\mathbb{R}^n$ and given $1 \leq p < \infty$. The mapping*

$$\mathrm{Tr} : C^\infty(\overline{\Omega}) \to C^\infty(\partial\Omega), \quad \mathrm{Tr}\,(u) = u|_{\partial\Omega} \tag{1.5.8}$$

can be uniquely extended to a bounded surjective linear operator $W^{1,p}(\Omega) \to \mathrm{Tr}\,(W^{1,p}(\Omega)) \subset L^p(\partial\Omega)$. Furthermore, for all $\varphi \in (C^1(\mathbb{R}^n))^n$ and $u \in W^{1,p}(\Omega)$, we have

$$\int_\Omega u \mathrm{div}\,(\varphi)\,\mathrm{d}x = -\int_\Omega \nabla u \cdot \varphi\,\mathrm{d}x + \int_{\partial\Omega} (\nu\cdot\varphi)\mathrm{Tr}\,(u)\,\mathrm{d}\mathscr{H}^{n-1}, \tag{1.5.9}$$

where ν is the unit outer normal to $\partial\Omega$.

REMARK 1.5.26. Here we refer the advance monograph [**EG15**] for the precise meaning of ν, which is well-defined for $\mathscr{H}^{n-1}$-a.e. on $\partial\Omega$. The function $\mathrm{Tr}(f)$ given in (1.5.8) is called the *trace* of f on $\partial\Omega$. We usually still denote $\mathrm{d}\mathscr{H}^{n-1}$ by $\mathrm{d}S_x$. If there is no ambiguity, we sometime omit the notation the trace operator (1.5.8) and simply write (1.5.9) as

$$\int_\Omega u \mathrm{div}\,(\varphi)\,\mathrm{d}x = -\int_\Omega \nabla u \cdot \varphi\,\mathrm{d}x + \int_{\partial\Omega} (\nu\cdot\varphi) u\,\mathrm{d}S_x.$$

In view of Exercise 1.2.20, it is natural to consider the following definition:

DEFINITION 1.5.27. Let Ω be any open set in $\mathbb{R}^n$, then we denote $H^m(\Omega) := W^{m,2}(\Omega)$ for each $m \in \mathbb{N}$. In this case, the norm reads

$$\|u\|_{H^m(\Omega)} = \left(\sum_{|\alpha| \leq m} \|\partial^\alpha u\|_{L^2(\Omega)}^2 \right)^{\frac{1}{2}}.$$

EXERCISE 1.5.28. Use Exercise 1.2.19 to show that the corresponding scalar product is given by

$$(u,v)_{H^m(\Omega)} = \sum_{|\alpha| \leq m} (\partial^\alpha u, \partial^\alpha v)_{L^2(\Omega)}.$$

By introducing the gradient $\nabla u(x) = (\partial_1 u(x), \cdots, \partial_n u(x))$, we see that

$$\begin{aligned} \int_\Omega \nabla u(x) \cdot \nabla v(x) \, \mathrm{d}x &= \int_\Omega \sum_{i=1}^n \partial_i u(x) \partial_i v(x) \, \mathrm{d}x \\ &= \sum_{i=1}^n \int_\Omega \partial_i u(x) \partial_i v(x) \, \mathrm{d}x = \sum_{i=1}^n (\partial_i u, \partial_i v)_{L^2(\Omega)} \end{aligned}$$

Therefore it is convenient to define

$$(\nabla u, \nabla v)_{L^2(\Omega)} := \int_\Omega \nabla u(x) \cdot \nabla v(x) \, \mathrm{d}x, \quad \|\nabla u\|_{L^2(\Omega)} := \sqrt{(\nabla u, \nabla u)_{L^2(\Omega)}},$$

so the scalar product and norm on $H^1(\Omega)$ can be expressed as

$$(u,v)_{H^1(\Omega)} = (u,v)_{L^2(\Omega)} + (\nabla u, \nabla v)_{L^2(\Omega)}, \quad \|u\|_{H^1(\Omega)} = \left(\|u\|_{L^2(\Omega)}^2 + \|\nabla u\|_{L^2(\Omega)}^2 \right)^{\frac{1}{2}}.$$

In view of (1.1.9), we introduce the Hessian matrix $\nabla^{\otimes 2} u(x) \equiv \nabla \otimes \nabla u(x)$ with entries $\left(\nabla^{\otimes 2} u(x)\right)_{ij} = \partial_i \partial_j u(x)$. We see that

$$\begin{aligned} \int_\Omega \nabla^{\otimes 2} u(x) : \nabla^{\otimes 2} v(x) \, \mathrm{d}x &= \int_\Omega \sum_{i,j=1}^n \partial_i \partial_j u(x) \partial_i \partial_j v(x) \, \mathrm{d}x \\ &= \sum_{i,j=1}^n \int_\Omega \partial_i \partial_j u(x) \partial_i \partial_j v(x) \, \mathrm{d}x = \sum_{i,j=1}^n (\partial_i \partial_j u, \partial_i \partial_j v)_{L^2(\Omega)}. \end{aligned}$$

Therefore it is convenient to define

$$(\nabla^{\otimes 2} u, \nabla^{\otimes 2} v)_{L^2(\Omega)} := \int_\Omega \nabla^{\otimes 2}(x) : \nabla^{\otimes 2} v(x) \, \mathrm{d}x, \quad \|\nabla^{\otimes 2} u\|_{L^2(\Omega)} := \sqrt{(\nabla^{\otimes 2} u, \nabla^{\otimes 2} u)_{L^2(\Omega)}},$$

and

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

$$\|u\|_{H^2(\Omega)} = \left(\|u\|_{L^2(\Omega)}^2 + \|\nabla u\|_{L^2(\Omega)}^2 + \|\nabla^{\otimes 2} u\|_{L^2(\Omega)}^2 \right)^{\frac{1}{2}}. \tag{1.5.10}$$

The $\|\cdot\|_{H^2(\Omega)}$-norm given in (1.5.10) is actually equivalent to the $\|\cdot\|_{H^2(\Omega)}$-norm given in Definition 1.5.27 in the following sense:

DEFINITION 1.5.29. Let $\|\cdot\|_1$ and $\|\cdot\|_2$ are norms on the vector space X. We say that $\|\cdot\|_1$ and $\|\cdot\|_2$ are equivalent if there exists a constant $c>0$ such that

$$c^{-1}\|u\|_1 \leq \|u\|_2 \leq c\|u\|_1 \quad \text{for all } u \in X.$$

Similarly, by introducing the k-tensor $\nabla^{\otimes k}u(x)$ with entries $\left(\nabla^{\otimes k}u(x)\right)_{i_1 i_2 \cdots i_k} = \partial_{i_1}\partial_{i_2}\cdots\partial_{i_k}u(x)$, we see that

$$\begin{aligned}\int_\Omega \nabla^{\otimes k}u(x) \overset{(k)}{\bullet} \nabla^{\otimes k}v(x)\,\mathrm{d}x &= \int_\Omega \sum_{i_1,\cdots,i_k=1}^{n} \partial_{i_1}\partial_{i_2}\cdots\partial_{i_k}u(x)\partial_{i_1}\partial_{i_2}\cdots\partial_{i_k}v(x)\,\mathrm{d}x \\ &= \sum_{i_1,\cdots,i_k=1}^{n} \partial_{i_1}\partial_{i_2}\cdots\partial_{i_k}u(x)\partial_{i_1}\partial_{i_2}\cdots\partial_{i_k}v(x)\,\mathrm{d}x \\ &= \sum_{i_1,\cdots,i_k=1}^{n} (\partial_{i_1}\partial_{i_2}\cdots\partial_{i_k}u, \partial_{i_1}\partial_{i_2}\cdots\partial_{i_k}v)_{L^2(\Omega)}.\end{aligned}$$

It is convenient to define $\nabla^{\otimes 0}u := u$, and for each $k \in \mathbb{N}$ that

$$(\nabla^{\otimes k}u, \nabla^{\otimes k}v)_{L^2(\Omega)} := \int_\Omega \nabla^{\otimes k}(x) \overset{(k)}{\bullet} \nabla^{\otimes k}v(x)\,\mathrm{d}x, \quad \|\nabla^{\otimes k}u\|_{L^2(\Omega)} := \sqrt{(\nabla^{\otimes k}u, \nabla^{\otimes k}u)_{L^2(\Omega)}}.$$

We define the scalar products and norm by

$$(u,v)_{H^m(\Omega)} = \sum_{k=0}^{m} (\nabla^{\otimes k}u, \nabla^{\otimes k}v)_{L^2(\Omega)}, \quad \|u\|_{H^m(\Omega)} = \left(\sum_{k=0}^{m} \|\nabla^{\otimes k}u\|_{L^2(\Omega)}^2\right)^{\frac{1}{2}}. \tag{1.5.11}$$

The $\|\cdot\|_{H^m(\Omega)}$-norm given in (1.5.11) is actually equivalent to the $\|\cdot\|_{H^m(\Omega)}$-norm given in Definition 1.5.27. We recall the following fact, which already mentioned in Theorem 1.5.25:

THEOREM 1.5.30 (Trace theorem). *Let Ω be a bounded domain in $\mathbb{R}^n$ with $C^{0,1}$ boundary $\partial\Omega$, i.e. Ω is a bounded Lipschitz domain in $\mathbb{R}^n$. The mapping*

$$\mathrm{Tr} : C^\infty(\overline{\Omega}) \to C^\infty(\partial\Omega), \quad \mathrm{Tr}\,(u) := u|_{\partial\Omega}$$

extends to a unique bounded linear surjective mapping $H^1(\Omega) \to H^{\frac{1}{2}}(\partial\Omega)$, where $H^{\frac{1}{2}}(\partial\Omega) := \mathrm{Tr}\,(H^1(\Omega)) \subset L^2(\partial\Omega)$, which is a Hilbert space equipped with the quotient norm

$$\|g\|_{H^{\frac{1}{2}}(\partial\Omega)} = \inf_{u\in H^1(\Omega), \mathrm{Tr}\,(u)=g} \|u\|_{H^1(\Omega)}.$$

It is also possible to define the "traces" and "normal derivatives" on $\partial\Omega$ for H^m-functions (see Theorem 1.5.25), see also [**LM72**, Theorem 9.4] for similar results for higher order derivatives:

THEOREM 1.5.31 (Trace theorem, a special case of [**AH09**, Theorem 7.3.11]). *Let $m \in \mathbb{Z}_{\geq 2}$ and let Ω be a bounded $C^{m-1,1}$ domain in $\mathbb{R}^n$. The mapping*

$$u \in C^\infty(\overline{\Omega}) \mapsto (u|_{\partial\Omega}, \partial_\nu u|_{\partial\Omega}) \in C^\infty(\partial\Omega) \times C^\infty(\partial\Omega),$$

where $\partial_\nu u := \nu \cdot \nabla u$, extends to a unique bounded linear surjective mapping $H^m(\Omega) \to H^{m-\frac{1}{2}}(\partial\Omega) \times H^{m-\frac{3}{2}}(\partial\Omega)$, where for each $k = 1, 2, \cdots, m$ the space $H^{k-\frac{1}{2}}(\partial\Omega) := \mathrm{Tr}\,(H^k(\Omega)) \subset L^2(\partial\Omega)$, which is a Hilbert space equipped with the quotient norm

$$\|g\|_{H^{k-\frac{1}{2}}(\partial\Omega)} = \inf_{u \in H^k(\Omega), \mathrm{Tr}\,(u) = g} \|u\|_{H^k(\Omega)}.$$

The following is an immediate corollary of Lemma 1.5.17:

COROLLARY 1.5.32 (Density). *Let Ω be a bounded $C^{m-1,1}$ domain in $\mathbb{R}^n$ and let $m \in \mathbb{N}$. Given any $u \in H^m(\Omega)$, there exists a sequence $\{u_k\}_{k\in\mathbb{N}}$ in $C^\infty(\overline{\Omega})$ such that*

$$u_n \to u \text{ in } H^m(\Omega), \quad \text{that is,} \quad \lim_{k\to\infty} \|u_k - u\|_{H^m(\Omega)} = 0.$$

In view of Corollary 1.5.32, it is natural to consider the following subspace of $H^m(\Omega)$:

DEFINITION 1.5.33. For each $m \in \mathbb{N}$, we define $H_0^m(\Omega)$ be the closure of $C_c^\infty(\Omega)$ with respect to the norm $\|\cdot\|_{H^m(\Omega)}$.

The relation between $H_0^m(\Omega)$ and $H^m(\Omega)$ are given in the followings:

LEMMA 1.5.34. *If Ω is a bounded Lipschitz domain in $\mathbb{R}^n$, then*

$$H_0^1(\Omega) = \left\{ u \in H^1(\Omega) : u|_{\partial\Omega} = 0 \right\}.$$

If Ω is a bounded $C^{1,1}$ domain in $\mathbb{R}^n$, then

$$H_0^2(\Omega) = \left\{ u \in H^1(\Omega) : u|_{\partial\Omega} = \partial_\nu u|_{\partial\Omega} = 0 \right\}.$$

REMARK. The subscript 0 in $H_0^m(\Omega)$ means the zero boundary value. Therefore here we denote $C_c^\infty(\Omega)$ rather than $C_0^\infty(\Omega)$, which also used by many authors, to avoid confusion. Since $H_0^1(\Omega) \cap H^2(\Omega) = \left\{ u \in H^2(\Omega) : u|_{\partial\Omega} = 0 \right\}$ for any bounded $C^{1,1}$ domain $\Omega \subset \mathbb{R}^n$, then $H_0^1(\Omega) \cap H^2(\Omega) \neq H_0^2(\Omega)$.

CHAPTER 2

Characteristics of linear operators: Singular value decomposition

As mentioned in Sections 1.1 and 1.2, a linear operator can be represented by a matrix. We now turn to the study of its properties.

2.1. Determinants

Here we recall the formal definition of determinant. A geometric interpretation can be found in [**Tre26**, Chapter 3].

DEFINITION 2.1.1. Given $A \in \mathbb{C}^{n\times n}$, and let $\mathbf{v}_1, \cdots, \mathbf{v}_n$ be column vectors of A, and we denote

$$A = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{pmatrix}, \quad \mathbf{v}_k = \begin{pmatrix} A_{1k} \\ \vdots \\ A_{nk} \end{pmatrix}.$$

The determinant $\det(A)$ is defined by $\det(A) := \mathsf{D}(\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n)$ such that

(1) D **is a multilinear form**, that is,

$$\mathsf{D}(\mathbf{v}_1, \cdots, \alpha\mathbf{u}_k + \beta\mathbf{v}_k, \cdots, \mathbf{v}_n) = \alpha\mathsf{D}(\mathbf{v}_1, \cdots, \mathbf{u}_k, \cdots, \mathbf{v}_n) + \beta\mathsf{D}(\mathbf{v}_1, \cdots, \mathbf{v}_k, \cdots, \mathbf{v}_n)$$

for all $\alpha, \beta \in \mathbb{C}$.

(2) D **is antisymmetric**, that is, if one interchanges two columns, the determinant changes sign: $\mathsf{D}(\cdots, \mathbf{v}_j, \cdots, \mathbf{v}_k, \cdots) = -\mathsf{D}(\cdots, \mathbf{v}_k, \cdots, \mathbf{v}_j, \cdots)$.

(3) $\det(\mathrm{Id}) = 1$, where Id is $n\times n$ identity matrix given in (1.1.10).

We now exhibit an explicit formula for the determinant. An *n-permutation* is a bijection from $\{1, \cdots, n\}$ onto itself. We denoted by $\mathfrak{S}_n$ the set of n-permutation. To avoid unnecessary technicalities, we give an *informal* definition of the sign of a permutation (see, e.g., the classical monograph [**Hun80**] for detailed explanations). For $\sigma \in \mathfrak{S}_n$, the *sign* of σ, denoted by $\mathrm{sign}(\sigma)$, is defined to be 1 if and even number of interchanges is required to rearrange the ordered tuple $(1, 2, \cdots, n)$ into $(\sigma(1), \sigma(2), \cdots, \sigma(n))$, and -1 if the number of interchanges is odd. A permutation with $\mathrm{sign}(\sigma) = 1$ is called an *even permutation*, while a permutation with $\mathrm{sign}(\sigma) = -1$ is called an *odd permutation*. With this notation at hand, the determinant can be written explicitly as:

$$\det A = \sum_{\sigma \in \mathfrak{S}_n} \mathrm{sign}(\sigma) A_{\sigma(1),1} A_{\sigma(2),2} \cdots A_{\sigma(n),n}.$$

A complete discussion of these properties would take us too far afield, and they are typically covered in a standard linear algebra course. The interested reader may consult the monograph [**Ber09**] for further details. We close this section by the following fundamental facts:

THEOREM 2.1.2. *For any* $A, B \in \mathbb{C}^{n\times n}$, *we have* $\det(AB) = (\det A)(\det B)$.

THEOREM 2.1.3. $A \in \mathbb{C}^{n\times n}$ *is invertible if and only if* $\det(A) \neq 0$. *In this case,* $\det(A^{-1}) = (\det A)^{-1}$.

As a direct consequence of the previous two theorems, the determinant is preserved under conjugation:

$$\det(QAQ^{-1}) = \det(A)$$

for all invertible matrix $Q \in \mathbb{C}^{n\times n}$. This also suggests us to introduce the following definition:

DEFINITION 2.1.4. The matrices $A, B \in \mathbb{C}^{n\times n}$ are called *similar* if there exist an invertible matrix $Q \in \mathbb{C}^{n\times n}$ such that $A = QBQ^{-1}$.

2.2. Spectral theorem for general matrices

We now introduce several important characteristics of matrices:

DEFINITION 2.2.1. Let X be a linear space. A scalar $\lambda \in \mathbb{C}$ is called an eigenvalue of an operator $T : X \to X$ if there exists a nontrivial vector $u \in X$ such that $Tu = \lambda u$, that is,

$$u \in \ker(T - \lambda \mathrm{Id}) := \{u \in X : (T - \lambda \mathrm{Id})u = 0\}.$$

The linear space $\ker(T - \lambda \mathrm{Id})$ is called the *kernel* (also known as *null space*). The set of all eigenvalues of $T : X \to X$ is called spectrum, and usually denoted $\sigma(T) \equiv \sigma(T : X \to X)$.

REMARK 2.2.2. The set $\sigma(T) \equiv \sigma(T : X \to X)$ depends on the choice of linear space X.

We begin with the finite dimensional setting. As discussed in Section 1.1, *after fixing a set of basis*, it suffices to consider the following definition.

DEFINITION 2.2.3. A scalar $\lambda \in \mathbb{C}$ is called an eigenvalue of $A \in \mathbb{C}^{n\times n}$ if there exists a nontrivial vector $\mathbf{v} \in \mathbb{C}^n$ such that $A\mathbf{v} = \lambda \mathbf{v}$, that is,

$$\ker(A - \lambda \mathrm{Id}) := \{\mathbf{v} \in \mathbb{C}^n : (A - \lambda \mathrm{Id})\mathbf{v} = 0\}.$$

In this case, the kernel $\ker(A - \lambda \mathrm{Id})$ is also called the *eigenspace*.

In a typical linear algebra course, the point made in Remark 2.2.2 is often left implicit, since the choice of basis is usually fixed from the outset when a matrix representation is introduced. However, the matrix representing a given linear operator depends on the chosen basis. Consequently, if different bases are used, the corresponding matrix representations will differ, and so will the coordinate representations of their eigenspaces. The eigenvalues themselves, however, remain unchanged.

We note the following important corollary of Theorem 2.1.3:

COROLLARY 2.2.4. *Let $A \in \mathbb{C}^{n\times n}$. Then the following are equivalent:*

(1) $\lambda \in \sigma(A)$,
(2) $\dim\ker(A-\lambda\mathrm{Id}) \in \{1,\cdots,n\}$,
(3) $\det(A-\lambda I)=0$.

The space $\ker(A-\lambda\mathrm{Id})$ is often interpreted *geometrically* as a, for example, a line, plane, or hyperplane. For this reason:

DEFINITION 2.2.5. Its dimension $\dim\ker(A-\lambda\mathrm{Id}) \in \{1,\cdots,n\}$ is called the *geometric multiplicity*.

Note that $\det(A-\lambda I)$ is a polynomial in λ of degree n. Therefore, by the fundamental theorem of algebra (see, e.g., my lecture notes [**Kow23**] for a proof based on Liouville's theorem), there exists $\lambda_1,\cdots,\lambda_n \in \mathbb{C}$ such that

$$\det(A-\lambda I) = (-1)^n \prod_{k=1}^{n}(\lambda-\lambda_k) \equiv (-1)^n(\lambda-\lambda_1)(\lambda-\lambda_2)\cdots(\lambda-\lambda_n). \tag{2.2.1}$$

The numbers $\lambda_1,\cdots,\lambda_n$ may contain repetitions; that is, the eigenvalues are not necessarily distinct.

DEFINITION 2.2.6. The polynomial (2.2.1) in λ is called the *characteristic polynomial* of A.

Therefore, there exist distinct eigenvalues $\tilde{\lambda}_1,\cdots,\tilde{\lambda}_\ell \in \{\lambda_1,\cdots,\lambda_n\}$ and $m_1,\cdots,m_\ell \in \mathbb{N}$ such that

$$\det(A-\lambda I) = (-1)^n \prod_{k=1}^{\ell}(\lambda-\tilde{\lambda}_k)^{m_k} \quad \text{with } \sum_{k=1}^{\ell} m_\ell = n. \tag{2.2.2}$$

The integers $m_1,\cdots,m_\ell \in \mathbb{N}$ arise from the *algebraic* factorization of the characteristic polynomial (2.2.1). This motivates the following terminology.

DEFINITION 2.2.7. The integer m_k is called the *algebraic multiplicity* of the eigenvalue $\tilde{\lambda}_k$.

The following theorem is usually covered in an elementary linear algebra course, so we shall not discuss its proof here.

THEOREM 2.2.8 (Spectral theorem [**Tre26**, Theorems 2.1 and 2.8]). *The following are equivalent for $A \in \mathbb{C}^{n\times n}$:*

(a) *A is similar to diagonal matrix $D = \mathrm{diag}(\lambda_1,\cdots,\lambda_n)$, where $\lambda_1,\cdots,\lambda_n \in \mathbb{C}$ are given in* (2.2.1).
(b) $\mathbb{C}^n = \bigoplus_{k=1}^n \ker(A-\lambda_k\mathrm{Id})$, *where $\lambda_1,\cdots,\lambda_n \in \mathbb{C}$ are given in* (2.2.1).

(c) $\dim\ker(A-\tilde{\lambda}_k\mathrm{Id}) = m_k$ *for all* $k=1,\cdots,\ell$, *where* $\tilde{\lambda}_1,\cdots,\tilde{\lambda}_\ell \in \{\lambda_1,\cdots,\lambda_n\}$ *are given in* (2.2.2).

In more concrete terms, condition (a) states that A is diagonalizable. Condition (b) asserts that the eigenvectors of A form a basis of $\mathbb{C}^n$, while condition (c) states that the geometric multiplicity of each eigenvalue coincides with its algebraic multiplicity. Unfortunately, not every matrix is diagonalizable, as the following exercise illustrates:

EXERCISE 2.2.9. Show that the Jordan block $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ is not diagonalizable in the sense of Theorem 2.2.8.

One might want to overcome this difficulty based on the following lemma:

LEMMA 2.2.10 ([**HS99**, Lemma IV-1-4]). *Given any* $A \in \mathbb{C}^{n\times n}$, *there exists a sequence of diagonalizable matrix* B_k *which is converges to* A.

Before presenting another approach, we first recall the following definition:

DEFINITION 2.2.11. A matrix $N \in \mathbb{C}^{n\times n}$ is said to be *nilpotent* if $N^n = 0$.

Instead, we have the following theorem:

THEOREM 2.2.12 (Jordan-Chevalley decomposition [**HS99**, Theorem IV-1-11]). *Let* $A \in \mathbb{C}^{n\times n}$. *Then there exist a diagonalizable matrix* $D \in \mathbb{C}^{n\times n}$ *and a nilpotent matrix* $N \in \mathbb{C}^{n\times n}$ *such that*

$$A = D + N \quad \text{and} \quad DN = ND, \tag{2.2.3}$$

and the decomposition (2.2.3) *is unique. If* $A \in \mathbb{R}^{n\times n}$, *then* $D \in \mathbb{R}^{n\times n}$ *and* $N \in \mathbb{C}^{n\times n}$.

The Jordan-Chevalley decomposition can be viewed as a direct consequence of the Jordan canonical form. We now describe a general procedure for constructing the decomposition in Theorem 2.2.12 is shown as below:

Algorithm 1 Computation of D and N in Theorem 2.2.12

1: Input a matrix $A \in \mathbb{C}^{n\times n}$.
2: Compute the characteristic polynomial $p(z) = \det(zI - A)$.
3: Decompose $1/p(z)$ into partial fractions

$$\frac{1}{p(z)} = \sum_{j=1}^{k} \frac{Q_j(z)}{(z-\lambda_j)^{m_j}},$$

where for each j the quantity $Q_j(z)$ is a nonzero polynomial with $\deg(Q_j) \leq m_j - 1$ and $\lambda_1, \cdots, \lambda_k$ are distinct zeros of p (i.e. eigenvalues of A).
4: For each $j = 1, \cdots, n$, we define

$$P_j(A) := Q_j(A) \prod_{\ell \neq j} (A - \lambda_j I)^{m_\ell}.$$

5: Output $D = \lambda_1 P_1(A) + \cdots + \lambda_k P_k(A)$ and $N = A - D$.

In practice, Algorithm 1 is rarely implemented directly because of its high computational cost. The method based on Lemma 2.2.10 is also numerically unstable, as illustrated by the following example[1]:

EXAMPLE 2.2.13. For each $\varepsilon \geq 0$, we define

$$A_\varepsilon = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ \varepsilon & 0 & 0 & 1 \end{pmatrix}.$$

Note that A_0 is a Jordan block and therefore not diagonalizable. In contrast, A_ε is diagonalizable for all $\varepsilon > 0$:

$$A_\varepsilon = Q_\varepsilon D_\varepsilon Q_\varepsilon^{-1}$$

with $D_\varepsilon = \mathrm{diag}\,(1-\varepsilon^{1/4}, 1-\mathrm{i}\varepsilon^{1/4}, 1+\mathrm{i}\varepsilon^{1/4}, 1+\varepsilon^{1/4})$ and

$$Q_\varepsilon = \begin{pmatrix} -\varepsilon^{-3/4} & -\mathrm{i}\varepsilon^{-3/4} & \mathrm{i}\varepsilon^{-3/4} & \varepsilon^{-3/4} \\ \varepsilon^{-1/2} & -\varepsilon^{-1/2} & -\varepsilon^{-1/2} & \varepsilon^{-1/2} \\ -\varepsilon^{-1/4} & \mathrm{i}\varepsilon^{-1/4} & -\mathrm{i}\varepsilon^{-1/4} & \varepsilon^{-1/4} \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

(1) If $\varepsilon = 0.01$, then $\varepsilon^{1/4} \approx 0.3162$, meaning that the eigenvalue of the Jordan block A_0 is perturbed by approximately $\approx 32\%$.
(2) If $\varepsilon = 0.05$, then $\varepsilon^{1/4} \approx 0.4729$, meaning that the eigenvalue of the Jordan block A_0 is perturbed by approximately $\approx 47\%$.

[1] I would like to thank Yueh-Cheng Kuo for pointing this out.

In particular, if we consider $A_\varepsilon \in \mathbb{R}^{n\times n}$ defined by

$$(A_\varepsilon)_{ij} = \begin{cases} 1 & \text{if } i = j \text{ or } j+1 = j \\ \varepsilon & \text{if } i = n \text{ and } j = 1, \\ 0 & \text{otherwise,} \end{cases}$$

its eigenvalues are $1 + \xi_m \varepsilon^{1/n}$, where $\xi_1, \cdots, \xi_n \in \mathbb{C}$ are roots of $\xi^n = 1$, see [**TB97**, (39.5)]. For $n = 50$:

(1) If $\varepsilon = 0.01$, then $\varepsilon^{1/50} \approx 0.9120$, meaning that the eigenvalue of the Jordan block A_0 is perturbed by approximately $\approx 91\%$.
(2) If $\varepsilon = 10^{-8}$ (corresponding to the commonly used single-precision (32-bit)), then $\varepsilon^{1/50} \approx 0.6918$, meaning that the eigenvalue of the Jordan block A_0 is perturbed by approximately $\approx 69\%$.
(3) If $\varepsilon = 10^{-16}$ (corresponding to the commonly used double-precision (64-bit)), then $\varepsilon^{1/50} \approx 0.4786$, meaning that the eigenvalue of the Jordan block A_0 is perturbed by approximately $\approx 48\%$.
(4) If one requires that the eigenvalue of the Jordan block A_0 be perturbed by no more than 10%, that is, $\varepsilon^{1/50} \leq 0.1$, then we must have $\varepsilon \leq 10^{-50}$. This, in turn, necessitates the use of multiprecision arithmetic, which is computationally expensive, see [**FI07**].

In practice, one often encounters matrices of size $n = \mathcal{O}(10^3)$, or even $n = \mathcal{O}(10^6)$, for example when solving partial differential equations using the finite element method (FEM). In such regimes, the approach in Lemma 2.2.10 is not numerically feasible.

In particular, Example 2.2.13 is a special case of Lidskii's perturbation theory [**Lid66**], see also [**MBO97**]. Even if one assumes a priori that the matrix is invertible, algorithms based directly on Theorem 2.2.8 are rarely used in numerical computations for general matrices. This is because evaluating determinants, characteristic polynomials, and matrix inverses can be both computationally expensive and numerically unstable. However, if A has additional structure, it may admit more efficient decompositions, as shown below.

THEOREM 2.2.14 ([**Tre26**, Theorem 2.4]). *If $A \in \mathbb{C}^{n\times n}$ is normal, that is, $A^*A = AA^*$, then it is unitary equivalent to a diagonal matrix, that is,*

$$A = UDU^* \tag{2.2.4}$$

for some diagonal matrix $D = \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$ and a unitary matrix U, that is, $U^{-1} = U^*$.

At least for now, we do not need to explicitly compute inverse matrices. In particular, in practice, we often encounter the following special situation:

THEOREM 2.2.15 ([**Tre26**, Theorems 2.2 and 4.3]). *If $A \in \mathbb{C}^{n\times n}$ is Hermitian (or called self-adjoint), that is, $A = A^*$, then its eigenvalues $\lambda_1, \cdots, \lambda_n$ are real. Furthermore, if $A \in \mathbb{R}^{n\times n}$, then U*

can be chosen to be real. Ordering the eigenvalues as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, we obtain the following Courant max-min *and* min-max *characterizations:*

$$\lambda_k = \max_{\substack{X \leq \mathbb{C}^n \\ \dim X = k}} \min_{\mathbf{v}\in X, \|\mathbf{v}\|=1} (A\mathbf{v},\mathbf{v}) = \min_{\substack{X \leq \mathbb{C}^n \\ \dim X = k-1}} \max_{\mathbf{u}\perp X, \|\mathbf{u}\|=1} (A\mathbf{u},\mathbf{u}). \tag{2.2.5}$$

Here, the notation $X \leq \mathbb{C}^n$ means that X is a linear subspace of $\mathbb{C}^n$, and the notation $\mathbf{u} \perp X$ means $\mathbf{u}^*\mathbf{w} = 0$ for all $\mathbf{w} \in X$.

EXERCISE 2.2.16. Let $A \in \mathbb{C}^{n\times n}$ be a Hermitian matrix. Show that (2.2.4) is equivalent to

$$A\mathbf{x} = \sum_{k=1}^{m} \lambda_k (\mathbf{x},\mathbf{u}_k)_{\mathbb{C}^n} \mathbf{u}_k \quad \text{for all } \mathbf{x} \in \mathbb{C}^n. \tag{2.2.6}$$

EXERCISE 2.2.17. Let $A \in \mathbb{R}^{n\times n}$. Using *the method of Lagrange multipliers*, show that

$$\lambda_1 = \max_{\mathbf{u}\in\mathbb{R}^n, \|\mathbf{u}\|=1} (A\mathbf{u},\mathbf{u}),$$

and

$$\lambda_k = \max_{\substack{\mathbb{R}^n \ni \mathbf{u} \perp \operatorname{span}\{\mathbf{u}_1,\cdots,\mathbf{u}_{k-1}\} \\ \|\mathbf{u}\|=1}} (A\mathbf{u},\mathbf{u}) \quad \text{for all } k = 2,\cdots,n.$$

While many elementary methods (including Exercise 2.2.17) suffice for “paper-and-pencil” computations, practical numerical linear algebra must account for machine precision and therefore relies on more refined and stable algorithms to ensure trustworthy results. One important class of such methods is the Krylov subspace methods, which are widely used for computing eigenvalues of large matrices [**GVL13, TB97**]. Their main advantage is that they avoid explicit matrix operations and require only matrix-vector products, which improves efficiency and numerical stability. We shall not discuss these topics in detail here, instead, they are typically covered in courses on numerical linear algebra or matrix computations.

2.3. Schmidt decomposition and singular values: finite dimensional case

In view of Theorem 2.2.15, for arbitrary matrix $A \in \mathbb{C}^{m\times n}$, one may consider the matrices $A^*A \in \mathbb{C}^{n\times n}$ or $AA^* \in \mathbb{C}^{m\times m}$, both of which are Hermitian. It is sufficient to consider $A^*A \in \mathbb{C}^{n\times n}$. First, we see that

$$\mathbf{v}^*A^*A\mathbf{v} = (A\mathbf{v})^*A\mathbf{v} = \|A\mathbf{v}\|^2 \geq 0 \quad \text{for all } \mathbf{v} \in \mathbb{C}^n,$$

in other words, A^*A is *non-negative definite* (or called *positive semi-definite*). As an immediate corollary, all eigenvalues of the Hermitian matrix $A^*A \in \mathbb{C}^{n\times n}$ must be non-negative.

DEFINITION 2.3.1. By Theorem 2.2.15, we have $A^*A = UDU^*$ for some unitary matrix U and $D = \operatorname{diag}(\lambda_1,\cdots,\lambda_n)$ for some $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. The numbers $\sigma_k := \sqrt{\lambda_k}$ are called the *singular values* of A.

Note that Definition 2.3.1 applies to *arbitrary* matrices $A \in \mathbb{C}^{m\times n}$, not necessarily square matrices. Using Theorem 2.2.15, we immediate reach the following corollary.

COROLLARY 2.3.2 (Courant characterization of singular values). *Given an* arbitrary *matrix $A \in \mathbb{C}^{m\times n}$, its singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ can be characterized as follows:*

$$\sigma_k = \max_{\substack{X \leq \mathbb{C}^n \\ \dim X = k}} \min_{\mathbf{v}\in X, \|\mathbf{v}\|=1} \|A\mathbf{v}\| = \min_{\substack{X \leq \mathbb{C}^n \\ \dim X = k-1}} \max_{\mathbf{u}\perp X, \|\mathbf{u}\|=1} \|A\mathbf{u}\|.$$

By definition of the singular values (Definition 2.3.1), the numbers $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_n^2$ are the eigenvalues of A^*A. By Theorem 2.2.15, the corresponding eigenvectors $\mathbf{u}_1, \cdots, \mathbf{u}_n$ form an orthonormal basis of $\mathbb{C}^n$.

EXERCISE 2.3.3 (Singular value decomposition (SVD) for matrices). Given an *arbitrary* matrix $A \in \mathbb{C}^{m\times n}$. Suppose that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_\ell > 0$ for some $\ell \in \mathbb{N}$. Show that the vectors

$$\mathbf{v}_k := \frac{1}{\sigma_k} A\mathbf{u}_k \quad \text{for all } k = 1, 2, \cdots, \ell \tag{2.3.1}$$

form an orthonormal set $\{\mathbf{v}_1, \cdots, \mathbf{v}_\ell\}$. Show that

$$A\mathbf{x} = \sum_{k=1}^{\min\{m,n\}} \sigma_k (\mathbf{x}, \mathbf{u}_k)_{\mathbb{C}^n} \mathbf{v}_k, \tag{2.3.2}$$

where $\{\mathbf{v}_1, \cdots, \mathbf{v}_m\}$ is an orthonormal basis of $\mathbb{C}^m$ extending the family in (2.3.1). Show that (2.3.2) is equivalent to the matrix factorization (called the *singular value decomposition* (SVD) for matrices)

$$A = V\Sigma U^*,$$

where $\Sigma = \mathrm{diag}\,(\sigma_1, \cdots, \sigma_{\min\{m,n\}}) \in \mathbb{R}^{m\times n}$, and

$$U = \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_n \end{pmatrix} \in \mathbb{C}^{n\times n}, \quad V = \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_m \end{pmatrix} \in \mathbb{C}^{m\times m}$$

are unitary matrices.

By comparing Exercises 2.2.16 and 2.3.3, we see that the singular value decomposition can be viewed as a generalization of the diagonalization of Hermitian matrices. We emphasize again that the *SVD applies to arbitrary matrices*, and many numerical algorithms are therefore based on it rather than on diagonalization.

REMARK 2.3.4. Given a matrix $A \in \mathbb{C}^{m\times n}$, one practical strategy for computing its SVD is to first reduce A to bidiagonal form using the Golub-Kahan bidiagonalization and then compute the SVD of the resulting bidiagonal matrix via Lanczos-type methods. The effectiveness of these algorithms relies on the observation

$$(A^*A)\mathbf{v} = A^*(A\mathbf{v}). \tag{2.3.3}$$

Rather than forming the product A^*A explicitly (*a practice that is generally discouraged in numerical computations*), one applies the matrices successively to a vector. This avoids expensive matrix-matrix operations and typically leads to improved numerical stability. For a thorough account of these methods, see the monographs [**GVL13, TB97**].

The discussion so far should make it clear why many modern linear algebra courses no longer emphasize the Jordan canonical form. The singular value decomposition is a powerful tool for solving linear systems of the form $A\mathbf{x} = \mathbf{b}$. We first consider the case in which A is invertible. Even in this setting, *one rarely computes A^{-1} explicitly, since doing so is computationally expensive and often numerically undesirable*. If $A = V\Sigma U^*$ is the singular value decomposition of A, then all singular values are positive, and hence

$$\mathbf{x} = U\Sigma^{-1}V^*\mathbf{b}.$$

This provides an efficient and numerically stable way to solve the system. The same idea extends naturally to the case where A is singular or even rectangular. From $A\mathbf{x} = \mathbf{b}$, we obtain

$$\Sigma U^*\mathbf{x} = V^*\mathbf{b}.$$

Removing the rows of U^* corresponding to zero singular values (equivalently, removing the corresponding columns of U), we arrive at $\tilde{\Sigma}\tilde{U}^*\mathbf{v} = V^*\mathbf{b}$, and therefore

$$\mathbf{x} = \overbrace{\tilde{U}\tilde{\Sigma}^{-1}V^*}^{A^\dagger :=}\mathbf{b}.$$

This is precisely the Moore-Penrose pseudoinverse solution. Consequently, if the system $A\mathbf{x} = \mathbf{b}$ is undertermined, then $\mathbf{x}$ is the solution of minimal norm. If the system is overdetermined, then $\mathbf{x}$ minimizes the residual norm $\|A\mathbf{x} - \mathbf{b}\|$.

REMARK 2.3.5. In practice, the system $A\mathbf{x} = \mathbf{b}$ is often solved without explicitly computing the SVD. A widely used iterative approach is the *Generalized Minimal Residual Method* (*GMRES*), see the monographs [**GVL13, TB97**].

2.4. Schmidt decomposition and singular values for compact linear operators

We now turn to functional analysis and introduce the infinite-dimensional analogue of the singular value decomposition:

THEOREM 2.4.1 (Singular value decomposition for compact linear operators [**KRS25**, Proposition 2.3]). *Let $\mathscr{A} : X \to Y$ be a compact linear operator (Definition 1.5.19) between separable Hilbert spaces, and assume that X is infinite dimensional. Then there exists singular values*

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq 0 \quad \lim_{j\to\infty} \sigma_j = 0,$$

together with orthonormal bases $\{\varphi_j\}_{j\in\mathbb{N}}$ *of* X *and* $\{\psi_j\}_{j\in\mathbb{N}}$ *of* Y*, such that*

$$\mathscr{A}u = \sum_{j\in\mathbb{N}} \sigma_j (u,\varphi_j)_X \psi_j \quad \textit{for all } u \in X.$$

This representation is called the singular value decomposition *of* $\mathscr{A}$*, and* $\{\varphi_j\}_{j\in\mathbb{N}}$ *is called a* singular basis *for* $\mathscr{A}$*. In particular, for every* j *with* $\sigma_j > 0$*,*

$$\mathscr{A}\varphi_j = \sigma_j \psi_j, \quad \mathscr{A}^*\psi_j = \sigma_j \varphi_j.$$

The singular values $\sigma_j = \sigma_j(A) = \sigma_j(A^*)$ *satisfy the following properties:*

(a) $\sigma_1 = \|\mathscr{A}\|_{X\to Y} := \sup_{\|u\|_x=1} \|\mathscr{A}u\|_Y$

(b) *If* $\mathscr{A}: X \to Y$ *is injective, then* $\sigma_j > 0$ *for all* $j \in \mathbb{N}$*.*

(c) *If* $\mathscr{B}: X \to Y$ *is compact, then for all* $j,k \in \mathbb{N}$ *one has*

$$\sigma_{j+k-1}(\mathscr{A}+\mathscr{B}) \le \sigma_j(\mathscr{A}) + \sigma_k(\mathscr{B}).$$

(d) *If* $\mathscr{C}: X_1 \to X$ *is compact, then for all* $j,k \in \mathbb{N}$ *one has*

$$\sigma_{j+k-1}(\mathscr{A}\circ\mathscr{C}) \le \sigma_j(\mathscr{A})\sigma_k(\mathscr{C}).$$

(e) *If* $\Phi: X_1 \to X$ *and* $\Psi: Y \to Y_1$ *are bounded linear operators, one has*

$$\sigma_j(\mathscr{A}\circ\Phi) \le \sigma_j(\mathscr{A})\|\Phi\|_{X_1\to X}, \quad \sigma_j(\Psi\circ\mathscr{A}) \le \sigma_j(\mathscr{A})\|\Psi\|_{Y\to Y_1}.$$

Moreover, if Φ *and* Ψ *are isometries, that is,* $\|\Phi u\|_X = \|u\|_{X_1}$ *for all* $u \in X_1$ *and* $\|\Psi v\|_{Y_1} = \|v\|_Y$ *for all* $x \in Y$*, then*

$$\sigma_j(\mathscr{A}\circ\Phi) = \sigma_j(\Psi\circ\mathscr{A}) = \sigma_j(\mathscr{A}).$$

REMARK 2.4.2. The singular values depend not only on the operator $\mathscr{A}$, but also on the underlying Hilbert spaces X and Y. When necessary, we shall denote them by $\sigma_j(\mathscr{A}: X \to Y)$ to emphasize this dependence. This is analogous to the finite-dimensional situation, where the matrix representation of a linear operator depends on the choice of basis.

We shall not discuss these details here. Instead, we illustrate the theory through a number of examples. We first present an example related to acoustic waves.

EXAMPLE 2.4.3 ([**KSZ25**, Section 1.1]). The simplest form of an acoustic wave is a plane wave of frequency $\kappa > 0$, propagating in the direction $\omega \in \mathbb{S}^{d-1} := \{\omega \in \mathbb{R}^d : |\omega| = 1\}$:

$$x \mapsto e^{\mathrm{i}\kappa\omega\cdot x} \equiv \cos(\kappa\omega\cdot x) + \mathrm{i}\sin(\kappa\omega\cdot x).$$

The (up to a multiplicative scaling factor $\kappa^{\frac{d-1}{2}}$) superposition of such plane waves with density $f \in L^2(\mathbb{S}^{d-1})$ defines the corresponding Herglotz wave function, given by

$$(\mathscr{A}_\kappa f)(x) := \kappa^{\frac{d-1}{2}} \int_{\mathbb{S}^{d-1}} e^{\mathrm{i}\kappa\omega\cdot x} f(\omega)\,\mathrm{d}S(\omega) \equiv \kappa^{\frac{d-1}{2}} \left(f\mathscr{H}^{d-1}\lfloor_{\mathbb{S}^{d-1}}\right)^{\wedge}(-\kappa x) \quad \text{for all } x \in \mathbb{R}^n.$$

In fact,$\mathscr{A}_\kappa f$ is an entire function, as a consequence of the Paley-Wiener theorem for Fourier transforms of compactly supported distributions, see, e.g., [Kow25] for details. It is easy to see that

$$\mathscr{A}_\kappa : L^2(\mathbb{S}^{d-1}) \to L^2(B_1) \text{ is a bounded injective linear operator (Definition 1.5.19).} \tag{2.4.1}$$

Moreover, the Agmon-Hörmander estimate [AH76] refines this bound by showing that

$$\|\mathscr{A}_\kappa f\|_{L^2(B_1)} \le C(d)\|f\|_{L^2(\mathbb{S}^{d-1})}$$

where the constant $C(d)$ depends only on the dimension d and is independent of κ. One standard way to prove the compactness of 2.4.1 is to use Sobolev embedding: First, it is easy to see that

$$\mathscr{A}_\kappa : L^2(\mathbb{S}^{d-1}) \to H^1(B_1) \text{ is a bounded linear operator (Definition 1.5.19),}$$

the compactness of (2.4.1) then follows from Lemma 1.5.20 and Theorem 1.5.21. Using Theorem 2.4.1, the singular values $\sigma_j(\mathscr{A}_\kappa : L^2(\mathbb{S}^{d-1}) \to L^2(B_1))$ are well-defined and positive, since (2.4.1) is injective.

We now turn to another example arising in acoustic wave theory.

EXAMPLE 2.4.4. [KW26] We consider the time-harmonic acoustic wave with a nontrivial volume source f, which is a distribution on $\mathbb{R}^d$, with compact support $\operatorname{supp}(f)$:

$$\begin{cases} -(\Delta+\kappa^2)u = f & \text{in } \mathbb{R}^d, \\ \lim_{|x|\to\infty} |x|^{\frac{n-1}{2}}(\partial_r u - \mathrm{i}\kappa u) = 0 & \text{uniformly for } \hat{x} = \frac{x}{|x|} \in \mathbb{S}^{d-1}. \end{cases} \tag{2.4.2}$$

The second equation in (2.4.2) is known as the Sommerfeld radiation condition. Without loss of generality, we assume that $\operatorname{supp}(f) \subset \overline{B_1}$. We define

$$\Phi_\kappa(x) := \frac{\mathrm{i}\kappa^{\frac{d-2}{2}}}{4(2\pi)^{\frac{d-2}{2}}}|x|^{-\frac{d-2}{2}}H^{(1)}_{\frac{d-2}{2}}(\kappa|x|),$$

where $H^{(1)}$ is the Hankel function of the first kind. For each $y \in \mathbb{R}^d$, it is known that Φ_κ satisfies the above-mentioned Sommerfeld radiation condition and

$$-(\Delta+\kappa^2)\Phi_\kappa = \delta_0 \text{ in distribution sense}$$

where δ_0 is the standard Dirac delta. In other words, Φ_κ is the outgoing fundamental solution of the Helmholtz operator $-(\Delta+\kappa^2)$. Together with the Rellich uniqueness theorem [CK19, Hör73], we can see that the solution of (2.4.2) is unique, and the solution u is written explicitly by the convolution

$$u = \Phi_\kappa * f$$

which is understood as the convolution of two distributions, see, e.g., [**Kow25**]. Now its *far-field pattern* is defined by

$$u_\kappa^\infty[f](\hat{x}) := \lim_{|x|\to\infty} \gamma_{d,k}^{-1} |x|^{\frac{d-1}{2}} e^{-\mathrm{i}\kappa|x|} u(x) \quad \text{with} \quad \gamma_{d,\kappa} := \frac{e^{-\mathrm{i}\pi\frac{d-3}{4}}}{2(2\pi)^{\frac{d-1}{2}}} \kappa^{\frac{d-3}{2}}.$$

Such choice of $\gamma_{d,\kappa}$ guarantees that

$$\Phi_\kappa^\infty(\hat{x}, y) = e^{-\mathrm{i}\kappa\hat{x}\cdot y} \quad \text{for all } \hat{x} \in \mathbb{S}^{d-1} \text{ and } y \in \mathbb{R}^d,$$

see, for example, [**Yaf10**, Section 1.2.3], therefore, u_κ^∞ can be represented in terms of Fourier transform:

$$u_\kappa^\infty[f](\hat{x}) = \int_{\mathbb{R}^d} e^{-\mathrm{i}\kappa\hat{x}\cdot y} f(y)\,\mathrm{d}y = \int_{B_1} e^{-\mathrm{i}\kappa\hat{x}\cdot y} f(y)\,\mathrm{d}y = \mathscr{F}[f](\kappa\hat{x}) \quad \text{for all } \hat{x} \in \mathbb{S}^{d-1}.$$

We again remark that $\mathscr{F}[f]$ is analytic since f is a distirbution with compact support (Paley-Wiener theorem). Fix any $\mathscr{K} > 0$ and define $\mathbb{S}^{d-1}_{\mathscr{K}} := (0,\mathscr{K}) \times \mathbb{S}^{d-1}$. We consider the bounder linear operator

$$G_{\mathscr{K}} : L^2(B_1) \mapsto L^2(\mathbb{S}^{d-1}), \quad (G_{\mathscr{K}} f)(\kappa, \hat{x}) := \chi_{\mathbb{S}^{d-1}_{\mathscr{K}}}(\kappa, \hat{x}) u_\kappa^\infty[f](x) \text{ for a.e. } (\kappa, \hat{x}) \in \mathbb{S}^{d-1}_{\mathscr{K}}.$$

In fact, the above bounded linear operator is both injective and compact. Consequently, Theorem 2.4.1 applies.

Before proceeding to more interesting examples, we briefly review some preliminary material from [**KRS25**].

DEFINITION 2.4.5 ([**KRS25**, (3.8)]). Let X and Y be separable Hilbert spaces. The space $\mathrm{HS}(X,Y)$ consists of those compact linear operators $T : X \to Y$ for which the quantity

$$\|T\|^2_{\mathrm{HS}(X,Y)} = \sum_{j=1}^\infty \sigma_j^2 = \sum_{j=1}^\infty \|T\varphi_j\|_Y^2 = \sum_{j,k=1}^\infty |(T\varphi_j, \psi_k)_Y|^2$$

is finite and *independent of the choice* (it should be noted that this fact is not at all obvious) of orthonormal bases $\{\varphi_j\}_{j\in\mathbb{N}} \subset X$ and $\{\psi_j\}_{j\in\mathbb{N}} \subset Y$, where $\sigma_j = \sigma_j(T : X \to Y)$ are singular values (Theorem 2.4.1). This is a Hilbert space with scalar product

$$(S,T)_{\mathrm{HS}(X,Y)} = \sum_{j=1}^\infty (S\varphi_j, T\varphi_j)_Y.$$

REMARK 2.4.6 ([**KRS25**, (3.9)]). When $X = Y = L^2(\mathbb{S}^{d-1})$, any $T \in \mathrm{HS}(L^2(\mathbb{S}^{d-1}), L^2(\mathbb{S}^{d-1}))$ has Schwartz kernel $K_T \in L^2(\mathbb{S}^d \times \mathbb{S}^d)$, i.e.

$$(T\phi)(\hat{x}) = \int_{\mathbb{S}^{d-1}} K_T(\hat{x}, \hat{y})\phi(\hat{y})\,\mathrm{d}S_{\hat{y}},$$

satisfying

$$\|T\|_{\mathrm{HS}(L^2(\mathbb{S}^d), L^2(\mathbb{S}^d))} = \|K_T\|_{L^2(\mathbb{S}^d\times\mathbb{S}^d)}.$$

REMARK 2.4.7. The space of bounded linear operators, endowed with the operator norm, forms a Banach space. However, *it is generally not a Hilbert space*.

We now ready to state the following interesting example.

EXAMPLE 2.4.8 ([**KSZ25**, Section 1.2]). The following facts may be found, e.g., in [**Mel95, PSU10**]. Let $\kappa > 0$ be a fixed frequency and let $q \in C_c^\infty(\mathbb{R}^d)$ be an unknown scattering potential. We suppose that we probe the medium by sending an incoming Herglotz wave $u^{\rm inc} = P_\kappa f$, where $f \in L^2(\mathbb{S}^{n-1})$ and

$$(P_\kappa f)(x) := \int_{\mathbb{S}^{d-1}} e^{\mathrm{i}\kappa\omega\cdot x} f(\omega)\,\mathrm{d}S(\omega) \equiv \left(f\mathscr{H}^{d-1}\lfloor_{\mathbb{S}^{d-1}}\right)^{\wedge}(-\kappa x) \quad \text{for all } x \in \mathbb{R}^n.$$

This induces a unique total wave $u^{\rm tot} = u^{\rm inc} + u^{\rm sc}$ solving

$$(\Delta + \kappa^2 + q)u^{\rm tot} = 0 \quad \text{in } \mathbb{R}^d$$

where $u^{\rm sc}$ satisfies the outgoing Sommerfeld radiation condition (Example 2.4.4). We write $\mathscr{P}_\kappa(q)f := u^{\rm tot}$ and call $\mathscr{P}_\kappa(q)$ the *Poisson operator* that maps a boundary data f on the sphere at infinity to the corresponding solution $u^{\rm tot}$. The scattering measurement at frequency κ are encoded by the *scattering operator* $\mathscr{S}_\kappa(q)$, defined by

$$(\mathscr{S}_\kappa(q)f)(\theta) = \lim_{r\to\infty} r^{\frac{d-1}{2}} e^{-\mathrm{i}\kappa r}(\partial_r u^{\rm tot} + \mathrm{i}\kappa u^{\rm tot})(r\theta).$$

After multiplying $\mathscr{S}_\kappa(q)$ by a suitable normalizing constant, it becomes a unitary operator on $L^2(\mathbb{S}^{d-1})$ that satisfies the integral identity

$$([\mathscr{S}_\kappa(q) - \mathscr{S}_\kappa(0)]f, g)_{L^2(\mathbb{S}^{d-1})} = c_{d,\kappa}((q-0)\mathscr{P}_\kappa(q)f, \mathscr{P}_\kappa(0)g)_{L^2(\mathbb{R}^d)}.$$

The linearization of $\mathscr{S}_\kappa$ at $q = 0$ is readily obtained from the above identity. This linearization, which will be the main subject of the next chapter, is denoted by $\mathscr{F}_\kappa$ (after multiplying by a suitable constant). After a suitable normalization, it is given by

$$(\mathscr{F}_\kappa(h)f, g)_{L^2(\mathbb{S}^{d-1})} = \kappa^{\frac{d-1}{2}}(hP_\kappa f, P_\kappa g)_{L^2(\mathbb{R}^d)}.$$

In fact, $\mathscr{F}_\kappa : L^2(B_1) \to \mathrm{HS}(L^2(\mathbb{S}^{d-1}), L^2(\mathbb{S}^{d-1}))$ is an injective compact linear operator. Consequently, Theorem 2.4.1 applies.

CHAPTER 3

Linearization

Up to this point, we have considered only linear operators, which is clearly too restrictive for many applications. For a general nonlinear operator, it is natural to first study its linearization at a given point in order to understand its local behavior.

3.1. Limits and continuity

We first recall some fact in calculus, see e.g. my lecture note [**Kow24**] and the references therein for more details.

DEFINITION 3.1.1. A subset $\Omega \subset \mathbb{R}^N$ is said to be *open* if for each $x \in \Omega$ there exists $\varepsilon = \varepsilon(x) > 0$ such that $B_\varepsilon(x) \subset \Omega$. Here and after, the open ball $B_R(x)$ is defined by

$$B_R(x) := \left\{ \boldsymbol{y} = (y_1, \cdots, y_N) \in \mathbb{R}^N : |\boldsymbol{x} - \boldsymbol{y}| < R \right\},$$

where $|\boldsymbol{z}| = \sqrt{z_1^2 + \cdots + z_N^2}$ for all $\boldsymbol{z} = (z_1, \cdots, z_N) \in \mathbb{R}^N$.

DEFINITION 3.1.2. Let Ω be an open set in $\mathbb{R}^N$ with $\boldsymbol{x}_0 \in \Omega$ and we consider a function $f : \Omega \setminus \{\boldsymbol{x}_0\} \to \mathbb{R}$.

(1) We say that the *limit* $\lim_{\boldsymbol{x} \to \boldsymbol{x}_0} f(\boldsymbol{x}) = L$ exists if the following holds: Given any $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon) > 0$ such that

$$0 < |\boldsymbol{x} - \boldsymbol{x}_0| < \delta \text{ implies } |f(\boldsymbol{x}) - L| < \varepsilon.$$

In this case, we also say that the *limit* $\lim_{\boldsymbol{x} \to \boldsymbol{x}_0} f(\boldsymbol{x})$ exists in $\mathbb{R}$.

(2) We say that the *limit* $\lim_{\boldsymbol{x} \to \boldsymbol{x}_0} f(\boldsymbol{x}) = +\infty$ exists if the following holds: Given any $M > 0$, there exists $\delta = \delta(M) > 0$ such that

$$0 < |\boldsymbol{x} - \boldsymbol{x}_0| < \delta \text{ implies } f(\boldsymbol{x}) > M.$$

(3) We say that the *limit* $\lim_{\boldsymbol{x} \to \boldsymbol{x}_0} f(\boldsymbol{x}) = -\infty$ exists if the following holds: Given any $M > 0$, there exists $\delta = \delta(M) > 0$ such that

$$0 < |\boldsymbol{x} - \boldsymbol{x}_0| < \delta \text{ implies } f(\boldsymbol{x}) < -M.$$

We also unify the above notions by saying that the *limit* $\lim_{\boldsymbol{x} \to \boldsymbol{x}_0} f(\boldsymbol{x})$ exists in $[-\infty, +\infty]$.

The following are some basic properties of limits:

LEMMA 3.1.3. *Let Ω be an open set in $\mathbb{R}^N$ with $\boldsymbol{x}_0 \in \Omega$ and we consider functions $g_1 : \Omega \setminus \{\boldsymbol{x}_0\} \to \mathbb{R}$ and $g_2 : \Omega \setminus \{\boldsymbol{x}_0\} \to \mathbb{R}$. If both limits $\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} g_1(\boldsymbol{x})$ and $\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} g_2(\boldsymbol{x})$ exist in $\mathbb{R}$, then the following holds true:*

(1) *for each $c_1 \in \mathbb{R}$ and $c_2 \in \mathbb{R}$ the limit $\lim_{\boldsymbol{x}\to\boldsymbol{x}_0}(c_1 g_1(\boldsymbol{x}) + c_2 g_2(\boldsymbol{x}))$ exists in $\mathbb{R}$ and satisfies*

$$\lim_{\boldsymbol{x}\to\boldsymbol{x}_0}(c_1 g_1(\boldsymbol{x}) + c_2 g_2(\boldsymbol{x})) = c_1 \lim_{\boldsymbol{x}\to\boldsymbol{x}_0} g_1(\boldsymbol{x}) + c_2 \lim_{\boldsymbol{x}\to\boldsymbol{x}_0} g_2(\boldsymbol{x}) \quad \textit{(linearity)}.$$

(2) *if $g_1(\boldsymbol{x}) \leq g_2(\boldsymbol{x})$ for all $\boldsymbol{x} \in B_\varepsilon(\boldsymbol{x}_0) \setminus \{\mathbf{x}_0\}$ for some $\varepsilon > 0$, then*

$$\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} g_1(\boldsymbol{x}) \leq \lim_{\boldsymbol{x}\to\boldsymbol{x}_0} g_2(\boldsymbol{x}) \quad \textit{(monotonicity)}.$$

(3) *the limit $\lim_{\boldsymbol{x}\to\boldsymbol{x}_0}(g_1(\boldsymbol{x})g_2(\boldsymbol{x}))$ exists in $\mathbb{R}$ and satisfies*

$$\lim_{\boldsymbol{x}\to\boldsymbol{x}_0}(g_1(\boldsymbol{x})g_2(\boldsymbol{x})) = \left(\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} g_1(\boldsymbol{x})\right)\left(\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} g_2(\boldsymbol{x})\right).$$

(4) *if we additionally assume that $\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} g_2(\boldsymbol{x}) \neq 0$, then the limit $\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} \frac{g_1(\boldsymbol{x})}{g_2(\boldsymbol{x})}$ exists in $\mathbb{R}$ and satisfies*

$$\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} \frac{g_1(\boldsymbol{x})}{g_2(\boldsymbol{x})} = \frac{\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} g_1(\boldsymbol{x})}{\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} g_2(\boldsymbol{x})}.$$

We now ready to introduce the following notion:

DEFINITION 3.1.4. Let Ω be an open set in $\mathbb{R}^N$ and let $f : \Omega \to \mathbb{R}$ be a function. If

$$\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} f(\boldsymbol{x}) = f(\boldsymbol{x}_0) \in \mathbb{R} \quad \text{for some } \boldsymbol{x}_0 \in \Omega,$$

then we say that f is *continuous at $\boldsymbol{x}_0$*.

(1) If there exists an open set $\boldsymbol{x}_0 \in U \subset \Omega$ such that f is continuous at all point $\boldsymbol{x} \in U$, then we say that f is *continuous near $\boldsymbol{x}_0$*.
(2) If f is continuous at all points in Ω, then we say that f is *continuous on* Ω.

3.2. Linearization of functions and Landweber iteration: finite dimensional case

In mathematics, linearization (British English: linearisation) refers to the process of approximating a function by a linear map near a given point. More precisely, if $y = f(x)$, the linearization of f at $x = a$ provides a linear approximation of f near a, based on the value $f(a)$ and the derivative of f at a. In the one-dimensional setting, this approximation is given by the tangent line at $x = a$. Thus, studying the linearization of f is equivalent to studying its derivative at $x = a$. For the case when $N = 1$, the differentiation of the function $f : \Omega \subset \mathbb{R}^1 \to \mathbb{R}$ can be simply define by the limit

$$f'(x) := \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}. \tag{3.2.1}$$

However, the above definition cannot be directly extended to higher dimensional case by simply replace x and h by vectors $\boldsymbol{x}$ and $\boldsymbol{h}$, since the division of vectors are not well-defined. Let's us observe the above equation holds if and only if

$$\begin{aligned} 0 &= \lim_{h\to 0}\left|\frac{f(x+h)-f(x)}{h} - f'(x)\right| \\ &= \lim_{h\to 0}\left|\frac{f(x+h)-f(x)-f'(x)h}{h}\right| = \lim_{h\to 0}\frac{|f(x+h)-f(x)-f'(x)h|}{|h|}. \end{aligned}$$

Now it is natural to consider the differentiation of the function $f:\Omega\subset\mathbb{R}^N\to\mathbb{R}$ as follows:

DEFINITION 3.2.1. Let Ω be an open set in $\mathbb{R}^N$ and let $f:\Omega\to\mathbb{R}$ be a function. We say that f is *differentiable* at $\boldsymbol{x}_0\in\Omega$ if there exists a vector $\boldsymbol{L}\in\mathbb{R}^N$ such that

$$\lim_{\boldsymbol{h}\to\boldsymbol{0}}\frac{|f(\boldsymbol{x}+\boldsymbol{h})-f(\boldsymbol{x})-\boldsymbol{L}\cdot\boldsymbol{h}|}{|\boldsymbol{h}|}=0, \tag{3.2.2}$$

where $\boldsymbol{L}\cdot\boldsymbol{h}=L_1h_1+\cdots+L_Nh_N$. In this case, the total derivative $\boldsymbol{D}f(\boldsymbol{x}_0)$ of f at $\boldsymbol{x}_0$is defined by the vector $\boldsymbol{D}f(\boldsymbol{x}_0):=\boldsymbol{L}$.

(1) If there exists an open set $\boldsymbol{x}_0\in U\subset\Omega$ such that f is differentiable at all point $\boldsymbol{x}\in U$, then we say that f is *differentiable near* $\boldsymbol{x}_0$.
(2) If f is differentiable at all points in Ω, then we say that f is *differentiable on* Ω.

It is not so obvious that whether the vector $\boldsymbol{L}$ in (3.2.2) is unique or not. Suppose that (3.2.2) holds true for $\boldsymbol{L}=\boldsymbol{L}_1$ and $\boldsymbol{L}=\boldsymbol{L}_2$, then

$$\begin{aligned} \frac{|(\boldsymbol{L}_1-\boldsymbol{L}_2)\cdot\boldsymbol{h}|}{|\boldsymbol{h}|} &= \frac{|(f(\boldsymbol{x}+\boldsymbol{h})-f(\boldsymbol{x})-\boldsymbol{L}_1\cdot\boldsymbol{h})-(f(\boldsymbol{x}+\boldsymbol{h})-f(\boldsymbol{x})-\boldsymbol{L}_2\cdot\boldsymbol{h})|}{|\boldsymbol{h}|} \\ &\leq \frac{|f(\boldsymbol{x}+\boldsymbol{h})-f(\boldsymbol{x})-\boldsymbol{L}_1\cdot\boldsymbol{h}|-|f(\boldsymbol{x}+\boldsymbol{h})-f(\boldsymbol{x})-\boldsymbol{L}_2\cdot\boldsymbol{h}|}{|\boldsymbol{h}|}\to 0 \quad \text{as } |\boldsymbol{h}|\to 0_+. \end{aligned}$$

By restricting the limit along the straight line $\boldsymbol{h}=h\boldsymbol{e}_i$ ($h\in\mathbb{R}$) for all $i=1,\cdots,N$, we conclude $\boldsymbol{L}_1=\boldsymbol{L}_2$. Here and after, $\boldsymbol{e}_i$ denotes the i^{th} column of the identity matrix, that is,

$$\begin{aligned} \boldsymbol{e}_1 &= (1,0,0,\cdots,0), \\ \boldsymbol{e}_2 &= (0,1,0,\cdots,0), \\ &\vdots \\ \boldsymbol{e}_N &= (0,\cdots,0,0,1). \end{aligned}$$

EXERCISE 3.2.2. Show that each differentiable function is also continuous.

After exhibiting an abstract definition of differentiation of a function f, we are now asking how to compute its total derivative $\boldsymbol{D}f(\boldsymbol{x})$ at each point $\boldsymbol{x}$. If f is differentiable at $\boldsymbol{x}_0\in\Omega$, then we may restrict the limit (3.2.2) on the straight line $\{h\boldsymbol{e}_i : h \in \mathbb{R}\}$ to see that

$$\begin{aligned}
0 &= \lim_{\boldsymbol{h}\to\boldsymbol{0}} \frac{|f(\boldsymbol{x}+\boldsymbol{h}) - f(\boldsymbol{x}) - \boldsymbol{L}\cdot\boldsymbol{h}|}{|\boldsymbol{h}|} = \lim_{\boldsymbol{h}=h\boldsymbol{e}_i\to\boldsymbol{0}} \frac{|f(\boldsymbol{x}+\boldsymbol{h}) - f(\boldsymbol{x}) - \boldsymbol{L}\cdot\boldsymbol{h}|}{|\boldsymbol{h}|} \\
&= \lim_{h\to 0} \frac{|f(\boldsymbol{x}+h\boldsymbol{e}_i) - f(\boldsymbol{x}) - L_i h|}{|h|},
\end{aligned}$$

which implies

$$L_i = \lim_{h\to 0} \frac{f(\boldsymbol{x}+h\boldsymbol{e}_i) - f(\boldsymbol{x})}{h},$$

which is exactly identical to the definition of the differentiation of functions with only one variable (3.2.1). Now it is natural to consider the following notion.

DEFINITION 3.2.3. Let Ω be an open set in $\mathbb{R}^N$ and let $f : \Omega \to \mathbb{R}$ be a function. For each $i = 1, \cdots, N$, the i^{th} partial derivative of f at $\boldsymbol{x}_0 = (x_1, \cdots, x_N)$ is defined by

$$\frac{\partial}{\partial x_i} f(x_1, \cdots, x_{i-1}, x, x_{i+1}, \cdots, x_N)\bigg|_{x=x_i} = \partial_i f(\boldsymbol{x}_0) := \lim_{h\to 0} \frac{f(\boldsymbol{x}_0 + h\boldsymbol{e}_i) - f(\boldsymbol{x}_0)}{h}.$$

If all partial derivatives at $\boldsymbol{x}_0$ exist, then we define the gradient $\nabla f(\boldsymbol{x}_0) := (\partial_1 f(\boldsymbol{x}_0), \cdots, \partial_N f(\boldsymbol{x}_0)) \in \mathbb{R}^N$.

We now put the above discussions in the following lemma.

THEOREM 3.2.4. *Let Ω be an open set in $\mathbb{R}^N$ and let $f : \Omega \to \mathbb{R}$ be a function. If f is differentiable at $\boldsymbol{x}_0 \in \Omega$, then $\boldsymbol{D}f(\boldsymbol{x}_0) = \nabla f(\boldsymbol{x}_0)$.*

REMARK 3.2.5. The existence of the gradient $\nabla f(\boldsymbol{x}_0)$ does not guarantee the differentiability of f at $\boldsymbol{x}_0$.

The following sufficient condition is often been used to check the differentiability of a function.

THEOREM 3.2.6 ([Apo74, Theorem 12.11]). *Let Ω be an open set in $\mathbb{R}^N$ and let $f : \Omega \to \mathbb{R}$ be a function. If the point $\boldsymbol{x}_0 \in \Omega$ satisfies the following two conditions:*

- *there exists $\varepsilon > 0$ such that all partial derivatives $\partial_1 f, \cdots, \partial_N f$ exist on $B_\varepsilon(\boldsymbol{x}_0)$; and*
- *all partial derivatives $\partial_1 f, \cdots, \partial_N f$ are continuous at $\boldsymbol{x}_0$;*

then f is differentiable at $\boldsymbol{x}_0$.

The above theorem suggested the following definition.

DEFINITION 3.2.7. Let Ω be an open set in $\mathbb{R}^N$. We denote $C^1(\Omega)$ be the collection of differentiable functions $f : \Omega \to \mathbb{R}$ such that all partial derivatives $\partial_1 f, \cdots, \partial_N f : \Omega \to \mathbb{R}$ are continuous.

We often use the following consequence of Theorem 3.2.6 since it is much easy to remember:

COROLLARY 3.2.8. *Let Ω be an open set in $\mathbb{R}^N$. If $f \in C^1(\Omega)$, then $f : \Omega \to \mathbb{R}$ is differentiable and hence $\boldsymbol{D} f(\boldsymbol{x}) = \nabla f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \Omega$.*

The above corollary allows us to solve multidimensional case by using technique in 1-dimensional case.

We now describe a method that uses linearization to solve nonlinear equations, known as the Landweber iteration. This method was originally introduced by Landweber in 1951 for a class of linear inverse problems [**Lan51**]. Here we present a nonlinear version of the iteration, following [**HNS95**]. We consider a system of nonlinear equations:

$$f_i(\boldsymbol{x}) = y_i \quad \text{for } i = 1, \cdots, N,$$

equivalently, by writing $\boldsymbol{f} = (f_1, \cdots, f_N)$,

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{y}. \tag{3.2.3}$$

Our goal is to solve $\boldsymbol{x}$ from the measurement $\boldsymbol{y}$. Suppose that $\boldsymbol{f} = (f_1, \cdots, f_N)$ is differentiable, its first derivative is given by

$$(\nabla \otimes \boldsymbol{f})(\boldsymbol{x}) = \begin{pmatrix} \nabla f_1(\boldsymbol{x}) & \nabla f_2(\boldsymbol{x}) & \cdots & \nabla f_N(\boldsymbol{x}) \end{pmatrix},$$

that is,

$$((\nabla \otimes \boldsymbol{f})(\boldsymbol{x}))_{ij} = \partial_i f_j(\boldsymbol{x}) \quad \text{for all } i, j = 1, \cdots, N.$$

Suppose that (3.2.3) admits a unique solution $\boldsymbol{x}_\infty$. Under suitable assumptions on $\boldsymbol{f}$, this solution can be computed iteratively as follows [**HNS95**]:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + ((\nabla \otimes \boldsymbol{f})(\boldsymbol{x}_k))(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_k)) \tag{3.2.4}$$

with some "good" initial guess $\boldsymbol{x}_0$. Note that (3.2.4) is nothing but the classical gradient descent iteration.

3.3. Basic Properties of Derivatives

3.3.1. Differentiation rules. Here we only give some special cases which are often used in practical applications. One can refer to the monographs [**Apo74, Rud87**] for the results which are much more optimal.

LEMMA 3.3.1. *Let Ω be an open set in $\mathbb{R}^N$ and let $f_1, f_2 : \Omega \to \mathbb{R}$.*

(a) ***Linearity.*** *If both $f_1, f_2 \in C^1(\Omega)$, then for each $c_1, c_2 \in \mathbb{R}$, the function*

$$c_1 f_1 + c_2 f_2 : \Omega \to \mathbb{R}, \quad (c_1 f_1 + c_2 f_2)(\boldsymbol{x}) := c_1 f_1(\boldsymbol{x}) + c_2 f_2(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in \Omega$$

is also in $C^1(\Omega)$, and satisfying

$$\nabla(c_1 f_1 + c_2 f_2)(\boldsymbol{x}) = c_1 \nabla f_1(\boldsymbol{x}) + c_2 \nabla f_2(\boldsymbol{x}) \quad \text{for all } \boldsymbol{x} \in \Omega.$$

(b) ***Product rule.*** *If both $f_1, f_2 \in C^1(\Omega)$, then the function*

$$f_1 f_2 : \Omega \to \mathbb{R}, \quad (f_1 f_2)(\boldsymbol{x}) := f_1(\boldsymbol{x}) f_2(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in \Omega$$

is also in $C^1(\Omega)$, and satisfying

$$\nabla(f_1 f_2)(\boldsymbol{x}) = f_2(\boldsymbol{x}) \nabla f_1(\boldsymbol{x}) + f_1(\boldsymbol{x}) \nabla f_2(\boldsymbol{x}) \quad \text{for all } \boldsymbol{x} \in \Omega.$$

LEMMA 3.3.2 (chain rule [**Apo74**, Theorem 12.7]). *Let Ω be an open set in $\mathbb{R}^N$ and let $f_1, \cdots, f_m : \Omega \to \mathbb{R}$ be functions which is differentiable at a point $\boldsymbol{x}_0 \in \Omega$. We denote the vector valued function*

$$\boldsymbol{f} : \Omega \to \mathbb{R}^m, \quad f(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \cdots, f_m(\boldsymbol{x})) \text{ for all } \boldsymbol{x} \in \Omega,$$

and its range is defined by $\boldsymbol{f}(\Omega) := \{\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^m : \boldsymbol{x} \in \Omega\}$. Let U be an open set in $\mathbb{R}^m$ such that $U \supset \boldsymbol{f}(\Omega)$ and let $g : U \to \mathbb{R}$ be a function which is differentiable at $\boldsymbol{f}(\boldsymbol{x}_0)$. Then the composition of functions

$$g \circ \boldsymbol{f} : \Omega \to \mathbb{R}, \quad g \circ \boldsymbol{f}(\boldsymbol{x}) := g(\boldsymbol{f}(\boldsymbol{x})) \quad \text{for all } \boldsymbol{x} \in \Omega$$

is also differentiable at $\boldsymbol{x}_0 \in \Omega$ and its partial derivatives are given by

$$\begin{aligned} \frac{\partial}{\partial x_i}(g \circ \boldsymbol{f})(\boldsymbol{x}) &= \nabla_{\boldsymbol{y}} g(\boldsymbol{y})|_{\boldsymbol{y}=f(\boldsymbol{x})} \cdot \frac{\partial}{\partial x_i} \boldsymbol{f}(\boldsymbol{x}) \\ &= \sum_{j=1}^m \frac{\partial}{\partial y_j} g(\boldsymbol{y}) \bigg|_{\boldsymbol{y}=\boldsymbol{f}(\boldsymbol{x})} \frac{\partial}{\partial x_i} f_j(\boldsymbol{x}) \quad \text{for all } \boldsymbol{x} \in \Omega. \end{aligned} \tag{3.3.1}$$

In practice, we often use the following corollary, which says that the composition of C^1 functions is also in C^1:

COROLLARY 3.3.3. *Let Ω be an open set in $\mathbb{R}^N$ and let $\boldsymbol{f} \in (C^1(\Omega))^N$. Let U be an open set in $\mathbb{R}^m$ such that $U \supset \boldsymbol{f}(\Omega)$ and let $g \in C^1(U)$. Then the composition of functions $g \circ \boldsymbol{f} \in C^1(\Omega)$ and satisfies* (3.3.1)*.*

3.3.2. Second order derivatives. Let Ω be an open set in $\mathbb{R}^N$. The first order derivative of $f : \Omega \to \mathbb{R}$ at each $\boldsymbol{x} \in \Omega$ is given by the vector

$$\nabla f(\boldsymbol{x}) = (\partial_1 f(\boldsymbol{x}), \cdots, \partial_N f(\boldsymbol{x})) \in \mathbb{R}^N.$$

We now further assume that $\partial_i f : \Omega \to \mathbb{R}$ is differentiable for all $i = 1, \cdots, N$, then the first order derivative of each $\partial_i f$ at each point $\boldsymbol{x} \in \Omega$ is given by the vector

$$\nabla \partial_i f(\boldsymbol{x}) = (\partial_1 \partial_i f(\boldsymbol{x}), \cdots, \partial_N \partial_i f(\boldsymbol{x})) \in \mathbb{R}^N.$$

This suggests that the second order derivative of $f:\Omega\to\mathbb{R}$ at each $\boldsymbol{x}\in\Omega$ should be the following 2-tensor (i.e. matrix)

$$\nabla^{\otimes 2}f(\boldsymbol{x})\equiv\nabla\otimes\nabla f(\boldsymbol{x}):=\begin{pmatrix}\partial_1\partial_1 f(\boldsymbol{x}) & \partial_1\partial_2 f(\boldsymbol{x}) & \cdots & \partial_1\partial_N f(\boldsymbol{x})\\ \partial_2\partial_1 f(\boldsymbol{x}) & \partial_2\partial_2 f(\boldsymbol{x}) & \cdots & \partial_2\partial_N f(\boldsymbol{x})\\ \vdots & \vdots & \ddots & \vdots\\ \partial_N\partial_1 f(\boldsymbol{x}) & \partial_N\partial_2 f(\boldsymbol{x}) & \cdots & \partial_N\partial_N f(\boldsymbol{x})\end{pmatrix},$$

that is,

$$\left(\nabla^{\otimes 2}f(\boldsymbol{x})\right)_{ij}:=\partial_i\partial_j f(\boldsymbol{x})\quad\text{for all } i,j=1,\cdots,N.$$

We call $\nabla^{\otimes 2}f(\boldsymbol{x})$ the *Hessian matrix*. The notation $\otimes$ comes from the juxtaposition $\boldsymbol{u}\otimes\boldsymbol{v}$ of vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ is defined by $\boldsymbol{u}\otimes\boldsymbol{v}:=\boldsymbol{u}\boldsymbol{v}^{\intercal}$, that is,

$$(\boldsymbol{u}\otimes\boldsymbol{v})_{ij}:=u_i v_j.$$

REMARK 3.3.4. Similarly, the third derivatives of f should be the 3-tensor $\nabla^{\otimes 3}f(\boldsymbol{x})$ given by

$$\left(\nabla^{\otimes 3}f(\boldsymbol{x})\right)_{ijk}\equiv(\nabla\otimes\nabla\otimes\nabla f(\boldsymbol{x}))_{ijk}:=\partial_i\partial_j\partial_k f(\boldsymbol{x})\quad\text{for all } i,j,k=1,\cdots,N.$$

Inductively, the m^{th} derivatives of f should be the m-tensor $\nabla^{\otimes m}f(\boldsymbol{x})$ given by

$$\left(\nabla^{\otimes m}f(\boldsymbol{x})\right)_{i_1 i_2\cdots i_m}:=\partial_{i_1}\partial_{i_2}\cdots\partial_{i_m}f(\boldsymbol{x})\quad\text{for all } i_1,i_2,\cdots,i_m=1,\cdots,N.$$

In practice, it is not convenient to work with nonsymmetric Hessian matrix $\nabla^{\otimes 2}f(\boldsymbol{x})$. Luckily this is not the usual case:

THEOREM 3.3.5 ([Apo74, Theorem 12.13]). *Let Ω be an open set in $\mathbb{R}^N$ and let $f:\Omega\to\mathbb{R}$ be a differentiable function. If there exists $\boldsymbol{x}_0\in\Omega$ and $i,j\in\{1,\cdots,N\}$ such that both $\partial_i\partial_j f:\Omega\to\mathbb{R}$ and $\partial_j\partial_i f:\Omega\to\mathbb{R}$ exist and continuous at $\boldsymbol{x}_0$, then*

$$\partial_i\partial_j f(\boldsymbol{x}_0)=\partial_j\partial_i f(\boldsymbol{x}_0).$$

This theorem suggested us to consider the following space:

DEFINITION 3.3.6. Let Ω be an open set in $\mathbb{R}^N$. We denote $C^2(\Omega)$ be the collection of functions $f:\Omega\to\mathbb{R}$ such that all partial derivatives

$$\partial_i f:\Omega\to\mathbb{R},\quad \partial_i\partial_j f:\Omega\to\mathbb{R}\text{ for all } i,j=1,\cdots,N$$

exist and continuous.

In practice, we often use the following corollary of Theorem 3.3.5.

COROLLARY 3.3.7. *Let Ω be an open set in $\mathbb{R}^N$. If $f\in C^2(\Omega)$, then for each $\boldsymbol{x}\in\Omega$ the Hessian matrix $\nabla^{\otimes 2}f(\boldsymbol{x})$ is symmetric.*

3.3.3. Extreme values. We first begin with some definitions.

DEFINITION 3.3.8. Let Ω be an open set in $\mathbb{R}^N$ and let $f : \Omega \to \mathbb{R}$ be a differentiable function. If $\nabla f(\boldsymbol{x}_0)$ is a zero vector for some $\boldsymbol{x}_0 \in \Omega$, then we refer such point $\boldsymbol{x}_0$ a *critical point* or *stationary point*.

DEFINITION 3.3.9. Let S be a set (not necessarily open) in $\mathbb{R}^N$, let $f : S \to \mathbb{R}$ be a function and let $\boldsymbol{x}_0 \in S$.

(1) If there exists $\varepsilon > 0$ such that $f(\boldsymbol{x}_0) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in S \cap B_\varepsilon(\boldsymbol{x}_0)$, then we call $\boldsymbol{x}_0$ a *local minimizer* of $f : S \to \mathbb{R}$.
(2) If there exists $\varepsilon > 0$ such that $f(\boldsymbol{x}_0) \geq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in S \cap B_\varepsilon(\boldsymbol{x}_0)$, then we call $\boldsymbol{x}_0$ a *local maximizer* of $f : S \to \mathbb{R}$.
(3) If $f(\boldsymbol{x}_0) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in S$, then we call $\boldsymbol{x}_0$ a *global minimizer* of $f : S \to \mathbb{R}$.
(4) If $f(\boldsymbol{x}_0) \geq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in S$, then we call $\boldsymbol{x}_0$ a *global maximizer* of $f : S \to \mathbb{R}$.

It is easy to prove the followings:

LEMMA 3.3.10. *Let Ω be an open set in $\mathbb{R}^N$ and let $f : \Omega \to \mathbb{R}$ be a differentiable function. If f has a local maximum or local minimum at $\boldsymbol{x}_0 \in \Omega$, then $\boldsymbol{x}_0$ is a critical point.*

It is also possible to define the notion for "positive" symmetric matrices:

DEFINITION 3.3.11. Let $A \in \mathbb{R}^{N \times N}$ be a symmetric matrix.

(1) We say that A is *positive definite*, denoted as $A \succ 0$, when $\boldsymbol{\xi}^\intercal A \boldsymbol{\xi} > 0$ for all $\boldsymbol{\xi} \in \mathbb{R}^N \setminus \{\mathbf{0}\}$.
(2) We say that A is *negative definite*, denoted as $A \prec 0$, when $\boldsymbol{\xi}^\intercal A \boldsymbol{\xi} < 0$ for all $\boldsymbol{\xi} \in \mathbb{R}^N \setminus \{\mathbf{0}\}$.

REMARK 3.3.12. The above definition also make sense when $N = 1$ by simply identify $\mathbb{R} \cong \mathbb{R}^{1 \times 1}$

It is remarkable that the second derivative test also works for higher dimensional case as well:

THEOREM 3.3.13. *Let Ω be an open set in $\mathbb{R}^N$ and let $f \in C^2(\Omega)$.*

(a) *If $\nabla f(\boldsymbol{x}_0) = \mathbf{0}$ and $\nabla^{\otimes 2} f(\boldsymbol{x}_0) \succ 0$ hold for some $\boldsymbol{x}_0 \in \Omega$, then $\boldsymbol{x}_0$ is a local minimizer of $f : \Omega \to \mathbb{R}$.*
(b) *If $\nabla f(\boldsymbol{x}_0) = \mathbf{0}$ and $\nabla^{\otimes 2} f(\boldsymbol{x}_0) \prec 0$ hold for some $\boldsymbol{x}_0 \in \Omega$, then $\boldsymbol{x}_0$ is a local maximizer of $f : \Omega \to \mathbb{R}$.*

3.4. Linearization of operators and Landweber iteration: infinite dimensional case

Let X and Y are Hilbert spaces. Similar to Definition 3.2.1, it is natural to consider the differentiation of the operator $\mathscr{A} : X \to Y$ as follows:

DEFINITION 3.4.1. Let X and Y are separable Hilbert spaces and let $\mathscr{A} : X \to Y$. We say that $\mathscr{A}$ is *Fréchet differentiable* at $u_0 \in X$ if there exists a linear operator $\mathscr{L}[u_0] : X \to Y$ such that

$$\lim_{\|h\|_X \to 0} \frac{\|\mathscr{A}(u_0+h) - \mathscr{A}(u_0) - \mathscr{L}_{u_0} h\|_Y}{\|h\|_X} = 0. \tag{3.4.1}$$

For each $u_0 \in X$, the linear operator $\mathscr{L}_{u_0} : X \to Y$ is uniquely determined (if exists), see the following exercise. Therefore we often denote $(\mathrm{d}\mathscr{A})_{u_0} := \mathscr{L}_{u_0}$, called the *Fréchet derivative* of $\mathscr{A} : X \to Y$ at $u_0 \in X$.

EXERCISE 3.4.2. Show that the linear operator $\mathscr{L}(u_0) : X \to Y$ is uniquely determined by (3.4.1) (if exists).

We now consider the Landweber iteration. We consider a nonlinear functional equation

$$\mathscr{A}(u) = v. \tag{3.4.2}$$

Our goal is to solve u from the measurement y. Suppose that (3.4.2) admits a unique solution u_∞. Under suitable assumptions on $\mathscr{A}$, this solution can be computed iteratively as follows [**HNS95**]:

$$u_{k+1} = u_k + [(\mathrm{d}\mathscr{A})(u_k)]^* (v - \mathscr{A}(u_k)), \tag{3.4.3}$$

with some "good" initial guess u_0, where $[(\mathrm{d}\mathscr{A})(u_k)]^* : Y \to X$ is the Hilbert space adjoint of $(\mathrm{d}\mathscr{A})(u_k) : X \to Y$ given by

$$\left(\phi, [(\mathrm{d}\mathscr{A})(u_k)]^* \psi\right)_X = ([(\mathrm{d}\mathscr{A})(u_k)]\,\phi, \psi)_Y \quad \text{for all } \phi \in X \text{ and } \psi \in Y.$$

When $\mathscr{A}$ is linear, then (3.4.3) coincides with Landweber's original algorithm [**Lan51**].

Bibliography

[AH76] S. Agmon and L. Hörmander. Asymptotic properties of solutions of differential equations with simple characteristics. *J. Analyse Math.*, 30:1–38, 1976. MR0466902, Zbl:0335.35013, doi:10.1007/BF02786703.

[Apo74] T. M. Apostol. *Mathematical analysis*. Addison-Wesley Publishing Co., second edition, 1974. MR0344384, Zbl:0309.26002.

[AH09] K. Atkinson and W. Han. *Theoretical numerical analysis. A functional analysis framework*, volume 39 of *Texts Appl. Math.* Springer, Dordrecht, third edition, 2009. MR2511061, Zbl:1181.47078, doi:10.1007/978-1-4419-0458-4.

[Axl24] S. Axler. *Linear algebra done right*. Undergrad. Texts Math. Springer, Cham, fourth edition, 2024. MR4696768, Zbl:1531.15001, doi:10.1007/978-3-031-41026-0.

[Ber09] D. S. Bernstein. *Matrix mathematics. Theory, facts, and formulas*. Princeton University Press, Princeton, NJ, second edition, 2009. MR2513751, Zbl:1183.15001, doi:10.1515/9781400833344.

[Bre11] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011. MR2759829, Zbl:1220.46002, doi:10.1007/978-0-387-70914-7.

[Bur99] V. Burenkov. Extension theorems for Sobolev spaces. In *The Maz'ya anniversary collection, Vol. 1 (Rostock, 1998)*, volume 109 of *Oper. Theory Adv. Appl.*, pages 187–200. Birkhäuser, Basel, 1999. MR1747873, Zbl:0948.46026, doi:10.1007/978-3-0348-8675-8_13.

[CK19] D. Colton and R. Kress. *Inverse acoustic and electromagnetic scattering theory*, volume 93 of *Applied Mathematical Sciences*. Springer, Cham, fourth edition, 2019. MR3971246, Zbl:1425.35001, doi:10.1007/978-3-030-30351-8.

[EG15] L. C. Evans and R. Gariepy. *Measure theory and fine properties of functions*. Textbooks in Mathematics. CRC Press, Boca Raton, FL, revised edition, 2015. MR3409135, Zbl:1310.28001, doi:10.1201/b18333.

[FI07] H. Fujiwara and Y. Iso. Application of multiple-precision arithmetic to direct numerical computation of inverse acoustic scattering. *J. Phys.: Conf. Ser.*, 73, 2007. Paper No. 012007, 9 pages. doi:10.1088/1742-6596/73/1/012007.

[GVL13] G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins Stud. Math. Sci. Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013. MR3024913, Zbl:1268.65037.

[GS97] D. Gottlieb and C.-W. Shu. On the Gibbs phenomenon and its resolution. *SIAM Rev.*, 39(4):644–668, 1997. MR1491051, Zbl:0885.42003, doi:10.1137/S0036144596301390.

[HNS95] M. Hanke, A. Neubauer, and O. Scherzer. A convergence analysis of the Landweber iteration for nonlinear ill-posed problems. *Numer. Math.*, 72(1):21–37, 1995. MR1359706, Zbl:0840.65049, doi:10.1007/s002110050158.

[HH79] E. Hewitt and R.E. Hewitt. The Gibbs-Wilbraham phenomenon: an episode in Fourier analysis. *Arch. Hist. Exact Sci.*, 21(2):129–160, 1979. MR0555102, doi:10.1007/BF00330404.

[Hör73] L. Hörmander. Lower bounds at infinity for solutions of differential equations with constant coefficients. *Israel J. Math.*, 16:103–116, 1973. MR0340793, doi:10.1007/BF02761975.

[HS99] P.-F. Hsieh and Y. Sibuya. *Basic theory of ordinary differential equations*. Universitext. Springer-Verlag, New York, 1999. MR1697415, Zbl:0924.34001, doi:10.1007/978-1-4612-1506-6.

[Hun80] T. W. Hungerford. *Algebra*, volume 73 of *Grad. Texts in Math.* Springer-Verlag, New York-Berlin, reprint edition, 1980. MR0600654, Zbl:0442.00002.

[KRS25] H. Koch, A. Rüland, and M. Salo. On instability mechanisms for inverse problems. *Ars Inven. Anal.*, 2025. Revision series, Paper No. 1, 94 pages. MR4462475 (2021 version), Zbl:1567.35302, doi:10.15781/kvf0-ds86, `arXiv:2012.01855`.

[Kow23] P.-Z. Kow. *Complex Analysis*. National Chengchi University, Taipei, 2023. https://puzhaokow1993.github.io/homepage.

[Kow24] P.-Z. Kow. *Calculus*. National Chengchi University, Taipei, 2024. https://puzhaokow1993.github.io/homepage.

[Kow25] P.-Z. Kow. *Fourier analysis and distribution theory*. National Chengchi University & University of Jyväskylä, 2025. https://puzhaokow1993.github.io/homepage.

[KSZ25] P.-Z. Kow, M. Salo, and S. Zou. Increasing resolution and instability for linear inverse scattering problems. *J. Funct. Anal.*, 289(1), 2025. Paper No. 110923, 37 pages. MR4878830, Zbl:1561.35147, doi:10.1016/j.jfa.2025.110923, `arXiv:2404.18482`.

[KW26] P.-Z. Kow and J.-N. Wang. Frequency-dependent contraction rates for the Bayesian method to the inverse source problem. *Trans. London Math. Soc.*, 13(1), 2026. Paper No. e70031, 48 pages. MR5079486, Zbl:8208119, doi:10.1112/tlm3.70031.

[Lan51] L. Landweber. An iteration formula for fredholm integral equations of the first kind. *Amer. J. Math.*, 73:615–624, 1951. MR0043348, Zbl:0043.10602, doi:10.2307/2372313.

[Lid66] V. B. Lidskij. Perturbation theory of non-conjugate operators (Russian). *Ž. Vyčisl. Mat i Mat. Fiz.*, 6(1):52–60, 1966. MR0196930, Zbl:0166.40501, Mi:zvmmf7501. English version: U.S.S.R. Comput. Math. Math. Phys. 6(1):73–85, 1966. doi:10.1016/0041-5553(66)90033-4.

[LM72] J.-L. Lions and E. Magenes. *Non-homogeneous boundary value problems and applications. Vol. I*, volume 181 of *Die Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, New York-Heidelberg, 1972. MR0350177, Zbl:0223.35039.

[McL00] W. McLean. *Strongly elliptic systems and boundary integral equations*. Cambridge University Press, 2000. MR1742312, Zbl:0948.35001.

[Mel95] R. B. Melrose. *Geometric scattering theory*. Stanford Lectures. Cambridge University Press, Cambridge, 1995. MR1350074, Zbl:0849.58071.

[Mit18] D. Mitrea. *Distributions, partial differential equations, and harmonic analysis*. Universitext. Springer, Cham, second edition, 2018. MR3887685, Zbl:1425.46001, doi:10.1007/978-3-030-03296-8.

[MBO97] J. Moro, J. V. Burke, and M. L. Overton. On the Lidskii-Vishik-Lyusternik perturbation theory for eigenvalues of matrices with arbitrary Jordan structure. *SIAM J. Matrix Anal. Appl.*, 18(4):793–817, 1997. MR1471994, Zbl:0889.15016, doi:10.1137/S0895479895294666.

[Onn82] C. W. Onneweer. The Wilbraham-Gibbs phenomenon in Fourier analysis. *Math. Medley*, 10(2):47–54, 1982. MR0697668, Zbl:0517.01023, `https://sms.math.nus.edu.sg/volume-10-1982/`.

[PSU10] L. Päivärinta, M. Salo, and G. Uhlmann. Inverse scattering for the magnetic Schrödinger operator. *J. Funct. Anal.*, 259(7):1771–1798, 2010. MR2665410, Zbl:1254.81092, doi:10.1016/j.jfa.2010.06.002, `arXiv:0908.3977`.

[Rud87] W. Rudin. *Real and complex analysis*. McGraw-Hill Book Co., New York, third edition, 1987. MR0924157, Zbl:0925.00005.

[Rud91] W. Rudin. *Functional Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, Inc., New York, second edition, 1991. MR1157815, Zbl:0867.46001.

[Str08] W. A. Strauss. *Partial differential equations: An introduction*. John Wiley & Sons, Ltd., Chichester, second edition, 2008. MR2398759, Zbl:1160.35002.

[TB97] L. N. Trefethen and D. III Bau. *Numerical linear algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997. MR1444820, Zbl:0874.65013, doi:10.1137/1.9780898719574.

[Tre26] S. Treil. *Linear algebra done wrong*. Brown University, 2026. `https://sites.google.com/a/brown.edu/sergei-treil-homepage/linear-algebra-done-wrong`.

[WZ15] R. L. Wheeden and A. Zygmund. *Measure and integral. An introduction to real analysis*. Pure Appl. Math. CRC Press, Boca Raton, FL, second edition, 2015. MR3381284, Zbl:1326.26007.

[Yaf10] D. R. Yafaev. *Mathematical scattering theory. Analytic theory*, volume 158 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2010. MR2598115, doi:10.1090/surv/158.

[Zyg02] A. Zygmund. *Trigonometric series. Vol. I, II*. Cambridge Math. Lib. Cambridge University Press, Cambridge, third edition, 2002. MR1963498, Zbl:1084.42003, doi:10.1017/CBO9781316036587.